PERGAMON

Performance analysis of a single-server ATM queue with a priority scheduling

Joris Walraevens[*], Bart Steyaert, Herwig Bruneel

*SMACS Research Group, Ghent University, Vakgroep TELIN (TW07), Sint-Pietersnieuwstraat 41,
B-9000 Gent, Belgium*

Abstract

In this paper, we consider a discrete-time queueing system with head-of-line priority. First, we will give some general results on a GI-1-1 queue with priority scheduling. In particular, we will derive expressions for the probability generating function of the system contents and the cell delay. Some performance measures (such as mean, variance and approximate tail distributions) of these quantities will be derived, and used to illustrate the impact and significance of priority scheduling in an ATM output queueing switch.

Scope and purpose

Queueing theory is an important subject in computers and operations research. Buffers/queues are used, e.g., in telecommunication networks, to store information that cannot be transmitted instantaneously. The study of the buffer behavior is important since network performance is directly related to it. Queues with a priority scheduling discipline are an important subject in queueing theory. As a result, these type of queues are thoroughly studied in the past, especially in continuous time. In discrete-time queueing models on the other hand, this type of queues is not as widely studied. Discrete-time queueing models are suitable for the performance evaluation of asynchronous transfer mode (ATM) switches. In ATM, different types of traffic need different quality of service (QoS) standards. The delay characteristics of delay-sensitive traffic (e.g., voice) are more stringent than those of delay-insensitive traffic (e.g., data). We can thus give priority to delay-sensitive traffic over delay-insensitive traffic, thus trying to reduce the delay of the delay-sensitive traffic. This paper studies the impact of a priority scheduling on the buffer characteristics.

[*] Corresponding author. Tel.: +32-9-264-8902; fax: +32-9-264-4295.

E-mail addresses: jw@telin.rug.ac.be (J. Walraevens), bs@telin.rug.ac.be (B. Steyaert), hb@telin.rug.ac.be (H. Bruneel).

1. Introduction

In recent years, there has been much interest in ATM as a promising technology for transport of high-bandwidth applications. Especially its well-defined QoS guarantee makes it extremely suitable for multimedia applications. Different types of traffic need different QoS standards. For real-time applications, it is important that mean delay and delay jitter are not too large, while for non-real-time applications, the cell loss ratio (CLR) is the restrictive quantity.

In general, one can distinguish two priority categories, which will be referred to as delay priority and loss priority. Delay priority scheduling tries to reduce the delay of delay-sensitive traffic (such as voice). This is done by using a more sophisticated type of scheduling than the simple FIFO scheduling. Priority is given to delay-sensitive traffic over delay-insensitive traffic. Several types of delay priority (or cell scheduling) schemes (such as weighted-round-robin (WRR), weighted-fair-queueing (WFQ)) have been proposed and analyzed for ATM applications, each with their own specific algorithmic and computational complexity (see e.g. [1] and the references therein). On the other hand, loss-priority schemes attempt to reduce the cell loss of loss-sensitive traffic (such as data). Again, various types of loss-priority (or cell discarding) strategies for ATM (such as push-out buffer (POB), partial buffer sharing (PBS)) have been presented in the literature (see e.g. [2]). An overview of both types of priority can be found in [3].

In this paper, we will focus on the effect of head-of-line (HOL) (or non-preemptive) delay-priority scheduling. We assume that delay-sensitive traffic has absolute priority over delay-insensitive traffic, i.e., when a server becomes idle, a cell of delay-sensitive traffic, when available, will always be scheduled next. This is the most drastic type of delay-priority scheduling, but also the easiest one to implement. In the existing literature, there have been a number of contributions with respect to this priority scheme. In [4–11], HOL-priority queues have been analyzed with a wide variety of arrival and service time distributions.

In this paper, we use an analysis based on generating functions for assessing the performance of ATM buffers with a priority scheduling discipline. From these generating functions, we can then easily calculate expressions for some interesting performance measures, such as mean value, variance and approximations for the tail distribution of the buffer contents and cell delay. These closed-form expressions require virtually no computational effort at all, and are well suited for evaluating the impact of the various system parameters on the overall performance. We will also show that our results can be applied to the case of an ATM output-queueing switch with HOL priority scheduling. There have been a number of contributions with respect to switches with output queueing, in the case of a single traffic type and a FIFO scheduling discipline, such as [12–14].

The contribution of this paper concerns the model that is considered, the solution technique that is used, as well as the results that are generated. First, as far as the model is concerned, the main difference with the articles involved with HOL-priority queues listed above is that, for the case of a multiclass output-queueing switch, the arrival processes of the different types of cells are not mutually independent. Therefore, the different classes cannot be analyzed separately (i.e., as a model with server interruptions for low-priority cells as demonstrated in Section 5), which complicates the analysis. Secondly, we want to show that a generating-functions solution method is extremely suitable for analyzing this type of buffers with a priority scheduling discipline, whereas existing models are mainly based on matrix-analytic methods. Finally, determining the tail behavior of the buffer contents and cell delay is one of the main contributions of the paper. Although these are important quantities

in the evaluation of QoS of high- and low-priority cell streams, this has received only few attention up to now. We will also show that the distribution of the buffer contents and cell delay of low-priority cells not necessarily has a geometric asymptotic behavior.

The outline is as follows. First, we consider a single queue with a general arrival distribution. In the following section, we will introduce the mathematical model. In Sections 3 and 4 we will analyze the steady-state system contents and cell delay. In Section 5, we discuss the results derived in the former sections and we calculate the moments of the system contents and cell delay in Section 6. We study the tail behavior of the system contents and cell delay in Section 7. We apply the obtained results to an output queueing switch with Bernoulli arrivals, and discuss the impact of a HOL priority scheduling discipline in Section 8. Some conclusions are formulated in Section 9.

2. Mathematical model

We consider a discrete-time single-server queueing system with infinite buffer space. Time is assumed to be slotted, where 1 slot equals the transmission time of a cell. There are two types of traffic arriving in the system, namely cells of class-1 and cells 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 cell arrivals are assumed to be i.i.d. from slot to slot and are characterized by the joint probability mass function (pmf) $a(m, n)$,

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

and joint probability generating function (pgf) $A(z_1, z_2)$,

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

Notice that the number of cell arrivals from different classes (within a slot) can be correlated. Further, we denote the total number of arriving cells during slot k by $a_{T,k} \triangleq a_{1,k} + a_{2,k}$ and its pgf is defined as $A_T(z) \triangleq E[z^{a_{T,k}}] = A(z, z)$. In the same way, we define the marginal pgf's of the number of 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 $\lambda_j = A_j'(1)$ and the total arrival rate by $\lambda_T = A_T'(1) = A_1'(1) + A_2'(1)$. The system has one server that provides the transmission of cells, at a rate of 1 cell per slot. We assume a stable system, i.e., $\lambda_T < 1$.

Newly arriving cells can enter service at the beginning of the slot following their arrival slot at the earliest. Class-1 cells are assumed to have priority over class-2 cells, and within one class the service discipline is FCFS. Due to the priority scheduling mechanism, it is as if class-1 cells are stored in front of class-2 cells in the queue. So, if there are any class-1 cells in the queue at the beginning of a slot, the one with the longest waiting time will be served next. If, on the other hand, no class-1 cells are present in the queue at the beginning of a particular slot, the class-2 cell with the longest waiting time, if any, will be served.

3. System contents

In this section, we concentrate on the effect of the HOL priority scheduling discipline on the probability generating function of the steady-state system contents, which represents the number of

cells in the buffer. This was already done in [6]—for a more general queueing system—but it is useful to give the analysis in our special case. We denote the system contents of class-j at the beginning of slot k by $u_{j,k}$ ($j = 1,2$) and the total system contents at the beginning of slot k by $u_{T,k}$. Furthermore, we denote the joint pgf of $u_{1,k}$ and $u_{2,k}$ by $U_k(z_1, z_2)$, i.e.,

$$U_k(z_1, z_2) \triangleq E[z_1^{u_{1,k}} z_2^{u_{2,k}}].$$

The system contents of both types of cells is characterized by the following system equations:

$$u_{1,k+1} = [u_{1,k} - 1]^+ + a_{1,k},$$

$$u_{2,k+1} = \begin{cases} [u_{2,k} - 1]^+ + a_{2,k} & \text{if } u_{1,k} = 0, \\ u_{2,k} + a_{2,k} & \text{if } u_{1,k} > 0, \end{cases}$$

where $[.]^+$ denotes the maximum of the argument and 0. The first equation follows from the observation that class-1 cells are not influenced by class-2 cells. A class-2 cell on the other hand can only be served, if there are no class-1 cells in the system. This leads to the second equation. Using these system equations, we can form the following relation between $U_{k+1}(z_1, z_2)$ and $U_k(z_1, z_2)$

$$U_{k+1}(z_1, z_2) = A(z_1, z_2) \frac{z_2 U_k(z_1, z_2) + (z_1 - z_2)U_k(0, z_2) + z_1(z_2 - 1)U_k(0,0)}{z_1 z_2}. \tag{1}$$

Since we are interested in the steady-state distribution of the system contents, we define $U(z_1, z_2)$ as

$$U(z_1, z_2) \triangleq \lim_{k \to \infty} U_k(z_1, z_2).$$

Applying this limit in Eq. (1), we find the following expression for $U(z_1, z_2)$:

$$U(z_1, z_2) = A(z_1, z_2) \frac{(z_1 - z_2)U(0, z_2) + z_1(z_2 - 1)U(0, 0)}{z_2(z_1 - A(z_1, z_2))}. \tag{2}$$

There are two quantities yet to be determined in the right-hand side of Eq. (2), namely the function $U(0, z_2)$ and the constant $U(0, 0)$. Applying Rouché's theorem, it can be proven that for a given value of z_2 ($|z_2| \leqslant 1$), the equation $z_1 = A(z_1, z_2)$ has one solution in the unit circle for z_1, which will be denoted by $Y(z_2)$ in the remainder, and which is implicitly defined by $Y(z) \triangleq A(Y(z), z)$. Since $Y(z_2)$ is a zero of the denominator of the right-hand side of (2) and since a generating function remains finite in the unit circle, $Y(z_2)$ must be a zero of the numerator too. We thus find

$$U(0, z_2) = U(0, 0) \frac{Y(z_2)(z_2 - 1)}{z_2 - Y(z_2)}.$$

Substituting this result in Eq. (2) yields

$$U(z_1, z_2) = U(0, 0) \frac{A(z_1, z_2)(z_2 - 1)}{z_2 - Y(z_2)} \frac{z_1 - Y(z_2)}{z_1 - A(z_1, z_2)}. \tag{3}$$

$U(0, 0)$ can be found by applying the normalization condition $U(1, 1) = 1$. Using de l' Hopital's rule gives the expected result for the probability of having an empty system: $U(0, 0) = 1 - \lambda_T$. From Eq. (3), we easily obtain an expression for the pgf $U_T(z)$ describing the total system

contents

$$U_T(z) \triangleq \lim_{k \to \infty} E[z^{u_{T,k}}] = U(z,z)$$

$$= (1 - \lambda_T) \frac{A_T(z)(z-1)}{z - A_T(z)}. \tag{4}$$

We can also calculate the pgf $U_j(z)$ $(j = 1, 2)$ of the system contents of class-j, namely

$$U_1(z) \triangleq \lim_{k \to \infty} E[z^{u_{1,k}}] = U(z, 1)$$

$$= (1 - \lambda_1) \frac{A_1(z)(z-1)}{z - A_1(z)}, \tag{5}$$

$$U_2(z) \triangleq \lim_{k \to \infty} E[z^{u_{2,k}}] = U(1, z)$$

$$= (1 - \lambda_T) \frac{A_2(z)(z-1)}{z - Y(z)} \frac{1 - Y(z)}{1 - A_2(z)}. \tag{6}$$

We will discuss these results in Section 5.

4. Cell delay

The cell delay is defined as the total amount of time that a cell spends in the system, i.e., the number of slots between the end of the cell's arrival slot and the end of its departure slot. In this section, we will derive expressions for the pgf's of the cell delay of both classes.

We can analyze the cell delay of class-1 cells as if they are the only type of cells in the system. This is e.g., done in [15] and the pgf of the cell delay of class-1 cells is given by

$$D_1(z) = \frac{1 - \lambda_1}{\lambda_1} \frac{z(A_1(z) - 1)}{z - A_1(z)}. \tag{7}$$

The analysis of the cell delay of a class-2 cell is more complicated. Consider a logical equivalent queueing system where all high-priority cells are stored in front of the class-2 cells, and let us tag an arbitrary class-2 cell that arrives in the system. The amount of time it spends in the system equals

$$d_2 = \sum_{j=1}^{[u_{T,k}-1]^+ + f_{2,k}} v_j^0 + 1, \tag{8}$$

where slot k is assumed to be the arrival slot of the tagged cell, $f_{2,k}$ is defined as the total number of cells that arrive during the arrival slot of the tagged cell, but which have to be served before it, and v_j^0 represents the number of slots it takes for the tagged cell to move one position ahead in the queue, e.g., from position j to position $j - 1$ (see Fig. 1). In case of FIFO scheduling, v_j^0 would equal 1. For HOL priority scheduling, this is not necessarily the case, since new class-1 cells can arrive while the tagged cell is waiting in the queue and these class-1 cells have to be served before the tagged cell. More specific, assume that the tagged cell is stored in the jth position in the queue
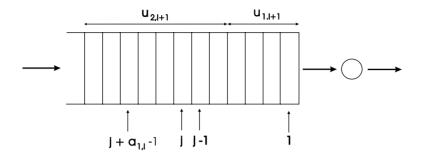
Fig. 1. The queueing system.

at the beginning of the lth slot ($0 < j \leqslant [u_{T,k} - 1]^+ + f_{2,k}$). If no class-1 cells arrive during slot l, v_j^0 equals 1. If $a_{1,l}$ (> 0) class-1 cells arrive during this slot on the other hand, the tagged cell will move back to position $j + a_{1,l} - 1$ in the queue at the beginning of slot $l + 1$, since these class-1 cells have to be served before all class-2 cells, and thus before the tagged one (Fig. 1). If we then define $v_{j,i}^1$ ($j \leqslant i \leqslant j + a_{1,l} - 1$) as the number of slots it takes the tagged cell to go from position i to position $i - 1$, it is clear that v_j^0 can be calculated as follows:

$$v_j^0 = \sum_{i=0}^{a_{1,l}-1} v_{j,j+i}^1 + 1. \tag{9}$$

Now, one can easily see that all v_j^0 and $v_{j,i}^1$ form a set of mutually independent random variables since they depend on the number of class-1 cell arrivals during different slots. From a stochastic point of view, these are i.i.d. variables and, as a result, are characterized by the same pgf $V(z)$. From Eq. (9), it can be seen that $V(z)$ satisfies

$$V(z) = zA_1(V(z)). \tag{10}$$

Furthermore, $f_{2,k}$ is the sum of all the class-1 cells that arrive during the same slot as the tagged one, and of the class-2 cells that have arrived before it during its arrival slot. The pgf of $f_{2,k}$ can be calculated taking into account that an arbitrary tagged cell is more likely to arrive in a larger bulk (e.g. [15]), yielding

$$F_2(z) = \frac{A_T(z) - A_1(z)}{\lambda_2(z - 1)}. \tag{11}$$

Using Eqs. (4) and (11) in the z-transform of Eq. (8) eventually gives us the steady-state pgf of d_2, i.e.,

$$D_2(z) = \frac{1 - \lambda_T}{\lambda_2} \frac{z(A_T(V(z)) - A_1(V(z)))}{V(z) - A_T(V(z))}, \tag{12}$$

where $V(z)$ is implicitly determined by Eq. (10).

5. Discussion of the results and special relations

In this section, we will discuss some of the results from the former sections. First, we notice that the pgf of the total system contents (Eq. (4)) is the same as for a single class system with an

identical cell arrival process described by $A_T(z)$. Indeed, since the service time is deterministic and equal to 1 slot for the two classes, the scheduling has no impact on the total system contents.

Second, we see that the system contents of class-1 cells (Eq. (5)) is not influenced by class-2 cells and furthermore that its pgf has the same structure as $U_T(z)$. This is of course due to the HOL priority scheduling. For class-1 cells, it seems as if no class-2 cells are present in the system. Consequently, since the scheduling is FIFO within class-1, $U_1(z)$ and $D_1(z)$ fulfill the following relation (see [16]):

$$U_1(z) = 1 - \lambda_1 + \lambda_1 D_1(z).$$

It is easily verified that indeed (5) and (7) satisfy this equation.

In the special case that the number of arrivals of class-1 and class-2 cells are uncorrelated, i.e., $A(z_1, z_2) = A_1(z_1)A_2(z_2)$, we can calculate the system contents of class-2 cells in an alternative way. Since class-2 cells can only be served when there are no class-1 cells in the system, we can model the system, with respect to class-2 cells, in terms of a system with server interruptions. The server is blocked for class-2 cells if there are class-1 cells waiting to be sent, and it is available if there are none. We can then calculate the pgf of the duration of busy and idle period of class-1 cells, i.e., the time period during which there are class-1 cells in the system (i.e., $u_1 > 0$) and the time period during which there are no such cells (i.e., $u_1 = 0$), respectively. It is easily verified that the duration of the idle period is geometrically distributed, i.e., its pgf is given by

$$I(z) = \frac{(1 - A_1(0))z}{1 - A_1(0)z}. \tag{13}$$

The calculation of the busy period is a bit more involved, and can be found in [15] for a general service time distribution. In case of deterministic service times of one slot, it is implicitly given by the following formula:

$$B(z) = \frac{A_1(z((1 - A_1(0))B(z) + A_1(0))) - A_1(0)}{1 - A_1(0)}. \tag{14}$$

Note that the lengths of consecutive busy and idle periods are statistically independent. It is clear that when the system is busy with respect to class-1 cells, it is blocked for class-2 cells. Therefore, with respect to class-2 cells, the system can be modelled as a single-server buffer with server interruptions, for which the lengths of consecutive available and blocked periods are i.i.d. and their respective pgf's are given by Eqs. (13) and (14). Such a queueing system has already been analyzed in [17]. Translating the results from this analysis to our case, the pgf of the system contents of class-2 cells becomes

$$U_2(z) = (1 - \lambda_T) \frac{A_2(z)(z - 1)}{z - X(z)} \frac{1 - X(z)}{1 - A_2(z)} \tag{15}$$

with

$$X(z) = A_1(X(z))A_2(z).$$

Eqs. (6) and (15) lead to the same result for $U_2(z)$, when $X(z) = Y(z)$. This is the case when the number of class-1 and class-2 arrivals during a slot are uncorrelated.

6. Calculation of moments

The functions $Y(z)$ and $V(z)$, defined in Sections 3 and 4, 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, the first derivatives are given by

$$Y'(1) = \frac{\lambda_2}{1 - \lambda_1}, \quad V'(1) = \frac{1}{1 - \lambda_1}.$$

Let us define λ_{ij} as

$$\lambda_{ij} \triangleq \left. \frac{\partial^2 A(z_1, z_2)}{\partial z_i \partial z_j} \right|_{z_1 = z_2 = 1}$$

with $i, j = 1, 2$. Now we can calculate the mean values of the system contents and cell delay of both classes by taking the first derivative of the respective pgf's for $z = 1$. We find

$$E[u_T] = \lambda_T + \frac{A_T''(1)}{2(1 - \lambda_T)} \tag{16}$$

for the mean value of total system contents,

$$E[u_1] = \lambda_1 + \frac{\lambda_{11}}{2(1 - \lambda_1)} \tag{17}$$

for the mean system contents of class-1 cells and

$$E[u_2] = \lambda_2 + \frac{2\lambda_{12} + \lambda_{22}}{2(1 - \lambda_T)} + \frac{\lambda_2 \lambda_{11}}{2(1 - \lambda_T)(1 - \lambda_1)} \tag{18}$$

for the mean system contents of class-2 cells. It is easily verified that Eqs. (16)–(18) satisfy $E[u_T] = E[u_1] + E[u_2]$.

Furthermore, from Eqs. (7) and (12), we derive the following expressions:

$$E[d_1] = 1 + \frac{\lambda_{11}}{2\lambda_1(1 - \lambda_1)} \tag{19}$$

and

$$E[d_2] = 1 + \frac{2\lambda_{12} + \lambda_{22}}{2\lambda_2(1 - \lambda_T)} + \frac{\lambda_{11}}{2(1 - \lambda_T)(1 - \lambda_1)} \tag{20}$$

for the mean cell delay of a class-1 and a class-2 cell, respectively. We can see from Eqs. (17)–(20) that Little's law $E[u_j] = \lambda_j E[d_j]$ ($j = 1, 2$) is fulfilled for both classes, as expected.

In a similar way, expressions for the variance of system contents and cell delay and some interesting correlation coefficients can be calculated by taking the appropriate derivatives of the respective generating functions as well. The expressions are nevertheless too exhaustive, but we will show them in some figures in Section 8.

7. Tail behavior

Not only the moments of the system contents and cell delay are important, but also, and especially, the tail distribution of these quantities, which are often used to impose statistical bounds on the guaranteed QoS for both classes.

From the generating functions of the total system contents, and of the system contents and cell delay of class-1 and class-2 cells derived in Sections 3 and 4, approximations of the tail probabilities can be derived using complex contour integration and residue theory. The procedure to find the corresponding probability mass function of a pgf, frequently used in the following of this section, is generally described in Appendix A.

In order to determine the asymptotic behavior of the tail distribution, the dominant singularity of the respective generating functions is important. In e.g. [18] (wherein a single-class ATM queue with a FIFO scheduling discipline is analyzed), it is proven that the dominant singularity lies on the positive real axis and is larger than 1.

First we concentrate on the total system contents. Provided that the pgf $A_T(z)$ exhibits no long-tail behavior, which is assumed to be the case here, the dominant singularity z_T of $U_T(z)$ is a zero of $z - A_T(z)$ and this singularity is a single pole. In the neighborhood of this pole, we can approximate $U_T(z)$ by

$$U_T(z) \approx \frac{K_T}{z_T - z}, \tag{21}$$

where K_T can be found by substituting $z = z_T$ in (21). Using residue theory, the tail probability is easily found to yield

$$\text{Prob}[u_T = n] \approx (1 - \lambda_T) \frac{z_T - 1}{A_T'(z_T) - 1} z_T^{-n} \tag{22}$$

for large enough n. The system contents of class-1 cells has an identical tail behavior:

$$\text{Prob}[u_1 = n] \approx (1 - \lambda_1) \frac{z_H - 1}{A_1'(z_H) - 1} z_H^{-n} \tag{23}$$

for large enough n, with z_H the dominant singularity on the positive real axis of $U_1(z)$, i.e., z_H is a zero of $z - A_1(z)$.

The tail behavior of the system contents of class-2 cells is a bit more involved, since it is not a priori clear what the dominant singularity is of $U_2(z)$. This is due to the occurrence of the function $Y(z)$ in (6), which is only implicitly defined. First we take a closer look at that function $Y(z)$. The first derivative of $Y(z)$ is given by

$$Y'(z) = \frac{A^{(2)}(Y(z), z)}{1 - A^{(1)}(Y(z), z)} \tag{24}$$

with $A^{(j)}(z_1, z_2) \triangleq \partial A(z_1, z_2) / \partial z_j$ ($j = 1, 2$). Consequently, $Y(z)$ has a singularity, denoted as z_B, where the denominator of $Y'(z)$ becomes 0, i.e., $A^{(1)}(Y(z_B), z_B) = 1$. Since $Y(z)$ remains finite in the neighborhood of z_B, this singularity is not a simple pole. Applying the results from [19] one can

show that in the neighborhood of z_B, $Y(z)$ is approximately given by

$$Y(z) \approx Y(z_B) - K_Y \sqrt{z_B - z} \tag{25}$$

with

$$K_Y = \sqrt{\frac{2A^{(2)}(Y(z_B), z_B)}{A^{(11)}(Y(z_B), z_B)}},$$

which can be found by taking the limit $z \to z_B$ of (25). $A^{(ij)}(z_1, z_2)$ is defined as $\partial^2 A(z_1, z_2)/\partial z_i \partial z_j$ (for $i, j = 1, 2$). From Eq. (25) it becomes obvious that z_B is a square-root branch point of $Y(z)$. $Y(z)$ has thus two real solutions when $z < z_B$ (the solution we are interested in is the one where $Y(z) < 1$, if $z < 1$), which coincide at z_B, and has no real solution when $z > z_B$. z_B is then of course also a branch point of $U_2(z)$. A second potential singularity z_L of $U_2(z)$ on the real axis is given by the positive zero of the denominator $z - Y(z)$, and is easily proven to be equal to z_T, if z_L exists.

The tail behavior of the system contents of class-2 cells is thus characterized by z_T or z_B, depending on which is the dominant (i.e., smallest) singularity. Three types of tail behavior may thus occur, namely when $z_L = z_T < z_B$, $z_L = z_T = z_B$ and z_L does not exist. In those three cases, $U_2(z)$ can be approximated in the neighborhood of its dominant singularity by

$$U_2(z) \approx \begin{cases} \dfrac{K_2^{(1)}}{z_T - z} & \text{if } z_L = z_T < z_B, \\[2ex] \dfrac{K_2^{(2)}}{\sqrt{z_B - z}} & \text{if } z_L = z_T = z_B, \\[2ex] U_2(z_B) - K_2^{(3)}\sqrt{z_B - z} & \text{if } z_L \text{ does not exist,} \end{cases}$$

where the constants $K_2^{(i)}$ ($i = 1, 2, 3$) can be found by investigating the behavior of $U_2(z)$ in the neighborhood of this dominant singularity. Using residue theory, we find the tail probabilities for the three possible cases:

$$\text{Prob}[u_2 = n] \approx \begin{cases} (1 - \lambda_T)\dfrac{A_2(z_T)(z_T - 1)^2}{z_T(A_2(z_T) - 1)(Y'(z_T) - 1)}z_T^{-n} \\[2ex] \dfrac{1 - \lambda_T}{K_Y}\sqrt{\dfrac{1}{z_B\pi}}\dfrac{A_2(z_B)(z_B - 1)^2}{A_2(z_B) - 1}n^{-1/2}z_B^{-n} \\[2ex] \dfrac{(1 - \lambda_T)K_Y}{2}\sqrt{\dfrac{z_B}{\pi}}\dfrac{A_2(z_B)(z_B - 1)^2}{(A_2(z_B) - 1)(z_B - Y(z_B))^2}n^{-3/2}z_B^{-n} \end{cases} \tag{26}$$

for large enough n, if $z_L = z_T < z_B$, if $z_L = z_T = z_B$, and if z_L does not exist, respectively. The first expression constitutes a typical geometric tail behavior, the third expression is a typical non-geometric tail behavior and the second expression gives a transition between geometric and non-geometric tail behavior. The latter two expressions are found from the approximations of the generating functions by using the theorem from Appendix B (which is a theorem stated in [20]).

Let us now consider the cell delay. The dominant singularity of $D_1(z)$ is the same as the one of $U_1(z)$, and we can thus approximate the tail behavior of the delay of class-1

cells by

$$\text{Prob}[d_1 = n] \approx \frac{(1 - \lambda_1)}{\lambda_1} \frac{z_H - 1}{A_1'(z_H) - 1} z_H^{-n} \tag{27}$$

for large enough n. The tail behavior of the delay of class-2 cells is again a bit more involved because of the appearance of the function $V(z)$ in (12), which is only implicitly known. The first derivative of $V(z)$ is given by

$$V'(z) = \frac{A_1(V(z))}{1 - zA_1'(V(z))}, \tag{28}$$

which, similar as before, indicates that $V(z)$ also has a square root branch point $\hat{z}_B$, with $\hat{z}_B A_1'(V(\hat{z}_B)) = 1$. In the neighborhood of $\hat{z}_B$, $V(z)$ is approximately given by

$$V(z) \approx V(\hat{z}_B) - K_V \sqrt{\hat{z}_B - z} \tag{29}$$

with

$$K_V = \sqrt{\frac{2A_1(V(\hat{z}_B))}{\hat{z}_B A_1''(V(\hat{z}_B))}}.$$

A second singularity of $D_2(z)$ is given by the dominant zero $\hat{z}_L$ of $V(z) - A_T(V(z))$ on the real axis and is easily proven to equal $z_T/A_1(z_T)$, if $\hat{z}_L$ exists.

So, $D_2(z)$ can be approximated in the neighborhood of his dominant singularity by

$$D_2(z) \approx \begin{cases} \dfrac{\hat{K}_2^{(1)}}{\dfrac{z_T}{A_1(z_T)} - z} & \text{if } \hat{z}_L = \dfrac{z_T}{A_1(z_T)} < \hat{z}_B, \\[4mm] \dfrac{\hat{K}_2^{(2)}}{\sqrt{\hat{z}_B - z}} & \text{if } \hat{z}_L = \dfrac{z_T}{A_1(z_T)} = \hat{z}_B, \\[4mm] D_2(\hat{z}_B) - \hat{K}_2^{(3)}\sqrt{\hat{z}_B - z} & \text{if } \hat{z}_L \text{ does not exist,} \end{cases}$$

where the constants $\hat{K}_2^{(i)}$ ($i = 1, 2, 3$) can be found by investigating $D_2(z)$ in the neighborhood of its dominant singularity. By using residue theory once again, the asymptotic behavior of $D_2(z)$ is given by

$$\text{Prob}[d_2 = n] \approx \begin{cases} \dfrac{1 - \lambda_T}{\lambda_2} \dfrac{(z_T - A_1(z_T))(A_1(z_T) - z_T A_1'(z_T))}{(A_1(z_T))^2 (A_T'(z_T) - 1)} \left(\dfrac{z_T}{A_1(z_T)} \right)^{-n} \\[4mm] \dfrac{1 - \lambda_T}{\lambda_2 K_V \sqrt{\hat{z}_B \pi}} \dfrac{\hat{z}_B A_T(V(\hat{z}_B)) - V(\hat{z}_B)}{A_T'(V(\hat{z}_B)) - 1} n^{-1/2} \hat{z}_B^{-n} \\[4mm] \dfrac{(1 - \lambda_T) K_V}{2\lambda_2 \sqrt{\pi/\hat{z}_B}} \dfrac{(\hat{z}_B - 1)(V(\hat{z}_B) A_T'(V(\hat{z}_B)) - A_T(V(\hat{z}_B)))}{(V(\hat{z}_B) - A_T(V(\hat{z}_B)))^2} n^{-3/2} \hat{z}_B^{-n}, \end{cases} \tag{30}$$

if $\hat{z}_L = z_T/A_1(z_T) < \hat{z}_B$, if $\hat{z}_L = z_T/A_1(z_T) = \hat{z}_B$, and if $\hat{z}_L$ does not exist, respectively. The first expression has a typical geometric tail behavior, the third expression has a typical non-geometric

tail behavior and the second expression gives a transition between geometric and non-geometric tail behavior.

A quantity of practical interest is the probability that a cell has a delay that exceeds a bound D. We find

$$\text{Prob}[d_1 > D] \approx \frac{\text{Prob}[d_1 = D + 1]\hat{z}_H}{\hat{z}_H - 1} \tag{31}$$

for the probability that the delay of a class-1 cell is larger than a bound D. This can be found by summing Eq. (27) for all appropriate values of n. Analogously, we can calculate the probability that a class-2 cell exceeds a bound D by summing Eq. (30) for the appropriate values of n. We find

$$\text{Prob}[d_2 > D] \approx \begin{cases} \dfrac{\text{Prob}[d_2 = D + 1]\hat{z}_L}{\hat{z}_L - 1} & \text{if } \hat{z}_L = \dfrac{z_T}{A_1(z_T)} < \hat{z}_B, \\[3mm] \dfrac{\text{Prob}[d_2 = D + 1]\hat{z}_B}{\hat{z}_B - 1} & \text{if } \hat{z}_L = \dfrac{z_T}{A_1(z_T)} = \hat{z}_B, \\[3mm] \dfrac{\text{Prob}[d_2 = D + 1]\hat{z}_B}{\hat{z}_B - 1} & \text{if } \hat{z}_L \text{ does not exist,} \end{cases}$$

where we used the approximation that $\sum_{n=D+1}^{\infty} n^{-a} z^{-n} \approx (D+1)^{-a} \sum_{n=D+1}^{\infty} z^{-n}$, with $a = \frac{1}{2}$ or $\frac{3}{2}$ and which holds for large enough D. Some similar expressions can be found for the probability that the system contents exceeds a certain bound.

Since the results obtained in this section are approximate (due to the dominant pole approximative method), the question remains if the expressions are accurate. From the analysis in [18], it follows that the approximation of the tail probabilities, obtained through the dominant pole method, are better when the dominant pole is more dominant (i.e., the higher the moduli of the other poles, the better the quality of the approximation) and when we go further in the tail of the distribution (i.e., when coefficient n in (22), (23), (26), (27) and (30) is higher). We will show in Section 8 that the approximate results for the tail probabilities obtained in this section are satisfactory.

8. Application

In this section, we apply our results from the former sections to an ATM output-queueing switch. We consider a non-blocking output-queueing switch with N inlets and N outlets (Fig. 2). We assume two types of traffic. Traffic of class-1 is delay sensitive (for instance voice) and traffic of class-2 is assumed to be delay insensitive (for instance data). We investigate the effect of HOL priority scheduling, as presented in the former of this paper.

The cell arrivals on each inlet are assumed to be i.i.d., and generated by Bernoulli process with arrival rate λ_T. An arriving cell is assumed to be of class-j with probability λ_j/λ_T ($j = 1, 2$) ($\lambda_1 + \lambda_2 = \lambda_T$). The incoming cells are then routed to the output queue corresponding to their destination, in an independent and uniform way. Therefore, the output queues behave identically and we can concentrate on the analysis of 1 output queue. In view of the previous, the arrivals of both types of cells to an output queue are generated according to a two-dimensional binomial process. It is fully
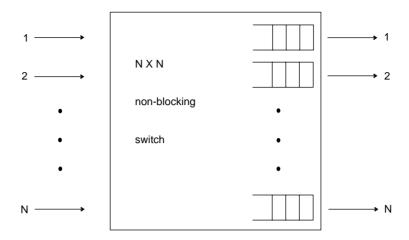
Fig. 2. An N × N output queueing switch.

characterized by the following joint pgf:

$$A(z_1, z_2) = \left(1 - \frac{\lambda_1}{N}(1 - z_1) - \frac{\lambda_2}{N}(1 - z_2) \right)^N. \tag{32}$$

Obviously, the number of class-1 and class-2 arrivals at an output queue during a slot are correlated. This is simply demonstrated by the following observation: when m class-1 cells arrive at the tagged queue during a slot ($0 \leqslant m \leqslant N$), the maximum number of class-2 arrivals during the same slot is limited by $N - m$. We note that for N going to infinity, the above expression becomes a product of two generating functions of Poisson distributions with means λ_1 and λ_2, respectively, and as a result, the arrival process becomes uncorrelated for both classes. In the following, we will investigate the effect of priority scheduling on some performance measures, such as mean value and variance of system contents and cell delay. We will, when possible, compare with a FIFO scheduling discipline to show the advantages and disadvantages of a priority scheduling discipline. In the remaining of this section, we assume a 16 × 16 switch ($N = 16$). We define α as the fraction of class-1 arrivals in the overall traffic mix (i.e., $\alpha = \lambda_1/\lambda_T$).

In Figs. 3 and 4, mean value and variance of the system contents of class-1 and class-2 cells is shown as a function of the total arrival rate, when $\alpha = 0.25$, 0.5 and 0.75, respectively. We have also shown the mean value and variance of the system contents for $\alpha = 0.5$ when a FIFO scheduling discipline is applied. These can be easily calculated because—in the special case of the arrival process characterized by (32)—the joint pgf of the system contents of both classes is given by $U_T(\alpha z_1 + (1 - \alpha)z_2)$ when a FIFO scheduling discipline is applied. One can easily see the influence of priority scheduling: the mean, as well as the variance of the number of class-1 cells in the system is severely reduced by the HOL priority scheduling; the opposite holds for class-2 cells. In addition, it also becomes apparent that increasing the fraction of high-priority cells in the overall mix increases the amount of high-priority traffic while decreasing the amount of low-priority traffic in the buffer. Finally, it is also clear that the impact of priority scheduling is more important if the load is high.
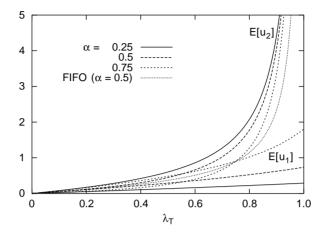
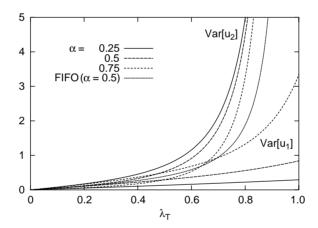Fig. 3. Mean value of system contents versus the total arrival rate.

Fig. 4. Variance of system contents versus the total arrival rate.

In Fig. 5, the correlation coefficient $\rho_{u_1 u_2}$, which quantifies the correlation between the number of class-1 and class-2 cells in the system during a slot, is shown as a function of the total arrival rate for $\alpha = 0.25, 0.5$ and 0.75. We see that $\rho_{u_1 u_2}$ increases when the fraction of class-1 cells increases (for a given total load). This can easily be understood by the priority scheduling. The influence of class-1 cells on class-2 cells will become more important, when the fraction of class-1 cells increases. A second observation is that $\rho_{u_1 u_2}$ is (slightly) negative when the total load is small, but becomes positive when the total load is large. The reason for this are two counteracting mechanisms. The first one is the switch structure. When more class-1 cells arrive at the switch, there will be less class-2 cells arriving at the same time (since the amount of inlets is limited), and vice versa. This negative correlation between cell arrivals of the two priority classes during a slot shows for small values of λ_T. For these parameter values, there is virtually no queueing and the buffer behavior is mainly determined by the number of arrivals during a single slot. The second influence is priority
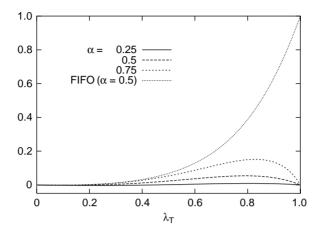
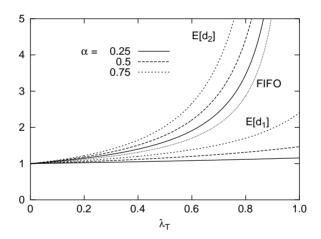Fig. 5. Correlation of system contents versus the total arrival rate.

Fig. 6. Mean value of cell delays versus the total arrival rate.

scheduling. As λ_T (and λ_1) further increases, more and more cells are being queued, and the presence of high-priority cells starts to seriously hinder the transmission of low-priority cells, thereby leading to a positive correlation between u_1 and u_2. Finally, when λ_T approaches 1, the total system contents (and the number of class-2 cells) approaches infinity, due to the system becoming unstable. As a result $\rho_{u_1 u_2}$ approaches 0. We have also shown the correlation coefficient for $\alpha = 0.5$, when a FIFO scheduling is applied. We see that the correlation coefficient in this case is always larger than when a priority scheduling discipline is applied. Since the system contents of both classes becomes infinite when λ_T approaches 1, $\rho_{u_1 u_2}$ approaches 1.

Figs. 6 and 7 show the mean value and the variance of the cell delay as a function of the total load for $\alpha = 0.25$, 0.5 and 0.75. In order to compare with FIFO scheduling, we have also shown the mean value and variance of the cell delay in that case. The cell delay is then of course the same
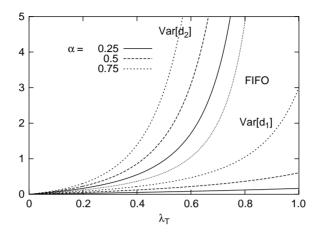
Fig. 7. Variance of cell delays versus the total arrival rate.

for class-1 and class-2 cells (independent of α), and can thus be calculated as if there is only one class arriving according to an arrival process with pgf $A(z,z)$. This has already been analyzed, e.g., in [13] for the special case of a multiserver output-queueing switch. We observe that the influence of HOL priority scheduling is quite large. Mean delay and delay jitter of class-1 cells reduces considerably compared to FIFO scheduling. The price to pay is of course a bigger mean delay and delay jitter for class-2 cells. If this kind of traffic is not delay sensitive, this is not too big a problem. Nevertheless, in a buffer with limited storage, an appropriate loss-priority scheme will have to be applied in order to avoid excessive cell loss of class-2 cells. Also note that if follows from these figures that increasing the fraction of high-priority cells in the overall traffic mix, increases the delay of high- and low-priority cells.

We have shown in Section 7, that the tails of class-2 system contents and cell delay can have three types of behavior, depending on which singularity of the respective pgf's is dominant. In case of the output queueing switch considered in this section, Fig. 8 shows for which combination of class-1 and class-2 arrival rates the transition type behavior occurs for the system contents and cell delay, i.e., for which combination of arrival rates the regular pole and the branch point coincide. Above the curves, the tail behavior is geometric, while below the curves the tail behavior is typically non-geometric. Note that in the area above the linear line (defined by $\lambda_1 + \lambda_2 = 1$) in Fig. 8, the total load is larger than 1, and as a result, the system becomes unstable.

Figs. 9 and 10 show the tail behavior of the system contents and cell delay of class-1 and class-2 cells if $\lambda_1 = 0.4$ and $\lambda_2 = 0.1$ (non-geometric behavior), approximately 0.21 (transition type behavior) and 0.4 (geometric behavior), respectively. Tail behavior of system contents and cell delay of class-1 cells is of course the same for the three cases, since the arrival process of class-1 cells does not change. We have our approximations also compared with simulation results (marks in the figures). The figures show that the approximations for the class-1, the geometric and transition type tail behavior of system contents and delay is very good in these cases. The approximations of the tails of the non-geometric case are not as good, but still satisfactory. The approximations of the tails are not good in all cases though, as is shown in Figs. 11 and 12. In these figures, the tail probabilities of the system contents and cell delay of class-2 cells is shown, with the parameters of
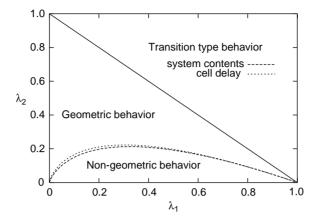
Fig. 8. Regions for tail behavior as a function of the arrival rates of both classes.
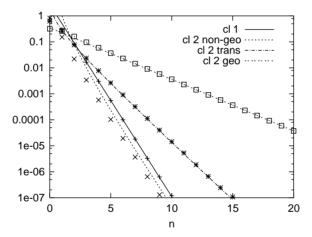

Fig. 9. Tail behavior of the high- and low-priority system contents for some combinations of class-1 and class-2 arrival rates.

the transition type behavior of the previous examples ($\lambda_1 = 0.4$ and $\lambda_2 = 0.21$). We have shown the three types of behavior, i.e., the tail behavior just before you have the transition from non-geometrical to geometrical tail behavior, the transition type tail behavior itself, and the tail behavior just after the transition. These tail probabilities should be very near to each other, but the figures show this is not the case. The incorrectness of the geometrical and non-geometrical approximations is due to the single-pole approximations. If both singularities lie near to each other, which is the case near the transition from non-geometric to geometric behavior, a single-pole approximation is clearly not good enough. More accurate approximations are necessary in those cases, but this lies outside the scope of this paper.

To conclude this section, we analyze the following case study. Consider two traffic classes generating cells that arrive in a common multiplexer buffer where they are temporarily stored before
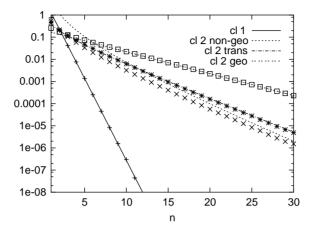
Fig. 10. Tail behavior of the high- and low-priority cell delay for some combinations of class-1 and class-2 arrival rates.
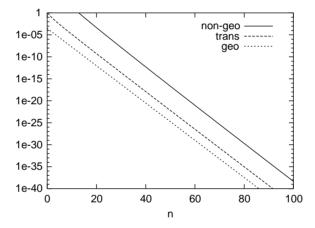

Fig. 11. Tail behavior of the low-priority system contents near the transition from non-geometrical to geometrical.

transmission. The cell arrival process of both classes is described by a joint pgf given by expression (32). For both classes, their respective cell delay must satisfy the constraint $\text{Prob}[d_j > T_j < 10^{-X_j}$, i.e., the fraction of cells of class-j that have a delay larger than the threshold T_j may not exceed 10^{-X_j}, where T_j and X_j depend on the application under consideration. It is assumed that class-1 cells are delay sensitive, implying that they are given priority over class-2 packets (and $T_1 < T_2$, since it makes no sense to have a higher delay threshold for delay-sensitive traffic). Class-2 traffic may be loss sensitive, and the amount of packets that is rejected due to a delay threshold being exceeded must be sufficiently small. Therefore, in the remainder we will set $X_2 = 9$ and $X_1 \equiv X$ (where the latter may be varied). It is clear that the performance of both traffic classes, in particular their delay characteristics, can be studied using the results derived throughout this paper.

The question we wish to answer is the following: what is the maximal load (denoted by $\rho_{T,\max}$), as a function of the traffic mix α, that still fulfills the two constraints? In Fig. 13, we show the
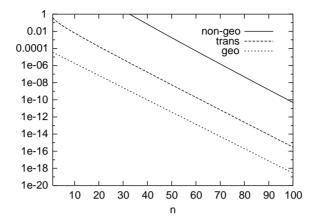
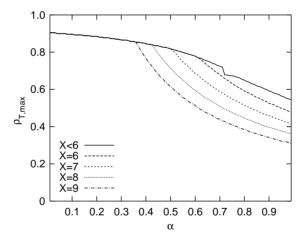Fig. 12. Tail behavior of the low-priority cell delay near the transition from non-geometrical to geometrical.

Fig. 13. Maximum load versus the fraction of class-1 arrivals for several values of X.

maximal load as a function of α when $T_1 = 10$, $T_2 = 100$ and $X = 1, \ldots, 9$. The constraint for the delay of class-2 cells is the same for all X, i.e., $\text{Prob}[d_2 > 100] < 10^{-9}$. For $X < 6$, we see that this constraint is the decisive one. We notice that the maximal load suddenly lowers a reasonable amount when α reaches approximately 0.7. At this point, the tail behavior changes from geometric to non-geometric tail behavior. The sudden change near 0.7 is probably due to the lack of accurateness in the tail behavior of the class-2 delay near the transition (as discussed earlier). Near this value for α, the maximal load we find is thus not that accurate, but one can see that the incorrectness is in the order of a few percentages. For higher X, the constraint for the delay of the high-priority traffic becomes decisive for high α, i.e., when more class-1 cells arrive. In Fig. 14, we show $\rho_{T,\max}$ as a function of α when $X = 4$, $T_2 = 100$ and $T_1 \geqslant 3$. The constraint for the delay of class-2 cells is again the same for all T_1. For $T_1 > 7$, we see that this constraint is the decisive one. For lower T_1, the constraint for the delay of the high-priority traffic becomes decisive for high α, i.e., when
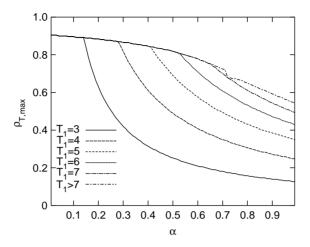
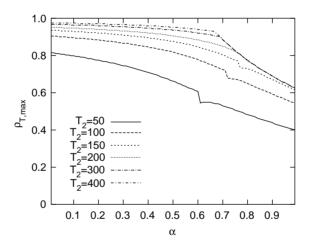Fig. 14. Maximum load versus the fraction of class-1 arrivals for several values of T_1.

Fig. 15. Maximum load versus the fraction of class-1 arrivals for several values of T_2.

more class-1 cells arrive. Finally, in Fig. 15, the maximum load as a function of α is shown, when $X = 4$, $T_1 = 10$ and several values of T_2. For low T_2, the constraint for the low-priority traffic is always the most stringent, while for $T_2 > 150$, the constraint for the high-priority traffic is decisive for high α. The behavior depicted in these three figures can be explained as follows. For $\alpha = 0$, the traffic mix consists of low-priority packets only, and $\rho_{T,\max}$ is relatively high, depending on the value of T_2. As α increases, $\rho_{T,\max}$ gradually decreases (but is still determined by T_2) since the growing fraction of high-priority packets causes the mean low-priority packet delay to rise. Then, as α further increases, a transition point is reached, which is defined as the value of α and ρ_T for which $\text{Prob}[d_1 > T_1] = 10^{-X}$ and $\text{Prob}[d_2 > T_2] = 10^{-9}$. Beyond this transition point, the bounding set by T_1 becomes predominant, and $\rho_{T,\max}$ further decreases due to the ever increasing presence of

high-priority packets in the traffic mix. These figures show that the maximum allowable load can strongly depend on the delay boundaries T_1 and T_2 set on the high- and low-priority packet delays, and the traffic mix α.

9. Conclusions

In this paper, we analyzed a queueing system with an HOL priority scheduling discipline. A generating functions approach was adopted, which led to close-form expressions of performance measure, such as mean and variance of the system contents and cell delay, and the correlation coefficient of the system contents of both types of cells, that are easy to evaluate. Furthermore, the tail behavior of system contents and cell delay is studied. We have shown that non-geometric tails can occur for system contents and cell delay of the low-priority traffic. The model included possible correlation between the number of arrivals of the two cell types during a slot. Therefore, the results could be used to evaluate the performance of a prioritized output-queueing switch with Bernoulli arrivals.

Acknowledgements

Appendix A. calculation of the probability mass function

Given a generating function $X(z) \triangleq \sum_{n=0}^{\infty} x(n)z^n$, the question is how to find an explicit, practically usable expression for its corresponding pmf $x(n)$. From the definition of $X(z)$ it follows that $x(n)$ is the coefficient of z^n in the expansion of $X(z)$ about $z = 0$, or equivalently the coefficient of z^{-1} in the expansion of $z^{-1-n}X(z)$ about $z = 0$. $x(n)$ is thus by definition the residue of the function $z^{-1-n}X(z)$ in the point $z = 0$. Since $z = 0$ is an n-multiple pole of $z^{-1-n}X(z)$, calculating the residue in $z = 0$ is nearly impossible for large n (since evaluating the residue in an n-multiple pole requires n derivations). Using the residue theorem of Cauchy however, it is proven that

$$x(n) = \text{Res}_{z=0}[X(z)z^{-1-n}]$$

$$= \frac{1}{2\pi i} \oint_{C_1} X(z)z^{-1-n}\,\mathrm{d}z - \sum_{j=0}^{m} \text{Res}_{z=z_j}X(z)z^{-1-n}$$

with $i = \sqrt{-1}$, C_1 a contour with infinite radius and z_j the poles of $X(z)$. The contour integral in the former expression is normally easy to calculate (in most cases the term equals zero). If we are only interested in the expression of $x(n)$ for large n, the sum of residues can be approximated by the residue in the dominant pole of $X(z)$ (the approximation is exact for $n \to \infty$). As a result, an easy, practically usable formula to calculate approximate tail probabilities is obtained.

Appendix B. inversion of $(1 - z)^{\alpha}$

Theorem B.1. *Assume that, with the sole exception of the singularity $z = 1$,*

$$F(z) \triangleq \sum_{n=1}^{\infty} f(n)z^n$$

is analytic in the domain

$$\Delta = \{z: |z| \leqslant 1 + \eta, |\mathrm{Arg}(z - 1)| \geqslant \theta\}$$

in which η is a positive real number and $0 < \theta < \pi/2$. Assume further that as z tends to 1 in Δ,

$$F(z) = K(1 - z)^{\alpha}$$

with $\alpha \notin \mathbb{N}$. Then, as $n \to \infty$,

$$f(n) = \frac{K}{\Gamma(-\alpha)} n^{-\alpha-1}.$$

References

[1] Liu KY, Petr DW, Frost VS, Zhu HB, Braun C, Edwards WL. Design and analysis of a bandwidth management framework for ATM-based broadband ISDN. IEEE Communications Magazine 1997;35(5):138–45.

[2] Van Mieghem P, Steyaert B, Petit GH. Performance of cell loss priority management schemes in a single server queue. International Journal of Communication Systems 1997;10(4):161–80.

[3] Bae JJ, Suda T. Survey of traffic control schemes and protocols in ATM networks. Proceedings of the IEEE 1991;79(2):170–89.

[4] Khamisy A, Sidi M. Discrete-time priority queues with two-state Markov modulated arrivals. Stochastic Models 1992;8(2):337–57.

[5] Rubin I, Tsai ZH. Message delay analysis of multiclass priority TDMA, FDMA, and discrete-time queueing systems. IEEE Transactions on Information Theory 1989;35(3):637–47.

[6] Sidi M, Segall A. Structured priority queueing systems with applications to packet-radio networks. Performance Evaluation 1983;3(4):265–75.

[7] Stanford DA. Interdeparture-time distributions in the non-preemptive priority $\sum M_i/G_i/1$ queue. Performance Evaluation 1991;12(1):43–60.

[8] Sugahara A, Takine T, Takahashi Y, Hasegawa T. Analysis of a nonpreemptive priority queue with SPP arrivals of high class. Performance Evaluation 1995;21(3):215–38.

[9] Takahashi Y, Hashida O. Delay analysis of discrete-time priority queue with structured inputs. Queueing Systems 1991;8(2):149–64.

[10] 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 1994;42(2–4):1837–45.

[11] Takine T. A nonpreemptive priority MAP/G/1 queue with two classes of customers. Journal of Operations Research Society of Japan 1996;39(2):266–90.

[12] Bruneel H, Steyaert B. Buffer requirements for ATM switches with multiserver output queues. Electronics Letters 1991;27(8):671–2.

[13] Bruneel H, Steyaert B, Desmet E, Petit GH. An analytical technique for the derivation of the delay performance of ATM switches with multiserver output queues. International Journal of Digital and Analog Communication Systems 1992;5:193–201.

[14] Hluchyj MG, Karol MJ. Queueing in high-performance packet switching. IEEE Journal on Selected Areas in Communications 1988;6(9):1587–97.

[15] Bruneel H, Kim BG. Discrete-time models for communication systems including ATM. Boston: Kluwer Academic Publishers, 1993.

[16] Xiong Y, Bruneel H. Buffer contents and delay for statistical multiplexers with fixed-length packet-train arrivals. Performance Evaluation 1993;17(1):31–42.

[17] Bruneel H. Analysis of buffer behaviour for an integrated voice-data system. Electronics Letters 1983;19(2):72–4.

[18] Bruneel H, Steyaert B, Desmet E, Petit GH. Analytic derivation of tail probabilities for queue lengths and waiting times in ATM multiserver queues. European Journal of Operational Research 1994;76(3):563–72.

[19] Drmota M. Systems of functional equations. Random Structures & Algorithms 1997;10(1–2):103–24.

[20] Flajolet P, Odlyzko A. Singularity analysis of generating functions. SIAM Journal on Discrete Mathematics 1990;3(2):216–40.

Joris Walraevens received the M.S. degree in Electrical Engineering from Ghent University, Belgium. In 1997, he joined the SMACS Research Group, Department for Telecommunications and Information Processing, at the same university. His main research interests include discrete-time queueing models and performance analysis of communication networks.

Bart Steyaert received the degrees of Licentiate in Physics and Licentiate in Computer Science from the University of Ghent, Belgium. He is a researcher at the SMACS Research Group, at the same university. His main research interests include discrete-time queueing models, traffic control, and stochastic modelling of high-speed communications networks.

Herwig Bruneel received the M.S. degree in Electrical Engineering, the degree of Licentiate in Computer Science, and the Ph.D. degree in Computer Science from Ghent University, Belgium. He is full time Professor in the Faculty of Applied Sciences and head of the Department of Telecommunications and Information Processing at the same university. He also leads the SMACS Research Group within this department and serves as the Academic Director for Research Affairs at Ghent University. His main interests include stochastic modelling and analysis of communication systems, discrete-time queueing theory, and the study of ARQ protocols. He has published more than 170 papers on these subjects and is coauthor of the book *H. Bruneel and B.G. Kim*, "*Discrete-Time Models for Communication Systems Including ATM*".