# Performance Analysis of Priority Queueing Systems in Discrete Time

Joris Walraevens, Dieter Fiems, and Herwig Bruneel

SMACS Research Group
Department for Telecommunications and Information Processing (IR07)
Ghent University - UGent
Sint-Pietersnieuwstraat 41, B-9000 Gent, Belgium
{jw,df,hb}@telin.UGent.be
http://telin.UGent.be/smacs

**Abstract.** The integration of different types of traffic in packet-based networks spawns the need for traffic differentiation. In this tutorial paper, we present some analytical techniques to tackle discrete-time queueing systems with priority scheduling. We investigate both preemptive (resume and repeat) and non-preemptive priority scheduling disciplines. Two classes of traffic are considered, high-priority and low-priority traffic, which both generate variable-length packets. A probability generating functions approach leads to performance measures such as moments of system contents and packet delays of both classes.

## 1 Introduction

In recent years, there has been much research devoted to the incorporation of multimedia applications in packet-based networks. Different types of traffic need different Quality of Service (QoS) standards. For real-time applications, it is important that mean delay and delay-jitter are bounded, while for non real-time applications, the throughput and loss ratio are the restrictive quantities. In order to guarantee acceptable delay boundaries to delay-sensitive traffic (such as voice/video), several scheduling schemes – for switches, routers, ...– have been proposed and analyzed, each with their own specific algorithmic and computational complexity. The most drastic in this respect is the strict priority scheduling. With this scheduling, as long as delay-sensitive (or high-priority) packets are present in the queueing system, this type of traffic is served. Delay-insensitive packets can thus only be transmitted when no delay-sensitive traffic is present in the system. As already mentioned, this is the most drastic way to meet the QoS constraints of delay-sensitive traffic, but also the easiest to implement.

Within this tutorial paper, we focus on the analysis of queues with this priority scheduling discipline. We give an overview of the different types of priority scheduling disciplines and, for the most part, we show and explain some techniques to analytically analyze discrete-time queues with a priority scheduling discipline. Assume packets arriving to a buffer (located in a switch, router, multiplexer, ... ) being categorized in two distinct classes, the high-priority class and

D. Kouvatsos (Ed.): Next Generation Internet, LNCS 5233, pp. 203–232, 2011.

the low-priority class. In a queue with a priority scheduling discipline, the high-priority packets are transmitted ahead of the low-priority packets, i.e., when a server becomes available, a high-priority packet is always scheduled for transmission. Only, when there are no high-priority packets in the buffer at that time, a low-priority packet is selected for transmission. Priority scheduling disciplines come in two basic flavors, i.e., *non-preemptive* and *preemptive*. In the former, transmission of a packet is never interrupted once it is in service. So, if new high-priority packets arrive while a low-priority packet is served, they have to wait until the low-priority packet leaves the server. In a queue with a preemptive priority scheduling discipline on the other hand, those newly arriving high-priority packets interrupt transmission of the low-priority packet in service. Within the latter type of priority scheduling, two different strategies can be distinguished, depending on what happens when an interrupted low-priority packet re-enters the server. If the packet can resume its service where it was interrupted, i.e., when the part that was already transmitted before the interruption does not have to be retransmitted again, it is called a preemptive *resume* priority scheduling discipline. In a preemptive *repeat* priority scheduling on the other hand, the packet has to be retransmitted completely after the interruption.

In the literature, there have been a number of contributions with respect to priority scheduling. An overview of some basic priority queueing models in continuous time can be found in [1–3] and references therein. Discrete-time priority queues with deterministic service times equal to one slot have been studied in [4–16]. Khamisy and Sidi [4] analyze the system contents of the different classes, for a queue fed by a two-state Markov-modulated arrival process. Takine et al. [5] present the system content and delay for Markov-modulated high-priority arrivals and geometrically distributed low-priority arrivals. Laevens and Bruneel [6] analyze the system content and delay in the case of a multi-server queue. Choi et al. [7] and Walraevens et al. [12, 15] analyze a priority queue with train arrivals with resp. fixed, geometrically distributed and generally distributed train lengths. Walraevens et al. [8] study the system content and packet delay, in the special case of an output queueing switch with Bernoulli arrivals. Mehmet Ali and Song [9] examine a priority queue with on-off sources. Van Velthoven et al. [10] and Demoor et al. [13] tackle priority queues with finite (high-priority) buffer space. Kamoun [11] analyzes a priority queue with service interruptions. Finally, Walraevens et al. [14, 16] study the transient behavior and the output process resp. of a priority queue. All these papers have a single-slot service time in common; as a result no distinction has to be made between preemptive and non-preemptive priority scheduling.

Continuous-time *non-preemptive* priority queues have been considered in [17–26]. Discrete-time non-preemptive queues are the subject of [27–35]. Rubin and Tsai [27] study the mean waiting time, for a discrete-time queue fed by an i.i.d. arrival process. Hashida and Takahashi [28] analyze the packet delay by means of a delay-cycle analysis. Takine et al. [29] and Takine [30] study a discrete-time MAP/G/1 queue, using matrix-analytic techniques. Walraevens et al. examine the system content [31] and the packet delay [32] in a two-class

non-preemptive priority queue with i.i.d. number of per-slot arrivals and general service times using generating functions. The results presented in section 3 are largely based on the latter two papers. This analysis is furthermore extended to a general number of classes in [33]. Maertens and al. [34] investigate the tail behavior of the total content in a priority buffer. Finally, Demoor et al. [35] analyze a priority queue with finite capacity for high-priority customers.

Continuous-time *preemptive resume* priority queues have been analyzed in [36–47]. Discrete-time preemptive resume priority queues are the subject of [48–53]. Walraevens et al. [49] and Ndreca and Scoppola [52] analyze a two-class preemptive priority queue with geometric service times. Walraevens et al. [50] study a priority queue with general high-priority service times and geometric low-priority service times, while Lee [48] and Walraevens et al. [53] handle a two-class priority queue with generally distributed service times for both classes. Van Houdt and Blondia [51] analyze a three-class priority queue. Queues with a *preemptive repeat* priority scheduling discipline are studied less frequently than their non-preemptive and preemptive resume counterparts. Continuous-time models are studied in [54, 55]. Discrete-time preemptive repeat priority queues are the subject of [56–58]. Mukherjee et al. [56] study a preemptive repeat with resampling scheduling of voice traffic over data traffic in a ring-based LAN. Resampling, in this context, means that the length of a repeated service time is not necessarily equal to the length of the first (interrupted) service time. It is a new sample (with the same distribution). Walraevens et al. [57, 58] analyze resp. the preemptive repeat priority queue with resampling and without resampling. Queues with resampling and without resampling resp. are also known as preemptive repeat *different* and preemptive repeat *identical* priority queues.

Finally, Hong and Takagi [59] and Kim and Chae [60] analyze priority models which are combinations of non-preemptive and preemptive priority.

In this tutorial paper, we show some analytic techniques for analyzing the performance of queues with a preemptive or non-preemptive priority scheduling discipline. The analysis is largely based on the probability generating functions (pgfs) approach. We discuss two main methods to analyze priority queues. In the first method, a non-preemptive priority queue with two classes is analyzed. The joint pgf of the system contents of both classes and the pgfs of the delays of packets of both classes are calculated. Starting from these pgfs, it is shown how moments and approximate tail probabilities are calculated. In the second method, performance of low- and high-priority traffic is assessed separately in the case of a preemptive priority scheduling discipline. Here, a single-class model can be used to assess performance of the high-priority traffic as the preemptive priority discipline implies that high-priority traffic perceives the system as one without low-priority traffic. Low-priority traffic performance, on the other hand, is assessed with a single-class model with service interruptions. From the point of view of the low-priority class, the server is interrupted whenever high-priority packets are served and is available otherwise. We obtain a stochastic model for the perceived interruption process and present the analysis of the corresponding queueing model with interruptions.

So, queueing models with service interruptions are highly applicable for modeling the low-priority class in priority queues. To end this introduction, we will make a (short) literature overview of queueing models with service interruptions. Continuous-time queues with service interruptions are the subject of (a.o.) two recent papers [61, 62]. Research on discrete-time queues with service interruptions dates back to the 70's. Early papers include those by Hsu [63] and Heines [64]. Both authors treat the single-server system with Bernoulli server interruptions and a Poisson arrival process. The former considers queue contents at random slot boundaries whereas the latter considers queue contents at service completion times. A single-server system with an i.i.d. arrival and a correlated on/off server interruption process is treated in [65–67]. Woodside and Ho [66] and Yang and Mark [67] model the on- and off-periods as a series of i.i.d. shifted geometric random variables, whereas Bruneel [65] assumes that the series of consecutive on- and off-periods share a common general distribution. The only restriction in the latter contribution is that the common probability generating function of the on-periods must be rational. Alternatively, correlation in the interruption process is captured by means of a Markovian process by Lee [68]. In a more general setting – that is, no assumptions are made regarding the nature of the interruption process – relationships between queue contents at different time epochs are derived by Bruneel [69].

Georganas [70] and Bruneel [71] treat multi-server systems with i.i.d. customer arrival and server interruption processes. The latter extends the former in the sense that it does not assume that all outputs are either available or not. The delay analysis of the latter system is presented by Laevens and Bruneel [72]. A multi-server system with a correlated interruption process is considered by Bruneel [73]. The interruption process is modeled as an on/off process (geometrical on-periods). The number of available servers during the consecutive on-slots, are modeled by means of an i.i.d. series of non-negative random variables whereas no servers are available during off-periods.

Some contributions also allow a certain degree of correlation in the arrival process. Bruneel [74] assumes that both arrival and interruption processes are on/off processes with geometric on- and off-periods. A stochastic number of customers (an i.i.d. series) enters the system during arrival-on periods, whereas no customers arrive in the system during arrival-off periods. The interruption process is similar as the one analyzed by Yang and Mark [67] in the case of uncorrelated arrivals. This interruption process is also considered by Ali et al. [75] and by Kamoun [76]. The former authors assume that customer arrivals stem from a superposition of two-state Markovian on-off sources, while the latter author considers a so-called train-arrival process.

All the former discrete-time queueing models with service interruptions have a fixed customer service time of a single slot in common. A queueing system where customers have a fixed multiple-slot service-time, is considered by Inghelbrecht et al. [77]. The interruption process is again similar as the one treated by Yang and Mark [67]. The presence of multiple-slot service times and interruptions implies that a packet's transmission may be interrupted. The contribution considers

both the case that the packet transmission is continued after the interruption (CAI) and the case that transmission is repeated after the interruption (RAI). These modes correspond to preemptive resume and preemptive repeat priority scheduling, discussed above, respectively. Interruption models with generally distributed service times and a Bernoulli interruption process are considered by Fiems et al. [78, 79]. In [78], results for the CAI and RAI transmission modes are presented whereas some variants are considered in [79]. In particular we mention the repeat after interruption with resampling mode (in which the service time of an interrupted packet is resampled upon retransmission) and the partial repeat after interruption mode (in which only part of the packet has to be retransmitted after an interruption). The same authors consider CAI and RAI modes in the case of a Markovian interruption process [80] and in the case of a renewal-type interruption process [81]. The results presented in section 4 are based on the latter contribution.

The remainder of this paper is outlined as follows. In the next section we provide a more detailed description of the queueing model under consideration. In sections 3 and 4, we analyze the priority system in the case of a non-preemptive priority discipline and in the case of a preemptive priority discipline respectively. Some conclusions are drawn in section 5.

## 2 Mathematical Model

We consider a discrete-time single-server queueing system with infinite buffer space. Time is assumed to be slotted. There are two types of traffic arriving in the system, namely packets of class 1 and packets of class 2. We denote the number of arrivals of class $j$ during slot $k$ by $E_j^{(k)}$ $(j = 1, 2)$. Both types of packet arrivals are assumed to be i.i.d. from slot-to-slot and are characterized by the joint probability mass function $e(m, n) \triangleq \Pr[E_1^{(k)} = m, E_2^{(k)} = n]$, and joint probability generating function (pgf) $E(z_1, z_2) \triangleq \mathrm{E}[z_1^{E_1^{(k)}} z_2^{E_2^{(k)}}]$. Notice that the number of packet arrivals from different classes (within a slot) can be dependent. If necessary for the analysis though, we will loosen this condition and assume the number of arrivals of both classes in a slot mutually independent. Further, we define the marginal pgfs of the number of arrivals of class 1 and class 2 during a slot by $E_1(z) \triangleq \mathrm{E}[z^{E_1^{(k)}}] = E(z, 1)$ and $E_2(z) \triangleq \mathrm{E}[z^{E_2^{(k)}}] = E(1, z)$ respectively. We furthermore denote the arrival rate of class $j$ $(j = 1, 2)$ by $\overline{E}_j = E_j'(1)$. The variance of the number of per-slot arrivals of class-$j$ is given by $\sigma^2_{E_j} = E_j''(1) - (E_j'(1))^2 + E_j'(1)$.

The service times of the class-$j$ packets are assumed to be i.i.d. and are characterized by the probability mass function $s_j(m) \triangleq \Pr[\text{service of a class-}j\text{ packet takes } m \text{ slots}]$, $m \geq 1$, and pgf $S_j(z) = \sum_{m=1}^{\infty} s_j(m) z^m$, with $j = 1, 2$. We furthermore denote the mean and variance of the service time of a class-$j$ packet by $\overline{S}_j = S_j'(1)$ and $\sigma^2_{S_j} = S_j''(1) - (S_j'(1))^2 + S_j'(1)$. We define the arrival load offered by class-$j$ packets as $\rho_j \triangleq \overline{E}_j \overline{S}_j$ $(j = 1, 2)$. The total arrival load is then given by $\rho_T \triangleq \rho_1 + \rho_2$.

The system has one server that provides the transmission of packets. Class-1 packets are assumed to have priority over class-2 packets, and within one class the service discipline is First Come First Served (FCFS). So, if there are any class-1 packets in the queue when the server becomes empty, the one with the longest waiting time will be served next. If, on the other hand, no class-1 packets are present in the queue at that moment, the class-2 packet with the longest waiting time, if any, will be served next.

## 3 Non-preemptive Priority Queues

In this section, we analyze non-preemptive priority queues. We derive the joint pgf of the system contents of both priority classes and calculate the pgfs of the packet delays of both classes. From these pgfs, we show how to calculate moments and (approximate) tail probabilities of the respective stochastic variables.

### 3.1 System Content at Service Initiation Epochs

To be able to analyze the system content at random slot boundaries and the packet delays of both classes, we first analyze the system content at the beginning of so-called start slots. These are slots at the beginning of which a packet (if available) can enter the server. Note that every slot during which the system is empty, is also a start slot. We denote the system content of class-$j$ packets at the beginning of the $l$-th start slot by $U_{s,j}^{(l)}$ $(j = 1, 2)$. Clearly, the set $\{U_{s,1}^{(l)}, U_{s,2}^{(l)}\}$ forms a Markov chain, since the arrival process is i.i.d. and the buffer solely contains entire messages at the beginning of start slots. If $S^{(l)}$ indicates the service time of the packet that enters service at the beginning of start slot $l$ (which is - by definition - regular slot $k$) the following system equations can be established:

1. If $U_{s,1}^{(l)} = U_{s,2}^{(l)} = 0$:

$$U_{s,1}^{(l+1)} = E_1^{(k)}\,,\; U_{s,2}^{(l+1)} = E_2^{(k)}\,.$$

   The only packets present in the system at the beginning of start slot $l+1$ are the packets that arrived during the previous slot, i.e., start slot $l$.
2. If $U_{s,1}^{(l)} = 0$ and $U_{s,2}^{(l)} > 0$:

$$U_{s,1}^{(l+1)} = \sum_{i=0}^{S^{(l)}-1} E_1^{(k+i)}\,,\; U_{s,2}^{(l+1)} = U_{s,2}^{(l)} + \sum_{i=0}^{S^{(l)}-1} E_2^{(k+i)} - 1\,.$$

   The class-2 packet in service leaves the system just before start slot $l+1$. $S^{(l)}$ is characterized by probability mass function $s_2(m)$.
3. If $U_{s,1}^{(l)} > 0$:

$$U_{s,1}^{(l+1)} = U_{s,1}^{(l)} + \sum_{i=0}^{S^{(l)}-1} E_1^{(k+i)} - 1\,,\; U_{s,2}^{(l+1)} = U_{s,2}^{(l)} + \sum_{i=0}^{S^{(l)}-1} E_2^{(k+i)}\,.$$

$S^{(l)}$ is characterized by probability mass function $s_1(m)$.

We assume that the system is stable, implying that the equilibrium condition $\rho_T < 1$ is met. We define $U_s(z_1, z_2) \triangleq \lim_{l\to\infty} \mathrm{E}\left[z_1^{U_{s,1}^{(l)}} z_2^{U_{s,2}^{(l)}}\right]$. Using the system equations, we derive a functional equation for $U_s$:

$$[z_1 - S_1(E(z_1, z_2))]\, U_s(z_1, z_2) = \frac{z_1 S_2(E(z_1, z_2)) - z_2 S_1(E(z_1, z_2))}{z_2} U_s(0, z_2) + z_1 \frac{z_2 E(z_1, z_2) - S_2(E(z_1, z_2))}{z_2} U_s(0, 0)\,. \quad (1)$$

It now remains for us to determine the unknown function $U_s(0, z_2)$ and the unknown parameter $U_s(0, 0)$. This can be done in two steps. First, we notice that $U_s(z_1, z_2)$ must be analytic for all values of $z_1$ and $z_2$ such that $|z_1| < 1$ and $|z_2| < 1$. In particular, this should be true for $z_1 = Y(z_2)$, with $Y(z_2) \triangleq S_1(E(Y(z_2), z_2))$ and $|z_2| < 1$, since it follows from (an extension of) Rouché's theorem [82] that $z_1 = S_1(E(z_1, z_2))$ has exactly one solution $|Y(z_2)| < 1$ for all such $z_2$. Notice that $Y(1)$ equals 1. The above implies that if we insert $z_1 = Y(z_2)$ in equation (1), where $|z_2| < 1$, the left hand side of this equation vanishes. The same must then be true for the right hand side, yielding

$$U_s(0, z_2) = U_s(0, 0) \frac{z_2 E(Y(z_2), z_2) - S_2(E(Y(z_2), z_2))}{z_2 - S_2(E(Y(z_2), z_2))}\,. \quad (2)$$

The following expression for $U_s(z_1, z_2)$ can now be derived by combining equations (1) and (2):

$$U_s(z_1, z_2) = U_s(0, 0) \left[ \frac{z_1(z_2 E(z_1, z_2) - S_2(E(z_1, z_2)))}{(z_1 - S_1(E(z_1, z_2)))(z_2 - S_2(E(Y(z_2), z_2)))} + \frac{S_2(E(Y(z_2), z_2))(S_1(E(z_1, z_2)) - z_1 E(z_1, z_2))}{(z_1 - S_1(E(z_1, z_2)))(z_2 - S_2(E(Y(z_2), z_2)))} + \frac{E(Y(z_2), z_2)(z_1 S_2(E(z_1, z_2)) - z_2 S_1(E(z_1, z_2)))}{(z_1 - S_1(E(z_1, z_2)))(z_2 - S_2(E(Y(z_2), z_2)))} \right]. \quad (3)$$

Finally, in order to find an expression for $U_s(0, 0)$, we put $z_1 = z_2 = 1$ and use de l'Hôpital's rule in equation (3). Therefore, we need the first derivative of $Y(z)$ for $z = 1$ and this follows from its definition

$$Y'(1) = \overline{S}_1(\overline{E}_1 Y'(1) + \overline{E}_2) = \frac{\overline{E}_2 \overline{S}_1}{1 - \rho_1}\,. \quad (4)$$

We then obtain $U_s(0, 0)$:

$$U_s(0, 0) = \frac{1 - \rho_T}{1 - \rho_T + \overline{E}_1 + \overline{E}_2}\,. \quad (5)$$

Substituting the expression for $U_s(0, 0)$ in (3) gives a fully determined version of $U_s(z_1, z_2)$.

### 3.2 System Content at the Beginning of Arbitrary Slots

The system content of priority class $j$ at the beginning of a slot $k$ in steady state is denoted by $U_{r,j}^{(k)}$ $(j = 1, 2)$. Define the steady-state joint pgf $U_r(z_1, z_2) \triangleq \mathrm{E}[z_1^{U_{r,1}^{(k)}} z_2^{U_{r,2}^{(k)}}]$. In order to derive an expression for $U_r(z_1, z_2)$, we condition on the status of the server during slot $k$. There are three possibilities: the server can be idle, a low-priority or a high-priority packet can be in service during slot $k$. The server is idle during a slot if and only if the system was empty at the beginning of the slot. On the other hand, if the server is busy during slot $k$, a class-$j$ packet is being served with probability $\rho_j/\rho_T$ $(j = 1, 2)$. We relate the system content at the beginning of a random slot to the system content at the beginning of the preceding start slot. The elapsed service time of the packet in service (if any) during slot $k$ is given by $\tilde{S}$. The system content at the beginning of slot $k$ is a superposition of the system content at the beginning of the last preceding start slot and the arrivals during $\tilde{S}$, yielding

$$\begin{aligned} U_r(z_1, z_2) =& U_r(0,0) + (1 - U_r(0,0)) \left\{ \frac{\rho_2}{\rho_T} \frac{U_s(0, z_2) - U_s(0,0)}{U_s(0,1) - U_s(0,0)} \tilde{S}_2(E(z_1, z_2)) \right. \\ &\left. + \frac{\rho_1}{\rho_T} \frac{U_s(z_1, z_2) - U_s(0, z_2)}{1 - U_s(0,1)} \tilde{S}_1(E(z_1, z_2)) \right\} . \end{aligned} \tag{6}$$

Hereby is $\tilde{S}_j(z)$ $(j = 1, 2)$ defined as the pgf of the elapsed service time of the class-$j$ packet in service at the beginning of slot $k$. It is shown in e.g. [83] that

$$\tilde{S}_j(z) = \frac{S_j(z) - 1}{\overline{S}_j(z - 1)} , \tag{7}$$

for $j = 1, 2$. It now remains for us to determine the unknown parameter $U_r(0, 0)$. Keeping in mind that, if the server is idle during slot $k$, slot $k$ is a start slot, $U_r(0, 0)$ can easily be found as follows:

$$\begin{aligned} U_r(0,0) =& \Pr[U_{r,1}^{(k)} = U_{r,2}^{(k)} = 0] \\ =& \Pr[U_{s,1}^{(l)} = U_{s,2}^{(l)} = 0 | \text{ slot } k \text{ is a start slot}] \Pr[\text{slot } k \text{ is a start slot}] , \end{aligned}$$

with start slot $l$ the start slot directly preceding slot $k$. Conditioning on the possibilities of a slot being a start slot, we find

$$U_r(0,0) = U_s(0,0) \left[ U_r(0,0) + \frac{1 - U_r(0,0)}{\overline{S}_1} \frac{\rho_1}{\rho_T} + \frac{1 - U_r(0,0)}{\overline{S}_2} \frac{\rho_2}{\rho_T} \right] = 1 - \rho_T . \tag{8}$$

Using equations (3), (5), (7) and (8) in (6), we derive a fully determined version for $U_r(z_1, z_2)$:

$$U_r(z_1, z_2) = (1 - \rho_T) \left\{ \frac{S_1(E(z_1, z_2))(z_1 - 1)}{z_1 - S_1(E(z_1, z_2))} + \frac{E(Y(z_2), z_2) - 1}{E(z_1, z_2) - 1} \right.$$

$$\times \left[ \frac{z_1 S_2(E(z_1,z_2))(S_1(E(z_1,z_2))-1)}{(z_1 - S_1(E(z_1,z_2)))(z_2 - S_2(E(Y(z_2),z_2)))} \right.$$
$$+ \frac{z_1 z_2 (S_2(E(z_1,z_2)) - S_1(E(z_1,z_2)))}{(z_1 - S_1(E(z_1,z_2)))(z_2 - S_2(E(Y(z_2),z_2)))}$$
$$\left. + \frac{z_2 S_1(E(z_1,z_2))(1 - S_2(E(z_1,z_2)))}{(z_1 - S_1(E(z_1,z_2)))(z_2 - S_2(E(Y(z_2),z_2)))} \right] \Bigg\} . \tag{9}$$

From the two-dimensional pgf $U_r(z_1, z_2)$, we can easily derive expressions for the pgfs of the system contents of high- and low-priority packets at the beginning of an arbitrary slot - denoted by $U_{r,1}(z)$ and $U_{r,2}(z)$ respectively - yielding

$$U_{r,1}(z) \triangleq \lim_{k\to\infty} \mathrm{E}\left[z^{U_{r,1}^{(k)}}\right] = U_r(z,1)$$
$$= \frac{S_1(E_1(z))(z-1)}{z - S_1(E_1(z))} \left[1 - \rho_T + \overline{E}_2 \frac{S_2(E_1(z)) - 1}{E_1(z) - 1}\right] , \tag{10}$$
$$U_{r,2}(z) \triangleq \lim_{k\to\infty} \mathrm{E}\left[z^{U_{r,2}^{(k)}}\right] = U_r(1,z)$$
$$= (1-\rho_T) \frac{S_2(E_2(z))(z-1)}{z - S_2(E(Y(z),z))} \frac{E(Y(z),z) - 1}{E_2(z) - 1} . \tag{11}$$

### 3.3 Packet Delay

The packet delay is defined as the total time period a tagged packet spends in the system, i.e., the number of slots between the end of the packet's arrival slot and the end of its departure slot. We denote the steady-state delay of a tagged class-$j$ packet by $D_j$ and its pgf by $D_j(z)$ $(j = 1, 2)$. Before deriving expressions for $D_1(z)$ and $D_2(z)$, we first define some notions and stochastic variables we will frequently use in this subsection. We denote the arrival slot of the tagged packet by slot $k$. If slot $k$ is a start slot, it is assumed to be start slot $l$. If slot $k$ is not a start slot on the other hand, the last start slot preceding slot $k$ is assumed to be start slot $l$. We denote the number of class-$j$ packets that arrive during slot $k$, but which are served before the tagged packet by $\tilde{E}_j^{(k)}$ $(j = 1, 2)$. Since we only analyze the integer part of the delay, the precise time instant within the slot at which the tagged packet arrives, is not important. Only the order of service of all packets arriving in the same slot has to be specified. The class-1 packets will be serviced before the class-2 packets, and within a class the order of service is FCFS. We furthermore denote the service time of the tagged class-$j$ packet by $\hat{S}_j$ $(j = 1, 2)$. We finally denote the service time and the elapsed service time of the packet in service (if any) during the arrival slot of the tagged packet by $S$ and $\tilde{S}$ respectively. The latter random variable is the amount of service that the packet being served has already received at the beginning of the tagged packet's arrival slot. Assume $S$ and $\tilde{S}$ equal to 0 if no service is ongoing.

**Delay of High-Priority Packets.** We have that the delay of a tagged class-1 packet - arriving during slot $k$ - is given by

$$D_1 = (S - \tilde{S} - 1)^+ + \sum_{m=1}^{U_{s,1}^{(l)}-1} \check{S}_{1,m} + \sum_{i=1}^{\tilde{S}} \sum_{m=1}^{E_1^{(k-i)}} S_{1,m}^{(k-i)} + \sum_{m=1}^{\tilde{E}_1^{(k)}} S_{1,m}^{(k)} + \hat{S}_1 \,,$$

with $(x)^+ = \max(x, 0)$, the $S_{1,m}^{(k)}$'s the service times of the class-1 packets that arrived during slot $k$, but that are served before the tagged class-1 packet, the $S_{1,m}^{(k-i)}$'s $(0 \leq i \leq \tilde{S})$ the service times of the class-1 packets that arrived during slot $k - i$, and with $\check{S}_{1,m}$ the service times of the class-1 packets already in the queue at the beginning of the ongoing service (thus without the possible packet in service during slot $k$). We make the convention that a sum $\sum_{m=l}^{k}$ is 0 if $k < l$. Using this equation and conditioning on the type of the packet that is in service (no service, class 1 or class 2, we can derive an expression for $D_1(z)$:

$$\begin{aligned} D_1(z) = & \tilde{E}_1(S_1(z)) S_1(z) \left\{ 1 - \rho_T + \rho_2 \frac{S_2^*\left(\frac{E_1(S_1(z)))}{z}, z\right)}{z} \right. \\ & \left. + \rho_1 \frac{U_s(S_1(z), 1) - U_s(0, 1)}{(1 - U_s(0, 1)) S_1(z)} \frac{S_1^*\left(\frac{E_1(S_1(z)))}{z}, z\right)}{z} \right\}, \end{aligned} \tag{12}$$

with $\tilde{E}_1(z) \triangleq \mathrm{E}[z^{\tilde{E}_1^{(k)}}]$, $S_2^*(x, z) \triangleq E[x^{\tilde{S}} z^S | U_{s,1}^{(l)} = 0, U_{s,2}^{(l)} > 0]$ and $S_1^*(x, z) \triangleq E[x^{\tilde{S}} z^S | U_{s,1}^{(l)} > 0]$. The random variable $\tilde{E}_1^{(k)}$ can be shown to have the following pgf (see e.g. [83]):

$$\tilde{E}_1(z) = \frac{E_1(z) - 1}{\overline{E}_1(z - 1)} \,. \tag{13}$$

If a class-$j$ packet is in service during slot $k$, $S$ is characterized by the probability mass function $s_j(m)$ $(j = 1, 2)$. The conditional joint pgf of $\tilde{S}$ and $S$ when a class-$j$ packet is in service has the following form:

$$S_j^*(x, z) = \frac{S_j(xz) - S_j(z)}{\overline{S}_j(x - 1)} \,, \tag{14}$$

with $j = 1, 2$. We now obtain the following expression for $D_1(z)$ from equation (12) together with equations (3), (13) and (14):

$$D_1(z) = \frac{1}{\overline{E}_1} \frac{S_1(z)(z - 1)}{z - E_1(S_1(z))} \frac{E_1(S_1(z)) - 1}{S_1(z) - 1} \left(1 - \rho_T + \rho_2 \frac{S_2(z) - 1}{\overline{S}_2(z - 1)}\right) . \tag{15}$$

**Delay of Low-Priority Packets.** An expression for $D_2(z)$ is a bit more involved. We tag a class-2 packet that enters the buffer during slot $k$ (in steady state). Let us refer to the packets in the system at the end of slot $k$, but that have to be served before the tagged packet as the "primary packets". So, basically, the tagged class-2 packet can enter the server, when all primary packets and all class-1 packets that arrived after slot $k$ (i.e., while the tagged packet is waiting in the queue) are transmitted. In order to analyze the delay of the tagged class-2 packet, the number of class-1 packets and class-2 packets that are served between the arrival slot of the tagged class-2 packet and its departure slot is important, not their precise service order. Therefore, we consider an equivalent virtual system with an altered service discipline. We assume that, from slot $k$ on, the order of service for class-1 packets (those in the queue at the end of slot $k$ and newly arriving ones) is Last Come First Served instead of FCFS in the equivalent system (the transmission of class-2 packets remains FCFS). So, a primary packet can enter the server, when the system becomes free (for the first time) of class-1 packets that arrived during and after the service time of the primary packet that precedes it in the queue according to the new service discipline. Let $V_{1,m}^{(i)}$ denote the length of the time period during which the server is occupied by the $m$-th class-1 packet that arrives during slot $i$ and its class-1 "successors", i.e., the time period starting at the beginning of the service of that packet and terminating when the system becomes free (for the first time) of class-1 packets which arrived during and after its service time. Analogously, let $V_{2,m}^{(i)}$ denote the length of the time period during which the server is occupied by the $m$-th class-2 packet that arrives during slot $i$ and its class-1 "successors". The $V_{j,m}^{(i)}$'s $(j = 1, 2)$ are called sub-busy periods, initiated by the $m$-th class-$j$ packet that arrived during slot $i$. We have the following general expression for $D_2$:

$$D_2 = (S - \tilde{S} - 1)^+ + \sum_{i=1}^{S-\tilde{S}-1} \sum_{m=1}^{E_1^{(k+i)}} V_{1,m}^{(k+i)} + \sum_{j=1}^{2} \sum_{m=1}^{\tilde{E}_j^{(k)}} V_{j,m}^{(k)}$$
$$+ \sum_{j=1}^{2} \sum_{i=1}^{\tilde{S}} \sum_{m=1}^{E_j^{(k-i)}} V_{j,m}^{(k-i)} + \sum_{m=1}^{U_{s,1}^{(l)}-1} \check{V}_{1,m} + \sum_{m=1}^{U_2^{(l)}-1_{U_{s,1}^{(l)}=0}} \check{V}_{2,m} + \hat{S}_2 \,,$$

with the $\check{V}_{j,m}$'s the sub-busy periods, initiated by the $m$-th class-1 packet already in the queue at the beginning of start slot $l$ and $1_X$ the indicator function of $X$. It is clear that the length of the sub-busy periods initiated by class-1 packets are i.i.d. and thus have the same pgf $V_1(z)$. Also the length of the sub-busy periods initiated by class-2 packets are i.i.d., and their pgf is denoted by $V_2(z)$. Using the equation for $D_2$ and conditioning on which class is being served, we derive an expression for $D_2(z)$:

$$D_2(z) = \tilde{E}(V_1(z), V_2(z)) S_2(z) \left\{ 1 - \rho_T + \rho_2 \frac{U_s(0, V_2(z)) - U_s(0, 0)}{(U_s(0, 1) - U_s(0, 0)) V_2(z)} \right.$$

$$\times \frac{S_2^*\left(\frac{E(V_1(z),V_2(z))}{zE_1(V_1(z))}, zE_1(V_1(z))\right)}{zE_1(V_1(z))} + \frac{U_s(V_1(z),V_2(z)) - U_s(0,V_2(z))}{(1-U_s(0,1))V_1(z)}$$
$$\times \rho_1 \frac{S_1^*\left(\frac{E(V_1(z),V_2(z))}{zE_1(V_1(z))}, zE_1(V_1(z))\right)}{zE_1(V_1(z))}\Bigg\}, \tag{16}$$

with pgfs $\tilde{E}(z_1,z_2) \triangleq \mathrm{E}[z_1^{\tilde{E}_1^{(k)}} z_2^{\tilde{E}_2^{(k)}}]$, $S_2^*(x,z) \triangleq E[x^{\tilde{S}} z^S | U_{s,1}^{(l)} = 0, U_{s,2}^{(l)} > 0]$ and $S_1^*(x,z) \triangleq E[x^{\tilde{S}} z^S | U_{s,1}^{(l)} > 0]$. The random variables $\tilde{E}_1^{(k)}$ and $\tilde{E}_2^{(k)}$ have the following joint pgf (extension of a technique used in e.g. [83]):

$$\tilde{E}(z_1,z_2) = \frac{E(z_1,z_2) - E_1(z_1)}{\overline{E}_2(z_2-1)}\,. \tag{17}$$

The $S_j^*(x,z)$'s ($j=1,2$) are again given by equation (14). Finally, we have to find expressions for $V_1(z)$ and $V_2(z)$. These pgfs satisfy the following relations:

$$V_j(z) = S_j(zE_1(V_1(z)))\,, \tag{18}$$

with $j=1,2$. This can be understood as follows: when the $m$-th class-$j$ packet that arrived during slot $i$ enters service, $v_{j,m}^{(i)}$ consists of two parts: the service time of that packet itself, and the service times of the class-1 packets that arrive during its service time and of their class-1 successors. This leads to equation (18). Equation (16) together with equations (3), (14) and (17) leads to:

$$D_2(z) = \frac{1-\rho_T}{\overline{E}_2} \frac{S_2(z)(E(V_1(z),V_2(z)) - E_1(V_1(z)))}{zE_1(V_1(z)) - E(V_1(z),V_2(z))} \frac{1-zE_1(V_1(z))}{1-V_2(z)}\,, \tag{19}$$

with $V_j(z)$ ($j=1,2$) implicitly given by equation (18).

### 3.4 Calculation of Moments

The functions $Y(z)$, $V_1(z)$ and $V_2(z)$ can only be explicitly found in case of some simple arrival and service processes. Their derivatives for $z=1$, necessary to calculate the moments of the system content and the packet delay, on the contrary, can be calculated in closed form. For example, $Y'(1)$ is given by equation (4) and the first derivatives of $V_j(z)$ for $z=1$ are given by $V_j'(1) = \overline{S}_j/(1-\rho_1)$, $j=1,2$. Now, we can calculate the mean values of the system contents and packet delays of both classes by taking the first derivatives of the respective pgfs for $z=1$. We find

$$\overline{D}_1 = \frac{\overline{S}_1}{2} + \frac{(\sigma_{E_1}^2 \overline{S}_1 + \overline{E}_1^2 \sigma_{S_1}^2)}{2(1-\rho_1)\overline{E}_1} + \frac{\overline{E}_2(\sigma_{S_2}^2 + \overline{S}_2(\overline{S}_2-1))}{2(1-\rho_1)}\,, \tag{20}$$

$$\overline{D}_2 = \frac{\overline{S}_2}{2} + \frac{\sigma_{E_2}^2 \overline{S}_2}{2(1-\rho_T)\overline{E}_2} + \frac{\overline{E}_2 \sigma_{S_2}^2}{2(1-\rho_T)(1-\rho_1)} + \frac{\sigma_{E_1}^2 \overline{S}_1^2 + \overline{E}_1 \sigma_{S_1}^2}{2(1-\rho_T)(1-\rho_1)}$$

$$-\frac{\rho_1(\overline{S}_2-1)}{2(1-\rho_1)}+\frac{\overline{S}_1\sigma_{E_1E_2}}{(1-\rho_T)\overline{E}_2}\,. \tag{21}$$

$\sigma_{E_1E_2}$ is the covariance of $E_1$ and $E_2$. We only showed the expressions for the mean packet delay (as we will do throughout this paper), but the mean system content can be found in a similar way. Alternatively, one can always use the discretized version of Little's law [84] to calculate the mean system content from the mean packet delay. In a similar way, expressions for higher order moments can be calculated by taking the appropriate derivatives of the respective generating functions as well.

### 3.5 Tail Behavior

The tail distributions of system content and packet delay are often used to impose statistical bounds on the guaranteed QoS for both classes, and are therefore important performance measures. From the pgfs of the system contents and packet delays of class-1 and class-2 packets derived in subsections 3.2 and 3.3, approximations of the tail probabilities can be derived using complex contour integration and residue theory. In order to determine the asymptotic behavior of the tail distribution, the dominant singularity of the respective generating function is important. We concentrate on the packet delay when no long-tail behavior is encountered in numbers of per-slot arrivals or service times.

First, we concentrate on the class-1 packet delay. The dominant singularity $z_H$ of $D_1(z)$ is a zero of $z-E_1(S_1(z))$ (see equation (15)) and this singularity is a single pole. In the neighborhood of this pole, we can approximate $D_1(z)$ by

$$D_1(z)\approx\frac{K_1}{z_H-z}\,, \tag{22}$$

where $K_1$ is found by taking the limit $z\to z_H$ in (22). Using residue theory, we find, for large enough $n$,

$$\Pr[D_1=n]\approx\frac{1}{\overline{E}_1}\frac{S_1(z_H)(z_H-1)[(1-\rho_T)(z_H-1)+\overline{E}_2(S_2(z_H)-1)]}{z_H(S_1(z_H)-1)(E_1'(S_1(z_H))S_1'(z_H)-1)}z_H^{-n}\,. \tag{23}$$

The tail behavior of the class-2 delay is a bit more involved, since it is not a priori clear what the dominant singularity is of $D_2(z)$. This is due to the occurrence of the function $V_1(z)$ in (19), which is only implicitly defined. First we take a closer look at this function $V_1(z)$. The first derivative of $V_1(z)$ is given by

$$V_1'(z)=\frac{S_1'(zE_1(V_1(z)))E_1(V_1(z))}{1-zS_1'(zE_1(V_1(z)))E_1'(V_1(z))}\,. \tag{24}$$

Consequently, $V_1(z)$ has a singularity $z_B$, where the denominator of $V_1'(z)$ becomes 0. Thus $z_BS_1'(z_BE_1(V_1(z_B)))E_1'(V_1(z_B))=1$. Since $V_1(z)$ remains finite in the neighborhood of $z_B$, this singularity is not a simple pole. Application of the results from [85] $V_1(z)$ is, in the neighborhood of $z_B$, approximately given by

$$V_1(z) \approx V_1(z_B) - K_V\sqrt{z_B - z}\,, \tag{25}$$

with $K_V = \sqrt{\dfrac{2E_1(V_1(z_B))}{z_B[z_B^2(E_1'(V_1(z_B)))^3 S_1''(z_B E_1(V_1(z_B))) + E_1''(V_1(z_B))]}}$, which can be found by taking the limit $z \to z_B$ of (25) and using (18). From equation (25), it becomes obvious that $z_B$ is a square-root branch point of $V_1(z)$. $V_1(z)$ has thus two real solutions when $z < z_B$ (the solution we are interested in is the one where $V_1(z) < 1$, if $z < 1$), which coincide at $z_B$, and has no real solution when $z > z_B$. $z_B$ is also a branch point of $D_2(z)$. A second potential singularity $z_L$ of $D_2(z)$ on the real axis is given by the positive zero of the denominator which is a zero of $zE_1(V_1(z)) - E(V_1(z), V_2(z))$. The tail behavior of the class-2 packet delay is thus characterized by $z_L$ or $z_B$, depending on which is the dominant (i.e., smallest) singularity. It depends on the number of arrivals and service time distributions which singularity dominates. Three types of tail behavior may thus occur, namely when $z_L < z_B$, when $z_L = z_B$ and when $z_L$ does not exist. In those three cases, $D_2(z)$ can be approximated in the neighborhood of its dominant singularity by:

$$D_2(z) \approx \begin{cases} \dfrac{K_2^{(1)}}{z_L - z} & \text{if } z_L < z_B \\ \dfrac{K_2^{(2)}}{\sqrt{z_B - z}} & \text{if } z_L = z_B \\ D_2(z_B) - K_2^{(3)}\sqrt{z_B - z} & \text{if } z_L \text{ does not exist}\,, \end{cases}$$

where the constants $K_2^{(i)}$ $(i = 1, 2, 3)$ can be found by investigation of the behavior of $D_2(z)$ in the neighborhood of this dominant singularity. By using residue theory once again (see [86] for more details), the asymptotic behavior of $D_2$ is given by

$$\Pr[D_2 = n] \approx \begin{cases} \dfrac{K_2^{(1)}}{z_L} z_L^{-n} & \text{if } z_L < z_B \\ \dfrac{K_2^{(2)}}{\sqrt{z_B \pi}} n^{-1/2} z_B^{-n} & \text{if } z_L = z_B \\ \dfrac{K_2^{(3)}}{2}\sqrt{\dfrac{z_B}{\pi}}\, n^{-3/2} z_B^{-n} & \text{if } z_L \text{ does not exist}\,. \end{cases}$$

The first expression shows geometric tail behavior, while the second and third expressions show non-geometric tail behavior.

## 4 Preemptive Priority Queues

In this section, we consider the preemptive resume and preemptive repeat priority scheduling disciplines. For ease of analysis, we here additionally assume that there is no correlation between the number of class-1 and class-2 packets arriving

during the same slot, that is, $E(z_1, z_2) = E_1(z_1)E_2(z_2)$. This assumption allows us to study high-priority and low-priority performance separately by use of a single-class queueing system. The influence of class-1 traffic on class-2 traffic can be incorporated with interruptions.

The following subsection considers performance of class-1 traffic. The other sections then focus on performance of class-2 traffic. In subsection 4.2, we deduce an appropriate description of the interruption process perceived by class-2 traffic. The analysis of this queueing system with interruptions is then presented in subsections 4.3 to 4.5.

### 4.1 High-Priority Traffic

Preemptive priority implies that high-priority class-1 traffic is not influenced by low-priority class-2 traffic. That is, a class-1 packet receives service as if there is no low-priority traffic at all. Therefore, performance of the class-1 traffic can be assessed by means of a standard queueing model without priorities. In particular, the assumed nature of arrival and service processes yields that class-1 traffic can be assessed by the $Geo^X/G/1$ queueing model. This model is investigated by amongst others, Bruneel and Kim [83], by Takagi [87] and also by Hunter [88]. Alternatively, we may also retrieve our results from the results in the previous section by assuming that there is no class-2 traffic. That is, we assume: $E(z_1, z_2) = E_1(z_1)$. One easily verifies that the non-preemptive system then reduces to a single-class system. Substitution of the former expression in equations (10) and (15), then yields the pgf $U_{r,1}(z)$ of the class-1 system content at random slot boundaries,

$$U_{r,1}(z) = (1 - \rho_1)\frac{(z-1)S_1(E_1(z))}{z - S_1(E_1(z))}\,,$$

and the pgf $D_1(z)$ of the class-1 delay,

$$D_1(z) = \frac{1-\rho_1}{\overline{E}_1}\frac{E_1(S_1(z)) - 1}{z - E_1(S_1(z))}\frac{(z-1)S_1(z)}{1 - S_1(z)}\,,$$

respectively. The moment generating property of pgfs then yields e.g. following expression for mean class-1 packet delay $\overline{D}_1$,

$$\overline{D}_1 = \frac{\rho_1\,(1-\rho_1) + \sigma_{S_1}^2\,\overline{E}_1{}^2 + \overline{S_1}\sigma_{E_1}^2}{2(1-\rho_1)\overline{E}_1}\,. \tag{26}$$

### 4.2 Interruption Process

Consider low-priority class-2 traffic. Low-priority traffic is only served whenever there are no high-priority packets in the system. That is, a low-priority packet perceives the server as one that alternates between an available state and a blocked state. Slots during which no class-1 packets are served are called available slots or A-slots. Similarly, slots during which a class-1 receives service are called

blocked slots or B-slots. Contiguous periods of A-slots (B-slots) are referred to as A-periods (B-periods). One may verify that due to the nature of the class-1 arrival process, the consecutive A-periods as well as the consecutive B-periods constitute series of i.i.d. random variables.

If the high-priority queue is empty at the beginning of a slot, it remains empty during the next slot if there are no arrivals. That is, an A-period continues during the next slot with probability $\alpha = E_1(0)$. This implies that the consecutive A-periods share a common geometrical distribution. Let $A(z)$ denote the corresponding pgf, then we get, $A(z) = (1-\alpha)z/(1-\alpha z)$. Let the sub-busy period of a packet denote the number of slots between the first service slot of this packet and the beginning of the slot where for the first time the number of packets in the system is one less. Note that this definition of sub-busy period is essentially the same as in the preceding section. Clearly, a sub-busy period consists of the time the packet occupies the server (i.e., the packet length) and the sub-busy periods of all class-1 arrivals during this time. That is,

$$V_1 = S_1 + \sum_{i=1}^{S_1} \sum_{j=1}^{E_1^{(i)}} V_{ij} \, .$$

Here $V_1$ denotes a random class-1 packet's sub-busy period, $S_1$ denotes this packet's length, $E_1^{(i)}$ denotes the number of arrivals during the $i$-th service slot of this packet and $V_{ij}$ denotes the sub-busy period of the $j$-th arrival during the $i$-th service slot of the packet. Due to the nature of the arrival process, the sub-busy periods $V_{ij}$'s are independent random variables sharing the same pgf of the sub-busy period $V_1$. Some standard $z$-transform manipulations transform the former equation into

$$V_1(z) = S_1(zE_1(V_1(z))) \, . \tag{27}$$

The busy period of class-1 traffic – that is, the B-period for class-2 traffic – then equals the sum of the sub-busy periods of all arrivals during a slot, given that there is at least one arrival,

$$B(z) = \frac{E_1(V_1(z)) - E_1(0)}{1 - E_1(0)} \, .$$

The latter follows from the fact that a busy period starts with a non-empty batch of packets arriving in an empty system.

Although equation (27) only provides an implicit expression for $V_1(z)$, it allows to retrieve various moments by evaluation of the appropriate derivatives for $z = 1$ (as discussed in subsection 3.4). Therefore, one may retrieve moments of A- and B-periods as well. In particular, mean lengths of A- and B-periods are given by,

$$\overline{A} = \frac{1}{1-\alpha} \, , \; \overline{B} = \frac{\rho_1}{1-\rho_1} \frac{1}{1-\alpha} \, .$$

For preemptive resume priority scheduling, the transmission of the packet is resumed after interruptions. We will therefore further refer to this scheduling discipline as the *continue after interruption mode* (CAI). Similarly, as transmission is repeated in case of the preemptive repeat priority scheduling, we will further refer to this mode as the *repeat after interruption mode* (RAI). Note that the interruption process under investigation may find other applications as well. B-periods are an abstraction for some kind of server unavailability which does not necessary have to be linked with priority queueing models.

### 4.3 Effective Service Times

In a first step, we derive expressions for the pgfs of the effective service times of packets. First of all, for ease of explanation, we assume that a packet exists of a number of cells, where each cell needs 1 slot service time (so basically the number of cells in a packet is equal to the number of slots in that packet's service time). The effective service time of an arbitrary packet is defined as the time period elapsed (expressed in slots) between the beginning of the slot during which the first cell of a packet enters the service unit, and the end of the slot during which the last cell of the packet is served. In other words, the effective service time of a packet includes the slots during which the server is interrupted, and in case of RAI (preemptive repeat), the slots required for repeating service of certain cells. Due to the nature of the output process and the packet length distributions, the effective service times of consecutive packets also constitute a series of independent positive random variables, with distributions only depending on the state of the server – described by the availability of the server (A or B) together with the number of remaining B-slots in case the server is unavailable – during the slot preceding the start of the effective service time of the packet and on the operation mode under consideration. This implies that once we know the pgfs of the effective service times for the different operation modes, the evaluation of the system under consideration reduces to the evaluation of an equivalent system without server interruptions but with (state-dependent) service times given by the effective service times.

**Continue after Interruption.** Recall that the continue after interruption mode corresponds to the preemptive resume priority scheduling discipline. Let $t_{k,A}^{CAI}(n)$ denote the probability that the effective service time of a packet of length $k$ (in cells) equals $n$ slots given that the slot preceding the effective service time is an A-slot. The continue after interruption mode is a memoryless operation mode, in the sense that from a system point of view, once the first cell of a packet of length $k$ has been served, there is no difference between serving the remaining $k-1$ cells of this packet and servicing a new packet of length $k-1$. Therefore, conditioning on the state of the server during the first slot of the effective service time yields,

$$t_{k,A}^{CAI}(n) = \alpha t_{k-1,A}^{CAI}(n-1) + (1-\alpha)\sum_{j=1}^{\infty} b(j) t_{k-1,A}^{CAI}(n-j-1)\,, \qquad (28)$$

for $n \geq k$ and for $k > 1$ whereas for $n < k$ and $k > 1$ this probability equals 0. Let $T_{k,A}^{CAI}(z)$ denote the conditional pgf corresponding to $t_{k,A}^{CAI}(n)$, then, using standard $z$-transform manipulations, equation (28) easily transforms into,

$$T_{k,A}^{CAI}(z) = (\alpha z + (1-\alpha) z B(z))\, T_{k-1,A}^{CAI}(z)\,, \tag{29}$$

for $k > 1$. Clearly, equation (29) is also valid for $k = 1$ if one defines $T_{0,A}^{CAI}(z) = 1$, i.e., a zero-length packet requires no service time. Equation (29) then easily yields explicit expressions for the effective service time of a packet conditioned on the packet length and given that the server was available during the slot preceding the effective service time. Summation over all possible packet lengths with respect to their probabilities, then yields following expression for the pgf of the effective service time of a random packet given that the server was available during the slot preceding the effective service time,

$$T_{A}^{CAI}(z) = S\left(\alpha z + (1-\alpha) z B(z)\right)\,. \tag{30}$$

Finally, taking the appropriate derivatives of (30) yields expressions for the various moments of the corresponding random variable.

**Repeat after Interruption.** The memoryless property that was used in the previous section is not valid in case of RAI (preemptive repeat) as the server has to completely repeat transmission of the packet after an interruption. Consider an arbitrary slot that is part of a packet's effective service time. We define the remaining service time of a packet as the number of slots that are necessary to complete transmission of a packet in case there would be no interruptions. It is clear that in case of RAI (as opposed to CAI), the remaining service time for a particular packet is not a decreasing function in time, as after an interruption this value equals the packet length (in slots) again. Analogously, the remaining effective service time is defined as the number of slots it will effectively take to complete service (including interruptions and repetitions) at a certain point in time during a packet's effective service time.

Let $t_{k,l,A}^{RAI}(n)$ denote the probability that the remaining effective service time of a packet of length $k$ equals $n$ slots given that the remaining service time equals $l$ slots and that the slot preceding the remaining effective service time is an A-slot. Conditioning on the state of the server during the first slot of the remaining effective service time then yields,

$$t_{k,l,A}^{RAI}(n) = \alpha t_{k,l-1,A}^{RAI}(n-1) + (1-\alpha)\sum_{j=1}^{\infty} b(j)\, t_{k,k-1,A}^{RAI}(n-j-1)\,,$$

for $k, l > 1$ and for $n \geq l$, whereas the latter probability equals 0 for $k, l > 1$ and $n < l$. Let $T_{k,l,A}^{RAI}(z)$ denote the corresponding conditional pgf, then

$$T_{k,l,A}^{RAI}(z) = \alpha z T_{k,l-1,A}^{RAI}(z) + (1-\alpha) z B(z) T_{k,k-1,A}^{RAI}(z)\,, \tag{31}$$

for $k, l > 1$. It is easy to verify that the latter equation remains valid for $l = 1$ by defining $T^{RAI}_{k,0,A}(z) = 1$, i.e., if there are no more cells to send, the service ends in the current slot with probability 1. The former equation is a first order linear recursive equation and therefore easily solved. Substitution of $l = k - 1$ then determines the unknown function $T^{RAI}_{k,k-1,A}(z)$. In particular the pgf of the complete effective service time conditioned on the length of the packet and the state of the server during the slot preceding the effective service is then given by,

$$T^{RAI}_{k,k,A}(z) = \frac{(\alpha z)^{k-1}(1-\alpha z)(\alpha z + (1-\alpha)zB(z))}{1-\alpha z-(1-\alpha^{k-1}z^{k-1})(1-\alpha)zB(z)}\,, \tag{32}$$

for $k > 1$. One can easily verify that this expression remains valid for the trivial case of single slot service times ($k = 1$). Summation over all possible packet lengths (in slots) with respect to the packet length probabilities then yields the pgf of the effective service time given that the server is available during the preceding slot,

$$T^{RAI}_{A}(z) = \sum_{k=1}^{\infty} s_2(k) T^{RAI}_{k,k,A}(z)\,. \tag{33}$$

Note that this expression is in general not explicit due to the infinite sum. The moment-generating property of pgfs however, allows to determine the various moments of the effective service time explicitly by evaluation of the appropriate derivatives of the pgf for $z = 1$.

**Remarks.** Clearly, the server is not always available during the slot that precedes the effective service time. Therefore, let $T_{B,m}(z)$ denote the pgf of the effective service time of a random packet given that the server is blocked during the slot preceding the effective service and given that the server remains blocked for another $m$ slots after this slot (the server operates in one of the modes under consideration).

Consider now the decomposition of the effective service time of a packet in two components: the number of slots up to the first non-interrupted slot (i.e., the effective service of the first cell of the packet) and the remaining effective service time. Both components are independent random variables. It is clear that the first component (and its pgf) does not depend on the operation mode whereas the second component does not depend on the state of the server during the slot preceding the effective service as the last slot of the first component is by definition an A-slot. Let $X_A(z)$ and $X_{B,m}(z)$ denote the pgfs of the first component given the state during the slot preceding the effective service and let $Y_{(mode)}(z)$ denote the pgf of the second component only depending on the operation mode, then

$$T_A(z) = X_A(z)Y_{(mode)}(z)\,,\; T_{B,m}(z) = X_{B,m}(z)Y_{(mode)}(z)\,,$$

with

$$X_A(z) = \alpha z + (1-\alpha) z B(z)\,,\ X_{B,m}(z) = z^{m+1}\,.$$

The former expression follows from the fact that the first cell is either transmitted directly (with probability $\alpha$) or immediately after a interruption (with probability $(1-\alpha)$) in the case that the preceding slot is an A-slot. The latter expression follows from the fact that the first cell of a packet is transmitted immediately after the interruption in case the slot preceding the packet's effective service time is a B-slot followed by another $m$ B-slots. Elimination of $Y_{(mode)}(z)$ in the equations above then yields,

$$T_{B,m}(z) = \frac{z^m}{\alpha + (1-\alpha)B(z)} T_A(z)\,. \tag{34}$$

Equations (30) and (33) also imply that whereas for CAI the $n$-th moment of the effective service time depends on the moments of the underlying packet length distribution up to and including order $n$, this is not the case for RAI. For the latter operation modes, the $n$-th moment depends on the complete packet length distribution. In particular the first moments of the effective service time given that the slot preceding this effective service time is an A-slot, are given by,

$$\overline{T}_A^{CAI} = \frac{\overline{S}_2}{\sigma}\,, \tag{35}$$

$$\overline{T}_A^{RAI} = \frac{1}{\sigma}\frac{\alpha}{1-\alpha}\left(S_2\left(\frac{1}{\alpha}\right) - 1\right)\,, \tag{36}$$

for CAI and RAI respectively. Here $\sigma$ denotes the fraction of slots that the server is available, that is,

$$\sigma = \frac{\overline{A}}{\overline{A} + \overline{B}} = \frac{1}{1 + (1-\alpha)\overline{B}}\,. \tag{37}$$

Let us now assume the existence of all moments of the B-periods and assume that $\alpha$ is nonzero. For CAI, the existence of the $n$-th moment of the packet length in cells then implies the existence of the $n$-th moment of the effective service time, whereas this is not the case for the RAI operation mode. Let $R_{S_2}$ denote the radius of convergence of the pgf $S_2(z)$, then, one can verify that for RAI the $n$-th moment exists if $\alpha^{-n} < R_{S_2}$ and does not exist if $\alpha^{-n} > R_{S_2}$. For $\alpha^{-n} = R_{S_2}$, the existence depends on the behavior of $S_2(z)$ and its derivatives on their common radius of convergence. The additional condition for RAI also implies, that for finite radii of convergence and given $\alpha$, only a finite number of moments exist. In particular, one can easily verify that for $\alpha \in (R_{S_2}^{-1}, R_{S_2}^{-1/2})$ the respective effective service time distributions are heavy-tailed in case of RAI.

### 4.4 System Content

We now use the results of the preceding section to establish expressions for the pgf of the class-2 system content – i.e., the number of packets present in

the system – at packet departure times and at random slot boundaries. Since the effective service time of a packet includes interruptions and possible service repetitions of packets in case of RAI, results of the previous section allows a unified analysis for both operation modes.

**At Packet Departure Times.** Let $U_{d,2}^{(n)}$ denote the class-2 system content at the beginning of the slot following the departure slot of the $n$-th class-2 packet, i.e., at the departure time of the $n$-th class-2 packet. For positive $U_d^{(n)}$, service of the $(n+1)$-th class-2 packet can start immediately as this packet is already present in the system. Therefore, as the previous slot was an A-slot since there was a class-2 packet departure, it will take $T_A$ slots to the next departure, where $T_A$ denotes the random variable representing the effective service time of a class-2 packet given its effective service is preceded by an A-slot, and whose pgf is given by (30) or (33) depending on the operation mode under consideration. The system content $U_{d,2}^{(n+1)}$ is then given by

$$U_{d,2}^{(n+1)} = U_{d,2}^{(n)} - 1 + \sum_{j=1}^{T_A} E_2^{(j)}\,, \qquad \text{for } U_{d,2}^{(n)} > 0\,, \tag{38}$$

with $E_2^{(j)}$ the number of class-2 packets arriving in the system during the $j$-th slot of the effective service time of the $(n+1)$-th class-2 packet. If, on the other hand, the class-2 buffer is empty after the departure of the $n$-th class-2 packet, service of the next class-2 packet cannot start immediately. Let $w$ denote the first slot following the departure slot during which one or more packets arrive in the system, and let $E_{2,w}$ and $\Theta_w$ denote the number of class-2 arrivals and the state of the server during this slot respectively. As service of the $(n+1)$-th class-2 packet starts in the slot following slot $w$ and its effective service time is described by the random variable $T_{\Theta_w}$, $U_{d,2}^{(n+1)}$ is given by,

$$U_{d,2}^{(n+1)} = E_{2,w} - 1 + \sum_{j=1}^{T_{\Theta_w}} E_2^{(j)}\,, \qquad \text{for } U_{d,2}^{(n)} = 0\,, \tag{39}$$

with $E_2^{(j)}$ the number of packets arriving in the system during the $j$-th slot of the effective service time of the $(n+1)$-th packet. As the numbers of packets arriving during consecutive slots constitute a series of i.i.d. random variables, the common pgf of the $E_2^{(j)}$'s in (38) and (39) equals $E_2(z)$. Furthermore, as the only distinction regarding the number of arrivals between a random slot and the slot $w$ is that we are certain there arrives at least one packet in the system during slot $w$, the pgf of $E_{2,w}$ is given by

$$E_{2,w}(z) = \frac{E_2(z) - E_2(0)}{1 - E_2(0)}\,. \tag{40}$$

Now, let $q_{k,B,n}$ denote the probability that the $k$-th slot following an A-slot is a B-slot followed by another $n$ B-slots, and let $Q_B(x,z) = \sum_{k=1}^{\infty}\sum_{n=0}^{\infty} q_{k,B,n} x^k\, z^n$

denote the corresponding $z$-transform (note that this is not a pgf). Analogously, let $q_{k,A}$ denote the probability that the $k$-th slot following an A-slot is an A-slot and let $Q_A(x) = \sum_{k=1}^{\infty} q_{k,A} x^k$ denote the corresponding $z$-transform. Then, conditioning on the number of slots since the last preceding A-slot yields,

$$q_{k,A} = \alpha q_{k-1,A} + \sum_{j=1}^{k-1} b(j) q_{k-j-1,A} \,,$$

$$q_{k,B,n} = (1-\alpha) \sum_{j=n+1}^{n+k} b(j) q_{k+n-j,A} \,.$$

for $k \geq 1$ and for $n \geq 0$, whereas $q_{0,A} = 1$ and $q_{0,B,n} = 0$ for all $n \geq 0$. Standard $z$-transform manipulations then yield,

$$Q_A(x) = \frac{\alpha x + (1-\alpha) x B(x)}{1 - \alpha x - (1-\alpha) x B(x)} \,,$$

$$Q_B(x,z) = (1-\alpha) x (Q_A(x) + 1) \frac{B(x) - B(z)}{x - z} \,.$$

Due to the nature of the arrival process, slot $w$ (i.e., the first slot with at least one class-2 packet arrival after the departure of the $n$-th packet) is the $k$-th slot ($k \geq 1$) after the last departure slot with probability $g(k)$,

$$g(k) = E_2(0)^{k-1} (1 - E_2(0)) \,.$$

Furthermore, this slot is an A-slot (B-slot followed by $n$ B-slots) with probability $q_{k,A}$ ($q_{k,B,n}$) as the server is available during the last slot of the effective service time of the preceding packet. Summation over all possible values of $k$ with respect to the probabilities $g(k)$ yields the probabilities $\gamma_A$ and $\gamma_{B,n}$ that the server is available during slot $w$ or remains unavailable for another $n$ slots following slot $w$ respectively. Let $\Gamma_B(z)$ denote the $z$-transform of $\gamma_{B,n}$ then,

$$\begin{cases} \gamma_A = \dfrac{1 - E_2(0)}{E_2(0)} Q_A(E_2(0)) \,, \\ \Gamma_B(z) = \dfrac{1 - E_2(0)}{E_2(0)} Q_B(E_2(0), z) \,. \end{cases} \tag{41}$$

Now, assume the existence of a stationary distribution of the system contents, i.e., $U_{d,2}(z) = U_{d,2}^{(k+1)}(z) = U_{d,2}^{(k)}(z)$. From (34), (38) and (39), it then follows that the pgf of the class-2 system content at departure times is given by,

$$U_{d,2}(z) = \frac{U_{d,2}(0) T_A(E_2(z))}{z - T_A(E_2(z))} \left\{ \gamma_A E_{2,w}(z) + \frac{\Gamma_B(E(z)) E_{2,w}(z)}{\alpha + (1-\alpha) B(E_2(z))} - 1 \right\} , \tag{42}$$

where $T_A(z)$ is given by (30) or (33) depending on the operation mode. The unknown parameter $U_{d,2}(0)$ in (42) can then be determined by applying the normalization condition $U_{d,2}(1) = 1$, leading to

$$U_{d,2}(0) = \frac{\sigma}{\overline{E}_2 \gamma_A} (1 - E_2(0)) \left(1 - \overline{E}_2 \overline{T}_A\right) \,, \tag{43}$$

with $\overline{T}_A$ given by (35) or (36) for CAI and RAI operation modes respectively and with $\sigma$ given by expression (37). Substitution of equations (41) and (43) into (42) then yields the pgf of the steady-state class-2 system content at departure times,

$$U_{d,2}(z) = \frac{\sigma(1-\overline{T}_A\overline{E}_2)}{\overline{E}_2}\frac{E_2(z)}{Q_A(E_2(z))}\frac{T_A(E_2(z))}{T_A(E_2(z))-z}\,.$$

**Random Slot Boundaries.** Let $U_{r,2}(z)$ denote the pgf of the (stationary) system content at random slot boundaries and assume that there are no bulk arrivals (all arrivals occur at distinct epochs within slots). According to Bruneel [89], the pgf of the system content at random slot boundaries then relates to the pgf of the system content at arrival times $U_{a,2}(z)$ as,

$$U_{r,2}(z) = \frac{U_{a,2}(z)(z-1)\overline{E}_2}{E_2(z)-1}\,. \tag{44}$$

Again under the assumption that there are no bulk arrivals, the system content at arrival and departure times have the same distribution (see e.g. Kleinrock [90] or Takagi [3]), or equivalently, $U_{a,2}(z) = U_{d,2}(z)$, yielding,

$$U_{r,2}(z) = \frac{U_{d,2}(z)(z-1)\overline{E}_2}{E_2(z)-1}\,. \tag{45}$$

As both system content at random slot boundaries and system content at packet departure times do not depend on the exact arrival epochs within the consecutive slots, the former expression remains valid for systems with possible bulk arrivals.

**Remarks.** We assumed that the system under consideration reaches equilibrium. This is only the case if the buffer empties infinitely often during time, i.e., $U_r(0) > 0$, or equivalently, if the effective system load $\rho_{eff} = \overline{T}_A\,\overline{E_2}$ is less then the number of servers,

$$\rho_{eff} < 1\,. \tag{46}$$

Substitution of (35) or (36) then yields explicit conditions for the existence of the stationary distribution of the buffer contents for CAI and RAI respectively. Note that for CAI $\rho_{eff} = \rho_T$, as there are no retransmissions. For RAI, we get $\rho_{eff} \geq \rho_T$ as the effective load includes possible retransmissions.

The existence of a stationary distribution however does not imply the existence of moments of this distribution. Let us assume that all moments of the given distributions (number of arrivals in a slot of both classes, length of the packets of both classes) exist. Taking the first derivative of (42) or (45) reveals that the mean system content in both cases depends on both mean and variance of the effective service time, or in general, taking the appropriate derivatives reveals that the $n$-th moment of the stationary system content is a function

of the moments of the effective service times up to order $(n+1)$. This implies that where for CAI – due to our initial assumptions – the equilibrium condition guarantees a finite mean system content, this is not the case for RAI. In the latter case, the $n$-th moment of the system content distribution is finite as long as both the equilibrium condition and the condition for having a finite $(n+1)$th moment of the effective service time for RAI are satisfied (cfr section 4.3).

### 4.5 Unfinished Work and Packet Delay

Let $W_2^{(k)}$ denote the unfinished class-2 work at the beginning of slot $k$, i.e., the number of slots it would take to empty the class-2 buffer under the assumption that there are no new class-2 packet arrivals. Note that this definition implies that the unfinished work takes the interruptions and possible service repetitions into account. Consider now the unfinished work $W_2^{(k+1)}$ at the beginning of slot $(k+1)$. These random variables are related as,

$$W_2^{(k+1)} = (W_2^{(k)} - 1)^+ + \sum_{j=1}^{E_2^{(k)}} T^{(j)}\,, \tag{47}$$

where $E_2^{(k)}$ denotes the number of arriving class-2 packets in slot $k$ and $T^{(j)}$ denotes the effective service time of the $j$-th class-2 packet arriving in slot $k$. The unfinished work at the beginning of slot $(k+1)$ equals the unfinished work at slot $k$, diminished with the work done in slot $k$ (if there is any) and augmented with the additional work arriving in slot $k$. For each class-2 packet arriving in slot $k$, an additional number of slots, equal to its effective service time is necessary to completely empty the class-2 buffer.

If the class-2 buffer is not empty at the beginning of slot $k$, the effective service times of all packets entering the system in slot $k$ are preceded by an A-slot as the server was available during the last slot of the preceding class-2 packet's effective service time. This is also the case for all but the first packet entering the system during slot $k$ if the system is empty at the beginning of slot $k$. The state of the server preceding the first packet's effective service time is an A-slot with probability $\gamma_A$ or a B-slot followed by another $m$ B-slots with probability $\gamma_{B,m}$ as was shown in the previous section. Let $W_2^{(k)}(z)$ denote the pgf corresponding with $W_2^{(k)}$, then, from equation (47),

$$W_2^{(k+1)}(z) = W_2^{(k)}(0)\left(H(z) - \frac{E_2(T_A(z))}{z}\right) + W_2^{(k)}(z)\frac{E_2(T_A(z))}{z}\,,$$

with

$$H(z) = E_2(0) + (E_2(T_A(z)) - E_2(0))\left(\gamma_A + \sum_{m=0}^{\infty}\gamma_{B,m}\frac{T_{B,m}(z)}{T_A(z)}\right).$$

Now, assume that the system reaches equilibrium – i.e., the equilibrium condition (46) is satisfied – and let $W_2(z)$ denote the pgf of the stationary distribution, i.e.,

$W_2(z) = W_2^{(k)}(z) = W_2^{(k+1)}(z)$. As an empty buffer implies zero unfinished work and vice versa, i.e., $W(0) = U_r(0)$, the pgf of the unfinished work in equilibrium is given by,

$$W_2(z) = \frac{\sigma}{\gamma_A}(1-\rho_{eff})\frac{zH(z) - E_2(T_A(z))}{z - E_2(T_A(z))}\,. \tag{48}$$

Consider a particular (tagged) class-2 packet arrival. The packet delay $D_2$ is defined as the number of slots between the end of the arrival slot and the end of the departure slot of this packet. Let $W_{2,t}$ denote the unfinished work at the beginning of this packet's arrival slot and let $\tilde{E}_2$ denote the numbers of packets arriving in the same slot but before the tagged packet, then,

$$D_2 = (W_{2,t} - 1)^+ + \sum_{j=1}^{\tilde{E}_2+1} T^{(j)}\,, \tag{49}$$

with $T^{(j)}$ the effective service time of the $j$-th packet arriving in the system in the tagged packet's arrival slot.

The pgf of the unfinished work at the beginning of the tagged packet's arrival slot is given by (48) due to the i.i.d. nature of the arrival process. Furthermore – similar as equation (13) in the preceding section – the pgf $\tilde{E}_2(z)$ corresponding to $\tilde{E}_2$ is given by,

$$\tilde{E}_2(z) = \frac{E_2(z) - 1}{\overline{E}_2(z-1)}\,. \tag{50}$$

If the unfinished work $W_{2,t}$ is nonzero, all effective service times $T^{(j)}$ are preceded by an A-slot as service of these packets starts immediately after service of the preceding packet. This is also the case for all $T^{(j)}$ but $T^{(1)}$ when the queue is empty at the beginning of the tagged packet's arrival slot. For the latter, the preceding slot is again either an A-slot or a B-slot followed by another $m$ B-slots with probability $\gamma_A$ and $\gamma_{B,m}$ respectively. Let $D_2(z)$ denote the pgf corresponding to $D_2$, from (48) to (50) then follows,

$$D_2(z) = \frac{\sigma}{\overline{E}_2}(1-\rho_{eff})\frac{z}{Q_A(z)}\frac{E_2(T_A(z)) - 1}{E_2(T_A(z)) - z}\frac{T_A(z)}{T_A(z) - 1}\,. \tag{51}$$

The moment-generating property of pgfs then allows the calculation of explicit expressions for the moments of the class-2 packet delay. In particular mean class-2 packet delay is given by,

$$\begin{aligned}\overline{D}_2 =& \frac{\overline{T}_A\sigma^2_{E_2}}{2\overline{E}_2(1-\rho_{eff})} + \frac{\overline{E}_2\sigma^2_{T_A}}{2(1-\rho_{eff})} + (1-\alpha)\frac{\sigma\sigma_B^2}{2} + \frac{\overline{T}_A}{2} \\ & - (1-\alpha)(1-\alpha\sigma\overline{B})\frac{\overline{B}}{2}\,,\end{aligned} \tag{52}$$

Here $\sigma^2_{T_A}$ and $\sigma_B^2$ are the variances of $T_A$ and a $B$-period respectively.

## 5 Conclusions

In this paper, we analyzed the high- and low-priority system content and packet delay in a queueing system with a two-class priority scheduling discipline. Two basic types of priority scheduling are analyzed, namely, non-preemptive and preemptive priority scheduling. For each queueing system, a different analysis method was used. A generating-functions-approach was adopted in both, which led to closed-form expressions for some of the relevant performance measures. The results could be used to analyze performance of buffers in a packet-based networking context. Several extensions of the models and analyses are possible, such as a general number of priority classes, correlation in the arrival process, ....

**Acknowledgments.** The first two authors are Postdoctoral Fellows with the Research Foundation, Flanders (F.W.O.-Vlaanderen), Belgium.

## References

1. Miller, R.: Priority queues. Annals of Mathematical Statistics 31, 86–103 (1960)
2. Kleinrock, L.: Queueing systems. Computer applications, vol. II. John Wiley & Sons, New York (1976)
3. Takagi, H.: Queueing analysis: a foundation of performance evaluation, vacation and priority systems, part 1, vol. 1. North-Holland, Amsterdam (1991)
4. Khamisy, A., Sidi, M.: Discrete-time priority queues with two-state Markov Modulated arrivals. Stochastic Models 8(2), 337–357 (1992)
5. Takine, T., Sengupta, B., Hasegawa, T.: An analysis of a discrete-time queue for broadband ISDN with priorities among traffic classes. IEEE Transactions on Communications 42(2-4), 1837–1845 (1994)
6. Laevens, K., Bruneel, H.: Discrete-time multiserver queues with priorities. Performance Evaluation 33(4), 249–275 (1998)
7. 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)
8. 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)
9. Mehmet Ali, M., Song, X.: A performance analysis of a discrete-time priority queueing system with correlated arrivals. Performance Evaluation 57(3), 307–339 (2004)
10. Van Velthoven, J., Van Houdt, B., Blondia, C.: The impact of buffer finiteness on the loss rate in a priority queueing system. In: Horváth, A., Telek, M. (eds.) EPEW 2006. LNCS, vol. 4054, pp. 211–225. Springer, Heidelberg (2006)
11. Kamoun, F.: Performance analysis of a discrete-time queuing system with a correlated train arrival process. Performance Evaluation 63(4-5), 315–340 (2006)
12. Walraevens, J., Wittevrongel, S., Bruneel, H.: A discrete-time priority queue with train arrivals. Stochastic Models 23(3), 489–512 (2007)
13. Demoor, T., Walraevens, J., Fiems, D., Bruneel, H.: Mixed finite-/infinite-capacity priority queue with interclass correlation. In: Al-Begain, K., Heindl, A., Telek, M. (eds.) ASMTA 2008. LNCS, vol. 5055, pp. 61–74. Springer, Heidelberg (2008)

14. Walraevens, J., Fiems, D., Bruneel, H.: Time-dependent performance analysis of a discrete-time priority queue. Performance Evaluation 65(9), 641–652 (2008)
15. Walraevens, J., Wittevrongel, S., Bruneel, H.: Performance analysis of a priority queue with session-based arrivals and its application to E-commerce web servers. International Journal On Advances in Internet Technology 2(1), 46–57 (2009)
16. Walraevens, J., Fiems, D., Wittevrongel, S., Bruneel, H.: Calculation of output characteristics of a priority queue through a busy period analysis. European Journal of Operational Research 198(3), 891–898 (2009)
17. Stanford, D.: Interdeparture-time distributions in the non-preemptive priority $\Sigma M_i/G_i/1$ queue. Performance Evaluation 12(1), 43–60 (1991)
18. Sugahara, A., Takine, T., Takahashi, Y., Hasegawa, T.: Analysis of a nonpreemptive priority queue with SPP arrivals of high class. Performance Evaluation 21(3), 215–238 (1995)
19. Abate, J., Whitt, W.: Asymptotics for M/G/1 low-priority waiting-time tail probabilities. Queueing Systems 25(1-4), 173–233 (1997)
20. Takine, T.: The nonpreemptive priority MAP/G/1 queue. Operations Research 47(6), 917–927 (1999)
21. Isotupa, K., Stanford, D.: An infinite-phase quasi-birth-and-death model for the non-preemptive priority M/PH/1 queue. Stochastic Models 18(3), 387–424 (2002)
22. Drekic, S., Stafford, J.: Symbolic computation of moments in priority queues. INFORMS Journal on Computing 14(3), 261–277 (2002)
23. Bouallouche-Medjkoune, L., Aissani, D.: Quantitative estimates in an $M_2/G_2/1$ priority queue with non-preemptive priority: the method of strong stability. Stochastic Models 24, 626–646 (2008)
24. Iftikhar, M., Singh, T., Landfeldt, B., Caglar, M.: Multiclass G/M/1 queuing system with self-similar input and non-preemptive priority. Computer Communications 31, 1012–1027 (2008)
25. Al-Begain, K., Dudin, A., Kazimirsky, A., Yerima, S.: Investigation of the $M_2/G_2/1/\infty, N$ queue with restricted admission of priority customers and its application to HSDPA mobile systems. Computer Networks 53, 1186–1201 (2009)
26. Chen, Y., Chen, C.: Performance analysis of non-preemptive GE/G/1 priority queueing of LER system with bulk arrivals. Computers and Electrical Engineering 35, 764–789 (2009)
27. Rubin, I., Tsai, Z.: Message delay analysis of multiclass priority TDMA, FDMA, and discrete-time queueing systems. IEEE Transactions on Information Theory 35(3), 637–647 (1989)
28. Hashida, O., Takahashi, Y.: A discrete-time priority queue with switched batch Bernoulli process inputs and constant service time. In: Proceedings of ITC 13, Copenhagen, pp. 521–526 (1991)
29. Takine, T., Matsumoto, Y., Suda, T., Hasegawa, T.: Mean waiting times in nonpreemptive priority queues with Markovian arrival and i.i.d. service processes. Performance Evaluation 20, 131–149 (1994)
30. Takine, T.: A nonpreemptive priority MAP/G/1 queue with two classes of customers. Journal of Operations Research Society of Japan 39(2), 266–290 (1996)
31. Walraevens, J., Steyaert, B., Bruneel, H.: Performance analysis of the system contents in a discrete-time non-preemptive priority queue with general service times. Belgian Journal of Operations Research, Statistics and Computer Science (JORBEL) 40(1-2), 91–103 (2000)
32. 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)

33. Walraevens, J., Steyaert, B., Moeneclaey, M., Bruneel, H.: Delay analysis of a HOL priority queue. Telecommunication Systems 30(1-3), 81–98 (2005)
34. Maertens, T., Walraevens, J., Bruneel, H.: Priority queueing systems: from probability generating functions to tail probabilities. Queueing Systems 55(1), 27–39 (2007)
35. Demoor, T., Walraevens, J., Fiems, D., De Vuyst, S., Bruneel, H.: Analysis of a non-preemptive priority queue with finite high-priority capacity and general service times. In: Proceedings of the 4th International Conference on Queueing Theory and Applications (QTNA 2009), Singapore, ID12 (2009)
36. Miller, D.: Computation of steady-state probabilities for M/M/1 priority queues. Operations Research 29(5), 945–958 (1981)
37. Sandhu, D., Posner, M.: A priority M/G/1 queue with application to voice/data communication. European Journal of Operational Research 40(1), 99–108 (1989)
38. Takine, T., Hasegawa, T.: The workload in the MAP/G/1 queue with state-dependent services: its application to a queue with preemptive resume priority. Communications in Statistics - Stochastic Models 10(1), 183–204 (1994)
39. Takahashi, Y., Miyazawa, M.: Relationship between queue-length and waiting time distributions in a priority queue with batch arrivals. Journal of the Operations Research Society of Japan 37(1), 48–63 (1994)
40. Boxma, O., Cohen, J., Deng, Q.: Heavy-traffic analysis of the M/G/1 queue with priority classes. In: Proceedings of ITC 16, Edinburgh, pp. 1157–1167 (1999)
41. Sharma, V., Virtamo, J.: A finite buffer queue with priorities. Performance Evaluation 47(1), 1–22 (2002)
42. Takada, H., Miyazawa, M.: A Markov Modulated fluid queue with batch arrivals and preemptions. Stochastic Models 18(4), 529–652 (2002)
43. Liu, Y., Gong, W.: On fluid queueing systems with strict priority. IEEE Transactions on Automatic Control 48(12), 2079–2088 (2003)
44. Jin, X., Min, G.: Performance analysis of priority scheduling mechanisms under heterogeneous network traffic. Journal of Computer and System Sciences 73, 1207–1220 (2007)
45. Tarabia, A.: Two-class priority queueing system with restricted number of priority customers. AEÜ-International Journal of Electronics and Communications 61(8), 534–539 (2007)
46. Tzenova, E., Adan, I., Kulkarni, V.: Output analysis of multiclass fluid models with static priorities. Performance Evaluation 65(1), 71–81 (2008)
47. Horvath, A., Horvath, G., Telek, M.: A traffic based decomposition of two-class queueing networks with priority service. Computer Networks 53, 1235–1248 (2009)
48. Lee, Y.: Discrete-time $Geo^{x}$/G/1 queue with preemptive resume priority. Mathematical and Computer Modelling 34(3-4), 243–250 (2001)
49. Walraevens, J., Steyaert, B., Bruneel, H.: 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)
50. Walraevens, J., Steyaert, B., Bruneel, H.: A packet switch with a priority scheduling discipline: Performance analysis. Telecommunication Systems 28(1), 53–77 (2005)
51. Van Houdt, B., Blondia, C.: Analyzing priority queues with 3 classes using tree-like processes. Queueing Systems 54 (2), 99–109 (2006)
52. Ndreca, S., Scoppola, B.: Discrete-time GI/Geom/1 queueing system with priority. European Journal of Operational Research 189, 1403–1408 (2008)
53. Walraevens, J., Steyaert, B., Bruneel, H.: Analysis of a discrete-time preemptive resume priority buffer. European Journal of Operational Research 186(1), 182–201 (2008)

54. Sumita, U., Sheng, O.: Analysis of query processing in distributed database systems with fully replicated files: a hierarchical approach. Performance Evaluation 8(3), 223–238 (1988)
55. Yoon, C., Un, C.: Unslotted 1- and $p_i$-persistent CSMA-CD protocols for fiber optic bus networks. IEEE Transactions on Communications 42(2-4), 158–465 (1994)
56. Mukherjee, S., Saha, D., Tripathi, S.: A preemptive protocol for voice-data integration in ring-based LAN: performance analysis and comparison. Performance Evaluation 11(3), 339–354 (1995)
57. Walraevens, J., Steyaert, B., Bruneel, H.: A preemptive repeat priority queue with resampling: performance analysis. Annals of Operations Research 146(1), 189–202 (2006)
58. Walraevens, J., Fiems, D., Bruneel, H.: The discrete-time preemptive repeat identical queue. Queueing Systems 53(4), 231–243 (2006)
59. Hong, S., Takagi, H.: Analysis of transmission delay for a structured-priority packet-switching system. Computer Networks and ISDN Systems 29(6), 701–715 (1997)
60. Kim, K., Chae, K.: Discrete-time queues with discretionary priorities. European Journal of Operational Research 200(2), 473–485 (2010)
61. Fidler, M., Persaud, R.: M/G/1 priority scheduling with discrete pre-emption points: on the impacts of fragmentation on IP QoS. Computer Communications 27(12), 1183–1196 (2004)
62. Fiems, D., Maertens, T., Bruneel, H.: Queueing systems with different types of server interruptions. European Journal of Operational Research 188(3), 838–845 (2008)
63. Hsu, J.: Buffer behavior with Poisson arrival and geometric output processes. IEEE Transactions on Communications 22, 1940–1941 (1974)
64. Heines, T.: Buffer behavior in computer communication systems. IEEE Transactions on Communications 28, 573–576 (1979)
65. Bruneel, H.: A general treatment of discrete-time buffers with one randomly interrupted output line. European Journal of Operational Research 27(1), 67–81 (1986)
66. Woodside, C., Ho, E.: Engineering calculation of overflow probabilities in buffers with Markov-interrupted service. IEEE Transactions on Communications 35(12), 1272–1277 (1987)
67. Yang, O., Mark, J.: Performance analysis of integrated services on a single server system. Performance Evaluation 11, 79–92 (1990)
68. Lee, D.: Analysis of a single server queue with semi-Markovian service interruption. Queueing Systems 27(1–2), 153–178 (1997)
69. Bruneel, H.: Buffers with stochastic output interruptions. Electronics Letters 19(18), 735–737 (1983)
70. Georganas, N.: Buffer behavior with Poisson arrivals and bulk geometric output processes. IEEE Transactions on Communications 24(8), 938–940 (1976)
71. Bruneel, H.: A general model for the behaviour of infinite buffers with periodic service opportunities. European Journal of Operational Research 16, 98–106 (1984)
72. Laevens, K., Bruneel, H.: Delay analysis for discrete-time queueing systems with multiple randomly interrupted servers. European Journal of Operational Research 85, 161–177 (1995)
73. Bruneel, H.: A discrete-time queueing system with a stochastic number of servers subjected to random interruptions. Opsearch 22(4), 215–231 (1985)
74. Bruneel, H.: On buffers with stochastic input and output interruptions. International Journal of Electronics and Communications (AEU) 38(4), 265–271 (1984)

75. Ali, M., Zhang, X., Hayes, J.: A discrete-time queueing analysis of the wireless ATM multiplexing system. In: Lorenz, P. (ed.) ICN 2001. LNCS, vol. 2093, pp. 429–438. Springer, Heidelberg (2001)
76. Kamoun, F.: Performance evaluation of a queuing system with correlated packet-trains and server interruption. Telecommunication Systems 41(4), 267–277 (2009)
77. Inghelbrecht, V., Laevens, K., Bruneel, H., Steyaert, B.: Queueing of fixed-length messages in the presence of server interruptions. In: Proceedings Symposium on Performance Evaluation of Computer and Telecommunication Systems, SPECTS 2k, Vancouver, Canada (July 2000)
78. Fiems, D., Steyaert, B., Bruneel, H.: Performance evaluation of CAI and RAI transmission modes in a GI-G-1 queue. Computers and Operations Research 28(13), 1299–1313 (2001)
79. Fiems, D., Steyaert, B., Bruneel, H.: Randomly interrupted GI-G-1 queues, service strategies and stability issues. Annals of Operations Research 112, 171–183 (2002)
80. Fiems, D., Steyaert, B., Bruneel, H.: Analysis of a discrete-time GI-G-1 queueing model subjected to bursty interruptions. Computers and Operations Research 30(1), 139–153 (2002)
81. Fiems, D., Steyaert, B., Bruneel, H.: Discrete-time queues with generally distributed service times and renewal-type server interruptions. Performance Evaluation 55(3-4), 277–298 (2004)
82. Adan, I., Van Leeuwaarden, J., Winands, E.: On the application of Rouché's theorem in queueing theory. Operations Research Letters 34(3), 355–360 (2006)
83. Bruneel, H., Kim, B.: Discrete-time models for communication systems including ATM. Kluwer Academic Publisher, Boston (1993)
84. Fiems, D., Bruneel, H.: A note on the discretization of Little's result. Operations Research Letters 30(1), 17–18 (2002)
85. Drmota, M.: Systems of functional equations. Random Structures & Algorithms 10(1-2), 103–124 (1997)
86. Flajolet, P., Odlyzko, A.: Singularity analysis of generating functions. SIAM Journal on discrete mathematics 3(2), 216–240 (1990)
87. Takagi, H.: Queueing Analysis; A foundation of performance evaluation. Discrete-time systems, vol. 3. Elsevier Science Publishers, Amsterdam (1993)
88. Hunter, J.J.: Mathematical Techniques of Applied Probability. Operations Research and Industrial Engineering, vol. 2. Academic Press, New York (1983)
89. Bruneel, H.: Performance of discrete-time queuing systems. Computers and Operations Research 20, 303–320 (1993)
90. Kleinrock, L.: Queueing systems. Theory, vol. I. John Wiley & Sons, New York (1975)