Available online at www.sciencedirect.com

ScienceDirect

Operations Research Letters 35 (2007) 297 – 307

Operations Research Letters

www.elsevier.com/locate/orl

Pricing and distributed QoS control for elastic network traffic

Hans van den Berg[a, c], Michel Mandjes[b, d], Rudesindo Núñez-Queija[b, e, *]

[a] *TNO Information and Communication Technology, P.O. Box 5050, 2600 GB Delft, The Netherlands*
[b] *CWI, P.O. Box 94079, 1090 GB Amsterdam, The Netherlands*
[c] *Department of Electrical Engineering, Mathematics, and Computer Science, University of Twente, The Netherlands*
[d] *Korteweg-de Vries Institute for Mathematics, University of Amsterdam, The Netherlands*
[e] *Department of Mathematics & Computer Science, Eindhoven University of Technology, The Netherlands*

Received 24 December 2005; accepted 31 March 2006
Available online 27 June 2006

Abstract

We study a processor-sharing model in which users choose between a high- and a low-priority service, based on their utility functions and prices charged by the service provider. The latter aims at revenue maximization. The model is motivated by file transmissions in data networks with distributed congestion control.

Keywords: Pricing & distributed control; Size-based prioritization; Multiclass processor-sharing queue; Nash equilibrium; Internet traffic

1. Introduction

The vast majority of traffic on the Internet relates to the transfer of documents (web pages, audio/video downloads, file transfers, etc.), usually coordinated by transmission control protocol (TCP). TCP is designed to support the transmission of *elastic* jobs, i.e., jobs that tolerate some variations in the throughput. By noticing packet loss, the end-stations are provided with information on the level of congestion along the path, based on which they adapt their transmission rates.

A TCP-based data transfer starts with a slow-start phase that estimates the bandwidth available, followed by the so-called congestion-avoidance phase, during which all active users are assigned equal bandwidth (assuming they have identical access rates and round-trip times). This motivated the use of the so-called *processor sharing* (PS) queueing discipline to model the dynamic behavior of TCP flows sharing a common network link [15].

TCP tends to share the network resources fairly among the users. For instance, no distinction is made on the basis of the sizes of the documents to be transferred. It can be argued, however, that such an equal sharing policy has important drawbacks. Consider for instance the situation in the Internet where essentially two types of flows can be distinguished: there are many small flows (commonly referred to as *mice*), and a small number of extremely long flows (*elephants*), such that the elephants make up a significant fraction

[*] Corresponding author. CWI, P.O. Box 94079, 1090 GB Amsterdam, The Netherlands.
E-mail address: sindo@cwi.nl (R. Núñez-Queija).

0167-6377/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.orl.2006.03.018

of the offered load [2,6]. If resources are shared fairly, the performance experienced by (many) short flows is seriously degraded by the small number of long flows.

This problem can be avoided by using e.g., the following simple flow-size based priority discipline. Flows with sizes less than a threshold value t are transmitted with high priority, whereas flows larger than t get low-priority service. In practice, it may not be realistic that end users actively choose the best priority class for each document, but automated agents can be used to make these decisions for the individual users.

An attractive property of the above policy is that for short flows it outperforms the non-priority system considerably, whereas for long flows there is typically just a slight performance degradation. This can be understood as follows. Suppose the threshold is such that the short (high-priority) flows account for a load ρ_{h}, and the long (low-priority) flows for ρ_{l}, where $\rho := \rho_{\mathrm{h}} + \rho_{\mathrm{l}} < 1$ (the server's speed is normalized to 1). Then the short jobs see a PS queue with load ρ_{h}, rather than a PS queue with load ρ. On the other hand, as we will show below, the long jobs roughly experience a PS queue with load ρ_{l} and server speed $1 - \rho_{\mathrm{h}}$, which leads to a performance (expressed in terms of the transfer delay for a job of given size) comparable to a (single class) PS queue with total load ρ and server speed 1.

The idea of improving performance, particularly of short flows, by discriminating between jobs of different sizes has a rich history [25,12,11] resulting in *service disciplines* such as *shortest remaining processing time* (SRPT), *multi-level processor sharing* (MLPS) and *foreground–background processor sharing* (FBPS). The subject recently regained attention in the context of flow transfers in the Internet and file retrieval from web servers [1,3,4,8,26].

When remaining service requirements are known, SRPT assigns full capacity to the job(s) which have the least amount of service left. It is known that SRPT minimizes the number of jobs in the system (and, therefore, the mean delay) among all work-conserving policies [25]. FBPS is the stochastic counterpart of SRPT: full service is given to the jobs that so far have received the least amount of service. For service requirement distributions with *decreasing hazard rate functions*, FBPS minimizes the mean delay among all work-conserving and *non-anticipating* (i.e., without knowledge of residual service requirements) disciplines [23]. MLPS is a discrete-class analogue of FBPS, where jobs move to lower-priority classes as the amount of service received passes certain threshold values. Like FBPS, MLPS does not require information about the residual job sizes, but does keep track of the received amounts of service. Moreover, large jobs still affect short jobs through their service requirements up to the threshold levels. The simple two-class priority discipline sketched above, overcomes these problems if users are allowed to choose the service classes themselves. By imposing charges on network usage, incentives can be given such that short flows choose the high-priority class and long flows choose the low-priority class. Thus, appropriate pricing schemes provide the opportunity of distributed control: end users choose the priority class which determines the performance received, whereas the network elements' complexity can be kept low (the high-priority packets should be marked, and then the desired prioritization can be achieved by standard priority schedulers).

As indicated above, several papers have studied the impact of flow-size-based scheduling disciplines. In particular, [1] investigated an implementation of a related two-level priority rule in Internet routers. (There the priority of jobs is decreased when a certain size is attained.) The present paper adds the element of pricing, thus offering the opportunity of distributed control. There is a growing body of literature on QoS differentiation and pricing, see for instance [7,13], various chapters of [16] and the recent work by Hassin and Haviv [9] (in particular, pp. 86–87) and [10]. Early works on a game-theoretic approach to queueing systems are the seminal papers by Mendelson and Whang [17,18]. For given penalty functions, they find incentive-compatible prices that maximize the system's 'net value'. Stidham [27] considered a model in which users are heterogeneous with respect to both their utility functions and their sensitivity to congestion (e.g., delay). The *Paris metro pricing* approach discussed in [22] offers different levels of service by using logically separated networks with different prices. Using game-theoretic techniques, [5] argue that this mechanism does not work if there are multiple competing providers: in order to maximize profits the providers rather focus on one user type.

Our analysis is based on a fundamental queueing model, in which jobs arrive according to a Poisson process with rate λ, and jobs are i.i.d. distributed as a random variable B. The jobs choose between a high-priority and a low-priority queue, where *within* each queue the capacity is shared in a PS manner. The server speed of the queue is normalized to 1. To avoid that all users opt for high-priority service for (all of) their jobs, differentiated charges are imposed: users pay an amount proportional to the job size, where the price per, say, bit in the high-priority queue is higher than in the low-priority queue. Evidently, users choose between the queues based on the performance offered and the prices. Simultaneously, the network provider can choose the prices so as to optimize its profit.

We analyze this system in two stages. We first identify the Nash equilibrium for given prices, i.e., the situation in which no user has incentives to unilaterally change its policy. Next, knowing the users' reactions to the pricing structure, the network provider chooses the prices so as to maximize the total revenue. Notice that we do not a priori assume $\rho := \lambda \mathbb{E} B < 1$. Stability is ensured by the fact that users may choose to refrain from service ('balk') if prices are considered to be too high for the quality of service offered. This will surely be the case in an overload scenario, but even for loads below 1, it could be that users choose not to transmit a substantial part of the documents. (These could then opt for service elsewhere.) Our analysis shows that the revenue maximizing equilibrium is such that *medium-sized* flows balk. This is intuitively explained as follows: for small flows, the prioritization has a significant positive effect on their transfer delay making the relatively high price worth paying. The transfer delay of the large flows, on the other hand, is almost insensitive to prioritization of the small flows, offering the network provider the possibility to use his network resources efficiently by keeping the prices relatively low. The medium-sized flows are not sensitive enough to the prioritization to have any potential for large revenues and at the same time they are too sensitive to provide scope for efficient network usage. We expect similar results to hold when there are more than two priority classes. In the revenue optimizing equilibrium, the classes with higher priority will attract shorter flows, with possibly non-empty sets of flow sizes that are not transmitted because they 'fall between' two adjacent priority classes.

This paper is organized as follows. Section 2 sketches the model and describes the user behavior for given prices. Section 3 contains the main result: the structure of the revenue maximizing Nash equilibrium, where we observe that small flows opt for premium service and large flows for low-priority service, whereas medium-size flows potentially balk. Section 4 illustrates the theory through a numerical example. Finally, some concluding remarks are made in Section 5.

2. Model description

Before presenting the model, we note that our results would remain valid under less restrictive conditions. In Remark 2.1 we briefly discuss which are the essential properties that are used in the proofs.

To introduce the model, and for later comparisons, we first study the situation with no prioritization. In that case, the system is modeled as an M/G/1 PS queue. With prioritization, the model becomes a two-class M/G/1 PS queue with preemptive priority, as already mentioned in the introduction. For this model we first discuss the dynamics for static choices of the users. Then we discuss how the users' choices depend on the prices charged. The issue of how the system manager should set the prices so as to maximize the revenue is discussed in Section 3.

2.1. No prioritization

Requests for file transfers arrive according to a Poisson process of rate λ and the file sizes are distributed as the random variable B with distribution function $F(x)$, $x \geqslant 0$. We assume that $F(x)$ is continuous.

In the absence of prioritization, users only have two choices: either to transmit or not to transmit. The load of files that are transmitted is equal to ρ_* (the capacity of the system is normalized to 1).

Assuming the system is in statistical equilibrium, we denote the expected delay (transfer time) of a transmitted file size x by $D(x)$. In the sequel, the central measure of performance will be the *stretch* of a request: $S(x) := D(x)/x$. (The stretch is sometimes called "slowdown" [28].) It is well known, see [24,11, Section 4.4], that the stretch is independent of x:

$$S(x) \equiv \frac{1}{1-\rho_*}.$$

Suppose the price for transmitting a file of length x is $xp(x)$. The users' utilities are reflected in the *willingness to pay* function $w(s)$, which is the amount *per bit* that users are willing to pay for transmitting a file of size x in sx time units. To the best of our knowledge, this paper is the first to consider prices that depend on the stretch. We assume that $w(s)$ is non-increasing in s, differentiable and that $w(s)$ is independent of the file size x. Even with this simple structure the dynamics already lead to non-trivial behavior of the system, as we shall see below. In the concluding remarks we discuss the impact of the assumption that w does not depend on x. (The proofs in this paper do allow for some variation of w as a function of x, but we do not deal with these issues here.) A file of length x is transmitted if and only if

$$w(S(x)) - p(x) \geqslant 0.$$

2.2. *Static user choices*

Now suppose that users can choose between a high-priority and low-priority service, or they can choose not to transmit. We write $x \in H$ if users choose to transmit files of size x with high priority and $x \in L$ if the low-priority service is used. Since not necessarily all files are transmitted, it may be that $\mathbb{P}(x \in H) + \mathbb{P}(x \in L) < 1$.

Thus, service requests that are to be transmitted arrive according to two independent Poisson processes to the two service classes, the arrival rates for the high-priority and low-priority classes being $\lambda_{\mathrm{h}} = \lambda \mathbb{P}(B \in H)$ and $\lambda_{\mathrm{l}} = \lambda \mathbb{P}(x \in L)$, respectively. The mean file size of the users requesting the high-priority service equals

$$f_{\mathrm{h}} := \frac{1}{\mathbb{P}(B \in H)} \int_{x=0}^{\infty} x\, \mathbf{1}_{x \in H}\, \mathrm{d}F(x),$$

and the mean file size of those choosing the low-priority service equals

$$f_{\mathrm{l}} := \frac{1}{\mathbb{P}(B \in L)} \int_{x=0}^{\infty} x\, \mathbf{1}_{x \in L}\, \mathrm{d}F(x),$$

where $\mathbf{1}_{\mathrm{E}}$ is the indicator function, which equals 1 if expression E is true and it equals 0 otherwise. Note that, because of the possibility of balking, in general $\mathbb{E}\, B \geqslant \mathbb{P}(B \in H)\, f_{\mathrm{h}} + \mathbb{P}(B \in L)\, f_{\mathrm{l}}$. The total capacity of the system (for both classes) is again normalized to 1 and we denote the loads on the two service classes by $\rho_{\mathrm{h}} = \lambda_{\mathrm{h}} f_{\mathrm{h}}$ and $\rho_{\mathrm{l}} = \lambda_{\mathrm{l}} f_{\mathrm{l}}$, respectively. In addition we define $\rho_{\mathrm{h}}^{(2)} := \lambda_{\mathrm{h}} f_{\mathrm{h}}^{(2)}$, where $f_{\mathrm{h}}^{(2)}$ denotes the second moment of the file-size distribution of the users that choose the high-priority service:

$$f_{\mathrm{h}}^{(2)} := \frac{1}{\mathbb{P}(B \in H)} \int_{x=0}^{\infty} x^2\, \mathbf{1}_{x \in H}\, \mathrm{d}F(x).$$

We assume that the high-priority class has preemptive priority over the low-priority class. Thus, whenever there is at least one high-priority user active, the service capacity is devoted to the high-priority class only. Within each queue, the service discipline is PS.

We denote the expected delay (transfer time) of a newly arriving high-priority request of size x by $D_{\mathrm{h}}(x)$ and that of a low-priority request of size x by $D_{\mathrm{l}}(x)$. Thus, the stretch of a high-priority file transfer is given by $S_{\mathrm{h}}(x) := D_{\mathrm{h}}(x)/x$ and that of a low-priority transfer by $S_{\mathrm{l}}(x) := D_{\mathrm{l}}(x)/x$. Since the high-priority service does not notice the low-priority service, the stretch in the high-priority queue is again independent of x:

$$S_{\mathrm{h}}(x) \equiv S_{\mathrm{h}} := \frac{1}{1-\rho_{\mathrm{h}}}.$$

For the stretch in the low-priority queue the following asymptotic result was obtained in [20, Theorem 5.6.1]

$$S_{\mathrm{l}}(x) = \frac{1}{1-\rho_{\mathrm{h}}-\rho_{\mathrm{l}}} + \frac{\rho_{\mathrm{h}}^{(2)}/2}{x(1-\rho_{\mathrm{h}}-\rho_{\mathrm{l}})^2} + \mathrm{o}(x^{-1}),$$

as $x \to \infty$. In the case where the low-priority class has an exponential file-size distribution, the $\mathrm{o}(1/x)$ is known explicitly [21]:

$$\begin{aligned} S_{\mathrm{l}}(x) = \frac{1}{1-\rho_{\mathrm{h}}-\rho_{\mathrm{l}}} + \frac{\rho_{\mathrm{h}}^{(2)}}{2x(1-\rho_{\mathrm{h}})^2} & \times \left(1 + \rho_{\mathrm{l}} \times \frac{2(1-\rho_{\mathrm{h}})-\rho_{\mathrm{l}}}{(1-\rho_{\mathrm{h}}-\rho_{\mathrm{l}})^2} \right. \\ & \left. \times (1 - \mathrm{e}^{-(1-\rho_{\mathrm{l}}/(1-\rho_{\mathrm{h}}))x/f_{\mathrm{l}}}) \right). \end{aligned}$$

In the sequel we shall simply ignore the $\mathrm{o}(1/x)$ term in the expression for $S_{\mathrm{l}}(x)$ and assume that

$$S_{\mathrm{l}}(x) = \frac{1}{1-\rho_{\mathrm{h}}-\rho_{\mathrm{l}}} + \frac{\rho_{\mathrm{h}}^{(2)}/2}{x(1-\rho_{\mathrm{h}}-\rho_{\mathrm{l}})^2}, \tag{1}$$

holds with equality, for all x.

Remark 2.1. The results derived later in this paper allow for a more general framework than that described here. The two essential properties of the model that we use in the proofs of our results are: (i) the stretch in the high-priority class depends only on ρ_{h} (and not on the file size) and (ii) the stretch in the low-priority queue is decreasing in the file size.

2.3. Users adjust to prices

Users may submit their transfer requests to be serviced in either service class or refrain from service altogether, depending on the current prices, loads and traffic characteristics in the queues. Suppose the high-priority queue charges a price $xp_{\mathrm{h}}(x)$ for transmitting a file of length x, while the low-priority queue charges a price $xp_{\mathrm{l}}(x)$.

The arrival rates λ_{h} and λ_{l}, the loads ρ_{h} and ρ_{l} as well as the entire file-size distributions

$$F_{\mathrm{h}}(x) := \mathbb{P}(B \leqslant x | B \in H)$$

and

$$F_{\mathrm{l}}(x) := \mathbb{P}(B \leqslant x | B \in L)$$

of the files transmitted with high and low priority (and their means f_{h} and f_{l}) are consequences of the users' choices. A file of length x should be sent to the high-priority queue if both

$$w(S_{\mathrm{h}}) - p_{\mathrm{h}}(x) \geqslant 0 \quad \text{and}$$
$$w(S_{\mathrm{h}}) - p_{\mathrm{h}}(x) \geqslant w(S_{\mathrm{l}}(x)) - p_{\mathrm{l}}(x). \tag{2}$$

It is put in the low-priority queue if

$$w(S_{\mathrm{l}}(x)) - p_{\mathrm{l}}(x) \geqslant 0 \quad \text{and}$$
$$w(S_{\mathrm{l}}(x)) - p_{\mathrm{l}}(x) > w(S_{\mathrm{h}}) - p_{\mathrm{h}}(x), \tag{3}$$

whereas the file is not transmitted if

$$w(S_{\mathrm{h}}) - p_{\mathrm{h}}(x) < 0 \quad \text{and}$$
$$w(S_{\mathrm{l}}(x)) - p_{\mathrm{l}}(x) < 0. \tag{4}$$

The system is in *equilibrium* if these choices are consistent with the definitions of λ_{h}, λ_{l}, ρ_{h}, ρ_{l} and $\rho_{\mathrm{h}}^{(2)}$, i.e., if $x \in H$ if (2) is satisfied, $x \in L$ if (3) holds and $x \notin H \cup L$ in case (4) is true.

3. Revenue maximization

We assume that the operator managing the system aims at maximizing the revenue. Then, for all x, any maximizing price functions $p_{\mathrm{h}}^*(x)$ and $p_{\mathrm{l}}^*(x)$ satisfy one of the following three relations

$$w(S_{\mathrm{l}}(x)) - p_{\mathrm{l}}^*(x) \leqslant w(S_{\mathrm{h}}) - p_{\mathrm{h}}^*(x) = 0, \tag{5}$$

$$w(S_{\mathrm{h}}) - p_{\mathrm{h}}^*(x) < w(S_{\mathrm{l}}(x)) - p_{\mathrm{l}}^*(x) = 0, \tag{6}$$

$$\max\{w(S_{\mathrm{h}}) - p_{\mathrm{h}}^*(x), w(S_{\mathrm{l}}(x - p_{\mathrm{l}}^*(x))\} < 0. \tag{7}$$

These equations are obtained from (2)–(4) by increasing the prices until users are about to change their choices. Relation (5) corresponds to files transmitted using the premium service, (6) to files using the secondary service and (7) to files not transmitted. In all cases, a user with a file of size x is charged exactly what he is willing to pay for the service chosen.

We next determine which of the price functions satisfying the above relations achieve(s) maximum revenue. In fact, the problem is equivalent to finding a revenue maximizing partition $A_{\mathrm{h}} \cup A_{\mathrm{l}} \cup A_0$ of the real line $[0, \infty)$ such that files with sizes $x \in A_{\mathrm{h}}$ (A_{l}) are transmitted with high (low) priority and files with sizes $x \in A_0$ are not transmitted. Indeed, such a partition determines the loads and the service requirement distributions in both queues and, therefore (using (5)–(7)), the price functions: If $x \in A_{\mathrm{h}}$, $p_{\mathrm{h}}^*(x) = w(S_{\mathrm{h}})$ and $p_{\mathrm{l}}^*(x) > w(S_{\mathrm{l}}(x))$. If $x \in A_{\mathrm{l}}$, $p_{\mathrm{h}}^*(x) > w(S_{\mathrm{h}})$ and $p_{\mathrm{l}}^*(x) = w(S_{\mathrm{l}}(x))$. If $x \in A_0$, $p_{\mathrm{h}}^*(x) > w(S_{\mathrm{h}})$ and $p_{\mathrm{l}}^*(x) > w(S_{\mathrm{l}}(x))$.

We state the main result in the following proposition which relies on three lemmas that we shall prove subsequently.

Proposition 3.1. *There exist threshold values* $0 \leqslant t_{\mathrm{h}} \leqslant t_{\mathrm{l}} \leqslant \infty$ *so that the partition*

$$A_{\mathrm{h}}^* \cup A_0^* \cup A_{\mathrm{l}}^*$$

with $A_{\mathrm{h}}^* = [0, t_{\mathrm{h}})$, $A_{\mathrm{l}}^* = [t_{\mathrm{l}}, \infty)$ *and* $A_0^* = [t_{\mathrm{h}}, t_{\mathrm{l}})$ *represents a revenue optimizing Nash equilibrium.*

Proof. Follows from Lemmas 3.2, 3.3 and 3.4. □

Lemma 3.2. *In a revenue maximizing Nash equilibrium with corresponding partition* $A_{\mathrm{h}} \cup A_{\mathrm{l}} \cup A_0$ *it holds that* $x \geqslant y$ *for all* $x \in A_0$ *and* $y \in A_{\mathrm{h}}$.

Proof. Suppose we have a Nash equilibrium represented by the partition $A_{\mathrm{h}} \cup A_{\mathrm{l}} \cup A_0$. Unless the prices satisfy (5)–(7), this partition cannot be optimal. Let us therefore assume that these conditions are satisfied. (Thus, for example, for $x \in A_{\mathrm{h}}$ it holds that $p_{\mathrm{h}}(x) = w(S_{\mathrm{h}})$ and $p_{\mathrm{l}}(x) \geqslant w(S_{\mathrm{l}}(x))$.) We will show that if $x < y$ for $x \in A_0$ and $y \in A_{\mathrm{h}}$, then the partition does not maximize profit.

Let $\varDelta > 0$ and $x > 0$ be fixed and choose $\delta = \delta(\varDelta, x) > 0$ such that

$$\varDelta := \lambda \int_{u=x-\delta}^{x+\delta} u \, \mathrm{d}F(u).$$

This is possible when $\varDelta$ is small enough (at the end we let $\varDelta \to 0$). So we choose δ such that $\varDelta$ is the load associated with the neighborhood $(x - \delta, x + \delta)$.

Suppose $(x - \delta, x + \delta) \subset A_0$ and that we move this neighborhood to A_{h}. (Similar arguments apply if only a left or a right neighborhood of x belongs to A_0; we do not need to consider singletons $\{x\}$ because F is continuous.) The load in the high-priority queue increases to $\rho'_{\mathrm{h}} = \rho_{\mathrm{h}} + \varDelta$ resulting in an increased stretch $S'_{\mathrm{h}} = S_{\mathrm{h}} + \mathrm{O}(\varDelta)$, if $\varDelta \to 0$. Users that used the high-priority service prior to the shift note the increased load ρ'_{h} through the larger stretch S'_{h}. Suppose we reduce the prices in h to $w(S'_{\mathrm{h}})$ so that it is attractive for these users (as well as for the newly switched users) to choose the high-priority service. Thus, the revenue from the "old" users is reduced by $(w(S'_{\mathrm{h}}) - w(S_{\mathrm{h}}))\rho_{\mathrm{h}} = \mathrm{o}(1)$, as $\varDelta \to 0$, while the users that switch generate an increase in revenue of $w(S'_{\mathrm{h}})\varDelta = w(S_{\mathrm{h}})\varDelta + \mathrm{o}(\varDelta)$. Both mutations in revenue do not depend on x.

The users of the low-priority service also experience a larger stretch. Not only because $\rho'_{\mathrm{h}} > \rho_{\mathrm{h}}$, but also $\rho'^{(2)}_{\mathrm{h}} = \rho^{(2)}_{\mathrm{h}} + x\varDelta + \mathrm{o}(\varDelta)$. The load of low-priority file transfers is unaltered ($\rho'_{\mathrm{l}} = \rho_{\mathrm{l}}$). Suppose prices for the low-priority service are also adjusted so that its users neither switch to the high-priority service, nor decide to balk, and become $p'_{\mathrm{l}}(u) = w(S'_{\mathrm{l}}(u))$ for all $u \in A_{\mathrm{l}}$. Observe that, for all $u \in A_{\mathrm{l}}$,

$$S'_{\mathrm{l}}(u) - S_{\mathrm{l}}(u) = \frac{1}{1 - \rho'_{\mathrm{h}} - \rho'_{\mathrm{l}}} + \frac{\rho'^{(2)}_{\mathrm{h}}/2}{u(1 - \rho'_{\mathrm{h}} - \rho'_{\mathrm{l}})^2} - \frac{1}{1 - \rho_{\mathrm{h}} - \rho_{\mathrm{l}}} - \frac{\rho^{(2)}_{\mathrm{h}}/2}{u(1 - \rho_{\mathrm{h}} - \rho_{\mathrm{l}})^2}.$$

Note that the increase in stretch of flows using the low-priority service is more pronounced when x is larger because $\rho'^{(2)}_{\mathrm{h}}$ depends linearly on x. Thus, $S'_{\mathrm{l}}(u) - S_{\mathrm{l}}(u)$ is increasing in x and $\int_{u \in A_{\mathrm{l}}} (w(S_{\mathrm{l}}(u)) - w(S'_{\mathrm{l}}(u)))u \, \mathrm{d}F(u)$, the net decrease in revenue due to low-priority transmissions, is an increasing function of x because $\varDelta$ is fixed.

For fixed $\varDelta$, moving the neighborhood of x from A_0 to A_{h} is profitable if

$$w(S_{\mathrm{h}})\varDelta + \mathrm{o}(\varDelta) > (w(S'_{\mathrm{h}}) - w(S_{\mathrm{h}}) + \mathrm{o}(\varDelta))\rho_{\mathrm{h}} + \int_{u \in A_{\mathrm{l}}} (w(S_{\mathrm{l}}(u)) - w(S'_{\mathrm{l}}(u)))u \, \mathrm{d}F(u). \tag{8}$$

Neglecting the $\mathrm{o}(\varDelta)$ terms (i.e., dividing by $\varDelta$, letting $\varDelta \to 0$ and using that $w(s)$ is differentiable), the right-hand side is increasing in x, thus there is a threshold x' such that moving the neighborhood of x from A_0 to A_{h} is profitable for $x < x'$. Similarly, we can argue that moving a neighborhood of y with load $\varDelta$ from A_{h} to A_0 is profitable for $y > x'$.

Summarizing, if we have a partition with $x \in A_0$ and $y \in A_{\mathrm{h}}$ and $x < y$ then either $y > x'$ or $x < x'$, and thus the partition can be improved so as to increase the revenue. □

Lemma 3.3. *In a revenue maximizing Nash equilibrium with corresponding partition $A_{\mathrm{h}} \cup A_{\mathrm{l}} \cup A_0$ it holds that $x \leqslant y$ for all $x \in A_0$ and $y \in A_{\mathrm{l}}$.*

Proof. The proof follows along the same lines as that of Lemma 3.2. Again, suppose we have a partition with corresponding maximal price functions. Let $\varDelta > 0$ and $x > 0$ be fixed and choose $\delta = \delta(x, \varDelta) > 0$ such that

$$\varDelta = \lambda \int_{u=x-\delta}^{x+\delta} u \, \mathrm{d}F(u)$$

is the load associated with the neighborhood $(x - \delta, x + \delta)$. Suppose that this neighborhood is in A_0 and that we move it to A_{l}. Users of the high-priority service do not notice this shift. The users of the low-priority service, however, are worse off since the load in their service class increases to $\rho'_{\mathrm{l}} = \rho_{\mathrm{l}} + \varDelta$. Suppose that we adjust prices so as to compensate for this loss in utility, thus decreasing system revenue. The net decrease in revenue only depends on $\varDelta$ (not on x). The users that do switch generate an increase in revenue of

$$w(S_{\mathrm{l}}(x))\varDelta + \mathrm{o}(\varDelta),$$

which is non-increasing in x. Thus, (for fixed $\varDelta$) there is a threshold x' such that moving the neighborhood of x from A_0 to A_l is profitable whenever $x > x'$. The proof can now be completed as in Lemma 3.2. □

For revenue optimizing Nash equilibria with $A_0 \neq \emptyset$, Lemmas 3.2 and 3.3 imply the following lemma. In order to cover the case where possibly $A_0 = \emptyset$, we include an independent proof.

Lemma 3.4. *In a revenue maximizing Nash equilibrium with corresponding partition $A_\mathrm{h} \cup A_\mathrm{l} \cup A_0$ it holds that $x \geqslant y$ for all $x \in A_\mathrm{l}$ and $y \in A_\mathrm{h}$.*

Proof. Similar to the proofs of Lemmas 3.2 and 3.3 let $\varDelta > 0$ and $x > 0$ be fixed, with $(x-\delta, x+\delta) \subset A_\mathrm{l}$ and

$$\varDelta = \lambda \int_{u=x-\delta}^{x+\delta} u \, \mathrm{d}F(u).$$

We move the neighborhood $(x-\delta, x+\delta)$ to A_h. Then other users of the high-priority service experience an increase in load from ρ_h to $\rho'_\mathrm{h} = \rho_\mathrm{h} + \varDelta$; in order to keep these users in the high-priority service class, the prices must be reduced, thus reducing the revenue from these users. This decrease in revenue only depends on $\varDelta$ and not on x.

Users of low-priority transmissions also experience a larger stretch since $\rho_\mathrm{h}^{(2)}$ increases to ${\rho'}_\mathrm{h}^{(2)} = x\varDelta + \mathrm{o}(\varDelta)$. (Notice that the total load remains unchanged: $\rho_\mathrm{h} + \rho_\mathrm{l} = \rho_\mathrm{h} + \varDelta + \rho_\mathrm{l} - \varDelta = \rho'_\mathrm{h} + \rho'_\mathrm{l}$.) Thus, the net decrease in revenue due to users that do not switch is more pronounced for larger x (for fixed $\varDelta$). The users that do switch generate an increase in revenue of

$$w(S_\mathrm{h})\varDelta - w(S_\mathrm{l}(x))\varDelta + \mathrm{o}(\varDelta),$$

which is smaller for larger values of x. Thus, (for fixed $\varDelta$) there is a threshold x' such that moving the neighborhood of x from A_l to A_h is profitable for $x < x'$. The proof can now be completed as in the proofs of Lemmas 3.2 and 3.3. □

4. Numerical example

The value of the offered load $\rho \equiv \lambda \mathbb{E}B$ has a significant impact on the nature of the revenue-maximizing solution. For small values of the offered load ρ, it can be expected that it is beneficial for the provider to serve all arriving flows, a part in the high-priority class, and a part in the low-priority class (hence $t_\mathrm{h} = t_\mathrm{l}$). With increasing (offered) load, our results indicate that a part of the flows will not join. In this section we illustrate this effect through a numerical example. We also assess the impact of the flow-size distribution $F(\cdot)$.

In this example we choose the willingness to pay as $w(s) = (a-s)/a$, for some fixed number $a > 0$, so that stretch and price are directly interchangeable. The function $w(s)$ has been normalized so that users are willing to pay 1 price unit per bit if the transmission has zero delay. A large value of a indicates a high tolerance to delay (the users are willing to pay a positive amount as long as $s < a$), which may be reasonable for data transfer. For time critical transmissions, a small value of a is more appropriate. (In the case of more delay-sensitive traffic, it would be more reasonable if $w(s)$ were rather flat for small values of s followed by a steeper fall off; a simple quadratic form such as $w(s) = 1 - bs^2$ could capture this, but we do not include this in our numerical illustration here.) In our example we take $a = 2$. The results above indicate that A_h is of the form $[0, t_\mathrm{h}]$, and A_l of the form $[t_\mathrm{l}, \infty)$, for $0 \leqslant t_\mathrm{h} \leqslant t_\mathrm{l} < \infty$.

Now we can compute the operator's revenue $\mathbf{R}$ as a function of the thresholds t_h and t_l:

$$\begin{aligned}\mathbf{R}(t_\mathrm{h}, t_\mathrm{l}) &= \rho_\mathrm{h} w(S_\mathrm{h}) + \lambda \int_{t_\mathrm{l}}^{\infty} x w(S_\mathrm{l}(x)) \, \mathrm{d}F(x) \\ &= \frac{1}{a}\left(\rho_\mathrm{h}(a - S_\mathrm{h}(\rho_\mathrm{h})) + \lambda\rho_\mathrm{l}\left(a - \frac{1}{1-\rho_\mathrm{h}-\rho_\mathrm{l}}\right)\right. \\ &\quad \left. - \lambda \frac{\rho_\mathrm{h}^{(2)}(1 - F(t_\mathrm{l}))}{2(1-\rho_\mathrm{h}-\rho_\mathrm{l})^2}\right).\end{aligned}$$

This revenue should be maximized over all $0 \leqslant t_\mathrm{h} \leqslant t_\mathrm{l} < \infty$. We have performed this optimization for two (unit-mean) flow-size distributions, exponential and Pareto: $F_{\exp}(x) = 1 - e^{-x}$ and $F_{\mathrm{Par}}(x) = 1 - 1/(x+1)^2$. Notice that the Pareto distribution has a much heavier tail than the exponential distribution; in fact, the Pareto distribution chosen here has infinite variance, motivated by the large variance in file sizes observed in practice.

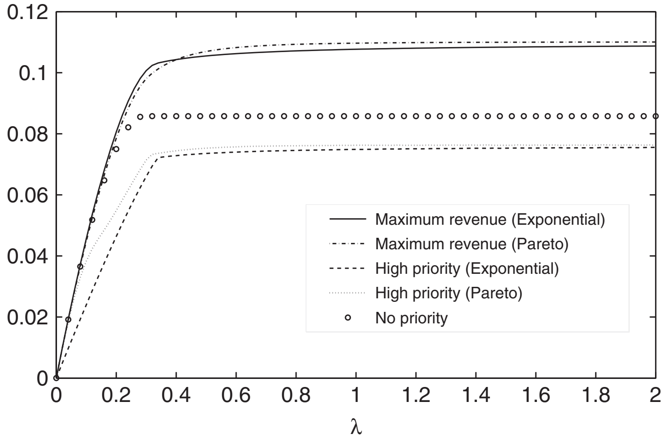

Fig. 1. Maximum revenue and contribution of the high-priority class as a function of λ.

We compare the revenues with those in the model *without* priority. Then the provider wishes to solve

$$\max_{A\subseteq[0,\infty)} \rho_A w\left(\frac{1}{1-\rho_A}\right),$$

where $\rho_A := \lambda\int_{x\in A} x\,\mathrm{d}F(x)$. It can be easily verified that the provider's optimum load is $\rho^\star := 1-1/\sqrt{a}$, irrespective of the flow-size distribution. If the offered load $\rho \equiv \lambda\mathbb{E}B$ is smaller than $\rho^\star$, then all flows will be served, and the (normalized) revenue is

$$\frac{\rho}{a}\left(a-\frac{1}{1-\rho}\right).$$

If $\rho \geqslant \rho^\star$ a part of the flows balks, yielding a revenue $1/a(\sqrt{a}-1)^2 = (a-2\sqrt{a}+1)/a$. For $a=2$, we find that $\rho^\star \approx 0.2929$ and for all $\rho \geqslant \rho^\star$ we obtain a revenue of $\frac{3}{2}-\sqrt{2} \approx 0.0858$.

In Fig. 1 we plot the maximum revenue in the non-prioritized system as well as the contributions of both priority classes to the maximum revenue in the prioritized system, as a function of the arrival rate. For the prioritized system we do so for both exponential and Pareto distributed flow sizes. In Fig. 2 we plot the corresponding optimal thresholds in the prioritized system. We observe that for both file-size distributions it is optimal to carry the complete load as long as the arrival rate is not too large. For exponentially distributed file sizes, "medium"-sized flows are rejected for $\lambda \geqslant 0.34$, for the Pareto distribution the turning point is a bit smaller at about $\lambda = 0.31$.

However, there are also significant differences between the exponential case and the Pareto case. Observe that in the exponential case both classes significantly contribute for all values of λ, whereas in the Pareto case, only the high-priority queue contributes to the maximum revenue for small arrival rates. Notice that this entails that there is no gain from the prioritization in case of small arrival rates when file sizes have a Pareto distribution. We indeed see from Fig. 1 that in those cases the non-prioritized system achieves an equally high maximum revenue. For exponential file sizes, the low-priority class offers the system manager the potential for efficient resource usage at all values of the offered traffic. We further observe that for larger traffic loads, the system specializes in very short and very long flows, with a growing range of flow sizes that will not be transmitted.

From the users' perspective, the acceptance rate and the corresponding carried loads are more relevant measures of performance. These are depicted in Figs. 3 and 4. With the Pareto distribution, for all values of λ, virtually all accepted transactions are carried in the high-priority queue, but (due to the heavy-tailed distribution) the load in the low-priority queue is nevertheless significant. When the file-size distribution is exponential, both the numbers of transactions as well

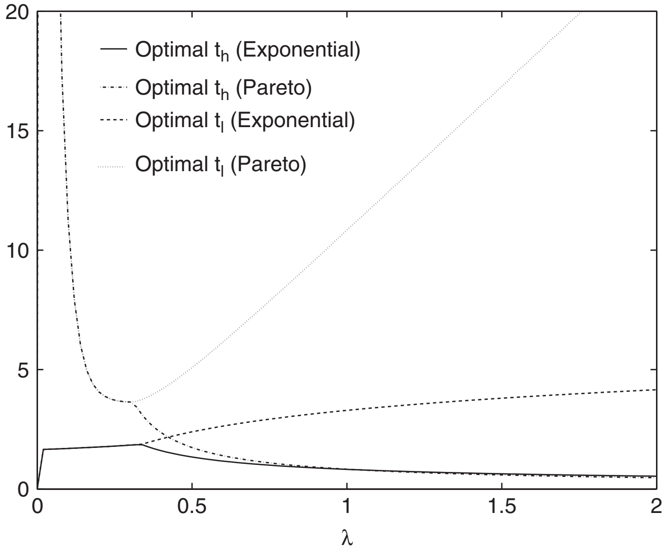

Fig. 2. Optimal thresholds as functions of λ.

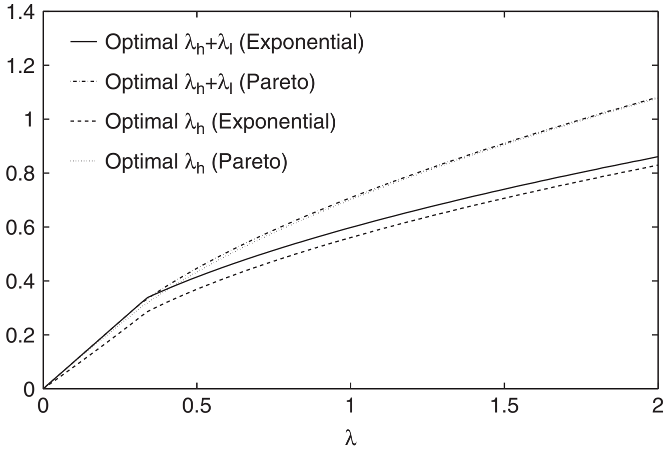

Fig. 3. Accepted rates of transactions.

as the corresponding loads on both queues are significant for all values of λ.

5. Concluding remarks

Even under the simplistic assumption that the willingness to pay function w does not depend on x, we have shown non-trivial behavior of the system. In general, for file transfers it is reasonable to assume that if $w(x, s)$ does depend on x, then it is non-increasing in x, since if there exists an x such that $w(u, s) \leqslant w(x, s)$ for all $u \leqslant x$, then a file of size x could be split in smaller ones, thus reducing the total charge. For a related discussion see [19].

Our assumption on w can be relaxed without affecting the proofs, as long as (8) holds. Imposing this condition (8) a priori, however, seems rather unnatural and therefore we chose not to aim

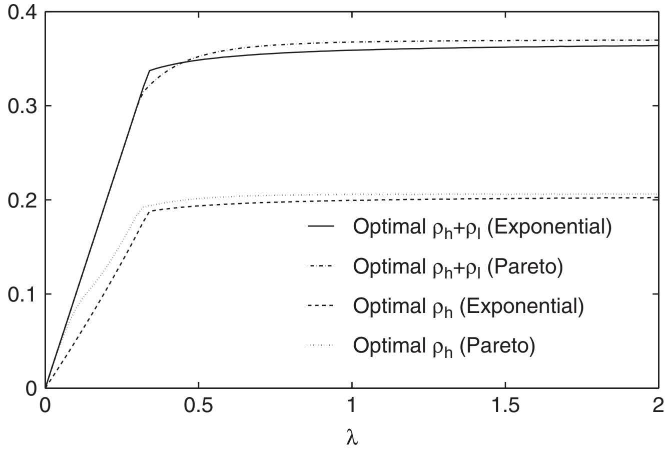

Fig. 4. Optimal loads as functions of λ.

for the largest generality under which our results hold, but rather focus on the qualitative behavior of the system for a particular, yet not unreasonable, choice for w.

We already argued in the introduction that we expect similar results to hold when there are more than two priority classes. In the revenue optimizing equilibrium, the classes with higher priority will attract shorter flows, with possibly non-empty sets of flow sizes that are not transmitted because they 'fall between' two adjacent priority classes.

We further expect that assuming that the high-priority class consists of *streaming* users (who use a fixed amount of bandwidth while in the system, unlike elastic users), will not qualitatively impact the results. Crucial properties in our analysis are that the highest priority service is not affected by other classes, and that it is only affected by the flows in the same class through the class load. However, for mimicking the results of the present paper, we need a result on the stretch of the low-priority files, to replace the estimate (1). Such a result is unfortunately not available for the situation in which the high-priority class consists of streaming flows.

Acknowledgment

This work was carried out in the project EQUANET, supported by the Dutch Department of Economic Affairs, via its agency SENTER/NOVEM. The research of the second and third author was also funded by the Dutch bsik/BRICKS project.

References

[1] K. Avrachenkov, U. Ayesta, P. Brown, E. Nyberg, Differentiation between short and long TCP flows: predictability of the response time, in: Proceedings of INFOCOM 2004, Hong Kong, 2004.

[2] M. Crovella, A. Bestavros, Self-similarity in World Wide Web traffic: evidence and possible causes, IEEE/ACM Trans. Networking 5 (1997) 362–373.

[3] S. Deb, A. Ganesh, P. Key, Resource allocation with persistent and transient flows, Proceedings Networking 2002, Lecture Notes in Computer Science, vol. 2345, 2002, pp. 455–466.

[4] H. Feng, V. Misra, Mixed scheduling disciplines for network flows, in: Proceedings of the Fifth Workshop on Mathematical Performance in Model Analogy. San Diego, CA, USA, 2003.

[5] R. Gibbens, R. Mason, R. Steinberg, Internet service classes under competition, IEEE J. Sel. Areas Commun. 18 (2000) 2490–2498.

[6] L. Guo, I. Matta, The war between mice and elephants, Proceedings of the Ninth IEEE International Conference on Network Protocols ICNP 2001, 2001.

[7] A. Gupta, D. Stahl, A. Whinston, A stochastic equilibrium model of Internet pricing, J. Econ. Dyn. Control 21 (1997) 697–722.

[8] M. Harchol-Balter, B. Schroeder, N. Bansal, M. Agrawal, Size-based scheduling to improve web performance, Trans. Comput. Syst. 21, 2003.

[9] R. Hassin, M. Haviv, To Queue or not to Queue—Equilibrium Behavior in Queueing Systems, Kluwer, Boston, 2003.

[10] R. Hassin, M. Haviv, Who should be given priority in a queue?, Oper. Res. Lett. 34 (2006) 191–198.

[11] L. Kleinrock, Queueing Systems, vol. II: Computer Applications, Wiley, New York, 1976.

[12] L. Kleinrock, R. Muntz, Processor-sharing queueing models of mixed scheduling disciplines for time-shared systems, J. ACM 19 (1972) 464–482.

[13] J. MacKie-Mason, H. Varian, Pricing congestible network resources, IEEE J. Sel. Areas Commun. 13 (1995) 1141–1149.

[15] L. Massoulié, J. Roberts, Bandwidth sharing: objectives and algorithms, in: Proceedings INFOCOM 1999, New York, NY, USA, 1999, pp. 1395–1403.

[16] L. McKnight, J. Bailey (Eds.), Internet Economics, MIT Press, Cambridge, MA, USA, 1997.

[17] H. Mendelson, Pricing computer services: queueing effects, Commun. ACM 28 (1985) 312–321.

[18] H. Mendelson, S. Whang, Optimal incentive-compatible priority pricing for the M/M/1 queue, Oper. Res. 38 (1990) 870–883.

[19] H. Moulin, Split-proof probabilistic scheduling, pre-print, Rice University, September, 2004.

[20] R. Núñez-Queija, Processor-sharing models for integrated-services networks, Ph.D. Thesis, Eindhoven University of Technology, ISBN 90-646-4667-8, 2000.

[21] R. Núñez-Queija, Sojourn times in a processor-sharing queue with service interruptions, Queueing Syst. 34 (2000) 351–386.

[22] A. Odlyzko, Paris metro pricing: the minimalist differentiated services solution, in: Proceedings Seventh International Workshop on Quality of Service (IWQoS'99), IEEE, 1999, pp. 159–161.

[23] R. Righter, J.G. Shanthikumar, G. Yamazaki, On extremal service disciplines in single-stage queueing systems, J. Appl. Probab. 27 (1990) 409–416.

[24] M. Sakata, S. Noguchi, J. Oizumi, An analysis of the M/G/1 queue under round-robin scheduling, Oper. Res. 19 (1971) 371–385.

[25] L. Schrage, L. Miller, The queue M/G/1 with the shortest remaining processing time discipline, Oper. Res. 14 (1966) 670–684.

[26] B. Schroeder, M. Harchol-Balter, Web servers under overload: how scheduling can help, Proceedings ITC 18, Berlin, Germany, 2003, pp. 171–180.

[27] S. Stidham Jr., Pricing and congestion management in a network with heterogeneous users, IEEE Trans. Autom. Control 49 (2004) 976–980.

[28] A. Wierman, M. Harchol-Balter, Classifying scheduling policies with respect to Unfairness in an M/GI/1, Proceedings ACM Sigmetrics, 2003.