Performance Analysis of a GI-G-1 Preemptive Resume Priority Buffer

Joris Walraevens, Bart Steyaert, and Herwig Bruneel

SMACS Research Group
Ghent University, Vakgroep TELIN (TW07V)
Sint-Pietersnieuwstraat 41, B-9000 Gent, Belgium.
Phone: 0032-9-2648902
Fax: 0032-9-2644295
{jw,bs,hb}@telin.rug.ac.be

Abstract. In this paper, we have analyzed a discrete-time $GI - G - 1$ queue with a preemptive resume priority scheduling and two priority classes. We have derived the joint generating function of the system contents of both classes and the generating functions of the delay of both classes. These pgf's are not explicitly found, but we have proven that the moments of the distributions can be found explicitly in terms of the system parameters. We have shown the impact of priority scheduling on the performance characteristics by some numerical examples.

1 Introduction

In recent years, there has been much interest devoted to incorporating multimedia applications in packet networks (e.g., IP networks). Different types of traffic need different QoS standards, but share the same network resources, such as buffers and bandwidth. For real-time applications, it is important that mean delay and delay-jitter are bounded, while for non real-time applications, the Loss Ratio (LR) is the restrictive quantity.

In general, one can distinguish two priority strategies, which will be referred to as Delay priority and Loss priority. Delay priority schemes attempt to guarantee acceptable delay boundaries to delay-sensitive traffic (such as voice/video). This can for instance be achieved by giving it HOL priority over non-delay-sensitive traffic. Several types of Delay priority (or scheduling) schemes have been proposed and analyzed, each with their own specific algorithmic and computational complexity (see e.g. [5, 8] and the references therein). On the other hand, Loss priority schemes attempt to minimize the packet loss of loss-sensitive traffic (such as data). An overview and classification of some Loss priority (or discarding) strategies can be found in [5, 3].

In this paper, we will focus on the effect of a specific Delay priority scheme, i.e., the preemptive resume priority scheduling discipline. We assume that delay-sensitive traffic has preemptive priority over delay-insensitive traffic, i.e., when the server becomes empty, a packet of delay-sensitive traffic, when available, will always be scheduled next. In the remaining, we will refer to the delay-sensitive

E. Gregori et al. (Eds.): NETWORKING 2002, LNCS 2345, pp. 745–756, 2002.

and delay-insensitive traffic as high and low priority traffic respectively. Newly arriving high priority traffic interrupts transmission of a low priority packet that has already commenced, and the interrupted low priority packet can resume its transmission when all the high priority traffic has left the system.

In the literature, there have been a number of contributions with respect to HOL priority scheduling. An overview of some basic HOL priority queueing models can be found in Kleinrock [4], Miller [7] and Takagi [9] and the references therein. Preemptive resume priority queues have been analyzed in Machihara [6], Takine et al. [10] and Walraevens et al. [11]. Machihara [6] analyzes waiting times when high priority arrivals are distributed according to a MAP process. Takine [10] studies the waiting times of customers arriving to a queue according to independent MAP processes. Finally, Walraevens [11] analyzes system contents and packet delay when the length of high priority packets are generally distributed and the length of low priority packets are geometrically distributed.

In this paper, we analyze the system contents and packet delay of high priority and low priority traffic in a discrete-time single-server buffer for a preemptive resume priority scheme and per-slot i.i.d. arrivals. The transmission times of the packets are assumed to be generally distributed. These distribution can be class-dependent, i.e., the transmission times of the high priority packets can be different from those of the low priority packets. We will demonstrate that an analysis based on generating functions is extremely suitable for modelling this type of buffers with a priority scheduling discipline. From these generating functions, expressions for some interesting performance measures (such as moments of system contents and packet delay of both classes) can be calculated.

The remainder of this paper is structured as follows. In the following section, we present the mathematical model. In sections 3 and 4, we will then analyze the steady-state system contents and packet delay of both classes. In section 5, we give expressions for some moments of the system contents and packet delay of both classes. Some numerical examples are treated in section 6. Finally, some conclusions are formulated in section 7.

2 Mathematical Model

We consider a discrete-time single-server system with infinite buffer space. Time is assumed to be slotted. There are two types of packets arriving to the system, namely packets of class 1 and packets of class 2. The number of arrivals of class j during slot k are i.i.d. and are denoted by $a_{j,k}$ $(j = 1, 2)$. Their joint probability mass distribution is defined as $a(m, n) \triangleq \text{Prob}[a_{1,k} = m, a_{2,k} = n]$. Note that the number of arrivals of both classes can be correlated during one slot. The joint probability generating function (pgf) of $a_{1,k}$ and $a_{2,k}$ is defined as $A(z_1, z_2) \triangleq E[z_1^{a_{1,k}} z_2^{a_{2,k}}] = \sum_{m=0}^{\infty} \sum_{n=0}^{\infty} a(m, n) z_1^m z_2^n$. The marginal pgf's of the number of arrivals of class j are denoted by $A_j(z)$ $(j = 1, 2)$ and are given by $A(z, 1)$ and $A(1, z)$ respectively. We will furthermore denote the mean arrival rate of class j packets during a slot by $\lambda_j \triangleq E[a_{j,k}] = A'_j(1)$ $(j = 1, 2)$.

The service times of the class j packets, i.e., the number of slots a class j packet is effectively being served, are i.i.d. and generally distributed and their pgf is denoted by $S_j(z)$ $(j = 1, 2)$. The mean service time of a class j packet is given by μ_j $(j = 1, 2)$.

The class 1 packets are assumed to have preemptive resume priority over the class 2 packets and within one class the scheduling is FCFS. The load offered by class j packets is given by $\rho_j \triangleq \lambda_j \mu_j$. The total load is then given by $\rho_T \triangleq \rho_1 + \rho_2$. We assume a stable system, i.e., $\rho_T < 1$.

3 System Contents

We denote the system contents of class j packets at the beginning of slot k by $u_{j,k}$ $(j = 1, 2)$. Their joint pgf is defined as $U_k(z_1, z_2) \triangleq E\left[z_1^{u_{1,k}} z_2^{u_{2,k}}\right]$. Since service times of both classes are generally distributed, the set $\{u_{1,k}, u_{2,k}\}$ does not form a Markov chain. Therefore, we introduce two new stochastic variables $r_{j,k}$ $(j = 1, 2)$ as follows: $r_{1,k}$ indicates the remaining number of slots needed to transmit the class 1 packet in service at the beginning of slot k, if $u_{1,k} > 0$, and $r_{1,k} = 0$ if $u_{1,k} = 0$; $r_{2,k}$ indicates the remaining number of slots service time of the class 2 packet longest in the system at the beginning of slot k, if $u_{2,k} > 0$, and $r_{2,k} = 0$ if $u_{2,k} = 0$. With these definitions, $\{r_{1,k}, u_{1,k}, r_{2,k}, u_{2,k}\}$ is easily seen to constitute a Markovian state description of the system at the beginning of slot k. If $s^*_{j,k}$ $(j = 1, 2)$ indicates the service time of the next class j packet to receive service at the beginning of slot k, the following system equations can be established:

1. If $r_{1,k} = 0$ (and hence $u_{1,k} = 0$):
   a) If $r_{2,k} = 0$ (and hence $u_{2,k} = 0$):

$$u_{j,k+1} = a_{j,k} \; ; \; r_{j,k+1} = \begin{cases} 0 & \text{if } a_{j,k} = 0 \\ s^*_{j,k} & \text{if } a_{j,k} > 0 \end{cases},$$

   with $j = 1, 2$. The only packets present in the system at the beginning of slot $k+1$ are the packets that arrive during the previous slot. If there have been new arrivals of class j packets during slot k, the remaining number of slots needed to service the first class j packet is that packet's full service time.
   b) If $r_{2,k} = 1$:

$$u_{1,k+1} = a_{1,k} \quad ; \quad u_{2,k+1} = u_{2,k} - 1 + a_{2,k};$$
$$r_{1,k+1} = \begin{cases} 0 & \text{if } a_{1,k} = 0 \\ s^*_{1,k} & \text{if } a_{1,k} > 0 \end{cases} ; \; r_{2,k+1} = \begin{cases} 0 & \text{if } u_{2,k} - 1 + a_{2,k} = 0 \\ s^*_{2,k} & \text{if } u_{2,k} - 1 + a_{2,k} > 0 \end{cases},$$

   i.e., the class 2 packet in service at the beginning of slot k leaves the system at the end of slot k.

c) If $r_{2,k} > 1$:

$$u_{1,k+1} = a_{1,k} \quad ; \quad u_{2,k+1} = u_{2,k} + a_{2,k};$$
$$r_{1,k+1} = \begin{cases} 0 & \text{if } a_{1,k} = 0 \\ s^*_{1,k} & \text{if } a_{1,k} > 0 \end{cases} ; \quad r_{2,k+1} = r_{2,k} - 1,$$

i.e., the class 2 packet in service at the beginning of slot k remains in the system (not necessarily in the server - because of the preemptive priority scheduling discipline, it can only remain in the server if there are no new class 1 arrivals). Its remaining service time is decreased by one.

2. If $r_{1,k} = 1$:

a) If $r_{2,k} = 0$ (and hence $u_{2,k} = 0$):

$$u_{1,k+1} = u_{1,k} - 1 + a_{1,k} \quad ; \quad u_{2,k+1} = a_{2,k};$$
$$r_{1,k+1} = \begin{cases} 0 & \text{if } u_{1,k} - 1 + a_{1,k} = 0 \\ s^*_{1,k} & \text{if } u_{1,k} - 1 + a_{1,k} > 0 \end{cases} ; \quad r_{2,k+1} = \begin{cases} 0 & \text{if } a_{2,k} = 0 \\ s^*_{2,k} & \text{if } a_{2,k} > 0 \end{cases},$$

i.e., the class 1 packet in service at the beginning of slot k, leaves the system at the end of slot k. There were no class 2 packets in the system at the beginning of slot k.

b) If $r_{2,k} > 0$:

$$u_{1,k+1} = u_{1,k} - 1 + a_{1,k} \quad ; \quad u_{2,k+1} = u_{2,k} + a_{2,k};$$
$$r_{1,k+1} = \begin{cases} 0 & \text{if } u_{1,k} - 1 + a_{1,k} = 0 \\ s^*_{1,k} & \text{if } u_{1,k} - 1 + a_{1,k} > 0 \end{cases} ; \quad r_{2,k+1} = r_{2,k},$$

i.e., the class 1 packet in service at the beginning of slot k, leaves the system at the end of slot k. The remaining service of the class 2 packet longest in the system stays the same.

3. If $r_{1,k} > 1$:

a) If $r_{2,k} = 0$ (and hence $u_{2,k} = 0$):

$$u_{1,k+1} = u_{1,k} + a_{1,k} \; ; \quad u_{2,k+1} = a_{2,k};$$
$$r_{1,k+1} = r_{1,k} - 1 \quad ; \quad r_{2,k+1} = \begin{cases} 0 & \text{if } a_{2,k} = 0 \\ s^*_{2,k} & \text{if } a_{2,k} > 0 \end{cases},$$

i.e., the class 1 packet in service at the beginning of slot k stays in the server at the beginning of slot $k+1$. Its remaining service is decreased by one.

b) If $r_{2,k} > 0$:

$$u_{j,k+1} = u_{j,k} + a_{j,k};$$
$$r_{1,k+1} = r_{1,k} - 1 \quad ; \quad r_{2,k+1} = r_{2,k},$$

with $j = 1, 2$. The difference with the previous case is that there was at least one class 2 packet in the system at the beginning of slot k.

We define $\mathrm{E}[X\{Y\}]$ as $\mathrm{E}[X|Y]\mathrm{Prob}[Y]$ in the remainder. We furthermore define $P_k(x_1, z_1, x_2, z_2)$ as the joint pgf of the state vector $(r_{1,k}, u_{1,k}, r_{2,k}, u_{2,k})$, i.e., $P_k(x_1, z_1, x_2, z_2) \triangleq E[x_1^{r_{1,k}} z_1^{u_{1,k}} x_2^{r_{2,k}} z_2^{u_{2,k}}]$. We assume that the system is stable (implying that the equilibrium condition requires that $\rho_T < 1$) and as a result $P_k(x_1, z_1, x_2, z_2)$ and $P_{k+1}(x_1, z_1, x_2, z_2)$ converge both to a common steady state value $P(x_1, z_1, x_2, z_2) = \lim_{k\to\infty} P_k(x_1, z_1, x_2, z_2)$. Using the system equations, we can constitute a relation between $P_k(x_1, z_1, x_2, z_2)$ and $P_{k+1}(x_1, z_1, x_2, z_2)$. By taking the $k \to \infty$ limit in this relation between $P_k(x_1, z_1, x_2, z_2)$ and $P_{k+1}(x_1, z_1, x_2, z_2)$ we obtain:

$$
\begin{aligned}
&[x_1 - A(z_1, z_2)]P(x_1, z_1, x_2, z_2) \\
&= \Bigg[x_1 A(0,0)(1 - S_1(x_1))(1 - S_2(x_2)) + \frac{x_1}{x_2} A(0, z_2)(1 - S_1(x_1))(x_2 S_2(x_2) - 1) \\
&\quad + A(z_1, 0)(x_1 S_1(x_1) - 1)(1 - S_2(x_2)) \\
&\quad + \frac{1}{x_2} A(z_1, z_2)(x_1 x_2 S_1(x_1) S_2(x_2) - x_1 S_1(x_1) - x_2 S_2(x_2) + x_2)\Bigg] P(0,0,0,0) \\
&\quad + x_1[A(0,0)(1 - S_1(x_1)) + A(z_1, 0) S_1(x_1)](1 - S_2(x_2)) R_2(0) \\
&\quad + x_1(A(0, z_2) - A(0,0))(1 - S_1(x_1))(S_2(x_2) - 1) R_1(0,0,0) \\
&\quad + (A(z_1, z_2) - A(z_1, 0))(S_2(x_2) - 1) P(x_1, z_1, 0, 0) \\
&\quad + x_1(A(z_1, z_2) - A(z_1, 0))(z_1 - S_1(x_1))(1 - S_2(x_2)) R_1(z_1, 0, 0) \\
&\quad + \frac{1}{x_2}[x_1 A(0, z_2)(1 - S_1(x_1)) + A(z_1, z_2)(x_1 S_1(x_1) - x_2)] P(0, 0, x_2, z_2) \\
&\quad + x_1[A(0, z_2)(1 - S_1(x_1)) + A(z_1, z_2) S_1(x_1)](S_2(x_2) - z_2) R_2(z_2) \\
&\quad + x_1 A(0, z_2)(1 - S_1(x_1)) R_1(0, x_2, z_2) + x_1 A(z_1, z_2)(S_1(x_1) - z_1) R_1(z_1, x_2, z_2)
\end{aligned}
$$

with functions $R_1(z_1, x_2, z_2) \triangleq \lim_{k\to\infty} \mathrm{E}\left[z_1^{u_{1,k}-1} x_2^{r_{2,k}} z_2^{u_{2,k}} \{r_{1,k} = 1\}\right]$ and $R_2(z_2) \triangleq \lim_{k\to\infty} \mathrm{E}\left[z_2^{u_{2,k}-1} \{r_{1,k} = u_{1,k} = 0, r_{2,k} = 1\}\right]$. It now remains for us to determine the functions $P(x_1, z_1, 0, 0)$, $P(0, 0, x_2, z_2)$, $R_2(z_2)$, $R_1(z_1, x_2, z_2)$ and the unknown parameters $P(0,0,0,0)$, $R_2(0)$ and $R_1(0,0,0)$. Using generating functions techniques (a.o. Rouché's theorem), we can ultimately calculate a fully determined version for $P(x_1, z_1, x_2, z_2)$ (calculations are omitted due to page limitations):

$$
\begin{aligned}
&P(x_1, z_1, x_2, z_2) = (1 - \rho_T) \\
&\Bigg[1 + \frac{x_1 z_1 (A(z_1, 0) - A(Y(0), 0))(S_1(x_1) - S_1(A(z_1, 0)))(1 - S_2(x_2))}{A(Y(0), 0)(x_1 - A(z_1, 0))(z_1 - S_1(A(z_1, 0)))(z_1 - S_1(A(z_1, z_2)))} \\
&+ x_1 z_1 \frac{(A(z_1, z_2) - A(Y(z_2), z_2))(S_1(x_1) - S_1(A(z_1, z_2)))}{(x_1 - A(z_1, z_2))(z_1 - S_1(A(z_1, z_2)))(z_2 - S_2(A(Y(z_2), z_2)))} \\
&\left\{\frac{S_2(A(Y(z_2), z_2))(z_2 - S_2(x_2))}{A(Y(z_2), z_2)} - z_2 \frac{(1 - x_2)(S_2(x_2) - S_2(A(Y(z_2), z_2)))}{x_2 - A(Y(z_2) z_2)}\right\} \\
&- x_2 z_2 \frac{(1 - A(Y(z_2), z_2))(S_2(x_2) - S_2(A(Y(z_2), z_2)))}{(x_2 - A(Y(z_2), z_2))(z_2 - S_2(A(Y(z_2), z_2)))}\Bigg], \qquad (1)
\end{aligned}
$$

with $Y(z)$ implicitly defined as $Y(z) \triangleq S_1(A(Y(z), z))$. From this pgf, several joint and marginal pgf's can be calculated. We can for instance calculate the joint pgf of the system contents of class j packets and the remaining service of the class j packet with the longest waiting time at the beginning of an arbitrary slot in steady-state defined as follows $P_j(x, z) \triangleq \lim_{k\to\infty} \mathrm{E}\left[x^{r_{j,k}} z^{u_{j,k}}\right]$, $j = 1, 2$. $P_1(x_1, z_1)$ ($P_2(x_2, z_2)$ respectively) can then be found from equation (1) by substituting x_2 and z_2 (x_1 and z_1 respectively) by 1. More importantly, we can calculate the joint pgf of the steady-state system contents of class 1 and class 2 packets from equation (1). It is given by:

$$\begin{aligned} U(z_1, z_2) &\triangleq \lim_{k\to\infty} \mathrm{E}\left[z_1^{u_{1,k}} z_2^{u_{2,k}}\right] = P(1, z_1, 1, z_2) \\ &= (1-\rho_T)\frac{S_2(A(Y(z_2), z_2))(z_2-1)}{z_2 - S_2(A(Y(z_2), z_2))} \\ &\quad \left[1 + z_1 \frac{(A(z_1, z_2) - A(Y(z_2), z_2))(S_1(A(z_1, z_2)) - 1)}{A(Y(z_2), z_2)(A(z_1, z_2) - 1)(z_1 - S_1(A(z_1, z_2)))}\right]. \end{aligned} \tag{2}$$

From the two-dimensional pgf $U(z_1, z_2)$, we can easily derive expressions for the pgf of the system contents of class 1 packets and class 2 packets at the beginning of an arbitrary slot from expression (2), yielding

$$\begin{aligned} U_1(z) &\triangleq \lim_{k\to\infty} \mathrm{E}\left[z^{u_{1,k}}\right] = U(z, 1) \\ &= (1-\rho_1)\frac{S_1(A_1(z))(z-1)}{z - S_1(A_1(z))}; \end{aligned} \tag{3}$$

$$\begin{aligned} U_2(z) &\triangleq \lim_{k\to\infty} \mathrm{E}\left[z^{u_{2,k}}\right] = U(1, z) \\ &= (1-\rho_T)\frac{A_2(z)}{A(Y(z), z)} \frac{S_2(A(Y(z), z)(z-1)}{z - S_2(A(Y(z), z))} \frac{1 - A(Y(z), z)}{1 - A_2(z)}. \end{aligned} \tag{4}$$

4 Packet Delay

The packet delay is defined as the total amount of time a packet spends in the system, more precisely, the number of slots between the end of the packet's arrival slot and the end of its departure slot. We can analyze the packet delay of class 1 packets as if they are the only packets in the system. This is e.g. done in [1] and the pgf of the packet delay of class 1 packets is given by

$$D_1(z) = \frac{1-\rho_1}{\lambda_1} \frac{S_1(z)(z-1)}{z - A_1(S_1(z))} \frac{1 - A_1(S_1(z))}{1 - S_1(z)}. \tag{5}$$

Because of the priority discipline, the analysis of the delay of the low priority class will be a bit more involved. We tag a class 2 packet that enters the buffer during slot k. Let us refer to the packets in the system at the end of slot k, but that have to be served before the tagged packet as the "primary packets". So, basically, the tagged class 2 packet can enter the server, when all primary

packets and all class 1 packets that arrived after slot k are transmitted. In order to analyse the delay of the tagged class 2 packet, the number of class 1 packets and class 2 packets that are served between the arrival slot of the tagged class 2 packet and its departure slot is important, not the precise order in which they are served. Therefore, in order to facilitate the analysis, we will consider an equivalent virtual system with an altered service discipline. We assume that from slot k on, the order of service for class 1 packets (those in the queue at the end of slot k and newly arriving ones) is LCFS instead of FCFS in the equivalent system (the transmission of class 2 packets remains FCFS). So, a primary packet can enter the server, when the system becomes free (for the first time) of class 1 packets that arrived during and after the service time of the primary packet that predecessed it according to the new service discipline. Let $v_{1,m}^{(i)}$ denote the length of the time period during which the server is occupied by the m-th class 1 packet that arrives during slot i and its class 1 "successors", i.e., the time period starting at the beginning of the service of that packet and terminating when the system becomes free (for the first time) of class 1 packets which arrived during and after its service time. Analogously, let $v_{2,m}^{(i)}$ denote the length of the time period during which the server is occupied by the m-th class 2 packet that arrives during slot i and its class 1 "successors". The $v_{j,m}^{(i)}$'s ($j = 1, 2$) are called sub-busy periods, caused by the m-th class j packet that arrived during slot i. The service time of the tagged class 2 packet is denoted by s_2^*.

When the tagged class 2 packet arrives, the system is in one of the following states:

1. $r_{1,k} = 0$ (and hence $u_{1,k} = 0$):
   a) $r_{2,k} = 0$ (and hence $u_{2,k} = 0$):

$$d_2 = \sum_{j=1}^{2} \sum_{m=1}^{f_{j,k}} v_{j,m}^{(k)} + s_2^* + \sum_{i=1}^{s_2^*-1} \sum_{m=1}^{a_{1,l_i}} v_{j,m}^{(l_i)}, \tag{6}$$

   with $f_{j,k}$ defined as the number of class j packets arriving during slot k, but that have to be served before the tagged packet. Slots l_i are defined as the slots during which the tagged packet receives service ($i = 1, .., s_2^*$). $f_{1,k}$ class 1 primary packets and $f_{2,k}$ class 2 primary packets that arrived during slot k and their class 1 successors have to be served before the tagged class 2 packet. During the service time of the tagged class 2 packet, new class 1 packets may arrive, which interrupt the tagged packet's service. The last two terms take this part of the delay into account.

   b) $r_{2,k} > 0$:

$$d_2 = (r_{2,k} - 1) + \sum_{i=1}^{r_{2,k}-1} \sum_{m=1}^{a_{1,n_i}} v_{1,m}^{(n_i)} + \sum_{j=1}^{2} \sum_{m=1}^{f_{j,k}} v_{j,m}^{(k)} + \sum_{m=1}^{u_{2,k}-1} \tilde{v}_{2,m} \tag{7}$$
$$+ s_2^* + \sum_{i=1}^{s_2^*-1} \sum_{m=1}^{a_{1,l_i}} v_{j,m}^{(l_i)},$$

with the n_i-th slots ($i = 1, .., r_{2,k}-1$) the slots (after slot k) that the class 2 packet longest in the server receives service and the $\tilde{v}_{2,m}$'s are defined as the sub-busy periods, caused by the m-th class 2 packet already in the queue at the beginning of start slot l. The residual service time of the packet in service during slot k contributes in the first term, the sub-busy periods of the class 1 packets arriving during the residual service time contribute in the second term, the sub-busy periods of the class 1 and class 2 packets arriving during slot k, but that have to be served before the tagged class 2 packet contribute in the third term, the sub-busy periods of the class 2 packets already in the queue at the beginning of slot k contribute in the fourth term and finally the service time of the tagged class 2 packet itself and the sub-busy periods of the class 1 packets arriving during this service time (except for its last slot) contribute in the last two terms.

2. $r_{1,k} > 0$:
   a) $r_{2,k} = 0$ (and hence $u_{2,k} = 0$):

$$d_2 = (r_{1,k} - 1) + \sum_{i=1}^{r_{1,k}-1} \sum_{m=1}^{a_{1,k+i}} v_{1,m}^{(k+i)} + \sum_{j=1}^{2} \sum_{m=1}^{f_{j,k}} v_{j,m}^{(k)} + \sum_{m=1}^{u_{1,k}-1} \tilde{v}_{1,m} \quad (8)$$
$$+ s_2^* + \sum_{i=1}^{s_2^*-1} \sum_{m=1}^{a_{1,l_i}} v_{j,m}^{(l_i)},$$

   with the $\tilde{v}_{1,m}$'s the sub-busy periods, caused by the m-th class 1 packet already in the queue at the beginning of slot k. The expression is almost the same as in the previous case, with the difference that a class 1 packet was being served during slot k.
   b) $r_{2,k} > 0$:

$$d_2 = (r_{1,k} - 1) + \sum_{i=1}^{r_{1,k}-1} \sum_{m=1}^{a_{1,k+i}} v_{1,m}^{(k+i)} + \sum_{j=1}^{2} \sum_{m=1}^{f_{j,k}} v_{j,m}^{(k)} + \sum_{m=1}^{u_{1,k}-1} \tilde{v}_{1,m} \quad (9)$$
$$+ r_{2,k} + \sum_{i=1}^{r_{2,k}} \sum_{m=1}^{a_{1,n_i}} v_{1,m}^{(n_i)} + \sum_{m=1}^{u_{2,l}-1} \tilde{v}_{2,m} + s_2^* + \sum_{i=1}^{s_2^*} \sum_{m=1}^{a_{1,l_i}} v_{j,m}^{(l_i)}.$$

   This case is a combination of the former two cases.

Due to the initial assumptions and since the length of different sub-busy periods only depends on the number of class 1 packet arrivals during different slots and the service times of the corresponding primary packets, the sub-busy periods associated with the primary packets of class 1 and class 2 form a set of i.i.d. random variables and their pgf will be presented by $V_1(z)$ and $V_2(z)$ respectively. Notice that $f_{1,k}$ and $f_{2,k}$ are correlated; in section 2 it was explained that $a_{1,k}$ and $a_{2,k}$ may be correlated as well. Once again, applying a z-transform technique to equations (6)-(9) and taking into account the previous remarks, we can ultimately derive an expression for $D_2(z)$:

$$D_2(z) = \frac{1-\rho_T}{\lambda_2} \frac{S_2(z)(A(V_1(z),V_2(z)) - A_1(V_1(z)))}{zA_1(V_1(z)) - A(V_1(z),V_2(z))} \frac{1-zA_1(V_1(z))}{1-V_2(z)}. \quad (10)$$

Finally, we have to find expressions for $V_1(z)$ and $V_2(z)$. These pgf's satisfy the following relations:

$$V_j(z) = S_j(zA_1(V_1(z))), \quad (11)$$

with $j = 1, 2$. This can be understood as follows: when the m-th class j packet that arrived during slot i enters service, $v_{j,m}^{(i)}$ consists of two parts: the service time of that packet itself, and the service times of the class 1 packets that arrive during its service time and of their class 1 successors. This leads to equation (11).

5 Calculation of Moments

The functions $Y(z)$, $V_1(z)$ and $V_2(z)$ can only be explicitly found in case of some simple arrival processes. Their derivatives for $z = 1$, necessary to calculate the moments of the system contents and the packet delay, on the contrary, can be calculated in closed-form. Let us define λ_{ij} and μ_{jj} as $\lambda_{ij} \triangleq \left.\frac{\partial^2 A(z_1,z_2)}{\partial z_i \partial z_j}\right|_{z_1=z_2=1}$ and $\mu_{jj} \triangleq \left.\frac{d^2 S_j(z)}{dz^2}\right|_{z=1}$, with $i, j = 1, 2$. Now we can calculate the mean system contents and the mean packet delay of both classes by taking the first derivatives of the respective pgf's for $z = 1$. We find

$$E[u_1] = \rho_1 + \frac{\lambda_{11}\mu_1 + \lambda_1^2\mu_{11}}{2(1-\rho_1)}, \quad (12)$$

for the mean system contents of class 1 packets and

$$E[u_2] = \rho_2 + \frac{\rho_1\lambda_2(\mu_2-1)}{1-\rho_1} + \frac{\lambda_{22}\mu_2}{2(1-\rho_T)} + \frac{\lambda_2^2\mu_{22}}{2(1-\rho_T)(1-\rho_1)} + \frac{\lambda_{12}\mu_1}{1-\rho_T} + \frac{\lambda_2(\lambda_{11}\mu_1^2 + \lambda_1\mu_{11})}{2(1-\rho_T)(1-\rho_1)}, \quad (13)$$

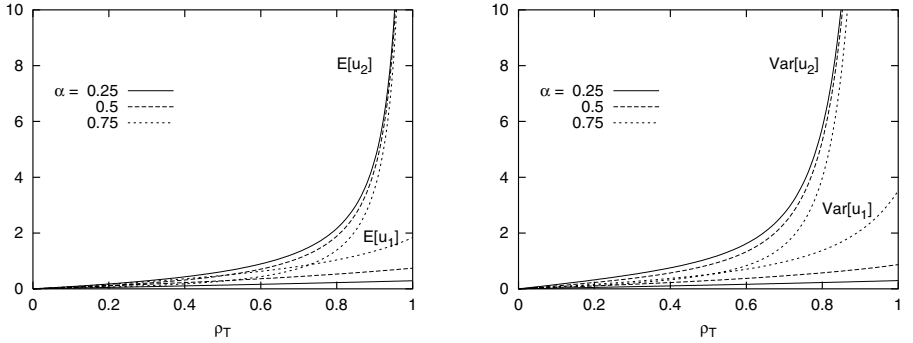
for the mean system contents of class 2 packets. The mean delay of both classes can also be found by taking the first derivatives of the respective pgf's for $z = 1$, and are given by $E[d_j] = E[u_j]/\lambda_j$. So, as expected, Little's law is satisfied.

In a similar way, expressions for the variance (and higher moments) can be calculated by taking the appropriate derivatives of the respective generating functions as well. These are nevertheless too elaborate to express them, but figures of the variance of system contents and packet delay of both classes will be shown in the next section.

6 Numerical Examples

In this section, we present some numerical examples. We assume the traffic of the two classes to be arriving according to a two-dimensional binomial process. Its two-dimensional pgf is given by $A(z_1, z_2) = (1 - \lambda_1(1 - z_1)/N - \lambda_2(1 - z_2)/N)^N$. The arrival rate of class j traffic is thus given by λ_j $(j = 1, 2)$. This arrival process occurs for instance at an output queue of a NxN output queueing switch/router fed by a Bernoulli process at the inlets. Notice also that if $N \to \infty$, the arrival process becomes a superposition of two independent Poisson streams. In the remainder of this section, we assume that $N = 16$. We furthermore denote the fraction of the high priority load in the total load by α, i.e., $\alpha = \rho_1/\rho_T$.

In Figure 1, the mean and variance of the system contents of class 1 and class 2 packets is shown as a function of the total load ρ_T, when service times of class 1 and class 2 packets are deterministically equal to 2 ($\mu_1 = \mu_2 = 2$) and α is 0.25, 0.5 and 0.75 respectively. We clearly see the influence of the priority scheduling. The mean and variance of the system contents of class 1 packets remains low, even if the fraction of class 1 packets is high. The mean value and variance of the system contents of class 2 packets on the other hand is large, especially when the system is heavily loaded.

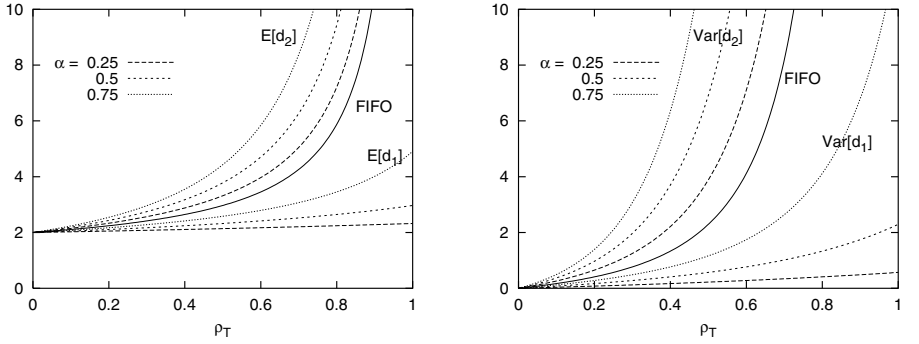


a. mean b. variance

Fig. 1. Mean and variance of the system contents versus the total load

In Figure 2, the mean value and variance of the packet delay of class 1 and class 2 packets is shown as a function of the total load ρ_T, when service times of both classes are deterministically equal to 2, i.e., $\mu_j = 2$ $(j = 1, 2)$ and α is, as before, 0.25, 0.5 and 0.75 respectively. In order to compare with FIFO scheduling, we have also shown the mean value and variance of the packet delay in that case. Since, in this example, the service times of the class 1 and class 2 packets are equally distributed, the packet delay is then of course the same for class 1 and class 2 packets, and can thus be calculated as if there is only one class of packets arriving according to an arrival process with pgf $A(z, z)$.

This has already been analyzed, e.g., in [2]. The influence of priority scheduling on the packet delay becomes obvious from these figures: mean and variance of the delay of class 1 packets reduces significantly. The price to pay is of course a larger mean value and variance of the delay of class 2 packets. If this kind of traffic is not delay-sensitive, as assumed, this is not a too big a problem. Also, the smaller the fraction of high priority load in the overall traffic mix, the lower the mean and variance of the packet delay of both classes will be.



a. mean b. variance

Fig. 2. Mean and variance of the packet delay versus the total load

Finally, Figure 3a. (Figure 3b. respectively) shows the mean delay of high and low priority packets when service time of the packets are deterministic, as a function of the mean service time of the low priority packets (high priority packets respectively), i.e., μ_2 (μ_1 respectively), when $\mu_1 = 2$ ($\mu_2 = 2$ respectively) and $\rho_T = 0.75$. α is, as before, 0.25, 0.5 and 0.75. The figures show that the mean packet delay of high-priority packets is not influenced by the mean service time of class 2 packets, while it is proportionally increasing with the mean service time of class 1 packets (when the load of high and low priority packets is kept constant). The mean packet delay of class 2 packets on the other hand is proportionally increasing with the mean service time of class 2 packets and with the mean service time of class 1 packets. Because of the preemptive priority scheduling discipline, mean delay of high priority packets is only influenced by its own arrival and service process, while the mean delay of low priority packets is influenced by the arrival and service processes of both classes.

7 Conclusion

In this paper, we have analyzed a discrete-time $GI - G - 1$ queue with a preemptive resume priority scheduling and two priority classes. We have derived the joint generating function of the system contents of both classes and the generating functions of the delay of both classes. These pgf's are not explicitly found,

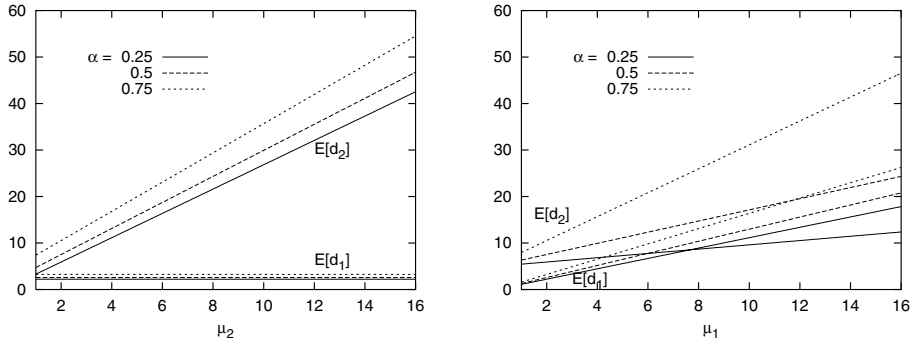


Fig. 3. Mean packet delay versus the mean service time of class 2 and class 1 packets

but we have proven that the moments of the distributions can be found explicitly in terms of the system parameters. We have shown the impact of priority scheduling on the performance characteristics by some numerical examples.

References

[1] H. Bruneel and B.G. Kim, *Discrete-time models for communication systems including ATM*, Kluwer Academic Publishers, Boston, 1993.

[2] H. Bruneel, *Performance of discrete-time queueing systems*, Computers and Operations Research, pp. 303-320, 1993.

[3] I. Cidon and R. Guérin, *On protective buffer policies*, Proceedings of Infocom '93 (San Francisco), pp. 1051-1058, 1993.

[4] L. Kleinrock, *Queueing systems volume II: Computer applications*, John Wiley & Sons, 1976.

[5] K. Liu, D.W. Petr, V.S. Frost, H. Zhu, C. Braun and W.L. Edwards, *Design and analysis of a bandwidth management framework for ATM-based broadband ISDN*, IEEE Communications Magazine, pp. 138-145, 1997.

[6] F. Machihara, *A bridge between preemptive and non-preemptive queueing models*, Performance Evaluation 23, pp. 93-106, 1995.

[7] R.G. Miller, *Priority queues*, Annals of Matematical. Statistics, pp. 86-103, 1960.

[8] S.P. Morgan, *Queueing disciplines and passive congestion control in byte-stream networks*, IEEE Transactions on Communications 39(7), pp. 1097-1106, 1991.

[9] H. Takagi, *Queueing analysis A foundation of Performance Evaluation Volume 1: Vacation and priority systems*, North-Holland, 1991.

[10] T. Takine and T. Hasegawa, *The workload in the MAP/G/1 queue with state-dependent services: its application to a queue with preemptive resume priority*, Commun. Statist.-Stochastic Models 10(1), pp. 183-204, 1994.

[11] J. Walraevens, B. Steyaert and H. Bruneel, Analysis of a preemptive resume priority buffer with general service times for the high priority class, Proceedings of the Africom 2001 Conference, Cape Town, May 28-30, 2001.