Optimal robust policies for bandwidth allocation and admission control in wireless networks

Vicent Pla*† Jorma Virtamo ‡ Jorge Martínez-Bauset†

July 16, 2008

Abstract

We consider joint strategies of bandwidth allocation and admission control for elastic users competing for a downlink data channel in a cellular network. For the sake of robustness and generality of the results we focus on the set of strategies whose performance does not depend on the detailed traffic characteristics beyond the traffic intensity. Performance is studied at the flow level in a dynamic setting were users come and go over time. A number of user classes are considered, which are characterized by their achievable bit rate, guaranteed throughput, arrival rate and mean flow size. We aim at characterizing a strategy which is optimal in the sense of having the lowest blocking probability. Such characterization provides some interesting insights into the optimal policy and its evolution as the system load increases. In some cases computing the optimal policy can be exceedingly complex except for lightly loaded systems. For those cases we propose a computationally feasible suboptimal policy that achieves a good relative performance. Finally, we show that in scenarios of practical interest, the loads of interest lie inside the region where the optimal policy can be efficiently computed.

1 Introduction

In spite of the enormous variety of traffic flows in the multiservice Internet and the corresponding difficulty in its characterization, the abstraction of flows pertaining to two broad classes (*elastic* and *streaming*) has proven to be simple yet practical for traffic engineering purposes [1]. Future wireless cellular networks are expected to provide not only voice service but also data services —mainly Internet access traffic— thus carrying the same traffic type than wired access networks, although not necessarily in the same proportion.

Streaming traffic corresponds to real-time audio or video and hence has rather stringent requirements on packet delay and jitter. Elastic traffic in turn can adapt to available bandwidth up

*Part of this work was done while visiting the Networking Laboratory at the Helsinki University of Technology.

†This author is with the Department of Communications, Universidad Politécnica de Valencia (UPV), ETSIT Camí de Vera s/n, 46022 Valencia, Spain.

‡This author is with the Networking Laboratory, Helsinki University of Technology, P.O. Box 3000, FI02015 TKK, Finland.

to a minimum. If available bandwidth drops below that minimum flows may abandon before completing the transaction leading to an unwanted waste of resources [2, 3]. Therefore, despite the adaptability of elastic flows it is advisable to enforce some type of admission control (AC) in order to guarantee a minimum bandwidth per flow and ensure an efficient use of resources.

In the quest of an equivalent of the Erlang formula for the Internet, Bonald and Proutiére, following the seminal work of Kelly [4], invented the concept of *balanced fairness* [5, 6], which is defined as the most efficient way to share network capacity among different flows so that (under some assumptions on the generation of flows) the resulting system is insensitive, i.e., the stationary distribution of the underlying queuing network does not depend on the detailed distribution of service requirements beyond their means. A further extension of the results on insensitive queuing network is presented in [7]. In [8] balanced fairness allocation is applied to several network types and its performance is compared with that of *max-min fairness* and *proportional fairness*. Insensitive scheduling from a classical teletraffic perspective is presented in [9].

Despite the packetized nature of data in 3G networks, performance analysis of wireless data networks at the flow level has attracted a considerable interest during the last years [10–18]. The vast majority of these studies rely to some extent on queuing models based on Processor Sharing (PS) and Discriminatory Processor Sharing (DPS) queues; see, for instance, the seminal paper [19] and [20–23].

In [10, 11] it is shown that the *Proportional Fair* scheduler, which is commonly deployed in downlink data channels, at the flow level can be satisfactorily evaluated by means of a Processor Sharing (PS) based queuing model. The resulting model has the advantage of being insensitive. Applications of such model to a single cell with no moving users can be found in [10–12]. Bonald et al. [13] study the flow level performance of mobility induced rate variations. While the exact analysis turns out to be exceedingly difficult, there are two limiting regimes, corresponding to infinitely fast and slow movement, that provide an upper and a lower performance bound. The models for the analysis of both limiting regimes are, again, insensitive. In [15] Liu and Virtamo extended the limiting regimes analysis to a finite population model. A multicell scenario where neighboring base stations interfere each other having a negative impact on the downlink data rate is examined in [14]. The multicell setting is modeled as network of multiclass PS queues. The exact analysis of the model proves to be unfeasible but after some simplifications the authors obtain a bound and approximations that turn out to be insensitive. The mutual interference between cells impacts negatively on the downlink data rates. Such a negative effect, however, can be lessened by means of intercell coordination and load balancing [24, 25]. Borst et al. [16] looked at a more complex scenario with multiple cells and intra and intercell mobility. Their study focuses on capacity (defined from a stability perspective) and the main conclusion is that mobility increases capacity even if the system works in a decentralized manner with no intercell scheduling coordination.

On the other hand, DPS based models are introduced to model the unequal capacity sharing arising in some situations. Unequal sharing may arise fundamentally due to TCP rate control or service differentiation enforced by packet schedulers. Under the assumption of time-scale separation Kessel et al. [26] show that the performance of a class with relatively slow dynamics

is insensitive to the DPS weights. Moreover, in normal load conditions, realizing service differentiation through a packet scheduler that operates in a DPS manner is rather ineffective [27]. Wu et al. [17] study insensitivity in a DPS system from a practical perspective. They conclude that if DPS weights are moderately asymmetric insensitivity can be assumed for practical purposes, whereas for highly skewed weights a significant degree of sensitivity to traffic details arises. Previously, Kessel et al. [26] have studied a DPS system in an asymptotic regime where the flow dynamic of the various classes occur on separate time scales and obtain insensitive performance measures. Furthermore, Bonald and Roberts [27] conclude that sensitivity of DPS is "slight for reasonable values of the weights, say $w_1/w_2 < 10$" (for a two-class case).

Integration of streaming and elastic traffic has also received some attention during the last years as the model for an integrated services network. Such scenario, though, results in a significantly more complex mathematical model and for which the theoretical support of insensitivity is lost. Exact analyses [28, 29] rely on the assumption of exponentially distributed size (duration) of elastic (streaming) flows and even then numerical methods and approximations can be considered useful [30]. In [31, 32] a more general distribution for streaming flows is allowed but then only an approximate analysis is feasible. In [18, 33–35] insensitive models are obtained by analyzing the system using a time-scale separation approximation: elastic flows see streaming flows evolve infinitely slowly (*quasi-stationary* regime) or fast (*fluid* regime), and vice versa. This approximation proves to be accurate for the time-scales of some practical scenarios and otherwise they yield upper and lower performance bounds. Finally, based on the results of [36] insensitive performance bounds are obtained in [37] for an integrated traffic scenario. These bounds, however, tend to be looser than those obtained using the quasi-stationary and fluid approximations [37].

In this paper we focus on elastic traffic carried over a wireless cellular network. Specifically, we address jointly the problem of bandwidth allocation (BWA) and admission control (AC). Our main goal is to characterize the optimal joint BWA-AC scheme —in the sense of having the lowest loss probability— among those that are insensitive to the distribution of the flow size, i.e., only depends on their mean values.

Our work inherits some of the ideas of a series of papers dealing with insensitive dynamic load balancing [38–40]. In all of them —and also here— the simplicity and robustness of insensitivity is an essential condition for the optimal policy that is sought. In [38] it is assumed that capacities are allocated according to balanced fairness and then the optimal routing policy is sought constrained to being balanced in order to preserve insensitivity. The optimality objective is to minimize the overall blocking probability or the maximum per-class blocking probability. A simple characterization of the optimal routing policy is obtained for the single-class traffic and also for the more general multi-class traffic. However, in the latter case the policy optimization is restricted to the set of *decentralized policies*, i.e., strategies where the routing decision for a class-k customer does not depend on the number of customers of other classes.

For the purpose of obtaining the insensitivity property it is not necessary that capacity allocation and routing are balanced separately. Actually, it was already noted in [5] that a better performance can be achieved if capacity allocation and routing are jointly balanced, which is a weaker requirement than separate balancing. This approach is followed in [39] and [40]. However

the performance advantage of joint balancing comes at the cost of higher complexity. In [39] the authors obtain and characterize the optimal joint allocation-routing policy in a single-class traffic scenario. To best of our knowledge no similar results exists for the multiclass traffic scenario. A multiclass traffic scenario with global information policies is studied in [40] but the aim is not a characterization of the optimal policy. An approach based on the theory of Markov Decision Processes (MDP) is used to formulate the optimization as a Linear Programming (LP) problem (see, for instance, [41]). The LP formulation allows more flexibility and hence a greater variety of problems, objective functions and constraints can be considered [40].

In this paper we address the more general problem of seeking a characterization of the optimal global policy in a multiclass traffic setting, but in turn we restrict ourselves to a simpler network topology — which fits a cellular scenario — than in all the aforementioned studies of this kind [38–40]: all traffic classes have a single feasible route into which they are allocated or otherwise blocked, and there is a single constraining resource, i.e the time-slotted wireless channel. We employ the same optimization approach as in [40] based on a MDP-LP formulation.

The remainder of the paper is structured as follows. Section 2 describes the model of the system and the optimization problem is formally stated. In Section 3 we present the characterization of the optimal policy and introduce several suboptimal policies. Numerical experiments illustrating the contents of this section and exploring scenarios of practical interest are shown in Section 4. Concluding remarks are given in Section 5.

2 Model Description and Problem Formulation

We model traffic at the flow level and ignore interactions at the packet level (scheduling, buffer management, TCP congestion control,...). The flow content is then viewed as a fluid which is transmitted as a continuous stream with rate changes occurring only at flow arrivals and departures. This is a widely used traffic model in the literature (see for instance [8] and its references). We focus on a single base station with a downlink channel allocated to data users. We consider that the downlink resources are time-shared among active users, i.e., flows. Transmission is done in a one-by-one fashion using time slots with duration much shorter than flow duration or flow inter-arrival times so that the validity of the flow level abstraction is maintained [12, 17].

Flows are classified into K different classes. Class-i flows arrive as a Poisson process with rate λ_i, their mean flow size (expressed in bits) is $1/\mu_i$ and require a minimum bit rate φ_i. Let C_i denote the feasible bit rate for class-i flows, i.e., the bit rate that is achieved during a slot assigned to one of such flows. Moreover, let us introduce $\rho_i = \lambda_i/(C_i\mu_i)$, $\rho = \sum_{i=1}^{K} \rho_i$ and $\lambda = \sum_{i=1}^{K} \lambda_i$. Classes can be defined by the feasible rates —which correspond to different locations within the cell [12, 42]—, flow types —having different mean flow sizes or minimum rate requirements—, or both.

Let $\boldsymbol{x} = (x_1, \ldots, x_K)$ denote the system state, where x_i is the number of active flows of the i-th class. The BWA-AC policy is described by $\phi_i(\boldsymbol{x})$ and $p_i(\boldsymbol{x})$: when the system is at state $\boldsymbol{x}$ arriving class-i flows are accepted with probability $p_i(\boldsymbol{x})$ and the ensemble of class-i flows is served with bit rate $\phi_i(\boldsymbol{x}) = C_i\tau_i(\boldsymbol{x})$, i.e., a fraction $\tau_i(\boldsymbol{x})$ of time-slots is assigned to class i;

note that

$$\sum_{i=1}^{K} \tau_i(\boldsymbol{x}) = 1. \tag{1}$$

Within a class, bandwidth is fairly shared among the flows. The bit rate seen by a class-i flow is $\phi_i(\boldsymbol{x})/x_i$ (if $x_i > 0$).

Subject to the minimum bit rate per flow requirements $\phi_i(\boldsymbol{x}) \geq x_i\varphi_i$ and the total capacity constraint (1), it is easily seen that the set of feasible states is

$$\mathcal{S} := \left\{ \boldsymbol{x} : \quad \sum_{i=1}^{K} \frac{x_i}{\beta_i} \leq 1 \right\}$$

where $\beta_i = C_i/\varphi_i$.

Denote by $\pi(\boldsymbol{x})$ the stationary state probabilities and by P_b the aggregate blocking probability. We want to find the insensitive BWA-AC policy that minimizes P_b while fulfilling the minimum rate requirements. More formally the problem can be stated as:

Find: $\phi_i(\boldsymbol{x})$ and $p_i(\boldsymbol{x})$ for $i = 1, \ldots, K$ and $\boldsymbol{x} \in \mathcal{S}$ that
Minimize: P_b
Subject to:
1. insensitivity with respect to the service time distribution;
2. minimum rate requirements: $\phi_i(\boldsymbol{x}) \geq x_i\varphi_i$, $\forall i$.

We formulate the optimization problem above as an MDP-LP. The state of the MDP consists of the system state $\boldsymbol{x}$, the admission decision $\boldsymbol{d}$ vector, and the bandwidth allocation b variable. The admission vector $\boldsymbol{d} = (d_1, \ldots, d_K) \in \{0,1\}^K$ codes which traffic classes will have their newly arriving flows accepted: if $d_i = 1$ new class-i flows are accepted, and rejected otherwise. The bandwidth allocation variable codes to which class the transmission capacity is allocated: $b = i$ means that transmission capacity is allocated to class i. Let $\pi(\boldsymbol{x}, \boldsymbol{d}, b)$ denote the MDP state probability, in other words, the probability that the system is at state $\boldsymbol{x}$, accepts only those new flows belonging to classes in the set $\{i : d_i = 1\}$, and the transmission capacity is allocated to ongoing class-b flows.

The system state probabilities $\pi(\boldsymbol{x})$, blocking probability P_b and policy parameters $\phi_i(\boldsymbol{x})$ and $p_i(\boldsymbol{x})$ can be expressed in terms of $\pi(\boldsymbol{x}, \boldsymbol{d}, b)$ as

$$\pi(\boldsymbol{x}) = \sum_{\boldsymbol{d} \in \{0,1\}^K} \sum_{b=1}^{K} \pi(\boldsymbol{x}, \boldsymbol{d}, b)$$

$$P_b = \sum_{i=1}^{K} \left(\frac{\lambda_i}{\lambda} \sum_{\boldsymbol{d}: d_i = 0} \sum_{\boldsymbol{x} \in \mathcal{S}} \sum_{b=1}^{K} \pi(\boldsymbol{x}, \boldsymbol{d}, b) \right)$$

$$\tau_i(\boldsymbol{x}) = \frac{\sum_{\boldsymbol{d}} \pi(\boldsymbol{x}, \boldsymbol{d}, i)}{\sum_{\boldsymbol{d}} \sum_b \pi(\boldsymbol{x}, \boldsymbol{d}, b)} \tag{2}$$

$$p_i(\boldsymbol{x}) = \frac{\sum_{\boldsymbol{d}: d_i = 1} \sum_b \pi(\boldsymbol{x}, \boldsymbol{d}, b)}{\sum_{\boldsymbol{d}} \sum_b \pi(\boldsymbol{x}, \boldsymbol{d}, b)} \tag{3}$$

The LP problem can now be written as follows

$$\min_{\pi(\boldsymbol{x},\boldsymbol{d},b)} \sum_i \left(\frac{\lambda_i}{\lambda} \sum_{\boldsymbol{d}:d_i=0} \sum_{\boldsymbol{x}} \sum_b \pi(\boldsymbol{x},\boldsymbol{d},b) \right) \tag{4}$$

s.t.

$$\pi(\boldsymbol{x},\boldsymbol{d},b) \geq 0 \qquad \forall \boldsymbol{x} \in \mathcal{S},\ \boldsymbol{d} \in \{0,1\}^K,\ b = 1,\ldots,K \tag{5}$$

$$\sum_{\boldsymbol{x}} \sum_{\boldsymbol{d}} \sum_b \pi(\boldsymbol{x},\boldsymbol{d},b) = 1 \tag{6}$$

$$\beta_i \sum_{\boldsymbol{d}} \pi(\boldsymbol{x},\boldsymbol{d},i) \geq x_i \sum_{\boldsymbol{d}} \sum_b \pi(\boldsymbol{x},\boldsymbol{d},b) \qquad \forall \boldsymbol{x} \in \mathcal{S},\ i = 1,\ldots,K \tag{7}$$

$$\rho_i \sum_{\boldsymbol{d}:d_i=1} \sum_b \pi(\boldsymbol{x}-\boldsymbol{e}_i,\boldsymbol{d},b) = \sum_{\boldsymbol{d}} \pi(\boldsymbol{x},\boldsymbol{d},i) \qquad \forall i = 1,\ldots,K, \quad \boldsymbol{x} \in \mathcal{S} : x_i > 0 \tag{8}$$

where $\boldsymbol{e}_i$ is the vector with a 1 in the i-th position and 0's elsewhere.

Equations (5) and (6) refer to probabilistic nature of $\pi(\boldsymbol{x},\boldsymbol{d},b)$, Eq. (7) represents the minimum rate requirement and Eq. (8) is the detailed balance condition. In the ordinary LP formulation of MDP theory, global balance conditions appear as linear constraints on the decision variables. In order to retain insensitivity, we impose stricter detailed balance conditions as constraints [40], which is equivalent to the balance condition [7, 40]

$$\frac{\psi_i(\boldsymbol{x}-\boldsymbol{e}_j)}{\psi_i(\boldsymbol{x})} = \frac{\psi_j(\boldsymbol{x}-\boldsymbol{e}_i)}{\psi_j(\boldsymbol{x})} \qquad i,j = 1,\ldots,K, \quad \boldsymbol{x} \in \mathcal{S} : x_i, x_j > 0$$

where

$$\psi_i(\boldsymbol{x}) = \rho_i \frac{p_i(\boldsymbol{x}-\boldsymbol{e}_i)}{\tau_i(\boldsymbol{x})}. \tag{9}$$

Note that the radio channel capacity constraint is implicitly included in the definition of $\pi(\boldsymbol{x},\boldsymbol{d},b)$. From (2) it readily follows that $\sum_{i=1}^K \tau_i(\boldsymbol{x}) = 1$, actually it also holds for $\boldsymbol{x} = (0,\ldots,0)$, although it has no physical sense.

3 Policy Characterization

For a given configuration, the LP formulated in the previous section can be numerically solved to obtain the values of $\pi(\boldsymbol{x},\boldsymbol{d},b)$ and by applying Eqs. (2)–(3), the BWA-AC parameters are obtained.

Our goal is to find a characterization for the optimal insensitive joint BWA-AC policy. In our quest we followed an inductive and rather experimental process: from the observation of particular solutions in rather simple scenarios we extracted and generalized the underlying characteristics of the optimal policy, which have been subsequently tested against a variety of more complex settings. In this section we describe the general form of the optimal insensitive joint BWA-AC policy. Since in some instances the general form may turn out to be excessively complicated for practical purposes, we also describe a simpler suboptimal form.

Denote by $\hat{\boldsymbol{\rho}} = (\hat{\rho}_1,\ldots,\hat{\rho}_K) = \rho^{-1}\boldsymbol{\rho}$ the traffic share across classes. Let us denote by letter ω with a subscript a BWA-AC policy, i.e., a set of values for $\{\tau_i(\boldsymbol{x}), p_i(\boldsymbol{x}) : \boldsymbol{x} \in \mathcal{S}, i = 1,\ldots,K\}$.

Let $\omega(\rho)$ represent the optimal policy as a function of the system load ρ. It has been found that, for a given traffic share $\hat{\boldsymbol{\rho}}$, there exists a finite number of thresholds for ρ

$$0 = \rho^{(0)} < \rho^{(1)} < \rho^{(1)} < \cdots < \rho^{(m)} = \infty$$

such that $\omega(\rho) = \omega_j$ for $\rho \in [\rho^{(j-1)}, \rho^{(j)}]$. Therefore, $\omega(\rho)$ (and thus $\tau_i(\boldsymbol{x})$ and $p_i(\boldsymbol{x})$) is a piecewise constant function of ρ. Moreover, as it will be seen below, the policy settings ω_j do not depend on the load conditions (ρ_i), they only depend on the values of C_i and φ_i. On the contrary the load thresholds $\rho^{(j)}$ do depend on the load conditions. In Section 3.1 we precisely specify the form of ω_1 and in Section 3.2 we describe the transformations that ω_1 undergoes as ρ increases giving rise to $\omega_2, \cdots, \omega_m$.

On the other hand, if the policy specification is available then the values of $\psi_i(\boldsymbol{x})$ can be easily computed (see Eq. (9)) and from these the system state probabilities easily follow as

$$\pi(\boldsymbol{x}) = \pi(\mathbf{0})\psi_{i_1}(\boldsymbol{e}_{i_1})\psi_{i_2}(\boldsymbol{e}_{i_1} + \boldsymbol{e}_{i_2}) \cdots \psi_{i_n}(\boldsymbol{x}). \tag{10}$$

Where $\mathbf{0} = (0, \ldots, 0)$, $n = \sum_{i=1}^{K} x_i$ is the number of flows in the state $\boldsymbol{x}$, and

$$\langle \mathbf{0}, \boldsymbol{e}_{i_1}, \boldsymbol{e}_{i_1} + \boldsymbol{e}_{i_2}, \ldots, \boldsymbol{e}_{i_1} + \cdots + \boldsymbol{e}_{i_n} = \boldsymbol{x} \rangle$$

is any direct path from state $\mathbf{0}$ to state $\boldsymbol{x}$. Note that the simple product-form above for the system state probabilities is another consequence of the detailed balance condition. The blocking probability can be then computed by

$$P_b = \sum_{i=1}^{K} \left(\frac{\lambda_i}{\lambda} \sum_{\boldsymbol{x} \in \mathcal{S}} (1 - p_i(\boldsymbol{x}))\, \pi(\boldsymbol{x}) \right).$$

In principle having the piecewise characterization of $\omega(\rho)$ does not save having to solve the LP since the load thresholds $\rho^{(j)}$ remain to be known, but it can be circumvented and the exact policy to apply can be determined as

$$\omega(\rho) = \arg\min_{\omega_j} P_b(\omega_j, \rho). \tag{11}$$

This approach can be especially convenient if working with suboptimal policies (see Section 3.3 below).

3.1 The *First Policy* ω_1

For a sufficiently low load the optimal policy is ω_1, i.e., $\omega(\rho) = \omega_1$ if $0 \leq \rho < \rho^{(1)}$. Here we describe the observed principles that characterize ω_1 and by applying those principles we obtain a method for computing the policy parameters.

Throughout this section we assume, without loss of generality, that $C_1 \geq C_2 \geq \cdots \geq C_K$. Define $\kappa(\boldsymbol{x}) = \max\{i : x_i > 0\}$.

At any state $\boldsymbol{x}$, the observed principles can be stated as:

1. The constraining resource (i.e., transmission time) is shared equally among flows unless this allocation fails to satisfy some class' rate requirement. In the latter case the throttled classes are allocated their minimum required rate ($x_i\varphi_i$) and the remaining capacity is equally shared among the flows of non-throttled classes. Hence for $\boldsymbol{x} \in \mathcal{S}$, $\tau_i(\boldsymbol{x})$ can be computed for classes in descending order as follows

$$\tau_K(\boldsymbol{x}) = \frac{x_K}{\min\left(\sum_{i=1}^{K} x_i, \beta_K\right)},$$

$$\tau_i(\boldsymbol{x}) = \max\left(\frac{x_i}{\beta_i}, \frac{x_i}{\sum_{j=1}^{i} x_j}\left(1 - \sum_{j=i+1}^{K} \tau_j(\boldsymbol{x})\right)\right).$$

2. Let i be a class such that all classes with lower feasible rates have no active flows, then, if accepting one more flow of this class leads to a feasible state, new flows are accepted with probability 1. In a more formal manner, for $i \geq \kappa(\boldsymbol{x})$ if $\boldsymbol{x} + \boldsymbol{e}_i \in \mathcal{S}$ then $p_i(\boldsymbol{x}) = 1$. Obviously, whatever the traffic class i, if $\boldsymbol{x} + \boldsymbol{e}_i \notin \mathcal{S}$, $p_i(\boldsymbol{x}) = 0$.

The first principle precisely specifies the BWA whereas the second one gives the AC probabilities only in some cases. Those cases not covered can be worked out by applying the fact that, since the system satisfies the detailed balance equations, it is reversible and, in particular, satisfies the *Kolmogorov's criterion* (see, for instance, [43, Chapter 10]). The method for doing so is detailed in what follows.

Through all discussion we assume that $\boldsymbol{x}, \boldsymbol{x} + \boldsymbol{e}_i \in \mathcal{S}$ and $i < \kappa(\boldsymbol{x})$, otherwise the value of $p_i(\boldsymbol{x})$ is already known: $p_i(\boldsymbol{x}) = 0$ if $\boldsymbol{x} + \boldsymbol{e}_i \notin \mathcal{S}$, and $p_i(\boldsymbol{x}) = 1$ if $\boldsymbol{x} + \boldsymbol{e}_i \in \mathcal{S}$ but $i \geq \kappa(\boldsymbol{x})$.

Define $\xi_i(\boldsymbol{x}) = \psi_i(\boldsymbol{x} + \boldsymbol{e}_i)/\rho_i = p_i(\boldsymbol{x})/\tau_i(\boldsymbol{x} + \boldsymbol{e}_i)$ and by applying the Kolmogorov's criterion to the cycle

$$\begin{array}{ccccccc}
\boldsymbol{x} - x_{\kappa(\boldsymbol{x})}\boldsymbol{e}_{\kappa(\boldsymbol{x})} + \boldsymbol{e}_i & \longleftarrow & \cdots & \longleftarrow & \boldsymbol{x} - \boldsymbol{e}_{\kappa(\boldsymbol{x})} + \boldsymbol{e}_i & \longleftarrow & \boldsymbol{x} + \boldsymbol{e}_i \\
\downarrow & & & & & & \uparrow \\
\boldsymbol{x} - x_{\kappa(\boldsymbol{x})}\boldsymbol{e}_{\kappa(\boldsymbol{x})} & \longrightarrow & \cdots & \longrightarrow & \boldsymbol{x} - \boldsymbol{e}_{\kappa(\boldsymbol{x})} & \longrightarrow & \boldsymbol{x}
\end{array}$$

we obtain

$$\xi_i(\boldsymbol{x}) = \xi_i(\boldsymbol{x} - x_{\kappa(\boldsymbol{x})}\boldsymbol{e}_{\kappa(\boldsymbol{x})}) \prod_{j=1}^{x_{\kappa(\boldsymbol{x})}} \frac{\xi_{\kappa(\boldsymbol{x})}\left(\boldsymbol{x} - j\boldsymbol{e}_{\kappa(\boldsymbol{x})} + \boldsymbol{e}_i\right)}{\xi_{\kappa(\boldsymbol{x})}\left(\boldsymbol{x} - j\boldsymbol{e}_{\kappa(\boldsymbol{x})}\right)}. \tag{12}$$

Since

$$\xi_{\kappa(\boldsymbol{x})}(\boldsymbol{x}) = \frac{1}{\tau_{\kappa(\boldsymbol{x})}(\boldsymbol{x} + \boldsymbol{e}_{\kappa(\boldsymbol{x})})} = \frac{1}{\max\left(\frac{x_{\kappa(\boldsymbol{x})}+1}{\beta_{\kappa(\boldsymbol{x})}}, \frac{x_{\kappa(\boldsymbol{x})}+1}{1+\sum_{m=1}^{K} x_m}\right)} = \frac{\min\left(\beta_{\kappa(\boldsymbol{x})}, 1 + \sum_{m=1}^{\kappa(\boldsymbol{x})} x_m\right)}{x_{\kappa(\boldsymbol{x})} + 1}$$

Eq. (12) becomes

$$\xi_i(\boldsymbol{x}) = \xi_i(\boldsymbol{x} - x_{\kappa(\boldsymbol{x})}\boldsymbol{e}_{\kappa(\boldsymbol{x})}) \frac{\min\left(\beta_{\kappa(\boldsymbol{x})}, 1 + \sum_{m=1}^{\kappa(\boldsymbol{x})} x_m\right)}{\min\left(\beta_{\kappa(\boldsymbol{x})}, 1 + \sum_{m=1}^{\kappa(\boldsymbol{x})-1} x_m\right)} \tag{13}$$

and by applying (13) recursively it follows that

$$\xi_i(\boldsymbol{x}) = \frac{\min\left(\beta_i, 1 + \sum_{m=1}^{i} x_m\right)}{x_i + 1} \prod_{j=i+1}^{\kappa(\boldsymbol{x})} \frac{\min\left(\beta_j, 1 + \sum_{m=1}^{j} x_m\right)}{\min\left(\beta_j, 1 + \sum_{m=1}^{j-1} x_m\right)}. \tag{14}$$

Finally, $p_i(\boldsymbol{x})$ can be computed as $p_i(\boldsymbol{x}) = \tau_i(\boldsymbol{x} + \boldsymbol{e}_i)\xi_i(\boldsymbol{x})$.

3.2 Policy Evolution

The first policy ω_1 can be considered, in a way, biased towards less-favored traffic classes in terms of feasible rate, which makes sense given the low load situation. As load increases, however, situation changes and optimal policy orientation shifts towards efficiency, limiting the access to the system of the more resource-consuming traffic classes. More precisely, we say that class i consumes more resources than class j if $\mu_i C_i < \mu_j C_j$. In other words, resource consumption of a class is measured as the flow mean sojourn time in the system considering there are no other active flows. Throughout this section it is assumed without loss of generality that $\mu_1 C_1 \geq \mu_2 C_2 \geq \cdots \geq \mu_K C_K$, i.e., traffic classes are sorted in ascending order according to resource consumption. It has been found that starting with ω_1, a series of transformations T_i, which penalize the most resource consuming classes and favor the least resource consuming ones, are successively applied as load increases

$$\omega_1 \xrightarrow{T_1} \omega_2 \xrightarrow{T_2} \cdots \xrightarrow{T_{m-1}} \omega_m.$$

The last policy ω_m is at the opposite side of ω_1, i.e., all resources are reserved for class 1, which is the least resource consuming class:

$$p_i(\boldsymbol{x}) = \begin{cases} 1 & \text{if } i = 1 \text{ and } \boldsymbol{x} + \boldsymbol{e}_1 \in \mathcal{S} \\ 0 & \text{otherwise} \end{cases}, \qquad \tau_i(\boldsymbol{x}) = \begin{cases} 1 & \text{if } i = 1 \text{ and } \boldsymbol{x} \in \mathcal{S} \\ 0 & \text{otherwise} \end{cases} \qquad i = 1, \ldots, K.$$

Next we describe the type of policy transformations T_i. Before doing so we need to introduce some additional notation. Let us define

$$\mathcal{R} := \left\{ \boldsymbol{x} = (0, x_2, \ldots, x_K) : \boldsymbol{x} \in \mathcal{S} \right\}$$

and introduce the order relation $\succ$ defined as follows: we say that $\boldsymbol{x} \succ \boldsymbol{y}$ if $x_j > y_j$ and $x_i = y_i$ for $i = j+1, \ldots, K$. Now consider that $\boldsymbol{y}_1, \boldsymbol{y}_2, \cdots, \boldsymbol{y}_{|\mathcal{R}|}$ is a sorted list of all the elements in $\mathcal{R}$, i.e.

$$\boldsymbol{y}_1 \succ \boldsymbol{y}_2 \succ \cdots \succ \boldsymbol{y}_{|\mathcal{R}|}.$$

Finally, for each $\boldsymbol{y}_i$ we define the set

$$\Delta_i := \left\{ \boldsymbol{x} = x_1 \boldsymbol{e}_1 + \boldsymbol{y}_i, \boldsymbol{x} \in \mathcal{S} \right\}.$$

Let us start with the policy ω_1. The set of feasible states for ω_1 is $\mathcal{S}_1 = \mathcal{S}$. The transformation T_1 will affect one or more states in Δ_1 in one of the following ways:

A: the bandwidth allocation to class K is set to its minimum, i.e., $\tau_K(\boldsymbol{x}) = \beta_K x_K$, and the released capacity is shared by the remaining classes

B: some admission probabilities of the least resource consuming class are set to 1 (if all the admission probabilities in Δ_1 of the least resource consuming class are already 1, the second least resource consuming class is considered and so on)

C: states in Δ_1 are made unfeasible by rejecting those flow arrival that would lead the system to a state in Δ_1, i.e., if $\boldsymbol{x} + \boldsymbol{e}_K \in \Delta_1$ then $p_K(\boldsymbol{x}) = 0$.

Note that since the detailed balanced condition has to be satisfied, changes applied to a state may also affect other neighboring states, which might be outside Δ_1. Before the type-C transformation is applied, none or several transformations of types A or B can be applied. Obviously after the type-C transformation no more transformations can target states in Δ_1 since these are not feasible anymore. After the type-C transformation, the set of feasible states becomes $\mathcal{S}_2 = \mathcal{S}_1 \setminus \Delta_1$, then none or several type-A,B transformations are applied to states in Δ_2 followed by the type-C transformation which clips the feasible state space to $\mathcal{S}_3 = \mathcal{S}_2 \setminus \Delta_2$,... This process is repeated until the feasible state space becomes $\mathcal{S}_M := \{(x_1, 0, \ldots, 0) : x_1 \leq \beta_1\}$, which corresponds to the last policy ω_m. Note that after some type-C transformations (more specifically after $\lfloor \beta_K \rfloor$ of them) no class-K flows are let into the system and class $(K-1)$ will then play the role of the most resource consuming class. Again, when class $(K-1)$ has been completely removed it will be substituted by class $(K-2)$, and so forth until only class 1 is let into the system.

3.3 Suboptimal Policies

The description of policy transformations given in previous section is not sufficient to obtain $\omega_{j\geq 2}$ from ω_1. That will require, at least, knowing how many and in which order type-A,B transformations occur between two type-C transformations. Unfortunately, in our experiments we could not observe any general and simple rule for the occurrence of transformations of types A and B. Considering the special case of equal minimum bit rates did not lead to any progress on this direction. Besides, even if we were able to determine the exact sequence of transformations the load values at which transformations occur ($\rho^{(j)}$) will remain unknown. As mentioned above, not knowing the thresholds $\rho^{(j)}$ can be circumvented by the approach of Eq. (11) but this requires computing P_b for each policy ω_j. Observe that if the system state probabilities have been computed under policy ω_j and T_j is of type C, the system state probabilities under policy ω_{j+1} can be recomputed by simply renormalizing.

Motivated by the aforementioned reasons we propose a set of suboptimal policies which are defined as follows: $\hat{\omega}_1 \equiv \omega_1$ and $\hat{\omega}_j \equiv \omega_1$(restricted to $\mathcal{S}_j$) for $j = 2, \ldots, M$, which is equivalent to say $\hat{\omega}_1 \equiv \omega_1$ and then only the type-C transformations are applied. By definition $\hat{\omega}_1 = \omega_1$, and also $\hat{\omega}_M = \omega_m$ but in general $\hat{\omega}_j$ is not necessarily included in $\{\omega_1, \ldots, \omega_m\}$ since, as noted previously, a transformation of type B or C may also affect states outside its target set of states Δ_l. For a given value of ρ the suboptimal policy can be obtained using the approach of Eq. (11), $\hat{\omega}(\rho) = \arg\min_{\hat{\omega}_j} P_b(\hat{\omega}_j, \rho)$.

A numerical evaluation example showing the good performance achieved by the suboptimal policies presented in this Section is given in Section 4.

3.4 Restriction to a *fair sharing state space*

In our policy optimization problem formulated as a linear program in (4)–(8) the optimization is performed over a state space $\mathcal{S}$, which is the widest possible: for states $\boldsymbol{x}$ lying outside $\mathcal{S}$ no bandwidth allocation satisfying the minimum rate requirements is possible; for states inside

$\mathcal{S}$ the minimum rate requirements are enforced through the linear constraint (7). It turns out that if the system load is sufficiently low ($0 < \rho < \rho^{(1)}$, i.e., ω_1 is optimal) the optimal policy shares the time in a fair fashion ($\tau_i(\boldsymbol{x}) = \frac{x_i}{\sum_{j=1}^{K} x_j}$) whenever it is possible, i.e., if the guaranteed minimum bit rates are satisfied: $C_i\tau_i(\boldsymbol{x}) \geq \varphi_i x_i,\ \forall i$. Denote by $\mathcal{S}_{FS}$ the set of those states, which is formally defined below. Moreover, new flow arrivals taking the system to a state where the fair sharing is possible are accepted with probability 1. These evidences suggest that if the optimization is constrained to $\mathcal{S}_{FS}$ the resulting policy may have a simpler form than in the general case. Of course, that simplicity will come at the price of poorer performance.

In this section we provide a characterization of the optimal policy over the restricted set $\mathcal{S}_{FS}$ and in Section 4 we show a numerical evaluation comparing its performance to that of the optimal policy (over $\mathcal{S}$) and the suboptimal policy presented in 3.3.

Denote by $A(\boldsymbol{x}) := \{i : x_i > 0\}$ the set of classes with active flows at state $\boldsymbol{x}$, then

$$\mathcal{S}_{FS} := \left\{ \boldsymbol{x} : \sum_{i=1}^{K} x_i \leq \min_{j \in A(\boldsymbol{x})} \beta_j \right\} \subseteq \mathcal{S}.$$

As expected the first policy ω_1^{FS} has a simple form

$$\tau_i(\boldsymbol{x}) = \frac{x_i}{\sum_{j=1}^{K} x_j}, \qquad \boldsymbol{x} \in \mathcal{S}_{FS} \setminus \mathbf{0},\ i = 1, \ldots, K$$

$$p_i(\boldsymbol{x}) = \begin{cases} 1 & \text{if } \boldsymbol{x} + \boldsymbol{e}_i \in \mathcal{S}_{FS} \\ 0 & \text{otherwise} \end{cases}, \qquad \boldsymbol{x} \in \mathcal{S}_{FS},\ i = 1, \ldots, K$$

The policy evolution

$$\omega_1^{FS} \xrightarrow{T_1} \omega_2^{FS} \xrightarrow{T_2} \cdots \xrightarrow{T_{n-1}} \omega_n^{FS}$$

proceeds in much the same way as before with the difference that now type-B transitions are not applicable. Note that $\omega_m = \omega_n^{FS}$. Even thought there are no type-B transitions, building the sequence of policies $\omega_2^{FS}, \ldots, \omega_n^{FS}$ by successive transformations of ω_1^{FS} can still be exceedingly complex. Therefore we study the suboptimal approach in this case too by applying only type-C transformations.

3.5 On the optimality region of ω_1

Here we obtain, under some assumptions, an analytical expression that characterizes the optimality region of ω^1 ($[0, \rho^{(1)}]$) and provides some insight into it.

As before, we assume without any loss of generality that $\mu_1 C_1 \geq \mu_2 C_2 \geq \cdots \geq \mu_K C_K$. Moreover, for the following we need to assume that $\partial\mathcal{S} = \partial^K\mathcal{S} = \{\boldsymbol{x}_b = (0, \ldots, 0, \lfloor \beta_K \rfloor)\}$, which implies that the first policy transformation T_1 will consist of removing state $\boldsymbol{x}_b$ from $\mathcal{S}$. This assumption will be met iff $\forall i < K,\ \boldsymbol{x}_b + \boldsymbol{e}_i \notin \mathcal{S}$, which is equivalent to

$$\frac{1}{\beta_i} > 1 - \frac{\lfloor \beta_K \rfloor}{\beta_K} \qquad i = 1, \ldots, K-1. \tag{15}$$

A sufficient condition for (15) is $\beta_K \in \mathbb{N}$.

Under the assumption above we can write

$$P_b(\omega_1,\rho) = \frac{\sum_{i=1}^{K}\lambda_i}{\lambda}\pi_{\omega_1}(\boldsymbol{x}_b) + B_1 = \pi_{\omega_1}(\boldsymbol{x}_b) + B_1,$$
$$P_b(\omega_2,\rho) = \frac{\lambda_K}{\lambda}\pi_{\omega_2}(\boldsymbol{x}_b - \boldsymbol{e}_K) + B_2 = \frac{1}{1-\pi_{\omega_1}(\boldsymbol{x}_b)}\left(\frac{\lambda_K}{\lambda}\frac{\pi_{\omega_1}(\boldsymbol{x}_b)}{\rho_K} + B_1\right),$$

where B_1 is the contribution to P_b, under policy ω_1, of all states except $\boldsymbol{x}_b$; and B_2 is the contribution to P_b, under policy ω_2, of all states except $\boldsymbol{x}_b$ and arrivals of type K in state $\boldsymbol{x}_b - \boldsymbol{e}_K$. Since $P_b(\omega_1,\rho) \le P_b(\omega_2,\rho)$ iff $\rho \le \rho^{(1)}$ it follows that ρ lies in the optimality region of ω_1 iff

$$\frac{\lambda}{C_K\mu_K}\big(1 - P_b(\omega_1,\rho)\big) \le 1, \tag{16}$$

where λ can be expressed as $\lambda = \rho\sum_{i=1}^{K}\mu_i C_i\hat{\rho}_i$.

A possible interpretation of (16) is that if, for a given ρ, the total admitted rate under policy ω_1 was offered to a server with capacity equal to the most recourse consuming traffic class, and that such system is stable, then ω_1 is optimal.

4 Numerical Examples

In this section some numerical experiments are presented to illustrate the description of previous section and its potential applicability to practical scenarios. First, a rather simple configuration is studied with the purpose of exemplify the appearance of optimal and suboptimal policies, their evolution when load increases and the performance comparison among them. Next, a set of more realistic, albeit more complex, configurations is studied.

4.1 Basic Configuration

Let $(C_1, C_2) = (5, 3)$; $(\varphi_1, \varphi_2) = (1/2, 1/2)$; $(\mu_1, \mu_2) = (1, 1)$; $\hat{\boldsymbol{\rho}} = (2/5, 3/5)$;.

First Policy. The first policy ω_1, which is obtained as described in Section 3.1, is given by

$$[\tau_1(i,j)]_{ij} = \begin{bmatrix} & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1/2 & 1/3 & 1/4 & 1/5 & 1/6 & \cdot \\ 1 & 2/3 & 1/2 & 2/5 & 1/3 & \cdot & \cdot \\ 1 & 3/4 & 3/5 & 1/2 & 1/3 & \cdot & \cdot \\ 1 & 4/5 & 2/3 & 1/2 & \cdot & \cdot & \cdot \\ 1 & 5/6 & 2/3 & 1/2 & \cdot & \cdot & \cdot \\ 1 & 5/6 & 2/3 & \cdot & \cdot & \cdot & \cdot \\ 1 & 5/6 & \cdot & \cdot & \cdot & \cdot & \cdot \\ 1 & 5/6 & \cdot & \cdot & \cdot & \cdot & \cdot \\ 1 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 1 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \end{bmatrix} \qquad [\tau_2(i,j)]_{ij} = \begin{bmatrix} & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1/2 & 2/3 & 3/4 & 4/5 & 5/6 & \cdot \\ 0 & 1/3 & 1/2 & 3/5 & 2/3 & \cdot & \cdot \\ 0 & 1/4 & 2/5 & 1/2 & 2/3 & \cdot & \cdot \\ 0 & 1/5 & 1/3 & 1/2 & \cdot & \cdot & \cdot \\ 0 & 1/6 & 1/3 & 1/2 & \cdot & \cdot & \cdot \\ 0 & 1/6 & 1/3 & \cdot & \cdot & \cdot & \cdot \\ 0 & 1/6 & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 1/6 & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \end{bmatrix}$$

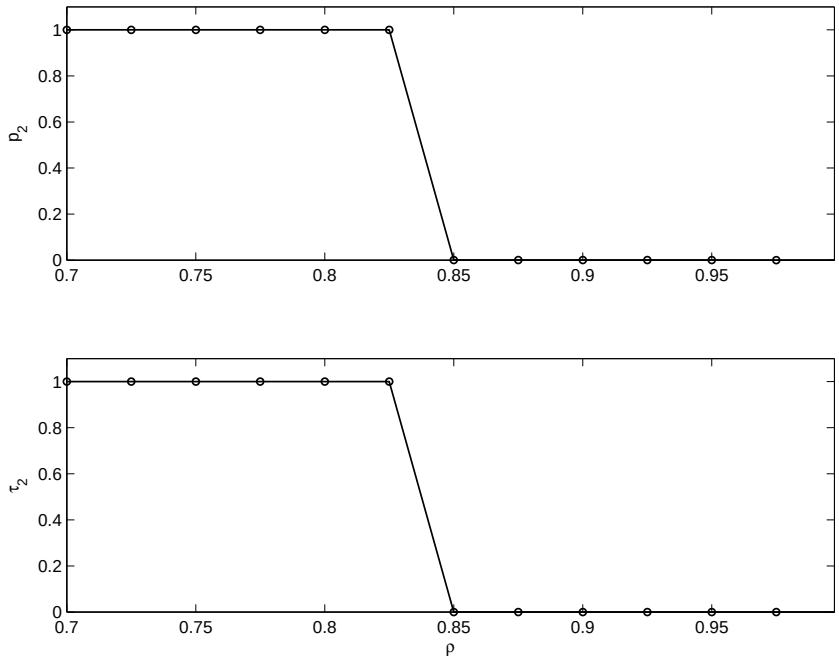


Figure 1: $p_2(0,4), \tau_2(0,5)$

$$[p_1(i,j)]_{ij} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 & \cdot \\ 1 & 1 & 1 & 1 & 2/3 & \cdot & \cdot \\ 1 & 1 & 1 & 3/4 & 0 & \cdot & \cdot \\ 1 & 1 & 4/5 & 3/5 & \cdot & \cdot & \cdot \\ 1 & 5/6 & 2/3 & 0 & \cdot & \cdot & \cdot \\ 1 & 5/6 & 0 & \cdot & \cdot & \cdot & \cdot \\ 1 & 5/6 & \cdot & \cdot & \cdot & \cdot & \cdot \\ 1 & 0 & \cdot & \cdot & \cdot & \cdot & \cdot \\ 1 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \end{bmatrix} \quad [p_2(i,j)]_{ij} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 & 0 & \cdot \\ 1 & 1 & 1 & 1 & 1 & \cdot & \cdot \\ 1 & 1 & 1 & 1 & 0 & \cdot & \cdot \\ 1 & 1 & 1 & 1 & \cdot & \cdot & \cdot \\ 1 & 1 & 1 & 0 & \cdot & \cdot & \cdot \\ 1 & 1 & 0 & \cdot & \cdot & \cdot & \cdot \\ 1 & 1 & \cdot & \cdot & \cdot & \cdot & \cdot \\ 1 & 0 & \cdot & \cdot & \cdot & \cdot & \cdot \\ 1 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \end{bmatrix}$$

Policy Evolution. Figures 1–6 show the evolution of the policy parameters as the load increases. Figure 7(a) depicts the blocking probability and a "summary" of the policy evolution. For each value of x_2 the admission probabilities for class 2, $p_2(x_1, x_2)$, have been averaged over those values of x_1 such that $(x_1, x_2 + 1) \in \mathcal{S}$, i.e., $p_2(x_1, x_2) > 0$ in ω_1. The resulting curves show the relative position (loadwise) of policy changes affecting a "row" of states (x_2 constant), and in particular values of ρ at which such rows are removed from the feasible states.

Figure 7(b) shows the same type of plot as Fig. 7(a) but now $\hat{\boldsymbol{\rho}} = (1/3, 2/3)$ has been varied. From the shape of the curves we observe that, as expected, varying $\hat{\boldsymbol{\rho}}$ changes the values $\rho^{(j)}$ but not the set of optimal policies ω_j.

In order to see the effect of modifying the resource consumption ordering we set $\mu_2 = 2$. Now $C_1 = 5 > 3 = C_2$ but $\mu_1 C_1 = 5 < 6 = \mu_2 C_2$, so the optimal policy evolves limiting the access of

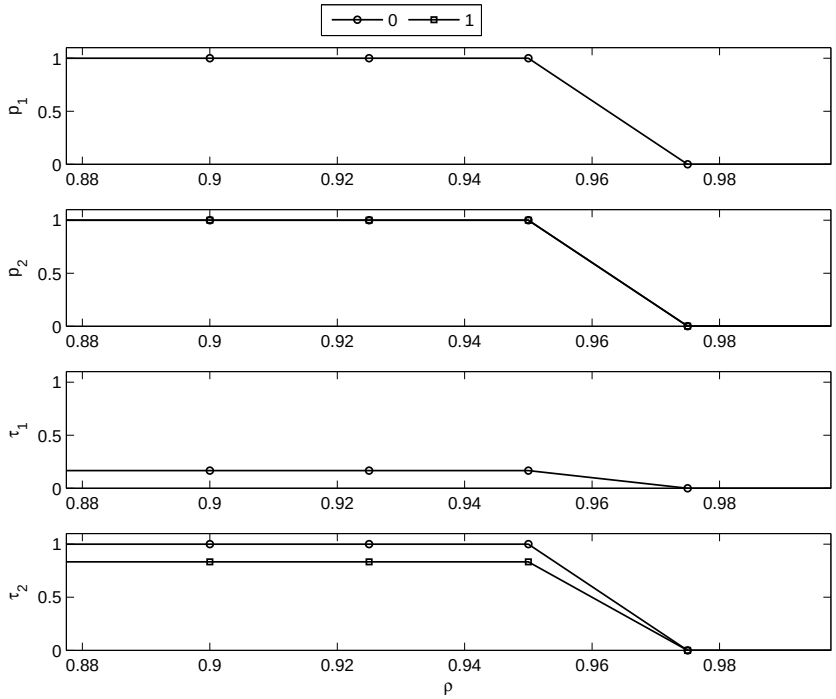


Figure 2: $p_1(x_1,5), p_2(x_1,4), \tau_1(x_1,5), \tau_2(x_1,5)$; curves are parameterized by x_1

class-1 traffic as shown in Fig. 8.

Optimal Policy in $\mathcal{S}_{FS}$. The first policy ω_1^{FS} is given by

$$[\tau_1(i,j)]_{ij} = \begin{bmatrix} & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1/2 & 1/3 & 1/4 & 1/5 & 1/6 & \cdot \\ 1 & 2/3 & 1/2 & 2/5 & 1/3 & \cdot & \cdot \\ 1 & 3/4 & 3/5 & 1/2 & \cdot & \cdot & \cdot \\ 1 & 4/5 & 2/3 & \cdot & \cdot & \cdot & \cdot \\ 1 & 5/6 & \cdot & \cdot & \cdot & \cdot & \cdot \\ 1 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 1 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 1 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 1 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 1 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \end{bmatrix} \quad [\tau_2(i,j)]_{ij} = \begin{bmatrix} & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1/2 & 2/3 & 3/4 & 4/5 & 5/6 & \cdot \\ 0 & 1/3 & 1/2 & 3/5 & 2/3 & \cdot & \cdot \\ 0 & 1/4 & 2/5 & 1/2 & \cdot & \cdot & \cdot \\ 0 & 1/5 & 1/3 & \cdot & \cdot & \cdot & \cdot \\ 0 & 1/6 & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \end{bmatrix}$$

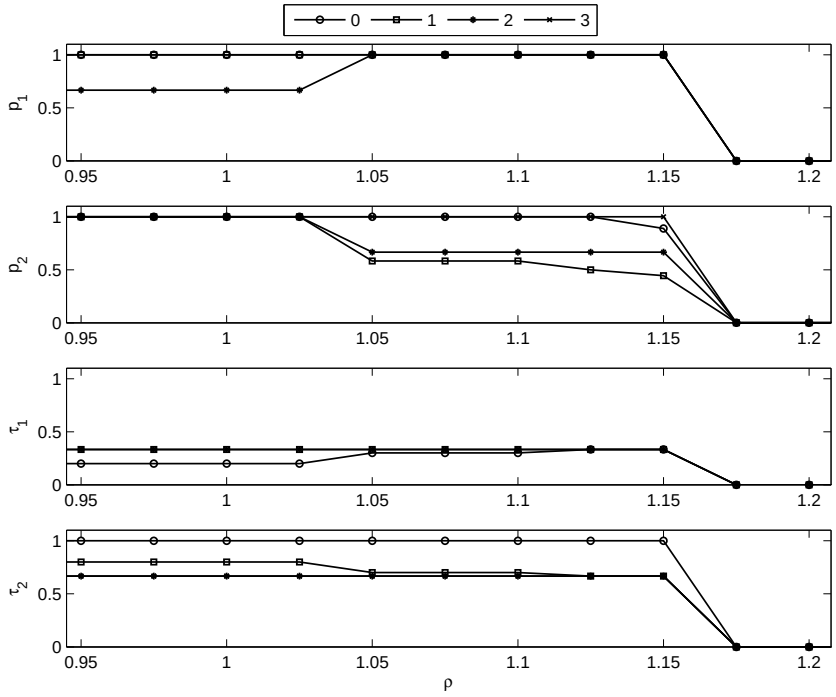


Figure 3: $p_1(x_1,4), p_2(x_1,3), \tau_1(x_1,4), \tau_2(x_1,4)$; curves are parameterized by x_1

$$[p_1(i,j)]_{ij} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 & 0 & \cdot \\ 1 & 1 & 1 & 1 & 0 & \cdot & \cdot \\ 1 & 1 & 1 & 0 & \cdot & \cdot & \cdot \\ 1 & 1 & 0 & \cdot & \cdot & \cdot & \cdot \\ 1 & 0 & \cdot & \cdot & \cdot & \cdot & \cdot \\ 1 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 1 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 1 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 1 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \end{bmatrix} \quad [p_2(i,j)]_{ij} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 & 0 & \cdot \\ 1 & 1 & 1 & 1 & 0 & \cdot & \cdot \\ 1 & 1 & 1 & 0 & \cdot & \cdot & \cdot \\ 1 & 1 & 0 & \cdot & \cdot & \cdot & \cdot \\ 1 & 0 & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \end{bmatrix}$$

Figure 9 shows the policy evolution as load increases. Observe that the curves corresponding to $p_1(\cdot,\cdot)$ (Fig. 9(a)) show a simpler behavior ($p_1(\cdot,\cdot)$ is either 0 or 1) than in the general case, which is due to the fact that there are no type-B policy transformations.

Comparison of policies. The curves in Fig. 10 represent the relative value of P_b for the different policies taking the optimal insensitive policy as the reference. The first policies (ω_1 and ω_1^{FS}) show important degradations as the load moves away from their optimality regions so it does not seem advisable to keep using ω_1 far beyond $\rho^{(1)}$. The optimal and suboptimal policies in $\mathcal{S}_{FS}$ (ω^{FS}, $\hat{\omega}^{FS}$) show a poor performance except for considerably high loads, where all policies tend to converge and the actual value of P_b is probably too high for such loads being considered a practical point of operation. Despite having little interest due to its poor performance compared

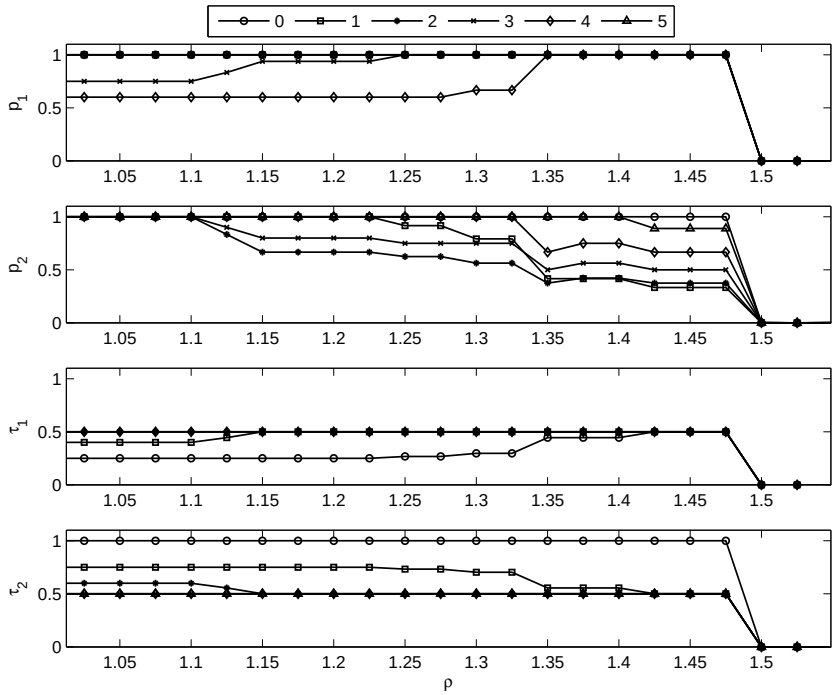


Figure 4: $p_1(x_1,3), p_2(x_1,2), \tau_1(x_1,3), \tau_2(x_1,3)$; curves are parameterized by x_1

to ω, it is noticeable that the policy $\hat{\omega}^{FS}$ is actually quite good as an approximation, i.e., it is only slightly above the policy ω^{FS}. Likewise $\hat{\omega}$ is an excellent approximation to ω, which is the targeted optimum, so its relative performance is also excellent; in this scenario the maximum deviation of $P_b(\hat{\omega})$ from $P_b(\omega)$ is a 1.3%. We also plotted a curve corresponding to the optimal (non-necessarily insensitive) policy, which exhibits an important gain over the more restrictive case of insensitive policies. For this curve, though, the validity of the results is limited to the case where the flow sizes are exponentially distributed. In order to compute P_b for the optimal policy the set of equations corresponding to the detailed balanced condition in the linear program (see Eq. (8)) were substituted by the global balance equations

$$\sum_i \left(I_{\{\boldsymbol{x}+\boldsymbol{e}_i \in \mathcal{S}\}} \lambda_i \sum_{\boldsymbol{d}:d_i=1} \sum_b \pi(\boldsymbol{x},\boldsymbol{d},b) + I_{\{\boldsymbol{x}-\boldsymbol{e}_i \in \mathcal{S}\}} C_i \mu_i \sum_{\boldsymbol{d}} \pi(\boldsymbol{x},\boldsymbol{d},i) \right.$$
$$\left. -\lambda_i \sum_{\boldsymbol{d}:d_i=1} \sum_b \pi(\boldsymbol{x}-\boldsymbol{e}_i,\boldsymbol{d},b) - C_i \mu_i \sum_{\boldsymbol{d}} \pi(\boldsymbol{x}+\boldsymbol{e}_i,\boldsymbol{d},i) \right) = 0 \qquad \forall \boldsymbol{x} \in \mathcal{S}$$

where $I_{\{\cdot\}}$ is the indicator function and by convention $\pi(\boldsymbol{x},\boldsymbol{d},b) = 0$ if $\boldsymbol{x} \notin \mathcal{S}$.

In Fig. 11 the sensitivity of the relative suboptimal performance to different configuration parameters is analyzed. All curves but the last one ($\mu_2/\mu_1 = 0.5$) display an excellent performance of $\hat{\omega}$. Further experiments in that direction revealed that it is indeed the imbalance between $\mu_1 C_1$ and $\mu_2 C_2$ that is the cause of the performance degradation.

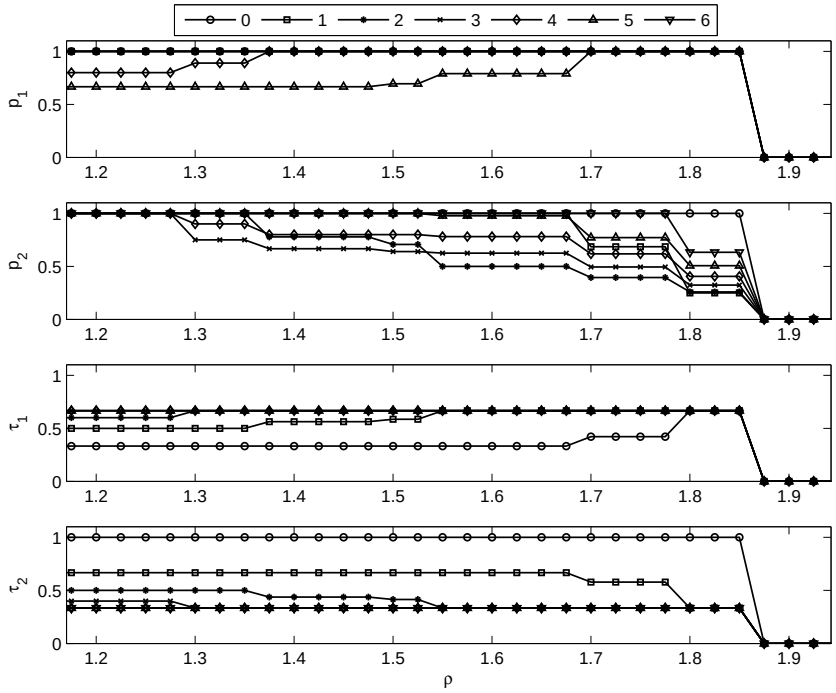


Figure 5: $p_1(x_1, 2), p_2(x_1, 1), \tau_1(x_1, 2), \tau_2(x_1, 2)$; curves are parameterized by x_1

4.2 Practical Scenarios

Now we consider a typical scenario with where $C_1/C_2 = 10$, and the mean flow sizes and minimum bit rates are equal for both flow classes, i.e., $\mu_1 = \mu_2$ and $\varphi_1 = \varphi_2$. We assume an ideal propagation model with no obstacles so that the cell areas corresponding to feasible rates $C1$ and $C2$ are two concentric ring of external radius r_1 and r_2 respectively. Moreover [12],

$$\frac{C_1}{C_2} = \left(\frac{r_2}{r_1}\right)^{\alpha}, \tag{17}$$

where α is the path loss exponent. We also assume that users are homogeneously distributed over the service area. Then the arrival rate in area is proportional to its surface. Therefore, the relative the relative load of the outer ring with respect to the inner one is given by

$$\frac{\rho_2}{\rho_1} = \frac{\lambda_2}{\lambda_1}\frac{C_1}{C_2} = \frac{\pi(r_2^2 - r_1^2)}{\pi r_1^2}\frac{C_1}{C_2} = \frac{C_1}{C_2}\left[\left(\frac{C_1}{C_2}\right)^{2/\alpha} - 1\right] = 10\left[(10)^{2/\alpha} - 1\right].$$

In the above derivation we employed Eq. (17) which implicitly assumes that the target E_b/N_0 (energy per bit to noise density ratio) was constant for all rates, what results in a linear feasible rate versus SINR (signal to interference-plus-noise ratio) dependency. In real systems the target is not necessarily constant but nevertheless the load distribution remains approximately the same if real values are used [12].

The feasible rates for the outer rings are in the order of tens to a few hundred of *kbps* (see Table 1 in [12]). We consider that a typical value for the minimum bit rates would be, at the very least, in the order of a few tens of *kbps*. Thus, maximum number of class-2 users that can be accepted ($\beta_2 = C_2/\varphi_2$) will typically take values below 10. Given a value of β_2, the value of β_1 is computed as $\beta_1 = 10\beta_2 \leq 100$.

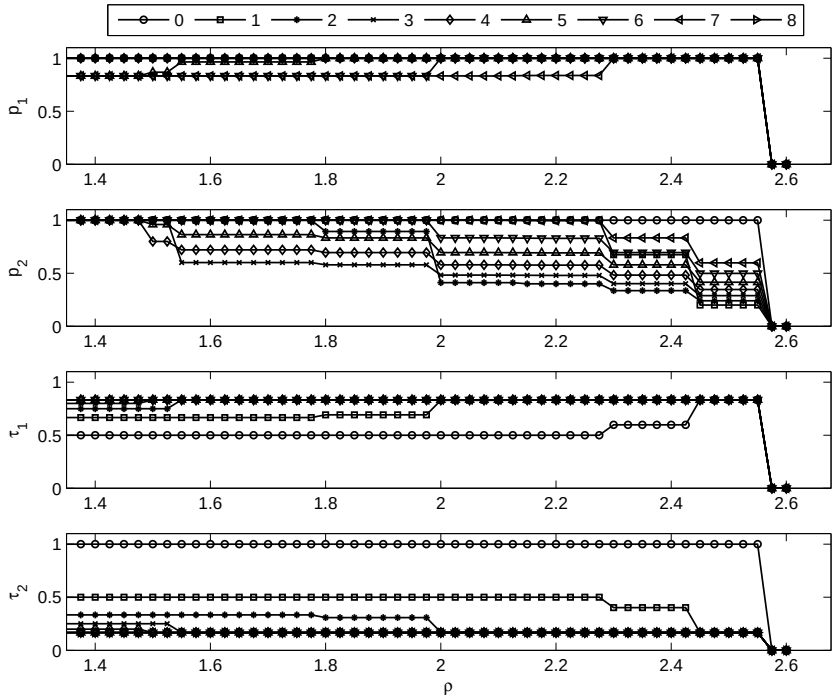
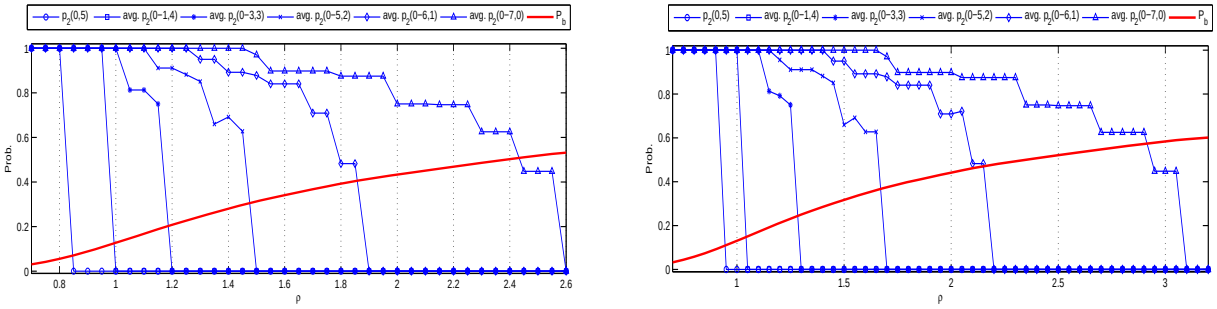


Figure 6: $p_1(x_1,1), p_2(x_1,0), \tau_1(x_1,1), \tau_2(x_1,1)$; curves are parameterized by x_1



(a) $\hat{\boldsymbol{\rho}} = (2/5, 3/5)$ (b) $\hat{\boldsymbol{\rho}} = (1/3, 2/3)$

Figure 7: Blocking probability and policy variations

All in all, the cardinality of the state space associated to the scenario specified above is far to high for getting any insight by inspecting the shape of the optimal policy. Furthermore, the size of the LP problem also poses a daunting problem in terms of accuracy. On the other hand, however, the optimality region of the *first policy*, ω_1, turns out to be wide enough to cover the region were the performance takes values which are of interest from a practical perspective.

By solving (16) (actually is it solved as an equality) one obtains $\rho^{(1)}$ and $P_b\big(\rho^{(1)}\big)$. Since $P_b(\rho) > P_b\big(\rho^{(1)}\big)$ if $\rho > \rho^{(1)}$, a value of $P_b\big(\rho^{(1)}\big)$ high enough entails that, for practical purposes, knowing ω_1 is enough. Figure 12 plots the value of $P_b\big(\rho^{(1)}\big)$ for the setting of our scenario. As shown, in most cases the load region of interest is included within the optimality region of ω_1. Only when both α and β_2 take the highest values, a load higher than $\rho^{(1)}$ would yield still a performance that is acceptable.

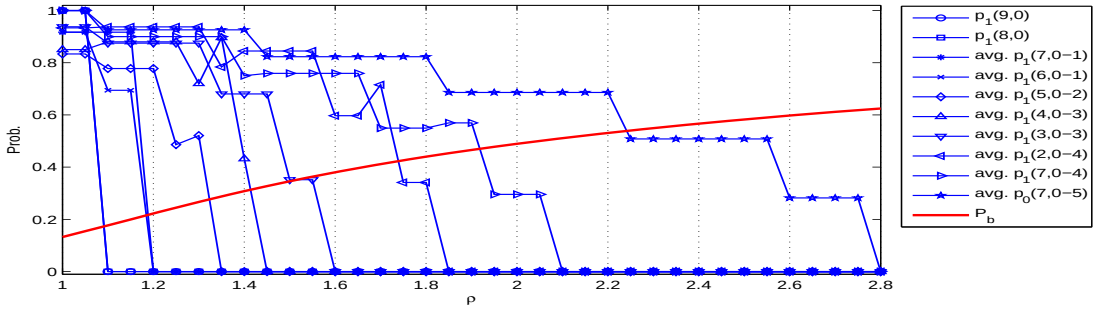
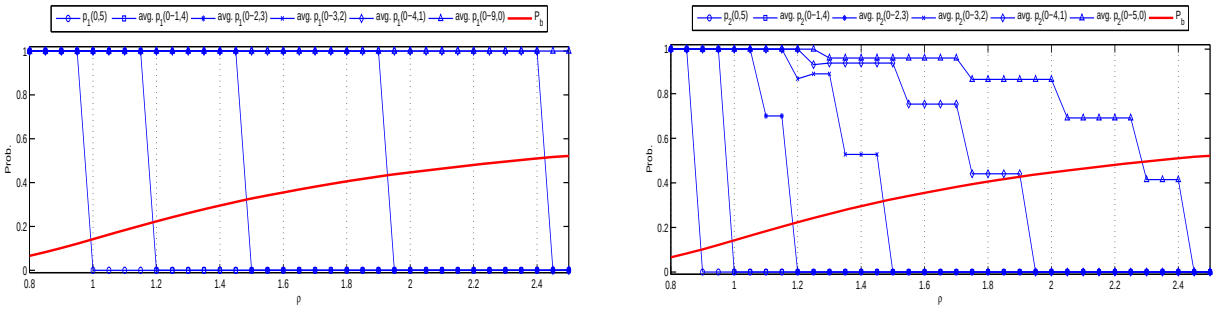


Figure 8: Blocking probability and policy variations; $\mu_2 = 2$



(a) $p_1(\cdot,\cdot)$ (b) $p_2(\cdot,\cdot)$

Figure 9: Blocking probability and policy variations

5 Conclusions

We have considered the joint optimization of bandwidth allocation and admission control for elastic users competing for a downlink data channel in a cellular network. Robustness and generality of the results were main concerns in our research and so we focused on those strategies that are insensitive to the detailed traffic characteristics beyond mean values. The optimization problem has been formulated using a *Markov Decision Process-Linear Programming* approach. A characterization of the optimal policy has been obtained inductively. It has been found that the optimal policy is a piecewise constant function of the system load having only finitely

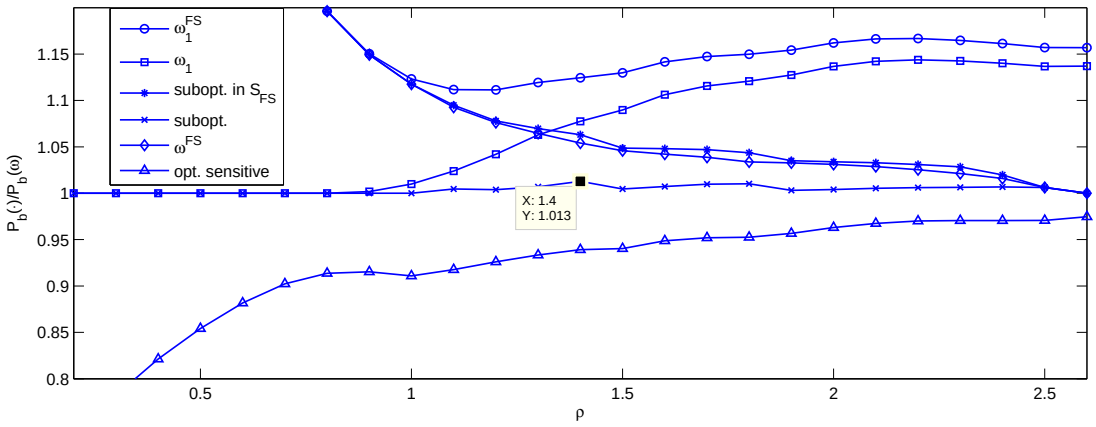


Figure 10: Relative performance: basic configuration.

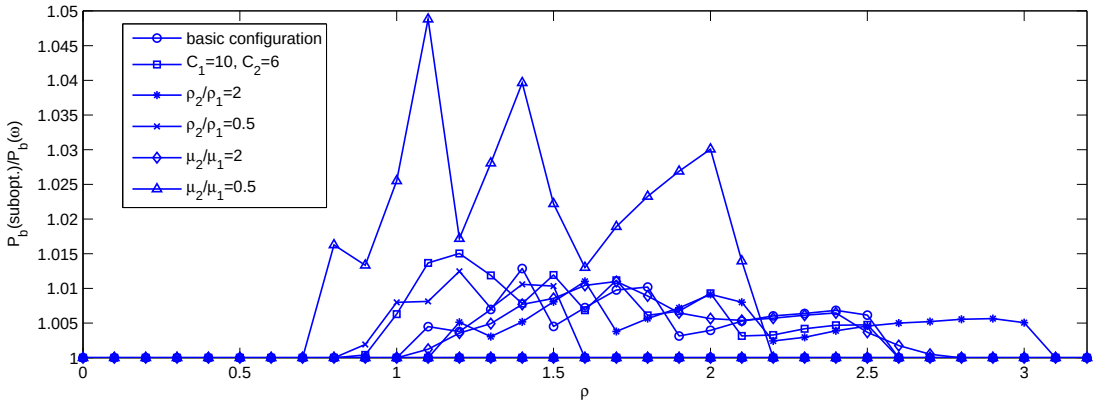


Figure 11: Relative performance of $\hat{\omega}$: sensitivity to configuration parameters.

many pieces. Moreover, the policy settings for each piece do only depend on the minimum rate requirements and feasible rates, in particular they are not dependent on the arrival rates. These features confer additional robustness to the solution.

We observed that in some cases, the complexity of computing the optimal policy can render this policy impractical except for low loads. As an alternative to those cases we proposed a much simpler suboptimal policy that satisfies the same requirements and achieves a good relative performance unless the values $\mu_i C_i$ (reciprocal of the mean sojourn time if a class-i user was alone in the system) for the different user classes are significantly imbalanced.

Finally, we have shown that in most scenarios of practical interest, outside the load region where the optimal policy can be efficiently computed the performance is too poor for being practical.

References

[1] Jim W. Roberts. Traffic theory and the Internet. *Communications Magazine, IEEE*, 39(1):94–99, 2001.

[2] L. Massoulié and J.W. Roberts. Arguments in favour of admission control for TCP flows. In *Proceedings of ITC 16*, 1999.

[3] T. Bonald and J.W. Roberts. Congestion at flow level and the impact of user behaviour. *Computer Networks*, 42:521–536, 2003.

[4] F.P. Kelly. *Reversibility and Stochastic Networks*. Wiley New York, 1979.

[5] T. Bonald and A. Proutière. Insensitivity in processor-sharing networks. *Performance Evaluation*, 49(1-4):193–209, 2002.

[6] T. Bonald and A. Proutière. Insensitive bandwidth sharing in data networks. *Queueing Systems: Theory and Applications*, 44(1):69–100, 2003.

[7] T. Bonald. Insensitive queueing models for communication networks. In *Valuetools '06: Proceedings of the 1st international conference on Performance evaluation methodolgies and tools*, page 57, New York, NY, USA, 2006. ACM Press.

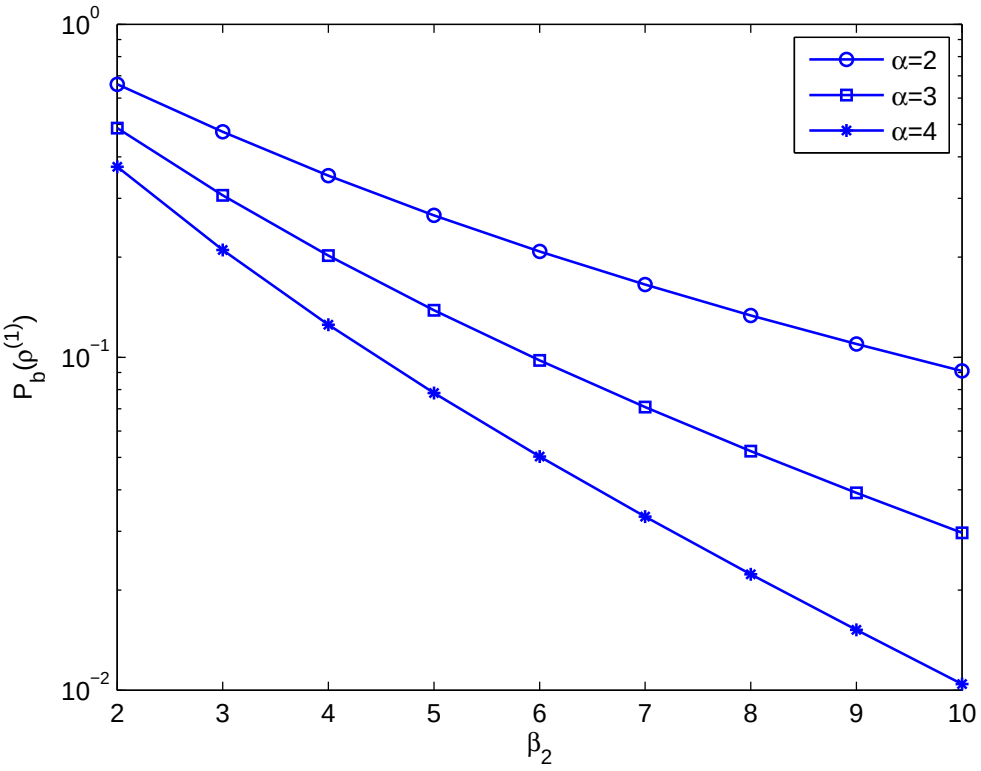


Figure 12: Maximum value of P_b within the optimality region of ω_1.

[8] T. Bonald, L. Massoulié, A. Proutière, and J. Virtamo. A queueing analysis of max-min fairness, proportional fairness and balanced fairness. *Queueing Systems: Theory and Applications*, 53(1-2):65–84, 2006.

[9] Villy Bæk Iversen. Reversible fair scheduling: the teletraffic theory revisited. In *Proceedings of the 20th International Teletraffic Congress ITC 20*, 2007.

[10] S.C. Borst, K.L. Clarkson, J.M. Graybeal, H. Viswanathan, and P.A. Whiting. User-level QoS and traffic engineering for 3G wireless 1xEV-DO systems. *Bell Labs Technical Journal*, 8(2):33–47, 2003.

[11] Sem Borst. User-level performance of channel-aware scheduling algorithms in wireless data networks. *IEEE/ACM Transactions on Networking*, 13(3):33–47, June 2005.

[12] Thomas Bonald and Alexandre Proutiere. Wireless downlink data channels: user performance and cell dimensioning. In *MobiCom '03: Proceedings of the 9th annual international conference on Mobile computing and networking*, pages 339–352, New York, NY, USA, 2003. ACM Press.

[13] T. Bonald, S.C. Borst, and A. Proutiere. How mobility impacts the flow-level performance of wireless data systems. In *INFOCOM 2004*, volume 3, pages 1872–1881. IEEE, 2004.

[14] T. Bonald, S. Borst, N. Hegde, and A. Proutiére. Wireless data performance in multi-cell scenarios. In *SIGMETRICS '04/Performance '04: Proceedings of the joint international*

conference on Measurement and modeling of computer systems, pages 378–380, New York, NY, USA, 2004. ACM.

[15] S. Liu and J. Virtamo. Performance analysis of wireless data systems with a finite population of mobile users. In *Proceedings of the 19th International Teletraffic Congress ITC 19*, pages 1295–1304, 2005.

[16] S. Borst, A. Proutiere, and N. Hegde. Capacity of Wireless Data Networks with Intra-and Inter-Cell Mobility. In *INFOCOM 2006*, pages 1–12. IEEE, 2006.

[17] Yujing Wu, Carey Williamson, and Jingxiang Luo. On processor sharing and its applications to cellular data network provisioning. *Performance Evaluation*, 64(9–12):892–908, October 2007.

[18] S. Borst and N. Hegde. Integration of streaming and elastic traffic in wireless networks. In *INFOCOM 2007*, pages 1884–1892. IEEE, 2007.

[19] G. Fayolle, I. Mitrani, and R. Iasnogorodski. Sharing a processor among many job classes. *J. ACM*, 27(3):519–532, 1980.

[20] K. Avrachenkov, U. Ayesta, P. Brown, and R. Núñez-Queija. Discriminatory processor sharing revisited. In *Proceedings of IEEE INFOCOM 2005*, volume 2, pages 784–795, 2005.

[21] E. Altman, K. Avrachenkov, and U. Ayesta. A survey on discriminatory processor sharing. *Queueing Systems*, 53(1):53–63, 2006.

[22] Samuli Aalto, Urtzi Ayesta, Sem Borst, Vishal Misra, and Rudesindo Núñez-Queija. Beyond processor sharing. *SIGMETRICS Performance Evaluation Review*, 34(4):36–43, 2007.

[23] S. F. Yashkov and A. S. Yashkova. Processor sharing: A survey of the mathematical theory. *Autom. Remote Control*, 68(9):1662–1731, 2007.

[24] T. Bonald, S. Borst, and A. Proutiere. Inter-cell coordination in wireless data networks. *European transactions on telecommunications*, 17(3):303–312, 2006.

[25] S. Liu and J. Virtamo. Inter-cell coordination with inhomogeneous traffic distribution. In *Proc. of the 2nd Conference on Next Generation Internet Design and Engineering (NGI 06)*, 2006.

[26] G. van Kessel, R. Núñez-Queija, and S. Borst. Differentiated bandwidth sharing with disparate flow sizes. In *Proceedings of IEEE INFOCOM 2005*, 2005.

[27] Thomas Bonald and James Roberts. Scheduling network traffic. *SIGMETRICS Performance Evaluation Review*, 34(4):29–35, 2007.

[28] Rudesindo Núñez-Queija, J. L. Berg, and Michel R.H. Mandjes. Performance evaluation of strategies for integration of elastic and stream traffic. Technical report, CWI (Centre for Mathematics and Computer Science), Amsterdam, The Netherlands, 1999.

[29] Remco Litjens, Hans van den Berg, and Richard J. Boucherie. Throughputs in processor sharing models for integrated stream and elastic traffic. *Performance Evaluation*, 65(2):152–180, 2008.

[30] Sándor Rácz, Balázs Péter Gerö, and Gábor Fodor. Flow level performance analysis of a multi-service system supporting elastic and adaptive services. *Performance Evaluation*, 49(1-4):451–469, 2002.

[31] N. Antunes, C. Fricker, F. Guillemin, and Ph. Robert. Integration of streaming services and TCP data transmission in the Internet. *Performance Evaluation*, 62(1-4):263–277, 2005.

[32] Sai Rajesh Mahabhashyam and Natarajan Gautam. On queues with markov modulated service rates. *Queueing Syst. Theory Appl.*, 51(1-2):89–113, 2005.

[33] N. Benameur, S. Ben Fredj, S. Oueslati-Boulahia, and J. W. Roberts. Quality of service and flow level admission control in the internet. *Comput. Networks*, 40(1):57–71, 2002.

[34] F. Delcoigne, A. Proutière, and G. Régnié. Modeling integration of streaming and data traffic. *Performance Evaluation*, 55(3-4):185–209, 2004.

[35] OJ Boxma, AF Gabor, R. Núñez-Queija, and H.P. Tan. Performance analysis of admission control for integrated services with minimum rate guarantees. In *Proceedings of NGI'06*, pages 41–47, 2006.

[36] T. Bonald and A. Proutière. On stochastic bounds for monotonic processor sharing networks. *Queueing Syst. Theory Appl.*, 47(1-2):81–106, 2004.

[37] Thomas Bonald and Alexandre Proutière. On performance bounds for the integration of elastic and adaptive streaming flows. In *SIGMETRICS '04/Performance '04: Proceedings of the joint international conference on Measurement and modeling of computer systems*, pages 235–245, New York, NY, USA, 2004. ACM Press.

[38] T. Bonald, M. Jonckheere, and A. Proutiére. Insensitive load balancing. In *SIGMETRICS '04/Performance '04: Proceedings of the joint international conference on Measurement and modeling of computer systems*, pages 367–377, New York, NY, USA, 2004. ACM Press.

[39] M. Jonckheere and J. Virtamo. Optimal insensitive routing and bandwidth sharing in simple data networks. In *SIGMETRICS '05: Proceedings of the 2005 ACM SIGMETRICS international conference on Measurement and modeling of computer systems*, pages 193–204, New York, NY, USA, 2005. ACM Press.

[40] Juha Leino and Jorma Virtamo. Insensitive load balancing in data networks. *Computer Networks*, 50(8):1059–1068, June 2006.

[41] Sheldon M. Ross. *Applied probability models with optimization applications*. Holden-Day, 1970.

[42] Rudesindo Núñez-Queija and Hwee-Pink Tan. Location-based admission control for differentiated services in 3G cellular networks. In *Proceedings of the 9th ACM international symposium on Modeling analysis and simulation of wireless and mobile systems (MSWiM'06)*, pages 322–329, New York, NY, USA, 2006. ACM.

[43] R. Nelson. *Probability, Stochastic Processes and Queueing Theory.* Springer-Verlag, 1995.