Mixed Finite-/Infinite-Capacity Priority Queue with Interclass Correlation

Thomas Demoor*, Joris Walraevens, Dieter Fiems,
and Herwig Bruneel

SMACS Research Group,
Department of Telecommunications and Information Processing,
Ghent University,
St.-Pietersnieuwstraat 41,
B-9000 Gent, Belgium
Tel.: +3292648902
{thdemoor,jw,df,hb}@telin.ugent.be

Abstract. We consider a discrete-time queueing system with two priority classes and absolute priority scheduling. In our model, we capture potential correlation between the arrivals of the two priority classes. For practical use, it is required that the high-priority queue is of (relatively) small size and we hence use a model with finite high-priority queue capacity. We obtain expressions for the probability mass functions of the steady-state system content and delay of the high-priority class as well as for the probability generating functions and moments of the steady-state system content and delay of the low-priority class. The results are compared to those of a similar system, but with an infinite capacity for high priority packets, and it is shown that the latter can be inaccurate. We also investigate the effect of correlation between the arrivals of both priority classes on the performance of the system.

Keywords: Queueing Systems and Network Models, Performance Modelling.

1 Introduction

The huge difference between the Quality of Service (QoS) demands for real-time traffic flows and best-effort traffic flows emburdens packet-based telecommunication networks, such as the Internet. Real-time traffic, such as Voice over IP, can often endure some packet loss but requires low delays and/or low delay jitter. Best-effort traffic benefits from low packet loss, hence avoiding retransmissions, but has less stringent delay characteristics. Therefore the packets are distributed into classes according to their QoS requirements. In the nodes (routers, ...) of the network, packets typically have to wait before being transmitted to the next node and thus constitute a queueing process. The order in which packets are transmitted is based on class-dependent priority rules. This approach to QoS

* Corresponding author.

K. Al-Begain, A. Heindl, and M. Telek (Eds.): ASMTA 2008, LNCS 5055, pp. 61–74, 2008.

differentiation is applied in the DiffServ architecture for Internet Protocol (IP), successfully implemented in corporate networks and debated on as one of the possible approaches to QoS in the future Internet [1].

In this paper, we study a queueing system with a single server supplying two queues, one per priority class, and an Absolute Priority scheduling algorithm, in order to minimize the delay of the high-priority (real-time) packets. The low-priority packets (class 2) are only served if there are no high-priority packets (class 1) in the system. This is the most drastic scheduling method, minimizing class-1 delay at the cost of class-2 performance, but is easy to implement. Analytic studies of queueing systems generally assume infinite queue capacity facilitating the mathematical analysis of the system. In the setting under consideration however, the class-1 packets are delay-sensitive so we require the class-1 queue capacity to be as small as possible while still meeting the required packet loss constraints for this traffic. Note that class-1 traffic that does not arrive at its destination in time is of no use and can be considered lost. We therefore consider a system with finite class-1 queue capacity. On the other hand, the loss-sensitivity of the class-2 packets results in a class-2 queue capacity as large as practically feasible, justifying the assumption of an infinite class-2 queue capacity. Notice that the number of arrivals of both classes can be correlated. A single user often generates packets of both classes simultaneously or no packets at all, yielding positive correlation. On the other hand, negative correlation arises when the number of sources that generate (class-1 or class-2) packets is limited as will be further investigated in the study of an output-queueing switch in the applications.

In the literature, priority queues have been discussed with various arrival and service processes, such as in the contributions [2,3,4,5,6,7]. The presented paper complements [2] where both queues are presumed to be of infinite capacity. As the class-1 queue capacity must be small in order to obtain a low class-1 delay the assumption that this queue has infinite capacity can lead to inaccurate results. Assessing the impact of the finite class-1 queue capacity on the performance measures is the purpose of the present contribution. Finite queue capacity is considered in [5] as well but only the packet loss is investigated profoundly and the delay is not analysed.

This paper is constituted as follows. First the model under consideration will be thoroughly described. In section 3 we investigate the system content for both classes and the class-1 packet loss ratio. Section 4 handles the delay of both classes. Afterwards, the results are applied in some numerical examples. We finally formulate our conclusions.

2 Model

We consider a discrete-time single-server priority queueing system with 2 classes, finite class-1 queue capacity N and an infinite class-2 queue. Class-1 packets are served with absolute priority over class-2 packets and within a class the queueing discipline is First-Come-First-Served (FCFS). The Tail Drop queue management

algorithm is used for the class-1 queue, hence the system accepts packets until the corresponding queue is entirely filled and packets that arrive at a full queue are dropped by the system. Time is divided into fixed-length slots corresponding to the transmission time of a packet. A packet can only enter the server at the beginning of a slot, even if it arrives in an empty system, and its service takes until the end of that slot (deterministic). The system can contain up to $N + 1$ class-1 packets simultaneously in a slot, N in the queue and 1 in the server. Consequently, there are at most N class-1 packets in the system at the beginning of a slot. Also note that a class-1 packet thus resides in the system for at most N slots, which bounds its delay.

We assume that for both classes the number of arrivals in consecutive slots form a sequence of independent and identically distributed (i.i.d.) random variables. We define $a_{i,k}$ as the number of class-i ($i = 1, 2$) packet arrivals during slot k. The arrivals of both classes are characterized by the joint probability mass function (pmf)

$$a(m, n) = \text{Prob}[a_{1,k} = m, a_{2,k} = n] \ , \tag{1}$$

and joint probability generating function (pgf)

$$A(z_1, z_2) = \text{E}[z_1^{a_{1,k}} z_2^{a_{2,k}}] \ . \tag{2}$$

Note that the arrival process allows correlation between both classes. Let the mean number of class-i arrivals per slot (class-i arrival load) be

$$\lambda_i = \text{E}[a_{i,k}] = \left. \frac{\partial A(z_1, z_2)}{\partial z_i} \right|_{z_1=1, z_2=1} , \quad (i = 1, 2) \ . \tag{3}$$

The total arrival load equals $\lambda_T = \lambda_1 + \lambda_2$.

We also define the pgf of the class-2 arrivals in a slot with i (i or more) class-1 arrivals as $A_i(z)$ ($A_i^*(z)$), yielding

$$\begin{aligned} A_i(z) &= \text{E}[z^{a_{2,k}} \mathbf{1}_{a_{1,k}=i}] \ , \\ A_i^*(z) &= \sum_{l=i}^{\infty} A_l(z) \ . \end{aligned} \tag{4}$$

Note that the indicator function $\mathbf{1}_{x=i}$ is 1 if $x = i$ and equals 0 otherwise.

The aim is to express the system content and delay of both classes in terms of the arrival process. The system content at the beginning of a slot is the number of packets contained by the system, thus by the queue or by the server, before packets arrive in the considered slot. The delay of a packet is the number of slots between its arrival slot and the slot after its departure. The class-1 packet loss ratio, this is the fraction of packets that arrive at the system but are not accepted into the system because the class-1 queue is entirely filled, is to be obtained as well.

3 System Content

Let the class-i system content at the beginning of slot k be denoted by $u_{i,k}$. The corresponding joint pgf is referred to as

$$U_k(z_1, z_2) = \mathrm{E}[z_1^{u_{1,k}} z_2^{u_{2,k}}] \ . \tag{5}$$

The (partial) pgf of the class-2 system content in a slot with class-1 system content equal to i is defined as

$$U_{i,k}(z) = \mathrm{E}[z^{u_{2,k}} \mathbf{1}_{u_{1,k}=i}] \ . \tag{6}$$

Note that

$$U_k(z_1, z_2) = \sum_{i=0}^{N} U_{i,k}(z_2) z_1^i \ . \tag{7}$$

Relating the system contents at the beginning of slots k and $k+1$ yields

$$\begin{aligned} u_{1,k+1} &= (u_{1,k} - 1)^+ + a^e_{1,k}, \\ u_{2,k+1} &= \begin{cases} (u_{2,k} - 1)^+ + a_{2,k}, & \text{if } u_{1,k} = 0, \\ u_{2,k} + a_{2,k}, & \text{if } u_{1,k} > 0, \end{cases} \end{aligned} \tag{8}$$

where $(x)^+$ denotes the maximum of x and 0. Due to the finite class-1 capacity, we only take the effectively admitted class-1 arrivals into account. The number of effective class-1 arrivals in slot k, denoted by $a^e_{1,k}$, is clearly influenced by the class-1 system content in slot k. This can be expressed as

$$a^e_{1,k} = \min(a_{1,k}, N - (u_{1,k} - 1)^+) \ . \tag{9}$$

Standard z-transform techniques enable the expression of (8) in terms of pgfs. We establish the system of equations

$$\begin{aligned} U_{i,k+1}(z) =& \frac{1}{z} U_{0,k}(z) A_i(z) + \frac{z-1}{z} U_{0,k}(0) A_i(z) \\ &+ \left(\sum_{j=1}^{i+1} U_{j,k}(z) A_{i-j+1}(z) \right), \quad i = 0 \ldots N-1 \ , \\ U_{N,k+1}(z) =& \frac{1}{z} U_{0,k}(z) A^*_N(z) + \frac{z-1}{z} U_{0,k}(0) A^*_N(z) \\ &+ \left(\sum_{j=1}^{N} U_{j,k}(z) A^*_{N-j+1}(z) \right) \ . \end{aligned} \tag{10}$$

The impact of the finite class-1 queue capacity is apparent when the queue is entirely filled due to i extra effective arrivals. These effective arrivals can correspond with i arrivals, or with $i+1$ arrivals of which one is dropped because the queue is full, or with $i+2$ arrivals of which two are dropped, ... This leads to the appearance of the pgfs $A^*_i(z)$ in the last equation of (10).

Under the assumption that the system reaches steady state, on which we will elaborate at the end of this section, let us define

$$\begin{aligned} U_i(z) &= \lim_{k\to\infty} U_{i,k}(z) = \lim_{k\to\infty} U_{i,k+1}(z), \quad i = 0 \dots N \ , \\ U(z_1, z_2) &= \lim_{k\to\infty} U_k(z_1, z_2) = \sum_{i=0}^{N} U_i(z_2) z_1^i \ . \end{aligned} \tag{11}$$

In steady-state the system of equations (10) becomes

$$\begin{aligned} U_i(z) =& \frac{1}{z} U_0(z) A_i(z) + \frac{z-1}{z} U_0(0) A_i(z) \\ &+ \left(\sum_{j=1}^{i+1} U_j(z) A_{i-j+1}(z) \right), \quad i = 0 \dots N-1 \ , \\ U_N(z) =& \frac{1}{z} U_0(z) A_N^*(z) + \frac{z-1}{z} U_0(0) A_N^*(z) \\ &+ \left(\sum_{j=1}^{N} U_j(z) A_{N-j+1}^*(z) \right) . \end{aligned} \tag{12}$$

We now define the $(N+1) \times (N+1)$ matrix

$$\mathbf{X}(z) = \begin{pmatrix} A_0(z) & A_1(z) & A_2(z) & \cdots & A_{N-1}(z) & A_N^*(z) \\ A_0(z)z & A_1(z)z & A_2(z)z & \cdots & A_{N-1}(z)z & A_N^*(z)z \\ 0 & A_0(z)z & A_1(z)z & \cdots & A_{N-2}(z)z & A_{N-1}^*(z)z \\ 0 & 0 & A_0(z)z & \cdots & A_{N-3}(z)z & A_{N-2}^*(z)z \\ \vdots & \vdots & \ddots & \cdots & \vdots & \vdots \\ 0 & \cdots & \cdots & \cdots & A_0(z)z & A_1^*(z)z \end{pmatrix} , \tag{13}$$

and the row vectors of $N+1$ elements

$$\mathbf{Y}(z) = \begin{pmatrix} A_0(z) \\ A_1(z) \\ \vdots \\ A_{N-1}(z) \\ A_N^*(z) \end{pmatrix}^T , \quad \mathbf{U}(z) = \begin{pmatrix} U_0(z) \\ U_1(z) \\ \vdots \\ U_{N-1}(z) \\ U_N(z) \end{pmatrix}^T . \tag{14}$$

In view of these definitions, the system of equations (12) is equivalent with

$$\mathbf{U}(z) \Big(z\mathbf{I}_{N+1} - \mathbf{X}(z) \Big) = (z-1) U_0(0) \mathbf{Y}(z) \ . \tag{15}$$

Here $\mathbf{I}_n$ denotes the $n \times n$ identity matrix. We have expressed $\mathbf{U}(z)$ in terms of known quantities and the unknown constant $U_0(0)$. For $z = 1$ this yields

$$\mathbf{U}(1) \Big(\mathbf{I}_{N+1} - \mathbf{X}(1) \Big) = \begin{pmatrix} 0 & 0 & \cdots & 0 & 0 \end{pmatrix} \ . \tag{16}$$

As $\mathbf{X}(1)$ is a right stochastic matrix we find

$$\text{Rank}\Big(\mathbf{I}_{N+1} - \mathbf{X}(1)\Big) = N \ . \tag{17}$$

We thus require an additional relation in order to determine the $N+1$ unknowns in the vector $\mathbf{U}(1)$. From (6) it is clear that $U_i(1) = \text{Prob}[u_1 = i]$. As the class-1 system content is normalised over the $N+1$ possible states we establish

$$\sum_{i=0}^{N} U_i(1) = 1 \ . \tag{18}$$

By replacing a relation in equation (16) by the normalisation condition we obtain the pmf of the class-1 system content as

$$\mathbf{U}(1) = \big(0\ 0 \cdots\ 0\ 1\big) \left(\big[\mathbf{I}_{N+1} - \mathbf{X}(1) \big| \mathbf{1}_{N+1}\big] \right)^{-1} . \tag{19}$$

Note that $\mathbf{1}_{N+1}$ is the column vector consisting of $N+1$ ones and that $[\mathbf{A}|\mathbf{B}]$ equals the matrix $\mathbf{A}$ with the last column replaced by $\mathbf{B}$.

We now determine the unknown constant $U_0(0)$, the probability that the system is empty. In steady state, the average number of packets accepted by a system equals the average number of packets leaving that system. Class-1 traffic is not affected by class-2 traffic and consequently class-1 can be seen as an independent system. The mean number of class-1 packets accepted by the system during a slot is denoted by λ_1^e. A class-1 packet leaves the system when the class-1 system content is larger than 0. This leads to $\lambda_1^e = 1 - U_0(1)$. The same reasoning for the system containing both queues yields $\lambda_1^e + \lambda_2 = 1 - U_0(0)$. Bringing these two equations together provides

$$U_0(0) = U_0(1) - \lambda_2 \ . \tag{20}$$

The pgf of the class-2 system content can now be found from (15). The moment-generating property of pgfs enables determination of the moments of the system content. Application of matrix properties significantly expedites the computation of these moments by expressing them in terms of the derivates of the pgfs of the arrival process.

From the class-1 system content we easily obtain the class-1 packet loss ratio plr_1. This is the fraction of class-1 packets that arrive at the system but are dropped. We have

$$plr_1 = \frac{\lambda_1 - \lambda_1^e}{\lambda_1} = 1 - \frac{1 - U_0(1)}{\lambda_1} \ . \tag{21}$$

As the class-1 queue has finite capacity and excess packets are thus dropped the class-1 system is always stable. For the entire system to reach steady state it is imperative that the average number of class-2 packets that is served exceeds the average number of class-2 arrivals, or that $\lambda_2 < 1 - \lambda_1^e$. Notice that requiring that $U_0(0) > 0$ is an equivalent stability constraint.

4 Packet Delay

We tag an arbitrary class-i packet. Let the delay of the packet be denoted by d_i. The arrival slot of the packet is assumed to be slot k. As stated earlier, the class-1 packets are not affected by class-2 traffic. Consequently, the delay of a class-1 packet can easily be obtained from the system content using the distributional form of Little's Theorem [8]. For the pmf of the class-1 delay this leads to

$$d_1(n) = \frac{U_n(1)}{1 - U_0(1)}, \quad n = 1 \dots N \ . \tag{22}$$

For a class-2 packet the analysis is more elaborate. Some preliminary work is performed before we tackle the delay. We first determine the (remaining) class-1 busy period. Next, the extended service completion time of a class-2 packet is defined. We finally establish the number of class-i packets in the system at the end of slot k to be served before the tagged class-2 packet.

The remaining class-1 busy period in slot k, $r_{1,k}$, corresponds with the number of slots until the next slot with class-1 system content equal to 0. Recall that class-1 traffic is unaffected by class-2 traffic. Relating $r_{1,k}$ and $r_{1,k+1}$ and letting k go to infinity results in a system of equations for $R_{1,n}(z)$, the conditional pgf of the remaining class-1 busy period in steady state, at the beginning of a slot during a busy period, if the class-1 system content at the beginning of that slot equals n. We obtain

$$\begin{aligned} R_{1,n}(z) =& z \sum_{m=0}^{N-n} R_{1,n-1+m}(z) A_m(1) \\ &+ z R_{1,N}(z) A^*_{N-n+1}(1), \quad n = 1 \dots N \ . \end{aligned} \tag{23}$$

Note that $R_{1,0}(z) = 1$ as the class-1 busy period ends when the class-1 queue is empty. Again the pgfs $A^*_i(z)$ appear when the class-1 queue is completely filled. Let us define the $N \times N$ matrix

$$\mathbf{M} = \begin{pmatrix} A_1(1) & A_0(1) & 0 & 0 & \cdots & 0 \\ A_2(1) & A_1(1) & A_0(1) & 0 & \cdots & 0 \\ A_3(1) & A_2(1) & A_1(1) & A_0(1) & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ A_{N-1}(1) & A_{N-2}(1) & A_{N-3}(1) & A_{N-4}(1) & \cdots & A_0(1) \\ A^*_N(1) & A^*_{N-1}(1) & A^*_{N-2}(1) & A^*_{N-3}(1) & \cdots & A^*_1(1) \end{pmatrix} , \tag{24}$$

and the row vectors of N elements

$$\mathbf{L} = \begin{pmatrix} A_0(1) \\ 0 \\ \vdots \\ 0 \end{pmatrix}^T , \quad \hat{\mathbf{R}}(z) = \begin{pmatrix} R_{1,1}(z) \\ R_{1,2}(z) \\ \vdots \\ R_{1,N}(z) \end{pmatrix}^T . \tag{25}$$

In matrix notation the system of equations (23) leads to

$$\hat{\mathbf{R}}(z) = z\mathbf{L}\Big(\mathbf{I}_N - z\mathbf{M}\Big)^{-1} . \tag{26}$$

Note that the relation between $R_{1,0}(z)$ and $R_{1,1}(z)$ is expressed in $\mathbf{L}$. We have determined $\hat{\mathbf{R}}(z)$ and we now extend this vector with $R_{1,0}(z) = 1$ resulting in the row vector of $N + 1$ elements

$$\mathbf{R}(z) = \begin{pmatrix} R_{1,0}(z) \\ R_{1,1}(z) \\ \vdots \\ R_{1,N}(z) \end{pmatrix}^T = \begin{pmatrix} 1 \\ \hat{\mathbf{R}}(z) \end{pmatrix}^T . \tag{27}$$

Notice that a class-1 busy period is simply the remaining class-1 busy period in a random slot preceded by a slot with an empty class-1 system content at the beginning of the slot and a number of arrivals larger than 0. Thus we obtain the pgf of the steady-state class-1 busy period as

$$B_1(z) = \frac{\sum_{m=1}^{N-1} R_{1,m}(z)A_m(1) + R_{1,N}(z)A_N^*(1)}{1 - A_0(1)} . \tag{28}$$

The extended service completion time of a class-2 packet, denoted by t_2, starts at the slot where the packet starts service and lasts until the next slot wherein a class-2 packet can be serviced [10]. If no class-1 packets arrive during the service-slot of the packet, the server can handle another class-2 packet in the next slot. If there are class-1 arrivals, we have to wait for a class-1 busy period after the service-slot until the service of another class-2 packet can start. We can thus express the pgf of the extended service completion time in steady state as

$$T_2(z) = A_0(1)z + \big(1 - A_0(1)\big)B_1(z)z . \tag{29}$$

The number of class-i packets in the system at the end of slot k that have to be served before the tagged class-2 packet is denoted by $v_{i,k}$. Let $u_{i,k}^*$ denote the number of class-i packets that remain in the system during slot k. This equals the class-i system content at the beginning of slot k diminished by 1 if a class-i packet is in service during slot k. As all class-1 packets that arrive during slot k are to be served before the tagged packet it is clear that $v_{1,k} = u_{1,k}^* + a_{1,k}^e$. The class-2 packets that arrive during slot k but after the tagged packet are not to be served before it. Consequently $v_{2,k} = u_{2,k}^* + \hat{a}_{2,k}$ where $\hat{a}_{2,k}$ denotes the number of class-2 arrivals during slot k to be served before the tagged packet. We will now determine some corresponding pgfs.

Foremost we define the steady-state pgfs

$$U_i^*(z) = \lim_{k\to\infty} \mathrm{E}[z^{u_{2,k}^*}\mathbf{1}_{u_{1,k}^*=i}], \quad i = 0 \ldots N-1 . \tag{30}$$

Standard z-transform techniques lead to

$$\begin{aligned} U_0^*(z) =& U_0(0)\frac{z-1}{z} + \frac{U_0(z)}{z} + U_1(z) \ , \\ U_i^*(z) =& U_{i+1}(z), \quad i = 1 \ldots N-1 \ . \end{aligned} \tag{31}$$

The corresponding column vector of N elements is denoted by

$$\mathbf{U}^*(z) = \begin{pmatrix} U_0^*(z) \\ U_1^*(z) \\ \vdots \\ U_{N-1}^*(z) \end{pmatrix} \ . \tag{32}$$

Determination of the number of class-2 arrivals before the tagged packet is a bit more involved. If the arrivals of both classes are correlated it is clear that $a_{1,k}$ and $\hat{a}_{2,k}$ are correlated as well. We again define steady-state pgfs

$$\begin{aligned} \hat{A}_i(z) =& \lim_{k\to\infty} \mathrm{E}[z^{\hat{a}_{2,k}} \mathbf{1}_{a_{1,k}=i}] \ , \\ \hat{A}_i^*(z) =& \sum_{l=i}^{\infty} \hat{A}_l(z) \ . \end{aligned} \tag{33}$$

Taking into account that the tagged class-2 packet is more likely to arrive in a slot with more arrivals [11] the pmf of the number of class-1 and class-2 arrivals in the arrival slot of a tagged class-2 packet is given by

$$\tilde{a}(m,n) = \frac{n a(m,n)}{\lambda_2} \ . \tag{34}$$

The pmf of the number of class-2 arrivals before the tagged packet in a slot with m class-1 arrivals is given by

$$\hat{a}(m,n) = \sum_{l=n+1}^{\infty} \frac{\tilde{a}(m,l)}{l} = \sum_{l=n+1}^{\infty} \frac{a(m,l)}{\lambda_2} \ . \tag{35}$$

Now it is straightforward that

$$\hat{A}_i(z) = \sum_{n=0}^{\infty} \hat{a}(m,n) z^n = \frac{A_i(z) - A_i(1)}{\lambda_2 (z-1)} \ . \tag{36}$$

Analogously we find that

$$\hat{A}_i^*(z) = \frac{A_i^*(z) - A_i^*(1)}{\lambda_2 (z-1)} \ . \tag{37}$$

Let us then define the $(N+1) \times N$ matrix

$$\hat{\mathbf{A}}(z) = \begin{pmatrix} \hat{A}_0(z) & 0 & 0 & \cdots & 0 & 0 \\ \hat{A}_1(z) & \hat{A}_0(z) & 0 & \cdots & 0 & 0 \\ \hat{A}_2(z) & \hat{A}_1(z) & \hat{A}_0(z) & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \hat{A}_{N-2}(z) & \hat{A}_{N-3}(z) & \hat{A}_{N-4}(z) & \cdots & \hat{A}_0(z) & 0 \\ \hat{A}_{N-1}(z) & \hat{A}_{N-2}(z) & \hat{A}_{N-3}(z) & \cdots & \hat{A}_1(z) & \hat{A}_0(z) \\ \hat{A}^*_N(z) & \hat{A}^*_{N-1}(z) & \hat{A}^*_{N-2}(z) & \cdots & \hat{A}^*_2(z) & \hat{A}^*_1(z) \end{pmatrix} . \tag{38}$$

We can now finally describe the class-2 delay. The number of slots a class-2 packet spends in the system equals

$$d_2 = r_{1,k+1} + \sum_{i=1}^{v_{2,k}} t_2 + 1 \ . \tag{39}$$

Keeping in mind that $r_{1,k+1}$ is completely defined by $v_{1,k}$ and that the $u^*_{i,k}$ are independent of the $a_{i,k}$ we find that

$$\begin{aligned} D_2(z) =& \mathrm{E}[z^{d_2}] = \sum_{i=0}^{N} \mathrm{E}[z^{d_2} \mathbf{1}_{v_{1,k}=i}] \\ &= \sum_{i=0}^{N-1} z R_{1,i}(z) \sum_{j=0}^{i} \hat{A}_{i-j}(T_2(z)) U^*_j(T_2(z)) \\ &\quad + z R_{1,N}(z) \sum_{j=0}^{N-1} \hat{A}^*_{N-j}(T_2(z)) U^*_j(T_2(z)) \ . \end{aligned} \tag{40}$$

This can be equivalently expressed as

$$D_2(z) = z \mathbf{R}(z) \hat{\mathbf{A}}(T(z)) \mathbf{U}^*(T(z)) \ . \tag{41}$$

By taking proper derivatives, moments of the class-2 delay can be calculated.

5 Applications

With the formulas at hand we study an output-queueing switch with S inlets and S outlets and two types of traffic as in [2]. On each inlet a batch arrives according to a Bernouilli process with parameter μ_T. A batch contains b (fixed) packets of class 1 with probability μ_1/μ_T or b packets of class 2 with probability μ_2/μ_T (with $\mu_1 + \mu_2 = \mu_T$). The incoming packets are then routed uniformely to the outlets where they arrive at a queueing system as described in this paper.

Therefore all the outlets can be considered identical and analysis of one of them is sufficient. The arrival process at the queueing system can consequently be described by the pmf

$$a(bn, bm) = \frac{S!\left(\frac{\mu_1}{S}\right)^n \left(\frac{\mu_2}{S}\right)^m \left(1 - \frac{\mu_T}{S}\right)^{S-n-m}}{n!m!(S-n-m)!}, \quad n+m \leq S \ , \tag{42}$$

and by $a(p, q) = 0$, for other values of p and q. Obviously the number of arrivals of class-1 and class-2 are negatively correlated. For instance in a slot with x class-1 arrivals there can be no more than $Sb - x$ class-2 arrivals. For increasing values of S the correlation increases and for S going to infinity the numbers of arrivals of both types become uncorrelated.

We now study an 8×8 output-queueing switch. Assume $b = 4$ and $\mu_1 = \mu_2 = 0.1$ yielding $\lambda_1 = \lambda_2 = 0.4$. On average the system thus receives the same amount of packets of both classes. In Fig. 1 the mean and the standard deviation (jitter) of the delay of both classes are plotted versus the class-1 queue capacity N. We clearly see the effect of the priority scheduling. The low mean and standard deviation for the class-1 delay give us the performance required for real-time traffic at the cost of the class-2 performance measures. The values increase for increasing N, as the number of dropped class-1 packets decreases. For larger N the values clearly converge to the values corresponding with the infinite system [2], represented by the dashed lines. However, the convergence is rather slow, especially for the class-2 delay.

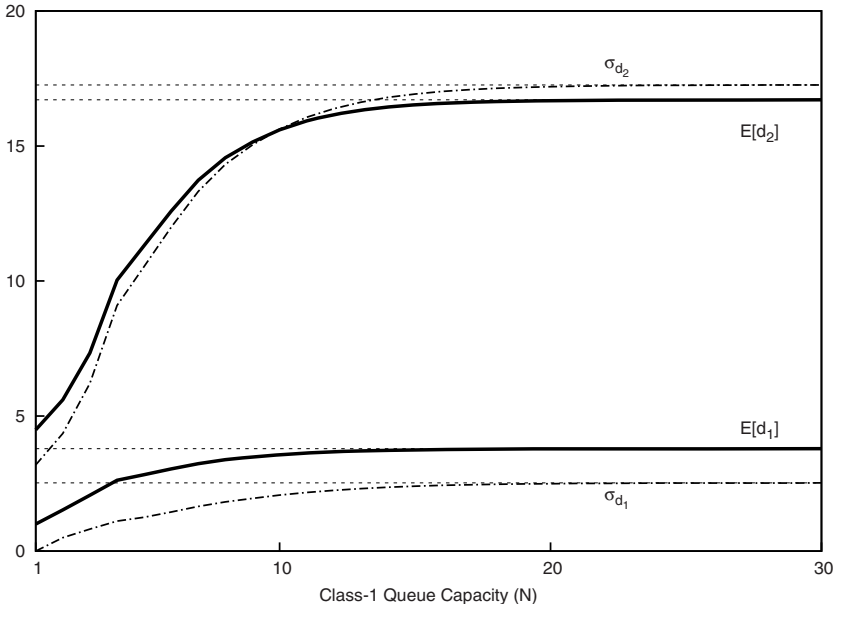

Fig. 1. Delays versus class-1 queue capacity

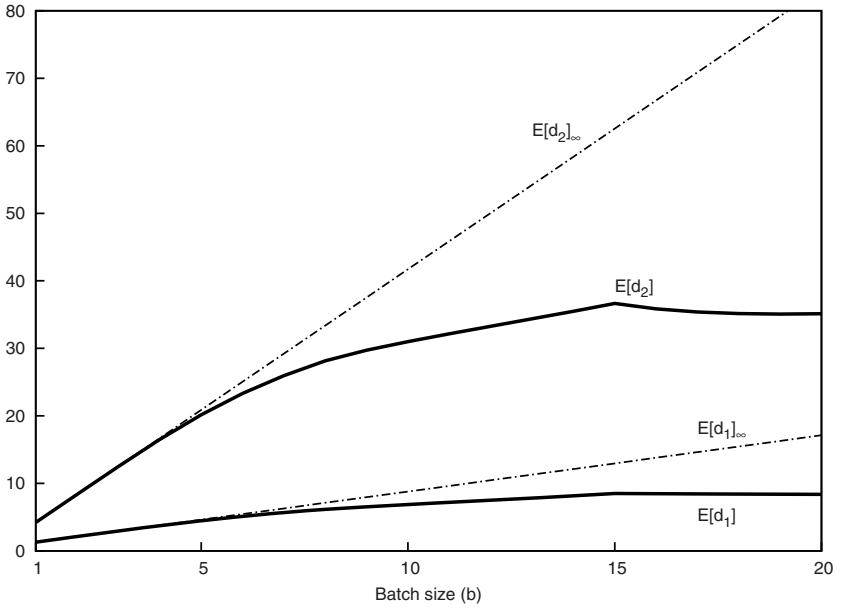

Fig. 2. Mean delays versus batch size

Now assume $N = 15$ and $\lambda_1 = \lambda_2 = 0.4$. We increase the batch size b while adjusting the μ_i accordingly in order to keep the λ_i constant. For increasing b the system thus receives the same amount of packets but the variance of the number of arrivals increases. In Fig. 2, we depict the mean delay of both classes versus the batch size b (as well as the mean delays of the infinite system). We clearly see that the delay increases and that the infinite system leads to inaccurate results when the variance in the arrival process increases. Since in practice arrival processes with high variance are very common, this proves that the infinite model can be imprecise. The decrease of the mean delays for $b > 15$ can be attributed to a high loss rate since for $b > 15$ the batch size exceeds the class-1 queue capacity.

In the $S \times S$ output-queueing switch only a moderate amount of (negative) correlation is present. In order to study the correlation between both classes profoundly, we end this section with the results for a very simple arrival process. A batch of class i arrives according to a Bernoulli distribution with parameter μ_i. A batch contains b (fixed) packets and thus $\lambda_i = b\mu_i$. The joint pmf is given by

$$\begin{aligned} a(0,0) =& 1 - \mu_1 - \mu_2 + c, \\ a(b,0) =& \mu_1 - c, \\ a(0,b) =& \mu_2 - c, \\ a(b,b) =& c. \end{aligned} \tag{43}$$

Notice that this arrival process allows the arrival of a batch of each class in a slot. The concurrence of the arrivals of both classes is controlled by the parameter c.

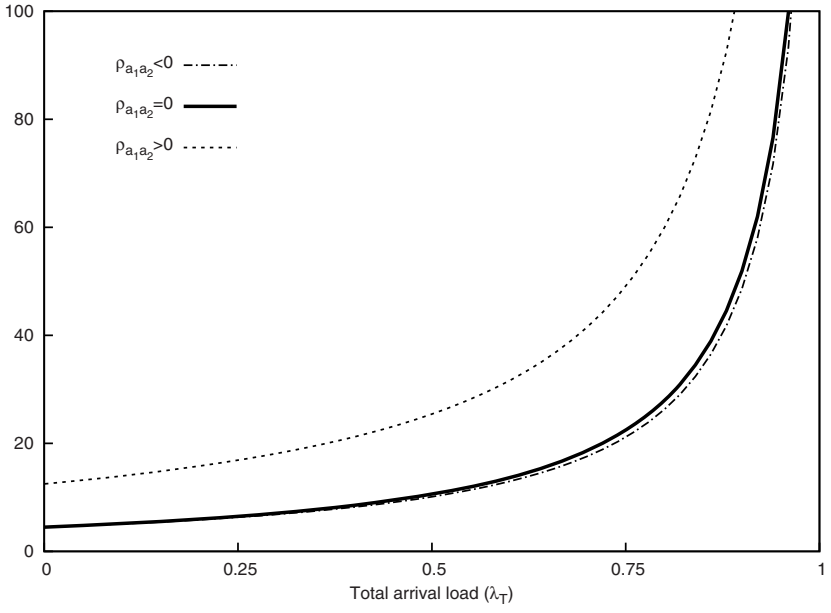

Fig. 3. Mean class-2 delay versus total load for various values of $\rho_{a_1 a_2}$

The correlation factor is given by

$$\rho_{a_1 a_2} = \frac{c - \mu_1 \mu_2}{\sqrt{\mu_1 \mu_2 (1 - \mu_1)(1 - \mu_2)}}. \tag{44}$$

By varying the value of c, while keeping the μ_i constant, we can alter the correlation between both classes. For $c = 0$ there are no slots in which a batch of each class arrives and thus the correlation is minimal ($\rho_{a_1 a_2} < 0$). For $c = \mu_1 \mu_2$ there is no correlation ($\rho_{a_1 a_2} = 0$), while for $c = \min(\mu_1, \mu_2)$ a batch of the class with the lowest arrival rate always arrives in a slot wherein a batch of the other class arrives, yielding (maximum) positive correlation ($\rho_{a_1 a_2} > 0$).

In Fig. 3 we depict the mean class-2 delay versus the total arrival load (λ_T) for the three values for c mentioned above. Assume $N = 15$, $b = 8$ and $\lambda_1 = \lambda_2$. The increase in mean delay between the uncorrelated case and the positively correlated case is remarkable, especially for higher values of λ_T. This follows from the fact that positive correlation between the arrivals of both classes increases the probability that a class-2 packet arrives in the same slot as a class-1 batch, its delay then more frequently includes service of an entire class-1 batch. For negative correlation the inverse effect is established but its influence is less noticeable in this example.

6 Conclusion

We have determined the probability mass functions of the high-priority (class-1) system content and delay and the probability generating functions of the

low-priority (class-2) system content and delay in a two-class priority queue with finite capacity for the high-priority packets. The class-1 packet loss ratio was also obtained. From these formulas it was shown that the infinite class-1 queue capacity approximation, that is frequently used, can yield inaccurate results. The presented model takes the exact class-1 queue capacity into account allowing the determination of precise values for the performance measures even when the class-1 queue capacity is small. In practice one needs to compromise between delay and allowed packet loss in order to determine a suitable class-1 queue capacity N. Once N is chosen the performance measures of both classes can be obtained as in this paper. It is also apparent that correlation between the arrivals of the different classes can have a huge impact on the performance measures and thus should not be considered negligible.

Acknowledgement. The second and third authors are Postdoctoral Fellows with the Fund for Scientific Research, Flanders (F.W.O.-Vlaanderen), Belgium.

References

1. Carpenter, B.E., Nichols, K.: Differentiated services in the Internet. Proceedings of the IEEE 90(9), 1479–1494 (2002)
2. Walraevens, J., Steyaert, B., Bruneel, H.: Performance analysis of a single-server ATM queue with a priority scheduling. Computers & Operations Research 30(12), 1807–1829 (2003)
3. Takine, T., Sengupta, B., Hasegawa, T.: An analysis of a discrete-time queue for broadband ISDN with priorities among traffic classes. IEEE Transactions on Communications 42(2-4), 1837–1845 (1994)
4. Takine, T.: A nonpreemptive priority MAP/G/1 queue with two classes of customers. Journal of Operations Research Japan 39(2), 266–290 (1996)
5. Van Velthoven, J., Van Houdt, B., Blondia, C.: The impact of buffer finiteness on the loss rate in a priority queueing system. In: Horváth, A., Telek, M. (eds.) EPEW 2006. LNCS, vol. 4054, pp. 211–225. Springer, Heidelberg (2006)
6. Mehmet Ali, M., Song, X.: A performance analysis of a discrete-time priority queueing system with correlated arrivals. Performance Evaluation 57(3), 307–339 (2004)
7. Sidi, M., Segall, A.: Structured priority queueing systems with applications to packet-radio networks. Performance Evaluation 3(4), 265–275 (1983)
8. Vinck, B., Bruneel, H.: Delay analysis for single server queues. Electronics Letters 32(9), 802–803 (1996)
9. Fiems, D., Steyaert, B., Bruneel, H.: Discrete-time queues with generally distributed service times and renewal-type server interruptions. Performance Evaluation 55(3-4), 277–298 (2004)
10. Fiems, D.: Analysis of discrete-time queueing systems with vacations. PhD thesis. Ghent University (2003)
11. Bruneel, H., Kim, B.G.: Discrete-time models for communication systems including ATM. Kluwer Academic Publishers, Dordrecht (2004)