The discrete-time preemptive repeat identical priority queue

Joris Walraevens*, Dieter Fiems and Herwig Bruneel

SMACS Research Group

Department of Telecommunications and Information Processing (IR07)

Ghent University - UGent

Sint-Pietersnieuwstraat 41, B-9000 Gent, Belgium.

Phone: +32 9 264 89 02, Fax: +32 9 264 42 95

E-mail: {jw,df,hb}@telin.UGent.be

Abstract

Priority queueing systems come natural when customers with diversified delay requirements have to wait to get service. The customers that cannot tolerate but small delays get service priority over customers which are less delay-sensitive. In this contribution, we analyze a discrete-time two-class preemptive repeat identical priority queue with infinite buffer space and generally distributed service times. Newly arriving high-priority customers interrupt the on-going service of a low-priority customer. After all high-priority customers have left the system, the interrupted service of the low-priority customer has to be repeated completely. By means of a probability generating functions approach, we analyze the system content and the delay of both types of customers. Performance measures (such as means and variances) are calculated and the impact of the priority scheduling is discussed by means of some numerical examples.

Key words: preemptive repeat priority, discrete-time queue

*Corresponding author

1 Introduction

In this paper, we present the analysis of a discrete-time preemptive repeat identical (PRI) priority queue. Time is divided into slots and the initiation of service is synchronized with respect to slot boundaries. Customers of two classes (class-1 and class-2) arrive in a single-server queueing system and the customers of class-1 are scheduled for service with *priority* over class-2 customers. So, when the server becomes available, a class-1 customer is served next (if any). If no class-1 customers are present, a class-2 customer starts service (if any). The scheduling type is *preemptive* which means that newly arriving class-1 customers interrupt an on-going service of a class-2 customer. Furthermore, an interrupted class-2 customer has to *repeat* its complete service upon returning in the server (after all class-1 customers have left the system). In the preemptive repeat *identical* priority queue, the service time of a particular class-2 customer is the same in all its service attempts (as opposed to the preemptive repeat different (PRD) priority queue where the service time is resampled in each new service attempt, see e.g. [1]).

Although this type of priority scheduling is mentioned - not analyzed - in earlier works like [2,3] and in e.g. the standard work of [4], analyses of queues with this type of priority scheduling discipline are much more scarce than analyses considering its more popular non-preemptive (NP) and preemptive resume (PR) counterparts. One of the reasons is that its analysis is much more involved than that of the NP, PR (and PRD) priority queues. This is basically because it is a *non-work-conserving* scheduling discipline (as opposed to the NP and PR priority disciplines), since an interrupted customer's service time has to be fully repeated. In other words, the load incorporates not only the service time of class-2 customers, but also their possible repeats. As such, the PRD can be seen as a 'simplified' version of the PRI priority discipline (from the viewpoint of the queueing analyst), since one can omit keeping track of the service time of the class-2 customer in service in the PRD priority queue as service times are resampled whenever interrupted (as opposed to PRI).

Continuous-time queues with a preemptive repeat priority discipline are analyzed in a.o. [5–8]. In [5] a PRD queue is analyzed in the context of a database system. The arrival process is assumed to be a Poisson process, while the service times are generally distributed. The use of a PRI priority scheduling discipline in CSMA-CD protocols for fiber optical bus networks is described in [6]. A general finite number (M) of stations, each with an infinite queueing capacity, are connected by an optical bus network. A station has priority of accessing the bus network over its downstream stations possibly overwriting information of the downstream stations. Therefore, this is modeled by a PRI priority queue with M priority classes. The arrival processes are modeled by Poisson processes. The transient behavior of a PRI priority queue with a Poisson arrival process and exponential service times using the randomization solution form and lattice path combinatorics is analyzed in [7]. Finally, an unslotted optical MAN (Metropolitan Area Network) ring operating with asynchronous variable

length optical packets is modeled as a PRI priority queue in [8]. In [1, 9, 10], discrete-time queues with a preemptive repeat priority scheduling discipline are studied. A preemptive repeat protocol for voice-data integration in a ring-based LAN is studied in [9]. A number of stations is connected by a ring network and each station is either a voice or a data station. Voice stations can overwrite the information of the data stations. The data stations can only put data on the network when no information of the other stations is passing by. This system is thus modeled as a preemptive repeat priority queueing system with deterministic service times (a number of slots so that the lengths of the service times equal the round trip time) for the low-priority customers. Note that the data and voice queues are all analyzed separately and that the queueing model for the low-priority queue is given by a queueing model with *service interruptions*. In [10], preemptive priority queues are studied. This is also done by means of queues with service interruptions. The technique of the *effective service times* is proposed for analyzing PR, PRD and PRI queues commonly. The interruptions are incorporated in the service times of the customers. These effective service times are obtained for the three priority disciplines separately, but once these are calculated a common queueing analysis is performed. This approach however requires that the number of arrivals of different classes in one slot are mutually independent. Finally, a discrete-time two-class PRD queue is analyzed in [1]. The numbers of arrivals are i.i.d. from slot-to-slot and the service times are generally distributed. Using probability generating functions, the moments of the system content and delay of both classes are obtained.

In this paper, we analyze the system content and delay of class-1 (high-priority) and class-2 (low-priority) customers in a discrete-time single-server buffer with a PRI priority scheme and per-slot i.i.d. arrivals. The numbers of class-1 and class-2 arrivals in a slot may however be correlated random variables (which is not the case in [9, 10]). This type of arrival correlation is for instance implied when the total numbers of arrivals in a slot has some natural bound. In this case more class-1 arrivals in a slot implies less possible class-2 arrivals. Notice that such bounds follow naturally from e.g. limits on the total number of packet sources or of inlets to a switch in a telecommunication system. This type of correlation also occurs when an incoming job (customer) can be split into two parts, an urgent and an non-urgent part, where the urgent part is of class-1, while the non-urgent part is of class-2. Taking this correlation in the arrival process into account leads to more accurate performance results, which is one of the contributions of the current paper. Furthermore, the service times of the customers are assumed to be generally distributed (in [9], the service times are deterministic). These distributions are class-dependent, i.e., the service times of the class-1 customers can be different from those of the class-2 customers (which reflects the case where different classes represent different applications). We will demonstrate that an analysis based on probability generating functions and on the supplementary variable technique is extremely suitable for modeling this type of buffers with a priority scheduling discipline and for calculating the relevant performance measures.

The remainder of this paper is structured as follows. In the following section, we present the mathematical model under consideration. In sections 3 and 4, the steady-state system content and customer delay of both classes are analyzed whereas section 5 concerns the calculation of various moments of the steady-state stochastic variables. We discuss the obtained results in section 6 and some numerical examples illustrate our approach in section 7. Finally, some conclusions are formulated in section 8.

2 Mathematical model

We consider a discrete-time single-server system with infinite buffer space. There are two types of customers arriving to the system, namely customers of class-1 and customers of class-2. The numbers of arrivals of class- j during slot k constitute a series of i.i.d. random variables and are denoted by $a_{j,k}$ ($j = 1, 2$). The common joint probability generating function (pgf) of $a_{1,k}$ and $a_{2,k}$ is defined as

$$A(z_1, z_2) \triangleq \mathbb{E} [z_1^{a_{1,k}} z_2^{a_{2,k}}] . \quad (1)$$

The marginal pgf's $A(z, 1)$ and $A(1, z)$ of the number of per-slot arrivals of class-1 and class-2 are denoted by $A_1(z)$ and $A_2(z)$ respectively. We will furthermore denote the mean arrival rate of class- j customers by $\lambda_j \triangleq \mathbb{E}[a_{j,k}] = A'_j(1)$ ($j = 1, 2$).

The service times of the class- j customers constitute a series of i.i.d. random variables and their pgf is denoted by $S_j(z)$ ($j = 1, 2$). The mean service time of a class- j customer is denoted by $\mu_j = S'_j(1)$ ($j = 1, 2$). For further use, we define the arrival loads of class-1 and class-2 as $\rho_1 = \lambda_1 \mu_1$ and $\rho_2 = \lambda_2 \mu_2$ respectively and the total arrival load as $\rho_T = \rho_1 + \rho_2$.

The class-1 customers are assumed to have preemptive repeat identical priority over the class-2 customers and within one class the scheduling is FCFS.

3 System content

In this paper, we use the *supplementary variable technique* (see [11]) to analyze the PRI priority queue. We denote the system content of class- j customers at the beginning of slot k by $u_{j,k}$ ($j = 1, 2$). The set $\{(u_{1,k}, u_{2,k}), k \geq 1\}$ does not form a Markov-chain in the case of general service times for both classes. Therefore, we introduce some additional stochastic variables. Firstly, we define r_k as follows: r_k indicates the residual service time at the beginning of slot k , i.e., the remaining number of slots needed to serve the customer in service from the beginning of slot k on, if $u_{T,k} > 0$, and $r_k = 0$ if $u_{T,k} = 0$; $u_{T,k} \triangleq u_{1,k} + u_{2,k}$ denotes the total system content at the beginning of slot k . Secondly, we define $t_{2,k}$ as the complete service time of the oldest class-2 customer at the beginning of slot k ,

if $u_{2,k} > 0$, and $t_{2,k} = 0$ if $u_{2,k} = 0$. (The oldest class- j customer in the system at a certain time instant is defined as the class- j customer that arrived first of all the class- j customers present in the system at that time instant.) This is necessary, since this customer's service can be interrupted and afterwards its complete service has to be repeated. With these definitions, $\{(r_k, u_{1,k}, t_{2,k}, u_{2,k}), k \geq 1\}$ forms a Markov-chain. A sample of the time-axis is shown in Figure 1 to demonstrate the PRI priority scheduling discipline and the stochastic variables involved. In this example, a class-2 service time of 4 slots is preempted in slot $k - 2$ by a newly arriving class-1 customer (with a service time of 3 slots) and is repeated after this class-1 customer's service time.

[Figure 1 about here.]

We first calculate the joint pgf of the steady-state version of $(r_k, u_{1,k}, t_{2,k}, u_{2,k})$. Let $s_{j,k}^*$ ($j = 1, 2$) denote the service time of the next class- j customer to receive service after slot k . The following system equations are established. Firstly we have

$$\begin{aligned} u_{1,k+1} &= u_{1,k} - \mathbf{1}_{r_k=1 \wedge u_{1,k} > 0} + a_{1,k}; \\ u_{2,k+1} &= u_{2,k} - \mathbf{1}_{r_k=1 \wedge u_{1,k}=0 \wedge u_{2,k} > 0} + a_{2,k} \end{aligned}$$

for the system content of class-1 and class-2. Here, $\mathbf{1}_X$ denotes the indicator function of X , which evaluates to 1 if X is true and to 0 if X is false. Indeed, the system content at the beginning of slot $k + 1$ equals the superposition of the system content at the beginning of the previous slot and the new arrivals during that slot minus the possible customer in service at the beginning of slot k , provided it was his last service slot (i.e., $r_k = 1$). When there were class-1 customers in the system at the beginning of slot k , the leaving customer is of class-1, otherwise he is a class-2 customer. Furthermore, we have

$$t_{2,k+1} = \begin{cases} 0 & \text{if } u_{2,k+1} = 0 \\ s_{2,k}^* & \text{if } u_{2,k+1} > 0 \wedge (u_{2,k} = 0 \vee (u_{2,k} > 0 \wedge u_{1,k} = 0 \wedge r_k = 1)) \\ t_{2,k} & \text{otherwise} \end{cases}.$$

This is understood as follows. First, $t_{2,k+1} = 0$ iff $u_{2,k+1} = 0$, by definition. Otherwise, the oldest class-2 customer at the beginning of slot $k + 1$ is a 'new' oldest class-2 customer when he arrived during slot k or when the 'old' oldest class-2 customer has left the system at the end of slot k . These

cases are summarized by the condition in the second line of the above expression. Finally, we find

$$r_{k+1} = \begin{cases} 0 & \text{if } u_{1,k+1} = u_{2,k+1} = 0 \\ r_k - 1 & \text{if } r_k > 1 \wedge \neg(u_{1,k} = 0 \wedge u_{1,k+1} > 0) \\ s_{1,k}^* & \text{if } u_{1,k+1} > 0 \wedge (r_k \leq 1 \vee (r_k > 1 \wedge u_{1,k} > 0)) \\ t_{2,k+1} & \text{if } u_{1,k+1} = 0 \wedge u_{2,k+1} > 0 \wedge r_k \leq 1 \end{cases}.$$

Firstly, we notice that $r_{k+1} = 0$ iff $u_{1,k+1} = u_{2,k+1} = 0$ by definition. The customer in service during slot k is still served during slot $k+1$ if $r_k > 1$, except when class-1 customers arrive in slot k while a class-2 customer is in service. In that case, the latter customer's service is interrupted. When a new customer enters the server at the beginning of slot k the remaining service time equals $s_{1,k}^*$ or $t_{2,k+1}$, depending on whether it is a class-1 or class-2 customer respectively.

We assume that the system is stable and will analyze the system in steady-state (i.e., for $k \rightarrow \infty$). Note that obtaining the stability condition of this system is not straight-forward due to the repeats of class-2 service times (e.g., it is not just $\rho_T < 1$). We will comment on the stability condition later in subsection 6.1. Defining

$$P(x, z_1, y_2, z_2) \triangleq \lim_{k \rightarrow \infty} E[x^{r_k} z_1^{u_{1,k}} y_2^{t_{2,k}} z_2^{u_{2,k}}],$$

using the system equations and letting $k \rightarrow \infty$, leads to the following expression for $P(x, z_1, y_2, z_2)$:

$$\begin{aligned} P(x, z_1, y_2, z_2) = & \frac{1}{x - A(z_1, z_2)} \left\{ xA(0, 0)(1 - S_2(xy_2) + S_1(x)(S_2(y_2) - 1)) \right. \\ & \times [P(0, 0, 0, 0) + R_1(0, 0, 0) + R_2(1, 0)] \\ & + [A(0, z_2)(xS_2(xy_2) - 1 + xS_1(x)(1 - S_2(y_2))) \\ & + (A(z_1, z_2) - A(z_1, 0))(xS_1(x) - 1)(S_2(y_2) - 1)]P(0, 0, 0, 0) \\ & + xA(z_1, 0)S_1(x)(1 - S_2(y_2))R_2(1, 0) \\ & + xA(0, z_2)[S_2(xy_2) - 1 + S_1(x)(1 - S_2(y_2))]R_1(0, 0, 0) \\ & + x(A(z_1, z_2) - A(z_1, 0))(S_1(x) - z_1)(S_2(y_2) - 1)R_1(z_1, 0, 0) \\ & + (A(z_1, z_2) - A(z_1, 0))(S_2(y_2) - 1)P(x, z_1, 0, 0) + xA(0, z_2)S_2(xy_2)R_2(1, z_2) \\ & + x(A(z_1, z_2) - A(0, z_2))S_1(x)S_2(y_2)R_2(1, z_2) - xA(0, z_2)S_1(x)R_1(0, y_2, z_2) \\ & - xz_2[A(0, z_2) + (A(z_1, z_2) - A(0, z_2))S_1(x)]R_2(y_2, z_2) \\ & + x(A(z_1, z_2) - A(0, z_2))S_1(x)P(1, 0, y_2, z_2) - (A(z_1, z_2) - A(0, z_2))P(x, 0, y_2, z_2) \\ & \left. + xA(0, z_2)R_1(0, xy_2, z_2) + xA(z_1, z_2)(S_1(x) - z_1)R_1(z_1, y_2, z_2) \right\}, \end{aligned} \quad (2)$$

with

$$R_1(z_1, y_2, z_2) \triangleq \lim_{k \rightarrow \infty} \mathbb{E} \left[z_1^{u_{1,k}-1} y_2^{t_{2,k}} z_2^{u_{2,k}} \mathbf{1}_{r_k=1 \wedge u_{1,k} > 0} \right]$$

$$R_2(y_2, z_2) \triangleq \lim_{k \rightarrow \infty} \mathbb{E} \left[y_2^{t_{2,k}} z_2^{u_{2,k}-1} \mathbf{1}_{r_k=1 \wedge u_{1,k}=0} \right].$$

In the remainder, we retrieve an expression for P for its arguments inside the unit circle (the ultimate expression is given in (11)). We therefore have to determine the boundary functions $P(x, z_1, 0, 0)$, $P(x, 0, y_2, z_2)$, $R_1(z_1, y_2, z_2)$ and $R_2(y_2, z_2)$ and the unknown constants $P(0, 0, 0, 0)$, $R_1(0, 0, 0)$ and $R_2(1, 0)$. This can be done in a few steps. Some of these steps are however quite standard and are therefore omitted here. We refer to [1, 12] for similar analyses. We give a short sketch of the different steps and go into more detail when describing parts of the analysis which are specific for the PRI priority queue.

Firstly, the following expressions can be found for $P(x, z_1, 0, 0)$ and $P(x, 0, y_2, z_2)$ by replacing z_2 (z_1 respectively) by 0 in equation (2) and by using the definition of r_k and $t_{2,k}$:

$$P(x, z_1, 0, 0) = \frac{\left\{ \begin{array}{l} [x(1 - S_1(x)) + A(z_1, 0)(xS_1(x) - 1)]P(0, 0, 0, 0) \\ + xA(z_1, 0)[S_1(x)R_2(1, 0) + (S_1(x) - z_1)R_1(z_1, 0, 0)] \end{array} \right\}}{x - A(z_1, 0)} \quad (3)$$

$$P(x, 0, y_2, z_2) = \frac{\left\{ \begin{array}{l} [x(1 - S_2(xy_2)) + A(0, z_2)(xS_2(xy_2) - 1)]P(0, 0, 0, 0) \\ + xA(0, z_2)[S_2(xy_2)R_2(1, z_2) - z_2R_2(y_2, z_2) \\ + (S_2(xy_2) - 1)R_1(0, 0, 0) + R_1(0, xy_2, z_2)] \end{array} \right\}}{x - A(0, z_2)}. \quad (4)$$

Substituting these expressions in expression (2) allows us to eliminate $P(x, z_1, 0, 0)$, $P(x, 0, y_2, z_2)$ and $P(1, 0, y_2, z_2)$.

Secondly, we extract as much information as possible from expression (2). This can be done by substituting a) x by $A(z_1, z_2)$ and b) x by $A(Y_1(z_2), z_2)$ and z_1 by $Y_1(z_2)$ in expression (2), by using the property that pgfs are bounded for their arguments inside and on the unit circle and (in step b)) also using Rouché's theorem. Y_1 is a pgf defined by

$$Y_1(z) \triangleq S_1(A(Y_1(z), z)). \quad (5)$$

These two substitutions yield two extra equations in the unknown functions.

Thirdly, we return to expressions (3) and (4). Since these are obviously also bounded for their arguments inside the unit circle, similar techniques can be used as before. Substituting a) x by $A(z_1, 0)$ and b) x by $A(Y_1(0), 0)$ and z_1 by $Y_1(0)$ in expression (3), $P(x_1, z_1, 0, 0)$ is found as a function of $P(0, 0, 0, 0)$. Furthermore, substituting x by $A(0, z_2)$ in expression (4), the following functional

equation is found for R_2 in terms of the remaining unknown constant $P(0, 0, 0, 0)$:

$$R_2(A(0, z_2)y_2, z_2) = \frac{\left\{ \begin{aligned} &(1 - A(0, z_2))S_2(A(0, z_2)y_2)[(A(Y_1(z_2), z_2) - 1)P(0, 0, 0, 0) \\ &+ A(Y_1(z_2), z_2)R_2(1, z_2)] + A(0, z_2)z_2(A(Y_1(z_2), z_2) - 1)R_2(y_2, z_2) \end{aligned} \right\}}{z_2(A(Y_1(z_2), z_2) - A(0, z_2))}. \quad (6)$$

Defining the following partial conditional pgf

$$R_{2,i}(z_2) \triangleq \mathbb{E} [z_2^{u_2-1} \mathbf{1}_{r=1 \wedge u_1=0} | t_2 = i],$$

where r , u_1 , t_2 and u_2 are the steady state versions of r_k , $u_{1,k}$, $t_{2,k}$ and $u_{2,k}$ respectively. $R_2(y_2, z_2)$ is expressed as a function of the $R_{2,i}(z_2)$ as follows:

$$R_2(y_2, z_2) = \sum_{i=1}^{\infty} \text{Prob}[t_2 = i | r = 1 \wedge u_1 = 0] y_2^i R_{2,i}(z_2). \quad (7)$$

Note that a class-2 customer leaves the system at the end of slots at the beginning of which $r = 1$ and $u_1 = 0$. Since t_2 equals the service time of the customer that leaves the system (the oldest one) and since every customer leaves the system just once, we find

$$\text{Prob}[t_2 = i | r = 1 \wedge u_1 = 0] = s_2(i),$$

with $s_2(i)$ the probability mass function of the class-2 service times. Expression (7) thus becomes

$$R_2(y_2, z_2) = \sum_{i=1}^{\infty} s_2(i) y_2^i R_{2,i}(z_2). \quad (8)$$

Substituting this expression in expression (6) yields

$$\begin{aligned} &z_2(A(Y_1(z_2), z_2) - A(0, z_2)) \sum_{i=1}^{\infty} s_2(i) (A(0, z_2)y_2)^i R_{2,i}(z_2) \\ &= (1 - A(0, z_2)) \sum_{i=1}^{\infty} s_2(i) (A(0, z_2)y_2)^i \\ &\quad \times [(A(Y_1(z_2), z_2) - 1)P(0, 0, 0, 0) + A(Y_1(z_2), z_2)R_2(1, z_2)] \\ &\quad + A(0, z_2)z_2(A(Y_1(z_2), z_2) - 1) \sum_{i=1}^{\infty} s_2(i) y_2^i R_{2,i}(z_2), \end{aligned}$$

where we have also substituted $S_2(A(0, z_2)y_2)$ by its power series expansion. Since this equation has to hold for all y_2 ($|y_2| \leq 1$), the coefficients of y_2^i in both sides of the expression have to be equal (and

this for all i). This leads to

$$R_{2,i}(z_2) = \frac{(1 - A(0, z_2))A(0, z_2)^i[(A(Y_1(z_2), z_2) - 1)P(0, 0, 0, 0) + A(Y_1(z_2), z_2)R_2(1, z_2)]}{z_2[(A(Y_1(z_2), z_2) - A(0, z_2))A(0, z_2)^i - A(0, z_2)(A(Y_1(z_2), z_2) - 1)]},$$

$i \geq 1$. Substituting this in (8) yields

$$R_2(y_2, z_2) = \frac{B(y_2, z_2)}{z_2} \left(\frac{A(Y_1(z_2), z_2) - 1}{A(Y_1(z_2), z_2)} P(0, 0, 0, 0) + R_2(1, z_2) \right), \quad (9)$$

with

$$B(y, z) \triangleq \sum_{i=1}^{\infty} \frac{s_2(i)y^i(1 - A(0, z))A(Y_1(z), z)A(0, z)^i}{(A(Y_1(z), z) - A(0, z))A(0, z)^i - A(0, z)(A(Y_1(z), z) - 1)}. \quad (10)$$

Substituting y_2 by 1 in expression (9), gives an expression of $R_2(1, z_2)$ as a function of $P(0, 0, 0, 0)$.

Since all unknown functions and constants in expression (2) are now basically found as functions of $P(0, 0, 0, 0)$, we find the following expression for $P(x, z_1, y_2, z_2)$

$$\begin{aligned} P(x, z_1, y_2, z_2) = & P(0, 0, 0, 0) \left[1 + xz_1 \frac{(A(z_1, 0) - A(Y_1(0), 0))(S_1(A(z_1, 0)) - S_1(x))(S_2(y_2) - 1)}{A(Y_1(0), 0)(x - A(z_1, 0))(z_1 - S_1(A(z_1, 0)))} \right. \\ & + xz_1 \frac{(A(z_1, z_2) - A(Y_1(z_2), z_2))(S_1(x) - S_1(A(z_1, z_2)))(z_2 B(y_2, z_2) - S_2(y_2)B(1, z_2))}{A(Y_1(z_2), z_2)(x - A(z_1, z_2))(z_1 - S_1(A(z_1, z_2)))(z_2 - B(1, z_2))} \\ & \left. + xz_2 \frac{Q(x, y_2, z_2)}{A(Y_1(z_2), z_2)(x - A(0, z_2))(z_2 - B(1, z_2))} \right]. \quad (11) \end{aligned}$$

Finally, in order to find an expression for $P(0, 0, 0, 0)$, we use the normalization condition $P(1, 1, 1, 1) =$

1. Substituting x_1, z_1, x_2 and z_2 by 1 and using de l'Hôpital's rule in equation (11), we obtain

$$P(0, 0, 0, 0) = 1 - \rho_{T,eff}, \quad (12)$$

with

$$\rho_{T,eff} \triangleq \rho_1 + \lambda_2 \mu_{2,eff},$$

and

$$\mu_{2,eff} \triangleq \frac{S_2(1/A_1(0)) - 1}{1/A_1(0) - 1}. \quad (13)$$

$\rho_{T,eff}$ denotes the effective load (including repeats of class-2 customers' service times) offered to the system. Using this result in equation (11), we finally obtain $P(x, z_1, y_2, z_2)$.

From this joint pgf $P(x, z_1, y_2, z_2)$ several marginal pgf's can be calculated. The most important one is the joint pgf of the system content of class-1 and class-2 customers:

$$U(z_1, z_2) \triangleq \mathbb{E}[z_1^{u_1} z_2^{u_2}] = P(1, z_1, 1, z_2) \\ = (1 - \rho_{T,eff}) \frac{Y_2(z_2)(z_2 - 1)}{z_2 - Y_2(z_2)} \left\{ 1 + z_1 \frac{(A(z_1, z_2) - A(Y_1(z_2), z_2))(S_1(A(z_1, z_2)) - 1)}{A(Y_1(z_2), z_2)(A(z_1, z_2) - 1)(z_1 - S_1(A(z_1, z_2)))} \right\}, \quad (14)$$

with $Y_2(z)$ given by

$$Y_2(z) \triangleq B(1, z) = \sum_{i=1}^{\infty} \frac{s_2(i)(1 - A(0, z))A(Y_1(z), z)A(0, z)^i}{(A(Y_1(z), z) - A(0, z))A(0, z)^i - A(0, z)(A(Y_1(z), z) - 1)}. \quad (15)$$

We give a stochastic interpretation of this function $Y_2(z)$ and of the function $Y_1(z)$ - which are pgfs - in section 6.

From the two-dimensional pgf $U(z_1, z_2)$, we derive the expressions for the marginal pgf of the total, class-1 and class-2 system content, respectively given by

$$U_T(z) \triangleq \lim_{k \rightarrow \infty} \mathbb{E}[z^{u_{1,k} + u_{2,k}}] = U(z, z) \\ = (1 - \rho_{T,eff}) \frac{Y_2(z)(z - 1)}{z - Y_2(z)} \left\{ 1 + z \frac{(A_T(z) - A(Y_1(z), z))(S_1(A_T(z)) - 1)}{A(Y_1(z), z)(A_T(z) - 1)(z - S_1(A_T(z)))} \right\} \quad (16)$$

$$U_1(z) \triangleq \lim_{k \rightarrow \infty} \mathbb{E}[z^{u_{1,k}}] = U(z, 1) \\ = (1 - \rho_1) \frac{S_1(A_1(z))(z - 1)}{z - S_1(A_1(z))} \quad (17)$$

$$U_2(z) \triangleq \lim_{k \rightarrow \infty} \mathbb{E}[z^{u_{2,k}}] = U(1, z) \\ = (1 - \rho_{T,eff}) \frac{A_2(z)(A(Y_1(z), z) - 1)}{A(Y_1(z), z)(A_2(z) - 1)} \frac{Y_2(z)(z - 1)}{z - Y_2(z)}. \quad (18)$$

One may note that the pgf of the class-1 system content corresponds to that of a single-class FIFO queue. This is expected since the preemptive scheduling implies that the presence of low-priority customers does not influence the performance of the high-priority customers.

4 Delay

In this section, we analyze the delay of customers, i.e., the number of slots customers stay in the system. Since the class-1 characteristics in preemptive priority queues are independent of the class-2 customers, the delays of class-1 customers in preemptive (repeat) priority queues are identical to the delays in a single-class queue. The common pgf of these steady-state delays can e.g. be found in [13]:

$$D_1(z) = \frac{1 - \rho_1}{\lambda_1} \frac{S_1(z)(z - 1)}{z - A_1(S_1(z))} \frac{A_1(S_1(z)) - 1}{S_1(z) - 1}. \quad (19)$$

In the remainder of this section, we focus on the calculation of the common pgf of the class-2 delays. We tag a random class-2 customer that enters the buffer, say during slot k . We use the notion of *sub-busy periods* to analyze the class-2 delay. Two different kinds of sub-busy periods are defined, i.e., *sub-busy periods initiated by a class-1 customer* and *sub-busy periods initiated by a class-2 customer*. The first type is defined as follows: it starts at the beginning of the slot the initiating class-1 customer enters the server and ends when the number of class-1 customers in the system is one less (for the first time) than when the initiating class-1 customer entered the server. A sub-busy period initiated by a class-2 customer starts at the beginning of the slot the initiating class-2 customer enters the server (for the first time) and it ends at the beginning of which a new class-2 customer can enter the server (if there is any).

Let us refer to the customers in the system at the end of slot k , but that have to be served before the tagged customer as the “primary customers”. So, the tagged class-2 customer enters the server for the first time, when *all primary customers* and *all class-1 customers that arrived after slot k* (i.e., while the tagged customer is waiting in the queue) are served. All primary class- j customers (except for the oldest class-1 and class-2 customer) add a complete class- j sub-busy period to the delay of the tagged customer. Let $\tilde{v}_{j,m}$ denote the length of the sub-busy period added to the tagged customer’s delay by the m -th class- j customer already in the queue at the beginning of slot k and let $v_{j,m}^{(i)}$ denote the length of the sub-busy period added to the delay of the tagged class-2 customer by the m -th class- j customer that arrives during slot i . Finally, $\tilde{v}_2$ denotes the sub-busy period initiated by the tagged class-2 customer itself. During the tagged customer’s arrival slot, the system is in one of the following states:

1. $u_{1,k} = u_{2,k} = 0$:

$$d_2 = \sum_{j=1}^2 \sum_{m=1}^{f_{j,k}} v_{j,m}^{(k)} + \left(\tilde{v}_2 - \sum_{m=1}^{a_{1,k+d_2}} v_{1,m}^{(k+d_2)} \right), \quad (20)$$

with $f_{j,k}$ the number of class- j customers arriving during slot k , but that have to be served before the tagged customer. All of the $f_{j,k}$ class- j primary customers ($j = 1, 2$) add a class- j sub-busy period to d_2 . This leads to the first term of the right-hand side of equation (20). The delay further includes the sub-busy period of the tagged customer, minus the sub-busy periods initiated by the class-1 customers arriving in the slot preceding the departure of the tagged customer (slot $k + d_2$). This accounts for the negative part in the last term of the right-hand side of the above expression.

2. $u_{1,k} = 0, u_{2,k} > 0$:

$$d_2 = (v_{2,k}^+ - 1) + \sum_{m=1}^{u_{2,k}-1} \tilde{v}_{2,m} + \sum_{m=1}^{f_{2,k}} v_{2,m}^{(k)} + \left(\tilde{v}_2 - \sum_{m=1}^{a_{1,k}+d_2} v_{1,m}^{(k+d_2)} \right), \quad (21)$$

with $v_{2,k}^+$ the remaining part of the sub-busy period initiated by the oldest class-2 customer at the beginning of slot k . The difference with the former case, is that (multiple) class-2 customers are present in the system when the tagged class-2 customer arrives and thus have to be served before the tagged customer. All these class-2 customers initiate their own sub-busy periods. Notice further that the sub-busy periods added by the $f_{1,k}$ class-1 customers that arrive during slot k are part of $v_{2,k}^+$.

3. $u_{1,k} > 0, u_{2,k} = 0$:

$$d_2 = (r_k - 1) + \sum_{i=1}^{r_k-1} \sum_{m=1}^{a_{1,k}+i} v_{1,m}^{(k+i)} + \sum_{m=1}^{u_{1,k}-1} \tilde{v}_{1,m} + \sum_{j=1}^2 \sum_{m=1}^{f_{j,k}} v_{j,m}^{(k)} + \left(\tilde{v}_2 - \sum_{m=1}^{a_{1,k}+d_2} v_{1,m}^{(k+d_2)} \right). \quad (22)$$

The residual service time of the class-1 customer in service during slot k contributes in the first term, the sub-busy periods added to d_2 by the class-1 customers arriving during the residual service time contribute in the second term, the sub-busy periods added by the class-1 customers already in the queue at the beginning of slot k contribute in the third term, the sub-busy periods added by the class-1 and class-2 customers arriving during slot k , but that have to be served before the tagged class-2 customer contribute in the fourth term and finally the service time of and the sub-busy period initiated by the tagged class-2 customer itself (minus the sub-busy periods initiated by the class-1 customers arriving in the slot preceding the class-2 customer's departure) contribute in the last term.

4. $u_{1,k} > 0, u_{2,k} > 0$:

$$d_2 = (r_k - 1) + \sum_{i=1}^{r_k-1} \sum_{m=1}^{a_{1,k}+i} v_{1,m}^{(k+i)} + \sum_{m=1}^{u_{1,k}-1} \tilde{v}_{1,m} + \sum_{j=1}^2 \sum_{m=1}^{f_{j,k}} v_{j,m}^{(k)} + v_{2,k}^+ + \sum_{m=1}^{u_{2,k}-1} \tilde{v}_{2,m} + \left(\tilde{v}_2 - \sum_{m=1}^{a_{1,k}+d_2} v_{1,m}^{(k+d_2)} \right). \quad (23)$$

This is a combination of the two previous situations.

Clearly, the former equations relate the class-2 delay of a tagged customer to the (Markovian) state of the system at the beginning of this tagged customer's arrival slot. Due to the i.i.d. nature of the per-slot arrivals, the stochastic characteristics are the same as those at the beginning of a random slot. I.e., P also denotes the pgf of the state of the system at the beginning of a random customer's arrival slot. We now relate the pgf D_2 of the class-2 customer delay to the joint pgf P .

$D_2(z)$ is given - by conditioning on whether the class-1 and/or the class-2 systems are empty - by

$$\begin{aligned} D_2(z) = & \mathbb{E} [z^{d_2} \mathbf{1}_{u_{1,k}=u_{2,k}=0}] + \mathbb{E} [z^{d_2} \mathbf{1}_{u_{1,k}=0 \wedge u_{2,k}>0}] \\ & + \mathbb{E} [z^{d_2} \mathbf{1}_{u_{1,k}>0 \wedge u_{2,k}=0}] + \mathbb{E} [z^{d_2} \mathbf{1}_{u_{1,k}>0 \wedge u_{2,k}>0}]. \end{aligned} \quad (24)$$

Before calculating the four terms of expression (24) using the system equations (20)-(23), we first take a closer look at the pgf's of the sub-busy periods. It can easily be seen that the sub-busy periods initiated by the primary customers of class-1 (class-2 respectively) form a set of i.i.d. random variables and their common pgf is represented by $V_1(z)$ ($V_2(z)$ respectively).

The pgf $V_1(z)$ satisfies $V_1(z) = S_1(zA_1(V_1(z)))$. This can be understood as follows. The sub-busy period initiated by a class-1 customer consists of two parts, the service time of that customer itself, and the sub-busy periods initiated by the class-1 customers that arrive during its service time. This leads to the above relation for $V_1(z)$. The calculation of $V_2(z)$ is more complicated because of the PRI priority scheduling. We denote the conditional pgf of the sub-busy period of a class-2 customer with a service time of i slots by $V_{2,i}(z)$. We thus have

$$V_2(z) = \sum_{i=1}^{\infty} s_2(i) V_{2,i}(z). \quad (25)$$

We first calculate an expression for the $V_{2,i}(z)$. When a class-2 customer enters the server, two events can occur: the class-2 customer can either be completely served in the first attempt, or the service of the class-2 customer is interrupted by newly arriving class-1 customers. A class-2 service is not interrupted if no class-1 customers arrive during the i slots of the class-2 service time, with exception of the last slot (since the class-2 customer leaves the system at the end of that slot independent of whether class-1 customers arrive during that slot). The possible class-1 customers that enter the system during that last service slot add class-1 sub-busy periods to the class-2 sub-busy period. We thus have for the first case:

$$\mathbb{E}[z^{v_2} \mathbf{1}_{\text{no interruption}} | s_2 = i] = \frac{(zA_1(0))^i A_1(V_1(z))}{A_1(0)}, \quad (26)$$

with v_2 and s_2 the sub-busy period added by and the service time of the class-2 initiating customer respectively and with $A_1(0)$ the probability that no class-1 customers arrive in a random slot. The class-2 service is interrupted on the other hand when class-1 packets arrive during one of the service slots of the class-2 packet (excluding the last one). These class-1 packets all add a class-1 sub-busy period. The class-2 packet is put back in the queue and has to wait in the queue until all class-1 packets are served before having another service attempt. This again initiates a new class-2 sub-busy

period with pgf $V_{2,i}(z)$. This leads to the following expression

$$E[z^{v_2} \mathbf{1}_{\text{interruption}} | s_2 = i] = \frac{(A_1(V_1(z)) - A_1(0))((A_1(0)z)^i - A_1(0)z)V_{2,i}(z)}{A_1(0)(A_1(0)z - 1)}. \quad (27)$$

Since adding expressions (26) and (27) yields $V_{2,i}(z)$, we find the following expression:

$$V_{2,i}(z) = \frac{(1 - A_1(0)z)A_1(V_1(z))(A_1(0)z)^i}{(A_1(V_1(z)) - A_1(0))(A_1(0)z)^i - A_1(0)(zA_1(V_1(z)) - 1)}, \quad (28)$$

Substituting this expression in (25) yields

$$V_2(z) = \sum_{i=1}^{\infty} \frac{s_2(i)(1 - A_1(0)z)A_1(V_1(z))(A_1(0)z)^i}{(A_1(V_1(z)) - A_1(0))(A_1(0)z)^i - A_1(0)(zA_1(V_1(z)) - 1)}. \quad (29)$$

Finally taking the z -transform of equations (20)-(23), we find D_2 as a function of P and R_2 , which are defined and calculated in the previous section. Substituting the results obtained in the previous section, we finally find (after some extensive mathematical manipulations)

$$\begin{aligned} D_2(z) = & \frac{1 - \rho_{T,eff}}{\lambda_2} \frac{V_2(z)}{A_1(V_1(z))} \left\{ \frac{(zA_1(V_1(z)) - 1)(A(0, V_2(z)) - A_1(0))}{(V_2(z) - 1)(A_1(0)z - A(0, V_2(z)))} \right. \\ & + \frac{(zA_1(V_1(z)) - A(Y_1(V_2(z)), V_2(z))) \sum_{i=1}^{\infty} s_2(i)(V_{2,i}(z) - 1)Y_{2,i}(V_2(z))}{A(Y_1(V_2(z)), V_2(z))(V_2(z) - Y_2(V_2(z)))(V_2(z) - 1)} \\ & \left. \times \frac{(A_1(0)A(V_1(z), V_2(z)) - A_1(V_1(z))A(0, V_2(z)))(z - 1)}{(A_1(0)z - A(0, V_2(z)))(zA_1(V_1(z)) - A(V_1(z), V_2(z)))} \right\}, \end{aligned} \quad (30)$$

with $Y_2(z)$ given by (15) and with

$$Y_{2,i}(z) \triangleq \frac{(1 - A(0, z))A(Y_1(z), z)A(0, z)^i}{(A(Y_1(z), z) - A(0, z))A(0, z)^i - A(0, z)(A(Y_1(z), z) - 1)}. \quad (31)$$

Notice that

$$Y_2(z) = \sum_{i=1}^{\infty} s_2(i)Y_{2,i}(z). \quad (32)$$

Note that although the mathematical derivations required to obtain expression (30) are quite extensive, these are primarily standard probability theory techniques. Therefore we have omitted (a large part of) them. We have included the most important calculations, which are the calculations of $V_1(z)$ and $V_2(z)$.

5 Calculation of moments

The functions $Y_j(z)$ and $V_j(z)$, $j = 1, 2$, defined in the previous two sections, can be explicitly found only in case of some specific arrival and service processes. Their derivatives for $z = 1$, necessary to calculate the moments of the system content and the delay, on the contrary, can be calculated in closed-form. By taking the necessary derivatives of the marginal pgf's, moments of the total, of the class-1 and of the class-2 system content are found, as well as the moments of the class-1 and class-2 delays. We show the expressions of the mean values in this section. The mean class-1 and class-2 system contents are given by

$$E[u_1] = U'_1(1) = \frac{\rho_1}{2} + \frac{\mu_1 \text{Var}[a_1]}{2(1 - \rho_1)} + \frac{\lambda_1^2 \text{Var}[s_1]}{2(1 - \rho_1)}, \quad (33)$$

and

$$\begin{aligned} E[u_2] = U'_2(1) = & \frac{\rho_{2,eff}}{2} + \frac{\mu_1^2 \lambda_2 \text{Var}[a_1]}{2(1 - \rho_{T,eff})(1 - \rho_1)} + \frac{\mu_{2,eff} \text{Var}[a_2]}{2(1 - \rho_{T,eff})} + \frac{\mu_1 \text{Cov}[a_1, a_2]}{1 - \rho_{T,eff}} \\ & + \frac{\lambda_2(\lambda_1 \text{Var}[s_1] + \lambda_2 \text{Var}[s_2]_{eff})}{2(1 - \rho_{T,eff})(1 - \rho_1)} + \frac{\rho_1 \lambda_2 (\mu_{2,eff} - 1)}{2(1 - \rho_1)}, \end{aligned} \quad (34)$$

respectively. $\text{Var}[a_j]$ ($j = 1, 2$) and $\text{Var}[s_1]$ denote the variance of the number of per-slot class- j arrivals and the variance of the class-1 service times respectively, $\text{Cov}[a_1, a_2]$ denotes the covariance of the numbers of per-slot class-1 and class-2 arrivals and $\text{Var}[s_2]_{eff}$ is given by

$$\begin{aligned} \text{Var}[s_2]_{eff} \triangleq & \mu_{2,eff} \left[\frac{2(1 - \rho_1)A^{(2)}(0, 1)}{\lambda_2 A_1(0)(1 - A_1(0))} \left\{ 1 - \frac{S'_2(1/A_1(0))}{\mu_{2,eff}} \right\} \right. \\ & \left. + \frac{2(S_2(1/A_1(0))^2 - S_2(1/A_1(0)))}{(1/A_1(0) - 1)(S_2(1/A_1(0)) - 1)} + \mu_{2,eff} - 1 \right]. \end{aligned} \quad (35)$$

The mean total system content is found by taking the first derivative of (16) and substituting z by 1. It is given by $E[u_1] + E[u_2]$.

The mean class-1 and class-2 delay can be found by taking the first derivative of $D_1(z)$ and $D_2(z)$ respectively and evaluating in 1. We find $E[d_j] = E[u_j]/\lambda_j$, $j = 1, 2$ in accordance with the discretized version of Little's law (see [14]). Note that higher moments of system contents and delays can also be explicitly calculated by taking higher derivatives of the respective pgf's (see numerical examples).

6 Discussion of some results

6.1 Stability

We briefly touch upon stability issues of this type of priority queues. We notice that $P(0, 0, 0, 0) = 0$ for $\rho_{T,eff} = 1$ (expression (12)). As a result the system is instable for $\rho_{T,eff} \geq 1$. Note that

$$\mu_{2,eff} \geq \mu_2.$$

This is proved by writing expression (13) in terms of power series, yielding

$$\mu_{2,eff} = \sum_{n=1}^{\infty} s_2(n) \frac{(1/A_1(0))^n - 1}{1/A_1(0) - 1}, \quad (36)$$

and by noting that

$$\frac{(1/A_1(0))^n - 1}{1/A_1(0) - 1} \geq n, \quad (37)$$

for all $n \geq 1$, since $A_1(0) \leq 1$. As a result the effective load is always larger than the arrival load in a PRI priority queue. This is also intuitively clear, since repeats of class-2 service times increase the total (effective) load of a system.

6.2 The functions $Y_1(z)$ and $Y_2(z)$

We will prove in this subsection that $Y_1(z)$ and $Y_2(z)$ (defined in expressions (5) and (15) respectively) are pgf's. More precisely, we will prove that $Y_j(z)$ ($j = 1, 2$) is the pgf of the number of class-2 arrivals during a sub-busy period initiated by a class- j customer.

If at the beginning of a random slot a class-1 customer with service time s_1 enters the server, a new sub-busy period starts (with length denoted by v_1). Denoting the number of class-2 customers that arrive during this sub-busy period by y_1 and denoting the number of class- j arrivals during the m -th slot of this sub-busy period by $a_j^{(m)}$ ($j = 1, 2$), we get

$$y_1 = \sum_{m=1}^{v_1} a_2^{(m)} \quad (38)$$

$$= \sum_{m=1}^{s_1} \left(a_2^{(m)} + \sum_{l=1}^{a_1^{(m)}} y_{1,l}^{(m)} \right), \quad (39)$$

with $y_{1,l}^{(m)}$ the number of class-2 customers that arrive during the sub-busy period initiated by the l -th class-1 customer that arrives during the m -th slot of the service time s_1 . Naturally, all $y_{1,l}^{(m)}$ have the same distribution as y_1 (since the lengths of class-1 sub-busy periods are all identically distributed)

and therefore the pgf of y_1 is thus indeed given by (5) as immediately follows from (39) assuming that a stationary regime is reached.

If at the beginning of a slot a class-2 customer with service time s_2 enters the server, a new sub-busy period starts (with length denoted by v_2). Denoting the number of class-2 customers that arrive during this sub-busy period by y_2 and denoting the number of class- j arrivals during the m -th slot of this sub-busy period by $a_j^{(m)}$ ($j = 1, 2$), we get

$$y_2 = \sum_{m=1}^{v_2} a_2^{(m)}. \quad (40)$$

We condition on the lengths of the service times. Thus

$$\mathbb{E}[z^{y_2}] = \sum_{i=1}^{\infty} s_2(i) \mathbb{E}[z^{y_2} | s_2 = i], \quad (41)$$

with s_2 the service time of the initial class-2 customer. When this class-2 customer enters the server for the first time, two events may occur: the class-2 customer is either completely served in the first attempt, or the service of the class-2 customer is interrupted by newly arriving class-1 customers. Transforming all this into the z -domain and using a similar reasoning as done in obtaining expression (28), we find the following expression for $\mathbb{E}[z^{y_2} | s_2 = i]$

$$\mathbb{E}[z^{y_2} | s_2 = i] = \frac{(1 - A(0, z))A(Y_1(z), z)A(0, z)^i}{(A(Y_1(z), z) - A(0, z))A(0, z)^i - A(0, z)(A(Y_1(z), z) - 1)}. \quad (42)$$

Substituting this in expression (41), we indeed find expression (15).

Note that when the number of class-1 and class-2 arrivals in a slot are independent stochastic variables, it can be seen that v_j and the $a_2^{(m)}$ are also independent variables. From expressions (38) and (40), it then follows that

$$Y_j(z) = V_j(A_2(z)), \quad (43)$$

($j = 1, 2$) in this case. Note that this latter expression is not generally valid when the number of class-1 and class-2 arrivals in a slot are correlated, since v_j and $a_2^{(m)}$ both depend on $a_1^{(m)}$ in this case, and this for all m .

6.3 Special case: uncorrelated number of per-slot class-1 and class-2 arrivals

In this case $A(z_1, z_2) = A_1(z_1)A_2(z_2)$. The second term of the right-hand sides of expression (30) is equal to zero in this case and $D_2(z)$ equals

$$D_2(z) = \frac{1 - \rho_{T,eff}}{\lambda_2} \frac{V_2(z)(zA_1(V_1(z)) - 1)}{A_1(V_1(z))(V_2(z) - 1)} \frac{A_2(V_2(z)) - 1}{z - A_2(V_2(z))}. \quad (44)$$

Note that in the case of PR and PRD, the same expression is obtained in case of uncorrelated per-slot class-1 and class-2 arrivals (see [12] and [1] respectively). So, in this case, the distribution of class-2 sub-busy periods is still different for the three priority queues (PR, PRD and PRI), but the relationship between the delay of a class-2 customer and the sub-busy periods of class-2 customers is identical for the three cases. In [10], a similar relationship is found. In this paper, the low-priority characteristics of priority queues are analyzed by using queues with server interruptions. Indeed, from the point-of-view of class-2 customers the server is interrupted when class-1 customers are being served. In [10], these interruptions are incorporated in the service times and are called *effective* service times. These are analyzed separately for the PR, PRD and PRI queue respectively. Once the distributions of these effective service times are found however, the PR, PRD and PRI queues are analyzed in a uniform manner. The effective service times in that dissertation are thus closely related to the sub-busy periods initiated by class-2 customers defined in this paper.

Note that such a uniform approach does not seem to be possible in case the numbers of per-slot class-1 and class-2 arrivals are correlated. This is basically because in this case the interruption process (which is defined by the class-1 arrivals) depends on the arrivals to the system (which are the class-2 arrivals). When the number of per-slot class-1 and class-2 arrivals are not correlated, the interruption process and the arrivals to the queue are totally independent which considerably simplifies the analysis. This is also illustrated by the fact that expression (30) of $D_2(z)$ for the PRI priority queue becomes much more complicated in the latter case.

7 Numerical examples

7.1 An $N \times N$ output-queueing switch

In this section, we briefly discuss some numerical examples. We apply the obtained results to the special case of an output-queueing packet switch as depicted in Figure 2. We assume two types of traffic. Traffic of class-1 is delay-sensitive whereas traffic of class-2 is assumed to be delay-insensitive. We investigate the effect of a preemptive repeat identical priority scheduling discipline, as presented in the preceding sections.

We replace the generic term customers by the telecommunication term packets in this subsection. The packet arrivals on each inlet are assumed to be i.i.d., and generated by a Bernoulli process with arrival rate λ_T . An arriving packet is assumed to be of class- j with probability λ_j/λ_T ($j = 1, 2$) ($\lambda_1 + \lambda_2 = \lambda_T$). The incoming packets are then routed to the output queue corresponding to their destination, in an independent and uniform way. Therefore, the output queues behave identically and we can concentrate on the analysis of 1 output queue. As a result, the arrivals of both types of packets to an output queue are generated according to a two-dimensional binomial process. It is fully characterized by the following joint pgf

$$A(z_1, z_2) = \left(1 - \frac{\lambda_T}{N} + \frac{\lambda_1}{N}z_1 + \frac{\lambda_2}{N}z_2\right)^N. \quad (45)$$

In the remainder of this section, we assume $N = 16$. Furthermore, we denote the fraction of the class-1 arrival load in the total arrival load by α , i.e., $\alpha \triangleq \rho_1/\rho_T$.

[Figure 2 about here.]

Since the class-1 characteristics are the same as in a single-class FIFO system with only class-1 arrivals, we focus on the class-2 performance measures in this section. More specifically, we will compare the class-2 characteristics in the PRI priority queue with those in the preemptive resume (PR) and preemptive repeat different (PRD) priority queues. In the PR priority queue, a preempted class-2 service time is later resumed instead of completely repeated and the reader is referred to [12] for the performance measures of this queue. On the other hand in the PRD priority queue the preempted class-2 service time is resampled and repeated (see [1]). This will give us the opportunity to study the influence of the identical repeats of class-2 packets.

In Figure 3, we show the mean value and the variance of the class-2 system content as functions of the total load, with the class-1 and class-2 service times deterministically equal to 20 slots and α equal to 0.25, 0.5 and 0.75. In both figures, curves for the PR and the PRI priority queues are depicted. We furthermore show - on this figure and the other figures presented in this section - the vertical asymptotes of the curves in case of PRI, which equal the values of the *arrival* load for which the *effective* load equals 1. On the right of these asymptotes, the PRI priority queue is unstable. It is seen that the mean value and the variance of the class-2 system content can be considerably larger in case of the PRI case. This is because of the extra load that is added due to the repeats of class-2 packets' service times.

[Figure 3 about here.]

Figure 4a. depicts the mean class-2 delays for the PR and PRI priority queues as functions of the total load, with deterministic service times, $\mu_1 = 2$, $\mu_2 = 20$ and with α equal to 0.25, 0.5 and 0.75.

It is seen that the mean delay of the class-2 packets is significantly higher in the case of the PRI priority scheduling discipline, even for an average arrival load. Again this is due to the repeats of the class-2 packets. Furthermore, Figure 4b. shows the mean class-2 packet delay in case of the PRD and PRI priority scheduling disciplines for $\alpha = 0.25, 0.5$ and 0.75 respectively. The class-1 service times are deterministically equal to 2 slots and the class-2 service times are equal to 10 or 30 slots, each with probability $1/2$. It is seen that the PRD priority queue performs better in terms of mean class-2 delays than the PRI priority queue when the variance of the class-2 service times is larger than 0. This is because in the PRD queue a long class-2 service time (30 slots) which is interrupted is resampled to 10 slots with a probability equal to 0.5 in the next service attempt. Obviously, a service time of 10 slots can also be resampled to one with 30 slots when interrupted, but since the probability of an interruption occurring in a service time of 30 slots is larger than in one of 10 slots, the resampling of the class-2 service times decreases the (mean) class-2 packet delay.

[Figure 4 about here.]

Finally, Figure 5 shows the mean class-2 delays for the PR and PRI priority queues as functions of the class-1 service time μ_1 , when class-2 service times equal 20 slots, the total arrival load is 0.75 and for α equal to 0.25, 0.5 and 0.75. The service times of both classes are deterministic. In case of PR scheduling, the mean delay increases with increasing μ_1 . Longer class-1 service times increase the build-up periods for class-2 packets in the queue (i.e., longer periods when the server is busy with class-1 packets), thereby increasing the mean class-2 delay. In case of the PRI scheduling, this effect is also observed for high μ_1 . For low μ_1 on the other hand, it is seen that the mean class-2 delay increases dramatically with decreasing μ_1 . Smaller class-1 service times (while keeping the class-1 arrival load constant) imply *more* class-1 arrivals, thus increasing the probability of a class-2 packet's service getting preempted and having to be repeated. For very low μ_1 , many class-1 packets arrive, and as a result the class-2 packets' services will have to be repeated many times, thus adding a lot of extra load and dramatically increasing the (mean) class-2 delay. So, in the case of PRI there is an optimum for μ_1 for which the mean class-2 delay becomes minimal.

[Figure 5 about here.]

7.2 Impact of the correlation between both classes

In this subsection, we will briefly touch upon the impact of the correlation factor of the number of per-slot arrivals of both classes on the mean delay of class-2 customers. Therefore, we assume a simple arrival process as follows: the number of per-slot class- j arrivals is Bernoulli distributed and thus given by: $\text{Prob}[a_{j,k} = 0] = 1 - \lambda_j$ and $\text{Prob}[a_{j,k} = 1] = \lambda_j$, $j = 1, 2$. The joint pgf $A(z_1, z_2)$ of the numbers

of arrivals of both classes is given by

$$A(z_1, z_2) = 1 - \lambda_1 - \lambda_2 + q_{12} + (\lambda_1 - q_{12})z_1 + (\lambda_2 - q_{12})z_2 + q_{12}z_1z_2,$$

with q_{12} a parameter. The correlation factor $\rho_{a_1 a_2}$ of the number of per-slot class-1 and class-2 arrivals is given by

$$\rho_{a_1 a_2} = \frac{q_{12} - \lambda_1 \lambda_2}{\sqrt{\lambda_1 \lambda_2 (1 - \lambda_1)(1 - \lambda_2)}}.$$

By varying q_{12} , this correlation factor can be varied while keeping the arrival rates of both classes constant (λ_1 and λ_2 respectively). In Figure 6, we show the mean delay of class-2 packets versus the total load for three values of q_{12} (i.e., for three different correlation factors), for $\alpha = 0.25$ and deterministic service times of 20 slots for each class (i.e. $\mu_1 = \mu_2 = 20$). We vary q_{12} from its minimal value $q_{12} = 0$ over $q_{12} = \lambda_1 \lambda_2$ to its maximal value $q_{12} = \min(\lambda_1, \lambda_2)$. The corresponding correlation factors equal

$$\frac{-\lambda_1 \lambda_2}{\sqrt{\lambda_1 \lambda_2 (1 - \lambda_1)(1 - \lambda_2)}}, 0, \frac{\min(\lambda_1, \lambda_2)(1 - \max(\lambda_1, \lambda_2))}{\sqrt{\lambda_1 \lambda_2 (1 - \lambda_1)(1 - \lambda_2)}},$$

respectively. In other words, the number of per-slot arrivals of class-1 and class-2 are negatively correlated for the lower curve of Figure 6, uncorrelated for the middle curve and positively correlated for the upper curve. Since the two lower curves are basically on top of each other, the influence of a negative correlation factor is marginal in this example. We see however a significant influence of a positive correlation factor, especially when the total load is high. The reason for the higher mean class-2 delay when the number of arrivals of both classes are more correlated is the increasing probability that in the arrival slot of a particular class-2 packet a class-1 packet arrives as well (which could also interrupt service of an ongoing class-2 service time). The class-2 packet is certainly served after this class-1 packet, thus increasing its mean delay. Neglecting this correlation could thus lead to a considerable underestimation of the (mean) delay of class-2 packets. Note that in case of negative or zero correlation between the numbers of per-slot class-1 and class-2 arrivals, the mean class-2 delay equals 20 in case of $\rho_T \rightarrow 0$. Clearly, for $\rho_T \rightarrow 0$, the tagged packet arrives in an empty queue with probability 1. Further, the negative correlation implies that no class-1 packets arrive in the same slot as the tagged packet. The total delay thus equals the packet's service time of 20 slots. When the correlation between the numbers of per-slot class-1 and class-2 arrivals is positive however, the mean class-2 delay is larger than 20 slots for $\rho_T \rightarrow 0$. Although the buffer is still empty at the beginning of the arrival slot of the tagged class-2 packet with probability 1 in this case, the positive correlation implies that a class-1 packet arrives during the arrival slot of the tagged packet with

positive probability. Since this class-1 packet has to be served before the tagged class-2 packet, the mean delay of a class-2 packet is larger than its own service time.

[Figure 6 about here.]

8 Conclusions

In this paper, we analyzed a preemptive repeat identical priority queue using a probability generating functions approach. We defined supplementary variables - besides the system content of both priority classes - in order to construct a Markov-chain. We calculated the joint pgf of the steady-state version of the stochastic variables appearing in the Markov-chain. This pgf was the starting point of all calculations of the various performance measures. Note that an infinite sum is part of the solution in the class-2 pgf's (system content and delay). This infinite sum does however not give problems when evaluating the means and variances of these variables, but the calculation of higher moments could turn out to be of a more complex nature in practice - although it is feasible in theory. We further compared the performance measures of the PR and PRD priority queue - analyzed in earlier works - and of the PRI priority queue studied in this paper.

Acknowledgement

This work has been supported by the InterUniversity Attraction Poles Programme - Belgian Science Policy. The authors would further like to thank the anonymous referees and the associate editor for their constructive suggestions which led to the improvement of this paper.

References

- [1] J. Walraevens, B. Steyaert, and H. Bruneel. Analysis of a preemptive repeat priority buffer with resampling. In *Proceedings of the International Network Optimization Conference (INOC 2003)*, pages 581–586, Evry, October 2003.
- [2] H. White and L. Christie. Queuing with preemptive priorities or with breakdown. *Operations Research*, 6(1):79–95, 1958.
- [3] R. Miller. Priority queues. *Annals of Mathematical Statistics*, 31:86–103, 1960.
- [4] L. Kleinrock. *Queueing systems volume II: Computer applications*. John Wiley & Sons, 1976.
- [5] U. Sumita and O. Sheng. Analysis of query processing in distributed database systems with fully replicated files: a hierarchical approach. *Performance Evaluation*, 8(3):223–238, 1988.

- [6] C. Yoon and C. Un. Unslotted 1- and p_i -persistent CSMA-CD protocols for fiber optic bus networks. *IEEE Transactions on Communications*, 42(2-4):158–465, 1994.
- [7] A. Krinik, D. Marcus, R. Shiflett, and L. Chu. Transient probabilities of a single server priority queueing system. *Journal of Statistical Planning and Inference*, 101(1-2):185–190, 2002.
- [8] H. Castel and G. Hebuterne. *Performance analysis of an optical MAN ring for asynchronous variable length packets*, pages 214–220. LNCS 3124 (Proceedings of ICT 2004). Springer Verlag, 2004.
- [9] S. Mukherjee, D. Saha, and S. Tripathi. A preemptive protocol for voice-data integration in ring-based LAN: performance analysis and comparison. *Performance Evaluation*, 11(3):339–354, 1995.
- [10] D. Fiems, B. Steyaert, and H. Bruneel. Discrete-time queues with generally distributed service times and renewal-type server interruptions. *Performance Evaluation*, 55(3-4):277–298, 2004.
- [11] D. Cox. The analysis of non-Markovian stochastic processes by the inclusion of supplementary variables. *Proceedings of the Cambridge Philosophical Society*, 51:433–441, 1955.
- [12] J. Walraevens, B. Steyaert, and H. Bruneel. *Performance analysis of a GI-G-1 preemptive resume priority buffer*, pages 745–756. LNCS 2345 (Proceedings of the Networking 2002 Conference, Pisa). Springer Verlag, 2002.
- [13] H. Bruneel and B. Kim. *Discrete-time models for communication systems including ATM*. Kluwer Academic Publisher, Boston, 1993.
- [14] D. Fiems and H. Bruneel. A note on the discretization of Little’s result. *Operations Research Letters*, 30(1):17–18, 2002.

List of Figures

1	Sample of the time-axis	25
2	An output-queueing packet switch	26
3	Moments of the class-2 system content versus the total load for the PR (lower curves) and the PRI (thick upper curves) priority queues	27
4	Mean class-2 delays versus the total load	28
5	Mean class-2 packet delays versus the class-1 service time for both the PR (lower curves) and the PRI (thick upper curves) priority queues	29
6	Mean packet delay of class-2 versus the total load for different $\rho_{a_1 a_2}$	30

Figure 1: Sample of the time-axis

Figure 2: An output-queueing packet switch

Figure 3: Moments of the class-2 system content versus the total load for the PR (lower curves) and the PRI (thick upper curves) priority queues

a. Comparison of PR (lower curves) and PRI (thick upper curves) priority

b. Comparison of PRD (lower curves) and PRI (thick upper curves) priority

Figure 4: Mean class-2 delays versus the total load

Figure 5: Mean class-2 packet delays versus the class-1 service time for both the PR (lower curves) and the PRI (thick upper curves) priority queues

Figure 6: Mean packet delay of class-2 versus the total load for different $\rho_{a_1 a_2}$