Service differentiation through priority jumps

Tom Maertens, Joris Walraevens, and Herwig Bruneel
Ghent University – UGent
Department of Telecommunications and Information Processing (IR07)
SMACS Research Group
Sint-Pietersnieuwstraat 41, B-9000 Gent, Belgium
Phone: +32-9-2648901
Fax: +32-9-2644295
E-mail: {tmaerten,jw,hb}@telin.ugent.be

Abstract

In this paper, we consider a queueing system with a priority scheduling scheme with priority jumps. Expressions for the probability generating functions of the queue contents and the packet delays have been derived in a previous paper. In the current paper, we determine expressions for some performance measures, i.e., mean values, and approximate tail distributions. These performance measures are furthermore used to illustrate the impact of the priority scheme on the performance of an output queue in a packet switch. We thereby compare the dynamic priority scheme with the FIFO scheme. At the end, we show that the results of this paper can be used in the performance study of more complicated models.

Keywords

discrete-time, queueing theory, dynamic priority scheduling, performance analysis

1 Introduction

One of the main keys to a succesful telecommunication network nowadays is the ability to efficiently support different services. Different services generate different types of traffic, and different types of traffic have extremely diverse *Quality-of-Service* (QoS) requirements. For *real-time* traffic for example, it is important that mean delay and delay jitter are not too large, while some loss is allowed. For *non-real-time traffic* on the other hand, the packet loss ratio is the restrictive quantity, and not so much the delay. In this paper, we further focus on delay as QoS measure. Taking into account their different delay requirements, we then categorize real-time traffic as *delay-sensitive*, and non-real-time traffic as *delay-tolerant*.

To support both types of traffic in a telecommunication network, they are scheduled according to a *priority* scheme. Two priority levels are provided, i.e., the *high- and low-priority* level, and some scheduling rules are introduced between both levels. In the Head-Of-Line (HOL) priority scheme for instance, the priority is always given to the delay-sensitive traffic. This priority scheme is not very efficient. Indeed, the HOL scheme provides relatively low delays for the delay-sensitive traffic, but when the system is highly loaded and a large portion of the system traffic consists of delay-sensitive traffic, it can cause excessive delays for the delay-tolerant traffic (see e.g., [1, 5, 13, 15, 20, 31]). Although this type of traffic tolerates a certain amount of delay, extreme values have to be avoided as much as possible. The Transmission Control Protocol (TCP) for example, could consider a delay-tolerant packet with a too big delay as lost, and would consequently decrease its transmission rate. This would decrease the throughput – which is particularly detrimental to data services – but is thus unnecessary since the packet is not lost. The service differentation between both types of traffic may thus be too drastic in some cases. This is because the HOL scheme is *static*: the priority levels never change in time, and packets of the low-priority level are only served when there are no high-priority packets present in the system.

Dynamic priority schemes aim for a more *gradual* service differentiation. The priority levels of both types of traffic can for instance be *varied dynamically* with time. When there is too much delay-tolerant traffic in the system, this type of traffic gets service priority for a while (see e.g., [4, 6, 7, 10, 11, 12, 16, 19]). Another way to reduce performance degradation for the delay-tolerant traffic, is to serve the priority levels in a *weighted* order. The priority levels do not change during time here, but packets of the low-priority level are with a certain regularity scheduled for service before the high-priority packets (see e.g., [17, 9, 25, 26, 28, 29, 30, 33]). In a third class of dynamic priority schemes, packets of the low-priority level can in the course of time *jump* to the high-priority level. From the server's point of view, such a scheme is then similar to the static HOL scheme: the server always chooses the packet for service at the head of the highest non-empty priority level. Many criteria can be used to decide when low-priority packets jump to the high-priority level: a maximum queueing delay in the low-priority queue (see [21]), a queue-length-threshold of the high- or low-priority queue (see [14, 23]), a random jumping probability per time unit (see [22]), the arrival characteristics of one type of traffic (see [24]), ...

In this paper, we consider a system that adopts a scheduling scheme with priority jumps. Particularly, we introduce a parameter β, which gives the probability that the total content of the low-priority queue jumps to the end of the high-priority queue. We opt for this straightforward model, so that we can analytically study the effect of priority jumps, and the influence of the system parameters on the performance of the system. The introduction of a jumping parameter β makes the model moreover very efficient. Indeed, the value of β can be chosen in such a way that the delay-tolerant traffic stays within its delay requirements: e.g., the more stringent the delay requirement, the larger the value of β. The service differentiation between both types of traffic can thus be adjusted by changing the value of the jumping parameter.

In a previous paper (see [22]), the authors have tackled the problem of obtaining analytical results for the probability generating functions of the joint pgf of the high- and low-priority queue, the marginal pgfs of the contents of the queues separately, and the pgfs of the delays of both types of packets. In the present paper, we concentrate on the derivation of expressions for some performance measures, such as the mean values and (approximate) tail probabilities. It is thereby shown that tail behaviour of a performance characteristic is not necessarily geometric. We further use these performance measures to illustrate the impact of the priority scheme with priority jumps on the performance of a specific queueing system. The results of this paper can moreover be applied to predict the performance of more complicated queueing systems.

The outline of the paper is as follows. In the following section, we summarise the results of [22]. In Sections 3 and 4, we calculate the moments of the queue contents and packet delays and study the tail behaviour of these quantities. We apply the obtained results to an output-queueing switch, and discuss the impact of the scheduling scheme in Section 5. Some conclusions are finally formulated in Section 6.

2 Previous results

We consider a *discrete-time* queueing system with *one server* and *two queues* of *infinite capacity*. The service time of a packet is one slot. We assume that two types of traffic arrive at the system: packets of type 1, representing the delay-sensitive traffic, and packets of type 2, which are delay-tolerant. Both types of traffic enter the system in separate queues. The numbers of arrivals are independent and identically distributed (i.i.d.) from slot-to-slot, but can be correlated in one slot. This dependence is expressed in their joint probability generating function (pgf) $A(z_1, z_2)$. We further define the marginal pgfs $A_1(z)$, $A_2(z)$, and $A_T(z)$, of the number of type-1 arrivals, of the number of type-2 arrivals, and of the total number of arrivals, in one slot. The corresponding arrival rates are then $\lambda_j = A_j'(1)$ $(j = 1, 2)$, and $\lambda_T = A_T'(1)$ (with $\lambda_T = \lambda_1 + \lambda_2$).

Newly arriving packets can enter service at the beginning of the slot following their arrival slot at the earliest. At the end of each slot, the content of the queue in which type-2 packets are originally stored, jumps with probability β to the other queue, where they join type-1 packets and previously jumped type-2 packets. The packets in the latter queue have service priority, and only when there are no packets present in this queue, type-2 packets in the other queue can be served. Within a queue, the service discipline is FIFO. For convenience, we further denote the queues by the high- and low-priority queue respectively. Note that the jumps occur at the end of a slot, and that the jumping packets are thus stored after the content of the high-priority queue at the end of the slot.

In [22], the authors have derived expressions for the pgfs of the total system content at the beginning of a slot, for the pgfs of the contents of both queues at the beginning of a slot, and for the pgfs of the delays of both types of packets. For the total system content and the queue contents, they have found

$$U_T(z) = \frac{(1-\lambda_T)A_T(z)(z-1)}{z - A_T(z)}, \tag{1}$$

$$U_H(z) = \frac{(1-\lambda_T)\beta A_T(z)z(z-1)}{(z-A_T(z))(z-(1-\beta)A_1(z))} - \frac{(1-\lambda_T)(1-\beta)\beta A_T(Y(1))A_1(z)(z-1)}{(z-(1-\beta)A_1(z))(Y(1)-A_T(Y(1)))}, \tag{2}$$

$$\begin{aligned} U_L(z) = &\frac{(1-\lambda_T)(1-\beta)A_2(z)(z-1)(1-Y(z))}{(1-(1-\beta)A_2(z))(z-Y(z))}\frac{Y(z)-(1-\beta)A_T(Y(z))}{Y(z)-A_T(Y(z))} \\ &+ \frac{\beta}{1-(1-\beta)A_2(z)}, \end{aligned} \tag{3}$$

with $Y(z)$ implicitly defined as $(1-\beta)A(Y(z), z)$. For the packet delays, following pgfs have been determined:

$$D_1(z) = \frac{\beta(1-\lambda_T)z(A_1(z)-1)(Y(1)A_T(z)-A_T(Y(1))z)}{\lambda_1(Y(1)-A_T(Y(1)))(z-A_T(z))(z-(1-\beta)A_1(z))}. \tag{4}$$

$$\begin{aligned} D_2(z) = &\frac{\beta(1-\lambda_T)}{\lambda_2}\frac{z(A_T(z)-A_1(z))}{(z-A_T(z))(1-(1-\beta)A_1(z))} \\ &+ \frac{(1-\beta)(1-\lambda_T)}{\lambda_2}\frac{z(A_T(V_0(z))-A_1(V_0(z)))(1-A_1(z))}{(V_0(z)-A_T(V_0(z)))(1-(1-\beta)A_1(z))}, \end{aligned} \tag{5}$$

with $V_0(z)$ implicitly given by $(1-\beta)zA_1(V_0(z))$. When we choose $\beta = 0$ in all these expressions, we obtain the same pgfs of the studied quantities as in [31]. This is expected, because when $\beta = 0$, type-2 packets never jump to the high-priority queue, and we thus have the same situation as in the static HOL priority scheme.

3 Calculation of the moments

In this section, we give expressions for the mean values of the studied quantities. Expressions for higher moments can be obtained in a similar way, but are omitted because of their size. We however illustrate them in figures in Section 5. To make the expressions more readable, we define λ_{11} and λ_{TT} as $\lambda_{11} \triangleq \left.\frac{\partial^2 A(z_1,z_2)}{\partial z_1 \partial z_1}\right|_{z_1=z_2=1}$ and $\lambda_{TT} \triangleq \left.\frac{\partial^2 A_T(z)}{\partial z^2}\right|_{z=1}$ respectively. By taking the first derivative of the respective pgfs for $z=1$, we obtain

$$\begin{aligned}
\mathrm{E}\left[u_T\right] =& \lambda_T + \frac{\lambda_{TT}}{2(1-\lambda_T)},\\
\mathrm{E}\left[u_H\right] =& 1+\lambda_2 - \frac{1-\lambda_1}{\beta} + \frac{\lambda_{TT}}{2(1-\lambda_T)} - \frac{(1-\lambda_T)(1-\beta)A_T(Y(1))}{Y(1)-A_T(Y(1))},\\
\mathrm{E}\left[u_L\right] =& \frac{(1-\beta)\lambda_2}{\beta} + \frac{(1-\lambda_T)(1-\beta)(Y(1)-(1-\beta)A_T(Y(1)))}{\beta(Y(1)-A_T(Y(1)))}.
\end{aligned}$$

It is easily verified that these expressions satisfy $\mathrm{E}\left[u_T\right] = \mathrm{E}\left[u_H\right] + \mathrm{E}\left[u_L\right]$, as expected. For the mean values of the packet delays, we find

$$\begin{aligned}
\mathrm{E}\left[d_1\right] =& 1+\lambda_2 - \frac{1-\lambda_1}{\beta} + \frac{\lambda_{TT}\lambda_1 + \lambda_{11} - \lambda_{11}\lambda_T}{2(1-\lambda_T)\lambda_1} - \frac{(1-\lambda_T)A_T(Y(1))}{Y(1)-A_T(Y(1))},\\
\mathrm{E}\left[d_2\right] =& 1-\lambda_1 + \frac{\lambda_1}{\beta} + \frac{(1-\lambda_1)\lambda_{TT} - (1-\lambda_T)\lambda_{11}}{2(1-\lambda_T)\lambda_2} - \frac{(1-\beta)(1-\lambda_T)(A_T(V_0(1)) - A_1(V_0(1)))\lambda_1}{\lambda_2(V_0(1)-A_T(V_0(1)))\beta}.
\end{aligned}$$

Notice that $\mathrm{E}\left[u_H\right] \neq \lambda_1 \mathrm{E}\left[d_1\right]$ and that $\mathrm{E}\left[u_L\right] \neq \lambda_2 \mathrm{E}\left[d_2\right]$, as one would - at first - expect according to Little's law. The reason for this is that in the calculation of the system content, packets of the low-priority queue jump to the high-priority queue and from that moment on, they are treated as part of the content of the high-priority queue. This is of course not the case in the calculation of the packet delay of a type-2 packet. So basically, the system content is analyed on a "queue"-basis, while the packet delays are analyzed on "packet"-basis.

4 Calculation of the tail probabilities

Another important performance characteristic, besides the moments, is the (tail) distribution of the studied quantities. The tail probabilities, i.e., the probability mass function (pmf) for large values, typically represent the 'exceptional' situations in a queueing system. The probability that the delay is larger than a given value N, or the packet loss, are examples of interesting performance measures for which the calculation of the tail probability is usually sufficient. The tail distribution is thus often used to impose statistical bounds on the guaranteed QoS for both types of traffic.

Exact theoretical solutions for this inversion problem make use of the probability generating property of pgfs, or of residue theory. However, since these solution methods need a lot of derivations, they are often quite unpractical. We will therefore use an *approximate* solution technique, which is known to be quite popular: the dominant-singularity method. In [3] for example, it has been shown that the pmf $x(n)$ of a discrete variable X is - for high n - dominated by the contribution of the singularity of the corresponding pgf $X(z)$ with the smallest absolute value. Because of a property of pgfs, this *dominant* singularity is necessarily positive real and larger than 1. In this section, we derive expressions for the tail probabilities of the total system content, of the contents of the high- and low-priority queue separately, and of the delay of a type-1 packet, by using this dominant-singularity approximation method and Darboux's theorem (see Appendix A).

It is assumed in the remainder that the pgfs $A_T(z)$, $A_1(z)$, and $A_2(z)$, and their derivatives go to infinity for z equal to their radii of convergence or for $z \to \infty$. This includes all 'usual' arrival processes, and is thus not a restrictive assumption. We furthermore suppose that $\beta > 0$. For $\beta = 0$ (i.e., the static HOL scheme), we refer to [31].

4.1 Content of the total system

The tail behaviour of the total system content has also been investigated in [31]. The following approximation is there found:

$$\text{Prob}\,[u_T = n] \approx \frac{(1-\lambda_T)(s_T - 1)s_T^{-n}}{A_T'(s_T) - 1}, \tag{6}$$

with s_T respresenting the dominant singularity of the pgf $U_T(z)$. It is the (dominant) positive real zero larger than 1 of $z - A_T(z)$, i.e., the numerator of $U_T(z)$.

4.2 Content of the high-priority queue

Let us further concentrate on the tail behaviour of the content of the high-priority queue. Two singularities may play a role here, namely the dominant positive real zeros larger than 1 of $z - A_T(z)$ and $z - (1-\beta)A_1(z)$. We denote them by s_T and s_1 respectively. For z positive real, larger than 1, and in the mutual regions of convergence of $A_T(z)$ and $A_1(z)$, we can however easily verify that $A_T(z) > (1-\beta)A_1(z)$. So s_T is always smaller than s_1, and as a consequence, s_T is the dominant singularity of $U_H(z)$.

Since the first derivative of $U_H(z)$ stays finite for $z = s_T$, this singularity is a pole with multiplicity one. In the neighbourhood of this pole, we can approximate $U_H(z)$ by $K_{U_H}/(s_T - z)$. The constant K_{U_H} is obtained by calculating $\lim\limits_{z \to s_T} U_H(z)(s_T - z)$. By using Darboux's theorem (see Appendix A), we subsequently find

$$\text{Prob}\,[u_H = n] = K_{U_H} s_T^{-n-1}. \tag{7}$$

4.3 The function $Y(z)$

The tail behaviour of the content of the low-priority queue is not so straightforward. This is in the first place due to the appearance of $Y(z)$ in the expression of $U_L(z)$ (see (3)). We first take a closer look at this implicitly defined function on the positive real axis.

As z increases along the positive real axis, a branch point s_B is encountered where $Y(z)$ stays finite, but where $Y'(z) \to \infty$ (see e.g., [18] and [31] for similar cases). s_B is thus the solution of

$$\left\{ \begin{array}{l} Y(s_B) = (1-\beta)A(Y(s_B), s_B) \\ Y'(s_B) \to \infty \end{array} \right. \Rightarrow \left\{ \begin{array}{l} Y(s_B) - (1-\beta)A(Y(s_B), s_B) = 0 \\ (1-\beta)A^{(1)}(Y(s_B), s_B) = 1 \end{array} \right. . \tag{8}$$

For values of z beyond s_B, $Y(z)$ is no longer properly defined. Note that $Y(z)$ is a solution of the functional equation $x - (1-\beta)A(x,z) = 0$. This equation has another positive real solution $Y^*(z)$, which decreases as z increases (see Figure 1). Both solutions then coincide for $z = s_B$. By applying the results of [8], one can show that in the neighbourhood of s_B, $Y(z)$ is approximately given by

$$Y(z) \approx Y(s_B) - K_Y(s_B - z)^{1/2}, \tag{9}$$

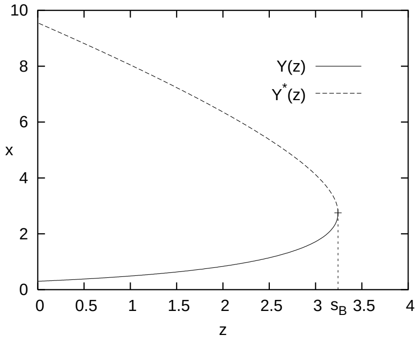
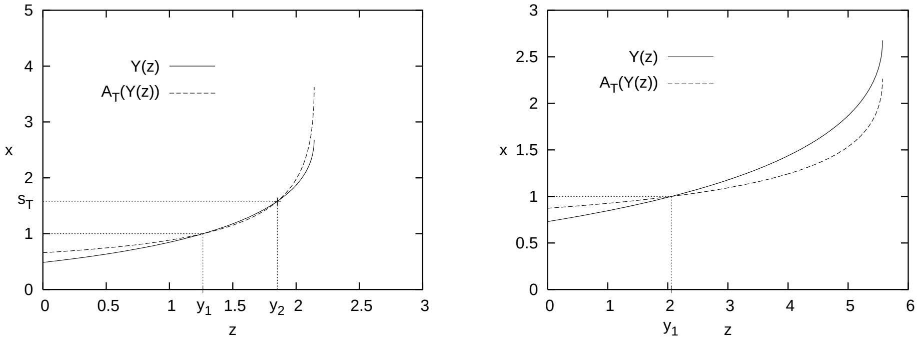


Figure 1: Solutions of $x - (1-\beta)A(x,z) = 0$



a. *y_1 and y_2* b. *y_1 (y_2 does not exist)*

Figure 2: The zero(s) of $Y(z) - A_T(Y(z))$

with

$$K_Y = \sqrt{\frac{2A^{(2)}(Y(s_B), s_B)}{A^{(11)}(Y(s_B), s_B)}}. \tag{10}$$

The constant K_Y is found by substituting $z = s_B$ in (9). Since $Y(z)$ appears in the expression of $U_L(z)$, s_B is also a singularity of this pgf, and may thus play a role in the tail behaviour of the content of the low-priority queue.

4.4 Content of the low-priority queue

Secondly, we show that none of the (dominant) positive real solutions larger than 1 of $1 - (1 - \beta)A_2(z)$ - denoted by s_2 - and $z - Y(z)$ - denoted by s_Y - are potential singularities of $U_L(z)$.

Since s_2 is a zero of $1 - (1-\beta)A(1,z)$, it is easily seen that $(x,z) = (1, s_2)$ is a solution of $x - (1-\beta)A(x,z) = 0$. s_2 is thus smaller than or equal to s_B, because the latter equation has no solution for $z > s_B$ (see previous subsection). We have also shown that this equation has two positive real solutions for $z \le s_B$, namely $(x,z) = (Y(z), z)$ and $(x,z) = (Y^*(z), z)$. So when $z = s_2$, there are two possibilities for x, i.e., $x = Y(s_2)$ and $x = Y^*(s_2)$. One of them must equal 1. Since $Y(s_B) > 1$ and $Y^*(z) > Y(s_B)$, $Y(s_2) = 1$. The numerator of $U_L(z)$ however also vanishes for $z = s_2$ when $Y(s_2) = 1$, which means that s_2 is not a singularity of $U_L(z)$. Likewise, choosing $z = s_Y$ in the numerator of $U_L(z)$ and using the definition of $Y(z)$ (see Section 2), also leads to zero. s_Y is thus not a singularity of $U_L(z)$ as well.

Finally, we look at the zeros of $Y(z) - A_T(Y(z))$, i.e., the last factor of the denominator which may possibly yield a potential singularity. Note that the zero(s) of this factor are smaller than or equal to s_B, since $Y(z)$ appears in the factor and since $Y(z)$ does not exist for $z > s_B$. The equation $x - A_T(x) = 0$ has $x = 1$ and $x = s_T$ as solutions on the positive real axis (with $s_T > 1$). We can thus easily verify that the smallest positive real zero of $Y(z) - A_T(Y(z))$ satisfies $Y(z) = 1$ (see Figure 2a.). This zero, denoted by y_1, coincides with s_2, and we have already shown that this is not a singularity of $U_L(z)$. On the other hand, the second positive real zero y_2 of $Y(z) - A_T(Y(z))$, which satisfies $Y(z) = s_T$ (see Figure 2a.), seems to be a potential singularity, since y_2 is not a zero of the numerator. y_2 does however not always exist. When $s_T > A_T(Y(s_B))$, the functions $Y(z)$ and $A_T(Y(z))$ cease to exist before they intersect once more (see Figure 2b.). In this case, they have only point of intersection, namely y_1.

In summary, the tail behaviour of the contents of the low-priority queue is dominated by the singularities s_B or y_2, depending on whether y_2 exists or not. Three cases can occur: y_2 exists and $y_2 < s_B$, y_2 exists and $y_2 = s_B$, or y_2 does not exist. In the first case, the singularity y_2 is dominant. This singularity is a pole with multiplicity one. Consequently, $U_L(z) \approx \dfrac{K_{U_L}^{(1)}}{y_2 - z}$ for $z \to y_2$. The constant $K_{U_L}^{(1)}$ can be obtained by determining $\lim\limits_{z \to s_T} U_L(z)(y_2 - z)$:

$$K_{U_L}^{(1)} = \frac{(1-\lambda_T)\beta(1-\beta)A_2(y_2)(y_2-1)(s_T-1)s_T}{Y'(y_2)(A_T'(s_T)-1)(1-(1-\beta)A_2(y_2))(y_2-s_T)}, \tag{11}$$

where we have used the fact that $Y(y_2) = s_T$. In the second case, y_2 and s_B coincide, and are so-called *co-dominant*. Using expression (9) in (3), and taking into account the fact that $Y(s_B) = s_T$, yields

$$U_L(z) \approx \frac{\left\{ \begin{array}{l} (1-\lambda_T)(1-\beta)A_2(z)(z-1)\left(1 - s_T + K_Y(s_B-z)^{1/2}\right) \\ \times \left(s_T - K_Y(s_B-z)^{1/2} - (1-\beta)s_T + (1-\beta)K_Y A_T'(s_T)(s_B-z)^{1/2}\right) \end{array} \right\}}{(s_B-z)^{1/2}K_Y(A_T'(s_T)-1)(1-(1-\beta)A_2(z))\left(z - s_T + K_Y(s_B-z)^{1/2}\right)}. \tag{12}$$

The pgf $U_L(z)$ can thus be approximated by $\dfrac{K_{U_L}^{(2)}}{(s_B-z)^{1/2}}$ in the neighbourhood of $y_2 = s_B$, with

$$K_{U_L}^{(2)} = \frac{(1-\lambda_T)(1-\beta)\beta s_T A_2(s_B)(s_B-1)(1-s_T)}{K_Y(A_T'(s_T)-1)(1-(1-\beta)A_2(s_B))(s_B-s_T)}. \tag{13}$$

In the third case, when y_2 does not exist, the branch point s_B is dominant. By substituting expression (9) in (3), we obtain

$$U_L(z) \approx \frac{\left\{ \begin{array}{l} (1-\lambda_T)(1-\beta)A_2(z)(z-1)\Big(1 - Y(s_B) + K_Y(s_B-z)^{1/2}\Big) \\ \times\Big(Y(s_B) - (1-\beta)A_T(Y(s_B)) + K_Y(s_B-z)^{1/2}\left((1-\beta)A_T'(Y(s_B)) - 1\right)\Big) \\ \times\Big(z - Y(s_B) - K_Y(s_B-z)^{1/2}\Big) \\ \times\Big(Y(s_B) - A_T(Y(s_B)) - K_Y(s_B-z)^{1/2}(A_T'(Y(s_B))-1)\Big) \end{array} \right\}}{\left\{ \begin{array}{l} (1-(1-\beta)A_2(z))\Big((z-Y(s_B))^2 - K_Y^2(s_B-z)\Big) \\ \times\Big((Y(s_B) - A_T(Y(s_B)))^2 - K_Y^2(s_B-z)(A_T'(Y(s_B))-1)^2\Big) \end{array} \right\}}. \tag{14}$$

This expression leads to $U_L(z) \approx U_L(s_B) - K_{U_L}^{(3)}(s_B - z)^{1/2}$ in the neighbourhood of s_B, with

$$K_{U_L}^{(3)} = \frac{\left\{ \begin{array}{l} (1-\lambda_T)(1-\beta)\Big(\beta(1-Y(s_B))(s_B - Y(s_B))(Y(s_B)A_T'(Y(s_B)) - A_T(Y(s_B))) \\ -(s_B - 1)(Y(s_B) - A_T(Y(s_B)))(Y(s_B) - (1-\beta)A_T(Y(s_B)))\Big)(s_B - 1)A_2(s_B) \end{array} \right\}}{(s_B - Y(s_B))^2\,(Y(s_B) - A_T(Y(s_B)))^2\,(1-(1-\beta)A_2(s_B))}. \tag{15}$$

We now have approximate expressions for $U_L(z)$ in the neighbourhood of its dominant singularity, for the three possible cases. By using Darboux's theorem with these approximations, we find the corresponding tail probabilities:

$$\mathrm{Prob}\,[u_L = n] = \begin{cases} K_{U_L}^{(1)} y_2^{-n-1} & \text{if } y_2 < s_B \\ \dfrac{K_{U_L}^{(2)} n^{-1/2} s_B^{-n}}{\sqrt{\pi s_B}} & \text{if } y_2 = s_B \\ \dfrac{K_{U_L}^{(3)} n^{-3/2} s_B^{-n}}{2\sqrt{\pi / s_B}} & \text{if } y_2 \text{ does not exist} \end{cases}, \tag{16}$$

where the constants $K_{U_L}^{(i)}$ $(i = 1, 2, 3)$ are given by (11), (13) and (15) respectively. The first expression constitutes a typical geometric tail behaviour, while the others are of a non-geometric nature.

4.5 Delay of a type-1 packet

The dominant singularity of $D_1(z)$ is the same as the dominant singularity of $U_H(z)$ (also with multiplicity one). In the neighbourhood of s_T, $D_1(z)$ is approximated by

$$D_1(z) \approx \frac{\beta(1-\lambda_T)s_T^2(A_1(s_T) - 1)}{\lambda_1(s_T - (1-\beta)A_1(z))(A_T'(s_T) - 1)(s_T - z)}. \tag{17}$$

For the tail probabilities of the delay of a type-1 packet, we obtain

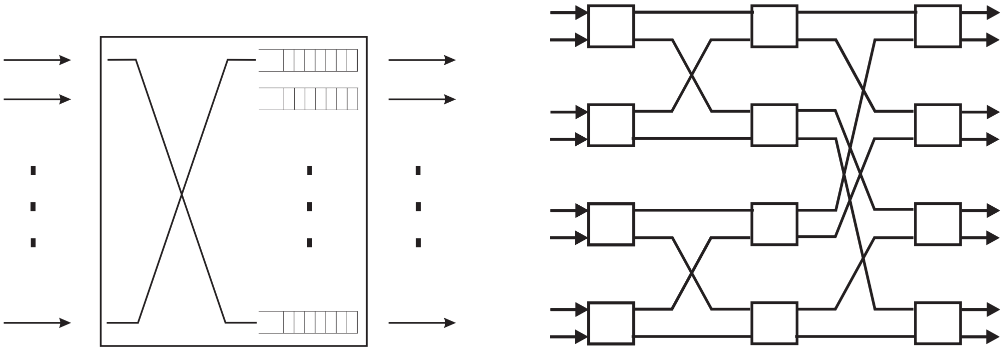
$$\mathrm{Prob}\,[d_1 = n] = \frac{\beta(1-\lambda_T)s_T^{1-n}(A_1(s_T) - 1)}{\lambda_1(s_T - (1-\beta)A_1(s_T))(A_T'(s_T) - 1)}. \tag{18}$$

4.6 Delay of a type-2 packet

The tail behaviour of the delay of type-2 packet is again more complicated. It can namely be characterised by four singularities: the positive real zeros larger than 1 of $z - A_T(z)$, $V_0(z) - A_T(V_0(z))$, and $1 - (1-\beta)A_1(z)$, plus the branch point of $V_0(z)$. We may thus have quite a lot of different cases with respect to the dominant singularity of $D_2(z)$. A special paper is therefore devoted to this (see [23]).

5 Application

The results obtained in the former sections are now applied to an output-queueing switch (see Figure 3a.). This output-queueing switch has N inlets and N outlets and we assume that two



a. *An NxN output-queueing switch* b. *An 8x8 self-routing 3-stage switching network*

Figure 3: The switching environment

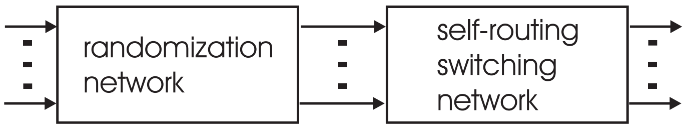


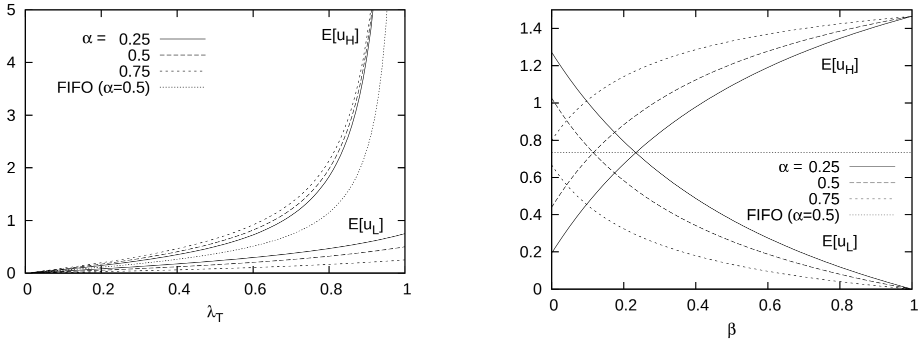
Figure 4: A switching network with traffic randomization

types of traffic arrive at the switch: traffic of type 1, which is delay-sensitive, and traffic of type 2, respresenting delay-tolerant traffic. The packet arrivals on the inlets are generated by independent and identically distributed (i.i.d.) Bernoulli processes with arrival rate λ_T. An arriving packet is assumed to be of type j with probability λ_j/λ_T $(j = 1, 2)$. So $\lambda_1 + \lambda_2 = \lambda_T$. The incoming packets are routed to the output queue corresponding to their destination, in an independent and uniform way. The output queues thus all behave identically, and we can concentrate on the study of one. The numbers of arriving packets to an output queue in one slot are generated according to a two-dimensional binomial process, which is fully characterized by the joint pgf

$$A(z_1, z_2) = \left(1 - \frac{\lambda_1}{N}(1 - z_1) - \frac{\lambda_2}{N}(1 - z_2)\right)^N . \tag{19}$$

Obviously, the numbers of packets entering an output queue are correlated within one slot. When m type-1 packets arrive during a slot $(0 \leq m \leq N)$, the maximum number of type-2 arrivals during the same slot is limited by $N - m$ (because there are only N inlets). An output queue is furthermore assumed to exist of two logical queues. Type-1 packets arrive to the first queue, and type-2 packets arrive to the second queue. The packets of the first queue have service priority over the packets of the second, and in each slot, the contents of the second jumps to the first with probability β. The results obtained in the former sections can thus be used to study the performance of an output queue in a switch.

The choice for Bernoulli arrivals on the inlets of the switch is motivated as follows. An NxN switching element, as described above, is the smallest building block of a PxP self-routing switching network (see e.g., [27] and [32]). The number of stages in the switching network, denoted by K, is then equal to $log_N P$. In Figure 3b., we for example see a 8x8 self-routing 3-stage switching network, consisting of 12 2x2 switching elements. In [27], the authors state that in the case of random (i.e., uncorrelated in time) input traffic, a Bernoulli proces is a reasonably good candidate to represent the arrival process on the inlets of the switching elements. However, due to the integrated traffic (voice and data) in real networks, the input traffic is rather bursty (i.e., correlated). It is known that the traffic burstiness (or correlation) adversely affects the performance. By adding a randomization network in front of the switching network (see Figure 4), the performance of the



a. *versus the total arrival rate for* $\beta = 0.5$ b. *versus* β *for* $\lambda_T = 0.7$

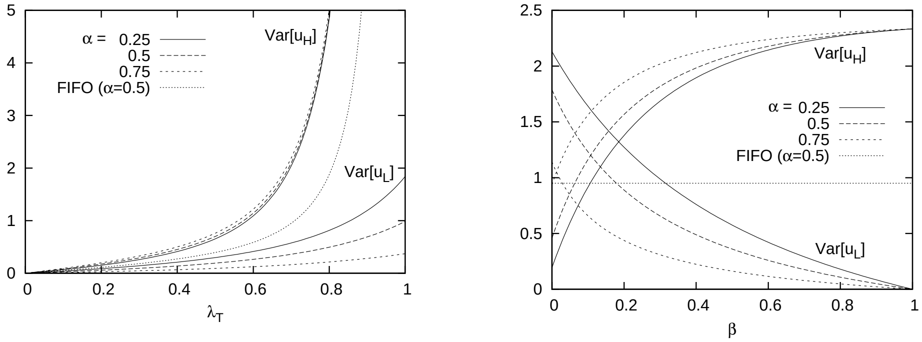
Figure 5: Mean values of the queue contents

switching network can be significantly improved. A randomazation network distributes the bursty input traffic among all the inlets of the self-routing network. As a consequence, the performance of the switching network is made less sensitive to the bursty input traffic (see e.g. [32]). For the ideally randomization network, its output traffic is assumed to be random traffic, independent of the input traffic. This brings us back to a Bernoulli arrival proces.

Let us now study the impact of the priority scheme with priority jumps on the performance of an output queue in a switch. We therefore consider some performance measures, such as the mean values and the variances of the queue contents and the packet delays. The performance study is focused on the comparison between queues with the dynamic priority scheme and the FIFO scheme. Note that we assume a 16x16-switch ($N = 16$). We finally define α as the fraction of type-1 arrivals in the overall traffic mix (i.e., $\alpha \triangleq \lambda_1/\lambda_T$).

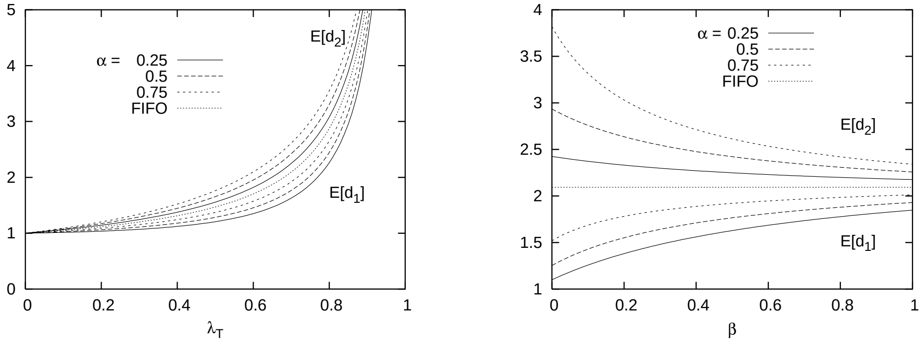
In Figure 5a., the mean values of the contents of the high- and low-priority queue are shown as functions of the total arrival rate λ_T, for $\beta = 0.5$ and $\alpha = 0.25$, 0.5 and 0.75 respectively. In order to compare the dynamic priority scheme with the FIFO scheme, we have applied a FIFO scheduling on a joint queue in which the packets of both types of traffic are mixed up (according to their arrivals). We have plotted the mean number of packets present in the system of one type of traffic, since for $\alpha = 0.5$, the mean number of type-1 packets in the system equals the mean number of type-2 packets. We can easily see that $\mathrm{E}\left[u_H\right]$ is larger for the dynamic priority scheme than for the FIFO scheme. For $\mathrm{E}\left[u_L\right]$, the opposite holds. This can be explained as follows: packets of the high-priority queue have priority over packets of the low-priority queue. So without priority jumps, the low-priority queue would build up as long as there are packets in the high-priority queue. Because of the priority jumps, the content of the low-priority queue however jumps once every two slots to the high-priority queue, thereby leaving the low-priority queue totally empty. As a consequence, $\mathrm{E}\left[u_H\right]$ is larger than $\mathrm{E}\left[u_L\right]$. The figure also shows that $\mathrm{E}\left[u_H\right]$ increases when α increases. This is expected, since a higher value of α means a larger fraction of type-1 packets in the arrival stream. The opposite again holds for $\mathrm{E}\left[u_L\right]$.

Figure 5b. shows the mean values of the queue contents as functions of β, for $\lambda_T = 0.7$ and $\alpha = 0.25$, 0.5 and 0.75 respectively. The influence of β is quite obvious: larger β means more jumps (on average), resulting into a higher $\mathrm{E}\left[u_H\right]$ and a lower $\mathrm{E}\left[u_L\right]$. The figure also shows that the two curves for $\alpha = 0.25$ and the two curves for $\alpha = 0.5$ intersect each other - and the FIFO curve - for certain values of β. This means that from those β-values on, $\mathrm{E}\left[u_H\right]$ is larger than $\mathrm{E}\left[u_L\right]$ for the respective values of α. For $\alpha = 0.75$, $\mathrm{E}\left[u_H\right]$ is always larger than $\mathrm{E}\left[u_L\right]$ (when $\lambda_T = 0.7$). When $\beta = 1$, we can easily see that $\mathrm{E}\left[u_L\right] = 0$. In each slot, the newly arriving type-2 packets immediately jump to the high-priority queue. The low-priority queue is thus always empty then. For the variances of the queue contents (see Figure 6), the same conclusions can be drawn as for the mean values.



a. *versus the total arrival rate for* $\beta = 0.5$ b. *versus* β *for* $\lambda_T = 0.7$

Figure 6: Variances of the queue contents

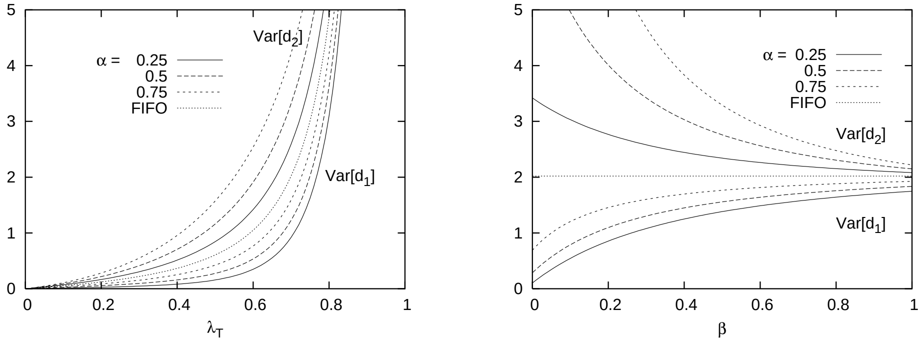


a. *versus the total arrival rate for* $\beta = 0.5$ b. *versus* β *for* $\lambda_T = 0.7$

Figure 7: Mean values of the packet delays

In Figure 7a., we depict the mean values of the packet delays of both types of traffic as functions of the total arrival rate, for $\beta = 0.5$ and $\alpha = 0.25$, 0.5 and 0.75 respectively. We also show the mean value of the packet delay for the FIFO scheme. The packet delays are then the same for type-1 and type-2 traffic (independent of α), and can thus be calculated as if there is only one type of traffic arriving at the system, according to an arrival process with pgf $A(z, z)$ (see [2]). The influence of the priority scheduling is quite obvious: $\mathrm{E}\,[d_1]$ is smaller for the dynamic priority scheme than for the FIFO scheme. For $\mathrm{E}\,[d_2]$, the opposite holds. The reason is clear: the type-1 packets have priority over the type-2 packets. The influence of the dynamic priority scheme is however limited. The mean delay of a type-1 packet reduces only moderately in comparison with the mean delay for the FIFO scheme, while the price to pay, a higher mean delay of a type-2 packet, is also rather small. Note further that it follows from this figure that increasing the fraction of type-1 packets in the overall traffic mix (i.e., increasing α), increases the mean delay of both types of packets. Indeed, the smaller amount of type-2 packets suffer from larger delays, and thus give cause for to a larger $\mathrm{E}\,[d_2]$ as well.

In Figure 7b., the mean values of the packet delays are shown as functions of β, for $\lambda_T = 0.7$ and $\alpha = 0.25$, 0.5 and 0.75. A larger value of β implies more jumps, and as a consequence, a lower negative effect from the priority scheduling on $\mathrm{E}\,[d_2]$. The price to pay is a higher $\mathrm{E}\,[d_1]$. We can derive similar conclusions with respect to the delay jitter (see Figure 8). We can here conclude that the dynamic priority scheme does what it is designed for: lowering the delay of the type-1 packets (which are delay-sensitive), but in contrast with the static HOL priority scheme (see [31]), taking into consideration the delay of the type-2 packets (being delay-tolerant). The parameter β can be chosen depending on the delay guarantees for both types of traffic. A low β will highly favour the delay-sensitive traffic, while choosing β higher will give the delay-sensitive traffic only a small delay



a. *versus the total arrival rate for* $\beta = 0.5$ b. *versus* β *for* $\lambda_T = 0.7$

Figure 8: Variances of the packet delays

reduction (compared to the FIFO scheme).

All numerical examples until now, assume that the number of arrivals of one type of traffic is correlated with the number of arrivals of the other type of traffic in one slot (as described in Section 2). Our model however does not include correlation amongst consecutive slots. The results of the current paper are therefore not directly useful to describe a queueing system with correlation in time. In the following, we propose an approximation for a system with correlated arrivals based on our results of the system with no correlation in time.

Before we propose our approximation, we describe the modelling of the correlation that is used to validate the approximation. The queueing system is modelled with a single server and two priority queues of infinite capacity. The queues are fed by m_j number of two-state (i.e., an idle and an active state) independent Markov sources, with $j = 1, 2$ denoting the high- and low-priority queue respectively. The arrival process to each queue is thus correlated within itself, but the two arrival processes are independent in this paragraph. For type-j sources, a transition from idle state to active state occurs with probability $1 - \gamma_j$, while the probability of a transition from active to idle state occurs with probability $1 - \alpha_j$. We assume that an active source generates one packet per slot, whereas an idle source generates no packets during a slot. σ_j represents the fraction of time a type-j user is in the active state, and α_j and γ_j are selected in accordance with

$$\text{mean active period of a type-}j\text{ user} = \frac{1}{1-\alpha_j} = \frac{K}{1-\sigma_j}, \tag{20}$$

and

$$\text{mean passive period of a type-}j\text{ user} = \frac{1}{1-\gamma_j} = \frac{K}{\sigma_j}. \tag{21}$$

Note that K is hereby defined as the burstiness factor of both types of traffic.

Now, the basic idea of our approximation is the fact that the influence of time-correlation on the performance measures is similar in case of the FIFO scheme as in the case of the HOL-PJ (Head-Of-Line with Priority Jumps) scheme. More precisely, we calculate the difference between $\mathrm{E}[d_1]$ for the HOL-PJ scheme and $\mathrm{E}[d]$ for FIFO scheme, and we assume that this difference is independent from the burstiness factor K. Or, said in a more readable manner:

$$\mathrm{E}[d]_{\text{FIFO}} - \mathrm{E}[d_1]_{\text{HOL-PJ}} \approx \text{independent from } K. \tag{22}$$

Table 1: Validation of simulation results versus approximate results for $E[d_1]$

	$K = 5$		$K = 10$		$K = 20$	
λ_T	$E[d_1]_{\text{sim}}$	$E[d_1]_{\text{approx}}$	$E[d_1]_{\text{sim}}$	$E[d_1]_{\text{approx}}$	$E[d_1]_{\text{sim}}$	$E[d_1]_{\text{approx}}$
0.1	1.394	1.402	1.906	1.888	2.973	2.860
0.2	1.892	1.909	3.509	3.003	5.415	5.190
0.3	2.558	2.556	4.555	4.441	8.675	8.191
0.4	3.453	3.450	6.575	6.367	12.870	12.200
0.5	4.754	4.698	9.457	9.073	18.906	17.823
0.6	6.751	6.590	13.785	13.153	26.810	26.278
0.7	10.099	9.778	20.920	19.986	43.001	40.403
0.8	16.967	16.232	35.778	33.732	72.657	68.732
0.9	37.399	35.836	79.012	75.211	163.005	153.961

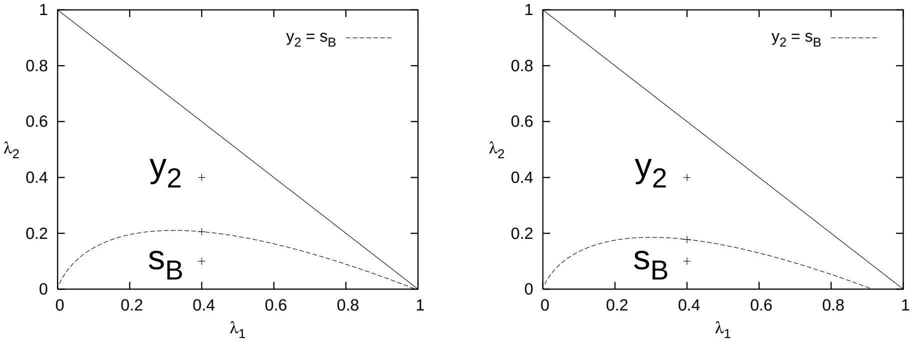
Table 2: Validation of simulation results versus approximate results for $E[d_2]$

	$K = 5$		$K = 10$		$K = 20$	
λ_T	$E[d_2]_{\text{sim}}$	$E[d_2]_{\text{approx}}$	$E[d_2]_{\text{sim}}$	$E[d_2]_{\text{approx}}$	$E[d_2]_{\text{sim}}$	$E[d_2]_{\text{approx}}$
0.1	1.482	1.452	1.988	1.938	2.968	2.910
0.2	2.085	2.017	3.226	3.110	5.501	5.298
0.3	2.853	2.741	4.811	4.616	8.706	8.366
0.4	3.868	3.704	6.909	6.620	12.996	12.454
0.5	5.445	5.048	9.850	9.423	18.994	18.173
0.6	7.407	7.058	14.261	13.620	28.147	26.745
0.7	10.883	10.396	21.541	20.605	43.239	41.021
0.8	17.864	17.047	36.302	34.547	72.760	69.547
0.9	38.412	36.919	79.851	76.294	162.865	155.044

This independency then leads to

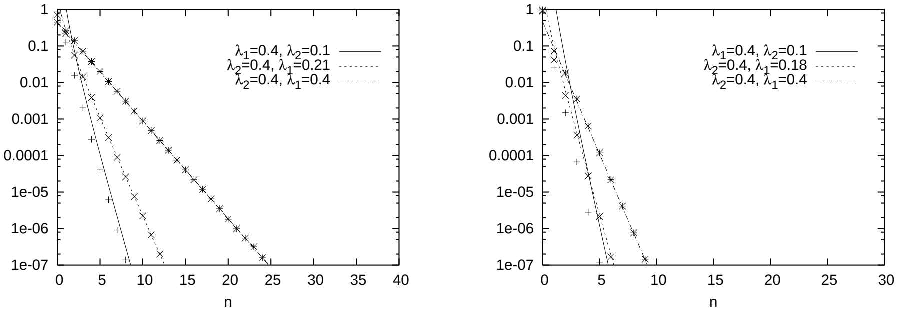
$$\mathrm{E}[d_1]_{\text{HOL-PJ, general } K} \approx \mathrm{E}[d]_{\text{FIFO,general } K} - (\mathrm{E}[d]_{\text{FIFO,}K\,=\,1} - \mathrm{E}[d_1]_{\text{HOL-PJ,}K\,=\,1}). \tag{23}$$

All quantities in the right-hand side of (23) can be explicitly calculated. $E[d]_{\text{FIFO,general } K}$ and $E[d]_{\text{FIFO,}K\,=\,1}$ can be easily obtained from [2], where a single-class FIFO queue with the arrival process as described in this paragraph is analysed. Further, the quantity $E[d_1]_{\text{HOL-PJ,}K\,=\,1}$ of the left-hand side of the equation, is determined in the current paper. To validate our approximation, we have compared the approximate results with simulation results, for $(m_1, m_2) = (8, 8)$ and various β. Table 1 shows the simulation results and the approximate results, for $\beta = 0.5$, and $K = 5$, 10 and 20 respectively. The table shows that our approximations are very good. We can thus predict the behaviour of a queueing system with the HOL-PJ scheduling scheme and with correlated arrivals, by combining the results obtained for a queue with the HOL-PJ scheme and uncorrelated arrivals, and the results for a FIFO queue and correlated arrivals. Simulations for other values of β learn that for lower β these approximations are slightly worse, but still very reasonable, while for higher β even better approximations are found. Note finally that a similar discussion can be followed with respect to the mean packet delay of type-2 packets (see Table 2).



a. *for* $\beta = 0.1$ b. *for* $\beta = 0.75$

Figure 9: Regions for tail behavior of $U_2(z)$ as a function of the arrival rates



a. *for* $\beta = 0.1$ b. *for* $\beta = 0.75$

Figure 10: Tail behavior of the queue-2 contents for some combinations of type-1 and type-2 arrival rates

Let us finally take a look at the tail behaviour of the content of the low-priority queue. We have shown that three types of behaviour are encountered, depending on which singularity is dominant. In case of the output-queueing switch considered in this section, Figure 9 shows for which combinations of type-1 and type-2 arrival rates, $y_2 = s_B$ (i.e., the singularities y_2 and s_B coincide). The curve splits the (λ_1, λ_2)-space in 2 regions: a region in which y_2 does not exist, making s_B dominant (below the curve), and a region in which y_2 is dominant (above the curve) - since y_2, when it exists, is smaller than s_B. Note that in the area above the linear line (defined by $\lambda_1 + \lambda_2 = 1$), the total arrival rate is larger than 1, resulting in an unstable system. When we compare Figure 9a. ($\beta = 0.1$) with Figure 9b. ($\beta = 0.75$), we notice the role of β: the region below the curve, where y_2 does not exist, becomes smaller for increasing β. Although the role of β is limited, we can conclude that the values of *all* system parameters have an influence on the tail behavior of $U_2(z)$.

Figure 10a. and 10b. then show the tail probabilities of the content of the low-priority queue, for the (λ_1, λ_2)-combinations indicated by the marks in Figure 9a. and Figure 9b. respectively. We have compared our approximations with simulation results (marks in Figures 10a. and 10b.). The figures show that all approximations are excellent.

6 Conclusions

In this paper, we have considered a queueing system with a priority scheme with priority jumps. We have derived explicit expressions for the mean values of the queue contents and the packet delays, and determined approximate expressions for the tail probabilities of the studied quantities. It is thereby shown that non-geometric tails can occur for the content of the low-priority queue. In the numerical examples, we have furthermore illustrated the impact of the priority scheme on the performance of an output queue in a packet switch. The results of this paper can moreover be used to predict the performance of a queueing system with the HOL-PJ scheme and with time-correlation in the arrival process.

Appendix A: Darboux's theorem

Theorem 1.1 *Suppose* $X(z) = \sum_{n=0}^{\infty} x(n)z^n$ *with positive real coefficients* $x(n)$ *is analytic near* 0 *and has only algebraic singularities* α_k *on its circle of convergence* $|z| = R$*, in other words, in a neighbourhood of* α_k *we have*

$$X(z) \sim (1 - \frac{z}{\alpha_k})^{-\omega_k} G_k(z), \tag{24}$$

where $\omega_k \neq 0, -1, -2, \ldots$ *and* $G_k(z)$ *denotes a nonzero analytic function near* α_k*. Let* $\omega = max_k Re(\omega_k)$ *denote the maximum of the real parts of the* ω_k*. Then we have*

$$x(n) = \sum_j \frac{G_j(\alpha_j)}{\Gamma(\omega_j)} n^{\omega_j - 1} \alpha_j^{-n} + o(n^{\omega-1} R^{-n}), \tag{25}$$

with the sum taken over all j *with* $Re(\omega_j) = \omega$ *and* $\Gamma(\omega)$ *the Gamma-function of* ω *(with* $\Gamma(n) = (n-1)!$ *for* n *discrete).*

Acknowledgement

We want to thank the referees for their comments which lead to a considerable improvement of this paper. Note also that the second author is a Postdoctoral Fellow with the Fund for Scientific Research, Flanders (F.W.O.-Vlaanderen), Belgium.

References

[1] J.J. Bae and T. Suda. Survey of traffic control schemes and protocols in ATM networks. *ACM Transactions in Networking*, 2(5):508–519, 1994.

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

[3] H. Bruneel, B. Steyaert, E. Desmet, and G.H. Petit. Analytic derivation of tail probabilities for queue lengths and waiting times in ATM multiserver queues. *European Journal of Operational Research*, 76:563–572, 1994.

[4] R. Chipalkatti, J.F. Kurose, and D. Towsley. Scheduling policies for real-time and nonreal-time traffic packet switching node. In *IEEE INFOCOM '89*, pages 774–783, 1989.

[5] B.D. Choi, B.I. Choi, Y. Lee, and D.K. Sung. Priority queueing system with fixed length packet-train arrivals. *IEE Proceedings-Communications*, 145(2):331–341, 1998.

[6] B.I. Choi and Y. Lee. Performance analysis of a dynamic priority queue for traffic control of bursty traffics in ATM networks. *IEE Proceedings-Communications*, 148(3):181–187, 2001.

[7] D.I. Choi, B.D. Choi, and D.K. Sung. Performance analysis of priority leaky bucket scheme with queue-length-treshold scheduling policy. *IEE Proceedings-Communications*, 145(6), 1998.

[8] M. Drmota. Systems of functional equations. *Random Structures & Algorithms*, 10(1-2):103–124, 1997.

[9] A. Francini, F.M. Chiussi, R.T. Clancy, K.D. Drucker, and N.E. Idirene. Enhanced weighted round robin schedulers for accurate bandwidth distribution in packet networks. *Computer Networks*, 37(5):561–578, 2001.

[10] S. Fratini. Analysis of a dynamic priority queue. *Commun. Statist.-Stochastic Models*, 6(3):415–444, 1990.

[11] P.A. Ganos, M.N. Koukias, and G.K. Kokkinakis. ATM switch with multimedia traffic priority control. *European Transactions on Telecommunications*, 7(6):527–540, 1996.

[12] E. Gelenbe. Approximate analysis of coupled queueing in ATM networks. *IEEE Communications Letters*, 3(2):31–33, 1999.

[13] A. Gravey and G. Hebuterne. Mixing time and loss priorities in a single server queue. In *13th International Teletraffic Congress*, pages 147–152, 1991.

[14] J.S. Jang, S.H. Schim, and B.C. Shin. Analysis of DQLT scheduling policy for an ATM multiplexer. *IEEE Communications Letters*, 1(6):175–177, 1997.

[15] L. Kleinrock. *Queueing Systems*. Wiley, New York, 1975.

[16] C. Knessl, D.I. Choi, and C. Tier. A dynamic priority queue model for simultaneous service of voice and data traffic. *SIAM Journal on Applied Mathematics*, 63(2):398–422, 2002.

[17] A. Kuurne and A.P. Miettinen. Weighted round robin scheduling strategies in (e)gprs radio interface. In *Proceedings of the Vehicular Technology Conference (VTC2006-Fall)*, volume 5, pages 3155–3159, 2004.

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

[19] J.T. Lee and Y.H. Kim. Performance analysis of a hybrid priority control scheme for input and output queueing ATM switches. In *Proceedings INFOCOM '98*, 1998.

[20] Y. Lee and B.D. Choi. Queueing system with multiple delay and loss priorities for ATM networks. *Information Sciences*, 138(1-4):7–29, 2001.

[21] Y. Lim and J.E. Kobza. Analysis of a delay-dependent priority discipline in an integrated multiclass traffic fast packet switch. *IEEE Transactions on Communications*, 38(5):659–685, 1990.

[22] T. Maertens, J. Walraevens, and H. Bruneel. On priority queues with priority jumps. *Performance Evaluation*, 63(12):1235–1252, 2006.

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

[24] T. Maertens, J. Walraevens, M. Moeneclaey, and H. Bruneel. Performance analysis of a discrete-time queueing system with priority jumps. *International Journal of Electronics and Communications (AEÜ)*. Accepted for publication.

[25] M.B. Mamoun, J.-M. Fourneau, and N. Pekergin. Analyzing weighted round robin policies with a stochastic comparison approach. *Computers and Operations Research*, 35(8):2420–2431, 2008.

[26] A.K. Parekh and R.G. Gallager. A generalized processor sharing approach to flow control in integrated services network: the single node case. *IEEE/ACM Transactions on Networking*, 1(3):344–357, 1993.

[27] G.H. Petit, A. Buchheister, A. Guerrero, and P. Parmentier. Performance evaluation methods applicable to an ATM multi-path self-routing switching network. In *Proceedings of the Thirteenth International Teletraffic Congress (ITC 13)*, volume 14, pages 917–922, 1991.

[28] H. Shimonishi, M. Yoshida, F. Ruixue, and H. Suzuki. An improvement of weighted round robin cell scheduling in ATM networks. In *Proceedings of the 1997 Global Telecommunications Conference (GLOBECOM 1997)*, volume 2, pages 1119–1123, 1997.

[29] M. Shreedhar and G. Varghese. Efficient fair queueing using deficit round robin. *ACM SIGCOMM Computer Communication Review*, 25(4):231, 1995.

[30] D. Stiliadis and A. Varma. Latency-rate servers: a general model for analysis of traffic scheduling algorithms. *IEEE/ACM Transactions on Networking*, 6(5):611–624, 1998.

[31] J. Walraevens, B. Steyaert, and H. Bruneel. Performance analysis of a single-server ATM queue with a priority scheduling. *Computers and Operations Research*, 30(12):1807–1829, 2003.

[32] Y. Xiong, H. Bruneel, and G.H. Petit. Performance study of an ATM self-routing multistage switch with bursty traffic: simulation and analytic approximation. *European Transactions on Telecommunications*, 4(4):443–453, 1993.

[33] Y.-S. Yen, W. Chen, J.-C. Zhhuang, and H.-H. Chao. A novel sliding weighted fair queueing scheme for multimedia transmission. In *Proceedings of the 19th International Conference on Advanced Information Networking and Applications (AINA 2005)*, volume 1, pages 15–20, 2005.