# A Hybrid Analytical/Simulation Optimization of Generalized Processor Sharing

Jasper Vanlerberghe*, Tom Maertens*, Joris Walraevens*, Stijn De Vuyst† and Herwig Bruneel*
*Stochastic Modelling and Analysis of Communication Systems Research Group (SMACS)
Department of Telecommunications and Information Processing (TELIN)
Ghent University (UGent)
Sint-Pietersnieuwstraat 41, B-9000 Gent, Belgium
Email: {jpvlerbe,tmaerten,jw,hb}@telin.UGent.be
†Supply Networks & Logistics Research Center (SNLRC)
Department of Industrial Management
Ghent University (UGent)
Technologiepark 903, B-9052 Zwijnaarde, Belgium
Email: Stijn.DeVuyst@UGent.be

***Abstract*—With Generalized Processor Sharing (GPS), packets of different applications are backlogged in different queues and the different queues are served according to predetermined weights. It is well-established that GPS is a viable approach to provide different QoS for different applications. However, since the analysis of systems with GPS is a notoriously hard problem, it is not easy to find the weights that optimize GPS for some given objective function. The latter is important from a practical point of view. In this paper, we assume the objective function to be some weighted combination of (non-linear) increasing functions of the mean delays. We use results from strict priority scheduling (which can be regarded as a special case of GPS) to establish some exact theoretical bounds on when GPS is more optimal than strict priority. Some important case studies are included, thereby resorting to Monte-Carlo estimation to find the optimal weights for GPS systems.**

***Keywords*—Generalized Processor Sharing (GPS), strict priority, optimization, queueing***

## I. Introduction

Modern telecommunication networks must be capable of supporting a wide variety of *heterogeneous services*, such as traditional data, video, and voice services, but in addition also more demanding interactive multimedia services like online gaming and video conferencing. Different services, however, have different traffic characteristics and may have extremely diverse *Quality-of-Service* (QoS) requirements. Interactive services, for example, tolerate only a minimal *delay*, while data services allow for more delay. The integration of heterogeneous services with different QoS requirements raises the need for *service differentiation*. Strict priority scheduling [27] is the most drastic way to provide this differentiation: delay-tolerant packets can only be transmitted when there are no delay-sensitive packets in the system. A more flexible scheduling is Generalized Processor Sharing (GPS). It was developed as an efficient and fair scheduling discipline providing manageable service differentiation in computer and telecommunication networks (see, e.g., [17], [18]).

With GPS, each traffic class is assigned a *weight*, and the link capacity is shared according to the weights of the (backlogged) traffic classes. In this way, each traffic class can be guaranteed a minimum service rate, even though other traffic classes may be greedy in demanding service. This property prevents individual traffic classes from experiencing service starvation, as can be the case with strict priority scheduling. Note, furthermore, that assigning all capacity to a single class (weights of other classes are zero) implies that other classes can only be served if there is no traffic of this single class in the system. Strict priority scheduling is thus a special case of GPS, emphasizing the flexibility of the latter scheduling discipline. A key problem with GPS, however, is the *optimal* choice of the weight values. Optimal weights are defined as the weights that optimize some *objective function*, which is, in the context of scheduling disciplines, usually constructed in terms of (mean) delays or holding times, or throughput and/or loss characteristics (see, e.g., [4], [5], [7], [16], [23], [24], [29], and references therein). The main cause of this problem is the notorious complexity of queueing analyses of GPS systems (see, e.g., [9]–[11], and [28]), as opposed to queueing analyses of strict priority systems.

In this paper, we show how to use the results of queueing analyses of strict priority systems to put some theoretical bounds on when GPS (with non-zero weights) is more optimal than strict priority in a two-class system. We conclude from our theoretical study, which is valid for very general arrival processes, that strict priority is always optimal for certain forms of objective functions. In particular, we prove that when the objective function is a weighted combination of concave increasing functions of the mean delays of both classes, strict priority (for one of the two classes) is always optimal. On the other hand, when the objective function is a weighted combination of convex increasing functions of the mean delays, strict priority is not always optimal and GPS comes into the picture as optimal choice. In this case, it depends on the system and objective function parameters whether (i) strict priority for the first class is optimal, (ii) strict priority for the second class is optimal, or (iii) GPS (with some weights different from zero) is optimal. We prove, however, that objective functions can be categorized in these three groups by means of the results of queueing analyses of strict priority systems only (i.e., there

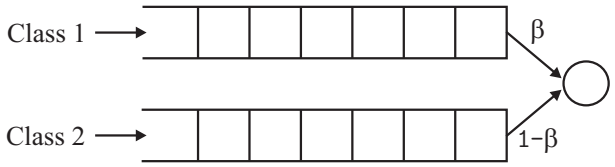


Fig. 1. GPS system at hand

is no need to analyse the GPS system). As a result, one can first easily investigate whether GPS is more optimal than strict priority scheduling before actually optimizing GPS. This is a big advantage in practice.

## II. Aim and model

In the remainder, we consider the discrete-time, probabilistic version of a two-class GPS queueing model with one server and general structured arrival process [11]. The GPS weights of both classes are normalized: *class* 1 is assigned weight $\beta$, while *class* 2 is assigned weight $1-\beta$, with $0 \leq \beta \leq 1$ (see Fig. 1). Assuming both classes have their own queue, this means that if both queues are non-empty at the beginning of a slot, queue 1 is served with probability $\beta$ and queue 2 is served with probability $1-\beta$. If one of the queues is empty, the other queue is served with probability 1. The cases $\beta = 0$ and $\beta = 1$ reduce to strict priority scheduling, which is well-studied and allows, in many important cases, for an explicit analytical solution, mainly in terms of a multi-dimensional probability generating function (pgf) (see, e.g., [22], [25]–[27]). For $0 < \beta < 1$, however, it is generally known that such a queueing system gives rise to a hard-to-solve functional equation for this pgf (see, e.g., [10], [28]). One (exact) technique is to transform the functional equation into a so-called Riemann-Hilbert boundary value problem (see, e.g., [9]–[11]). Unfortunately, this technique requires considerable numerical and mathematical efforts, and can only be solved in specific cases (for example, the symmetric case $\beta = 1/2$ with identical arrival and service characteristics for both classes leads to notable simplifications [10]). As a consequence, several alternative approaches have been proposed over the years. We refer to [28] for a survey and to [1] for a comparison between these approaches.

Each of the approaches to analyse GPS systems has its limitations, computationally and/or accuracy-wise. Hence, they are not directly applicable for optimization purposes. In comparison with the queueing analyses of GPS systems with fixed weights, the *optimization* of GPS has therefore received considerably less attention in the literature. Some exceptions are [16], [20], and [24]. In [24], the authors derive some monotonicity properties for weighted combinations of the mean queue contents (holding times) for systems with more general scheduling policies. Lieshout and Mandjes [16] consider an objective function based on loss probabilities. Their results suggest that the choice of weights is not that critical, and that strict priority may suffice for practical purposes. Finally, in [20], the authors optimize the average sum of the $n$-th powers of the queue contents. They conclude that the objective function tends to balance the performance of the queues for $n \to \infty$.

In absence of exact results for general $\beta$, we opt for Monte-Carlo simulation to estimate the characteristics of GPS systems. Simulations can be quite time-consuming and problems might arise when using simulation for optimization. However, we limit these problems by using the 'common random numbers' (CRN) variance reduction technique (see, e.g., [3], [21]). Moreover, we identify the cases for which strict priority scheduling (for one of the two classes) is optimal and, hence, for which we do not even need to simulate the GPS system. Specifically, by means of explicit analytical results of strict priority scheduling and from the observation that the total unfinished work is independent of the scheduling discipline (as long as it is work-conserving), we prove in Sect. III in which cases (objective function, values of the system parameters) the optimal scheduling discipline is simply strict priority, i.e., $\beta$ equal to 0 or 1. In Sect. IV, we provide some details on the Monte-Carlo simulation used to estimate GPS system characteristics. In Sect. V, finally, we apply our analysis on some important case studies.

## III. Analytical framework

As already mentioned, we consider a discrete-time queueing model with one server and two traffic classes, i.e., class 1 and class 2. The mean arrival rates of class-$j$ are denoted by $\lambda_j$, $j = 1, 2$, while we define the total arrival rate as $\lambda_T$ $(= \lambda_1 + \lambda_2)$. Packets of class 1 arrive in queue 1; packets of class 2 enter queue 2. Packets from queue $j$ ($j = 1, 2$) need service for a (shifted) geometrically distributed number of slots with mean $1/\mu_j$. If one of the queues is empty at the beginning of a slot, a packet of the other queue is served. If both queues are non-empty at the beginning of a slot, a packet of queue 1 is served with probability $\beta$ and a packet of queue 2 is served with probability $1-\beta$ (see Fig. 1). At the moment, we do not specify the arrival process; we only make some standard assumptions, namely (i) that non-negative integer numbers of class-1 and class-2 packets arrive during each slot, (ii) that the arrival process is stationary, and that the arrival process is such (iii) that the system is stable (i.e., $\lambda_T < 1$) and (iv) that the probability that packets are present in both queues simultaneously is non-zero.

We will write the analytical framework in terms of the mean unfinished work in both queues in steady state, where the unfinished work in a queue at the beginning of a slot is defined as the sum of the residual service times of the packets present in the queue at that moment. The framework can be easily translated to mean packet delays and mean queue contents; we will comment on this later. It is obvious that these performance measures depend on the parameter $\beta$. Therefore, we denote the mean unfinished work in queue $j$ at the beginning of a random slot in steady state as $\bar{w}_j(\beta)$ ($j = 1, 2$). Functions $\bar{w}_1(\beta)$ and $\bar{w}_2(\beta)$ have two straightforward, yet important properties. Their proofs can be found in the appendix.

**Proposition 1.** Function $\bar{w}_1(\beta) + \bar{w}_2(\beta)$ is independent of $\beta$.

**Proposition 2.** Functions $\bar{w}_2(\beta)$ and $\bar{w}_1(\beta)$ are strictly increasing and strictly decreasing, respectively, with respect to (w.r.t.) $\beta$.

Basically, Propositions 1 and 2 follow from the observation that GPS is a work-conserving scheduling discipline and that class 2 is given 'less priority' (diminishing share of the bandwidth) with increasing $\beta$ (see also Fig. 1), respectively. It

should be noted that the above propositions are in terms of mean values. This is sufficient for the aim of this paper. However, the propositions can also be stated and proved in stronger terms (e.g., stochastic variables being independent of $\beta$, stochastic domination in $\beta$, ...).

Now we turn to the objective function that we wish to minimize. We assume the objective function to be a weighted combination of strictly increasing functions of the mean unfinished work in both queues, i.e.,

$$F(\beta,\gamma) \triangleq \gamma g_1(\bar{w}_1(\beta)) + (1-\gamma) g_2(\bar{w}_2(\beta)), \tag{1}$$

with $0 \leq \gamma \leq 1$. The parameter $\gamma$ expresses the relative importance that is given to $\bar{w}_1(\beta)$: the higher $\gamma$, the more important $\bar{w}_1(\beta)$. When $\gamma = 0$, the objective function only takes into account $\bar{w}_2(\beta)$; when $\gamma = 1$, only $\bar{w}_1(\beta)$ plays a role. In these cases, it is obvious that the strict priority scheduling discipline ($\beta = 0$ and $\beta = 1$, respectively) minimizes $F(\beta,\gamma)$, because of Proposition 2 and because of the assumption that $g_1$ and $g_2$ are increasing functions. Function $g_j$ ($j = 1, 2$) expresses how increments in the mean unfinished work in queue $j$ are penalized. For convex functions, for instance, increments of high values are more penalized than increments of low values. For concave functions, it is the other way around.

Before continuing, it is important to note that by choosing appropriate functions $g_j$, one can easily switch to an objective function in terms of mean steady-state queue contents and/or packet delays. Since service times of class $j$ are geometrically distributed with parameter $\mu_j$, the mean content of queue $j$, defined as $\bar{u}_j(\beta)$, can be easily expressed as a function of the mean unfinished work in queue $j$:

$$\bar{u}_j(\beta) = \mu_j \bar{w}_j(\beta). \tag{2}$$

Furthermore, by applying Little's law, we get that

$$\bar{d}_j(\beta) = \frac{\mu_j \bar{w}_j(\beta)}{\lambda_j}, \tag{3}$$

with $\bar{d}_j(\beta)$ the mean packet delay for class $j$. So by choosing $g_j = \hat{g}_j \circ h_j$ in equation (1), with $h_j(x) = \mu_j x/\lambda_j$ and $\hat{g}_j$ an increasing function independent of the system parameters, we obtain an objective function in terms of the mean packet delays for both classes. In Section V, where we illustrate our results, we make this switch. Notice that this is the main reason why we allowed $g_1$ and $g_2$ to be different. In practice, $\hat{g}_1$ will be chosen equal to $\hat{g}_2$ (linear, quadratic, logarithmic, ...; see also [20], [24]), and $g_1$ and $g_2$ will be different through $h_1$ and $h_2$ only.

Let us first prove an important lemma.

**Lemma 1.** Assume $g_j'(x) > 0, \forall x \geq 0$ ($g_j(x)$ is strictly increasing for $x \geq 0$). Then $\dfrac{\partial F(\beta,\gamma)}{\partial \beta} > (<,=)\, 0$ iff $\gamma < (>,=)\, \phi(\beta)$ with

$$\phi(\beta) \triangleq \frac{g_2'(\bar{w}_2(\beta))}{g_2'(\bar{w}_2(\beta)) + g_1'(\bar{w}_1(\beta))}. \tag{4}$$

Furthermore, $\phi(\beta) \in ]0,1[$.

*Proof:* By taking the partial derivative of $F(\beta,\gamma)$ w.r.t. $\beta$ and using Proposition 1, we obtain

$$\frac{\partial F(\beta,\gamma)}{\partial \beta} = [g_2'(\bar{w}_2(\beta)) - \gamma(g_1'(\bar{w}_1(\beta)) + g_2'(\bar{w}_2(\beta)))]\bar{w}_2'(\beta). \tag{5}$$

Then the lemma follows from the assumption that the functions $g_j$ are strictly increasing functions and from Proposition 2 which states that $\bar{w}_2'(\beta) > 0$. ∎

Lemma 1 is important, as it separates, for given $\beta$, the interval $\gamma = [0,1]$ in three parts, $[0, \phi(\beta)[$, the point $\phi(\beta)$, and $]\phi(\beta), 1]$, where the objective function $F(\beta,\gamma)$ increases, is constant, and decreases, respectively. In particular, for $\beta = 0$, $F(\beta,\gamma)$ is increasing (decreasing) when $\gamma < (>)\, \phi(0)$; for $\beta = 1$, the objective function is increasing (decreasing) when $\gamma < (>)\, \phi(1)$. It is easily seen that $\phi(0)$ and $\phi(1)$ can be calculated from results of strict priority scheduling, see (4). The values of $\phi(0)$ and $\phi(1)$ are of primordial importance. Indeed, we prove that for certain important classes of functions $g_j$ and depending on the value of $\gamma$ w.r.t. $\phi(0)$ and $\phi(1)$, strict priority scheduling (either $\beta = 0$ or $\beta = 1$) will be optimal.

Specifically, we analyse three particular classes for the (increasing) functions $g_j$, namely linear functions, convex functions, and concave functions. The derivative of $\phi(\beta)$ will play an important role in these analyses, so we first calculate this derivative from equation (4):

$$\phi'(\beta) = \frac{[g_2''(\bar{w}_2(\beta))g_1'(\bar{w}_1(\beta)) + g_2'(\bar{w}_2(\beta))g_1''(\bar{w}_1(\beta))]\bar{w}_2'(\beta)}{[g_2'(\bar{w}_2(\beta)) + g_1'(\bar{w}_1(\beta))]^2}. \tag{6}$$

**Theorem 1.** Assume $g_j'(x) > 0$ and $g_j''(x) = 0, \forall x \geq 0$ (i.e., $g_j(x)$ is a linear increasing function). Then $\phi(\beta) = \phi_C$ is a constant function. As a result, strict priority is always optimal, namely

(i) for $\gamma < \phi_C$, $\beta = 0$ (strict priority for class 2) is optimal,

(ii) for $\gamma > \phi_C$, $\beta = 1$ (strict priority for class 1) is optimal, and

(iii) for $\gamma = \phi_C$, all values of $\beta$ are equally optimal.

*Proof:* From equation (6), it follows that $\phi(\beta)$ is a constant (say $\phi_C$). Then the theorem follows directly from Lemma 1: (i) if $\gamma < \phi_C$, $F(\beta,\gamma)$ is strictly increasing w.r.t. $\beta$ and has its minimum in $\beta = 0$; (ii) if $\gamma > \phi_C$, $F(\beta,\gamma)$ is strictly decreasing w.r.t. $\beta$ and reaches its minimum in $\beta = 1$; and, (iii) if $\gamma$ is equal to the constant $\phi_C$, $F(\beta,\gamma)$ is identical for all values of $\beta$. ∎

From this theorem, it follows that strict priority is always optimal when the objective function is a weighted combination of linear functions of the mean unfinished work in both queues. Similar observations have been made in the past (cf. the $c\mu$-rule [24]).

We now turn to convex increasing functions $g_j$.

**Theorem 2.** Assume $g_j'(x) > 0$ and $g_j''(x) > 0, \forall x \geq 0$ (i.e., $g_j(x)$ is strictly increasing and strictly convex). Then $\phi(\beta)$ is a strictly increasing function. As a result,

(i) for $\gamma \leq \phi(0)$, $\beta = 0$ (strict priority for class 2) is optimal,

(ii) for $\gamma \geq \phi(1)$, $\beta = 1$ (strict priority for class 1) is optimal, and

(iii) for $\phi(0) < \gamma < \phi(1)$, $\beta_{\text{opt}}(\gamma) = \phi^{-1}(\gamma)$ is optimal, with $\phi^{-1}$ the inverse function of $\phi$. The function $\beta_{\text{opt}}(\gamma)$ has the following properties: it is different from 0 or 1 (i.e., we have 'true' GPS) and it is strictly increasing.

*Proof:* The right-hand side of equation (6) is positive, due to the assumptions on the functions $g_j$ and Proposition 2. As a consequence, $\phi(\beta)$ is a strictly increasing function. It then follows directly that if $\gamma \leq \phi(0)$, $\gamma < \phi(\beta)$ for $0 < \beta \leq 1$. From Lemma 1, we have that $F(\beta, \gamma)$ is in this case non-decreasing w.r.t. $\beta$ and reaches its minimum in $\beta = 0$. If $\gamma \geq \phi(1)$, on the other hand, $\gamma > \phi(\beta)$ for $0 \leq \beta < 1$. In this case, $F(\beta, \gamma)$ is non-increasing w.r.t. $\beta$ and reaches its minimum in $\beta = 1$. Finally, when $\phi(0) < \gamma < \phi(1)$, there is a unique $\beta \in ]0, 1[$ so that $\phi(\beta) = \gamma$ (since $\phi(\beta)$ is a strictly increasing function). We denote this $\beta$ by $\beta_{\text{opt}}(\gamma)$ (i.e., $\beta_{\text{opt}}(\gamma) = \phi^{-1}(\gamma)$). Then Lemma 1 indicates that $F(\beta, \gamma)$ decreases with $\beta$ in the interval $[0, \beta_{\text{opt}}(\gamma)[$ and increases in the interval $]\beta_{\text{opt}}(\gamma), 1]$. As a result, $F(\beta, \gamma)$ is unimodal, and $\beta_{\text{opt}}(\gamma)$ is optimal and lies between 0 and 1 ($0 < \beta_{\text{opt}}(\gamma) < 1$). Finally, $\beta_{\text{opt}}(\gamma)$ being strictly increasing follows from $\beta_{\text{opt}}(\gamma) = \phi^{-1}(\gamma)$ and $\phi(\beta)$ strictly increasing. ■

From this theorem, it follows that GPS might be optimal for weighted combinations of convex increasing functions of the mean unfinished work in both queues. Moreover, the theorem sets bounds on the values of $\gamma$ for which GPS is optimal. These bounds can be calculated from results of queueing analyses of a *strict priority* system. This is a big advantage, as the analysis of a strict priority system is usually much easier than the analysis of a GPS system. Outside the bounds, strict priority is optimal, so one does not have to search for the optimal $\beta$. Only inside the bounds, one should search for the optimal $\beta$. The optimal $\beta$ is in fact $\phi^{-1}(\gamma)$, with $\phi(\beta)$ defined in (4). In order to calculate $\phi(\beta)$, however, we need to find the mean unfinished work in both queues for that $\beta$, which is the hard part. To find the optimal $\beta$, one can, for example, adopt the 'golden section search' algorithm, as explained in Section V.

Finally, we consider concave increasing functions $g_j$.

**Theorem 3.** Assume $g_j'(x) > 0$ and $g_j''(x) < 0, \forall x \geq 0$ (i.e., $g_j(x)$ is strictly increasing and strictly concave). Then $\phi(\beta)$ is a strictly decreasing function. As a result, strict priority is always optimal, namely

(i) for $\gamma \leq \phi(1)$, $\beta = 0$ (strict priority for class 2) is optimal,

(ii) for $\gamma \geq \phi(0)$, $\beta = 1$ (strict priority for class 1) is optimal, and

(iii) for $\phi(1) < \gamma < \phi(0)$, $\beta = 0$ ($\beta = 1$) is optimal if $F(0, \gamma) < (>) F(1, \gamma)$; if $F(0, \gamma) = F(1, \gamma)$, both $\beta = 0$ and $\beta = 1$ are optimal.

*Proof:* From equation (6) and Proposition 2, it follows that $\phi(\beta)$ is strictly decreasing for strictly increasing and strictly concave functions $g_j$. By similar arguments as in the previous proof, we verify cases (i) and (ii). When $\phi(1) < \gamma < \phi(0)$, $F(\beta, \gamma)$ first increases w.r.t. $\beta$, until it reaches a maximum, and then decreases w.r.t. $\beta$. Consequently, the minimum of the objective function is reached for $\beta = 0$ or $\beta = 1$, depending on which is lowest. This results in case (iii). ■

This theorem states that for weighted combinations of concave increasing functions of the mean unfinished work in both queues, strict priority scheduling is always optimal. One merely has to compare the objective functions for $\beta = 0$ and $\beta = 1$ (priority for class 2 and class 1, respectively).

## IV. Simulation details

Before we dive into some case studies and numerical examples, we describe how we estimate the objective function $F(\beta, \gamma)$ for different values of $\beta$ ($0 \leq \beta \leq 1$). Remember that $F(\beta, \gamma)$ is assumed to be a weighted combination of strictly increasing functions of some characterizing steady-state mean performance measures of GPS systems (see equations (1)-(3)). So to estimate $F(\beta, \gamma)$, we need estimates for these mean performance measures. Therefore, we run a Monte-Carlo simulation over a number of slots (say $K$), according to a particular arrival and service process.

Monte-Carlo simulations, and thus also estimates of $F(\beta, \gamma)$, for different values of $\beta$, are furthermore coupled in two ways. First, every simulated trajectory (or sample path) starts from an initial state where the queues are empty. The simulated trajectories therefore exhibit a transient period before reaching steady-state behaviour. Secondly, the sequences of class-1 and class-2 arrivals and all service times are in all simulations identical. As a consequence, the cycles (busy periods) are in every trajectory identical as well. Additionally, the sequences of uniform random variables used to decide which queue to serve in a slot with two non-empty queues are also the same. In this way, we make sure that $\bar{w}_1(\beta)$ ($\bar{w}_2(\beta)$) is a strictly decreasing (increasing) function w.r.t. $\beta$, as described in the proof of Proposition 2 and an important property in the proofs of Theorems 1-3.

Generating identical arrival sequences and service times in subsequent simulations was done by initialising the seed of the random number generator to the same value at the beginning of each simulation. This also guarantees that the sequence of uniform random variables used for deciding on server allocation is the same every time, because exactly one such value is generated per slot, regardless of whether a decision needs to be made or not. This method, referred to as 'common random numbers' (CRN) or correlated sampling (see, e.g., [3], [21]), is known to reduce the variance of the estimates considerably, which assures smooth curves for $F(\beta, \gamma)$, and, moreover, a unique solution for $\beta_{\text{opt}}(\gamma)$. Adopting CRN, however, comes at a price. To see this, let us denote by $\boldsymbol{u}$ the sequence of independent random numbers that we used to generate a trajectory over $K$ slots. From this trajectory, estimates $F(\beta, \gamma, K, \boldsymbol{u})$ of the objective function, for different values of $\beta$, and $\beta_{\text{opt}}(\gamma, K, \boldsymbol{u})$ of the optimal $\beta$ are obtained. It is then clear that such estimates are *biased* in two ways: due to the transient period and, more importantly, because of the particular choice of $\boldsymbol{u}$. For another random sequence $\boldsymbol{u}'$, the estimates $F(\beta, \gamma, K, \boldsymbol{u}')$, $0 \leqslant \beta \leqslant 1$, may be different, as well as the point $\beta_{\text{opt}}(\gamma, K, \boldsymbol{u}')$ where it is minimal.

Nevertheless, both types of bias are not systematic and will disappear if very long trajectories are used. So, even with CRN, the estimates are asymptotically consistent. The queueing process studied here is ergodic if $\lambda_T < 1$ and indeed strongly mixing, which guarantees that whatever arrival and server allocation sequences and service times are used (i.e., whatever $\boldsymbol{u}$), the generated time averages will converge to the steady-state ensemble averages. Choosing $K = 10^8$ is more than enough to establish this, resulting in nearly unbiased and almost variance-free estimates.

## V. Important case studies

In this final section, we apply and support the analytical derivations of Section III to some interesting case studies. Specifically, we study three interesting (classes of) objective functions $F(\beta, \gamma)$, corresponding to the three cases handled in Theorems 1-3, and this for a specific arrival and service process.

In particular, we assume one-slot service times, i.e., $\mu_j = 1$ ($j = 1, 2$), and a general structured arrival process which is independent and identically distributed (i.i.d.) from slot to slot. Such an arrival process, characterized by the pgf $A(z_1, z_2)$ of the numbers of class-1 and class-2 arrivals during a slot, allows for correlation between these numbers within one slot. Throughout this section, the formulas will be given for general $A(z_1, z_2)$. However, for the figures illustrating the formulas and the theoretical results, we consider the popular two-dimensional binomial arrival process with pgf

$$A(z_1, z_2) = \left(1 - \frac{\lambda_1}{N}(1 - z_1) - \frac{\lambda_2}{N}(1 - z_2)\right)^N. \quad (7)$$

This is the arrival process in a queue of an $N$x$N$ output-queueing switch with Bernoulli arrivals at its inlets and with independent and uniform routing towards the outlets. The parameter $N$ expresses the maximum total number of arriving packets during a slot and is assumed to be 16 in the numerical examples. In the numerical examples, additionally, we switch to packet delays instead of unfinished work (see equation (3)), as, in our opinion, objective functions in packet delays are practically most relevant in the context of heterogeneous services.

A queueing model with such an independent structured arrival process, with i.i.d. and geometrically distributed service times, and with class 1 having strict priority over class 2 (i.e., $\beta = 1$ in the GPS setting) has been studied in detail in [25]. In [25], the authors derive expressions for the pgfs of the queue contents and the packet delays. These pgfs further lead to expressions for some interesting performance measures involving these quantities (e.g., mean values, variances, tail probabilities, ...). By adopting deterministic service times of one slot, we obtain

$$\bar{w}_1(1) = \lambda_1 + \frac{\lambda_{11}}{2(1 - \lambda_1)}, \quad (8)$$

and

$$\bar{w}_2(1) = \lambda_2 + \frac{\lambda_{TT}}{2(1 - \lambda_T)} - \frac{\lambda_{11}}{2(1 - \lambda_1)}, \quad (9)$$

for the mean unfinished work in both queues, where $\lambda_j$ ($j = 1, 2$) denotes the class-$j$ arrival rate, $\lambda_T$ represents the total arrival rate (i.e., $\lambda_T = \lambda_1 + \lambda_2$) and is assumed to be strictly less than one, $\lambda_{jj} \triangleq \left.\dfrac{\partial^2 A(z_1, z_2)}{\partial z_j^2}\right|_{z_1=1, z_2=1}$ ($j = 1, 2$), and $\lambda_{TT} \triangleq \left.\dfrac{d^2 A(z, z)}{dz^2}\right|_{z=1}$. For convenience, we introduce an additional parameter $\alpha$ indicating the fraction of class-1 packets in the overall arrival stream, i.e., $\alpha \triangleq \lambda_1/\lambda_T$. To find the formulas for the same model but with class 2 having strict priority over class 1 (i.e., $\beta = 0$ in the GPS setting), one just has to replace $A(z_1, z_2)$ by $A(z_2, z_1)$ in the expressions for the pgfs and take the appropriate derivatives:

$$\bar{w}_1(0) = \lambda_1 + \frac{\lambda_{TT}}{2(1 - \lambda_T)} - \frac{\lambda_{22}}{2(1 - \lambda_2)}, \quad (10)$$

and

$$\bar{w}_2(0) = \lambda_2 + \frac{\lambda_{22}}{2(1 - \lambda_2)}. \quad (11)$$

As described in the previous section, these expressions are sufficient to calculate $\phi(0)$ and $\phi(1)$, which identify the values of $\gamma$ for which either priority ($\beta = 0$ or $\beta = 1$) or GPS ($0 < \beta < 1$) is optimal.

### A. Linear functions: $g_j(x) = a_j x$

We start with the case where the functions $g_j$ are linear increasing functions. Assume $g_j(x) = a_j x$ ($j = 1, 2$), with $a_j$ constants. Using equation (4) then yields

$$\phi(\beta) = \frac{a_2}{a_2 + a_1}. \quad (12)$$

As proved in Theorem 1, $\phi(\beta)$ is constant in this case. As a result, the objective function $F(\beta, \gamma)$ is strictly increasing w.r.t. $\beta$ for $\gamma < \phi(\beta)$, constant for $\gamma = \phi(\beta)$, and decreasing for $\gamma > \phi(\beta)$. This is demonstrated in Fig. 2, where we depict $F(\beta, \gamma)$ as a function of $\beta$, for $a_j = 1/\lambda_j$, $\alpha = 0.8$, $\lambda_T = 0.9$, and five different values of $\gamma$ (including $\phi(\beta) = 0.8$). As already mentioned, $F(\beta, \gamma)$ cannot be calculated exactly for $0 < \beta < 1$. Estimates for $F(\beta, \gamma)$ (here, and further in this section), for 101 equidistant values of $\beta$ in the range $[0, 1]$, are obtained through coupled Monte-Carlo simulations, as described in the previous section.

From theorem 1, it follows that $\beta = 0$ is optimal if $\gamma < \phi(\beta)$, or, by means of formula (12), if

$$a_1 \gamma < a_2(1 - \gamma). \quad (13)$$

This can be viewed as an occurrence of the $c\mu$-rule which states that the queue with the highest $c \cdot \mu$ should be given the highest priority [24], with $c$ the weight given to the mean holding time of that queue in the objective function and $\mu$ the service rate of that queue. In our case, $\gamma$ and $1 - \gamma$ play the role of $c$ (the weights in the objective function) and $a_j = \mu_j$ if we are concerned with holding times (queue contents). For the special case $\mu_j = 1$, we find that class 2 should be given priority if $\gamma < 1/2$. If we are interested in minimizing the weighted sum of the mean delays, class 2 should be given priority if $\gamma < \alpha$. This can also be observed from Fig. 2. Note, finally, that all curves intersect at the same point. For the corresponding value of $\beta$, the objective function is independent of $\gamma$. This is only possible if $\bar{d}_1(\beta) = \bar{d}_2(\beta)$, i.e., if the mean delays for both classes are balanced.

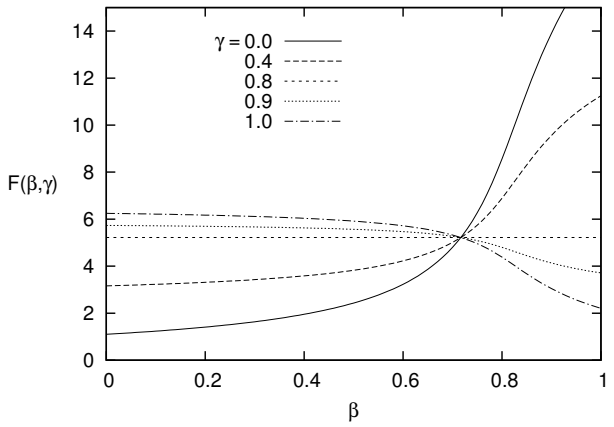


Fig. 2. Objective functions $F(\beta,\gamma)$ as a function of $\beta$, for $g_j(x) = x/\lambda_j$ $(j = 1, 2)$, $\alpha = 0.8$, and $\lambda_T = 0.9$

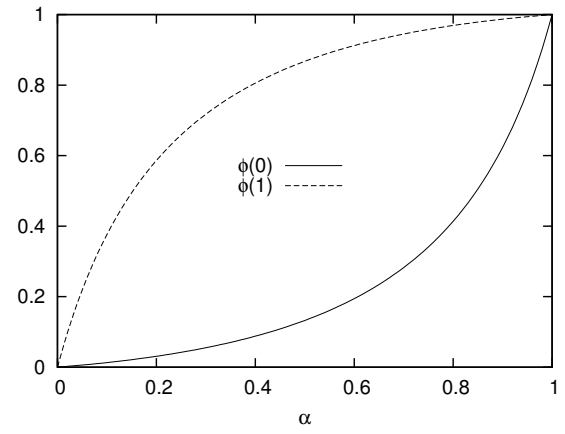


Fig. 3. $\phi(0)$ and $\phi(1)$ as a function of $\alpha$, for $a_j = 1/\lambda_j$ $(j = 1, 2)$ and $\lambda_T = 0.9$

### B. *Convex functions:* $g_j(x) = (a_j x)^n$, $n \geq 2$

Secondly, we demonstrate the case where the functions $g_j$ are convex increasing functions. As mentioned earlier, the rationale for convex $g_j$ is that increments of high values of the considered performance measures are more penalized than increments of low values. Assume $g_j(x) = (a_j x)^n$ $(j = 1, 2)$, with $n$ discrete and larger than 1, and $a_j$ constants. First, we treat the case $n = 2$; later, we consider general values of $n$.

*1)* $n = 2$*:* By using equation (4) and expressions (8)-(11), we find that

$$\phi(0) = \frac{a_2^2(\lambda_{22} + 2\lambda_2(1-\lambda_2))(1-\lambda_T)}{\left\{ \begin{array}{l} (1-\lambda_T)(a_2^2 - a_1^2)\lambda_{22} + a_1^2(1-\lambda_2)\lambda_{TT} \\ +2(1-\lambda_T)(1-\lambda_2)(a_2^2\lambda_2 + a_1^2\lambda_1) \end{array} \right\}}, \tag{14}$$

and

$$\phi(1) = \frac{a_2^2 \left\{ \begin{array}{l} (1-\lambda_1)\lambda_{TT} - (1-\lambda_T)\lambda_{11} \\ +2\lambda_2(1-\lambda_T)(1-\lambda_1) \end{array} \right\}}{\left\{ \begin{array}{l} (1-\lambda_T)(a_1^2 - a_2^2)\lambda_{11} + a_2^2(1-\lambda_1)\lambda_{TT} \\ +2(1-\lambda_T)(1-\lambda_1)(a_2^2\lambda_2 + a_1^2\lambda_1) \end{array} \right\}}, \tag{15}$$

respectively. In this case, $\phi(0)$ and $\phi(1)$ are not equal. In fact, it follows from Theorem 2 that $\phi(0) < \phi(1)$. From (14)-(15), this can also be proved for this specific case. In Fig. 3, we illustrate $\phi(0)$ and $\phi(1)$ as a function of $\alpha$, for the two-dimensional binomial arrival process with $\lambda_T = 0.9$. For $\gamma \geq \phi(1)$ (i.e., at the upper left region of the graph), $\beta_{\text{opt}}(\gamma) = 1$ according to Theorem 2. When $\gamma \leq \phi(0)$ (i.e., at the lower right region of the graph), $\beta_{\text{opt}}(\gamma) = 0$. When $\phi(0) < \gamma < \phi(1)$, Theorem 2 states that the objective function $F(\beta,\gamma)$ reaches a minimum for some $\beta$ between 0 and 1. So here, $\beta_{\text{opt}}(\gamma) \in ]0, 1[$.

We illustrate these results by choosing a specific value for $\alpha$ and looking at the behaviour of $F(\beta,\gamma)$ for that $\alpha$. Assume $\alpha = 0.8$. Then $\phi(0) = 0.41$ and $\phi(1) = 0.96$. Fig. 4 shows $F(\beta,\gamma)$ as a function of $\beta$, for $\alpha = 0.8$, $\lambda_T = 0.9$, and five different values of $\gamma$ (including 0.41 and 0.96). We clearly see that only when $\gamma \in ]\phi(0), \phi(1)[$, $F(\beta,\gamma)$ reaches a minimum for some $\beta$ different from 0 and 1. For all other values of

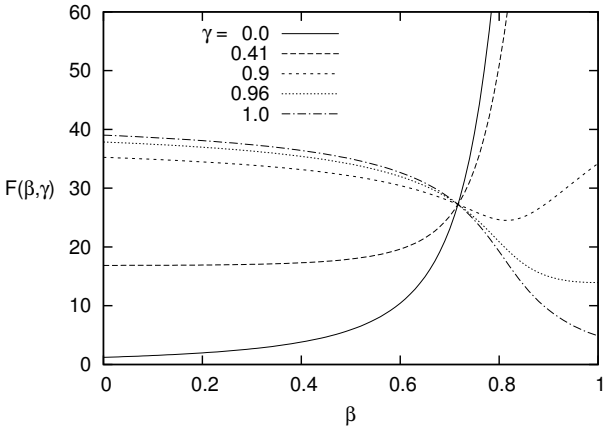


Fig. 4. Objective functions $F(\beta,\gamma)$ as a function of $\beta$, for $g_j(x) = (x/\lambda_j)^2$ $(j = 1, 2)$, $\alpha = 0.8$, and $\lambda_T = 0.9$

$\gamma$ (i.e., $\gamma \in [0, \phi(0)]$ and $\gamma \in [\phi(1), 1]$), $F(\beta,\gamma)$ is non-decreasing and non-increasing w.r.t. $\beta$ and, hence, reaches a minimum in $\beta = 0$ and $\beta = 1$, respectively. For $\gamma = 0.41$ and $\gamma = 0.96$, it is observed that the objective functions have horizontal asymptotes in $\beta = 0$ and $\beta = 1$, respectively, as proved in Lemma 1. Furthermore, it can be seen from the curve for $\gamma = 0.9$ that the difference between GPS with $\beta = \beta_{\text{opt}}(\gamma)$ and the strict priority cases $\beta = 0$ and $\beta = 1$ can be considerable.

When $\gamma \in ]\phi(0), \phi(1)[$, we need to search for the optimal value of $\beta$. Since we have no explicit analytical results for the mean values of characterizing quantities in GPS systems, this search boils down to the determination of the minimum of a function which cannot be computed exactly and thus has to be estimated. Here, however, we know that the function is strictly unimodal in $\beta$, and, luckily, there are some well-known techniques for this optimization problem. In particular, 'golden section search', 'successive parabolic interpolation', or a combination of these two, and 'stochastic approximation' (e.g., Robbins-Monro and Kiefer-Wolfowitz), have proven their merit in the past (see, e.g., [8], [13], [14] for more information).

We have opted for the 'golden section search' algo-

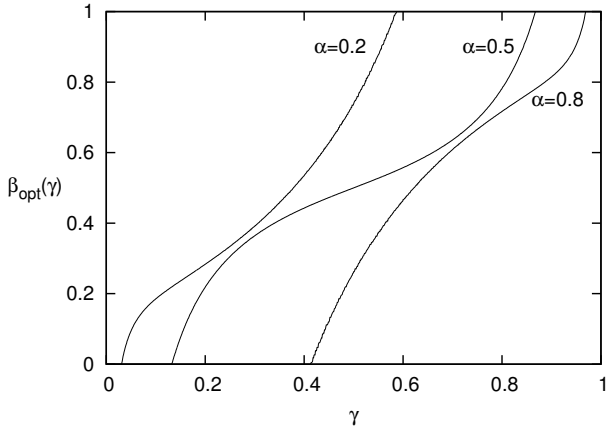


Fig. 5. Optimal values of $\beta$ as a function of $\gamma$, for $\lambda_T = 0.9$ and $g_j(x) = (x/\lambda_j)^2$ $(j = 1, 2)$

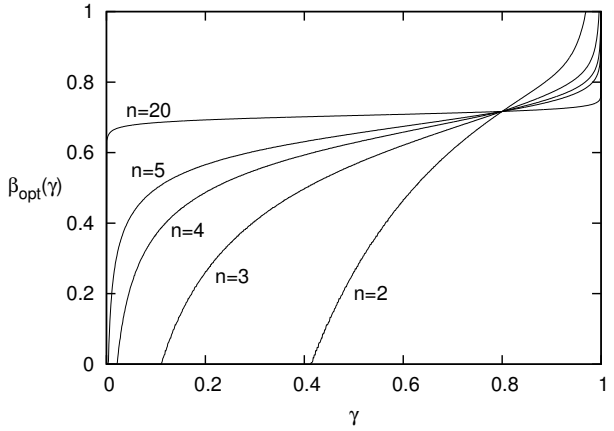


Fig. 6. Optimal values of $\beta$ as a function of $\gamma$, for $\alpha = 0.8$, $\lambda_T = 0.9$, and $g_j(x) = (x/\lambda_j)^n$ $(j = 1, 2)$

rithm. This algorithm is known to be reliable [15] and easy to implement. Convergence to the global minimum is relatively slow but certain. The algorithm [19] can be summarized in our setting as follows. In each iteration, we start with a triplet $(\beta_1, \beta_2, \beta_3)$ of values of $\beta$ $(\beta_1 < \beta_2 < \beta_3)$ with $F(\beta_2, \gamma) < F(\beta_1, \gamma)$ and $F(\beta_2, \gamma) < F(\beta_3, \gamma)$. So the minimum of $F(\beta, \gamma)$ lies inside the interval $[\beta_1, \beta_3]$ and the current estimation of this minimum is $\beta_2$. The next $\beta$ (say $\beta_4$) is chosen in the largest of the two intervals $[\beta_1, \beta_2]$ and $[\beta_2, \beta_3]$, at a distance of 0.38197 (golden section) times this interval from $\beta_2$. Then, $F(\beta, \gamma)$ is estimated for $\beta = \beta_4$ via Monte-Carlo simulation. Depending on the value of $F(\beta_4, \gamma)$ relative to that of $F(\beta_2, \gamma)$, the triplet $(\beta_1, \beta_2, \beta_3)$ is updated ($\beta_4$ becomes an element of this triplet), leading to a narrower search interval. As termination condition for this algorithm, we choose to test the gaps between the values of $\beta_1$ and $\beta_3$. The algorithm terminates when the relative accuracy bounds $|\beta_3 - \beta_1| < \tau$, where $\tau$ indicates a tolerance parameter. It is important to remark that estimates of $F(\beta, \gamma)$ are still obtained through coupled Monte-Carlo simulations for different values of $\beta$, as described in the previous section.

Fig. 5 illustrates the optimal values of $\beta$ as a function of $\gamma$, for $\lambda_T = 0.9$, and three different values of $\alpha$. Note that $\beta_{\text{opt}}(\gamma)$ was produced for 1001 equidistant values of $\gamma$ and that the tolerance parameter $\tau$ was assumed to be 0.0001. We observe that $\beta_{\text{opt}}(\gamma)$ increases the most for $\gamma$ close to either $\phi(0)$ or $\phi(1)$, which shows that if $\gamma \in ]\phi(0), \phi(1)[$, the optimal $\beta$ is likely to be not that close to 0, nor to 1. This will be more pronounced for general powers $n$, which we briefly comment on next.

*2) General $n$:* Fig. 6 illustrates the optimal values of $\beta$ as a function of $\gamma$, for $\alpha = 0.8$, $\lambda_T = 0.9$, and several values of $n$. The optimal values of $\beta$ are obtained in the same way as in the previous figure (i.e., with the 'golden section search' algorithm). We observe that the curves for $\beta_{\text{opt}}(\gamma)$ become steeper in the neighbourhoods of $\phi(0)$ and $\phi(1)$ with increasing $n$. This can be understood for $n \to \infty$, in particular. For $n \to \infty$, the value of $\gamma$ becomes meaningless; the minimization of the objective function comes down to the minimization of the maximum of the mean packet delays for both classes, i.e., the balancing of the mean packet delays. In other words, for

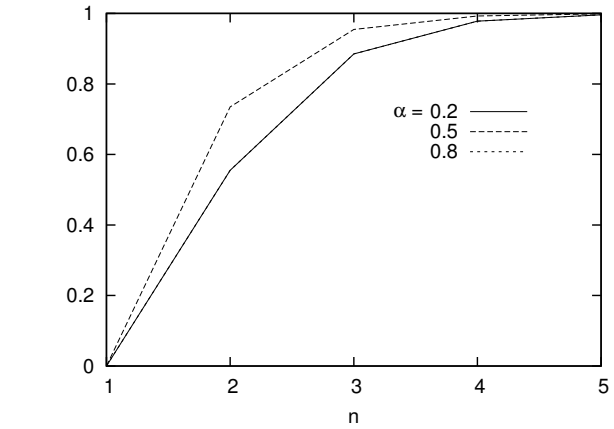


Fig. 7. $\phi(1) - \phi(0)$ as a function of $n$, for $\lambda_T = 0.9$ and $g_j(x) = (x/\lambda_j)^n$ $(j = 1, 2)$

$n \to \infty$, $\beta_{\text{opt}}(\gamma)$ becomes independent of $\gamma$ (except for $\gamma = 0$ and $\gamma = 1$ where $\beta = 0$ and $\beta = 1$ are still optimal) and this $\beta_{\text{opt}}(\gamma) \equiv \beta_{\text{opt}}$ makes $\bar{d}_1(\beta_{\text{opt}}) = \bar{d}_2(\beta_{\text{opt}})$. This effect is also observed in [20] in a different context. Note, furthermore, that $\beta_{\text{opt}}$ is in fact the same $\beta$ for which all curves in Figs. 2 and 4 intersect.

A second observation from Fig. 6 is that $\phi(0)$ decreases and $\phi(1)$ increases with $n$. This is also clear from Fig. 7, where we show the difference between $\phi(1)$ and $\phi(0)$ as a function of $n$, for $\lambda_T = 0.9$ and three different values of $\alpha$ (the curves for 0.2 and 0.8 are identical, due to the apparent symmetry). The difference indeed increases (rapidly) with $n$. For $n = 1$, the difference is 0, cf. subsection V-A; for $n \to \infty$, the difference tends to 1. So for larger $n$, GPS is more frequently the optimal scheduling discipline. For $\gamma = 0.2$ and $n = 2$, for example, strict priority (with $\beta = 0$) is optimal; for $\gamma = 0.2$ and $n = 3$, GPS is optimal. This basically means that, to determine the optimal $\beta$, the need for simulation increases when $n$ increases. To counterbalance this, the optimal $\beta$ becomes less sensitive to $\gamma$, if $\gamma$ is not too close to either 0 or 1 (see Fig. 6).

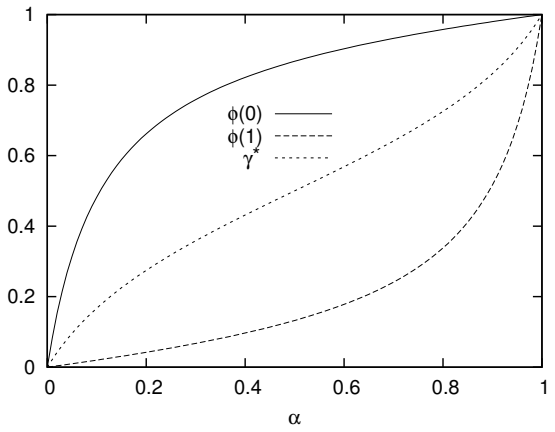


Fig. 8. $\phi(0)$, $\phi(1)$, and $\gamma^*$ as a function of $\alpha$, for $g_j(x) = \ln(x/\lambda_j)$ and $\lambda_T = 0.9$

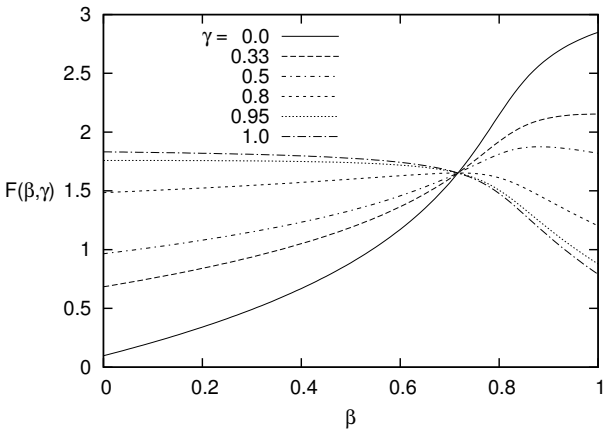


Fig. 9. Objective functions $F(\beta, \gamma)$ as a function of $\beta$, for $\alpha = 0.8$, $\lambda_T = 0.9$, and $g_j(x) = \ln(x/\lambda_j)$ $(j = 1, 2)$

### C. *Concave functions:* $g_j(x) = \ln(a_j x)$

We end this section with the case where the functions $g_j$ are concave increasing functions. In contrast with convex functions $g_j$, increments of low values of the considered performance measures are here more penalized than increments of high values. Assume $g_j(x) = \ln(a_j x)$ $(j = 1, 2)$, with $a_j$ a constant. In the same way as above, we can easily obtain formulas for $\phi(0)$ and $\phi(1)$. According to Theorem 3, $\phi(0) > \phi(1)$. This is demonstrated in Fig. 8, where we depict $\phi(0)$ and $\phi(1)$ as a function of $\alpha$, for $\lambda_T = 0.9$. When $\gamma \geq \phi(0)$, $\beta_{\text{opt}}(\gamma) = 1$; when $\gamma \leq \phi(1)$, $\beta_{\text{opt}}(\gamma) = 0$. When $\phi(1) < \gamma < \phi(0)$, it depends on the values of $F(0, \gamma)$ and $F(1, \gamma)$ whether $\beta = 0$ or $\beta = 1$ is optimal. By using equation (1) and expressions (8)-(11), we can determine the conditions on $\gamma$ so that $F(0, \gamma) < (>, =) F(1, \gamma)$. Indeed, solving the equation $F(0, \gamma) = F(1, \gamma)$ for $\gamma$ leads to a formula for the $\gamma$ for which both $\beta = 0$ and $\beta = 1$ are optimal. Let us denote this value of $\gamma$ by $\gamma^*$. The formula for $\gamma^*$ is omitted here because of its size. However, we illustrate the behaviour of $\gamma^*$ in Fig. 8. When $\gamma \geq \gamma^*$, $\beta_{\text{opt}}(\gamma) = 1$. When $\gamma \leq \gamma^*$, on the other hand, $\beta_{\text{opt}}(\gamma) = 0$.

Finally, Fig. 9 shows $F(\beta, \gamma)$ as a function of $\beta$, for $\alpha = 0.8$, $\lambda_T = 0.9$, and six different values of $\gamma$. Amongst those values of $\gamma$ are $\phi(0)$ $(= 0.95)$ and $\phi(1)$ $(= 0.33)$. For these system and objective function parameters, furthermore, $\gamma^* = 0.72$. We clearly see that when $\gamma < 0.72$, $F(\beta, \gamma)$ reaches its minimum in $\beta = 0$. For $\gamma > 0.72$, $\beta = 1$ minimizes $F(\beta, \gamma)$. Furthermore, we observe that the curves corresponding with $\gamma \in ]\phi(1), \phi(0)[$ reach a maximum for some $\beta \in ]0, 1[$, as reasoned in the proof of Theorem 3.

## VI. Conclusions

In this paper, we have obtained criteria to separate important objective functions for a two-class queueing system into two groups: either strict priority is optimal or Generalized Processor Sharing (GPS) is optimal. These criteria are formulated in terms of mean performance measures of the queueing system with strict priority, which can be calculated explicitly in many cases. This is the main contribution of our paper: in some important cases, we can exclude GPS as optimal scheduling discipline, without having to analyse the queueing system with GPS, which is, even for easy models, extremely hard. We have proved that strict priority is always optimal if the objective function is a weighted combination of linear or concave functions of the mean unfinished work (and/or mean holding times/packet delays). Furthermore, we have derived bounds on the weights in the objective function for which GPS is optimal, in case of a weighted combination of convex functions of the mean unfinished work. In case GPS is optimal, we have proposed a simulation algorithm based on 'common random numbers' and 'golden section search'. We have illustrated our approach with some important case studies, in case of an uncorrelated arrival process. However, this can be easily extended to correlated arrival processes; the only necessary result is the mean per-class unfinished work of the corresponding queueing model with strict priority scheduling.

In the future, we want to extend our hybrid analytical/simulation optimization approach. We are positive that more analytic results can aid in minimizing (the number of) simulations. For instance, in [28], derivatives of the mean unfinished work in $\beta = 0$ and $\beta = 1$ are calculated. We will investigate how these can be used to our advantage. Furthermore, the queueing model can be extended as well: more than two classes and general service times. Note that the latter is not as straightforward as it seems, as the relation (2) is no longer valid. Of course, the objective function can be generalized as well. A weighted combination of unimodal functions of the mean unfinished work with a minimum in target values, for instance, is practically interesting. Extensions with inclusion of higher moments of the unfinished work would be intriguing as well. Finally, we plan to consider other scheduling disciplines, such as Discriminatory Processor Sharing [2], [12] or multi-class Random-Order-of-Service [6].

## Appendix

*Proof of Proposition 1:* Since GPS is a work-conserving scheduling discipline, the total unfinished work in the system at the beginning of a slot is independent of this discipline and thus also of the value of $\beta$. Hence, the mean total unfinished work $\bar{w}_1(\beta) + \bar{w}_2(\beta)$ is independent of $\beta$ as well. ■

*Proof of Proposition 2:* This can be proved through a coupling argument. Assume two GPS systems, one with

$\beta = \beta_1$ and one with $\beta = \beta_2$. These systems are completely coupled as follows: both systems are empty at the beginning and the numbers of class-1 and class-2 arrivals are equal in each slot in both systems. All service times are identical as well. Furthermore, we define the i.i.d. continuous random variables $r_k$, $k \geq 1$, equal in both systems and uniformly distributed in $[0, 1]$. These variables are used to decide which queue to serve in slot $k$ if packets are present in both queues: if $r_k \leq \beta$, queue 1 is served; otherwise, queue 2 is served. We prove that $\bar{w}_2(\beta_1) < \bar{w}_2(\beta_2)$ if $\beta_1 < \beta_2$, i.e., $\bar{w}_2(\beta)$ is strictly increasing. Then Proposition 1 immediately implies that $\bar{w}_1(\beta)$ is a strictly decreasing function.

First, it is not hard to see that the mean unfinished work in queue 2 is non-decreasing w.r.t. $\beta$. For the completely coupled GPS systems, the unfinished work in queue 2 in the system with $\beta = \beta_2$ will always be at least as high as the unfinished work in queue 2 in the system with $\beta = \beta_1$. So $\bar{w}_2(\beta_1) \leq \bar{w}_2(\beta_2)$ for $0 \leq \beta_1 < \beta_2 \leq 1$. Secondly, we prove that $\bar{w}_2(\beta_1) < \bar{w}_2(\beta_2)$. For this, we will prove that $\text{Prob}[w_2(\beta_1) < w_2(\beta_2)] > 0$, with $w_2(\beta_j)$ the unfinished work in queue 2 at the beginning of a random slot in the system with $\beta = \beta_j$, $j = 1, 2$. Look at such a random slot. If $w_2(\beta_1) < w_2(\beta_2)$, we are done. If not, assume that we picked a slot where both queues contain packets (this occurs with positive probability, see the assumptions of the arrival process). So during the slot, one unit of work is performed, either of queue 1 or queue 2, depending on the value of a random variable $r$ (stationary version of $r_k$). If $\beta_1 < r < \beta_2$, a unit of work of queue 2 is performed in the system with $\beta = \beta_1$, whereas in the system with $\beta = \beta_2$, a unit of work of queue 1 is performed. This obviously leads to a difference in the unfinished work in queue 2 in both systems at the beginning of the next slot. Since $\text{Prob}[\beta_1 < r < \beta_2] = \beta_2 - \beta_1 > 0$, this occurs with positive probability, concluding this proof. ■

## ACKNOWLEDGMENT

This research has been co-funded by the Interuniversity Attraction Poles (IAP) Programme initiated by the Belgian Science Policy Office.

## REFERENCES

[1] I.J.B.F. Adan, O.J. Boxma, and J.A.C. Resing. Queueing models with multiple waiting lines. *Queueing Systems: Theory and Applications*, 37(1–3):65–98, 2001

[2] E. Altman, K. Avrachenkov, and U. Ayesta. A survey on discriminatory processor sharing. *Queueing Systems: Theory and Applications*, 53:53–63, 2010.

[3] S. Asmussen, P.W. Glynn. *Stochastic Simulation*. Springer, 2007.

[4] B. Ata and T.L. Olsen. Near-optimal dynamic lead-time quotation and scheduling under convex-concave customer delay costs. *Operations Research*, 57(3):753–768, 2009.

[5] B. Ata and M.H. Tongarlak. On scheduling a multiclass queue with abandonments under general delay costs. *Queueing Systems: Theory and Applications*, in press.

[6] U. Ayesta, A. Izagirre, and I.M. Verloop. Heavy traffic analysis of the discriminatory randomorderofservice discipline. *SIGMETRICS Performance Evaluation Review*, 39(2):41–43, 2011.

[7] C.F. Bispo. The single-server scheduling problem with convex costs. *Queueing Systems: Theory and Applications*, 73:261–294, 2013.

[8] R.P. Brent. *Algorithms for Minimization without Derivatives*. Prentice Hall, 1973.

[9] J.W. Cohen. Boundary value problems in queueing theory. *Queueing Systems: Theory and Applications*, 3(2):97–128, 1988.

[10] J.W. Cohen and O.J. Boxma. *Boundary value problems in queueing system analysis*. North-Holland, Amsterdam, 1983

[11] G. Fayolle and R. Iasnogorodski. Two coupled processors: the reduction to a Riemann-Hilbert problem. *Probability Theory and Related Fields*, 47(3):325–351, 1979.

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

[13] M. Heath. *Scientific Computing: An Introductory Survey, 2nd Edition*. McGraw-Hill, 2002.

[14] H.J. Kushner and G.G. Yin. *Stochastic Approximation and Recursive Algorithms and Applications, 2nd Edition*. Springer, 2003.

[15] K. Lange. *Numerical Analysis for Statisticians, 2nd Edition*. Springer, 2010.

[16] P. Lieshout and M. Mandjes. Generalized processor sharing: characterization of the admissible region and selection of optimal weights. *Computers & Operations Research*, 35:2497–2519, 2008.

[17] A.K. Parekh and R.G. Gallager. A generalised processor sharing approach to flow control in integrated services networks: the single-node case. *IEEE/ACM Transactions on Networking*, 1(3):344–357, 1993

[18] A.K. Parekh and R.G. Gallager. A generalised processor sharing approach to flow control in integrated services networks: the multiple-node case. *IEEE/ACM Transactions on Networking*, 2(2):137–150, 1994.

[19] W.H. Press, S.A. Teukolsky, W. T. Vetterling, and B.P. Flannery. *Numerical Recipes: The Art of Scientific Computing, 3rd Edition*. Camebridge University Press, 2007.

[20] B. Rengarajan, C. Caramanis, and G. de Veciana. Analyzing queuing systems with coupled processors through semidefinite programming. In *INFORMS: Applied Probability Session*, November 2008

[21] J.C. Spall. *Introduction to stochastic search and optimization — estimation, simulation and control*. Wiley, 2003.

[22] H. Takagi. *Queueing analysis: Vacation and priority systems, part 1*. North-Holland, 1991.

[23] J.A. Van Mieghem and P. Van Mieghem. Price-coupled scheduling for differentiated services: Gc versus GPS. *International Journal of Communication Systems*, 15(5):429–452, 2002.

[24] I.A. Verloop, U. Ayesta and S. Borst. Monotonicity properties for multi-class queueing systems. *Discrete Event Dynamic Systems - Theory and Applications*, 20(4):473–509, 2010.

[25] J. Walraevens, B. Steyaert, and H. Bruneel. Performance analysis of a GI-Geo-1 buffer with a preemptive resume priority scheduling discipline. *European Journal of Operational Research*, 157(1):130–151, 2004.

[26] J. Walraevens, S. Wittevrongel, and H. Bruneel. A discrete-time priority queue with train arrivals. *Stochastic Models*, 23(3):489–512, 2007.

[27] J. Walraevens, B. Steyaert, and H. Bruneel. Analysis of a discrete-time preemptive resume priority buffer. *European Journal of Operational Research*, 186(1):182–201, 2008.

[28] J. Walraevens, J.S.H. van Leeuwaarden, and O.J. Boxma. Power series approximations for two-class generalized processor sharing systems. *Queueing Systems: Theory and Applications*, 66(2):107–130, 2010.

[29] J. Walraevens, T. Maertens, and H. Bruneel. A semi-preemptive priority scheduling discipline: performance analysis. *European Journal of Operational Research*, 224(2):324–332, 2013.