Analysis of Packet Delay in a GI-G-1 Queue with Non-preemptive Priority Scheduling

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
E-mail: {jw,bs,hb}@telin.rug.ac.be

Abstract. Priority scheduling for packets is becoming a hot topic, as attemps are being made to integrate voice services in existing IP data networks. In this paper, we consider a discrete-time queueing system with head-of-line (HOL) non-preemptive priority scheduling. Two classes of traffic will be considered, i.e., high priority and low priority traffic, which both generate variable-length packets. We will derive expressions for the Probability Generating Function (pgf) of the packet delay of the high priority traffic and the low priority traffic. From these, some performance measures (such as the mean value) will be derived. These will be used to illustrate the significance of priority scheduling and the effect of non-preemptive scheduling on the high priority traffic.

1 Introduction

In recent years, there has been much interest devoted to incorporating multimedia applications in IP networks. Different types of traffic need different QoS standards. 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 Time Priority and Space Priority. Time priority schemes attempt to guarantee acceptable delay boundaries to delay-sensitive traffic (such as voice/video). This is achieved by giving it HOL priority over non-delay-sensitive traffic, and/or by sharing access to the server among the various traffic classes in such a way so that each can meet its own specific delay requirements. Several types of Time priority (or scheduling) schemes (such as Weighted-Round-Robin (WRR), Weighted-Fair-Queueing(WFQ)) have been proposed and analyzed, each with their own specific algorithmic and computational complexity (see e.g. [6] and the references therein). On the other hand, Space Priority schemes attempt to minimize the packet loss of loss-sensitive traffic (such as data). Again, various types of Space Priority (or discarding) strategies (such as Push-Out Buffer (POB), Partial Buffer Sharing (PBS)) have been presented in the literature (see

e.g. [15]), mainly in the context of ATM buffers. An overview of both types of priority schemes can be found in [1].

In the existing 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 Jaiswal [3], Takacs [10] and Takagi [11] and the references therein. Khamisy et al. [4], Laevens et al. [5], Takine et al. [13] and Walraevens et al. [16] have studied discrete-time HOL priority queues with deterministic service times equal to one slot. Furthermore, non-preemptive HOL priority queues have been considered by Rubin et al. [7], Stanford [8], Sugahara et al. [9] and Takine et al. [12, 14]. Rubin [7] studies the mean waiting time, for a queue fed by an i.i.d. arrival process. Stanford [8] analyses the interdeparture time distribution in a queue fed by a Poisson process. In Sugahara [9], a non-preemptive queue in continuous time is presented, with a Switched Poisson Process arrival process for the high priority packets. Finally, Takine [12, 14] studies a discrete-time MAP/G/1 queue, using matrix-analytic techniques.

In this paper, we analyse the packet delay of high and low priority traffic in a discrete-time single-server buffer for a non-preemptive HOL priority scheme and per-slot i.i.d. arrivals. The transmission times of the packets generated by both types are assumed to be generally distributed. We will demonstrate that an analysis based on generating functions is extremely suitable for modelling this type of buffers with priority scheduling.

2 Mathematical Model

We consider a discrete-time single-server queueing system with infinite buffer space. Time is assumed to be slotted. There are 2 types of traffic arriving in the system, namely packets of class 1 and packets of class 2. We denote the number of arrivals of class j during slot k by $a_{j,k}$ $(j = 1, 2)$. Both types of packet arrivals are assumed to be i.i.d. from slot-to-slot and are characterized by the joint probability mass function $a(m, n)$ and joint probability generating function (pgf) $A(z_1, z_2)$. Notice that the number of packet arrivals from different classes (within a slot) can be correlated. Further, we define the marginal pgf's of the arrivals from class 1 and class 2 during a slot by $A_1(z) \triangleq E[z^{a_{1,k}}] = A(z, 1)$ and $A_2(z) \triangleq E[z^{a_{2,k}}] = A(1, z)$ respectively. We furthermore denote the arrival rate of class j $(j = 1, 2)$ by $\lambda_j = A_j'(1)$.

The service times of the class j packets are assumed to be i.i.d. and are characterized by the probability mass function $s_j(m)$ and probability generating function $S_j(z)$ $(j = 1, 2)$. We furthermore denote the mean service time of a class j packet by $\mu_j = S_j'(1)$. We define the load offered by class j packets as $\rho_j \triangleq \lambda_j \mu_j$ $(j = 1, 2)$. The total load is then given by $\rho \triangleq \rho_1 + \rho_2$.

The system has one server that provides the transmission of packets. Class 1 packets are assumed to have non-preemptive priority over class 2 packets, and within one class the service discipline is FCFS. Due to the priority scheduling mechanism, it is as if class 1 packets are stored in front of class 2 packets in the queue. So, if there are any class 1 packets in the queue when the server becomes

idle, the one with the longest waiting time will be served next. If, on the other hand, no class 1 packets are present in the queue at that moment, the class 2 packet with the longest waiting time, if any, will be served next. Since the priority scheduling is non-preemptive, service of a packet will not be interrupted by newly arriving packets.

3 System Contents

To be able to analyze the packet delay, we will first analyse the system contents at the beginning of so-called start slots, i.e., slots at the beginning of which a packet (if available) can enter the server. Note that every slot during which the system is empty, is also a start slot. We denote the system contents of class j packets at the beginning of the l-th start slot by $n_{j,l}$ $(j = 1, 2)$. Their joint pgf will be denoted by $N_l(z_1, z_2)$. Clearly, the set $\{(n_{1,l}, n_{2,l})\}$ forms a Markov chain, since the arrival process is i.i.d. and only random variables during start slots are involved. If s^* indicates the service time of the packet that enters service at the beginning of start slot l (which is - by definition - regular slot k) the following system equations can be established:

1. If $n_{1,l} = n_{2,l} = 0$:

$$n_{1,l+1} = a_{1,k} \ ; \ n_{2,l+1} = a_{2,k}, \tag{1}$$

   i.e., the only packets present in the system at the beginning of start slot $l+1$ are the packets that arrived during the previous slot, i.e., start slot l.

2. If $n_{1,l} = 0$ and $n_{2,l} > 0$:

$$n_{1,l+1} = \sum_{i=0}^{s^*-1} a_{1,k+i} \ ; \ n_{2,l+1} = n_{2,l} + \sum_{i=0}^{s^*-1} a_{2,k+i} - 1, \tag{2}$$

   i.e., the class 2 packet in service leaves the system just before start slot $l+1$. s^* is characterized by probability mass function $s_2(m)$, since a class 2 packet enters the server at the beginning of start slot l.

3. If $n_{1,l} > 0$:

$$n_{1,l+1} = n_{1,l} + \sum_{i=0}^{s^*-1} a_{1,k+i} - 1 \ ; \ n_{2,l+1} = n_{2,l} + \sum_{i=0}^{s^*-1} a_{2,k+i}, \tag{3}$$

   i.e., the class 1 packet in service leaves the system just before start slot $l+1$. s^* is characterized by probability mass function $s_1(m)$, since a class 1 packet enters the server at the beginning of start slot l.

In the remainder, we define $\mathrm{E}[X\{Y\}]$ as $\mathrm{E}[X|Y]\mathrm{Prob}[Y]$. The system equations (1)-(3) can now be translated into relations between z-transforms. Exploiting the statistical independence of the (set of) random variables s^*, $(n_{1,l}, n_{2,l})$ and $(a_{1,k+i}, a_{2,k+i}), i \geq 0$, respectively, this leads to the following relation:

$$N_{l+1}(z_1, z_2) \triangleq \mathrm{E}\left[z_1^{n_{1,l+1}} z_2^{n_{2,l+1}}\right] = \mathrm{E}\left[z_1^{n_{1,l+1}} z_2^{n_{2,l+1}} \{n_{1,l} = n_{2,l} = 0\}\right]$$

$$+ \mathrm{E}\left[z_1^{n_{1,l+1}} z_2^{n_{2,l+1}} \{n_{1,l} = 0, n_{2,l} > 0\}\right] + \mathrm{E}\left[z_1^{n_{1,l+1}} z_2^{n_{2,l+1}} \{n_{1,l} > 0\}\right]$$
$$= A(z_1, z_2) N_l(0,0) + \frac{S_2(A(z_1,z_2))}{z_2} \left[N_l(0,z_2) - N_l(0,0)\right] \quad (4)$$
$$+ \frac{S_1(A(z_1,z_2))}{z_1} \left[N_l(z_1,z_2) - N_l(0,z_2)\right].$$

We assume that the system is stable (implying that the equilibrium condition requires that $\rho < 1$) and as a result $N_l(z_1, z_2)$ and $N_{l+1}(z_1, z_2)$ converge both to a common steady-state limit denoted by $N(z_1, z_2)$. By taking the $l \to \infty$ limit of equation (4), we obtain:

$$[z_1 - S_1(A(z_1,z_2))] N(z_1,z_2) = z_1 \frac{z_2 A(z_1,z_2) - S_2(A(z_1,z_2))}{z_2} N(0,0) \quad (5)$$
$$+ \frac{z_1 S_2(A(z_1,z_2)) - z_2 S_1(A(z_1,z_2))}{z_2} N(0,z_2).$$

It now remains for us to determine the unknown function $N(0, z_2)$ and the unknown parameter $N(0, 0)$. This can be done in two steps. First, we notice that $N(z_1, z_2)$ must be bounded for all values of z_1 and z_2 such that $|z_1| \leq 1$ and $|z_2| \leq 1$. In particular, this should be true for $z_1 = Y(z_2)$, with $Y(z_2) \triangleq S_1(A(Y(z_2), z_2))$ and $|z_2| \leq 1$, since it follows from Rouché's theorem that there is exactly one solution $|Y(z_2)| \leq 1$ for all such z_2. Notice that $Y(1)$ equals 1. The above implies that if we choose $z_1 = Y(z_2)$ in equation (5), where $|z_2| \leq 1$, the left hand side of this equation vanishes. The same must then be true for the right hand side, yielding

$$N(0,z_2) = N(0,0) \frac{z_2 A(Y(z_2),z_2) - S_2(A(Y(z_2),z_2))}{z_2 - S_2(A(Y(z_2),z_2))}. \quad (6)$$

Finally, in order to find an expression for $N(0, 0)$, we put $z_1 = z_2 = 1$ and use de l'Hospital's rule in equation (5). Therefore, we need the first derivative of $Y(z)$ for $z = 1$ and this is given by

$$Y'(1) = \mu_1(\lambda_1 Y'(1) + \lambda_2) = \frac{\lambda_2 \mu_1}{1 - \rho_1}. \quad (7)$$

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

$$N(0,0) = \frac{1-\rho}{1 - \rho + \lambda_1 + \lambda_2}. \quad (8)$$

A fully determined expression for $N(z_1, z_2)$ can now be derived by combining equations (5) and (6):

$$N(z_1,z_2) = N(0,0) \left[\frac{z_1(z_2 A(z_1,z_2) - S_2(A(z_1,z_2)))}{(z_1 - S_1(A(z_1,z_2)))(z_2 - S_2(A(Y(z_2),z_2)))} \right.$$
$$\left. + \frac{S_2(A(Y(z_2),z_2))(S_1(A(z_1,z_2)) - z_1 A(z_1,z_2))}{(z_1 - S_1(A(z_1,z_2)))(z_2 - S_2(A(Y(z_2),z_2)))} \right] \quad (9)$$

$$+ \frac{A(Y(z_2),z_2)(z_1 S_2(A(z_1,z_2)) - z_2 S_1(A(z_1,z_2)))}{(z_1 - S_1(A(z_1,z_2)))(z_2 - S_2(A(Y(z_2),z_2)))}\Bigg],$$

with $N(0,0)$ given by equation (8) and $Y(z)$ implicitly defined by $Y(z) = S_1(A(Y(z),z))$.

4 Packet Delay

The packet delay is defined as the total time period a tagged packet spends in the system, i.e., the number of slots between the end of the packet's arrival slot and the end of its departure slot. We denote the delay of a tagged class j packet by d_j and its pgf by $D_j(z)$ $(j = 1, 2)$. Before deriving expressions for $D_1(z)$ and $D_2(z)$, we first define some stochastic variables we will frequently use in this section. We denote the arrival slot of the tagged packet by slot k. If slot k is a start slot, it is assumed to be start slot l. If slot k is not a start slot on the other hand, the last start slot preceeding slot k is assumed to be start slot l. We denote the number of class j packets that arrive during slot k, but which are served before the tagged packet by f_j $(j = 1, 2)$. We denote the service time of the tagged class j packet by s_j^* $(j = 1, 2)$. We finally denote the service time and the elapsed service time of the packet in service (if any) during the arrival slot of the tagged packet by s^* and s^+ respectively.

4.1 Delay of Class 1 Packets

We tag a class 1 packet. There are 3 possibilities when the tagged packet arrives:

1. The server is idle during slot k, yielding

$$d_1 = \sum_{m=1}^{f_1} s_{1,m}^{(k)} + s_1^*, \tag{10}$$

   with the $s_{1,m}^{(k)}$'s the service times of the class 1 packets that arrived during slot k, but that are served before the tagged class 1 packet.

2. A class 2 packet is in service during slot k (implying that $n_{1,l} = 0$, $n_{2,l} > 0$), yielding

$$d_1 = (s^* - s^+ - 1) + \sum_{i=1}^{s^+} \sum_{m=1}^{a_{1,k-i}} s_{1,m}^{(k-i)} + \sum_{m=1}^{f_1} s_{1,m}^{(k)} + s_1^*, \tag{11}$$

   with the $s_{1,m}^{(k-i)}$'s $(0 \le i \le s^+)$ the service times of the class 1 packets that arrived during slot $k - i$. The residual service time of the packet in service during slot k contributes in the first term, the service times of the class 1 packets in the system at the beginning of slot k contribute in the second term, the service times of the class 1 packets arrived during slot k, but served before the tagged class 1 packet contribute in the third term, and finally the service time of the tagged class 1 packet itself contributes in the last term.

3. A class 1 packet is in service during slot k (i.e., $n_{1,l} > 0$), yielding

$$d_1 = (s^* - s^+ - 1) + \sum_{m=1}^{n_{1,l}-1} \tilde{s}_{1,m} + \sum_{i=1}^{s^+} \sum_{m=1}^{a_{1,k-i}} s_{1,m}^{(k-i)} + \sum_{m=1}^{f_1} s_{1,m}^{(k)} + s_1^*. \quad (12)$$

The difference with the previous situation is that there may be multiple high priority packets in the buffer (apart from the one in service) at the beginning of slot l, which will contribute to the tagged packet's delay. If we denote by $\tilde{s}_{1,m}$ the service times of the class 1 packets already in the queue at the beginning of the ongoing service (thus without the packet in service during slot k), then this condition is quantified by the second term in the right-hand side of the above expression.

Again, equations (10)-(12) can be z-transformed. Taking the sum then eventually leads to an expression for $D_1(z)$:

$$\begin{aligned} D_1(z) \triangleq \mathrm{E}[z^{d_1}] &= \mathrm{E}\left[z^{d_1}\{\text{no service}\}\right] + \mathrm{E}\left[z^{d_1}\{\text{service class 2 packet}\}\right] \\ &\quad + \mathrm{E}\left[z^{d_1}\{\text{service class 1 packet}\}\right] \\ &= F_1(S_1(z))S_1(z)\left\{1 - \rho + \rho_2 \frac{S_2^*\left(\frac{A_1(S_1(z)))}{z}, z\right)}{z}\right. \\ &\quad \left. + \rho_1 \frac{S_1^*\left(\frac{A_1(S_1(z)))}{z}, z\right)}{z} \frac{N(S_1(z),1) - N(0,1)}{(1 - N(0,1))S_1(z)}\right\}, \end{aligned} \quad (13)$$

with $F_1(z) \triangleq E[z^{f_1}]$, $S_2^*(x,z) \triangleq E[x^{s^+} z^{s^*} | n_{1,l} = 0, n_{2,l} > 0]$ and $S_1^*(x,z) \triangleq E[x^{s^+} z^{s^*} | n_{1,l} > 0]$. The random variable f_1 can be shown to have the following pgf (see e.g. [2]):

$$F_1(z) = \frac{A_1(z) - 1}{\lambda_1(z-1)}. \quad (14)$$

If a class j packet is in service during slot k, s^* is characterized by the probability mass function $s_j(m)$ $(j = 1, 2)$. Notice that the distributions of s^* and s^+ are correlated, since s^+ is the elapsed part of the service time s^* at the beginning of slot k. Considering these observations, one can derive the following expression for $S_j^*(x,z)$:

$$S_j^*(x,z) = \frac{S_j(xz) - S_j(z)}{\mu_j(x-1)}, \quad (15)$$

with $j = 1, 2$. Substitution of (9), (14) and (15) into equation (13) finally leads to a closed-form version of $D_1(z)$:

$$D_1(z) = \frac{(1-\rho)(z-1) + \lambda_2(S_2(z) - 1)}{\lambda_1(S_1(z) - 1)} \frac{S_1(z)(A_1(S_1(z)) - 1)}{z - A_1(S_1(z))}. \quad (16)$$

4.2 Delay of Class 2 Packets

Because of the priority discipline, an expression for d_2 will be a bit more involved. We now 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.

When the tagged class 2 packet arrives, there are 3 possibilities:

1. The server is idle during slot k, yielding

$$d_2 = \sum_{j=1}^{2} \sum_{m=1}^{f_j} v_{j,m}^{(k)} + s_2^*, \tag{17}$$

   i.e., f_1 class 1 primary packets and f_2 class 2 primary packets that arrived during slot k and their class 1 successors have to be served before the tagged class 2 packet.

2. A class 2 packet is in service during slot k, yielding

$$d_2 = (s^* - s^+ - 1) + \sum_{i=1}^{s^*-s^+-1} \sum_{m=1}^{a_{1,k+i}} v_{1,m}^{(k+i)} + \sum_{j=1}^{2} \sum_{m=1}^{f_j} v_{j,m}^{(k)} \tag{18}$$
$$+ \sum_{j=1}^{2} \sum_{i=1}^{s^+} \sum_{m=1}^{a_{j,k-i}} v_{j,m}^{(k-i)} + \sum_{m=1}^{n_{2,l}-1} \tilde{v}_{2,m} + s_2^*,$$

   with the $\tilde{v}_{2,m}$'s 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 1 and class 2 packets that arrived during the elapsed service time contributes in the fourth term, the sub-busy period of the class 2 packets already in the queue at the beginning of start slot l contributes in the fifth term and finally the service time of the tagged class 2 packet itself contributes in the last term.

3. A class 1 packet is in service during slot k, yielding

$$d_2 = (s^* - s^+ - 1) + \sum_{i=1}^{s^*-s^+-1} \sum_{m=1}^{a_{1,k+i}} v_{1,m}^{(k+i)} + \sum_{j=1}^{2} \sum_{m=1}^{f_j} v_{j,m}^{(k)} + \sum_{j=1}^{2} \sum_{i=1}^{s^+} \sum_{m=1}^{a_{j,k-i}} v_{j,m}^{(k-i)} + \sum_{m=1}^{n_{1,l}-1} \tilde{v}_{1,m} + \sum_{m=1}^{n_{2,l}} \tilde{v}_{2,m} + s_2^*, \tag{19}$$

with the $\tilde{v}_{j,m}$'s $(j = 1, 2)$ the sub-busy periods, caused by the m-th class j packet already in the queue at the beginning of start slot l. The expression is virtually the same as in the previous case, with an additional term that takes into account the sub-busy periods of the class 1 packets already in the system when the transmission of the class 1 packet currently in the server started (i.e., at the beginning of slot l).

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 and f_2 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 (17)-(19) and taking into account the previous remarks, we can derive an expression for $D_2(z)$:

$$\begin{aligned} D_2(z) \triangleq \mathrm{E}[z^{d_2}] &= \mathrm{E}\left[z^{d_2}\{\text{no service}\}\right] + \mathrm{E}\left[z^{d_2}\{\text{service class 2 packet}\}\right] \\ &\quad + \mathrm{E}\left[z^{d_2}\{\text{service class 1 packet}\}\right] \\ &= F(V_1(z), V_2(z)) S_2(z) \Bigg\{ 1 - \rho + \rho_2 \frac{S_2^*\left(\frac{A(V_1(z),V_2(z))}{zA_1(V_1(z))}, zA_1(V_1(z))\right)}{zA_1(V_1(z))} \\ &\quad \frac{N(0,V_2(z)) - N(0,0)}{(N(0,1) - N(0,0))V_2(z)} + \rho_1 \frac{S_1^*\left(\frac{A(V_1(z),V_2(z))}{zA_1(V_1(z))}, zA_1(V_1(z))\right)}{zA_1(V_1(z))} \\ &\quad \frac{N(V_1(z),V_2(z)) - N(0,V_2(z))}{(1 - N(0,1))V_1(z)} \Bigg\}, \end{aligned} \tag{20}$$

with $F(z_1, z_2) \triangleq E[z_1^{f_1} z_2^{f_2}]$, $S_2^*(x,z) \triangleq E[x^{s^+} z^{s^*} | n_{1,l} = 0, n_{2,l} > 0]$ and $S_1^*(x,z) \triangleq E[x^{s^+} z^{s^*} | n_{1,l} > 0]$. The random variables f_1 and f_2 can be shown to have the following joint pgf (extension of a technique used in e.g. [2]):

$$F(z_1, z_2) = \frac{A(z_1, z_2) - A_1(z_1)}{\lambda_2(z_2 - 1)}. \tag{21}$$

The $S_j^*(x,z)$'s $(j = 1, 2)$ are again given by equation (15). Finally, we have to find expressions for $V_1(z)$ and $V_2(z)$. These pgfs satisfy the following relations:

$$V_j(z) = S_j(zA_1(V_1(z))), \tag{22}$$

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 (22). Equation (20) together with equations (21) and (15) leads to a fully determined version for $D_2(z)$:

$$D_2(z) = \frac{1-\rho}{\lambda_2} \frac{S_2(z)(A(V_1(z), V_2(z)) - A_1(V_1(z)))}{zA_1(V_1(z)) - A(V_1(z), V_2(z))} \frac{1 - zA_1(V_1(z))}{1 - V_2(z)}. \tag{23}$$

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 cell delay, on the contrary, can be calculated in closed-form. For example, $Y'(1)$ is given by equation (7) and the first derivatives of $V_j(z)$ for $z = 1$ are given by

$$V_j'(1) = \frac{\mu_j}{1 - \rho_1},$$

with $(j = 1, 2)$. Let us define λ_{ij} and μ_{jj} as

$$\lambda_{ij} \triangleq \left. \frac{\partial^2 A(z_1, z_2)}{\partial z_i \partial z_j} \right|_{z_1 = z_2 = 1} ; \; \mu_{jj} \triangleq \left. \frac{d^2 S_j(z)}{dz^2} \right|_{z=1},$$

with $i, j = 1, 2$. Now we can calculate the mean values of the packet delay of both classes by taking the first derivatives of the respective pgfs for $z = 1$. We find

$$E[d_1] = \mu_1 + \frac{1}{2} \frac{\mu_1 \lambda_{11}}{\lambda_1(1 - \rho_1)} + \frac{1}{2} \frac{\lambda_1 \mu_{11} + \lambda_2 \mu_{22}}{1 - \rho_1},$$

for the mean value of the packet delay of a class 1 packet and

$$E[d_2] = \mu_2 + \frac{1}{2} \frac{\mu_1^2 \lambda_{11}}{(1 - \rho)(1 - \rho_1)} + \frac{1}{2} \frac{2\mu_1 \lambda_{12} + \mu_2 \lambda_{22}}{\lambda_2(1 - \rho)} + \frac{1}{2} \frac{\lambda_1 \mu_{11} + \lambda_2 \mu_{22}}{(1 - \rho)(1 - \rho_1)},$$

for the mean value of the packet delay of a class 2 packet. In a similar way, expressions for the variance (or higher order moments) can be calculated as well by taking the appropriate derivatives of the respective generating functions.

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:

$$A(z_1, z_2) = (1 - \frac{\lambda_1}{N}(1 - z_1) - \frac{\lambda_2}{N}(1 - z_2))^N. \tag{24}$$

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 switch fed by a Bernoulli process at the inlets (see [16]). 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 will furthermore assume deterministic service times for both classes.

In Fig. 1., the mean value of the packet delay of class 1 packets and class 2 packets is shown as a function of the total load ρ, when $\mu_1 = \mu_2 = 2$. The fraction of class 1 arrivals is 0.25, 0.5 and 0.75 respectively of the total number of arrivals. In order to compare with FIFO scheduling, we have also shown the mean value of the packet delay in that case. Since, in this example, the service times of the class 1 and class 2 packets are equal, 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 situation has already been analyzed, e.g., in [2]. One can observe the influence of priority scheduling: mean delay of class 1 packets reduces significantly. The price to pay is of course a larger mean delay for 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 packets in the overall traffic mix, the lower the mean packet delay of both classes will be.

Fig. 2. shows the mean value of the packet delay of class 1 packets as a function of the total load, when $\lambda_1 = 0.25$, $\mu_1 = 2$ and $\mu_2 = 1, 2, 4, 8, 16$. This figure shows the influence of the non-preemptive priority scheduling. When the service time of a class 2 packet is assumed to be deterministically equal to 1 slot, i.e., $\mu_2 = 1$, the preemptive priority scheduling has the same effect as the non-preemptive priority scheduling. If $\mu_2 > 1$, the non-preemptive priority has worse performance than the preemptive priority scheduling in terms of the mean packet delay for class 1 packets. Furthermore, for a given value of the low priority packet length, the mean high priority packet delay increases proportional to the total load ρ.

7 Conclusions

In this paper, we analyzed the packet delay in a queueing system with non-preemptive HOL priority scheduling. A generating-functions-approach was adopted, which led to closed-form expressions of performance measures, such as mean of packet delay of both classes, that are easy to evaluate. The model included

possible correlation between the number of arrivals of the two classes during a slot and general service times for packets of both classes. Therefore, the results can be applied to analyse the performance of traffic streams in an environment with delay and loss sensitive traffic, such as an integrated voice/data IP network.

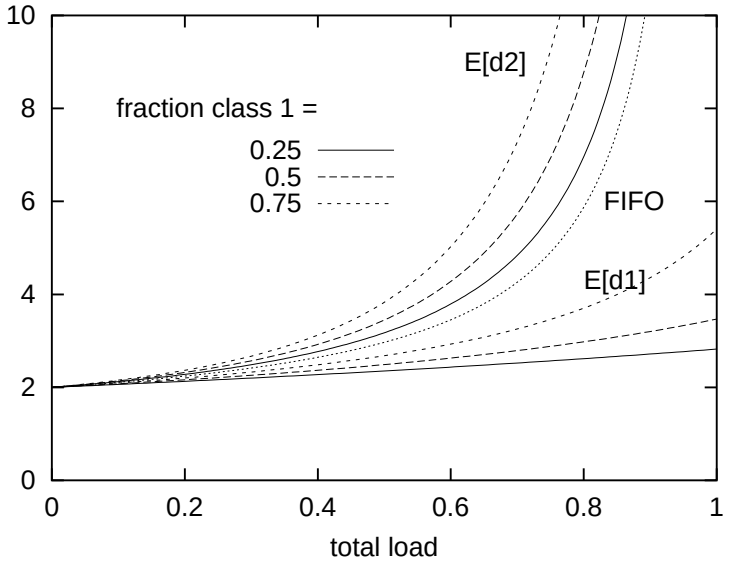

Fig. 1. Mean packet delay when the fraction of class 1 arrivals equals 0.25, 0.5 and 0.75

References

[1] J.J. Bae and T. Suda, *Survey of traffic control schemes and protocols in ATM networks*, Proceedings of the IEEE 79(2), pp. 170-189, 1991.

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

[3] N.K. Jaiswal, *Priority queues*, Academic Press, New York, 1968.

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

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

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

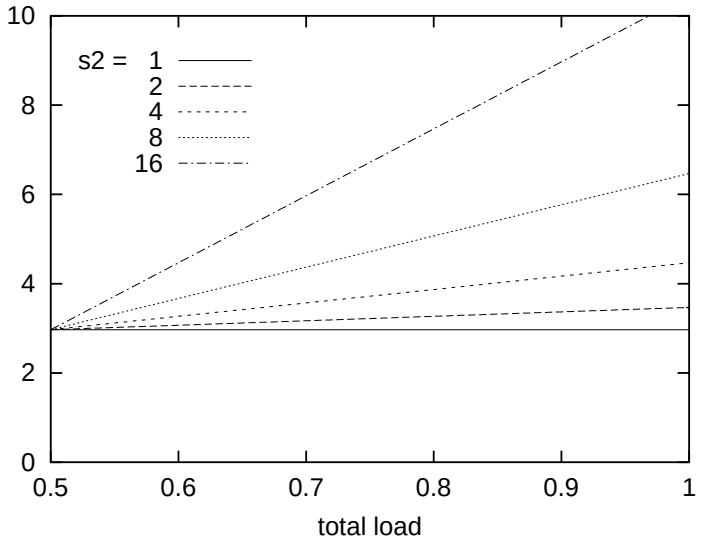

Fig. 2. Mean packet delay of class 1 packets when the service time of class 2 packets equals 1, 2, 4, 8, 16

[7] I. Rubin and Z. Tsai, *Message delay analysis of multiclass priority TDMA, FDMA, and discrete-time queueing systems*, IEEE Transactions on Information Theory 35(3), pp. 637-647, 1989.

[8] D.A. Stanford, *Interdeparture-time distributions in the non-preemptive priority* $\sum M_i/G_i/1$ *queue*, Performance Evaluation 12, pp. 43-60, 1991.

[9] A. Sugahara, T. Takine, Y. Takahashi and T. Hasegawa, *Analysis of a nonpreemptive priority queue with SPP arrivals of high class*, Performance Evaluation 21, pp. 215-238, 1995.

[10] L. Takacs, *Priority queues*, Operations Research 12, pp. 63-74, 1964.

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

[12] T. Takine, Y. Matsumoto, T. Suda and T. Hasegawa, *Mean waiting times in nonpreemptive priority queues with Markovian arrival and i.i.d. service processes*, Performance Evaluation 20, pp. 131-149, 1994.

[13] 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), pp. 1837-1845, 1994.

[14] T. Takine, *A nonpreemptive priority MAP/G/1 queue with two classes of customers*, Performance Evaluation 20, pp. 266-290, 1996.

[15] P. Van Mieghem, B. Steyaert and G.H. Petit, *Performance of cell loss priority management schemes in a single server queue*, International Journal of Communication Systems 10, pp. 161-180, 1997.

[16] J. Walraevens and H. Bruneel, *HOL priority in an ATM output queueing switch*, Proceedings of the seventh IFIP workshop on performance modelling and evaluation of ATM/IP networks, Antwerp, 28-30 june, 1999.