The use of relative priorities in optimizing the performance of a queueing system

Refael Hassin[1] Justo Puerto[2] Francisco R. Fernández [3]

November 15, 2007

Abstract

Relative priorities in an n-class queueing system can reduce server and customer costs. This property is demonstrated in a single server Markovian model where the goal is to minimize a non-linear cost function of class expected waiting times. Special attention is given to minimizing server's costs when the expected waiting time of each class is restricted.

1 Introduction

Control of queueing systems to maximize profits or welfare has been the subject of numerous papers. The common methods are by setting adequate price and priority regimes. (See [10] for a survey of such models.) In other cases, the service provider also sets and advertises waiting time standards [1]. The common priority regime is that of (preemptive or non-preemptive) *absolute priorities*, where the customer classes are ranked and customers are called to be served according to this order.

There is a voluminous literature analyzing and comparing different priority disciplines, see for instance the survey texts by Gelenbe and Mitrani [8] and Kleinrock [13]. A notable generalization of this concept was offered by Federgruen and Groenevelt [7] who considered work conserving priority rules. For each rule there corresponds a *performance vector* giving the expected waiting time of each customer class under the given rule. The *performance space* consists of the collection of performance vectors achievable by the available rules. Federgruen and Groenevelt showed that the performance space is the convex hull of the points corresponding to the regimes. Thus, each point in this polyhedron is achievable. However, the natural way of obtaining a given point in the performance space is, for example, by randomizing between a set of absolute priority rules, assuming that the outcome of this randomization can be hidden from the customers. The latter condition may often be hard to implement.

For a linear objective function of the system, that depends on the performance vector, there is an optimal extreme point rule, in absolute priorities. For other functions this is not true, and therefore it is of interest to identify technically feasible priority rules that optimize a nonlinear objective over the performance space.

We consider an alternative approach, that of *relative priorities*, where the priority given to a class also depends on state variables associated with other classes. We demonstrate several new possible uses of such regimes. In particular, we show that every point in the performance space can be achieved by a suitable choice of relative priorities. Thus, we offer a new method for optimizing nonlinear system objective functions without the need to conceal from the customers the details of the priority rule.

[1] Department of Statistics and Operations Research, Tel Aviv University, Tel Aviv 69987, Israel. Research supported by Israel Science Foundation Grant no. 237/02. Email: hassin@post.tau.ac.il

[2] Department of Statistics and Operations Research, University of Seville. 41012 Sevilla, Spain. Research partially supported by Spanish Ministry of Science and Technology grant no. BFM2004:0909, P06-FQM-01366 and MTM2007-67433-C02-01. E-mail: puerto@us.es

[3] Department of Statistics and Operations Research, University of Seville. 41012 Sevilla, Spain. Research partially supported by HI2003:0189. E-mail: fernande@us.es

We consider a single server and several customers. Customer i submits jobs to be processed by the server according to a Poisson process with rate λ_i. The service rate is exponential with mean $1/\mu$. A function $f(W_1, \dots, W_n)$ gives the cost incurred by the system when i-jobs have expected waiting time of W_i, $i = 1, \dots, n$. (By waiting time we mean the time in the system including in service: often called sojourn time.) We also consider a variation of this model where the service rate is a decision variable, and the cost function is extended to include the cost associated with the chosen service rate. In both cases we give conditions under which relative priorities reduce costs.

We elaborate on a special case of the above model, where customer i requires that the expected time his jobs stay in the system is bounded by a constant t_i. The server is free to choose the service rate μ and a priority rule. The server incurs a cost $C(\mu)$ per unit of time if the chosen service rate is μ. The function C is monotone non-decreasing. We investigate the optimal choices to be made by the server, and show that the server can profit by using relative priorities.

We consider the priority scheme called *discriminatory processor sharing* (DPS). Under this model there exist nonnegative parameters $x_i \in (0,1)$, $\sum_{i=1}^n x_i = 1$ representing *relative priority* of customers of the classes. If n_i customers are present in the system, $i = 1, \dots, n$, an i-customer receives a fraction $x_i \left(\sum_{i=1}^n n_i x_i\right)^{-1}$ of the service capacity. In particular, the total capacity dedicated to class-i is $n_i x_i \left(\sum_{i=1}^n n_i x_i\right)^{-1}$. Of course, the limit case when $x_i \to 1$ means that the class i obtains absolute priority.

The DPS discipline is used in several queueing models in the computer science and communication literature. In these cases firms cater to multiple customer classes or market segments with the help of shared service facilities or processes, so as to exploit pooling benefits. Different customer classes typically have rather disparate sensitivities to the delays encountered. Conversely, from the firm's perspective it is vital to offer differentiated levels of service to different customer classes so as to maximize (long run) profits. In many service industries, waiting time standards are used as a primary advertised competitive instrument. For example, most major electronic brokerage firms, (e.g., Ameritrade, Fidelity, E-trade) prominently feature the average or median execution speed per transaction which is monitored by independent firms. Thus, in order to improve waiting time standards often firms segment their costumers in classes and some firms go as far as to provide an individual execution time score card as part of the customer's personal account statements. [2, 3, 12, 13, 17, 18].

Clearly, DPS gives more options than can be achieved by absolute priorities, and one may claim that it is expected that by applying DPS a server should be able to achieve better performance or profit than otherwise. However, at least in one notable case this assumption turns to be false. Hassin and Haviv [11] considered two customer classes and a single server who sets both prices and relative priority. They observed that it follows from Mendelson and Whang [15] that when the server is not restricted in choosing these variables, there exists an optimal solution with absolute priorities and thus the application of DPS doesn't improve the welfare achieved by the system. However, they also showed that if the server is restricted to a given set of prices, or if the server must set a common price to both classes, then relative priorities may be used to increase profits. Thus, it is a question of interest to identify other settings where the use of relative priorities can be helpful.

In Section 2 we analyze how to reduce system costs by using DPS as opposed to the use of absolute priorities. We give conditions that ensure, for a given cost function, when DPS outperforms FCFS. Section 3 considers a model where each class fixes its aspiration level on the waiting time and the problem is to ensure these levels at a minimum service rate. (Here the customers are those who set the waiting time standards, and the firm adopts itself to minimize its costs, whereas in [1] the standards are choice variables set by the firm to maximize its profits.) We provide explicit forms for the service rate requirements under different priority regimes: FCFS, absolute preemptive priorities and DPS. The main results proved in this section are: 1) A comparison of service rate requirement under different priority regimes; 2) A general result that characterizes the existence of a DPS policy satisfying given aspiration levels for any number of classes; 3) For $n = 2$ and any given aspiration levels $t_1,\ t_2$, we explicitly determine the optimal priority parameters minimizing the service rate under DPS; 4) We show that for $n = 2$, using DPS improves the service rate regarding the service rate under FCFS, whenever $t_1 \neq t_2$.

2 Optimizing the cost of the system using DPS

Let x_i denote the relative priority given to the ith-class. The problem is:

$$\min_{x \in S_n} f(W_1, \ldots, W_n), \tag{1}$$

where f is a monotone nondecreasing function of its arguments and $S_n = \{x \in \mathbb{R}^n : \sum_{i=1}^n x_i = 1,\ x_i \geq 0, \forall i\}$. Note that although x does not appear explicitly in the function to be minimized the expected waiting times $W_i,\ i = 1, \ldots, n$ depends on the relative priority x_i given to the i-th class. At times, when it is necessary to understand the problem, we will make explicit the dependence of the expected waiting times on the different parameters.

2.1 The achievable waiting times

To investigate *qualitative* properties of this problem we proceed to obtain the functional dependence of $W_i\ i = 1, \ldots, n$. A *mixing priority discipline* consists of multiplexing a finite set of priority disciplines in such a way that each of them will operate during a desired percentage of time.

Denote by $\Pi(N)$ the set of permutations of the finite set $N = \{1, \ldots, n\}$. Take $\pi \in \Pi(N)$ to be an ordering of the n classes. Here, $\pi(i)$ represents the position which has been assigned to class i. The smaller the position index, the higher the priority associated to the class. We denote by W_i^π the expected waiting time in the system for class i under π. It is well-known (see, for instance Gross and Harris (1998)) that for $\mu > \lambda := \sum_{i=1}^n \lambda_i$, the value for a $M/M/1$ system is:

$$W_i^\pi = \frac{\mu}{(\mu - \sum_{j:\pi(j)<\pi(i)} \lambda_j)(\mu - \sum_{j:\pi(j)\leq\pi(i)} \lambda_j)}.$$

We denote by W^π the vector whose coordinates are given by $W_i^\pi\ i = 1, \ldots, n$, and

$$\mathcal{F}(N) = \operatorname{conv}\{W^\pi \in \mathbb{R}^n : \pi \in \Pi(N)\}.$$

The following theorem states a geometrical characterization of the performance space by the family of DPS policies when the number of classes is at least three ($n > 2$).

Theorem 2.1 *The performance space achievable by the family of DPS policies coincides with the relative interior of $\mathcal{F}(N)$. This set is contained in a hyperplane of $\mathbb{R}^n$.*

Proof: It is known (see [6, Theorem 2]) that the entire set of performance waiting time vectors that are achievable by some scheduling strategy coincides with $\mathcal{F}(N)$.[4] Moreover, according to

[4] A scheduling strategy is the specification of the order in which the customers are served, with the only restriction that sequencing decisions are not based on advanced knowledge of remaining service times.

[16, Theorem 3], DPS policies are almost complete with respect to the waiting time vectors of scheduling strategies:[5] This implies that the performance space achievable by DPS policies is $\mathcal{F}(N)$ without its boundary.

Since DPS strategies are work conserving and do not use advance information about individual service times, their achievable waiting times fulfill Kleinrock's conservation law:

$$\sum_{i=1}^{n} \rho_i W_i = \frac{1}{1-\rho} \sum_{k=1}^{n} \frac{\lambda_k}{\mu^2}, \tag{2}$$

where $\rho_i = \lambda_i/\mu$ and $\rho = \lambda/\mu$. Hence, any achievable waiting time vector by DPS policies must be included in the hyperplane defined by this law. ■

The above result describes the geometry of the performance space for $n > 2$. The case $n = 2$ is slightly simpler since this is the unique case where the extreme preemptive strategies $(1, 2)$ and $(2, 1)$[6] coincide with DPS policies $(1, 0)$ and $(0, 1)$[7], respectively. Hence, the performance space achievable by DPS policies coincide with $\mathcal{F}(\{1, 2\})$.

For the case of two priority classes, after some algebra, the expression (2) results in:

$$AW_1 + BW_2 - D = 0,$$

where $A = \lambda_1(\mu - \lambda)$, $B = \lambda_2(\mu - \lambda)$, and $D = \lambda$.

The bounds on W_1 and W_2 are obtained by setting $x_1 = 0, 1$. The performance space, for a given μ, is given in Corollary 2.2 and illustrated in Figure 1.

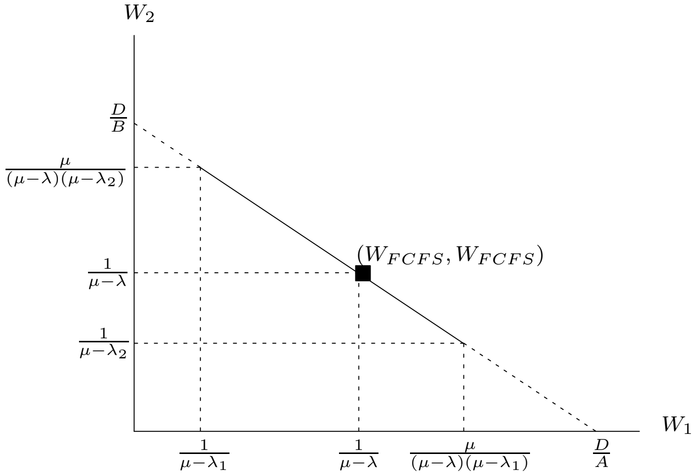
W_2
$\frac{D}{B}$
$\frac{\mu}{(\mu-\lambda)(\mu-\lambda_2)}$
(W_{FCFS}, W_{FCFS})
$\frac{1}{\mu-\lambda}$
$\frac{1}{\mu-\lambda_2}$
W_1
$\frac{1}{\mu-\lambda_1}$
$\frac{1}{\mu-\lambda}$
$\frac{\mu}{(\mu-\lambda)(\mu-\lambda_1)}$
$\frac{D}{A}$

Figure 1: The performance space for $n = 2$

Corollary 2.2 *For any fixed μ, the performance space is a segment in the plane (W_1, W_2) with extreme points $[LO(\mu), UP(\mu)]$, where*

$$LO(\mu) = \left(\frac{1}{\mu - \lambda_1}, \frac{\mu}{(\mu - \lambda)(\mu - \lambda_1)} \right), \qquad \lambda < \mu < +\infty,$$

[5] A family of policies Ψ is almost complete for a given set of performance vectors H whenever H_Ψ, the set of performance vectors achievable by policies in Ψ, satisfies that H_Ψ equals H without its boundary.

[6] The standard notation for preemptive strategies specifies the permutation which gives the preemption sequence on the different classes. Thus, $(2, 1)$ means that any job of class 2 will be completed before any job of class 1.

[7] The notation for DPS policies gives in the i-th coordinate the relative probability assigned to class i.

and

$$UP(\mu) = \left(\frac{\mu}{(\mu-\lambda)(\mu-\lambda_2)}, \frac{1}{\mu-\lambda_2} \right), \qquad \lambda < \mu < +\infty.$$

Computing the performance space for a given DPS policy is in general a hard problem. To date, there exists a closed formula only for the case of two priority classes. The following result is due to Fayolle, Mitrani and Iasnogorodski [9]. Let $\lambda = \lambda_1 + \lambda_2$ and $\Lambda = \lambda_1 x_1 + \lambda_2 x_2$, then

$$W_i = \frac{1}{\mu-\lambda} \frac{\mu - \lambda x_i}{\mu - \Lambda}, \qquad i = 1, 2. \tag{3}$$

It is of interest to compare the waiting times under DPS with $W_{FCFS} = \frac{1}{\mu-\lambda}$ obtained under the First-Come First-Served (FCFS) discipline. Inserting $x_1 = \frac{1}{2}$ in (3) we obtain that $\Lambda = \lambda$ and $W_1 = W_2 = W_{FCFS}$. Therefore, the best result obtained under DPS is at least as good as that obtained under FCFS. The point $(W_1, W_2) = (W_{FCFS}, W_{FCFS})$ is marked in Figure 1.

2.2 Optimal DPS policies

Using the characterization in Theorem 2.1 for $n > 2$, Problem (1) can be rewritten as:

$$\begin{array}{ll} \min & f(W_1, \ldots, W_n) \\ \text{s.t.} & \\ & \displaystyle\sum_{\pi \in \Pi(N)} \alpha_\pi = 1 \\ & \displaystyle W_i - \sum_{\pi \in \Pi(N)} \alpha_\pi W_i^\pi = 0, \quad i = 1, \ldots, n \\ & \alpha_\pi \geq 0, \forall\, \pi \in \Pi(n). \end{array} \tag{4}$$

For the Markovian $M/M/1$ system any feasible solution of (4) must satisfy Kleinrock's conservation law (2). The linear dependence in (2) implies that any feasible solution of (4) can be represented by at most n out of the $n!$ α-coefficients. Moreover, any solution that lies in the relative boundary of $\mathcal{F}(N)$ can be represented by at most $(n-1)$ non-null α-coefficients. These relative boundary points cannot be properly achieved by DPS policies. (But they can be arbitrarily approximated up to any given accuracy.)

In the case of two priority classes we can give a more accurate answer. If the optimum in Problem (1) is not attained at the extreme points of the interval then there exists a DPS policy that outperforms the absolute priorities. Therefore natural candidates to have optimal solutions in DPS policies are convex cost functions (and certainly concave functions never give optimal solutions in relative priorities).

Some interesting particular instances of the above result are given below.

1. If $f(W_1, W_2) = C_1 W_1 + C_2 W_2$ then there is always an optimal solution in absolute priorities. In addition, only if $\frac{C_1}{C_2} = \frac{A}{B}$ there also exist solutions in non absolute priorities. In fact in this case any $x_1 \in [0, 1]$ is an optimal solution. (See [13] to find classical examples of linear objective functions in the control of queues.)

2. Suppose that $f(W_1, W_2) = \max\{C_1 W_1, C_2 W_2\}$, $C_i > 0$, $i = 1, 2$. Usage of this objective function is justified when the server compensates users according to worst case performance, as for instance in emergency systems. Then:

   (a) If $\frac{C_1}{C_2} \geq \frac{1}{1-\rho}$ then the unique optimal solution is $x_1 = 1$.

(b) If $\frac{C_1}{C_2} \leq 1 - \rho$ then the unique optimal solution is $x_1 = 0$.

(c) If $1 - \rho < \frac{C_1}{C_2} < \frac{1}{1-\rho}$ then there is a unique optimal solution at some $x_1 \in (0, 1)$. This value of x_1 solves the following two equations: $AW_1 + BW_2 = D$ and $C_1W_1 = C_2W_2$.

2.3 The problem with variable μ

Once we have analyzed the optimization problem with fixed μ we focus on the problem with variable μ. With two priority classes, the problem is:

$$\min_{\substack{x_1 \in [0,1] \\ \lambda < \mu < +\infty}} f(\mu, W_1, W_2).$$

Figure 2 represents the domain of (W_1, W_2) for different values of μ. In particular the two curves are the geometrical loci of the extreme points of the segments $[LO(\mu), UP(\mu)]$ as a function of μ from $\mu = 3.5, \ldots, 10$.

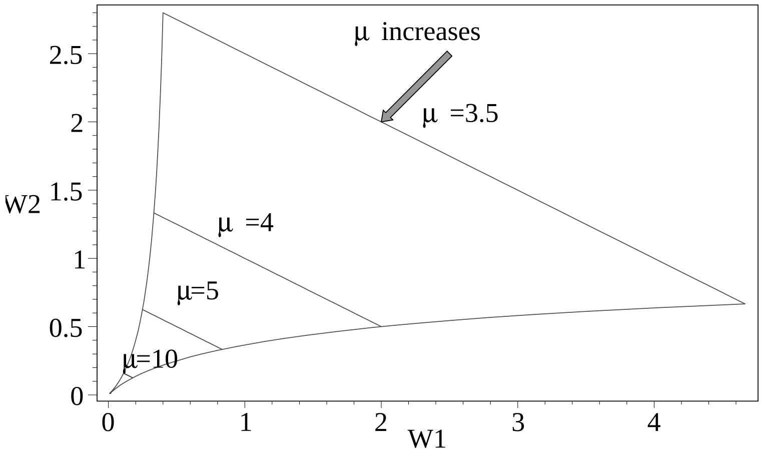

Figure 2: W_1 and W_2 as a function of μ for $\lambda_1 = 1$ and $\lambda_2 = 2$.

Are relative priorities also worth using if μ is a decision variable in the problem? The answer depends on the form of the cost function to be considered. A way to test the better performance of DPS is to check that its behavior outperforms the one in absolute priorities for any feasible μ value. Of course this is only a sufficient condition. Nevertheless, this argument can be applied in particular for $f(\mu, W_1, W_2) = C(\mu) + \max\{C_1W_1, C_2W_2\}$. For this cost function we always have that if

$$1 - \rho < \frac{C_1}{C_2} < \frac{1}{1 - \rho}, \quad \forall \mu,$$

then the optimal solution must be in non absolute priorities since it is the case for any μ. In particular, this condition always holds when $C_1 = C_2$. Therefore, DPS is worth using.

3 The aspiration problem: Minimizing the service rate

The goal of this section is to minimize the necessary service rate to ensure given aspiration levels t_i, $i = 1, \dots, n$ on the waiting times (of the different classes). Since improving service rate is not cost free, our goal induces a trade-off that should be solved up to optimality.

We assume that parameters t_i, $i = 1, \dots, n$ are given. Therefore, this induces the following cost function $f(W_1, \dots, W_n) = 0$ if and only if $W_i \le t_i$, for all $i = 1, \dots, n$. Otherwise $f(W_1, \dots, W_n) = \infty$. Clearly f is convex. Our goal is to compare service rate requirements under different priority regimes: FCFS, absolute preemptive priorities, and DPS.

Suppose first that the queue discipline is FCFS. The system's requirement is now $\frac{1}{\mu-\lambda} \le \min\{t_i : i = 1, \dots, n\}$, and the minimum service rate that satisfies these requirements is

$$\mu_{\text{FCFS}} = \lambda + \frac{1}{\max\{t_i : i = 1, \dots, n\}}. \tag{5}$$

To characterize the optimal service rate if we use absolute preemptive priorities, denote $a^i_\pi = \sum_{j:\pi(j)<\pi(i)} \lambda_j$ and $b^i_\pi = \sum_{j:\pi(j)\le\pi(i)} \lambda_j$ for $i = 1, \dots, n$, $\pi \in \Pi(n)$. In this case we look for the smallest $\mu > \sum_{i=1}^n \lambda_i$ that satisfies, for some permutation π, the following set of inequalities:

$$\frac{\mu}{(\mu - a^i_\pi)(\mu - b^i_\pi)} \le t_i, \qquad \forall i = 1, \dots, n.$$

For a given i, the condition is equivalent to $\mu^2 - \mu\left(a^i_\pi + b^i_\pi + \frac{1}{t_i}\right) + a^i_\pi b^i_\pi \ge 0$, which, since we also require $\mu \ge \sum_{i=1}^n \lambda_i$, gives

$$\mu \ge r^i_\pi = \frac{1}{2}\left\{a^i_\pi + b^i_\pi + \frac{1}{t_i} + \sqrt{\left(a^i_\pi + b^i_\pi + \frac{1}{t_i}\right)^2 - 4a^i_\pi b^i_\pi}\right\}.$$

The minimum service rate that can be achieved with absolute preemptive priorities is:

$$\mu_{\text{PR}} = \min_{\pi\in\Pi(N)} \max_{1\le i\le n} r^i_\pi. \tag{6}$$

3.1 The aspiration problem with relative priorities

Let μ_{DPS} denote the minimum value of the service rate that satisfies a given aspiration level vector $T = (t_1, \dots, t_n) > 0$ using DPS. For a given service rate μ and a permutation $\pi \in \Pi(N)$ let $W_i^{\mu,\pi}$ denote the expected waiting time of class i given the absolute priority regime π. By Theorem 2.1, μ_{DPS} is the infimum value of μ, greater than $\sum_{i=1}^n \lambda_i$, for which there exists a nonnegative vector $\alpha = (\alpha_\pi)$ such that

$$\sum_{\pi\in\Pi(N)} \alpha_\pi = 1 \text{ and } \sum_{\pi\in\Pi(N)} \alpha_\pi W_i^{\mu,\pi} \le t_i, \ i = 1, \dots, n. \tag{7}$$

For any given value of μ this is a linear set of constraints on the α variables. Consequently, if the system (7) has a solution then it has one with at most $n+1$ positive values of α_π.

The optimal value μ_{DPS} is the unique solution to the following problem.

$$\begin{aligned}
\min \quad & \mu && (8)\\
\text{s.t.} \quad & \textstyle\sum_{\pi\in\Pi(N)} \frac{\alpha_\pi \mu}{(\mu-a^i_\pi)(\mu-b^i_\pi)} \le t_i, \quad i = 1, \dots, n, && (9)\\
& \textstyle\sum_{i=1}^n \lambda_i \le \mu \\
& \textstyle\sum_{\pi\in\Pi(N)} \alpha_\pi = 1, \\
& \alpha_\pi \ge 0, \qquad \forall \pi \in \Pi(N).
\end{aligned}$$

It is assumed that the data $\{a^i_\pi\}$, $\{b^i_\pi\}$, $\{t_i\}$ are rational, where each rational data item is represented as a ratio of two integers. Let M denote the maximum of the absolute values of all integers in this representation.

The constraints of the problem are algebraic functions defined over the rationals. For $i = 1, \ldots, n$, the ith constraint can be converted to a polynomial in the variables μ and $\{\alpha_\pi\}$ by multiplying (9) by $\prod_\pi[(\mu - a^i_\pi)(\mu - b^i_\pi)]$.

It follows from [4] and the references cited therein that there is an algebraic optimal solution, μ_{DPS}, $\{\alpha^*_\pi\}$. In particular, there is a minimal univariate characteristic polynomial, say $P(z)$, with integer data, such that $P(\mu_{\text{DPS}}) = 0$. More specifically, from the nature of the above constraints, the degree of $P(z)$ is bounded above by $f(n) = 2n(n!)$, and the absolute value of each one of its integer coefficients is bounded above by $(n!)M^{2n(n!)}$.

For each real μ, testing whether $\mu_{\text{DPS}} \leq \mu$ or $\mu_{\text{DPS}} > \mu$ requires the solution of a set of $n+1$ linear constraint in the $n!$ nonnegative variables $\{\alpha_\pi\}$. If μ is rational with integer numerator and denominator bounded above by N, solving such a linear program can be done in $Q(n!, \log M, \log N)$ time, where Q is polynomial.

With the above machinery, using the results in [4] and [5], we conclude with the following:

Theorem 3.1 *[19] There is a bivariate polynomial function $G(x, y)$, such that the time to find the characteristic polynomial $P(z)$ of μ_{DPS}, and a rational interval $[a, b]$, such that μ_{DPS} is the unique root of $P(z)$ in this interval, is bounded by $G(n!, \log M)$.*

Theorem 3.1 finds an interval $[a, b]$ containing μ_{DPS}. The optimal value μ_{DPS} can be located by any search algorithm for the root of $P(z)$ in $[a, b]$ (for example, Newton's method).

Once the solution μ_{DPS} is found we have to check whether it is attainable by DPS policies or not. This depends on the number of non-null α^*_π variables in the optimal solution of Problem 8. (There are at most $n + 1$.) Recall that μ_{DPS} is attainable by DPS policies if W belongs to the relative interior of $\mathcal{F}(\mathcal{N})$.

Comparing the service rates requirements under the different priority regimes, simply consists of comparing the values obtained by (5), (6) and Theorem 3.1.

3.2 Two classes

Consider now the case of $n = 2$ customer classes. Suppose the server implements a DPS with $x_1, x_2 = 1 - x_1$. The service rate should be large enough to satisfy the requirements $W_i \leq t_i$ $i = 1, 2$. Consider first $i = 1$. By (3), the requirement amounts to

$$\mu - \lambda x_1 \leq (\mu - \lambda)(\mu - \Lambda)t_1,$$

and of course $\mu > \lambda$. (Recall that $\lambda = \lambda_1 + \lambda_2$ and $\Lambda = \lambda_1 x_1 + \lambda_2 x_2$.) Equivalently,

$$t_1\mu^2 - [t_1(\Lambda + \lambda) + 1]\mu + \lambda(t_1\Lambda + x_1) \geq 0.$$

Let

$$\begin{aligned} \Delta_1 &= t_1^2(\Lambda + \lambda)^2 + 1 + 2t_1(\Lambda + \lambda) - 4t_1\lambda(t_1\Lambda + x_1) \\ &= [t_1(\Lambda - \lambda) + 1]^2 + 4t_1\lambda x_2. \end{aligned}$$

The condition is now

$$\mu \geq \mu_1 = \frac{t_1(\Lambda + \lambda) + 1 + \sqrt{\Delta_1}}{2t_1} = \frac{\Lambda + \lambda}{2} + \frac{1 + \sqrt{\Delta_1}}{2t_1}. \tag{10}$$

Similarly, the condition $W_2 \leq t_2$ amounts to

$$\mu \geq \mu_2 = \frac{\Lambda + \lambda}{2} + \frac{1 + \sqrt{\Delta_2}}{2t_2}, \tag{11}$$

where $\Delta_2 = [t_2(\Lambda - \lambda) + 1]^2 + 4t_2\lambda x_1$. We note that Δ_1 (Δ_2) are functions of x_1 although we do not write explicitly this dependence in its definition to simplify notation.

To satisfy both requirements, the server chooses a rate $\mu = \max\{\mu_1, \mu_2\}$.

Clearly, μ_1 is a decreasing function of x_1, and μ_2 is an increasing function of x_1. Therefore, the best priority parameter is that which satisfies $\mu_1 = \mu_2$.

Figure 3 (left) illustrates the solution for some values of the parameters. The graphs shown give μ as a function of the priority parameter x_1. The part of the function to the left of the minimum is μ_1 and it decreases when customer 1 obtains higher priority. Similarly, the part to the right of the minimum gives μ_2 which increases when customer 1 obtains higher priority and thus customer 2 obtains lower priority. The optimal service rate is obtained at the point where $\mu_1 = \mu_2$. In this figure we see that a decrease in t_1, which amounts to higher standards required by customer 1, leads to a solution with a higher μ and x_1. Of course this result is expected. Similarly, in Figure 3 (right) we see that an increase in λ_1 leads to increased value of μ, and in this example it is coupled with a decrease in the priority allocated to this customer.

We also conclude from Figure 3 that $\mu_{\text{DPS}} < \mu_{\text{PR}}$ is possible, that is, using relative priorities, it may be possible to reduce the service rate relative to the best result that can be obtained by any permutation of absolute priorities. This conclusion results from the observation that the two relative priority regimes that are possible in our example are represented by the values of the graphs at the extreme points $x = 0$ and $x = 1$. However, we see that a lower service rate is possible if we use intermediate priority values.

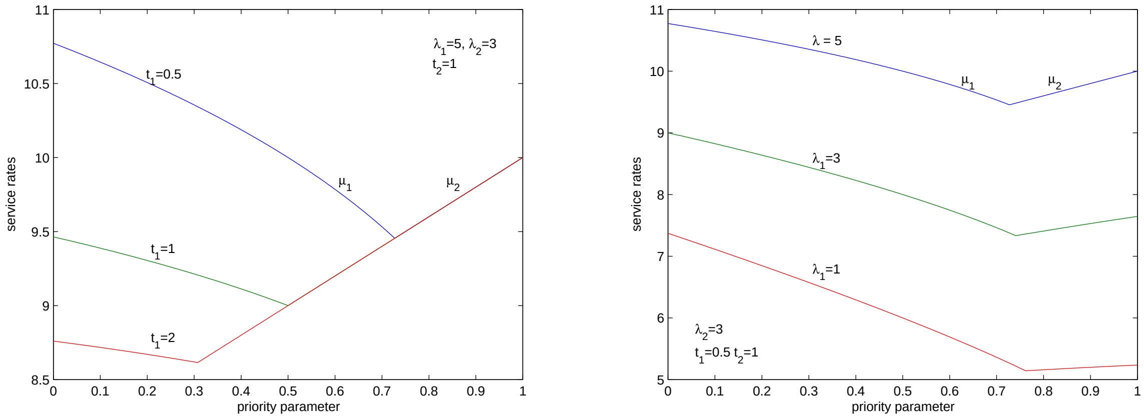

Figure 3: Required service rate as a function of x_1

As noted above, the minimum value of the system requirement under DPS is achieved when $\mu = \mu_1 = \mu_2$. This condition applied to (10) and (11) results in:

$$\frac{1 + \sqrt{\Delta_1}}{t_1} = \frac{1 + \sqrt{\Delta_2}}{t_2}. \tag{12}$$

After some algebra (manipulate equation (12) multiplying both sides by t_2, putting the 1 to the left side, raising to the power 2, and substituting $\Delta_2 = [t_2(\Lambda - \lambda) + 1]^2 + 4t_2\lambda x_1$, condition

(12) turns out to be:

$$t_2^2\left(\left[1+\sqrt{\Delta_1}\right]^2-(\Lambda-\lambda)^2t_1^2\right)-2t_2t_1\left(1+\sqrt{\Delta_1}+(\Lambda-\lambda)t_1+2\lambda x_1t_1\right)=0.$$

Since $t_2\neq 0$, the unique non-null root of the above equation is

$$t_2=2t_1\frac{1+\sqrt{\Delta_1}+(\Lambda-\lambda)t_1+2\lambda x_1t_1}{\left(1+\sqrt{\Delta_1}\right)^2-(\Lambda-\lambda)^2t_1^2}. \tag{13}$$

Lemma 3.2 *The function*

$$\phi(x_1)=2t_1\frac{1+\sqrt{\Delta_1}+(\Lambda-\lambda)t_1+2\lambda x_1t_1}{\left(1+\sqrt{\Delta_1}\right)^2-(\Lambda-\lambda)^2t_1^2}, \tag{14}$$

is continuous and increasing.

Proof: Let $\psi_1(x_1)=1+\sqrt{\Delta_1(x_1)}+\Lambda(x_1)-\lambda(1-2x_1)$ and $\psi_2(x_1)=1+\sqrt{\Delta_1(x_1)}+\Lambda(x_1)+\lambda(1-2x_1)$. (Notice that we have chosen in Δ_1 the appropriate root so that $\phi(x_1)$ goes to infinity when x_1 goes to 1.) Clearly, $\phi(x_1)=\frac{\psi_1(x_1)}{\psi_2(x_1)}$ for any $x_1\in[0,1)$ and its derivative $\phi'(x_1)$ is positive. Indeed, $\phi'(x_1)=2\lambda t_1^2[2-x_1+\lambda_1t_1(3-2x_1)+\lambda_2t_1(5-4x_1)+\lambda_1^2t_1^2(1-x_1)+\lambda_1\lambda_2t_1^2+(1-\lambda_2t_1)^2x_1+(2+\lambda t_1)\sqrt{\Delta_1(x_1)}]\Delta_1(x_1)^{-1/2}(1+\sqrt{\Delta_1(x_1)}-(\Lambda-\lambda)^2t_1^2)^{-2}>0$, since all the terms in the numerator are non-negative and some of them are strictly positive.

On the other hand, $0<\phi(0)<1$ and $\lim_{x_1\to 1^-}\phi(x_1)=+\infty$. Thus, ϕ is continuous, increasing monotone in the interval $[0,1)$.

■

Our next result gives the optimal priority value that ensures the aspiration levels and minimizes the service rate.

Corollary 3.3 *For any fixed value $t_1>0$ the optimal priority parameter x_1^*, as a function of t_2, is:*

$$x_1^*=\phi^{-1}(t_2).$$

Proof: The above properties (increasing monotonicity and continuity) of the function ϕ ensure that it has a proper inverse function and therefore the optimal priority parameter x_1^* can be computed by

$$x_1^*=\phi^{-1}(t_2).$$

■

Figure 4 shows x_1^* as a function of t_2. It assumes $t_1=1$, $\lambda_1=5$ and three values of λ_2. We note that the result is not very sensitive to the value of λ_2. Also note that when $t_2\to\infty$ we naturally have $x_1\to 1$, and that $x_1=0$ is obtained for positive values of t_2. The latter property is illustrated in the right part of Figure 4 which is a magnified section of the left part. Note that for $t_1=t_2$, $x_1^*=0.5$ even when $\lambda_1\neq\lambda_2$. With $t_2>t_1$ we have that x_1^* is monotone increasing with λ_2, and the opposite holds when $t_2<t_1$.

3.3 Comparing the disciplines

The rest of this section is devoted to comparing the required minimal service rate under the optimal DPS priority parameter, $\mu_{DPS}(x_1^*)$, with the same rate under FCFS, μ_{FCFS}.

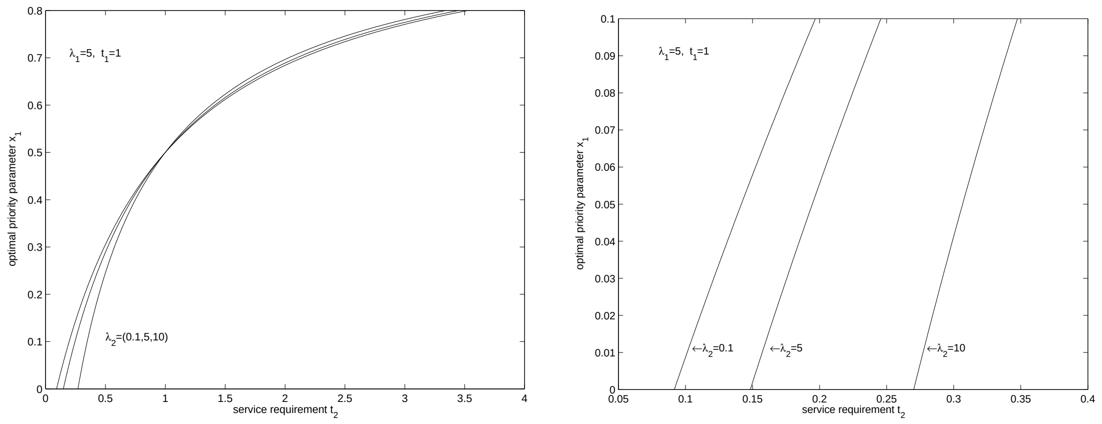

Figure 4: Optimal DPS priority parameter as a function of t_2

Theorem 3.4 *For $t_2 = t_1$, the minimal service rate required is the same for DPS and FCFS, but for $t_2 \neq t_1$ there is a priority parameter x_1^* that guarantees $\mu_{DPS}(x_1^*) < \mu_{FCFS}$.*

Proof: With $x_1 = \frac{1}{2}$, $\sqrt{\Delta_1} = 1 + t_1\frac{\lambda}{2}$ giving that the required service rate is

$$\mu_{DPS}(\frac{1}{2}) = \frac{3}{4}\lambda + \max_{i=1,2}\left\{\frac{2 + t_i\frac{\lambda}{2}}{2t_i}\right\} = \lambda + \max_{i=1,2}\left(\frac{1}{t_i}\right) = \mu_{FCFS},$$

where μ_{FCFS} is given in (5).

On the other hand, $t_2 = \phi(\frac{1}{2})$ if and only if $t_2 = t_1$ (substituting $x_1 = \frac{1}{2}$ in (14) gives $t_1 = t_2$). This means that if $t_1 \neq t_2$ then (12) is not satisfied for $x_1 = \frac{1}{2}$, meaning that it is not optimal and there is another value for x_1 that gives a strictly smaller value for μ. Since $x_1 = \frac{1}{2}$ gives the FCFS value we conclude the proof. ■

The minimal service rate requirements for DPS and FCFS are illustrated in Figure 5. This figure assumes that t_1 is fixed at 1 whereas t_2 varies. The FCFS requirement is determined by the minimum of t_1 and t_2 and therefore it is constant for $t_2 \geq 1$. We see that the two curves intersect when $t_2 = t_1$, but for any other value of t_2 selecting the right DPS parameter allows us to reduce the service rate - as proved in Theorem 3.4.

4 Concluding Remarks

Theorem 2.1 extends further to the case of $G/M/1$ systems because a work conservation law for the long-run expected amount of work in the system exists (see e.g. [7] and [13]). However, since no explicit formulas are known for the remaining elements in our analysis (e.g. W_i^π) in $G/M/1$ queues, the extension to that model, although meaningful, is currently an open question.

Acknowledgment

We thank Arie Tamir for providing us with a formal description of the existence of an algebraical optimal solution to Problem (8) through Theorem 3.1.

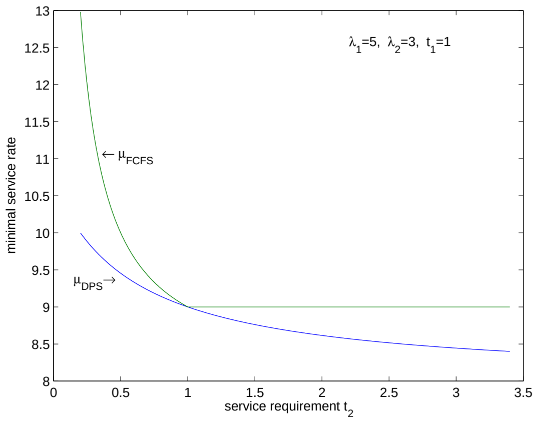

Figure 5: Minimal DPS and FCFS service requirements

References

[1] G. Allon and A. Federgruen (2006), “Competition in service industries with segmented markets.” http://ssrn.com/abstract=907322. (Jan. 2006). To appear in *Operations Research.*

[2] E. Altman, T. Jimenez and D. Kofman (2004), “DPS queues with stationary ergodic service times and performance of TCD in overloads,” in *Proceedings of INFOCOM*, 2004.

[3] T. Boland and L. Massoulie (2001), “Impact of fairness in Internet performance,” *Proceedings of Sigmetrics 2001*, 82-91.

[4] R. Chandrasekaran and A. Tamir (1984) “Optimization problems with algebraic solutions: quadratic fractional programs and ratio games,” Mathematical Programming **30** 326–339.

[5] R. Kannan, A.K. Lenstra and L. Lovász (1984) “Polynomial factorization and the nonrandomness of bits of algebraic and some transcendental numbers,” *Proc. 16th Symp. on Theory of Computing (STOC)* 191–200.

[6] E. G. Coffman and I. Mitrani (1980), “A characterization of waiting time performance realizable by single-server queues,” *Operations Research*, **28** 810–821

[7] A. Federgruen and H. Groenevelt (1988), “Characterization and optimization of achievable performance in general queueing systems,” *Operations Research*, **36** 733–741.

[8] E. Gelenbe and I. Mitrani (1980), *Analysis and Synthesis of Computer Systems*, Academic Press, London.

[9] G. Fayolle, I. Mitrani and R. Iasnogorodski (1980), “Sharing a processor among many classes,” *Journal of the Association for Computing Machinery*, **27** 519–532.

[10] R. Hassin and M. Haviv (2003), *To Queue or not to Queue: Equilibrium Behavior in Queueing Systems,* Kluwer International Series.

[11] R. Hassin and M. Haviv (2006), “Who should be given priority in a queue,” *Operations Research Letters*, **34** 191-198.

[12] M. Haviv and J. van der Wal (1997), ”Equilibrium strategies for processor sharing and queues with relative priorities,” *Probability in the Engineering and the Informational Sciences*, **11** 403-412.

[13] L. Kleinrock (1976), *Queueing Systems, Vol. 2: Computer Applications*, Willey Interscience, New York.

[14] B. Martos (1975), *Nonlinear programming. Theory and methods,* Amsterdam - Oxford: North-Holland Publishing Company; New York: American Elsevier Publishing Co., Inc.

[15] H. Mendelson and S. Whang (1990), “Optimal incentive-compatible priority pricing for the $M/M/1$ queue,” *Operations Research* **38** 870-883.

[16] I. Mitrani and J.H. Hine (1977), “Complete parametrized families of job scheduling strategies,” *Acta Informatica* **8** 61–73.

[17] H. Moulin and R. Stong (2002), “Fair queuing and other probabilistic allocation methods”, *Mathematics of Operations Research*, **27**, 1–30.

[18] K. M. Rege (1994), “A decomposition theorem and related results for the discriminatory processor sharing queue,” *Queueing Systems* **18** 333–351.

[19] A. Tamir (2006), Private communication.