Analysis of Priority Queues with Session-based Arrival Streams

Joris Walraevens, Sabine Wittevrongel and Herwig Bruneel
Department of Telecommunications and Information Processing (IR07)
Ghent University - UGent
Sint-Pietersnieuwstraat 41, B-9000 Gent, Belgium.
E-mail: {jw,sw,hb}@telin.UGent.be

Abstract

In this paper, we analyze a discrete-time priority queue with session-based arrivals. We consider a user population, where each user can start and end sessions. Sessions belong to one of two classes and generate a variable number of fixed-length packets which arrive to the queue at the rate of one packet per slot. The lengths of the sessions are generally distributed. Packets of the first class have transmission priority over the packets of the other class. The model is motivated by a web server handling delay-sensitive and delay-insensitive content. By using probability generating functions, some performance measures of the queue such as the moments of the packet delays of both classes are calculated. The impact of the priority scheduling discipline and of the session nature of the arrival process is shown by some numerical examples.

1 Introduction

Head-Of-the-Line (HOL) priority scheduling is one of the main scheduling types in network buffers to diversify the delay of traffic streams with different delay requirements. With this scheduling discipline, as long as delay-sensitive high-priority packets (packets of voice and video streams, gaming, …) are present in the buffer, they are transmitted. Best-effort low-priority packets can thus only be transmitted when no high-priority traffic is present. Another reason why one would like to diversify the delay characteristics of different applications is because one application may provide revenues for the provider while another does not (or to a lesser extent). It is then natural to give priority to the packets of the first application.

In the related literature, there have been a large number of contributions with respect to the analysis of HOL priority queues. In particular, discrete-time HOL priority queues with deterministic service times equal to one slot have been studied in [12, 14, 16, 19]. The steady-state buffer content and delay in the case of a multiserver queue with general independent arrivals are studied in [12]. Mehmet Ali and Song [14] analyze the buffer content in a multiplexer with two-state on-off sources. The steady-state buffer content and the delay for Markov modulated high-priority interarrival times and geometrically distributed low-priority interarrival times are analyzed in [16]. Walraevens et al. [19] study the steady-state buffer content and packet delay, in the special case of an output queueing switch with Bernoulli arrivals.

In this paper, we consider an arrival process induced by a two-layered structure. *Sessions* are started and terminated by users on the higher layer. These sessions inject trains of *packets* in the network. Since we perform a discrete-time analysis, we assume time is divided into slots of equal length and we assume that packets of a session arrive to the queue at the rate of one packet per slot. Note that this two-layered structure introduces *time correlation* in the packet arrival process. Indeed, since the packets in a session arrive in consecutive slots, the number of packet arrivals in one slot depends on the number of arrivals in previous slots. Session-based arrival processes are an adequate choice to model, e.g., the common segmentation of data files into packets before their transmission through a telecommunication network [8, 10].

In particular, the suggested arrival process is an ideal candidate to model the output buffer of a web server [9]. A web server is a computer system that accepts requests from users for a certain web page or embedded file and that responds by sending the requested file to the user. Traffic generated by a web server towards its output buffer can be described by a session-based arrival process. Moreover, if there is content on the web pages that is delay-sensitive, for instance multimedia content, priority can be given to the transmission of files containing this content over other downloads [24]. In the case of an e-commerce web server, it makes also sense to prioritize the downloads on a (potential) revenue base [17], that is, to give priority to the transmission of packets of content that is likely to provide (large) revenues.

First-In-First-Out (FIFO) queues with session-based arrivals are analyzed in [1, 2, 5, 6, 21, 22]. Somewhat related on/off-type arrival models are considered in [7, 11, 23], also for the FIFO case. Further in [4], sessions consisting of a fixed number of packets are considered in case of an uncorrelated packet arrival process. In view of the importance of priority scheduling in multimedia networks, HOL priority queues with session-based arrivals and either deterministic or geometric session lengths are studied in [3] and [20] respectively.

In the current paper, we further extend the previous analyses to a discrete-time priority queue with session-based arrivals and *generally* distributed session lengths. The distributions of the session lengths may be class-dependent, which reflects that different priority classes represent different applications. We analyze the buffer contents (i.e., the number of packets in the buffer) as well as the packet delays (i.e., the number of slots a packet stays in the buffer) of both the high-priority and low-priority class using *probability generating functions* (pgfs). In contrast with the specific session-length distributions studied in the past (see [3, 20]), an infinite-dimensional state vector has to be defined when dealing with generally distributed session lengths. This combined with the priority scheduling makes the analysis of the low-priority buffer content and packet delay far from straightforward. Nevertheless, closed-form formulas for the means of these stochastic variables (and in most cases also for higher moments) can be found by means of the analysis technique developed in this paper.

The remainder of the paper is structured as follows. In the next section, we present the mathematical model. In section 3, we construct a functional equation. This functional equation is the starting point of the analysis of the steady-state buffer content and packet delay, described in sections 4 and 5 respectively. Some numerical examples are treated in section 6, while we conclude this paper in section 7.

2 Queueing model

We consider a discrete-time single-server system with infinite buffer space. Time is assumed to be slotted. There are two types of sessions, namely sessions of class 1 and sessions of class 2. The numbers of newly generated class-j sessions during consecutive slots are i.i.d. (independent and identically distributed). The numbers of newly generated class-1 and class-2 sessions during slot k are denoted by $b_{1,k}$ and $b_{2,k}$ respectively. Their joint pgf is defined as $B(z_1, z_2) \triangleq \mathrm{E}\left[z_1^{b_{1,k}} z_2^{b_{2,k}}\right]$. Note that the numbers of sessions of both classes generated during a slot may be correlated. The corresponding marginal pgfs are denoted by $B_j(z)$ $(j = 1, 2)$ and are given by $B(z, 1)$ and $B(1, z)$ respectively.

Each class-j session lasts a random number of slots which is assumed generally distributed with pgf $L_j(z)$ and probability mass function (pmf) $l_j(i)$, $j = 1, 2$, $i \geq 1$. The packets of a session arrive back-to-back at the rate of one packet per slot. For further use, we define $p_j(n)$ as the probability that a class-j session that is going on for n slots continues at least one more slot, i.e.,

$$p_j(n) \triangleq \frac{1 - \sum_{i=1}^{n} l_j(i)}{1 - \sum_{i=1}^{n-1} l_j(i)}. \tag{1}$$

The total numbers of class-1 and class-2 packets arriving during slot k are denoted by $a_{1,k}$ and $a_{2,k}$ respectively and their joint pgf is defined as $A_k(z_1, z_2) \triangleq \mathrm{E}\left[z_1^{a_{1,k}} z_2^{a_{2,k}}\right]$. The transmission times of the packets equal one slot and per slot one packet is transmitted (if there is any).

Packets of class 1 have HOL priority over packets of class 2. This means that as long as there are class-1 packets in the buffer, they are transmitted. A class-2 packet can only be transmitted when there are no class-1 packets present.

On average $B_j'(1)$ class-j sessions are started in a random slot, each generating on average $L_j'(1)$ packets. Therefore the load generated by class-j packets equals

$$\rho_j = B_j'(1) L_j'(1), \tag{2}$$

$j = 1, 2$. We assume a stable system, i.e., the total load is smaller than 1:

$$\rho_T \triangleq \rho_1 + \rho_2 = B_1'(1)L_1'(1) + B_2'(1)L_2'(1) < 1. \tag{3}$$

3 Functional equation

In this section, we first construct a Markov chain description of the system. The arrival process is summarized by the random variables $e_{j,n,k}$, representing the number of class-j sessions that deliver their n-th packet during slot k. Indeed, the following relationships clearly hold:

$$\begin{aligned} &e_{j,1,k} = b_{j,k}; \\ &e_{j,n+1,k} = \textstyle\sum_{i=1}^{e_{j,n,k-1}} c_{j,n,k-1}^{(i)}, \qquad n \geq 1, \end{aligned} \tag{4}$$

$j = 1, 2$. For a given n, the $c_{1,n,k-1}^{(i)}$'s are i.i.d. random variables with values 0 or 1. The same holds for the $c_{2,n,k-1}^{(i)}$'s. The random variable $c_{j,n,k-1}^{(i)}$ equals 1 iff the i-th active session of class j that has sent the n-th packet during slot $k-1$ continues to send a packet in the next slot. The variable $a_{j,k}$, the total number of class-j packets arriving during slot k, can be expressed as

$$a_{j,k} = \sum_{n=1}^{\infty} e_{j,n,k}, \qquad j = 1, 2. \tag{5}$$

We further denote the buffer content of class-1 packets and class-2 packets at the beginning of slot k by $u_{1,k}$ and $u_{2,k}$ respectively. The following system equations then directly follow from the HOL priority scheduling of class-1 packets over class-2 packets:

$$\begin{aligned} u_{1,k+1} &= [u_{1,k}-1]^+ + a_{1,k}; \\ u_{2,k+1} &= [u_{2,k}-\mathbf{1}_{\mathbf{u_{1,k}=0}}]^+ + a_{2,k}, \end{aligned} \tag{6}$$

where $[.]^+$ denotes the maximum of the argument and 0 and with $\mathbf{1_X}$ the indicator function of X.

A Markovian state description of the system is given by $(e_{1,1,k-1}, e_{1,2,k-1}, \ldots, u_{1,k}, e_{2,1,k-1}, e_{2,2,k-1}, \ldots, u_{2,k})$ and equations (4)-(6) fully describe the behavior of the system. We introduce the joint pgf of the state vector:

$$P_k(x_{1,1}, x_{1,2}, \ldots, z_1, x_{2,1}, x_{2,2}, \ldots, z_2) \triangleq \mathrm{E}\left[\prod_{j=1}^{2}\left(\prod_{n=1}^{\infty} x_{j,n}^{e_{j,n,k-1}}\right) z_j^{u_{j,k}}\right].$$

It follows that

$$\begin{aligned} &P_{k+1}(x_{1,1}, x_{1,2}, \ldots, z_1, x_{2,1}, x_{2,2}, \ldots, z_2) = \\ &\mathrm{E}\left[\left(\prod_{j=1}^{2}\prod_{n=1}^{\infty}(x_{j,n}z_j)^{e_{j,n,k}}\right) z_1^{[u_{1,k}-1]^+} z_2^{[u_{2,k}-\mathbf{1}_{\mathbf{u_{1,k}=0}}]^+}\right] \\ &= B(x_{1,1}z_1, x_{2,1}z_2) \\ &\times \left\{\mathrm{E}\left[\left(\prod_{j=1}^{2}\prod_{n=1}^{\infty}(C_{j,n}(x_{j,n+1}z_j))^{e_{j,n,k-1}}\right)\right.\right. \\ &\left.\times z_2^{[u_{2,k}-1]^+}\mathbf{1}_{\mathbf{u_{1,k}=0}}\right] \\ &+ \mathrm{E}\left[\left(\prod_{j=1}^{2}\prod_{n=1}^{\infty}(C_{j,n}(x_{j,n+1}z_j))^{e_{j,n,k-1}}\right)\right. \\ &\left.\left.\times z_1^{u_{1,k}-1} z_2^{u_{2,k}}\mathbf{1}_{\mathbf{u_{1,k}>0}}\right]\right\}, \end{aligned}$$

by using the system equations (4)-(6) and by taking into account that $b_{1,k}$ and $b_{2,k}$ are statistically independent of the other random variables involved. Here,

$$C_{j,n}(z) \triangleq \mathrm{E}\left[z^{c^{(i)}_{j,n,k-1}}\right] = 1 - p_j(n) + p_j(n)z, \tag{7}$$

$n \geq 1, j = 1, 2$. This follows from the fact that the $c^{(i)}_{j,n,k-1}$'s are Bernoulli distributed random variables as mentioned before. We now use the property that a system void of class-j packets at the beginning of a slot implies that no class-j packets arrived in the system during the previous slot. Or in other words, using that $a_{j,k-1} = 0$ - or equivalently that $e_{j,n,k-1} = 0$ for all n - if $u_{j,k} = 0$, we find

$$\begin{aligned} &P(x_{1,1}, x_{1,2}, .., z_1, x_{2,1}, x_{2,2}, .., z_2) \\ &= \frac{B(x_{1,1}z_1, x_{2,1}z_2)}{z_1 z_2}[z_1(z_2-1)P(0, \ldots, 0) + z_2 \times \\ &P(C_{1,1}(x_{1,2}z_1), C_{1,2}(x_{1,3}z_1), .., z_1, C_{2,1}(x_{2,2}z_2), .., z_2) \\ &+ (z_1 - z_2)P(0, .., 0, C_{2,1}(x_{2,2}z_2), C_{2,2}(x_{2,3}z_2), .., z_2)]. \end{aligned} \tag{8}$$

with P the limiting function of P_k and P_{k+1} for $k \to \infty$. The functional equation (8) contains all information concerning the steady-state behavior of the system, although not in a transparent form. Nevertheless, several explicit results can be derived from it, which is the subject of the following sections. For future reference, we end this section with the definition of some joint pgfs concerning the class-1 and the total system content:

$$\begin{aligned} P_1(x_1, x_2, .., z) &\triangleq P(x_1, x_2, .., z, 1, .., 1), \\ P_T(x_{1,1}, x_{1,2}, .., x_{2,1}, x_{2,2}, .., z) &\\ &\triangleq P(x_{1,1}, x_{1,2}, .., z, x_{2,1}, x_{2,2}, .., z), \end{aligned}$$

and the corresponding functional equations:

$$\begin{aligned} P_1(x_1, x_2, .., z) = &\frac{B_1(x_1 z)}{z}[(z-1)P_1(0, .., 0) \\ &+ P_1(C_{1,1}(x_2 z), C_{1,2}(x_3 z), .., z)], \end{aligned} \tag{9}$$

$$\begin{aligned} &P_T(x_{1,1}, x_{1,2}, .., x_{2,1}, x_{2,2}, .., z) \\ &= \frac{B(x_{1,1}z, x_{2,1}z)}{z}[(z-1)P_T(0, .., 0) \\ &+ P_T(C_{1,1}(x_{1,2}z), C_{1,2}(x_{1,3}z), .., C_{2,1}(x_{2,2}z), .., z)]. \end{aligned} \tag{10}$$

4 Buffer content

For general $(x_{1,1}, x_{1,2}, .., z_1, x_{2,1}, x_{2,2}, .., z_2)$, the functional equation (8) is hard to solve. Therefore, we solve it for a specific set of these arguments and discuss how moments of the buffer content are calculated.

4.1 Solving the functional equation

We here select only those values of $x_{j,n}$ and z_j, $n \geq 1, j = 1, 2$, for which the P-functions on both sides of equation (8) have identical arguments (when non-zero), i.e., we choose $x_{j,n} = C_{j,n}(x_{j,n+1}z_j)$ for $j = 1, 2$, $n \geq 1$. By using (1) and (7) in this expression, $x_{j,n}$ can be solved in terms of z_j. Denoting this solution by $\chi_{j,n}(z_j)$, we find

$$\chi_{j,n}(z_j) = \frac{\sum_{i=n}^{\infty} l_j(i) z_j^{i-n}}{1 - \sum_{i=1}^{n-1} l_j(i)}, \qquad n \geq 1. \tag{11}$$

In particular, we have that $\chi_{j,1}(z_j) = L_j(z_j)/z_j$ and $\chi_{j,n}(1) = 1$, $n \geq 1$. Choosing $x_{j,n} = \chi_{j,n}(z_j)$ in (8), we obtain

$$\begin{aligned} &P(\chi_{1,1}(z_1), \chi_{1,2}(z_1), .., z_1, \chi_{2,1}(z_2), \chi_{2,2}(z_2), .., z_2) \\ &= \frac{B(L_1(z_1), L_2(z_2))}{z_2\left[z_1 - B(L_1(z_1), L_2(z_2))\right]}[z_1(z_2-1)P(0,..,0) \\ &\quad + (z_1 - z_2)P(0,..,0,\chi_{2,1}(z_2), \chi_{2,2}(z_2), .., z_2)]. \end{aligned}$$

$P(\chi_{1,1}(z_1), \chi_{1,2}(z_1), .., z_1, \chi_{2,1}(z_2), \chi_{2,2}(z_2), .., z_2)$ can be fully determined by applying Rouché's theorem and the normalization condition, as is e.g. done in [20]. This leads to

$$\begin{aligned} P(0,..,0,\chi_{2,1}(z_2), \chi_{2,2}(z_2), .., z_2) &= \frac{Y(z_2)(z_2-1)P(0,..,0)}{z_2 - Y(z_2)} \\ P(0,..,0) &= 1 - \rho_T \end{aligned}$$

and finally

$$\begin{aligned} &P(\chi_{1,1}(z_1), \chi_{1,2}(z_1), .., z_1, \chi_{2,1}(z_2), \chi_{2,2}(z_2), .., z_2) \\ &= (1-\rho_T)\frac{B(L_1(z_1), L_2(z_2))(z_2-1)}{z_1 - B(L_1(z_1), L_2(z_2))}\frac{z_1 - Y(z_2)}{z_2 - Y(z_2)}, \end{aligned} \tag{12}$$

with $Y(z)$ implicitly defined as

$$Y(z) \triangleq B(L_1(Y(z)), L_2(z)), \quad |Y(z)| < 1 \text{ if } |z| < 1. \tag{13}$$

Expression (12) can also be explained as follows: expression (11) denotes the pgf of the number of remaining packets that a class-j session sends after its n-th packet. Indeed, this session lasts for i slots with probability $l_j(i)/(1 - \sum_{m=1}^{n-1} l_j(m))$, $i = n, .., \infty$. The left-hand side of expression (12) thus equals the joint pgf of the buffer contents of both classes at the beginning of a slot in the steady state augmented with all packets of sessions that are already ongoing in the previous slot but which have not yet arrived. It is then observed that these quantities equal the buffer contents in a system with the same generation process of sessions but where all packets of a session arrive *simultaneously* in its slot of generation, i.e., when observing a system with batch arrivals instead of the one with train arrivals. This can be understood by noting that when an identical generation process of sessions is applied the service process of both systems is identical. (Note further that this would no longer be the case if the model was extended to e.g. not necessarily back-to-back packet arrivals in a session.) Such a priority queue with batch arrivals has already been analyzed in [19], leading to expression (12).

4.2 Performance measures

By substituting $x_{1,n}$ and $x_{2,n}$ ($n \geq 1$) by 1 in expression (8) a functional equation is found for the joint pgf of the buffer content of both classes. It does not seem to be possible to derive an explicit expression for this pgf from this functional equation. However, all moments of the class-1 and the total buffer content as well as the mean of the class-2 buffer content can be calculated from the results of subsection 4.1. The moments of the class-1 content can be calculated from (9) and (12) with $z_2 = 1$ (by taking appropriate derivatives, for more details on this we refer to [22]). Similarly, the moments of the total buffer content are calculated from (10) and (12) with $z_1 = z_2$. The mean class-2 buffer content is finally calculated as the difference between the mean total buffer content and the mean class-1 content.

Obtaining higher moments of the class-2 buffer content is still an open issue at the moment, since the dependency between the class-1 and class-2 buffer contents influences these. As discussed before, we are not able to characterize this dependency. However, we show in the following section that this does not prohibit us from obtaining the moments of the low-priority packet delay.

5 Packet delay

The delay of a packet is defined as the number of slots between the end of the packet's slot of arrival and the end of its departure slot (thus excluding its arrival slot and including its departure slot). Within each class, we assume that packets are transmitted in the order of their arrival. Recall that class-1 packets have HOL priority over class-2 packets. We analyze the class-1 and class-2 packet delays separately in the remainder of this section.

5.1 Class-1 packet delay

The analysis of the class-1 packet delay is rather easy once the observation is made that transmission of class-1 packets is not influenced by class-2 packets in the system, due to the HOL priority scheduling discipline. Due to a distributional form of Little's law being applicable here [18], $D_1(z)$, the pgf of the class-1 packet delay in the steady state, is expressed in terms of the pgf $P_1(1,..,z)$ of the buffer content of class 1 at the beginning of a random slot, as follows:

$$D_1(z) = \frac{P_1(1,..,z) - 1 + \rho_1}{\rho_1}. \tag{14}$$

We may thus derive the moments of the class-1 packet delay from the moments of the class-1 system content. E.g., the mean class-1 packet delay E$[d_1]$ is given by

$$\mathrm{E}[d_1] = 1 + \frac{\rho_1 B_1'(1)L_1''(1) + B_1''(1)(L_1'(1))^2}{2\rho_1(1-\rho_1)}.$$

5.2 Class-2 packet delay

The analysis of the steady-state class-2 packet delay is more involved, because of the HOL priority discipline. We tag a random class-2 packet and denote it by Q_2. We denote the slot during which Q_2 arrives by S_2. We first make the following key observation: if a class-1 packet is transmitted before Q_2, all packets of the same session of this class-1 packet are transmitted before Q_2 as well. Indeed, only other class-1 packets can be transmitted between the transmissions of two randomly chosen packets of a *same* class-1 session.

Furthermore, we denote the number of class-1 sessions that have sent their n-th packet during slot S_2 by $e^*_{1,n}$, and the total system content at the beginning of the following slot by u^*_T. Furthermore, let r_2 indicate the number of packets arriving during slot S_2 and to be transmitted after packet Q_2. Before writing down an expression for and analyzing the delay of Q_2, we first concentrate on the *virtual delay* w_2 of Q_2. This virtual delay is here defined as the delay when no *new sessions* are generated after slot S_2. Then w_2 equals

$$w_2 = u^*_T - r_2 + \sum_{n=1}^{\infty} \sum_{i=1}^{e^*_{1,n}} l^+_{1,n,i}, \tag{15}$$

with $l^+_{1,n,i}$ the number of packets arriving after slot S_2 of the i-th class-1 session that generated its n-th packet during slot S_2. The virtual delay thus equals the superposition of the buffer content just after slot S_2 and to be transmitted no later than Q_2 and the packets that arrive after slot S_2 of class-1 sessions which were already generating a packet during slot S_2. Note that the $l^+_{1,n,i}$'s are all independent of the system state just after slot S_2. Their pgf is given by $\chi_{1,n}(z)$ (see (11)). With the definition

$$Q(x_1, x_2, .., y, z) \triangleq \mathrm{E}\left[\left(\prod_{n=1}^{\infty} x_n^{e^*_{1,n}}\right) y^{r_2} z^{u^*_T}\right],$$

expression (15) leads to the pgf of w_2:

$$W_2(z) \triangleq \mathrm{E}[z^{w_2}] = Q(\chi_{1,1}(z), \chi_{1,2}(z), .., 1/z, z). \tag{16}$$

Relating the buffer content distribution just after the arrival slot of a random class-2 packet to the buffer content distribution at the beginning of a random slot (i.e., a manifestation of the typical renewal-theory paradox, see e.g. [15]), we find

$$\begin{aligned} &Q(x_1, x_2, .., y, z) \\ &\quad = \frac{P_T(x_1, x_2, .., 1, .., z) - P_T(x_1, x_2, .., y, .., z)}{\rho_2(1-y)}, \end{aligned} \tag{17}$$

with P_T the function calculated in section 4.

We now relate the delay d_2 and the virtual delay w_2 of packet Q_2. Obviously, the virtual delay is an integral part of the delay. During the transmission of a certain packet, say P, belonging to the virtual delay workload, new class-1 sessions may be generated, the transmission of their packets adding to the delay of Q_2. During the transmission of the packets of these class-1 sessions new class-1 sessions may in turn be generated, which further add to the delay of Q_2, etc. The total number of all packets of all these sessions (including packet P itself) is called the *sub-busy period* initiated by P. Summarizing, we can write

$$d_2 = \sum_{i=1}^{w_2-1} v_{1,i} + 1, \tag{18}$$

with $v_{1,i}$ the sub-busy period added by the i-th packet of the virtual delay workload. Note that these $v_{1,i}$'s are all i.i.d. with pgf denoted by $V_1(z)$. By z-transforming expression (18), we then obtain

$$D_2(z) \triangleq \mathrm{E}[z^{d_2}] = \frac{zW_2(V_1(z))}{V_1(z)}.$$

Using (16), we find

$$D_2(z) = \frac{zQ(\chi_{1,1}(V_1(z)), \chi_{1,2}(V_1(z)), .., 1/V_1(z), V_1(z))}{V_1(z)}. \tag{19}$$

The use of (17) in the latter expression provides us with an expression for $D_2(z)$ in terms of the P_T-function and $V_1(z)$. The P_T-function is characterized in sections 3 and 4. So what remains is the calculation of the function $V_1(z)$.

In order to do so, we note that the $v_{1,i}$'s in expression (18) can be expressed as

$$v_{1,i} = 1 + \sum_{m=1}^{b_{1,i}} \sum_{n=1}^{l^{(m)}_{1,i}} v^{(m,n)}_{1,i}, \tag{20}$$

with $b_{1,i}$ the number of new class-1 sessions generated during the transmission of the i-th packet of the virtual delay workload, $l^{(m)}_{1,i}$ the number of packets the m-th session of $b_{1,i}$ contains and $v^{(m,n)}_{1,i}$ the sub-busy period initiated by the n-th packet of the m-th session of $b_{1,i}$. Indeed, a sub-busy period initiated by a packet consists of the transmission slot of that packet and the sub-busy periods of all packets of all sessions that are generated during that slot. Note that the $v^{(m,n)}_{1,i}$'s are i.i.d. having the same pgf as the $v_{1,i}$'s, i.e., $V_1(z)$. Expression (20) then leads to the following implicit expression for $V_1(z)$:

$$V_1(z) = zB_1(L_1(V_1(z))). \tag{21}$$

Although this does not lead to an explicit formula for $V_1(z)$, its derivatives in $z = 1$ can be explicitly calculated due to the knowledge that $V_1(1) = 1$, since $V_1(z)$ is a pgf.

Expression (19) combined with expressions (17) and (21) enables us to calculate the moments of the class-2 packet delay as functions of (partial) derivatives of the P_T-function, evaluated for all arguments equal to 1. We have argued in the previous section that these derivatives can be calculated. In general, the calculations of the moments of the class-2 delay are however highly complex, since several partial derivatives of P_T have to be calculated, which is a non-trivial task. For instance, the first derivative of expression (19) evaluated in $z = 1$ leads to an expression containing (partial) derivatives of $\chi_{1,m}$, V_1 and P_T. These derivatives can in turn be calculated from expressions (11), (21) and (10) and (12) respectively. The following final expression for the mean class-2 packet delay can then be obtained

$$\mathrm{E}[d_2] = 1 + \frac{\rho_T L_2''(1)}{2L_2'(1)(1-\rho_T)} + \frac{B_2''(1)L_2'(1)}{2B_2'(1)(1-\rho_T)} + \frac{\frac{\partial^2 B}{\partial z_1 \partial z_2}(1,1)L_1'(1)}{B_2'(1)(1-\rho_T)} + \frac{B_1'(1)L_1''(1) + B_1''(1)(L_1'(1))^2}{2(1-\rho_1)(1-\rho_T)}.$$

Higher moments of the class-2 packet delay can be calculated as well.

6 Numerical examples

We illustrate our findings by means of a numerical example. We assume that class-1 and class-2 sessions are both generated according to independent Poisson processes with means λ_1 and λ_2 respectively. We thus have

$$B(z_1, z_2) = e^{\lambda_1(z_1-1)}e^{\lambda_2(z_2-1)}.$$

We are primarily interested in the influence of the variability of the session lengths on the performance of the system, i.e. on the mean packet delays of both classes (for the influence of the mean session lengths we refer to [3, 20]). Therefore, we firstly consider the example of negative binomially distributed class-j session lengths with parameters m_j and p_j, i.e., with pgf

$$L_j(z) = \left(\frac{p_j z}{1-(1-p_j)z}\right)^{m_j}.$$

By decreasing m_j while keeping $\mathrm{E}[l_j] = L_j'(1) = m_j/p_j$ constant, the variance of the session lengths $\mathrm{Var}[l_j] = m_j(1-p_j)/p_j^2$ can be increased while the mean value is kept constant. It may be noted that $m_j = 1$ corresponds to a geometric distribution, while $p_j = 1$ corresponds to deterministic session lengths.

Throughout this section, we consider the high-priority load to be a quarter of the total load, i.e., $\alpha \triangleq \rho_1/\rho_T = 0.25$. The means of the session lengths equal 16 slots for both classes.

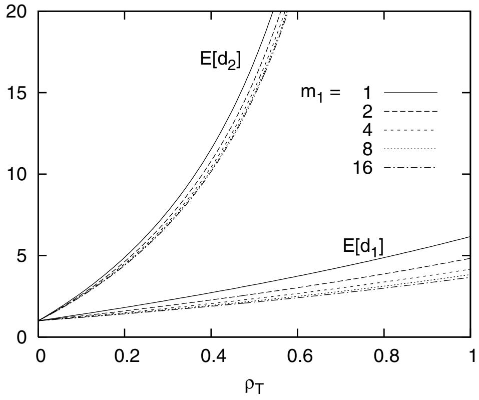

Figure 1. Mean packet delays of both classes versus the total load for $\alpha = 0.25$, E$[l_1] = 16$, E$[l_2] = 16$ and $m_2 = 2$.

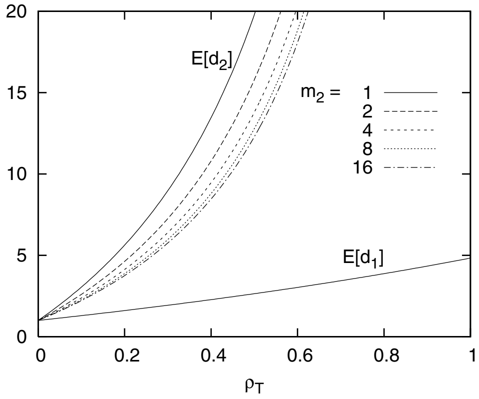

Figure 2. Mean packet delays of both classes versus the total load for $\alpha = 0.25$, E$[l_1] = 16$, E$[l_2] = 16$ and $m_1 = 2$.

In Figure 1 (Figure 2 respectively), we depict the mean delays of packets of both classes as functions of the total load ρ_T when $m_2 = 2$ ($m_1 = 2$ respectively) and for varying m_1 (m_2 respectively). Firstly, it can be concluded from these figures that priority scheduling indeed differentiates the delay characteristics of both classes. Secondly, we see that the mean delays of packets are influenced by the variance of the session lengths of their own class. Thirdly, it is shown that the mean delay of low-priority packets is also influenced by the variance of the high-priority session lengths, although not as much as by the variance of the lengths of the sessions of its own class. Obviously, the high-priority packet delay does not depend on the low-priority arrival process.

In the first two figures, we showed the mean delays when

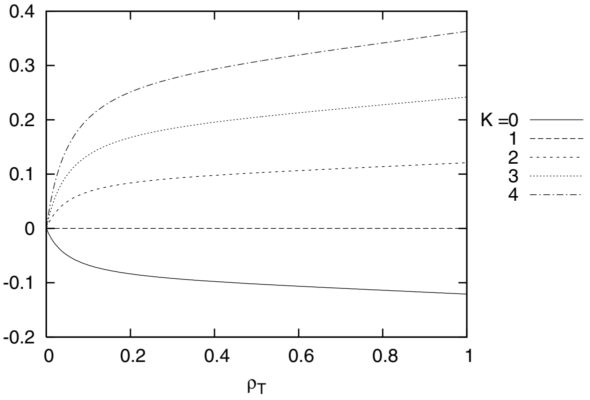

Figure 3. Relative deviation of the mean class-2 delay versus the total load for $\alpha = 0.25$, E$[l_j] = 16$, Var$[l_j] = K_j(16^2 - 16)$, $j = 1, 2$ and $K_1 = K, K_2 = 1$.

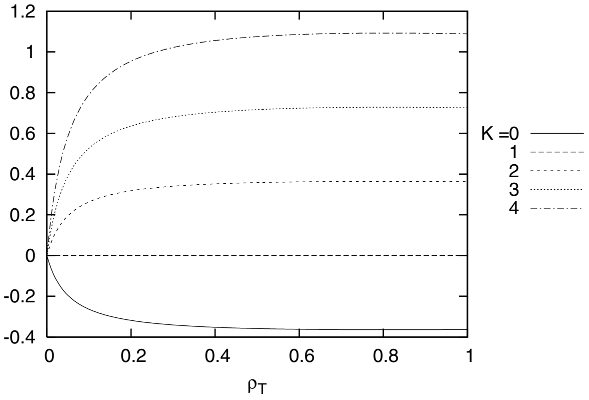

Figure 4. Relative deviation of the mean class-2 delay versus the total load for $\alpha = 0.25$, E$[l_j] = 16$, Var$[l_j] = K_j(16^2 - 16)$, $j = 1, 2$ and $K_1 = 1, K_2 = K$.

the variance of the session lengths was less than or equal to the variance of geometrically distributed session lengths (with the same mean value). To conclude, we show the impact of higher variances of the session lengths in Figures 3 and 4. In Figure 3, the class-2 session lengths are geometrically distributed, while the variance of the class-1 session lengths is assumed to equal $K_1(16^2 - 16)$. The relative deviation of the mean class-2 packet delay, defined as $(\mathrm{E}[d_2]_{K_1=K} - \mathrm{E}[d_2]_{K_1=1})/\mathrm{E}[d_2]_{K_1=1}$, is plotted for several values of K. Note that the reference case $K = 1$ corresponds to the geometric distribution. The case $K = 0$ corresponds to the deterministic case while $K > 1$ corresponds to distributions that have a larger variance than the geometric one. Note that a variance with $K > 1$ can easily be constructed by using a mix of geometric distributions. In Figure 4, the class-1 session lengths are geometrically distributed and the variance of the class-2 session lengths is assumed to equal $K_2(16^2 - 16)$. Now, the relative deviation $(\mathrm{E}[d_2]_{K_2=K} - \mathrm{E}[d_2]_{K_2=1})/\mathrm{E}[d_2]_{K_2=1}$ of the mean class-2 packet delay is plotted for several values of K. We once again see from both plots that the variances of the class-1 and class-2 session lengths have a non-negligible impact on the mean class-2 delay. Furthermore, we may conclude from Figure 4 that in this case $\mathrm{E}[d_2]_{K_2=K} = C(K).\mathrm{E}[d_2]_{K_2=1}$, with $C(K)$ nearly independent of the total load when the load is high. This is not the case when the high-priority lengths are varied. A linear relation between the relative deviation and K can still be envisaged though.

7 Conclusions

In this paper, we studied a discrete-time two-class priority queue with a two-layered arrival process. Packets of variable-length sessions of both classes arrive to the system at the rate of one packet per slot. The session lengths may have a general distribution. Since the arrival process is fairly general, the analysis is obviously non-trivial. Using probability generating functions, we have shown that explicit closed-form expressions for the mean values of the system contents and packet delays of both classes can be derived, as well as higher moments for the packet delays of both classes. We have finally shown the influence of the variance of the session lengths of both classes on the mean (low-priority) packet delay through some numerical examples.

This research can be extended in different ways. A non-exhaustive list is a) the calculation of tail probabilities of the packet delay, which is non-trivial for priority queues, see e.g. [12, 13]; b) the analysis of the session delay and c) the analysis of a model where the packets in a session do not necessarily arrive back-to-back, which would highly complicate the analysis since we used this assumption several times in this paper.

Acknowledgment. The first author is a Postdoctoral Fellow with the Fund for Scientific Research, Flanders (F.W.O.-Vlaanderen), Belgium.

References

[1] H. Bruneel. Packet delay and queue length for statistical multiplexers with low-speed access lines. *Computer Networks and ISDN Systems*, 25(12):1267–1277, 1993.

[2] H. Bruneel. Calculation of message delays and message waiting times in switching elements with slow access lines. *IEEE Transactions on Communications*, 42(2/3/4):255–259, 1994.

[3] B. Choi, D. Choi, Y. Lee, and D. Sung. Priority queueing system with fixed-length packet-train arrivals. *IEE Proceedings-Communications*, 145(5):331–336, 1998.

[4] I. Cidon, A. Khamisy, and M. Sidi. Delay, jitter and threshold crossing in ATM systems with dispersed messages. *Performance Evaluation*, 29(2):85–104, 1997.

[5] J. Daigle. Message delays at packet-switching nodes serving multiple classes. *IEEE Transactions on Communications*, 38(4):447–455, 1990.

[6] S. De Vuyst, S. Wittevrongel, and H. Bruneel. Statistical multiplexing of correlated variable-length packet trains: an analytic performance study. *Journal of the Operational Research Society*, 52(3):318–327, 2001.

[7] K. Elsayed and H. Perros. The superposition of discrete-time Markov renewal processes with an application to statistical multiplexing of bursty traffic sources. *Applied Mathematics and Computation*, 115(1):43–62, 2000.

[8] G. Heijenk, M. E. Zarki, and I. Niemegeers. Modelling of segmentation and reassembly processes in communication networks. In *Proceedings of ITC14*, pages 513–524, Antibes, 1994.

[9] L. Hoflack, S. De Vuyst, S. Wittevrongel, and H. Bruneel. Analytic traffic model of web server. *Electronics Letters*, 44(1):61–63, 2008.

[10] H. Inai and J. Yamakita. A two-layer queueing model to predict performance of packet transfer in broadband networks. *Annals of Operations Research*, 79:349–371, 1998.

[11] F. Kamoun. Performance analysis of a discrete-time queuing system with a correlated train arrival process. *Performance Evaluation*, 63(4-5):315–340, 2006.

[12] K. Laevens and H. Bruneel. Discrete-time multiserver queues with priorities. *Performance Evaluation*, 33(4):249–275, 1998.

[13] T. Maertens, J. Walraevens, and H. Bruneel. Priority queueing systems: from probability generating functions to tail probabilities. *Queueing Systems*, 55(1):27–39, 2007.

[14] M. Mehmet Ali and X. Song. A performance analysis of a discrete-time priority queueing system with correlated arrivals. *Performance Evaluation*, 57(3):307–339, 2004.

[15] I. Mitrani. *Modelling of Computer and Communication Systems*. Cambridge University Press, Cambridge, 1987.

[16] T. Takine, B. Sengupta, and T. Hasegawa. 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.

[17] T. Tan, K. Moinzadeh, and V. Mookerjee. Optimal processing policies for an e-commerce web server. *INFORMS Journal on Computing*, 17(1):99–110, 2005.

[18] B. Vinck and H. Bruneel. A note on the system contents and cell delay in FIFO ATM-buffers. *Electronics Letters*, 31(12):952–954, 1995.

[19] J. Walraevens, B. Steyaert, and H. Bruneel. Performance analysis of a single-server ATM queue with a priority scheduling. *Computers & Operations Research*, 30(12):1807–1829, 2003.

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

[21] S. Wittevrongel. Discrete-time buffers with variable-length train arrivals. *Electronics Letters*, 34(18):1719–1721, 1998.

[22] S. Wittevrongel and H. Bruneel. Correlation effects in ATM queues due to data format conversions. *Performance Evaluation*, 32(1):35–56, 1998.

[23] Y. Xiong and H. Bruneel. Buffer behavior of statistical multiplexers with correlated train arrivals. *International Journal of Electronics and Communications (AEÜ)*, 51(3):178–186, 1997.

[24] T. Yu and K. Lin. QCWS: an implementation of QoS-capable multimedia web services. *Multimedia Tools and Applications*, 30(2):165–187, 2006.