Performance Analysis of a Priority Queue with Session-based Arrivals and its Application to E-commerce Web Servers

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 a session-based arrival process. We consider an infinitely large 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 packets of the other class. The model is motivated by E-commerce web servers and web servers handling delay-sensitive and delay-insensitive content. By using probability generating functions, 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 demonstrated. We furthermore use our analysis to provide specific results for an E-commerce web server.

Keywords-priority; session arrivals; E-commerce; web server; queueing analysis

1 Introduction

We analyze a two-class discrete-time Head-Of-the-Line (HOL) priority queue with a session-based arrival process.

HOL priority scheduling is one of the main scheduling types in network buffers to diversify the delays of traffic streams with different delay requirements [24].When 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 the following: one application might provide revenues for the provider while another does not (or to a lesser extent). It is then natural (and profitable) to give priority to the packets of the first application.

Besides priority scheduling, there are numerous other scheduling types proposed in the literature to diversify the Quality-of-Service (QoS) of different applications. A (theoretical) scheduling discipline is Generalized Processor Sharing (GPS). With this discipline, the 'transmission unit' spends weighted fractions of its capacity on the different classes. Delay-sensitive traffic gets a larger weight than delay-insensitive traffic, so that it gets some kind of preferential treatment. One possible implementation of GPS is Weighted Fair Queueing. For a further overview of such scheduling disciplines, we refer to [12]. However, priority scheduling is still one of the most popular scheduling types, since it is relatively easy to implement and to operate.

In the current paper, we further 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 [14, 17].

In particular, the suggested arrival process is an ideal candidate to model the output buffer of a web server [15]. 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. In the case of an E-commerce web server, it makes sense to prioritize the downloads on a (potential) revenue base [27], that is, to give priority to the transmission of packets of content that is likely to provide (large) revenues. Furthermore, most web pages contain content that is delay-sensitive, for instance multimedia content. Priority can then also be given to the transmission of files containing this content over other downloads [37].

From a queueing-analysis point of view, the combination of a priority scheduling discipline and a session-based arrival process with generally distributed session lengths forms the main novelty of this paper. We thereby extend previous analyses [5, 32] where the session lengths were assumed to have a specific distribution (deterministic and geometric respectively). The distributions of the session lengths may further 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 [5, 32]), 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. From a networking point of view, the added value lies in the application of our results to an E-commerce web server. We finally note that this paper is the extended version of [1].

The remainder of the paper is structured as follows. In the next two sections, we describe some related literature and present the mathematical model respectively. In section 4, we construct a functional equation. This functional equation is the starting point of the analysis of the steady-state number of arrivals per slot, the steady-state buffer content and steady-state packet delay, described in sections 5, 6 and 7, respectively. Numerical examples are treated in section 8, while we apply our results to an E-commerce web server in section 9. We finally conclude this paper in section 10.

2 Related literature

A first property of our model is the HOL priority scheduling. There have been a large number of contributions in the related literature with respect to the performance analysis of HOL priority queues. In particular, discrete-time HOL priority queues with deterministic service times equal to one slot have been studied in [4, 9, 11, 13, 19, 20, 22, 25, 26, 28, 30, 31, 35]. Hashida and Takahashi [13] analyze a two-class priority system, where the high-priority arrivals and low-priority arrivals are governed by a two-state Markov-modulated Batch Bernoulli Process and a Batch Bernoulli Process respectively. The numbers of per-slot arriving high-priority packets are governed by an underlying Markov chain and the numbers of per-slot low-priority arrivals are independent and identically distributed (i.i.d.). Application of a conservation law leads to expressions for the mean delays of both classes. Takine et al. [26] analyze the same model as in [13] by means of matrix-analytic techniques. Moments of high-priority, low-priority and total system contents and moments of high-priority and low-priority delay are calculated. In [25], bounds for the delay distribution are given in a multi-server queue with a rather general arrival process. Xabier Albizuri et al. [35] study the delay of the low-priority traffic in a multi-server queue by assuming that the number of servers available for the low-priority traffic is variable (depending on the number of high-priority packets served at the time). Mehmet Ali and Song [22] analyze a queue with the arrival process existing of a number of two-state Markovian sources and by using probability generating functions. In [19], priority queueing systems with a general number of priority classes are analyzed. The distribution of the number of per-slot arrivals depends on the state of a two-state Markov chain. In [4, 20], two-class multiserver queues are analyzed with the number of arrivals i.i.d. from slot to slot. The joint pgf of the system contents of both classes is calculated in both papers (although the analysis in [4] is more tedious than in [20]). The pgfs of the delays of both types of packets are also calculated in [20]. From these pgfs, moments of the analyzed stochastic variables are calculated in both papers. In [4], the corresponding probabilities are furthermore numerically determined using Fast Fourier Transforms, while these probabilities are analytically approximated for high values of the stochastic variable (tail probabilities) in [20]. Walraevens et al. [30, 31] study the steady-state buffer content and packet delay in the special case of an output-queueing switch with Bernoulli arrivals and the transient buffer content respectively. Finally, in [9, 11, 28], different queueing models with finite buffer size are studied.

A second important characteristic of our model is the session-based nature of the arrival process. First-In-First-Out (FIFO) queues with session-based arrivals are analyzed in [2, 3, 8, 33, 34]. Bruneel [2, 3] and Wittevrongel [33] analyze different aspects of FIFO queues with a session-based arrival process and geometrically distributed session lengths. This model is further extended to generally distributed session lengths by Wittevrongel and Bruneel [33, 34]. De Vuyst et al. [8] further added a second correlation in the model (besides the session nature of the arrival

process) by introducing a two-state environment that determines the number of starting sessions. Somewhat related on/off-type arrival models are considered in [10,18,36], also for the FIFO case. Further in [6], 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, HOL priority queues with session-based arrivals have been studied as well. Daigle [7] calculates mean session delays in a continuous-time priority queue with session-based arrivals. Our current analysis is a direct extension of the analyses in [5] and [32] where discrete-time HOL priority queues are analyzed with deterministic and geometric session lengths respectively.

3 Framework and queueing model

We make extensive use of probability generating functions (pgfs) in this paper. The pgf of a generic discrete random variable X is defined as $X(z) \triangleq \mathrm{E}[z^X]$ with E[.] the expected-value operator. There is a one-to-one map between the probability mass function (pmf) $x(n) \triangleq \text{Prob}[X = n], n \geq 0$ and its pgf $X(z)$, as $X(z)$ is the z-transform of the sequence $\{x(n), n \geq 0\}$:

$$X(z) = \sum_{n=0}^{\infty} x(n) z^n. \tag{1}$$

$X(z)$ thus completely characterizes the random variable. Note that $X(1) = 1$. Furthermore, moments of the random variable are easily calculated by means of the moment-generating property of pgfs. For instance, the mean value of a random variable is given by taking the derivative of its pgf in 1: $\mathrm{E}[X] = X'(1)$. It is straightforward to extend the notion of pgfs to the joint pgf of more than one random variable.

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 independent and identically distributed (i.i.d.). 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

$$B(z_1, z_2) \triangleq \mathrm{E}\left[z_1^{b_{1,k}} z_2^{b_{2,k}}\right]. \tag{2}$$

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 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.,

$$p_j(n) \triangleq \frac{1 - \sum_{i=1}^{n} l_j(i)}{1 - \sum_{i=1}^{n-1} l_j(i)}. \tag{3}$$

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

$$A_k(z_1, z_2) \triangleq \mathrm{E}\left[z_1^{a_{1,k}} z_2^{a_{2,k}}\right]. \tag{4}$$

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 (the mean value of a random variable is given by the first derivative in 1 of the pgf of the variable). Therefore the load generated by class-j packets equals

$$\rho_j = B_j'(1) L_j'(1), \tag{5}$$

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

$$\rho_T \triangleq \rho_1 + \rho_2 = B_1'(1) L_1'(1) + B_2'(1) L_2'(1) < 1. \tag{6}$$

4 Start of the analysis

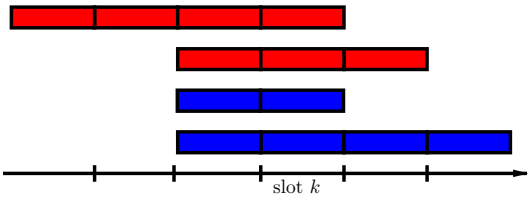
In this section, we first give a Markov-chain description of the system. In a second part, we construct a functional equation that summarizes this Markov chain and that is the starting point of further calculations in the next sections.

4.1 Markov-chain description

The arrival process is fully described 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 hold:

$$\begin{aligned} &e_{j,1,k} = b_{j,k}; \\ &e_{j,n+1,k} = \textstyle\sum_{i=1}^{e_{j,n,k-1}} c_{j,n,k-1}^{(i)}, \qquad n \geq 1, \end{aligned} \tag{7}$$

$j = 1, 2$. 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 if and only if the i-th



$e_{1,2,k} = 1, e_{1,4,k} = 1 \Rightarrow a_{1,k} = 2$

$e_{2,2,k} = 2 \Rightarrow a_{2,k} = 2$

$c^{(1)}_{1,2,k} = 1, c^{(1)}_{1,4,k} = 0$

$c^{(1)}_{2,2,k} = 0, c^{(2)}_{2,2,k} = 1$

Figure 1. Example illustrating the involved random variables of the arrival process. During slot k two high-priority sessions (red, on top) and two low-priority sessions (blue, bottom) are sending a packet. All non-zero random variables concerning slot k are given.

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 equations (7) can then be understood as follows: $e_{j,1,k}$ represents the number of class-j sessions that deliver their first packet during slot k and therefore equals the new number of sessions that start in that slot. The variable $e_{j,n+1,k}$ corresponds to the number of class-j packets that deliver their $(n+1)$-st packet and therefore equals the number of class-j packets that delivered their n-th packet in the previous slot ($e_{j,n,k-1}$) and that are still sending a packet during the current slot.

The variable $a_{j,k}$, the total number of class-j packets arriving during slot k, can be expressed as

$$a_{j,k} = \sum_{n=1}^{\infty} e_{j,n,k}, \qquad j = 1, 2. \tag{8}$$

The above defined variables are illustrated in Figure 1.

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:

$$\begin{aligned} u_{1,k+1} &= [u_{1,k} - 1]^+ + a_{1,k}; \\ u_{2,k+1} &= [u_{2,k} - \mathbf{1}_{\mathbf{u_{1,k}=0}}]^+ + a_{2,k}, \end{aligned} \tag{9}$$

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

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 (7)-(9) fully describe the behavior of the system.

4.2 Construction of the functional equation

We introduce the joint pgf of the state vector:

$$\begin{aligned} &P_k(x_{1,1}, x_{1,2}, \ldots, z_1, x_{2,1}, x_{2,2}, \ldots, z_2) \\ &\qquad \triangleq \mathrm{E}\left[\prod_{j=1}^{2}\left(\prod_{n=1}^{\infty} x_{j,n}^{e_{j,n,k-1}}\right) z_j^{u_{j,k}}\right]. \end{aligned} \tag{10}$$

It follows that

$$\begin{aligned} &P_{k+1}(x_{1,1}, x_{1,2}, \ldots, z_1, x_{2,1}, x_{2,2}, \ldots, z_2) = \\ &\mathrm{E}\left[\left(\prod_{j=1}^{2}\prod_{n=1}^{\infty}(x_{j,n}z_j)^{e_{j,n,k}}\right) z_1^{[u_{1,k}-1]^+} z_2^{[u_{2,k}-\mathbf{1}_{\mathbf{u_{1,k}=0}}]^+}\right] \\ &= \mathrm{E}\left[(x_{1,1}z_1)^{b_{1,k}}(x_{2,1}z_2)^{b_{2,k}}\right] \\ &\times \left\{\mathrm{E}\left[\left(\prod_{j=1}^{2}\prod_{n=2}^{\infty}\prod_{i=1}^{e_{j,n-1,k-1}}(x_{j,n}z_j)^{c^{(i)}_{j,n-1,k-1}}\right)\right.\right. \\ &\left.\times z_2^{[u_{2,k}-1]^+}\mathbf{1}_{\mathbf{u_{1,k}=0}}\right] \\ &+ \mathrm{E}\left[\left(\prod_{j=1}^{2}\prod_{n=2}^{\infty}\prod_{i=1}^{e_{j,n-1,k-1}}(x_{j,n}z_j)^{c^{(i)}_{j,n-1,k-1}}\right)\right. \\ &\left.\left.\times z_1^{u_{1,k}-1}z_2^{u_{2,k}}\mathbf{1}_{\mathbf{u_{1,k}>0}}\right]\right\} \\ &= B(x_{1,1}z_1, x_{2,1}z_2) \\ &\times \left\{\mathrm{E}\left[\left(\prod_{j=1}^{2}\prod_{n=1}^{\infty}(C_{j,n}(x_{j,n+1}z_j))^{e_{j,n,k-1}}\right)\right.\right. \\ &\left.\times z_2^{[u_{2,k}-1]^+}\mathbf{1}_{\mathbf{u_{1,k}=0}}\right] \\ &+ \mathrm{E}\left[\left(\prod_{j=1}^{2}\prod_{n=1}^{\infty}(C_{j,n}(x_{j,n+1}z_j))^{e_{j,n,k-1}}\right)\right. \\ &\left.\left.\times z_1^{u_{1,k}-1}z_2^{u_{2,k}}\mathbf{1}_{\mathbf{u_{1,k}>0}}\right]\right\}, \end{aligned} \tag{11}$$

by using the law of total probability, using system equations (7)-(9) and by taking into account that $b_{1,k}$ and $b_{2,k}$ are statistically independent of the other random variables involved. Here,

$$C_{j,n}(z) \triangleq \mathrm{E}\left[z^{c^{(i)}_{j,n,k-1}}\right] = 1 - p_j(n) + p_j(n)z, \tag{12}$$

$n \geq 1, j = 1, 2$. This follows from the fact that the $c^{(i)}_{j,n,k-1}$'s are Bernoulli-distributed random variables as

mentioned before (see Figure 1). 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

$$\begin{aligned}
&P_{k+1}(x_{1,1}, x_{1,2}, .., z_1, x_{2,1}, x_{2,2}, .., z_2) \\
&= \frac{B(x_{1,1}z_1, x_{2,1}z_2)}{z_1 z_2}[z_1(z_2-1)P_k(0,\ldots,0) + z_2 \times \\
&P_k(C_{1,1}(x_{1,2}z_1), C_{1,2}(x_{1,3}z_1), .., z_1, C_{2,1}(x_{2,2}z_2), .., z_2) \\
&+ (z_1 - z_2)P_k(0, .., 0, C_{2,1}(x_{2,2}z_2), C_{2,2}(x_{2,3}z_2), .., z_2)].
\end{aligned} \tag{13}$$

In steady state, P_k and P_{k+1} both converge to the same limiting function P. It then follows from equation (13) that this function must satisfy the following functional equation:

$$\begin{aligned}
&P(x_{1,1}, x_{1,2}, .., z_1, x_{2,1}, x_{2,2}, .., z_2) \\
&= \frac{B(x_{1,1}z_1, x_{2,1}z_2)}{z_1 z_2}[z_1(z_2-1)P(0,\ldots,0) + z_2 \times \\
&P(C_{1,1}(x_{1,2}z_1), C_{1,2}(x_{1,3}z_1), .., z_1, C_{2,1}(x_{2,2}z_2), .., z_2) \\
&+ (z_1 - z_2)P(0, .., 0, C_{2,1}(x_{2,2}z_2), C_{2,2}(x_{2,3}z_2), .., z_2)].
\end{aligned} \tag{14}$$

The functional equation (14) contains all information concerning the steady-state behavior of the system, although not in 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:

$$P_1(x_1, x_2, .., z) \triangleq P(x_1, x_2, .., z, 1, .., 1), \tag{15}$$

$$\begin{aligned}
P_T(x_{1,1}, x_{1,2}, .., &x_{2,1}, x_{2,2}, .., z) \\
&\triangleq P(x_{1,1}, x_{1,2}, .., z, x_{2,1}, x_{2,2}, .., z),
\end{aligned} \tag{16}$$

that is, P_1 equals P with arguments $x_{2,j}$ (for all $j \geq 1$) and z_2 equal to 1 and P_T equals P with arguments z_1 and z_2 both equal to z. The corresponding functional equations are

$$\begin{aligned}
P_1(x_1, x_2, .., z) =& \frac{B_1(x_1 z)}{z}[(z-1)P_1(0, .., 0) \\
&+ P_1(C_{1,1}(x_2 z), C_{1,2}(x_3 z), .., z)],
\end{aligned} \tag{17}$$

$$\begin{aligned}
&P_T(x_{1,1}, x_{1,2}, .., x_{2,1}, x_{2,2}, .., z) \\
=& \frac{B(x_{1,1}z, x_{2,1}z)}{z}[(z-1)P_T(0, .., 0) \\
&+ P_T(C_{1,1}(x_{1,2}z), C_{1,2}(x_{1,3}z), .., C_{2,1}(x_{2,2}z), .., z)].
\end{aligned} \tag{18}$$

5 Number of arrivals

Define the joint pgf $E(x_{1,1}, x_{1,2}, .., x_{2,1}, x_{2,2}, ..)$ as follows:

$$E(x_{1,1}, x_{1,2}, .., x_{2,1}, x_{2,2}, ..) \triangleq \lim_{k\to\infty} \mathrm{E}\left[\prod_{j=1}^{2}\prod_{n=1}^{\infty} x_{j,n}^{e_{j,n,k}}\right], \tag{19}$$

i.e., it is the joint pgf of the numbers of class-1 and class-2 sessions that deliver their n-th packet (for all $n \geq 1$) during an arbitrary slot in steady state. This pgf is given by

$$\begin{aligned}
&E(x_{1,1}, x_{1,2}, .., x_{2,1}, x_{2,2}, ..) \\
&\quad = P(x_{1,1}, x_{1,2}, .., 1, x_{2,1}, x_{2,2}, .., 1) \\
&\quad = B(x_{1,1}, x_{2,1}) \\
&\qquad \times E(C_{1,1}(x_{1,2}), C_{1,2}(x_{1,3}), .., C_{2,1}(x_{2,2}), ..).
\end{aligned} \tag{20}$$

The last step is found by putting $z_1 = z_2 = 1$ in (14). Successive applications of (20) then lead to the following explicit result for E:

$$\begin{aligned}
&E(x_{1,1}, x_{1,2}, .., x_{2,1}, x_{2,2}, ..) \\
&\qquad = \prod_{n=0}^{\infty} B(g_1^{(n)}(x_{1,n+1}), g_2^{(n)}(x_{2,n+1})),
\end{aligned} \tag{21}$$

with

$$g_j^{(n)}(x) \triangleq \sum_{i=1}^{n} l_j(i) + x\left(1 - \sum_{i=1}^{n} l_j(i)\right), \tag{22}$$

$j = 1, 2$. To obtain (21), we have used the following relationships, which can easily be derived from (3) and (12):

$$\begin{aligned}
&C_{j,1}(C_{j,2}(..C_{j,n}(x)..)) \\
&\qquad = \sum_{i=1}^{n} l_j(i) + x\left(1 - \sum_{i=1}^{n} l_j(i)\right),
\end{aligned} \tag{23}$$

$$\lim_{n\to\infty} C_{j,i}(C_{j,i+1}(..C_{j,n}(x)..)) = 1, \qquad i \geq 1, \tag{24}$$

$j = 1, 2$.

The joint pgf of the total numbers of arrivals of both classes during a random slot in steady state is given by

$$\begin{aligned}
A(z_1, z_2) =& E(z_1, z_1, .., z_2, z_2, ..) \\
&= \prod_{n=0}^{\infty} B(g_1^{(n)}(z_1), g_2^{(n)}(z_2)),
\end{aligned} \tag{25}$$

which is found from (21). Taking the necessary derivatives of this expression delivers all moments of the class-1, class-2 and total numbers of arrivals per slot in steady state. We find, for instance, that

$$\mathrm{E}[a_j] = B_j'(1)L_j'(1), \tag{26}$$

as expected.

6 Buffer content

For general $(x_{1,1}, x_{1,2}, .., z_1, x_{2,1}, x_{2,2}, .., z_2)$, the functional equation (14) is hard to solve. Therefore, we solve it for a specific set of these arguments and discuss how moments of the steady-state buffer content are calculated. We also comment on the consequences of the fact that we are not able to solve the functional equation for general arguments.

6.1 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 (14) 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 (3) and (12) 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

$$\chi_{j,n}(z_j) = \frac{\sum_{i=n}^{\infty} l_j(i) z_j^{i-n}}{1 - \sum_{i=1}^{n-1} l_j(i)}, \qquad n \geq 1. \tag{27}$$

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 (14), we obtain

$$\begin{aligned} &P(\chi_{1,1}(z_1), \chi_{1,2}(z_1), .., z_1, \chi_{2,1}(z_2), \chi_{2,2}(z_2), .., z_2) \\ &\quad = \frac{B(L_1(z_1), L_2(z_2))}{z_2\,[z_1 - B(L_1(z_1), L_2(z_2))]} [z_1(z_2 - 1)P(0, .., 0) \\ &\qquad + (z_1 - z_2)P(0, .., 0, \chi_{2,1}(z_2), \chi_{2,2}(z_2), .., z_2)]. \end{aligned} \tag{28}$$

$P(\chi_{1,1}(z_1), \chi_{1,2}(z_1), .., z_1, \chi_{2,1}(z_2), \chi_{2,2}(z_2), .., z_2)$ can be fully determined by applying Rouché's theorem and the normalization condition, as is e.g. done in [32]. This leads to

$$\begin{aligned} &P(0, .., 0, \chi_{2,1}(z_2), \chi_{2,2}(z_2), .., z_2) \\ &\qquad = \frac{Y(z_2)(z_2 - 1)P(0, .., 0)}{z_2 - Y(z_2)}, \end{aligned} \tag{29}$$

$$P(0, .., 0) = 1 - \rho_T \tag{30}$$

and finally

$$\begin{aligned} &P(\chi_{1,1}(z_1), \chi_{1,2}(z_1), .., z_1, \chi_{2,1}(z_2), \chi_{2,2}(z_2), .., z_2) \\ &\quad = (1 - \rho_T) \frac{B(L_1(z_1), L_2(z_2))(z_2 - 1)}{z_1 - B(L_1(z_1), L_2(z_2))} \frac{z_1 - Y(z_2)}{z_2 - Y(z_2)}, \end{aligned} \tag{31}$$

with $Y(z)$ implicitly defined as

$$Y(z) \triangleq B(L_1(Y(z)), L_2(z)), \quad |Y(z)| < 1 \text{ if } |z| < 1. \tag{32}$$

We note that $Y(z)$ is a pgf. As a result $Y(1) = 1$ and all derivatives of Y in 1 can be calculated from (32). The first derivative for instance is given by

$$Y'(1) = \frac{\rho_2}{1 - \rho_1}. \tag{33}$$

By putting $z_1 = z$ in (31) and by substituting z_2 by 1 and z respectively, we find

$$P_1(\chi_{1,1}(z), \chi_{1,2}(z), .., z) = (1 - \rho_1) \frac{B_1(L_1(z))(z - 1)}{z - B_1(L_1(z))}, \tag{34}$$

$$\begin{aligned} &P_T(\chi_{1,1}(z), \chi_{1,2}(z), .., \chi_{2,1}(z), \chi_{2,2}(z), .., z) \\ &\qquad = (1 - \rho_T) \frac{B(L_1(z), L_2(z))(z - 1)}{z - B(L_1(z), L_2(z))}, \end{aligned} \tag{35}$$

with P_1 and P_T defined in (15) and (16), respectively. Note that in order to obtain (34) from (31), l'Hôpital's rule has to be applied. This calculation further needs expression (33) for $Y'(1)$. Expressions (34) and (35) will be used in the next subsection and the following sections.

6.2 Calculation of moments

By substitution of $x_{1,n}$ and $x_{2,n}$ ($n \geq 1$) by 1 in expression (14), a functional equation is found for the joint pgf of the buffer contents 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 6.1. The moments of the class-1 content can be calculated from (17) and (34) by taking appropriate derivatives (for more details on this we refer to [34]). Similarly, the moments of the total buffer content are calculated from (18) and (35). 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.

7 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.

7.1 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 [29], $D_1(z)$, the pgf of the class-1 packet delay in 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:

$$D_1(z) = \frac{P_1(1,..,z) - 1 + \rho_1}{\rho_1}. \tag{36}$$

We may thus derive the moments of the class-1 packet delay from the moments of the class-1 system content. We argued in the previous section that we are able to calculate the latter. The mean class-1 packet delay $\mathrm{E}[d_1]$ is given by

$$\mathrm{E}[d_1] = D_1'(1) = 1 + \frac{\rho_1 B_1'(1) L_1''(1) + B_1''(1)(L_1'(1))^2}{2\rho_1(1-\rho_1)}. \tag{37}$$

The mean delay of a high-priority packet is thus influenced by the mean values and the second moments of the class-1 session lengths and of the number of starting sessions of class 1 in a slot.

7.2 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

$$w_2 = u_T^* - r_2 + \sum_{n=1}^{\infty} \sum_{i=1}^{e_{1,n}^*} l_{1,n,i}^+, \tag{38}$$

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 (27)). With the definition

$$Q(x_1, x_2, .., y, z) \triangleq \mathrm{E}\left[\left(\prod_{n=1}^{\infty} x_n^{e_{1,n}^*}\right) y^{r_2} z^{u_T^*}\right], \tag{39}$$

expression (38) leads to the pgf of w_2:

$$W_2(z) \triangleq \mathrm{E}[z^{w_2}] = Q(\chi_{1,1}(z), \chi_{1,2}(z), .., 1/z, z). \tag{40}$$

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. [23]), we find

$$Q(x_1, x_2, .., y, z) = \frac{P_T(x_1, x_2, .., 1, .., z) - P_T(x_1, x_2, .., y, .., z)}{\rho_2(1-y)}, \tag{41}$$

with P_T the function analyzed in section 6.

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 belonging to the virtual delay workload, say packet P, 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

$$d_2 = \sum_{i=1}^{w_2 - 1} v_{1,i} + 1, \tag{42}$$

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. and their common pgf is denoted by $V_1(z)$. By z-transforming expression (42), we then obtain

$$D_2(z) \triangleq \mathrm{E}[z^{d_2}] = \frac{z W_2(V_1(z))}{V_1(z)}. \tag{43}$$

Using (40), we find

$$D_2(z) = \frac{z Q(\chi_{1,1}(V_1(z)), \chi_{1,2}(V_1(z)), .., 1/V_1(z), V_1(z))}{V_1(z)}. \tag{44}$$

The use of (41) 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 by (18) and (35). So what remains is the calculation of the function V_1.

In order to do this, we note that the $v_{1,i}$'s in expression (42) can be expressed as

$$v_{1,i} = 1 + \sum_{m=1}^{b_{1,i}} \sum_{n=1}^{l_{1,i}^{(m)}} v_{1,i}^{(m,n)}, \tag{45}$$

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, $l_{1,i}^{(m)}$ the number of packets the m-th session of $b_{1,i}$ contains and $v_{1,i}^{(m,n)}$ 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 $v_{1,i}^{(m,n)}$'s are i.i.d. having the same pgf as the $v_{1,i}$'s, i.e., V_1. Expression (45) then leads to the following implicit expression for V_1:

$$V_1(z) = zB_1(L_1(V_1(z))). \tag{46}$$

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

Expression (44) combined with expressions (41) and (46) 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 (44) 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 (27), (46) and (18) and (35) respectively. The following final expression for the mean class-2 packet delay can then be obtained

$$\mathrm{E}[d_2] = D_2'(1) = 1 + \frac{\rho_T L_2''(1)}{2L_2'(1)(1-\rho_T)} + \frac{B_2''(1)L_2'(1)}{2B_2'(1)(1-\rho_T)} + \frac{\frac{\partial^2 B}{\partial z_1 \partial z_2}(1,1)L_1'(1)}{B_2'(1)(1-\rho_T)} + \frac{B_1'(1)L_1''(1) + B_1''(1)(L_1'(1))^2}{2(1-\rho_1)(1-\rho_T)}. \tag{47}$$

The mean low-priority packet delay is thus influenced by the mean values and the second moments of the class-1 and class-2 session lengths and of the number of starting sessions of class 1 and class 2 in a slot. It further depends on the covariance between the number of class-1 and class-2 starting sessions in a slot (through $\frac{\partial^2 B}{\partial z_1 \partial z_2}(1,1)$).

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

8 Numerical examples

Our results can be used by practitioners to estimate the (mean) delay that high- and low-priority packets sustain in a particular network node. The influence of the correlation in the arrival process on the mean delays can also be characterized.

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

$$B(z_1, z_2) = e^{\lambda_1(z_1-1)} e^{\lambda_2(z_2-1)}. \tag{48}$$

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 [5, 32]). 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

$$L_j(z) = \left(\frac{p_j z}{1-(1-p_j)z} \right)^{m_j}. \tag{49}$$

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.

In Figure 2 (Figure 3 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

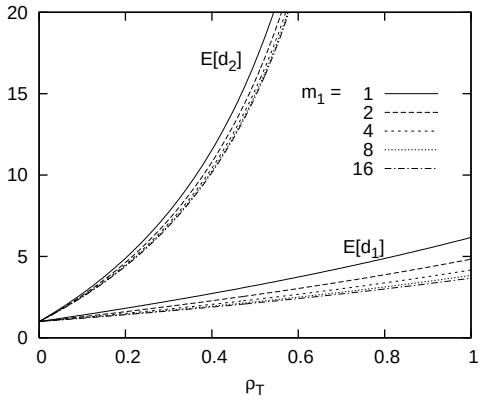


Figure 2. Mean packet delays of both classes versus the total load for $\alpha = 0.25$, $\mathrm{E}[l_1] = 16$, $\mathrm{E}[l_2] = 16$ and $m_2 = 2$. Higher m_1 means a lower variance of the session lengths of class 1.

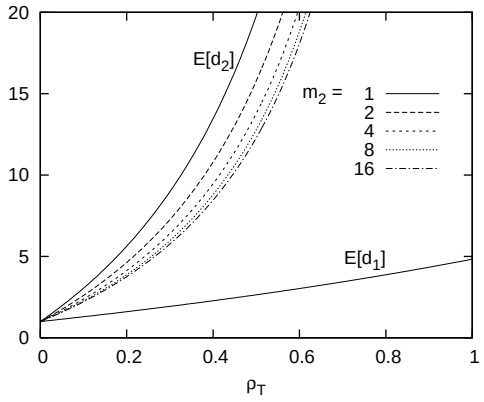


Figure 3. Mean packet delays of both classes versus the total load for $\alpha = 0.25$, $\mathrm{E}[l_1] = 16$, $\mathrm{E}[l_2] = 16$ and $m_1 = 2$. Higher m_2 means a lower variance of the session lengths of class 2.

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 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 4 and 5. In Figure 4, 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 5, 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. From both plots, it is once again concluded 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 conclude from Figure 5 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.

9 Performance of an E-commerce web server

We consider an E-commerce web server. Users request files and the web server responds by sending the requested files to the users. Two types of content are stored on the web server, content that provides revenues (class 1) and content that does not (class 2). We apply our model on the situation described in [16] and depicted in Figure 6. The web server is connected to the Internet through a gateway, which is considered the bottleneck. In the gateway, a buffer is therefore installed and packets of class 1 are transmitted, via the output channel, with priority over class-2 packets. Our analysis is used to calculate the mean delay that packets sustain in the gateway.

We use the model from this paper to analyze the performance of the web server. Therefore, we first assign values to some relevant model parameters. We assume that the output channel of the gateway has a bandwidth of 100 Mbit/s. Likewise, the packets of each session are transferred by the web server to the gateway at the rate of 100 Mbit/s. We assume further that each packet contains 100 bytes. Since it takes exactly one slot to transmit a packet, the slot length equals 8 μs. Sessions correspond to the requested files. The session length (i.e., the file sizes) distribution is taken from a real trace. The trace can be found at http://ita.ee.lbl.gov/html/contrib/EPA-HTTP.html[1], and contains the recordings of web requests of one day. We have rounded the byte sizes to the nearest multiple of 100 Bytes. The mean session size then equals 8502

[1]The logs were collected by Laura Bottomley (laurab@ee.duke.edu)

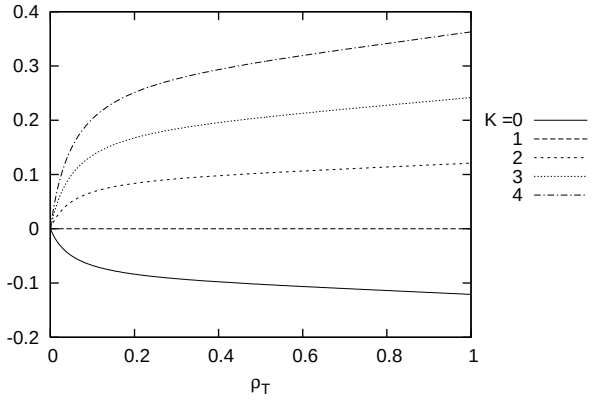


Figure 4. Relative deviation of the mean class-2 delay versus the total load for $\alpha = 0.25$, E$[l_j] = 16$, Var$[l_j] = K_j(16^2 - 16)$, $j = 1, 2$ and $K_1 = K, K_2 = 1$.

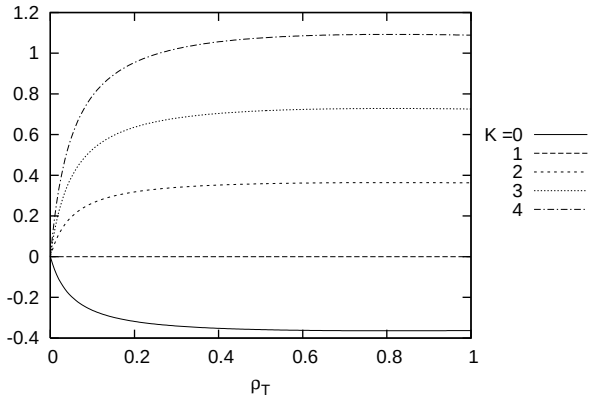


Figure 5. Relative deviation of the mean class-2 delay versus the total load for $\alpha = 0.25$, E$[l_j] = 16$, Var$[l_j] = K_j(16^2 - 16)$, $j = 1, 2$ and $K_1 = 1, K_2 = K$.

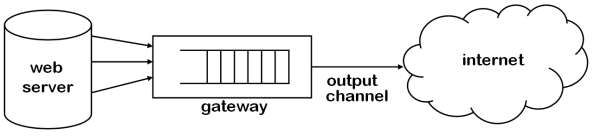


Figure 6. Conceptual scheme of a web server connected to the Internet through a gateway

	E$[d_1]$	E$[d_2]$	FIFO
$\alpha = 0.25$	9.724	16.622	14.897
$\alpha = 0.5$	11.448	18.347	14.897
$\alpha = 0.75$	13.173	20.072	14.897

Table 1. Mean class-1 and class-2 packet delays (in μs) in the E-commerce web server for some values of α. The packet delay in case of FIFO is included as reference value.

Bytes, with a variance of 5.004e9. We assume that the numbers of requests during a slot are distributed according to a Poisson process. The trace exists of 36677 (valid) requests over 24 hours, which leads to a mean number of 3.396e-6 requests per slot. Finally, we assume that each request is of class 1 with probability α and of class 2 with probability $1 - \alpha$ (independent of other requests).

In a first scenario, we assume that the file sizes of both classes have the same distribution, i.e., the distribution calculated from the trace. In Table 1, some values of the mean packet delays of both classes are given for three different values of α. As a reference value, we have also added the mean packet delay when FIFO scheduling is implemented instead of priority scheduling. In the latter case, the mean delay is independent of the class and of α, and equals about 7 μs more than the transmission time of a packet (8 μs). Thus, on average, a packet has to wait 7 μs in the gateway. It is seen that this value can be reduced to about 2 μs by giving priority to requests that provide revenues if these requests are only a small part (a quarter) of the total number of requests, and to about 5 μs if it constitutes a big part (3/4) of the requests. Of course, the price to pay is an increase of the mean low-priority packet delay, namely about 2 μs more for $\alpha = 0.25$ and more than 5 μs extra for $\alpha = 0.75$.

In a second scenario, we split the trace into two groups: group A contains all request files with size smaller than or equal to 1900 Bytes and group B consists of the request files that are larger than 1900 Bytes (1900 Bytes is the median of the request file size distribution, so groups A and B approximately exist out of the same number of requests). The request file sizes of group A have a mean of 734 Bytes and a variance of 2.453e5, while those of group B have a mean of 16369 Bytes and a variance of 9.950e9. In Table 2, we show the mean delay of both priority classes for two different cases: (a) class 1 equals group A and class 2 equals group B (small request files have priority), and (b) class 1 equals group B and class 2 equals group A (large request files have priority). We conclude that the advantage of a priority scheduling is much larger when the high-priority request files are generally small. Furthermore, giving priority is especially advantageous to packets of small request

	E[d_1]	E[d_2]
class 1 = group A	8.001	15.254
class 1 = group B	14.897	14.912

Table 2. Mean class-1 and class-2 packet delays (in μs) in the E-commerce web server for some class-dependent distributions of the packet sizes.

files: there is a difference of almost 7 μs between both cases for the mean packet delay of group-A packets, while there is only a minor difference of $.4$ μs for the mean packet delay of packets of group B.

10 Conclusion

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 of both classes can have general distributions and these distributions can be different for both classes. 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 shown the influence of the variance of the session lengths of both classes on the mean (low-priority) packet delay through numerical examples. We have finally applied our results to an E-commerce web server and showed how the performance of such a web server can be predicted by means of the results of our analysis.

Our main qualitative conclusions are: (i) give priority to only a small fraction of the requests to the web server, i.e., only to those applications that generate the largest revenues, and (ii) giving priority to applications with small request file sizes is more effective than giving priority to time-consuming applications.

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. [20, 21]; b) the extension to more than two priority classes; 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 Postdoctoral Fellow with the Research Foundation, Flanders (F.W.O.-Vlaanderen), Belgium.

References

[1] J. Walraevens, S. Wittevrongel, and H. Bruneel. Analysis of priority queues with session-based arrival streams. In *Proceedings of the Seventh International Conference on Networking (ICN 2008)*, pages 503–510, Cancun, April 2008.

[2] 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.

[3] 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.

[4] J. Chang and Y. Harn. A discrete-time priority queue with two-class customers and bulk services. *Queueing Systems*, 10:185–212, 1992.

[5] 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.

[6] 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.

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

[8] 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.

[9] T. Demoor, J. Walraevens, D. Fiems, and H. Bruneel. Mixed finite-/infinite-capacity priority queue with interclass correlation. In *Proceedings of the 15th International Conference on Analytical and Stochastic Modelling Techniques and Applications (ASMTA 2008), LNCS 5055*, pages 61–74, Nicosia, 2008.

[10] 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.

[11] D. Fiems, J. Walraevens, and H. Bruneel. Performance of a partially shared priority buffer with correlated arrivals. In *Proceedings of the 20th International Teletraffic Congress (ITC20), LNCS*, volume 4516, pages 582–593, Ottawa, 2007.

[12] R. Guérin and V. Peris. Quality-of-service in packet networks: basic mechanisms and directions. *Computer Networks*, 31(3):169–189, 1999.

[13] O. Hashida and Y. Takahashi. A discrete-time priority queue with switched batch Bernoulli process inputs and constant service time. In *Proceedings of ITC 13*, pages 521–526, Copenhagen, 1991.

[14] 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.

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

[16] L. Hoflack, S. De Vuyst, S. Wittevrongel, and H. Bruneel. Modeling web server traffic with session-based arrival streams. In *Proceedings of the 15th international conference on analytical and stochastic modelling techniques and applications (ASMTA 2008), LNCS 5055*, pages 47–60, Nicosia, 2008.

[17] 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.

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

[19] A. Khamisy and M. Sidi. Discrete-time priority queues with two-state markov modulated arrivals. *Stochastic Models*, 8(2):337–357, 1992.

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

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

[22] 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.

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

[24] J. Roberts. Internet traffic, QoS, and pricing. *Proceedings of the IEEE,* 92(9):1389–1399, 2005.

[25] S. Shakkottai and R. Srikant. Many-sources delay asymptotics with applications to priority queues. *Queueing Systems*, 39(2-3):183–2000, 2001.

[26] 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.

[27] 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.

[28] J. Van Velthoven, B. Van Houdt, and C. Blondia. The impact of buffer finiteness on the loss rate in a priority queueing system. *Lecture Notes in Computer Science*, 4054:211–225, 2006.

[29] 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.

[30] J. Walraevens, D. Fiems, and H. Bruneel. Time-dependent performance analysis of a discrete-time priority queue. *Performance Evaluation*, 65(9):641–652, 2008.

[31] 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.

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

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

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

[35] F. Xabier Albizuri, M. Graña, and B. Raducanu. Statistical transmission delay guarantee for nonreal-time traffic multiplexed with real-time traffic. *Computer Communications*, 26(12):1365–1375, 2003.

[36] 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.

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