On the Influence of High Priority Customers on a Generalized Processor Sharing Queue

Jasper Vanlerberghe, Joris Walraevens, Tom Maertens, 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
{jpvlerbe,jw,tmaerten,hb}@telin.UGent.be

Abstract. In this paper, we study a hybrid scheduling mechanism in discrete-time. This mechanism combines the well-known Generalized Processor Sharing (GPS) scheduling with strict priority. We assume three customer classes with one class having strict priority over the other classes, whereby each customer requires a single slot of service. The latter share the remaining bandwith according to GPS. This kind of scheduling is used in practice for the scheduling of jobs on a processor and in Quality of Service modules of telecommunication network devices. First, we derive a functional equation of the joint probability generating function of the queue contents. To explicitly solve the functional equation, we introduce a power series in the weight parameter of GPS. Subsequently, an iterative procedure is presented to calculate consecutive coefficients of the power series. Lastly, the approximation resulting from a truncation of the power series is verified with simulation results. We also propose rational approximations. We argue that the approximation performs well and is extremely suited to study these systems and their sensitivity in their parameters (scheduling weights, arrival rates, loads ...). This method provides a fast way to observe the behaviour of such type of systems avoiding time-consuming simulations.

Keywords: Generalized Processor Sharing (GPS), priority, queueing, scheduling, power series

1 Introduction

Numerous queueing systems in practice, have a high-priority bypass possibility. In this paper we study the influence of these high priority customers on a generalized processor sharing (GPS) queue. For instance, the processor of a computer system is shared by several jobs, whereby each class of jobs gets a time-share according to the weight of its class. However, the processor can also be interrupted, for hardware I/O for instance (i.e., the user pushes a key, requested data

from the harddisk becomes available ...), these are in fact short high-priority jobs, bypassing the normal scheduling mechanism.

An example from telecommunications is DiffServ [9]. DiffServ is short for Differentiated Services and is an architecture designed to deliver a different Quality of Service (QoS) grade to various services in telecommunication networks. It defines an Expedited Forwarding (EF) class of packets next to the Assured Forwarding (AF) class. EF packets have essentially high priority and are thus given strict priority over all other packets. The AF class of packets is divided into subclasses, and the scheduling amongst the subclasses is a GPS-based scheduling.

Cisco implemented this kind of scheduling mechanism in some of its gigabit switch routers. The brand names used are IP Realtime Transport Protocol (RTP) Priority and Low Latency Queueing (LLQ); both are based on a mixture of GPS-like scheduling with priority bypassing. They differ in the type of traffic they support, i.e., UDP vs TCP.

As a result of its practical application, this model also attracted attention from the research community, where it is frequently referred to as PQ-GPS. Jin et al. [4, 5] studied PQ-GPS under long-range dependent traffic by using a flow decomposition approach dividing the system into single-server single-queue (SSSQ) systems. They obtain analytical upper and lower bounds. Parveen [12] used the same SSSQ approach to study a system containing both long-range and short-range dependent traffic. After the single queue decomposition he however uses another technique resulting in a single approximation, as opposed to an upper and lower bound. Lastly, we mention Wang et al. [20] who studied a finite hybrid queueing model using PQ and Weighted Fair Queueing (WFQ). As WFQ is known to be a good approximation for GPS, it is also of interest here. Drawing up a Markov chain for the system and solving it for the steady-state probability, they conclude with a sensitivity analysis for the parameters of the system.

Next to studying hybrid scheduling models, most of the attention has gone to both individual models, i.e., either priority queueing or generalized processor sharing models. Priority queueing was, for instance, studied in [3,6,13,15,18,19]. Whereas, GPS was analyzed in [7, 8, 10, 11, 17, 21].

In this paper, we analyze a hybrid priority-GPS scheduling algorithm. We construct a functional equation for the probability generating function (pgf) of the queue contents in steady state. Subsequently, we develop an iterative procedure to calculate the coefficients of the power series of this pgf, whereby the power series is constructed in the GPS-weight. Due to practical restrictions, we use the truncated power series to construct approximations. Lastly, we evaluate the approximations using simulation results.

2 Mathematical model

We consider a discrete-time (i.e., time is assumed to be slotted) queueing system with three queues of infinite capacity and one transmission channel. Three classes of customers, named 1, 2 and 3, arrive to the system. Customers of class 1 have strict priority over the other customers. Consequently, the server always

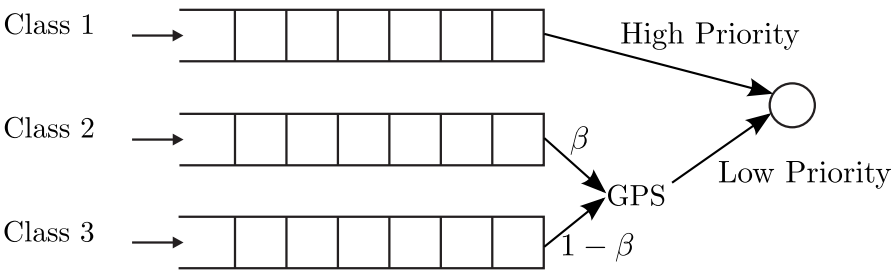

Fig. 1. Model

serves class 1 as long as this class is backlogged. If class 1 is not backlogged, class 2 and 3 customers are served according to a discrete-time implementation of GPS. As such, the server serves a class 2 customer with probability β and a class 3 customer with probability (w.p.) $1-\beta$, if both classes are backlogged. The weight parameter of the GPS scheduling is thus β and can be used to divide the bandwith among customers of class 2 and 3. Within each queue, the customers are served in FIFO order. This model is depicted in Fig. 1.

The number of arrivals of class j ($j = 1, 2, 3$) in slot k is denoted by $a_{j,k}$, where we assume $\{a_{j,k}, k > 0\}$ forms a sequence of independent and identically distributed random variables. The joint pgf of the arrivals of all classes is denoted as $A(z_1, z_2, z_3) \triangleq E[z_1^{a_{1,k}} z_2^{a_{2,k}} z_3^{a_{3,k}}]$. Furthermore, we define λ_j as the mean number of arrivals in queue j and λ_T as the mean total number of arrivals to the queueing system per time slot. Every customer requires a single slot of service. This means that the load ρ (i.e., the mean number of slots of work arriving to the system per slot) equals λ_T; subsequently, the stability condition for this queueing system is $\lambda_T < 1$.

In the next sections, we study the stationary distribution of the queue content in each of the queues. Therefore, we define $u_{j,k}$ as the queue content in queue j at the beginning of slot k and $U_k(z_1, z_2, z_3) \triangleq E[z_1^{u_{1,k}} z_2^{u_{2,k}} z_3^{u_{3,k}}]$ as the joint pgf of the queue content at the beginning of slot k. The stationary distribution is then $U(z_1, z_2, z_3) = \lim_{k\to\infty} U_k(z_1, z_2, z_3)$.

3 The functional equation

Let us first establish the system equations, relating $(u_{1,k}, u_{2,k}, u_{3,k})$ and $(u_{1,k+1}, u_{2,k+1}, u_{3,k+1})$, i.e., the state of the system at the beginning of slot k and the state of the system at slot $k+1$. We split the equations into several (sub)cases:

- **All queues empty,** i.e., $u_{j,k} = 0, j = 1, 2, 3$:

$$(u_{1,k+1}, u_{2,k+1}, u_{3,k+1}) = (a_{1,k}, a_{2,k}, a_{3,k}) \tag{1}$$

- **Queue 1 not empty,** i.e., $u_{1,k} > 0$:

$$(u_{1,k+1}, u_{2,k+1}, u_{3,k+1}) = (u_{1,k} - 1 + a_{1,k}, u_{2,k} + a_{2,k}, u_{3,k} + a_{3,k}) \tag{2}$$

– **Queue 1 empty,** i.e., $u_{1,k} = 0$:
  - queue 2 empty and queue 3 not empty i.e., $u_{2,k} = 0, u_{3,k} > 0$:

$$(u_{1,k+1}, u_{2,k+1}, u_{3,k+1}) = (a_{1,k}, a_{2,k}, u_{3,k} - 1 + a_{3,k}) \tag{3}$$

  - queue 2 not empty and queue 3 empty, i.e., $u_{2,k} > 0, u_{3,k} = 0$:

$$(u_{1,k+1}, u_{2,k+1}, u_{3,k+1}) = (a_{1,k}, u_{2,k} - 1 + a_{2,k}, a_{3,k}) \tag{4}$$

  - queue 2 and 3 both not empty, i.e., $u_{2,k} > 0, u_{3,k} > 0$:

$$(u_{1,k+1}, u_{2,k+1}, u_{3,k+1}) = \begin{cases} (a_{1,k}, u_{2,k} - 1 + a_{2,k}, u_{3,k} + a_{3,k}) \text{ w.p. } \beta \\ (a_{1,k}, u_{2,k} + a_{2,k}, u_{3,k} - 1 + a_{3,k}) \text{ w.p. } 1 - \beta \end{cases} \tag{5}$$

From these systems equations, we construct a relation between the pgfs $U_k(z_1, z_2, z_3)$ and $U_{k+1}(z_1, z_2, z_3)$:

$$\begin{aligned} U_{k+1}(z_1, z_2, z_3) = A(z_1, z_2, z_3) \Bigg(& U_k(0,0,0) \\ & + \frac{1}{z_1}\big(U_k(z_1, z_2, z_3) - U_k(0, z_2, z_3)\big) \\ & + \frac{1}{z_3}\big(U_k(0, 0, z_3) - U_k(0, 0, 0)\big) \\ & + \frac{1}{z_2}\big(U_k(0, z_2, 0) - U_k(0, 0, 0)\big) \\ & + \left(\frac{\beta}{z_2} + \frac{1-\beta}{z_3}\right)\Big(U_k(0, z_2, z_3) \\ & - U_k(0, 0, z_3) - U_k(0, z_2, 0) + U_k(0, 0, 0)\Big)\Bigg). \end{aligned} \tag{6}$$

In steady state, both U_k and U_{k+1} are equal. We denote $U(z_1, z_2, z_3) \triangleq \lim_{k\to\infty} U_k(z_1, z_2, z_3) = \lim_{k\to\infty} U_{k+1}(z_1, z_2, z_3)$ as the pgf of the queue content in steady state. By letting $k \to \infty$ in Equation (6) and solving the result for $U(z_1, z_2, z_3)$, we retrieve the following functional equation for $U(z_1, z_2, z_3)$:

$$U(z_1, z_2, z_3) = \frac{A(z_1, z_2, z_3)\left\{\begin{array}{l} \Big(z_2(z_1 - z_3) + \beta z_1(z_3 - z_2)\Big)U(0, z_2, z_3) \\ +(1-\beta)z_1(z_3 - z_2)U(0, z_2, 0) \\ -\beta z_1(z_3 - z_2)\Big)U(0, 0, z_3) \\ +\Big(z_1 z_3(z_2 - 1) + \beta z_1(z_3 - z_2)U(0, 0, 0) \end{array}\right\}}{z_2 z_3(z_1 - A(z_1, z_2, z_3))} \tag{7}$$

This functional equation still contains some unknowns that need to be determined to obtain full knowledge of the statistical distribution of the queue length.

Therefore, we need to calculate the unknown boundary functions $U(0, z_2, z_3)$, $U(0, z_2, 0)$, $U(0, 0, z_3)$ and $U(0, 0, 0)$. This last unknown is easily found as $U(0, 0, 0)$ and equals the probability that the system is empty in steady state ($u_1 = u_2 = u_3 = 0$). In queueing theory, this is a well-known result and is equal to $1 - \lambda_T$. For ease of notation, however, we will only do this substitution after eliminating the other boundary functions.

4 The Power Series Approximation

To eliminate the boundary functions, we write $U(z_1, z_2, z_3)$ as a power series in β, where we assume $U(z_1, z_2, z_3)$ is analytic in a neighborhood of $\beta = 0$. This approach was also used in [17] to analyze a two-queue GPS system. We write:

$$U(z_1, z_2, z_3) = \sum_{m=0}^{\infty} V_m(z_1, z_2, z_3)\beta^m. \tag{8}$$

In the remainder of this section, we use this power series and the functional equation from the previous section to derive an iterative procedure to calculate V_m from V_{m-1}.

4.1 Eliminating $V_m(0, 0, z_3)$

The first step is to replace $U(z_1, z_2, z_3)$ by its power series in (7). Subsequently, we can equate the coefficients of β^m on the right and left hand side of Equation (7). For the coefficient of $\beta^m, m \geq 0$, this yields

$$\begin{aligned}(z_2 z_3(z_1 - A(z_1, z_2, z_3))V_m(z_1, z_2, z_3) \qquad\qquad (9)\\ = A(z_1, z_2, z_3)\Big[z_1(z_3 - z_2)\Big(P_{m-1}(z_2, z_3) + V_m(0, z_2, 0) + V_{m-1}(0, 0, 0)\Big)\\ + z_2(z_1 - z_3)V_m(0, z_2, z_3) + z_1 z_3(z_2 - 1)V_m(0, 0, 0)\Big],\end{aligned}$$

where we defined $V_{-1}(z_1, z_2, z_3) \triangleq 0$ and $P_m(z_2, z_3) = V_m(0, z_2, z_3) - V_m(0, z_2, 0) - V_m(0, 0, z_3)$ and thus $P_{-1}(z_2, z_3) = 0$. Looking closely at Equation (9), we can see that only two of the remaining unknown boundary functions $V_m(0, z_2, 0)$ and $V_m(0, z_2, z_3)$ are needed to calculate $V_m(z_1, z_2, z_3)$, assuming $V_{m-1}(z_1, z_2, z_3)$ is known. By introducing the power series we effectively eliminated one of the unknown boundary functions.

4.2 Eliminating $V_m(0, z_2, z_3)$

By using a generalization of Rouché's theorem [1], we can prove that $z_1 - A(z_1, z_2, z_3)$ has one zero in the unit disk of z_1 for an arbitrary z_2 and z_3 in the unit disk. We denote this zero by $Y_{2,3}(z_2, z_3)$ and it is thus implicitly defined as $Y_{2,3}(z_2, z_3) = A(Y_{2,3}(z_2, z_3), z_2, z_3)$, with $|Y_{2,3}(z_2, z_3)| < 1$. As the left hand

side of Equation (9) is zero for $z_1 = Y_{2,3}(z_2, z_3)$ and $V_m(z_1, z_2, z_3)$ remains finite in the unit circle, the right hand side should also equal zero. This leads to

$$\begin{aligned} &z_2(z_3 - Y_{2,3}(z_2, z_3))V_m(0, z_2, z_3) \\ &\quad = Y_{2,3}(z_2, z_3)(z_3 - z_2)\Big(P_{m-1}(z_2, z_3) + V_m(0, z_2, 0) + V_{m-1}(0, 0, 0)\Big) \\ &\qquad + Y_{2,3}(z_2, z_3)z_3(z_2 - 1)V_m(0, 0, 0). \end{aligned} \tag{10}$$

4.3 Eliminating $V_m(0, z_2, 0)$

We can prove that $Y_{2,3}(z_2, z_3)$ is the pgf of a random variable of this system, see [18] for a similar example. Then by again using Rouché's theorem, we can prove that $z_3 - Y_{2,3}(z_2, z_3)$ has one zero in the unit disk of z_3 for an arbitrary z_2 in the unit disk. We denote this zero by $Y_2(z_2)$ and it is thus implicitly defined as $Y_2(z_2) = Y_{2,3}(z_2, Y_2(z_2)) = A(Y_2(z_2), z_2, Y_2(z_2))$, with $|Y_2(z_2)| < 1$. As the left hand side of Equation (10) is zero for $z_3 = Y_2(z_2)$ and $V_m(0, z_2, z_3)$ remains finite in the unit circle, the right hand side should also equal zero. This yields

$$V_m(0, z_2, 0) = -P_{m-1}(z_2, Y_2(z_2)) - V_{m-1}(0, 0, 0) + \frac{Y_2(z_2)(z_2 - 1)V_m(0, 0, 0)}{z_2 - Y_2(z_2)}. \tag{11}$$

Feeding this result back into Equation (10), we get that

$$\begin{aligned} &z_2(z_3 - Y_{2,3}(z_2, z_3))V_m(0, z_2, z_3) \\ &\quad = Y_{2,3}(z_2, z_3)(z_3 - z_2)\Big(Q_{m-1}(z_2, z_3) + \frac{Y_2(z_2)(z_2 - 1)V_m(0, 0, 0)}{z_2 - Y_2(z_2)}\Big), \end{aligned} \tag{12}$$

with

$$\begin{aligned} Q_m(z_2, z_3) &= P_m(z_2, z_3) - P_m(z_2, Y_2(z_2)) \\ &= V_m(0, z_2, z_3) - V_m(0, z_2, Y_2(z_2)) - V_m(0, 0, z_3) + V_m(0, 0, Y_2(z_2)). \end{aligned} \tag{13}$$

Lastly, as $U(0, 0, 0) = 1 - \lambda_T$ (shown before), we know that $V_0(0, 0, 0) = 1 - \lambda_T$ and $V_m(0, 0, 0) = 0$ for $m > 0$.

So by introducing the power series notation and the two implicitly defined functions $Y_{2,3}$ and Y_2, we found a solution for the boundary functions. Substituting, these solutions in Equation (9), we get (with $m > 0$) that

$$V_0(z_1, z_2, z_3) = \frac{(1 - \lambda_T)A(z_1, z_2, z_3)(z_2 - 1)(z_3 - Y_2(z_2))(z_1 - Y_{2,3}(z_2, z_3))}{(z_2 - Y_2(z_2))(z_3 - Y_{2,3}(z_2, z_3))(z_1 - A(z_1, z_2, z_3))}, \tag{14}$$

$$V_m(z_1, z_2, z_3) = \frac{A(z_1, z_2, z_3)(z_3 - z_2)Q_{m-1}(z_2, z_3)(z_1 - Y_{2,3}(z_2, z_3))}{z_2(z_3 - Y_{2,3}(z_2, z_3))(z_1 - A(z_1, z_2, z_3))}. \tag{15}$$

As a result, starting from V_0, V_m can be calculated from V_{m-1}. This concludes the iterative calculation procedure of $U(z_1, z_2, z_3)$.

As a test of our analysis, suppose we would want to study the joint probability generating function of u_1 and $u_2 + u_3$. We can do this by replacing both z_2 and z_3 by z, as $E[z_1^{u_1} z^{u_2+u_3}] = U(z_1, z, z)$. We subsequently get:

$$V_0(z_1, z, z) = \frac{(1 - \lambda_T)A(z_1, z, z)(z - 1)(z_1 - Y_{2,3}(z, z))}{(z - Y_{2,3}(z, z))(z_1 - A(z_1, z, z))}, \tag{16}$$

$$V_m(z_1, z, z) = 0. \tag{17}$$

As V_m equals zero for $m > 0$, $U(z_1, z, z) = V_0(z_1, z, z)$ and the pgf is independent of β, as expected. The result we get, is the pgf for a priority queueing system with 2 queues as can be found in [19]. This confirms our result.

5 Approximations of performance measures

In the previous section, we derived an iterative algorithm to calculate the joint pgf $U(z_1, z_2, z_3)$ of the queue content. More practical performance measures of the system, however, would for instance be the mean length of each of the three queues. These can be calculated from the power-series form of the pgf

$$\begin{aligned} E[u_j] &= \left.\frac{\partial U(z_1, z_2, z_3)}{\partial z_j}\right|_{z_1=z_2=z_3=1} \\ &= \sum_{m=0}^{\infty} \beta^m \left.\frac{\partial V_m(z_1, z_2, z_3)}{\partial z_j}\right|_{z_1=z_2=z_3=1}. \end{aligned} \tag{18}$$

We showed earlier that $V_m(z_1, 1, 1) = 0$ for $m > 0$, so $E[u_1] = V_0(1, 1, 1)$ is independent of β. This is of course expected, as the length of the high-priority queue should not depend on the scheduling of the packets of the lower priority queues.

A second conclusion follows from the fact that in the work conserving system presented here, the total backlog is a constant. This constant $E[u_T]$ is independent of β. As a result, we get:

$$E[u_T] = E[u_1] + E[u_2] + E[u_3], \tag{19}$$

$$E[u_T] - E[u_1] = E[u_2] + E[u_3] \tag{20}$$

$$= \sum_{m=0}^{\infty} \beta^m \left.\frac{\partial V_m(1, z_2, 1)}{\partial z_2}\right|_{z_2=1} + \sum_{m=0}^{\infty} \beta^m \left.\frac{\partial V_m(1, 1, z_3)}{\partial z_3}\right|_{z_3=1}. \tag{21}$$

The terms in the left hand side are constants, while $E[u_2]$ and $E[u_3]$ on the right hand side of the equation are a function of β, as can be seen from Equation (18). Subsequently, this means that for $m > 0$:

$$\left.\frac{\partial V_m(1, z_2, 1)}{\partial z_2}\right|_{z_2=1} = - \left.\frac{\partial V_m(1, 1, z_3)}{\partial z_3}\right|_{z_3=1}. \tag{22}$$

This result can significantly help speed up calculations, as we only need to calculate one of those derivatives.

With these results, we are able to calculate the exact mean queue lengths, or at least to an arbitrary precision. This is however only theoretically possible. In practice, the calculation of V_m is far from straightforward. The calculation of Q_{m-1} in (15) involves $V_m(0, z_2, Y_2(z_2))$ and $V_m(0, 0, Y_2(z_2))$, for which several applications of l'Hopital's rule are needed. The differentiation in l'Hopital's rule leads to very large expressions, quickly becoming infeasable for current computers. Calculating the mean queue length involves another differentiation and evaluation in 1 for all $z_j, j = 1..3$, leading to several more applications of l'Hopital's rule.

We, however, have another trick up our sleeve. We can also calculate the power series in $\beta = 1$ leading to:

$$U(z_1, z_2, z_3) = \sum_{m=0}^{\infty} (1-\beta)^m \tilde{V}_m(z_1, z_2, z_3) \tag{23}$$

So, because of the symmetry in the system, $\tilde{V}_m$ can be calculated from V_m, whereby class 3 customers are sent to queue 2 and class 2 customers to queue 3. In particular, $\tilde{V}_m$ can be calculated from Equation (15) with $A(z_1, z_2, z_3)$ replaced by $A(z_1, z_3, z_2)$. Subsequently, the mean lengths of queues 2 and 3 can be calculated as

$$E[u_j] = \sum_{m=0}^{\infty} \left. \frac{\partial \tilde{V}_m(z_1, z_2, z_3)}{\partial z_{5-j}} \right|_{z_1=z_2=z_3=1}, \qquad j = 2, 3. \tag{24}$$

Basically, in practice we can calculate the first M terms of the power series of $E[u_2]$ and $E[u_3]$, either in $\beta = 0$ or in $\beta = 1$, from the functions V_0 up to V_M. With these values we can construct approximations. We opt to approximate $E[u_2]$ and $E[u_3]$ by rational functions (Padé approximants) of the form

$$[L/N]_{E[u_j]}(\beta) = \frac{\sum_{l=0}^{L} v_{j,l} \beta^l}{\sum_{n=0}^{N} w_{j,n} \beta^n}, \tag{25}$$

whereby the coefficients $v_{j,l}$ and $w_{j,n}$ should be chosen such that the derivatives of $[L/N]_{E[u_j]}(\beta)$ in either 0 or 1 match the values obtained before. For $[L/N]_{E[u_j]}(\beta)$ to be unique, we need a normalization. Therefore, we choose $w_{j,0} = 1$. As we have $2(M+1)$ datapoints and $L + N + 1$ coefficients in $[L/N]_{E[u_j]}(\beta)$, we need to choose L and N such that $L + N = 2M + 1$.

The Padé approximants can introduce difficulties as the denominator can introduce poles for $\beta \in [0, 1]$. Furthermore, the result could be non-monotone; however, the mean queue length of class 2 (class 3) should decrease (increase) in β. Lastly, the performance of each approximant is different and varies with the parameters of the arrival process, so it is unclear which one performs best beforehand (see also the numerical examples in the next section). These problems are identical to the ones in [16], the solution presented therein can also be used

here to overcome these problems. This solution (in short) consists of disregarding the unfeasible approximants and averaging the remaining ones. As to keep this text self-contained and simple, we will restrict the discussion here to the Padé approximants (and in the remainder do not use the solution from [16].

6 Numerical Examples

In this section, we will compare our power series approximation for the mean queue length with simulation results. As the mean queue length for class 1 is not influenced by the other queues and could also easily be calculated from results for single-class FCFS queueing, we will not discuss it here. Furthermore, we only analyze queue 2, as the system is work conserving, results for queue 3 follow easily from (19).

We will use an arrival process with a joint pgf of the number of arrivals of the three classes of the form

$$A(z_1, z_2, z_3) = \left(1 + \frac{\lambda_1}{16}(z_1 - 1) + \frac{\lambda_2}{16}(z_2 - 1) + \frac{\lambda_3}{16}(z_3 - 1)\right)^{16}, \qquad (26)$$

where λ_j is the arrival rate of class j customers (as defined earlier). Furthermore, we define $\alpha_1 = \frac{\lambda_1}{\lambda_T}$ and $\alpha_2 = \frac{\lambda_2}{\lambda_T}$ as the fraction of class 1 and class 2 customers, respectively.

For the simulation results in this section, we have used Monte-Carlo simulations over 10^7 slots. This high number of slots is enough to eliminate bias from the transient phase. Additionally, each simulation uses exactly the same sequence of arrivals and decision variables, to minimize the variance between simulations for different parameters of the system. This is the well-known technique of the common random numbers [2, 14].

In Fig. 2, we show the mean length of queue 2 as a function of the weight β, with $\lambda_T = 0.9, \alpha_1 = 0.1$, and $\alpha_2 = 0.1$. The figure shows curves of the simulation result and the Padé approximants without poles. We can see that for these parameters the [2/3] Padé approximant is very accurate.

Secondly, we observe that the approximations perform best close to $\beta = 0$ and $\beta = 1$. This is expected as the available information is exactly the value up to the M-th order derivative in these points (in this case $M = 2$). Subsequently, the approximants are constructed to match this information, thus performing well near $\beta = 0$ and $\beta = 1$.

In our second numerical example, we study the influence of the amount of high-priority (i.e., class 1) customers. We keep the total load $\lambda_T = 0.9$ fixed and $\lambda_2 = \lambda_3$, while increasing α_1 from 0.1 to 0.6. The mean queue-2 length is depicted in Fig. 3 on the left, showing both the simulation results and the best performing Padé approximant. We can see that the performance of the approximation is still accurate though slightly deteriorates as α_1 decreases, this results from the choice of the approximant. For this graph, we chose the [3/2] approximant, which on average performs best for these curves, but for smaller α_1 the [2/3] approximant is actually better. Furthermore for $\beta = 1$, i.e., when the

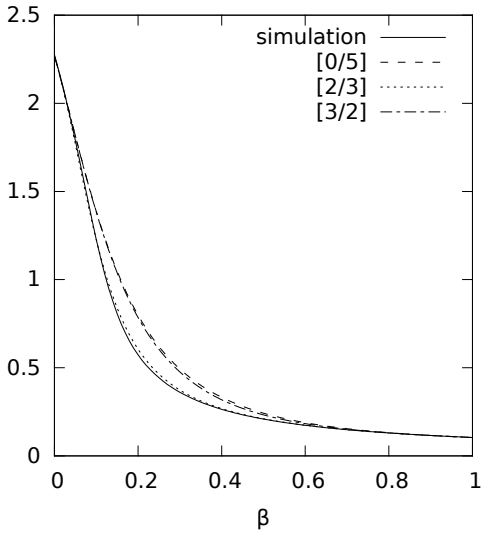

Fig. 2. Mean queue-2 length: comparison between simulation and Padé approximants.

queueing system is effectively a strict priority system with class 1 having highest priority, class 2 medium priority and class 3 low priority, higher α_1 barely makes a difference. This is mainly because there are few class 2 customers in the system as α_2 decreases from 0.1 to 0.056. On the other end for $\beta = 0$, we have a strict priority queueing system with class 1 high priority, class 3 medium priority and class 2 low priority. As class 2 is the lowest on the priority ladder, the influence of the bypassing (higher priority) class 3 and class 1 customers is greater. With α_2 small, however, queueing rarely happens and the influence is rather small.

Using Little's theorem, we also calculated the mean class-2 delay, it is depicted in Fig. 3 on the right. We saw before that as α_1 increases the mean queue-2 length decreases, mainly because α_2 decreases (we keep the total load and ratio between class-2 and 3 packets fixed). As we can see from the graph of the delay, for an increasing amount of high priority packets the class-2 packets have a larger delay. There are thus less class-2 packets in the system but they stay there longer.

In Fig. 4, we show $E[u_2]$ as a function of β for different values of the total load λ_T, with $\alpha_1 = \alpha_2 = 0.1$ fixed. As the load in the system increases, we observe the queue-2 length increases as well. This is a classical queueing result: a higher load always leads to higher congestion. As in the previous example (and for the same reason), we can see the effect at $\beta = 1$ is barely visible as opposed to at $\beta = 0$. Furthermore, we see that approximation is close to the simulated result. For $\lambda_T = 0.99$, we only depicted the approximation. Simulations over 10^7 slots do not converge for this high load, as the event of the system being empty becomes very rare.

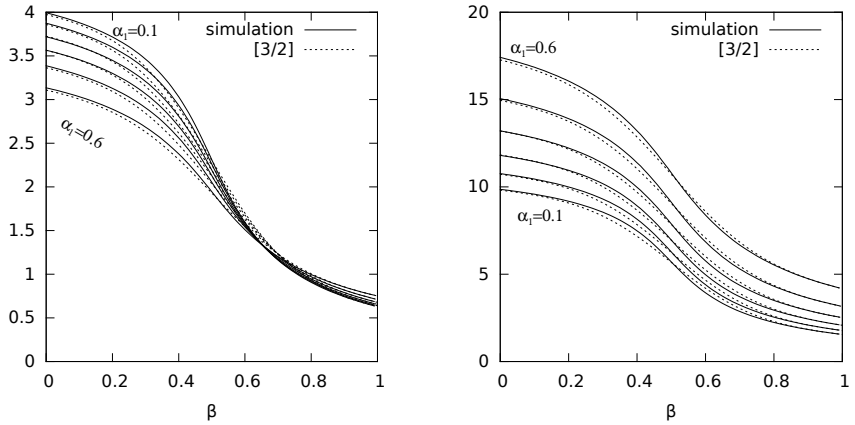

Fig. 3. Mean queue-2 length (left) and mean queue-2 delay (right): effect of increasing fraction of class 1 customers

Lastly, we look at the influence of the amount of class-2 customers while keeping the total load and the amount of high-priority packets constant. The results are depicted in Fig. 5 for $\lambda_T = 0.9, \alpha_1 = 0.1$ and α_2 ranging from 0.1 to 0.5. As the amount of class-2 packets increases the queue length increases, which was to be expected. Another observation is that the performance of the approximation deteriorates. In Fig. 5, we chose to show the [2/3] approximant. This is, however, not the best approximation for every parameter combination. For instance, for $\alpha_2 = 0.5$ Padé approximant [3/2] is the best one. However, even if we compare every simulation with the best fitting approximant, the performance still deteriorates.

7 Conclusions

In this paper, we derived an analytical method to calculate the joint probability generating function of a three-class queueing system with a hybrid GPS-priority scheduling. The iterative algorithm leads to solutions with arbitrary precision in theory. Unfortunately, in practice, we are limited by the capabilities of current computers in the derivation of performance measures. Using Padé approximants, we have presented a method to use partial information to construct approximations. These approximations were compared with results from simulation and prove to work well. As a result, this power series approximation leads to a very efficient method to study these kind of systems for the whole parameter space, avoiding very time and resource consuming simulations.

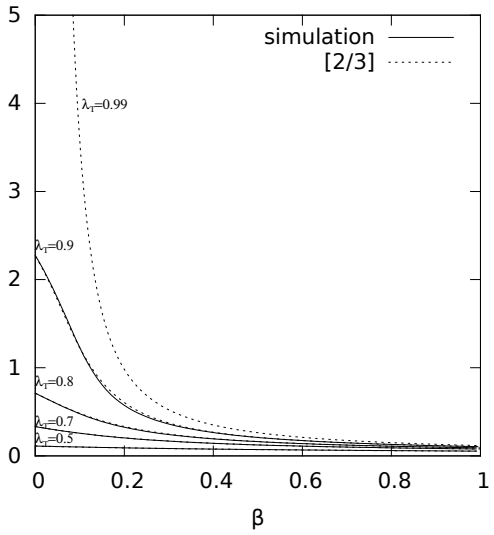

Fig. 4. Mean queue-2 length: effect of increasing total load

Acknowledgement

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

References

1. Adan, I.J., Van Leeuwaarden, J., Winands, E.M.: On the application of Rouché's theorem in queueing theory. Operations Research Letters 34(3), 355–360 (2006)
2. Asmussen, S., Glynn, P.W.: Stochastic Simulation: Algorithms and Analysis: Algorithms and Analysis, vol. 57. Springer (2007)
3. Choi, B., Choi, D., Lee, Y., Sung, D.: Priority queueing system with fixed-length packet-train arrivals. IEE Proceedings-Communications 145(5), 331–336 (1998)
4. Jin, X., Min, G.: Analytical modelling of hybrid PQ-GPS scheduling systems under long-range dependent traffic. In: Advanced Information Networking and Applications, 2007. AINA'07. 21st International Conference on. pp. 1006–1013. IEEE (2007)
5. Jin, X., Min, G.: Performance modelling of hybrid PQ-GPS systems under long-range dependent network traffic. Communications Letters, IEEE 11(5), 446–448 (2007)
6. Kim, K., Chae, K.C.: Discrete-time queues with discretionary priorities. European Journal of Operational Research 200(2), 473–485 (JAN 16 2010)
7. Lee, J.Y., Kim, S., Kim, D., Sung, D.K.: Bandwidth optimization for internet traffic in generalized processor sharing servers. Parallel and Distributed Systems, IEEE Transactions on 16(4), 324–334 (2005)

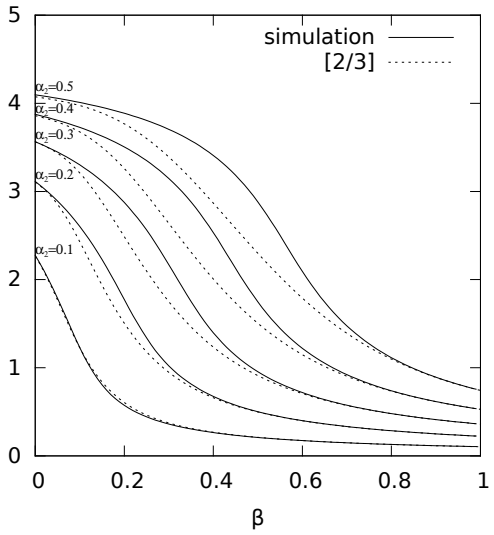

Fig. 5. Mean queue-2 length: effect of increasing fraction of class 2 customers

8. Lieshout, P., Mandjes, M.: Generalized processor sharing: Characterization of the admissible region and selection of optimal weights. Computers & Operations Research 35(8), 2497–2519 (2008)
9. Nichols, K., Blake, S., Baker, F., Black, D.: Definition of the differentiated services field (DS field) in the IPv4 and IPv6 headers. RFC 2474 (Proposed Standard) (dec 1998), http://www.ietf.org/rfc/rfc2474.txt, updated by RFCs 3168, 3260
10. Parekh, A.K., Gallager, R.G.: A generalized processor sharing approach to flow control in integrated services networks: the single-node case. IEEE/ACM Transactions on Networking (TON) 1(3), 344–357 (1993)
11. Parekh, A.K., Gallagher, R.G.: A generalized processor sharing approach to flow control in integrated services networks: the multiple node case. IEEE/ACM Transactions on Networking (TON) 2(2), 137–150 (1994)
12. Parveen, A.S.: A survey of an integrated scheduling scheme with long-range and short-range dependent traffic. International Journal of Engineering Sciences & Research Technology 3(1), 430–439 (2014)
13. Smith, P.J., Firag, A., Dmochowski, P.A., Shafi, M.: Analysis of the M/M/N/N queue with two types of arrival process: Applications to future mobile radio systems. Journal of Applied Mathematics 2012 (2012)
14. Spall, J.C.: Introduction to stochastic search and optimization: estimation, simulation, and control, vol. 65. John Wiley & Sons (2005)
15. Takine, T., Sengupta, B., Hasegawa, T.: An analysis of a discrete-time queue for broadband ISDN with priorities among traffic classes. Communications, IEEE Transactions on 42(234), 1837–1845 (1994)
16. Vanlerberghe, J., Walraevens, J., Maertens, T., Bruneel, H.: Approximating the optimal weights for discrete-time generalized processor sharing. In: Networking Conference, 2014 IFIP. pp. 1–9. IEEE (2014)

17. Walraevens, J., van Leeuwaarden, J., Boxma, O.: Power series approximations for two-class generalized processor sharing systems. Queueing systems 66(2), 107–130 (2010)
18. Walraevens, J., Steyaert, B., Bruneel, H.: Delay characteristics in discrete-time GI-G-1 queues with non-preemptive priority queueing discipline. Performance Evaluation 50(1), 53–75 (2002)
19. Walraevens, J., Steyaert, B., Bruneel, H.: Performance analysis of a single-server ATM queue with a priority scheduling. Computers & Operations Research 30(12), 1807–1829 (2003)
20. Wang, L., Min, G., Kouvatsos, D.D., Jin, X.: Analytical modeling of an integrated priority and WFQ scheduling scheme in multi-service networks. Computer Communications 33, S93–S101 (2010)
21. Zhang, Z.L., Towsley, D., Kurose, J.: Statistical analysis of the generalized processor sharing scheduling discipline. Selected Areas in Communications, IEEE Journal on 13(6), 1071–1080 (1995)