Mobile Networks and Applications 6, 511–524, 2001

Handoff and Optimal Channel Assignment in Wireless Networks

NOVELLA BARTOLINI
Dipartimento di Informatica, Sistemi e Produzione, Università di Roma "Tor Vergata", 00133 Roma, Italy

Abstract. In this paper, a non-preemptive prioritization scheme for access control in cellular networks is analyzed. Two kinds of users are assumed to compete for the access to the limited number of frequency channels available in each cell: the high priority users represent handoff requests, while the low priority users correspond to initial access requests originated within the same cell. Queueing of handoff requests is also considered. The research for the best access policy is carried out by means of a Markov decision model which allows us to study a very wide class of policies which includes some well known pure stationary policies, as well as randomized ones. The cutoff priority policy, consisting in reserving a certain number of channels to the high priority stream of requests, is proved to be optimal within this class while using an objective function in the form of a linear combination of some quality of service parameters, when no queueing device is considered. Numerical results confirm the optimality of the cutoff priority policy when queueing of handoff requests is allowed.

Keywords: cellular networks, handoff call, initial access request, optimal access control, cutoff priority policy, hysteresis policy, threshold priority policy, randomized policies, Markov decision process

1. Introduction

A very important result achieved in the field of wireless communication is the capability to support global roaming. The user, no longer tied to a particular fixed station, has ubiquitous access to a wide variety of services from voice communication to data exchange and elaboration, while roaming throughout the area covered by the wireless network. In order to allow frequency reuse, the geographic area covered by the network is divided into small zones called *cells*. The base station of each cell in the network has to serve two streams of requests, one coming from the cell itself (*initial access requests*), the other coming from neighbor cells (*hand-off calls*). Once a call has started, the mobile station (MS) might leave the initial cell, entering a neighbor one, while remaining connected to the network. Mobiles crossing the cell boundary cannot continue to use the same frequency channel because different channels are assigned to adjacent cells to avoid radio interferences in the shared transmission medium. Intercellular handoff is the procedure by which the user, while releasing the old frequency channel belonging to the initial cell, is provided with a new one by the base station of the destination cell. This procedure is fundamental to avoid any interruption of the initiated connections. If the destination cell does not have enough channels to support the handoff, the call is blocked. It is important to limit the probability of forced termination, because from a user's perspective, the forced termination of an ongoing call is less desirable than getting a busy signal due to the block of an initial access attempt.

The system can reduce the chances of unsuccessful handoff by assigning higher priority to handoff requests than that assigned to initial access requests. These handoff prioritizing schemes provide improved performance at the expense of a reduction in the total admitted traffic. The purpose of this paper is to propose a wide class of policies which includes some commonly studied as well as new policies. An original Markov decision model, which allows queueing of handoff requests, is proposed and the formulas for the most important Quality-of-Service parameters are also given. By means of this decision model, we search for an access control policy that gives a solution to the tradeoff between the high priority service of handoff requests and the risk of compromising the whole traffic because of an insufficient weight to initial attempts of connections. Recently various call admission control schemes have been proposed. In many of the proposed schemes, the control policy is strictly based on the number of free channels available in the cell.

- Under the *cutoff priority policy* (CPP) [1–5], priority to handoff 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.
- Under the *threshold priority policy* (TPP) [1,6], like with CPP, a handoff call is accepted as long as a channel is free. However, under TPP, an initial attempt is accepted only if the number of ongoing calls at the initial access is below a threshold value and a free channel is available. The concept of TPP was used earlier for congestion control and store-and-forward networks.
- Under the hysteresis policy (HysP) [7,8], a handoff 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 (more details will be given in section 2.1).

TPP, CPP and HysP, 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 in two steps. Linear programming methods permit to discard not stationary and randomized policies from the search for the optimum. The real optimization phase is instead realized through dynamic programming methods. The main contribution of this paper is the analytical proof of the optimality of CPP when the objective function gives higher priority to the handoff stream when queueing of requests is not allowed. Simulation results confirm the good behavior of CPP with respect to other policies in a more general condition, when a queueing device is used. 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 among access policies are either based on simulations [1,7] 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. In section 3 this model is uniformized and discretized, in order to apply methods and results of the theory of discrete-time processes, and it is proved the optimality of CPP when queueing of requests is not allowed. In section 4 the formulas of the most important quality of service parameters are illustrated. In section 5 some numerical results that confirm the analytical results achieved in the previous sections are introduced. Section 6 concludes the paper with some final remarks.

2. Analytical model

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 handoff calls and low priority for initial access requests. Arrivals are assumed to be generated according to Poisson processes with rates λ_H and λ_L for high and low priority users, respectively. Service requirements for both streams are identical and exponentially distributed. The assumption of exponentially distributed holding times has been justified by Guerin [2] and is required for the tractability of the model. Blocked initial requests are lost, while a blocked handoff call can wait in the handoff queue for a channel of the new cell by continuing to use a channel of the previous cell. The queueing scheme is briefly described as follows. No initial access request is granted a channel before the handoff requests in the queue are served. When a MS reaches the overlapping region between two adjacent cells, also called *handoff 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 handoff call and its departure from the queue. In the case of high demand for handoff, handoff calls will be denied queueing due to the limited size of the handoff queue. The queueing device has a finite number of places M_H; an upper bound to this number is the total number of channels available in the cells adjacent to the one we are considering.

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

(1) the conversation is completed (it may happen even with a queued handoff request, which thus abandons the queue);

(2) the MS goes out of cell;

(3) a waiting handoff 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.

The service requirements (channel holding time) for handoff and initial access requests could be different, but an average figure used in the model should be sufficiently accurate.

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 cell, otherwise the handoff procedure is initiated. The base station of the destination cell may decide to refuse the handoff request, to provide it with a new channel, if available, or to put it into the handoff queue while waiting for a new available channel. While in the handoff 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 handoff procedure unsuccessfully, or may also decide to conclude the call.

2.1. Markov models of access control policies

In this section, the algorithms of CPP, HysP and TPP are described and a Markov model is formulated to describe the behavior of the system under the application of these policies.

2.1.1. Cutoff priority policy

CPP reserves a fixed number of channels to the handoff stream. When the number of busy channels is less than a fixed cutoff value $T < C$, the system is considered in normal load condition and all streams of requests are served.

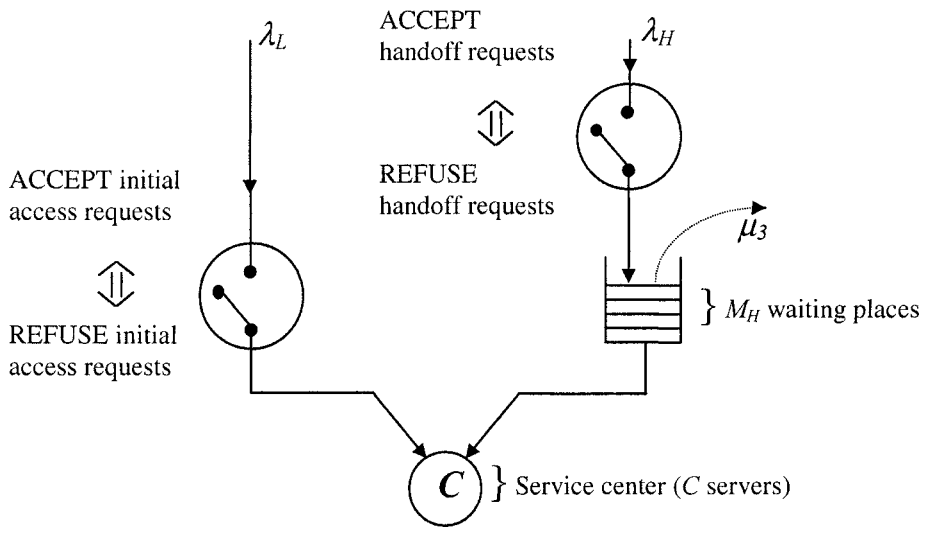

Figure 1. System configuration.

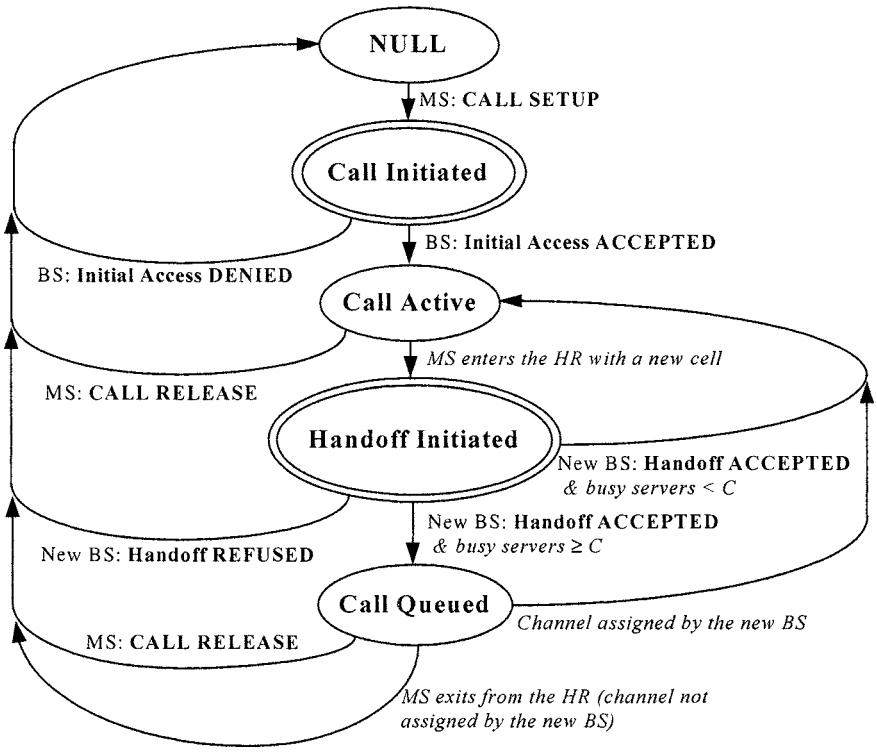

Figure 2. Call state model.

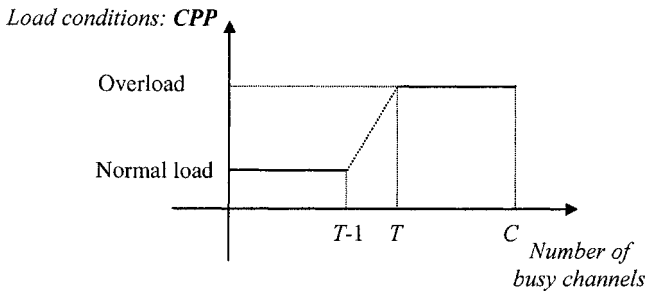

Figure 3. Load conditions (CPP).

When the workload exceeds the cutoff, the system enters the overload condition and begins to serve handoff requests only (figure 3).

The Markov chain of figure 4 represents the system under the application of CPP with cutoff value T. The states of the process are defined through an index i which corresponds to the sum of the number of busy servers S_b with the number of queued handoff requests. Notice that if $i < C$ then $i = S_b$, otherwise, if $i \geqslant C$, then $i = C + q_H$ where q_H is the occupancy level of the handoff queue. If the process is in the state k an arrival brings the system to state $(k + 1)$ with rate $(\lambda_H + \lambda_L)$ for $k < T$, and λ_H for $k \geqslant T$ because initial

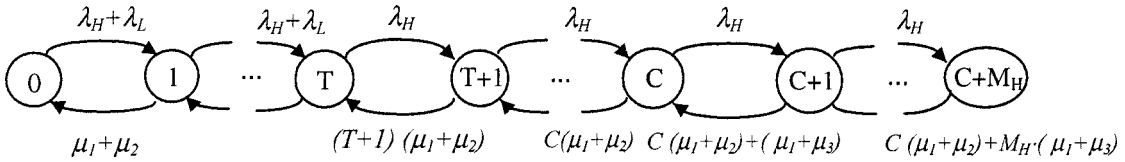

Figure 4. Transition diagram of CPP.

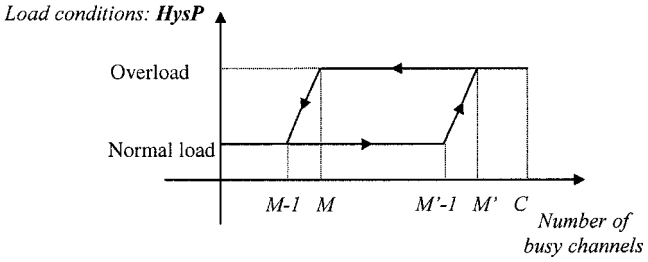

Figure 5. Load condition of HysP.

accesses are lost. The death rate in state k is $k(\mu_1 + \mu_2)$ for $0 \leqslant k \leqslant C$ and $C(\mu_1 + \mu_2) + (k - C)(\mu_1 + \mu_3)$ for $k > C$. The term $(k - C)(\mu_1 + \mu_3)$ comes from the rate that each mobile waiting in the handoff area either completes the call at rate μ_1 or goes out of the HR with rate μ_3.

2.1.2. Hysteresis policy

HysP takes into account the total occupancy level of the service center to determine the access control decision. Under HysP, if all channels are free, the load condition is considered normal and the system serves both streams of requests. The system remains in this condition until the number of busy channels reaches the threshold M'. When the system enters the overload condition, it accepts only requests which belong to the high priority stream. Once reached the threshold M', the system remains in overload condition until its occupancy level falls to another defined threshold value M, where $0 < M < M' \leqslant C$ (figure 5).

The Markovian model of HysP is shown in figure 6. The index associated to each state has the same meaning of the state index in the CPP model. However, now it is duplicated for those occupancy levels in which the system may behave differently according to its past history in correspondence to the same occupancy level.

2.1.3. Threshold priority policy

Under TPP a handoff call is accepted as long as a channel is free or a waiting place is available in the queueing device. An initial access is accepted if the number of ongoing calls at the initial access is below a threshold value $K \leqslant C$ and at least one channel is free. In the TPP Markov model of figure 7, each state is represented through a couple of indexes (i, t), with $t \leqslant i$, where the first index has the same meaning of the state index of the CPP and HysP models, and the second index t corresponds to the number of ongoing calls at their initial access. If the system is in the state (i, t), with $t < K$, the acceptance of a new request may bring the system to two different states with occupancy level $(i + 1)$. If the accepted call is an initial access, the system goes to $(i + 1, t + 1)$ with rate λ_L, while it goes to $(i + 1, t)$ with rate λ_H, if the arrival is a handoff request. If the system is in the state (i, t) and an ongoing call at its initial access disappears from the base station control, because either the call is completed or the MS goes out of cell, the system goes to the state $(i - 1, t - 1)$ with rate $t(\mu_1 + \mu_2)$ regardless of the value of i. If the occupancy level is $i < C$ and a handoff call disappears, because either the call is completed or the MS goes out of cell, the system goes to the state $(i - 1, t)$ with rate $(i - t)(\mu_1 + \mu_2)$, while if $i \geqslant C$ the handoff call disappears from the base station control, bringing the system to state $(i-1, t)$ with rate $(C-t)(\mu_1+\mu_2)+(i-C)(\mu_1+\mu_3)$. The contribution $(C-t)(\mu_1+\mu_2)$ denotes the rate that each mobile with an ongoing connection either completes the call with rate μ_1 or abandons the cell coverage area, with rate μ_2. The term $(i - C)(\mu_1 + \mu_3)$ comes from the rate that each mobile waiting in HR either completes the call with rate μ_1 or expires the residence time at rate μ_3.

2.2. A general decision model

The algorithms described in section 2.1 can be seen as particular instances of the general decision model we are introducing. In this model the definitions of the traffic parameters λ_H, λ_L, μ_1, μ_2 and μ_3 are the same used for the models of CPP, HysP and TPP. We consider a base station with C available channels and a queueing device for handoff calls with a finite number M_H of waiting places. The memoryless property of all probability distributions in a Markov process makes 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 handoff 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

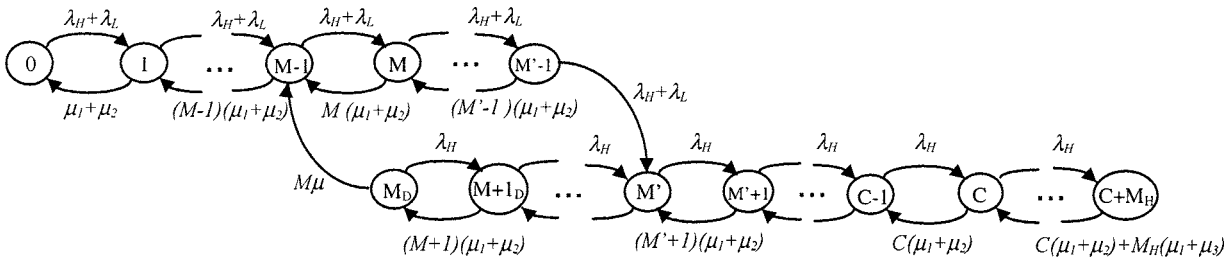

Figure 6. Transition diagram of HysP.

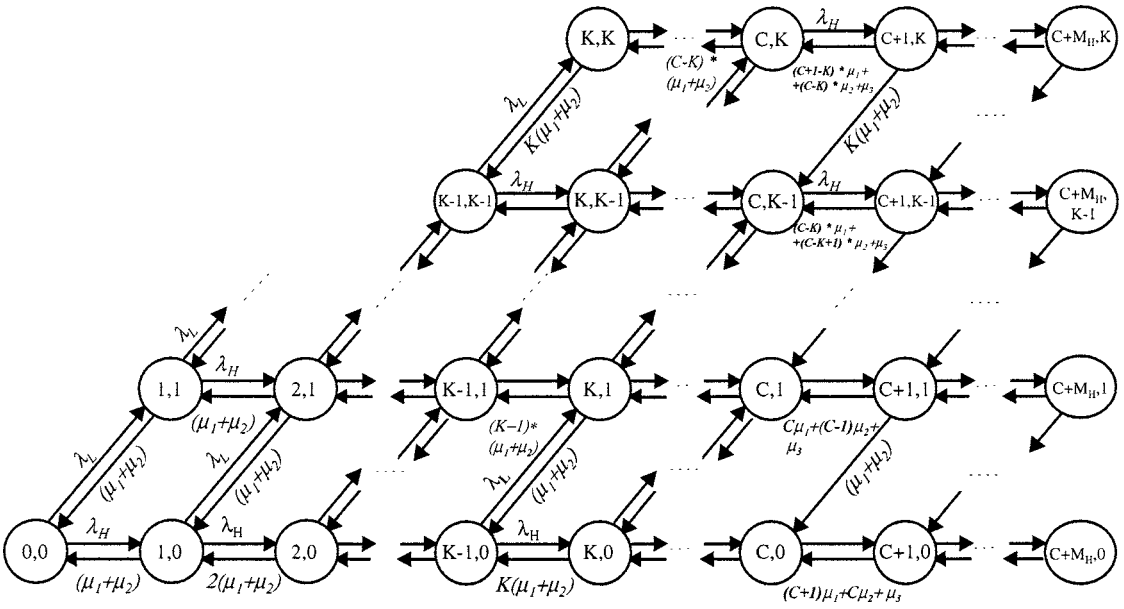

Figure 7. Transition diagram of TPP.

$n(\mathbf{s}) = i$. If $C \leqslant n(\mathbf{s}) < C + M_{\mathrm{H}}$, the set of feasible actions reduces to a_2, a_4 because we have no queueing device for initial access requests and to a_4 if $n(\mathbf{s}) = C + M_{\mathrm{H}}$.

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 $\lambda(a)\pi^{+}_{\mathbf{sq}}$, where

$$\lambda(a) = \begin{cases} \lambda_{\mathrm{H}} + \lambda_{\mathrm{L}} & \text{if } a = a_1, \\ \lambda_{\mathrm{H}} & \text{if } a = a_2, \\ \lambda_{\mathrm{H}} & \text{if } a = a_3, \\ 0 & \text{if } a = a_4. \end{cases} \tag{1}$$

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 $n(\mathbf{s}) \leqslant C$ and $\{n(\mathbf{s})\mu_1 + C\mu_2 + [n(\mathbf{s}) - C]\mu_3\}\pi^{-}_{\mathbf{sk}}$ if $n(\mathbf{s}) > C$, where $\sum_{\mathbf{k} \in Prec(\mathbf{s})} \pi^{-}_{\mathbf{sk}} = 1$.

The transition diagram of the process is represented in figure 8.

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

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

We now show how to set the parameters $\pi^{+}_{\mathbf{sq}}$, $\pi^{-}_{\mathbf{sq}}$ and the stationary state-decision associations to turn our general decision model into the models of figures 4, 6 and 7.

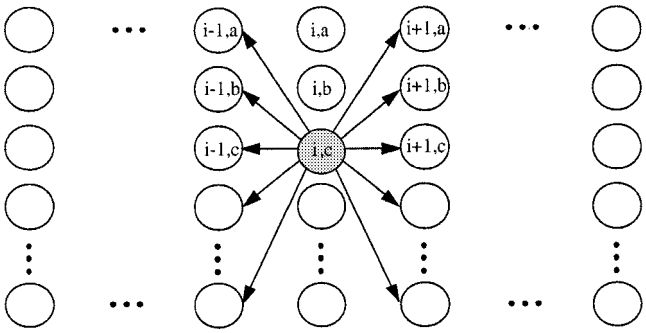

Figure 8. Possible transitions from state (i, k).

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 $T \leqslant i_s < C$ and the decision a_4 if $i_s = C$. HysP can be obtained by selecting two different tags $tag1$ and $tag2$; $\pi^+_{\mathbf{sk}} = \pi^-_{\mathbf{sk}} = 1$ if $j_s = j_k = tag1$ and i_s, lower than the threshold M', $\pi^+_{\mathbf{sk}} = \pi^-_{\mathbf{sk}} = 1$ if $j_s = j_k = tag2$ and i_s higher than the threshold M, $\pi^+_{\mathbf{sk}} = 1$ if $\mathbf{s} = (M' - 1, tag1)$ and $\mathbf{k} = (M', tag2)$, $\pi^-_{\mathbf{sk}} = 1$ if $\mathbf{s} = (M, tag2)$ and $\mathbf{k} = (M - 1, tag1)$ and else $\pi^+_{\mathbf{sk}} = \pi^-_{\mathbf{sk}} = 0$, and taking the decision a_1 for all the states $\mathbf{s}$ with $j_s = tag1$, the decision a_2 for the states with $j_s = tag2$ and $i_s < C$ and the decision a_4 in the state $(C, tag2)$ as shown in figure 6.

The TPP model with threshold K can be obtained as a particular instance of the general model by allowing the state tag t to represent the number of accepted initial accesses with an ongoing call. If $\mathbf{s} = (i, t)$ with $t \leqslant i \leqslant C$,

$$\pi^+_{\mathbf{sk}} = \begin{cases} \dfrac{\lambda_{\mathrm{L}}}{\lambda_{\mathrm{H}} + \lambda_{\mathrm{L}}} & \text{if } \mathbf{k} = (i+1, t+1) \text{ and } t < K, \\ \dfrac{\lambda_{\mathrm{H}}}{\lambda_{\mathrm{H}} + \lambda_{\mathrm{L}}} & \text{if } \mathbf{k} = (i+1, t) \text{ and } t < K, \\ 1 & \text{if } \mathbf{k} = (i+1, t) \text{ and } t \geqslant K, \\ 0 & \text{otherwise,} \end{cases} \tag{2}$$

and

$$\pi^-_{\mathbf{sk}} = \begin{cases} \dfrac{t}{i} & \text{if } \mathbf{k} = (i-1, t-1), \\ \dfrac{(i-t)}{i} & \text{if } \mathbf{k} = (i-1, t), \\ 0 & \text{otherwise,} \end{cases} \tag{3}$$

while if $C \leqslant i \leqslant C + M_{\mathrm{H}}$,

$$\pi^+_{\mathbf{sk}} = \begin{cases} 1 & \text{if } \mathbf{k} = (i+1, t) \text{ and } i < C + M_{\mathrm{H}}, \\ 0 & \text{otherwise,} \end{cases} \tag{4}$$

and

$$\pi^-_{\mathbf{sk}} = \begin{cases} \dfrac{t(\mu_1 + \mu_2)}{i\mu_1 + C\mu_2 + (i - C)\mu_3} & \text{if } \textit{condition\_A}, \\ 1 - \dfrac{t(\mu_1 + \mu_2)}{i\mu_1 + C\mu_2 + (i - C)\mu_3} & \text{if } \textit{condition\_B}, \\ 0 & \text{otherwise,} \end{cases} \tag{5}$$

where *condition_A* means that

$$\mathbf{k} = (i - 1, t - 1),$$

and *condition_B* means that

$$\mathbf{k} = (i - 1, t - 1) \quad \text{and} \quad i > t.$$

Under TPP the system takes the decision a_1 in state $\mathbf{s} = (i, t)$ if $i < C$ and $t < K$, the decision a_2 in any other state except if $i = C + M_{\mathrm{H}}$ where the system cannot choose anything but the decision a_4. It has to be noted that the generality of the defined model consists not only in the shape of its transition diagram, but also in its including both pure stationary and not stationary policies of the randomized kind. Examples of randomized policies that can be represented through this model can be obtained from the models of CPP and of HysP by allowing the system to take not deterministic decisions in correspondence to particular states.

In [3] a policy called *Limited Fractional Guard Channel* (LFG) is proved to be optimal among a restricted class of policies for the problem of minimizing the blocking probability of the initial attempt requests, with a strong constraint on the blocking probability of the handoff stream. LFG is a randomized modification of CPP, and consists in accepting all requests if the occupancy level is lower than a given cutoff value T, and in refusing initial attempts if the occupancy level is greater than T. However, if the occupancy level is exactly T, the decision is not deterministic and the system can accept requests coming from both streams with probability p, and accept only handoff requests with probability $(1 - p)$. A randomized version of HysP can also be obtained by allowing a not deterministic decision in the states $(M' - 1, tag1)$ and in $(M, tag2)$.

3. 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 it 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.

3.1. Discretization technique

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 $\phi(\mathbf{s}, a)\mathrm{e}^{\phi(\mathbf{s},a)t}$. 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 which characterizes the process is bounded by the maximum outgoing rate which is less than $C(\mu_1+\mu_2)+M_{\mathrm{H}}(\mu_1+\mu_3)+\lambda_{\mathrm{H}}+\lambda_{\mathrm{N}}$. 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 epoch 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 $\{X(t)\}$ [9,10,15].

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 $\Lambda = C(\mu_1+\mu_2)+M_{\mathrm{H}}(\mu_1+\mu_3)+\lambda_{\mathrm{H}}+\lambda_{\mathrm{H}}$.

The transition probabilities of the uniformized process are:

$$\widehat{p}^{\,a}_{\mathbf{sk}} = \begin{cases} \dfrac{\lambda(a)\pi^{+}_{\mathbf{sk}}}{\Lambda} & \text{if } c_1, \\ \dfrac{n(\mathbf{s})(\mu_1+\mu_2)\pi^{-}_{\mathbf{sk}}}{\Lambda} & \text{if } c_2, \\ \dfrac{\{n(\mathbf{s})\mu_1 + C\mu_2 + [n(\mathbf{s})-C]\mu_3\}\pi^{-}_{\mathbf{sk}}}{\Lambda} & \text{if } c_3, \\ \dfrac{\lambda - [n(\mathbf{s})(\mu_1+\mu_2)] + \lambda(a)]}{\Lambda} & \text{if } c_4, \\ \dfrac{\lambda - [n(\mathbf{s})\mu_1 + C\mu_2 + (n(\mathbf{s})-C)\mu_3 + \lambda(a)]}{\Lambda} & \text{if } c_5, \\ 0 & \text{if } c_6, \end{cases} \tag{6}$$

where

c_1 : $\mathbf{k} \in Succ(\mathbf{s})$,
c_2 : $\mathbf{k} \in Prec(\mathbf{s})$ and $0 \leqslant n(\mathbf{s}) \leqslant C$,
c_3 : $\mathbf{k} \in Prec(\mathbf{s})$ and $n(\mathbf{s}) > C$,
c_4 : $\mathbf{k} = n(\mathbf{s})$ and $n(\mathbf{s}) \leqslant C$,
c_5 : $\mathbf{k} = n(\mathbf{s})$ and $n(\mathbf{s}) > C$, and
c_6 : otherwise.

In order to give higher priority to the handoff 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 handoff 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 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. The penalty cost must be paid only in case of actual refusal, that is when the system decides to refuse a certain call which is actually already arrived, or when a MS, with a queued ongoing call, exits from the HR before being served. 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:

$$\widehat{r}(\mathbf{s}, a) = \begin{cases} 0 & \text{if } a = a_1, \\ \dfrac{L\lambda_{\mathrm{L}} + H\max\{0, [n(\mathbf{s}-C)]\mu_3\}}{\Lambda} & \text{if } a = a_2, \\ \dfrac{H\lambda_{\mathrm{H}}}{\Lambda} & \text{if } a = a_3, \\ \dfrac{L\lambda_{\mathrm{L}} + H\lambda_{\mathrm{H}} + H\max\{0, [n(\mathbf{s})-C]\mu_3\}}{\Lambda} & \text{if } a = a_4, \end{cases} \tag{7}$$

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

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

$$\lim_{T\to\infty} \frac{E\{\sum_{n=0}^{N(T)} r[X_n, u_n(X_n) \mid X_0 = i]\}}{T}. \tag{8}$$

We refer to [11] 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.

3.2. Linear programming formulation

The most important results of the theory of discrete-time Markov decision processes can be applied to the discretized

model formulated in section 3.1. In particular, observing the shape of the transition diagram of figure 8, 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 $D^*_{\mathbf{s}a}$ represents the stationary probability that, in correspondence to the state $\mathbf{s}$, the system takes the decision a. We can write

$$\begin{aligned} &D_{\mathbf{s}a} = P\{a_n = a \mid \mathbf{s}_n = \mathbf{s}\},\\ &D_{\mathbf{s}a} \geqslant 0 \quad \text{and}\\ &\sum_{a\in A_{\mathbf{s}}} D_{\mathbf{s}a} = 1, \quad \mathbf{s}\in E, \end{aligned}$$

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

$$z = \sum_{(\mathbf{s},a)\in S} D_{\mathbf{s},a}\, p_{\mathbf{s}} \widehat{r}(\mathbf{s},a), \tag{9}$$

where $p_{\mathbf{s}}$ denotes the stationary probability that the system is in the state $\mathbf{s}$, and the product $D_{\mathbf{s}a}\, p_{\mathbf{s}}$ represents the joined probability for the system to be in state $\mathbf{s}$ and contemporaneously to take the decision a. Substituting the expression of $\widehat{r}(\mathbf{s},a)$ given by equation (7) into the equation (9) the following expression for the objective function can be obtained:

$$\begin{aligned} z = {} & H\frac{\lambda_{\mathrm{H}}}{\Lambda} \sum_{[(\mathbf{s},a)\in S]\wedge(a=a_4)} p(\mathbf{s},a)\\ & + L\frac{\lambda_{\mathrm{L}}}{\Lambda} \sum_{[(\mathbf{s},a)\in S]\wedge(a=a_2|a=a_4)} p(\mathbf{s},a)\\ & + H \sum_{[(\mathbf{s},a)\in S]\wedge n(\mathbf{s})>C} \big[n(\mathbf{s}) - C\big]\frac{\mu_3}{\Lambda} p(\mathbf{s},a). \end{aligned} \tag{10}$$

Equation (10) shows how the objective function can be expressed in the form of a linear combination of some Quality-of-Service (QoS) parameters, the two blocking probabilities of the handoff and initial access streams, and the probability to see a handoff call escaping from the queueing device. The different weights of this linear combination confirm the different values of priority which has been given to the two request streams. In section 4 a more detailed presentation of these QoS parameters can be found. Furthermore, analyzing the topology of our transition diagram, we can also 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 $x_{\mathbf{s}a} \triangleq D_{\mathbf{s}a}\lambda_{\mathbf{s}}$, $\mathbf{s}a \in S$, and recalling that $D_{\mathbf{s}a} = x_{\mathbf{s}a}/p_{\mathbf{s}} = x_{\mathbf{s}a}/\sum_{\mathbf{j}\in A_{\mathbf{s}}} x_{\mathbf{j}a}$, $a \in A_{\mathbf{s}}$, the linear programming problem becomes:

$$\begin{aligned} &\text{maximize}\\ &\quad \sum_{(\mathbf{s},a)\in S} r(\mathbf{s},a) x_{\mathbf{s},a}\\ &\text{constrained to}\\ &\quad x_{\mathbf{s}a} \geqslant 0, \quad (\mathbf{s},a)\in S,\\ &\quad \sum_{(\mathbf{s})a\in S} x_{\mathbf{s}a} = 1,\\ &\quad \sum_{a\in A_{\mathbf{j}}} x_{\mathbf{j}a} = \sum_{(\mathbf{s})a\in S} p^a_{(\mathbf{s})\mathbf{j}} x_{\mathbf{s}a}, \quad \mathbf{j}\in E. \end{aligned} \tag{11}$$

Proposition 1. The linear programming problem (11) 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: $\sum_{a\in A_{\mathbf{s}}} \mathbf{x}^0_{\mathbf{s}a} > 0\ \forall \mathbf{s} \in E$. 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 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 nonredundant 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 $\sum_{a\in A_{\mathbf{s}}} x^0_{\mathbf{s}a} > 0\ \forall \mathbf{s} \in E$, we conclude that for all $\mathbf{s} \in E$ there will be exactly a decision a for which $x^0_{\mathbf{s}a} > 0$. This leads to conclude a very important result which is the existence of a pure stationary optimal policy. This result gives us the possibility to further restrict our consideration to policies for which $D_{\mathbf{s}a} \in \{0,1\}$. □

3.3. Optimization through dynamic programming methods

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 $V_N^\alpha(\mathbf{s})$. The discrete-time discount factor $a < 1$, which corresponds to the continuous-time discount coefficient $\eta > 0$, related to the not uniformized process, is

$$\alpha = \frac{\Lambda}{\eta + \Lambda}. \tag{12}$$

The optimal discounted cost function can be calculated with the following dynamic programming equation [12,14]:

$$V_K^\alpha(\mathbf{s}) = \min_{a \in A_\mathbf{s}} \left\{ \widehat{r}(\mathbf{s}, a) + \sum_{\mathbf{z} \in E} \alpha \widehat{p}_{\mathbf{sz}}^{\,a} V_{K-1}^\alpha(\mathbf{z}) \right\}. \tag{13}$$

$V_K^\alpha(\mathbf{s})$ 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. Substituting the known expression of the cost function (7), of the discount factor (12) and of the transition probabilities (6), the following equation is obtained for the total discounted cost.

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

$$\begin{aligned}
V_K^\alpha(\mathbf{s}) = {} & \frac{1}{\Lambda + \eta} \\
& \times \min \Bigg\{ \sum_{\mathbf{l} \in Prec(\mathbf{s})} n(\mathbf{s})(\mu_1 + \mu_2) \pi_{\mathbf{sl}}^- V_{K-1}^\alpha(\mathbf{l}) \\
& + \sum_{\mathbf{j} \in Succ(\mathbf{s})} (\lambda_\mathrm{L} + \lambda_\mathrm{H}) \pi_{\mathbf{sj}}^+ V_{K-1}^\alpha(\mathbf{j}) \\
& + \big[\big(C + M_\mathrm{H} - n(\mathbf{s})\big)\mu_1 \\
& + \big(C - n(\mathbf{s})\big)\mu_2 + M_\mathrm{H}\mu_3 \big] V_{k-1}^\alpha(\mathbf{s}); \\
& \lambda_\mathrm{L} L + \sum_{\mathbf{l} \in Prec(\mathbf{s})} n(\mathbf{s})(\mu_1 + \mu_2) \pi_{\mathbf{sl}}^- V_{K-1}^\alpha(\mathbf{l}) \\
& + \sum_{\mathbf{j} \in Succ(\mathbf{s})} \lambda_\mathrm{H} \pi_{\mathbf{sj}}^+ V_{K-1}^\alpha(\mathbf{j}) \\
& + \big[\big(C + M_\mathrm{H} - n(\mathbf{s})\big)\mu_1 + \big(C - n(\mathbf{s})\big)\mu_2 \\
& + M_\mathrm{H}\mu_3 + \lambda_\mathrm{L} \big] V_{K-1}^\alpha(\mathbf{s}); \\
& \lambda_\mathrm{H} H + \sum_{\mathbf{l} \in Prec(\mathbf{s})} n(\mathbf{s})(\mu_1 + \mu_2) \pi_{\mathbf{sl}}^- V_{K-1}^\alpha(\mathbf{l}) \\
& + \sum_{j \in Succ(\mathbf{s})} \lambda_L \pi_{\mathbf{sj}}^+ V_{K-1}^\alpha(\mathbf{j}) \\
& + \big[\big(C + M_\mathrm{H} - n(\mathbf{s})\big)\mu_1 \\
& + \big(C - n(\mathbf{s})\big)\mu_2 + M_\mathrm{H}\mu_3 + \lambda_\mathrm{H} \big] V_{K_1}^\alpha(\mathbf{s}); \\
& \lambda_\mathrm{H} H + \lambda_\mathrm{L} L \\
& + \sum_{\mathbf{l} \in Prec(\mathbf{s})} n(\mathbf{s})(\mu_1 + \mu_2) \pi_{\mathbf{sl}}^- V_{K-1}^\alpha(\mathbf{l}) \\
& + \big[\big(C + M_\mathrm{H} - n(\mathbf{s})\big)\mu_1 + \big(C - n(\mathbf{s})\big)\mu_2 \\
& + M_\mathrm{H}\mu_3 + \lambda_\mathrm{H} + \lambda_\mathrm{L} \big] V_{K-1}^\alpha(\mathbf{s}) \Bigg\}.
\end{aligned} \tag{14}$$

If $C \leqslant n(\mathbf{s}) \leqslant C + M_\mathrm{H}$,

$$\begin{aligned}
V_K^\alpha(\mathbf{s}) = {} & \frac{1}{\Lambda + \eta} \\
& \times \min \Bigg\{ \lambda_\mathrm{L} L + \big(n(\mathbf{s}) - C\big)\mu_3 H \\
& + \sum_{\mathbf{l} \in Prec(\mathbf{s})} \big[n(\mathbf{s})\mu_1 + C\mu_2 \\
& + \big(n(\mathbf{s}) - C\big)\mu_3 \big] \pi_{\mathbf{sl}}^- V_{K-1}^\alpha(\mathbf{l}) \\
& + \sum_{\mathbf{j} \in Succ(\mathbf{s})} \lambda_\mathrm{H} \pi_{\mathbf{sj}}^+ V_{K-1}^\alpha(\mathbf{j}) \\
& + \big[\big(C + M_\mathrm{H} - n(\mathbf{s})\big)\mu_1 \\
& + \big(M_\mathrm{H} - n(\mathbf{s})\big)\mu_3 + \lambda_\mathrm{L} \big] V_{K-1}^\alpha(\mathbf{s}); \\
& \lambda_\mathrm{L} L + \lambda_\mathrm{H} H + \big(n(\mathbf{s}) - C\big)\mu_3 H \\
& + \sum_{\mathbf{l} \in Prec(\mathbf{s})} \big[n(\mathbf{s})\mu_1 + C\mu_2 \\
& + \big(n(\mathbf{s}) - C\big)\mu_3 \big] \pi_{\mathbf{sl}}^- V_{K-1}^\alpha(\mathbf{l}) \\
& + \big[\big(C + M_\mathrm{H} - n(\mathbf{s})\big)\mu_1 + \big(M_\mathrm{H} - n(\mathbf{s})\big)\mu_3 \\
& + \lambda_\mathrm{L} + \lambda_\mathrm{H} \big] V_{K-1}^\alpha(\mathbf{s}) \Bigg\},
\end{aligned} \tag{15}$$

where we have to consider

$$V_K^\alpha(\mathbf{s}) = \begin{cases} 0 & \text{if } \mathbf{s} \in \bigcup_{n(\mathbf{z})=0} Prec(\mathbf{z}) \quad \forall K, \\ \infty & \text{if } \mathbf{s} \in \bigcup_{n(\mathbf{z})=C+M_\mathrm{H}} Succ(\mathbf{z}) \quad \forall K, \end{cases}$$

and $V_0^\alpha(\mathbf{s}) = 0$.

Proposition 2. $\forall \mathbf{s}$ and $\mathbf{z}$, such that $n(\mathbf{s}) = n(\mathbf{z})$, $V_K^\alpha(\mathbf{s}) = V_K^\alpha(\mathbf{z})$ $\forall K \in \mathcal{N}$.

Proof. By induction on the number of steps K (see appendix 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 the function $W(\cdot, \cdot)$ can be defined on the domain $\{0, 1, \ldots, C + M_\mathrm{H}] \times \mathcal{N}$, with the following property: $W^\alpha(n(\mathbf{s}), K) = V_K^\alpha(\mathbf{s}) = V_K^\alpha(\mathbf{z})$. Using the now stated property, the expression of $W^\alpha(i, K)$ can be written as follows.

If $i < C$,

$$\begin{aligned}
W^\alpha(i, K) = {} & \frac{1}{\Lambda + \eta} \\
& \times \Big\{ \big[(C + M_\mathrm{H} - i)\mu_1 \\
& + (C - i)\mu_2 + M_\mathrm{H}\mu_3 \big] W^\alpha(i, K-1) \\
& + i(\mu_1 + \mu_2) W^\alpha(i-1, K-1) + \lambda_\mathrm{L} W^\alpha(i, K-1) \\
& + \lambda_\mathrm{L} \min\big\{ L, W^\alpha(i+1, K-1) - W^\alpha(i, K-1) \big\} \\
& + \lambda_\mathrm{H} W^\alpha(i, K-1) \\
& + \lambda_\mathrm{H} \min\big\{ H, W^\alpha(i+1, K-1) - W^\alpha(i, K-1) \big\} \Big\},
\end{aligned} \tag{16}$$

while if $C \leqslant i \leqslant C + M_{\mathrm{H}}$,

$$W^{\alpha}(i,K) = \frac{1}{\Lambda+\eta} \times \Big\{ \big[(C+M_{\mathrm{H}}-i)\mu_1 + (M_{\mathrm{H}}-i)\mu_3\big] W^{\alpha}(i,K-1) + \big[i\mu_1 + C\mu_2 + (i-C)\mu_3\big] W^{\alpha}(i-1,K-1) + \lambda_{\mathrm{L}}\big[L + W^{\alpha}(i,K-1)\big] + (i-C)\mu_3 H + \lambda_{\mathrm{H}} W^{\alpha}(i,K-1) + \lambda_{\mathrm{H}} \min\big\{H, W^{\alpha}(i+1,K-1) - W^{\alpha}(i,K-1)\big\}\Big\}. \tag{17}$$

Equations (16) and (17) show that the choice whether to accept or not a given request depends on the value of the increment of the cost function:

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

Proposition 3. $W^{\alpha}(i,K)$ is not decreasing in i, thus, $0 \leqslant \Delta W_K^{\alpha}(i)$.

Proof. By induction on the number of steps K (see appendix for details). □

Proposition 4. If $M_{\mathrm{H}} = 0$, $W^{\alpha}(i,K)$ is also concave in the number of busy servers i.

Proof. By induction on the number of steps K (see appendix for details). □

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

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

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 [11], 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 $M_{\mathrm{H}} = 0$.

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

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

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

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

and

$$i_{\mathrm{H}} = \arg\min\big\{\Delta W^{\alpha}(i) > \big\}.$$

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

- initial access admitted for $i < i_{\mathrm{L}}$,
- initial access denied for $i \geqslant i_{\mathrm{L}}$ (that is the already described CPP),

while since $i_{\mathrm{H}} \geqslant i_{\mathrm{L}}$ the decision of refusing the high priority calls is obviously discarded.

Theorems of equivalence [11] 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 [13].

Theorem 6 (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 7. CPP is optimal under the average cost criterion, for the infinite horizon problem, when $M_{\mathrm{H}} = 0$.

One can also establish, by induction on the horizon length, the following result for the optimal cutoff value T [13].

Proposition 8. T is monotonically increasing in both L and λ_{L}, when $M_{\mathrm{H}} = 0$.

The proof that, if $M_{\mathrm{H}} = 0$, 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.

4. QoS parameters

From the point of view of the customer, 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 in section 3.2, or may be simply calculated through the following balance equations:

$$
\begin{aligned}
&k(\mu_1+\mu_2)P(k) = (\lambda_{\mathrm{L}}+\lambda_{\mathrm{H}})P(k-1) \\
&\quad \text{if } 1 \leqslant k \leqslant T, \\
&k(\mu_1+\mu_2)P(k) = \lambda_{\mathrm{H}}P(k-1) \\
&\quad \text{if } T < k \leqslant C, \\
&\big[C(\mu_1+\mu_2)+(k-C)(\mu_1+\mu_3)\big]P(k) = \lambda_{\mathrm{H}}P(k-1) \\
&\quad \text{if } C < k \leqslant C+M_{\mathrm{H}}, \\
&\sum_{k=0}^{C+M_{\mathrm{H}}} P(k) = 1.
\end{aligned}
$$

The solution is

$$
P(k) = \begin{cases}
\left(\dfrac{\lambda_{\mathrm{L}}+\lambda_{\mathrm{H}}}{\mu_1+\mu_2}\right)^k \dfrac{1}{k!}P(0) & \text{if } 1 \leqslant k \leqslant T, \\
\dfrac{\lambda_{\mathrm{H}}^{k-T}(\lambda_{\mathrm{H}}+\lambda_{\mathrm{L}})^T}{(\mu_1+\mu_2)^k k!}P(0) & \text{if } T < k \leqslant C, \\
\dfrac{\lambda_{\mathrm{H}}^{k-T}(\lambda_{\mathrm{H}}+\lambda_{\mathrm{L}})^T P(0)}{(\mu_1+\mu_2)^C \prod_{i=1}^{T}[C(\mu_1+\mu_2)+i(\mu_1+\mu_3)]C!} & \text{if } C < k \leqslant C+M_{\mathrm{H}},
\end{cases}
$$

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

Once calculated the steady state probabilities, the most important QoS parameters can be computed. For example, the probability B_{L} that an initial access is blocked is

$$B_{\mathrm{L}} = \sum_{i=T}^{C+M_{\mathrm{H}}} P(i), \tag{20}$$

while the probability B_{H} that a handoff call is blocked is equal to the probability that no place is available in the queueing device, that is, $B_{\mathrm{H}} - P(C+M_{\mathrm{H}})$. The mean queue length is

$$L_{\mathrm{H}} = \sum_{i=C+1}^{C+M_{\mathrm{H}}} (i-C)P(i). \tag{21}$$

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

$$B_{\mathrm{H\,out}} = \mu_3 \frac{L_{\mathrm{H}}}{(1-B_{\mathrm{H}})\lambda_{\mathrm{H}}}, \tag{22}$$

while the term $E_{\mathrm{H}} = \mu_3 L_{\mathrm{H}}$ is the unconditioned probability to see a handoff call escaping from the system before obtaining a channel. The terms B_{L}, B_{H} and E_{H} appear in linear combination in the formulation of the objective function for the average cost optimization problem, given by equation (10).

5. Numerical results

In the previous paragraphs the optimality of CPP is proved when the system has no queueing capability. CPP represents a tradeoff 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 handoff queueing is allowed.

In figures 9 and 10, the behavior of CPP with variable cutoff value T is compared with that of HysP, with $M' = C$ and $M = T$, for a system with $C = 10$ available channels, $\lambda_{\mathrm{L}} = 50$, $\lambda_{\mathrm{H}} = 75$, $\mu_1 = 1$, $\mu_2 = 2$, $\mu_3 = 10$ and $M_{\mathrm{H}} = 5$.

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

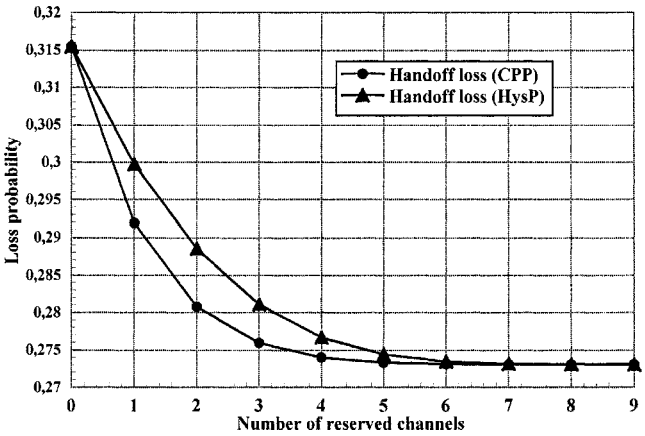

Figure 9. Handoff loss probability.

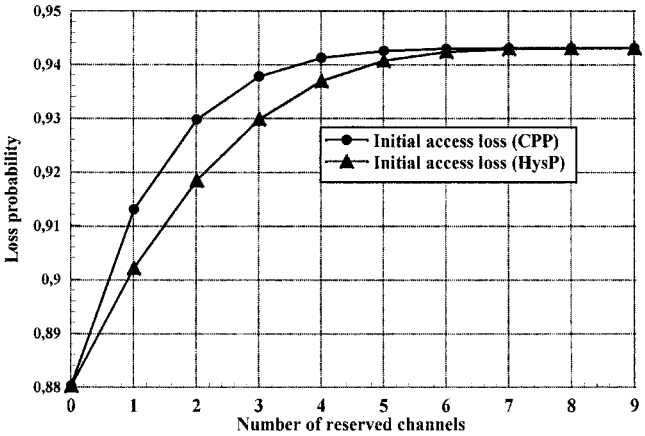

Figure 10. Initial access loss probability.

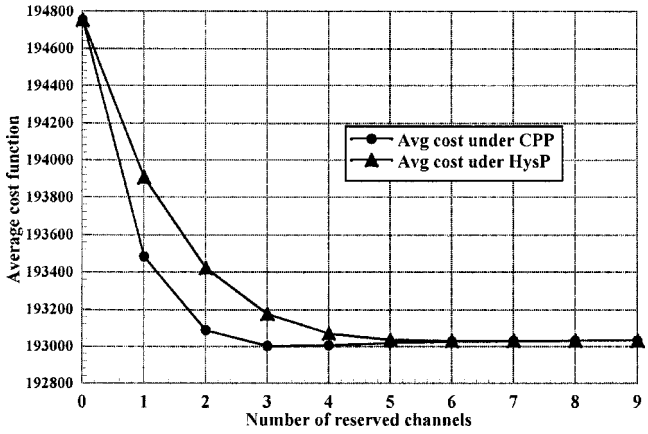

Figure 11. Average cost function.

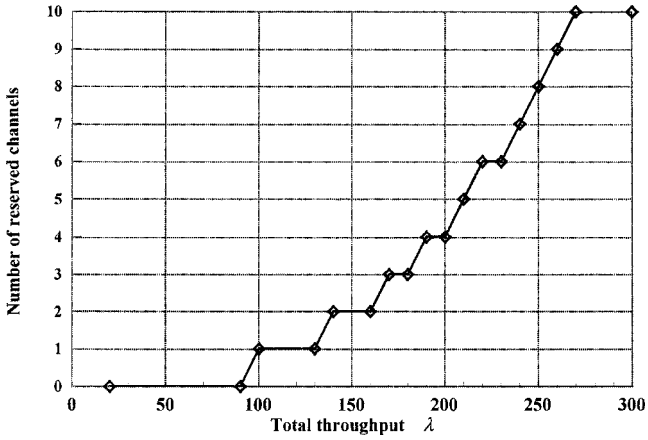

Figure 12. Variation of the number of reserved channels with total throughput.

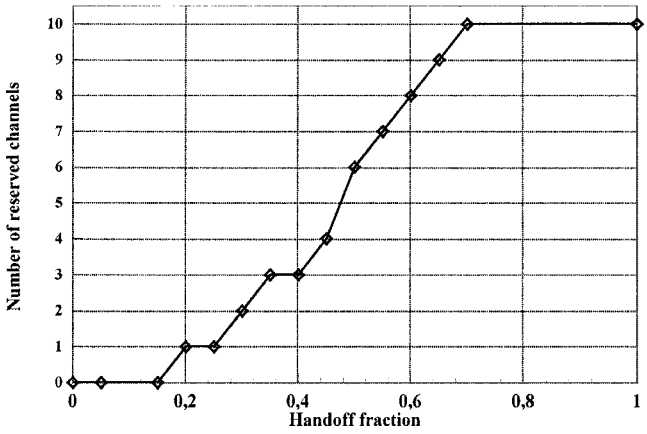

Figure 13. Variation of the number of reserved channels with the handoff fraction γ.

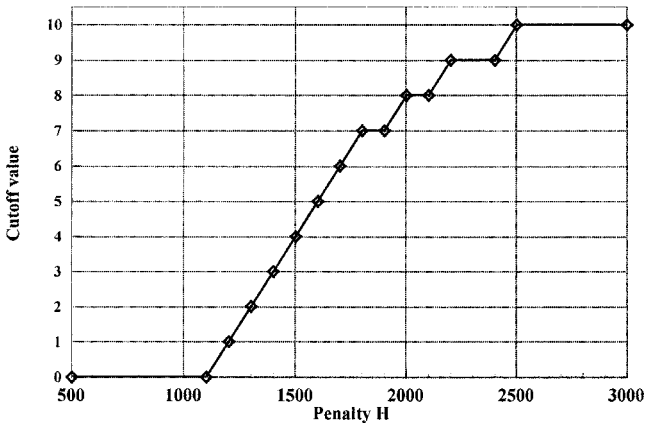

Figure 14. Variation of the cutoff value with the penalty H.

For the above described system, the average cost function with $H = 3000$ and $L = 1800$ is optimal under the application of CPP with cutoff value $T = 7$, that is with 3 reserved channels. It can be seen that, increasing the number of reserved 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 $T = 7$, a tradeoff solution is achieved.

Figure 11 shows the trend of the average cost function with the number of reserved channels under HysP and CPP. It can be seen that under HysP, the average cost is higher than with the application of CPP, even if the choice of the cutoff value is not optimal.

Figure 12 shows how the optimal cutoff value decreases or the number of reserved channels increases with the total arriving throughput $\lambda = \lambda_L + \lambda_H$ for a system with $C = 10$ available channels, $\gamma = \lambda_H/\lambda = 0.4$, $\mu_1 = 4$, $\mu_2 = 2$, $\mu_3 = 10$ and $M_H = 5$. This means that the system becomes more selective in accepting potentially unprofitable customers, if the arrival rate of requests grows.

The same behavior of the number of reserved channels is obtained if the number of handoff calls grows with respect to the number of initial attempts of connection (figure 13). In figure 13 the trend of the optimal number of reserved channels in function of the handoff fraction $\gamma = \lambda_H/\lambda$ is showed for a system with $C = 10$ available channels, total throughput $\lambda = 125$, $\mu_1 = 1$, $\mu_2 = 2$, $\mu_3 = 10$ and $M_H = 5$.

Figure 14 confirms the property stated in proposition 8, that the optimal cutoff value is increasing with the available channels, $\lambda_L = 50$, $\lambda_H = 75$, $\mu_1 = 4$, $\mu_2 = 2$, penalty value H. It represents a system with $C = 10$ $\mu_3 = 10$ and $M_H = 5$.

6. Conclusions

In this paper an optimization method of channel assignment is proposed. The model is based on a cost function which gives higher priority to handoff requests than to originating calls. The cost function has been studied through a decision Markov model characterized by a great generality. This model is able to represent both not stationary policies and randomized fractional policies. Moreover, thanks to the particular shape of its transition diagram, it allows us to study policies of great interest such as the threshold policy and algorithms with one or more cycles of hysteresis. The optimization analysis is carried out in two steps. Linear programming methods permit to discard not stationary and ran-

domized policies from the search for the optimum. The real optimization phase is instead realized through dynamic programming methods. We analytically prove 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. We also show that numerical results confirm the optimality of CPP even when handoff queueing is allowed.

Appendix

Proposition 2. $\forall(\mathbf{s})$ and $\mathbf{z}$ such that $n(\mathbf{s}) = n(\mathbf{z})$, $\forall_k^\alpha(\mathbf{s}) = V_k^\alpha(\mathbf{z})\ \forall k \in \mathcal{N}$.

Proof. By induction on k. If $k = 0$, $V_k^\alpha(\mathbf{s}) = 0\ \forall \mathbf{s} \in E$, so the proposition is trivially true, also $\forall \mathbf{s} \in En(\mathbf{s})$. Suppose that the proposition is true for $k-1$, i.e. $V_{k-1}^\alpha(\mathbf{s}) = V_{k-1}^\alpha(\mathbf{z})$ $\forall \mathbf{s}$ and $\mathbf{j}$ with $n(\mathbf{s}) = n(\mathbf{j})$.

Refer to $V_{k-1}^\alpha(Prec(\mathbf{s}))$ and to $V_{k-1}^\alpha(Succ(\mathbf{s}))$ using the inductive hypothesis, as to the value of the function $V_{k-1}^\alpha(\mathbf{x})$ in correspondence to any state $\mathbf{x}$ in the sets $Prec(\mathbf{s})$ and $Succ(\mathbf{s})$, respectively. Substituting the expressions $\sum_{\mathbf{q}\in Succ(\mathbf{s})} \pi_{\mathbf{sq}}^+ = 1$ and $\sum_{\mathbf{j}\in Prec(\mathbf{s})} \pi_{\mathbf{sj}}^- = 1$, $V_k^\alpha(\mathbf{s})$ can be written as follows:

$$\begin{aligned} V_k^\alpha(\mathbf{s}) = \frac{1}{\Lambda+\eta}\Big\{&\big[\big(C + M_{\mathrm{H}} - n(\mathbf{s})\big)\mu_1 \\ &+ \big(C - n(\mathbf{s})\big)\mu_2 + M_{\mathrm{H}}\mu_3\big]V_{k-1}^\alpha(\mathbf{s}) \\ &+ n(\mathbf{s})(\mu_1+\mu_2)V_{k-1}^\alpha\big(Prec(\mathbf{s})\big) \\ &+ (\lambda_{\mathrm{L}} + \lambda_{\mathrm{H}})V_{k-1}^\alpha(\mathbf{s}) \\ &+ \lambda_{\mathrm{L}} \min\big\{L,\, V_{k-1}^\alpha\big(Succ(\mathbf{s})\big) - V_{k-1}^\alpha(\mathbf{s})\big\} \\ &+ \lambda_{\mathrm{H}} \min\big\{H,\, V_{k-1}^\alpha\big(Succ(\mathbf{s})\big) - V_{k-1}^\alpha(\mathbf{s})\big\}\Big\}; \end{aligned}$$

while if $C \leqslant i \leqslant C + M_{\mathrm{H}}$,

$$\begin{aligned} V_k^\alpha(\mathbf{s}) = \frac{1}{\Lambda+\eta}\Big\{&\big[\big(C + M_{\mathrm{H}} - n(\mathbf{s})\big)\mu_1 \\ &+ \big(M_{\mathrm{H}} - n(\mathbf{s})\big)\mu_3 + (\lambda_{\mathrm{L}} + \lambda_{\mathrm{H}})\big]V_{k-1}^\alpha(\mathbf{s}) \\ &+ \lambda_{\mathrm{L}}L + \big[n(\mathbf{s})\mu_1 + C\mu_2 \\ &+ \big(n(\mathbf{s}) - C\big)\mu_3\big]V_{k-1}^\alpha\big(Prec(\mathbf{s})\big) \\ &+ \big(n(\mathbf{s}) - C\big)\mu_3 H \\ &+ \lambda_{\mathrm{H}} \min\big\{H,\, V_{k-1}^\alpha\big(Succ(\mathbf{s})\big) - V_{k-1}^\alpha(\mathbf{s})\big\}\Big\}. \end{aligned}$$

Using the expressions here obtained for $V_k^\alpha(\mathbf{s})$ and for $V_k^\alpha(\mathbf{z})$ for any $\mathbf{s}$ and $\mathbf{z}$ such that $n(\mathbf{s}) = n(\mathbf{z})$ we obtain $V_k^\alpha(\mathbf{s}) = V_k^\alpha(\mathbf{z})$. Thus, the proposition is still valid for k. □

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

Proof. Since in this paper we use this proposition only with $M_{\mathrm{H}} = 0$, in order to reduce the length of this proof we refer to this case, while the complete proof can easily be obtained as an extension of this one. If we use the notation $\mu \triangleq \mu_1 + \mu_2$, when $M_{\mathrm{H}} = 0$, $W^\alpha(i, K)$ can be written as follows:

$$\begin{aligned} W^\alpha(i, K) = \frac{1}{\Lambda+\eta}\Big\{&\lambda_{\mathrm{H}} \min\big\{W^\alpha(i+1, K-1); \\ &\qquad H + W^\alpha(i, K-1)\big\} \\ &+ \lambda_{\mathrm{L}} \min\big\{W^\alpha(i+1, K-1); \\ &\qquad L + W^\alpha(i, K-1)\big\} \\ &+ i\mu W^\alpha(i-1, K-1) \\ &+ (C-i)\mu W^\alpha(i, K-1)\Big\}. \end{aligned}$$

Omitting the common factor $\lambda_{\mathrm{H}}/(\Lambda+\eta)$ the first term in the expression of $\Delta W_k^\alpha(i)$ is

$$\begin{aligned} &\min\big\{W^\alpha(i+1, k-1);\, H + W^\alpha(i, k-1)\big\} \\ &\quad - \min\big\{W^\alpha(i, k-1);\, H + W^\alpha(i-1, k-1)\big\} \\ &\quad = W^\alpha(i, k-1) \\ &\qquad + \min\big\{W^\alpha(i+1, k-1) - W^\alpha(i, k-1);\, H\big\} \\ &\qquad - W^\alpha(i-1, k-1) \\ &\qquad - \min\big\{W^\alpha(i, k-1) - W^\alpha(i-1, k-1);\, H\big\} \geqslant \end{aligned}$$

using the inductive hypothesis that $W^\alpha(i, k-1)$ is monotonously not decreasing

$$\begin{aligned} &\geqslant W^\alpha(i, k-1) - W^\alpha(i-1, k-1) \\ &\qquad - \min\big\{W^\alpha(i, k-1) - W^\alpha(i-1, k-1);\, H\big\} \\ &\geqslant 0, \end{aligned}$$

again because $W^\alpha(i, k-1)$ is monotonous. The same argumentation can be made for the second term of $\Delta W_K^\alpha(i)$. Consider the last two terms in $\Delta W_k^\alpha(i)$ (omitting the common factor $\mu/(\Lambda+\eta)$):

$$\begin{aligned} &(C-i)W^\alpha(i, k-1) + iW^\alpha(i-1, k-1) \\ &\qquad - \big(C - (i-1)\big)W^\alpha(i-1, k-1) \\ &\qquad - (i-1)W^\alpha(i-2, k-1) \\ &\quad = (C-i)W^\alpha(i, k-1) - (C-i)W^\alpha(i-1, k-1) \\ &\qquad + (i-1)W^\alpha(i-1, k-1) - (i-1)W^\alpha(i-2, k-1) \\ &\quad \geqslant (i-1)W^\alpha(i-1, k-1) - (i-1)W^\alpha(i-2, k-1) \\ &\quad \geqslant iW^\alpha(i-1, k-1) - (i-1)W^\alpha(i-2, k-1) \\ &\quad \geqslant 0, \end{aligned}$$

where the inequalities derive from the monotonous behavior of $W^\alpha(i, k-1)$. Therefore, $\Delta W_k^\alpha(i) = W^\alpha(i, k) - W^\alpha(i-1, k) \geqslant 0$, so it follows that $W^\alpha(i, k)$ is monotonously not decreasing in i for all k. □

Proposition 4. If $M_{\mathrm{H}} = 0$, $W^\alpha(i, K)$ is also concave in the number of busy servers i.

Proof. The property of concavity of the function $W^\alpha(i, k)$ can also be written in the following way: $\Delta W_k^\alpha(i) < \Delta W^\alpha(i+1)$ for $0 \leqslant i < C-1$. Again we make use of induction on k. The basis step of the induction is for $k = 0$. $W^\alpha(i, 0) = 0$ for $1 \leqslant i \leqslant C$, while if $M_{\mathrm{H}} = 0$

then $W^{\alpha}(C+1,0)=\infty$, therefore, it satisfies the convexity property to be proved. Assume $W^{\alpha}(i,k-1)$ is convex and notice how this implies the same property for $W^{\alpha}(i,k)$. The first two terms in the expression of $W^{\alpha}(i,k)$ can be trivially proved concave by induction on k. From now on the notation $\delta(i)$ will be used to represent the sum of the third and fourth term in the expression of $W^{\alpha}(i,k)-W^{\alpha}(i-1,k)$. Thence

$$\begin{aligned}\delta(i) &= iW^{\alpha}(i-1,k-1)+(C-i)W^{\alpha}(i,k-1)\\ &\quad -(i-1)W^{\alpha}(i-2,k-1)\\ &\quad -\big(C-(i-1)\big)W^{\alpha}(i-1,k-1)\\ &= (C-i)W^{\alpha}(i,k-1)-(i-1)W^{\alpha}(i-2,k-1)\\ &\quad +(2i-1-C)W^{\alpha}(i-1,k-1).\end{aligned}$$

In order to prove the concavity of the last two terms in $W^{\alpha}(i,k)$, the inequality $\delta(i+1)-\delta(i)\geqslant 0$ will be proved:

$$\begin{aligned}&\delta(i+1)-\delta(i)\\ &\quad = (C-i-1)W^{\alpha}(i+1,k-1)\\ &\qquad -iW^{\alpha}(i-1,k-1)+(2i+1-C)W^{\alpha}(i,k-1)\\ &\qquad -(C-i)W^{\alpha}(i,k-1)+(i-1)W^{\alpha}(i-2,k-1)\\ &\qquad -(2i-1-C)W^{\alpha}(i-1,k-1)\\ &\quad = (C-i-1)W^{\alpha}(i+1,k-1)\\ &\qquad +(3i+1-2C)W^{\alpha}(i,k-1)\\ &\qquad +(C+1-3i)W^{\alpha}(i-1,k-1)\\ &\qquad +(i-1)W^{\alpha}(i-2,k-1)\\ &\quad = (C-i-1)W^{\alpha}(i+1,k-1)\\ &\qquad -2(C-i-1)W^{\alpha}(i,k-1)\\ &\qquad +(C-i-1)W^{\alpha}(i-1,k-1)\\ &\qquad +(i-1)W^{\alpha}(i,k-1)\\ &\qquad -2(i-1)W^{\alpha}(i-1,k-1)\\ &\qquad +(i-1)W^{\alpha}(i-2,k-1)\\ &\quad \geqslant 0\end{aligned}$$

for the concavity of $W^{\alpha}(i,k-1)$. Therefore, also the sum of the third and fourth term of the expression defining $W^{\alpha}(i,k)$ is a concave function. If follows that $W^{\alpha}(i,k)$ is itself a convex function, because it can be expressed as a sum of concave functions. □

Acknowledgements

Many thanks are due to V. Grassi for his indispensable contribution, and to M. Colajanni and S. Tucci for their careful review of the manuscript.

References

[1] B. Gavish and S. Shridar, Threshold priority policy for channels assignment in cellular networks, IEEE Transactions on Computers 46(3) (March 1997).
[2] R. Guerin, Queueing-blocking system with two arrival streams and guard channels, IEEE Transactions on Communications 36(2) (February 1988).
[3] R. Ramjeee, R. Nagarajan and D. Towsley, On optimal call admission control in cellular networks, in: *Proceedings of IEEE INFOCOM'96 Conference* (March 1996).
[4] D. Hong and S.S. Rappaport, Priority oriented channel access for cellular systems serving vehicular and portable radio telephones, IEEE Proceedings 136(5) (October 1989).
[5] Q. Zeng, K. Mukumoto and A. Fukuda, Performance analysis of mobile cellular radio systems with priority reservation handoff procedures, IEEE 0-7803-1927-3/94.
[6] S.S. Lam and M. Reiser, Congestion control of store-and-forward networks by input buffer limits – An analysis, IEEE Transactions on Communications 27 (January 1979) 127–133.
[7] D. McMillan, Delay analysis of a cellular mobile priority queueing system, IEEE/ACM Transactions on Networking 3(3) (June 1995).
[8] R.G. Scherer, On a cutoff priority queueing system with hysteresis and unlimited waiting room, Computer Networks and ISDN Systems 20 (1990).
[9] A.T. Bharucha-Reid, *Elements of the Theory of Markov Processes and Their Applications* (McGraw-Hill, 1960).
[10] J. Keilson, *Markov Chain Models. Rarity and Exponentiality* (Springer-Verlag, New York, 1979).
[11] D.P. Heyman and M.J. Sobel, *Stochastic Models in Operations Research*, Vols. 1 and 2 (McGraw-Hill, 1984).
[12] K. Hastings, *Introduction to the Mathematics of Operations Research* (Dekker, 1989).
[13] C. Derman, *Finite State Markovian Decision Processes* (Academic Press, 1970).
[14] H.C. Tijms, *Stochastic Modelling and Analysis. A Computational Approach* (Wiley, 1986).
[15] S. Ross, *Applied Probability Models with Optimization Applications* (Holden-Day, 1970).

N. Bartolini graduated with honors in 1997 and received her Ph.D. in computer engineering at the University of Rome, Italy. In 1997 she worked as a researcher at Fondazione Ugo Bordoni, and in 1999 she has been a visiting scholar at the CATSS center, University of Texas at Dallas, USA. She is now working as a researcher at the University of Rome, Italy. Her research interests lie in the area of computer networks and parallel computing.
E-mail: novella@uniroma2.it