Performance comparison of several priority schemes with priority jumps

Tom Maertens*, Joris Walraevens, 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 several discrete-time priority queues with priority jumps. In a priority scheduling scheme with priority jumps, real-time and non-real-time packets arrive in separate queues, i.e., the high- and low-priority queue respectively. In order to deal with possibly excessive delays however, non-real-time packets in the low-priority queue can in the course of time jump to the high-priority queue. These packets are then treated in the high-priority queue as if they were real-time packets. Many criteria can be used to decide when packets of the low-priority queue jump to the high-priority queue. Some criteria have already been introduced in the literature, and we first overview this literature. Secondly, we propose and analyse a new priority scheme with priority jumps. Finally, we extensively compare all cited schemes. The schemes all differ in their jumping mechanism, based on a certain jumping criterion, and thus all have a different performance. We show the pros and cons of each jumping scheme.

1 Introduction

An efficient priority scheme is of great importance in the design and construction of telecommunication networks. Modern telecommunication networks, i.e., originally data-oriented networks in which real-time applications are integrated, have to cope with the strict delay-related performance requirements of real-time traffic (e.g., voice and video). For this type of traffic, mean delay and delay jitter have to be small. For non-real-time traffic on the other hand, loss ratio and throughput are important performance metrics. Different types of traffic are thus characterised by different QoS (Quality of Service) standards. The ability to differentiate real-time, *delay-sensitive* traffic, and non-real-time, *delay-tolerant* traffic, is one of the main keys to a succesful telecommunication

*Corresponding author

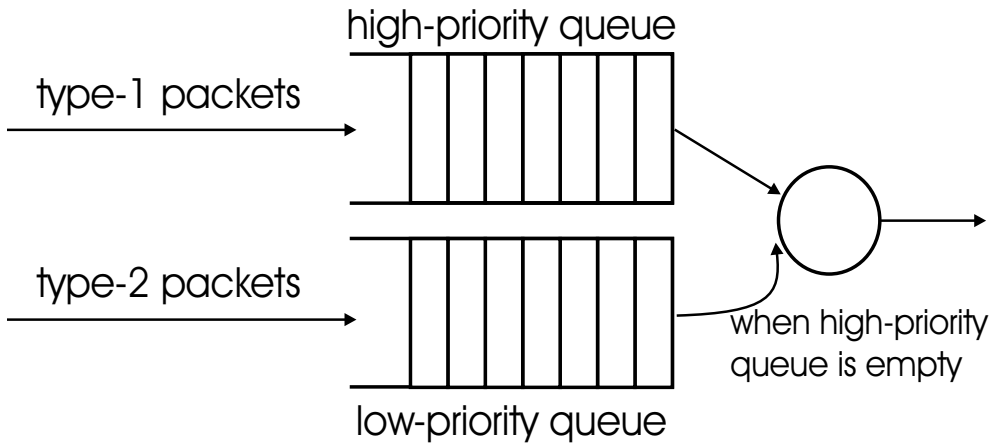


Figure 1: The static HOL priority scheme

network. Adopting priority scheduling schemes in routers, where multiple priority levels provide the transmission of different types of traffic, helps to achieve differentiation in the QoS constraints.

In the *static*, Head-Of-Line (HOL) priority scheduling scheme, transmission priority is always given to the delay-sensitive packets. This means that as long as there is delay-sensitive, *type-1* traffic present in the system, this traffic has transmission priority over delay-tolerant, *type-2* traffic. In the assumption that both types of traffic arrive in separate queues, packets of the low-priority queue are thus only transmitted when the high-priority queue is empty (see Figure 1). The HOL priority scheme does indeed provide low delays for the type-1 traffic (see e.g., [1, 3, 8]). The performance for type-2 traffic can however be severely degraded: the HOL priority scheme can cause excessive delays for the type-2 traffic when the network is highly loaded. Although type-2 traffic is delay-tolerant to a certain extent, excessive delays have to be avoided. Furthermore, some other negative effects can follow from excessive delays: the Transmission Control Protocol (TCP) e.g., could consider a type-2 packet with a too big delay as being lost, and would consequently decrease its transmission rate. This decreases the throughput, which is detrimental to data-applications. The decrease of the transmission rate is however unnecessary since the type-2 packet is not lost. The impact of the HOL priority scheme on the performance of a telecommunication network may thus be too disadvantageous in some cases. To deal with this so-called *starvation problem* of type-2 packets, several priority schemes with *priority jumps* have been proposed in the (recent) past.

In a priority scheme with priority jumps (in the remainder, called a *jumping scheme*), first introduced in [4], the priority level of packets can be adapted in the course of time. Concretely, packets of the low-priority queue can in time jump to the (tail of the) high-priority queue (see Figure

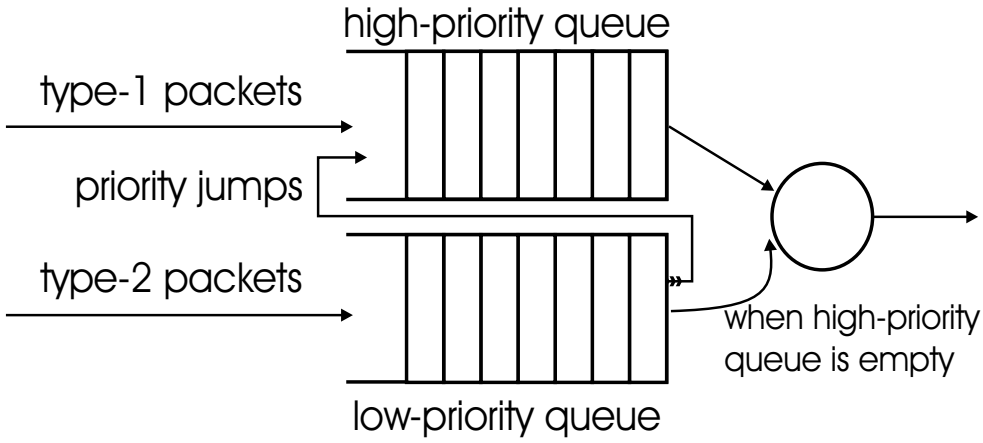


Figure 2: Priority schemes with priority jumps

2). Introducing jumping mechanisms in priority schemes tries to enhance priority scheduling by avoiding excessive delays for type-2 packets, while keeping the delay for type-1 traffic small. Many criteria can be used to decide when type-2 packets jump to the high-priority queue: a maximum queueing delay in the low-priority queue [4], a queue-length-threshold of the high- or low-priority queue [2, 6], a random jumping probability per time unit [5], the arrival characteristics of type-1 or type-2 traffic [7],...

In this paper, we consider several jumping schemes: the Head-Of-Line with Priority Jumps (HOL-PJ) scheme, the Head-Of-Line Merge-By-Probability (HOL-MBP) scheme, the Head-Of-Line Jump-Or-Serve (HOL-JOS) scheme, and the original Head-Of-Line Jump-If-Arrival (HOL-JIA^1) scheme. Most of these schemes are already analysed in the literature [2, 4, 5, 6, 7]. An overview of them is given in section 2. Note that we here consider discrete-time queueing models, i.e., time is assumed to be *slotted*. We furthermore propose and analyse a new scheme, namely the HOL-JIA^2 scheme. This is an improved version of the HOL-JIA^1 scheme studied in [7]. Via an analysis based on probability generating functions (pgfs), we derive the pgfs of the contents of the high- and low-priority queue, and the pgf of the delay of a type-1 packet. Moments are easily determined from the calculated pgfs. We also obtain the mean type-2 packet delay, although it seems difficult to determine an expression for the corresponding pgf. An extensive performance comparison of all considered jumping schemes is further presented in section 4.

The contribution of this paper first concerns the newly proposed JIA mechanism. Letting possible jumps depend on arriving type-2 packets makes HOL-JIA scheduling *self-adaptive*, i.e., the

arrival characteristics of type-2 traffic determine the effect of HOL-JIA scheduling on the performance of the telecommunication system. This self-adaptiveness seems to be promising, since no parameters have to be set by an operator. Secondly, for the HOL-JIA schemes, it appears to be complex to analyse the delay of a type-2 packet by deriving its corresponding pgf. However, a non-standard use of Little's theorem provides us a cunning trick to calculate the mean type-2 packet delay. Finally, the numerical examples clearly illustrate that *subtle* differences in the jumping mechanisms can yield considerable differences between their impact on the behaviour of a system. Finding the ideal jumping scheme is thus not straightforward.

The paper is organised as follows. In the following section, we briefly overview previous work on priority schemes with priority jumps. Section 3 contains the idea and the study of the newly proposed HOL-JIA^2 scheme. In section 4, we compare the performance of different priority jumping schemes. Conclusions and future work are formulated in section 5.

2 Overview of jumping schemes in the literature

2.1 The original HOL-PJ scheme

The original HOL-PJ jumping scheme was introduced in [4]. In this jumping scheme, a maximum queueing delay L is imposed on packets in the low-priority queue. Immediately after a packet's delay at the low-priority queue equals L, the packet jumps to the tail of the high-priority queue. An exact analysis of the queue contents and the delay distributions in this queueing system is very cumbersome, since it is necessary to keep track of the waiting times of the packets in the low-priority queue. In [4], the authors therefore develop a queueing model for calculating the average queueing delays of both types of traffic and for heuristically approximating the delay distributions. Possible disadvantages of this jumping scheme are the processing overhead required for monitoring packets for time-out, and the additional hardware necessary to keep timestamps of all the packets in the low-priority queue.

2.2 The HOL-MBP scheme

The Head-Of-Line Merge-By-Probability (HOL-MBP) jumping scheme (see [5]) was mainly proposed to evade the disadvantages of the HOL-PJ scheme. In the HOL-MBP scheme, a parameter

β is introduced, and defined as the probability that at the end of each slot the *total* content of the low-priority queue jumps to the tail of the high-priority queue. Or, in other words, β gives the probability that at the end of each slot the contents of the high- and low-priority queue are *merged*. Maertens et al [5] have derived the pgfs of the contents of the high- and low-priority queue, and the pgfs of the delays of both types of traffic. This pgf approach then easily led to expressions for performance measures (such as mean values and variances). A comparison study (see [5]) moreover shows that the (simulated) performance of the HOL-PJ scheme hardly differs from the performance of the HOL-MBP scheme. The latter can however be implemented more easily, and is analytically tractable.

2.3 The HOL-JOS scheme

In the HOL-MBP scheme, the total content of the low-priority queue can jump at the end of each slot. Since this could mean that lots of packets have to be moved simultaneously (especially when β is extremely low), we have proposed some other jumping schemes. In the Head-Of-Line Jump-Or-Serve (HOL-JOS) jumping scheme (see [6]), only the packet at the HOL-position of the low-priority queue can jump to the high-priority queue. This possible jump at the beginning of each slot depends on the content of the high-priority queue at the beginning of the slot, i.e., when this queue is non-empty, the packet jumps. When the high-priority queue is empty on the other hand, the *HOL-packet* of the low-priority queue is immediately transmitted (or, served). Maertens et al [6] have obtained the pgfs of the contents of the high- and low-priority queue, and the pgfs of the delays of both types of traffic. From these pgfs, again expressions for some interesting performance measures (such as mean values, variances, and approximate tail probabilities of the studied stochastic variables) are efficiently derived.

2.4 The HOL-JIA^1 scheme

The *flow* of delay-tolerant, type-2 traffic into the high-priority queue may be too drastic in the HOL-JOS scheme. To somehow restrict this flow, an extra jumping condition can be introduced. As in the HOL-JOS scheme, only the packet at the HOL-position of the low-priority queue can jump to the high-priority queue in the first Head-Of-Line Jump-If-Arrival (HOL-JIA^1) jumping scheme (see [7]). However, the possible jump at the end of a slot does not only depend on the contents

of the high-priority queue at the beginning of the slot, but also on the number of type-2 packets that arrive in that slot. Specifically, during a slot in which a packet of the high-priority queue is transmitted, the HOL-packet of the low-priority queue jumps to the high-priority queue if, and only if, type-2 packets arrive during that slot. Note that in this scheme, arriving type-2 packets are not allowed to jump immediately upon arrival (i.e., at the end of their arrival slot). In [7], we have derived the pgfs of the contents of the high-and low-priority queue, and the pgf of the delay of a type-1 packet. Related moments are then easily derived from the obtained pgfs.

3 The HOL-JIA^2 scheme

3.1 Idea

In the HOL-JIA^1 scheme, it is assumed that arriving type-2 packets are not allowed to jump at the end of their arrival slot. However, when few type-2 packets arrive at the system, the type-2 packet at the HOL-position of the low-priority queue may experience an excessive delay since it has to wait for another, rare type-2 arrival. To avoid this situation, we can allow type-2 packets to jump immediately upon arrival. The type-2 packet that jumps at the end of a slot to the high-priority queue is thus either a packet that was already in the low-priority queue at the beginning of that slot, or, when the low-priority queue was empty at the beginning of the slot, a packet that arrived during the slot. This jumping scheme is defined as the HOL-JIA^2 scheme. This is a newly proposed jumping scheme, and we will therefore first briefly describe its analysis. Since this analysis is rather similar to the analyses of the HOL scheme (see [8]) and the HOL-JOS scheme (see [6]), we only give a sketch of the analysis and further refer to [6] and [8] for more details. We derive the pgfs of the contents of the high- and low-priority queue, and the pgf of the delay of a type-1 packet. This allows us to calculate moments, such as mean values and variances. A procedure to calculate the mean delay of a type-2 packet is furthermore proposed.

3.2 Mathematical model

We consider a discrete-time queueing system with two queues of infinite capacity, and with one transmission channel. Two types of traffic arrive at the system: packets of type 1, which are stored in the first queue, and packets of type 2, which are stored in the second. The numbers of per-slot

type-1 and type-2 arrivals are characterised by their joint pgf $A(z_1, z_2)$, and the marginal pgf's $A_T(z) = A(z, z)$, $A_1(z) = A(z, 1)$ and $A_2(z) = A(1, z)$ (with $\lambda_j = A_j'(1)$ the arrival rate of type-j packets, and $\lambda_T = \lambda_1 + \lambda_2$ the total arrival rate). The transmission times equal one slot, and packets of the first queue have a higher priority than those of the second queue. So, whenever there are packets present in the high-priority queue, they have transmission priority; only when the high-priority queue is empty, packets of the low-priority queue can be transmitted (see Figure 2).

The system is finally influenced by the following jumping mechanism: at the end of each slot in which a packet of the high-priority queue is transmitted and in which type-2 packets arrive at the system, the packet at the HOL-position of the low-priority queue jumps to the high-priority queue. When the low-priority queue is empty at the beginning of such a slot, one of the newly arriving type-2 packets jumps to the high-priority queue. Since the jump occurs at the end of the slot, the jumping packet is queued behind the type-1 arrivals during the same slot.

3.3 Analysis of the system contents

Let us define $u_{1,k}$ and $u_{2,k}$ as the contents of the high- and low-priority queue at the beginning of slot k respectively, and $u_{T,k}$ as the total system content at the beginning of slot k. We hereby assume that the packet in transmission (if any) is part of the queue that is "served" in that slot. The joint pgf of $u_{1,k}$ and $u_{2,k}$ is denoted by $U_k(z_1, z_2) \triangleq \mathrm{E}\left[z_1^{u_{1,k}} z_2^{u_{2,k}}\right]$. The following system equations can be derived:

- if $u_{1,k} = 0$:

$$\begin{cases} u_{1,k+1} = a_{1,k} \\ u_{2,k+1} = [u_{2,k} - 1]^+ + a_{2,k} \end{cases}, \tag{1}$$

- if $u_{1,k} > 0$:

  - if $a_{2,k} = 0$:

$$\begin{cases} u_{1,k+1} = u_{1,k} - 1 + a_{1,k} \\ u_{2,k+1} = u_{2,k} \end{cases}, \tag{2}$$

– if $a_{2,k} > 0$:

$$\begin{cases} u_{1,k+1} = u_{1,k} + a_{1,k} \\ u_{2,k+1} = u_{2,k} + a_{2,k} - 1 \end{cases}, \tag{3}$$

where $[\cdots]^+$ denotes the maximum of the argument and zero. When the high-priority queue is empty at the beginning of slot k, a packet of the low-priority queue (if any) is transmitted during slot k (Eq. (1)). When the high-priority is non-empty at the beginning of slot k, a packet of the high-priority queue is transmitted. In this case, a low-priority packet jumps at the end of slot k to the high-priority queue iff $a_{2,k} > 0$ (Eqs. (2) and (3)). Introducing pgfs in the system equations and letting $k \to \infty$ establishes a *steady-state* relationship between $U(z_1, z_2)$, $U(0, z_2)$ and $U(0, 0)$:

$$U(z_1, z_2) = \frac{z_1(z_2 - 1)A(z_1, z_2)U(0, 0) + (z_1 - z_2)A(z_1, 0)U(0, z_2)}{z_1 z_2 - z_1 A(z_1, z_2) - (z_2 - z_1)A(z_1, 0)}. \tag{4}$$

Using the normalization condition and Rouché's theorem to obtain $U(0, 0)$ and $U(0, z_2)$ respectively (see e.g., [6] and [8] for a similar procedure), finally yields the joint pgf of the contents of both queues at the beginning of a random slot in the steady state:

$$U(z_1, z_2) = \frac{(1 - \lambda_T)(z_2 - 1)\Big(z_1 A(z_1, z_2)(z_2 - A(Y(z_2), z_2)) + (z_1 - z_2)A(z_1, 0)A(Y(z_2), z_2)\Big)}{\Big(z_2 - A(Y(z_2), z_2)\Big)\Big(z_1 z_2 - z_1 A(z_1, z_2) - (z_2 - z_1)A(z_1, 0)\Big)}, \tag{5}$$

with

$$Y(z) \triangleq \frac{Y(z)}{z} A(Y(z), z) + \frac{(z - Y(z))}{z} A(Y(z), 0). \tag{6}$$

Substituting z_1 and z_2 in (5) by the appropriate values, yields the marginal pgfs $U_T(z)$, $U_1(z)$ and $U_2(z)$ of the total system content, and of the contents of the high- and low-priority queue respectively:

$$\begin{aligned} U_T(z) &\triangleq \lim_{k \to \infty} \mathrm{E}\left[z^{u_{T,k}}\right] = U(z, z) \\ &= \frac{(1 - \lambda_T)A_T(z)(z - 1)}{z - A_T(z)}, \\ U_1(z) &\triangleq \lim_{k \to \infty} \mathrm{E}\left[z^{u_{1,k}}\right] = U(z, 1) \end{aligned} \tag{7}$$

$$= \frac{(A_2(0)-\lambda_1)(z-1)A(z,0)}{A_2(0)\Big(z-A(z,0)-z(A_1(z)-A(z,0))\Big)}, \tag{8}$$

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

$$= \frac{(1-\lambda_T)(z-1)\Big(A_2(z)(z-A(Y(z),z)) - A_2(0)(z-1)A(Y(z),z)\Big)}{\Big(z-A(Y(z),z)\Big)\Big(z-A_2(z)-(z-1)A_2(0)\Big)}, \tag{9}$$

with $Y(z)$ implicitly defined by (6). By taking the first derivatives of (6)-(9), for $z = 1$, and by making extensive use of de l'Hopital's rule, we get expressions for $Y'(1)$ (necessary for further derivations) and for $\mathrm{E}[u_T]$, $\mathrm{E}[u_1]$ and $\mathrm{E}[u_2]$, i.e., the mean values of the total system content, and of the contents of the high- and low-priority queue respectively. For future reference, we here give the expression for $\mathrm{E}[u_T]$:

$$\mathrm{E}[u_T] = \lambda_T + \frac{\lambda_{TT}}{2(1-\lambda_T)}, \tag{10}$$

with $\lambda_{TT} \triangleq A_T''(1)$. Note further that expressions for higher moments can be obtained by taking higher order derivatives of the respective pgfs, for $z = 1$.

3.4 Analysis of the packet delay

Since the possible jump of the HOL-packet of the low-priority queue takes place *at the end* of a slot, *all* type-1 packets that arrive during a particular slot k, including a "tagged" type-1 packet, are queued in front of the possibly jumping packet. The delay of the tagged type-1 packet, i.e., the number of slots between the end of the packet's arrival slot and the end of its departure slot, thus only depends on the content of the high-priority queue at the beginning of slot k ($u_{1,k}$). So, $D_1(z)$ (the pgf of the type-1 packet delay) can be easily expressed in terms of $U_1(z)$ (see e.g., [5, 8] for more details), for which an expression was found in the previous subsection (see Eq. (8)). This leads to

$$D_1(z) = \frac{(A_2(0)-\lambda_1)z(A_1(z)-1)(1-A_1(z)+A(z,0))}{\lambda_1 A_2(0)\Big(z-A(z,0)-z(A_1(z)-A(z,0))\Big)}. \tag{11}$$

By taking the first derivative of (11) for $z = 1$, we find an expression for $\mathrm{E}[d_1]$, i.e., the mean delay of a type-1 packet:

$$\mathrm{E}[d_1] = 1 + \frac{\lambda_1(\lambda_1 - A^{(1)}(1,0))}{A_2(0)(A_2(0) - \lambda_1)} + \frac{\lambda_{11}A_2(0)}{2\lambda_1(A_2(0) - \lambda_1)}, \tag{12}$$

with $A_2(0)$ the probability of having no type-2 arrivals in a slot, $A^{(1)}(1,0) \triangleq \left.\dfrac{\partial A(z_1,z_2)}{\partial z_1}\right|_{z_1=1,z_2=0}$, and $\lambda_{11} \triangleq A_1''(1)$. By taking higher order derivatives for $z = 1$, expressions for higher moments can also be obtained.

The total number of slots that a tagged type-2 packet spends in the system can be expressed as $d_2 = [u_{T,k} - 1]^+ + a_{1,k} + f_{2,k} + p + 1$, with k the arrival slot of the tagged type-2 packet, $u_{T,k}$ the total system content at the beginning of slot k, and $a_{1,k}$ and $f_{2,k}$ the number of type-1 and type-2 packets that arrive during slot k, but which have to be transmitted before the tagged packet. The quantity p represents the number of type-1 packets that arrive during slots following the tagged packet's arrival slot, but which have to be transmitted before the tagged one (because of the priority scheduling). In priority models, p is typically described as a sum of *sub-busy periods*, with a sub-busy period being defined as the number of type-1 arrivals during the time that the tagged packet is in a certain position in the low-priority queue. In this particular model however, these sub-busy periods depend on the evolution of the high-priority queue, which leads to correlation between subsequent sub-busy periods. This is usually not the case in previously studied priority models (see e.g., [5, 8]). As a consequence, an exact analysis of the delay of a tagged type-2 packet is rather complex, and still an open issue at the moment.

Although it seems complicated to derive an explicit expression for the pgf of the delay of a type-2 packet, it is however possible to calculate the mean delay of a type-2 packet. It should first be mentioned that $\mathrm{E}[u_j] = \lambda_j \mathrm{E}[d_j]$ $(j = 1, 2)$ does *not* hold, as one would at first expect according to Little's theorem. The reason for this is that in the calculations of the system contents jumped type-2 packets are treated as part of the content of the high-priority queue. This is not the case in the calculations of the packet delay. Or, in other words, Little's theorem does not hold with respect to each queue separately, because the system contents is defined on a "queue"-basis, while the packet delay is defined on a "packet"-basis. For the total system on the contrary, Little's theorem does hold: $\mathrm{E}[u_T] = \lambda_T \mathrm{E}[d]$ (with $\mathrm{E}[d]$ the mean delay of an arbitrary - type-1 or type-2

- packet). Substituting this relationship in $\mathrm{E}[d] = \frac{\lambda_1}{\lambda_T}\mathrm{E}[d_1] + \frac{\lambda_2}{\lambda_T}\mathrm{E}[d_2]$ (with $\frac{\lambda_j}{\lambda_T}$ the probability that a random arriving packet is of type j, and with $j = 1, 2$), we find

$$\mathrm{E}[d_2] = \frac{\mathrm{E}[u_T] - \lambda_1 \mathrm{E}[d_1]}{\lambda_2}. \tag{13}$$

Since $\mathrm{E}[u_T]$ as well as $\mathrm{E}[d_1]$ are calculated (see Eqs. (10) and (12) respectively), we are thus able to derive an expression for the mean delay of a type-2 packet.

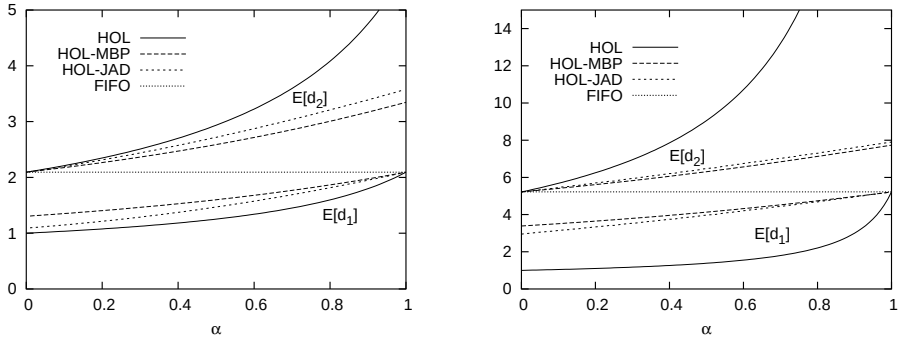
4 Performance comparison

In this section, we compare the performance of the various jumping schemes for a specific arrival process. We thereby especially focus on the comparison of the newly proposed HOL-JIA^2 scheme with the other jumping schemes. Since the jumping schemes were mainly introduced to lower the delay of delay-tolerant, type-2 traffic without having a too negative effect on the delay of delay-sensitive, type-1 traffic, we focus on the mean packet delays of both types of traffic to compare them. The performance comparison is done in three steps. We first briefly compare the jumping schemes that have a jumping parameter, i.e., the HOL-MBP scheme and the HOL-PJ scheme (see also [5]). Afterwards, we make an extensive comparison of the jumping schemes that do not have an extra jumping parameter: the HOL-JOS scheme, the HOL-JIA^1 scheme, and the HOL-JIA^2 scheme. We balance the pros and cons of each jumping scheme. Finally, we illustrate the (dis)advantages of having a jumping parameter. Note that we have also included the static HOL priority scheme (see e.g., [8]) and sometimes the plain First-In-First-Out (FIFO) scheme, for reference purposes.

Except for the jumping mechanism, the model of all schemes is chosen identical, in order to provide a fair and valid comparison. More precisely, the mathematical model of subsection 3.2 is adopted. Unless otherwise stated, we furthermore consider a two-dimensional binomial arrival process, fully characterised by the joint pgf

$$A(z_1, z_2) = \Big(1 - \frac{\lambda_1}{N}(1 - z_1) - \frac{\lambda_2}{N}(1 - z_2)\Big)^N, \tag{14}$$

with $N = 16$ in the figures. The arrival rate of type-j traffic is then given by λ_j ($j = 1, 2$), and the



a. for $\lambda_T = 0.7$ b. for $\lambda_T = 0.9$

Figure 3: Mean value of packet delays versus α when $L = 4$ ($\beta = 0.25$)

total arrival rate by $\lambda_T = \lambda_1 + \lambda_2$. We define α as the fraction of type-1 traffic in the overall traffic mix (i.e., $\alpha = \lambda_1/\lambda_T$). It should be noticed that (14) specifies the arrival process to a particular queue in an output-queueing packet switch with Bernoulli arrivals at its inlets, and with uniform routing. In the case of FIFO scheduling, the packet delay is the same for type-1 and type-2 packets (independent of α), and can thus be calculated as if only one type of traffic arrives according to an arrival process with pgf $A(z, z)$.

4.1 Initial remark

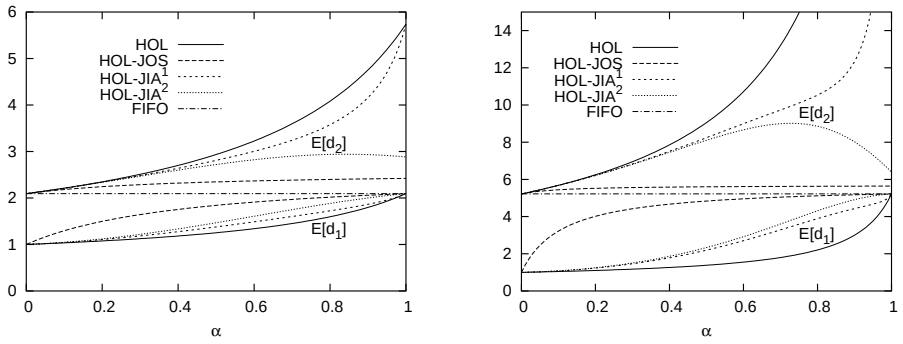
Since all considered schemes (i.e., the FIFO scheme, the static HOL scheme, and all jumping schemes) are work-conserving, and since all packets have the same transmission time (one slot), the mean total system content $\mathrm{E}[u_T]$ is the same for all schemes. According to Little's theorem, this means that also the mean delay of an arbitrary packet $\mathrm{E}[d]$ is the same for all considered schemes. As a consequence, the scheduling scheme has no influence on $\mathrm{E}[d]$. We furthermore know that for each scheme $\mathrm{E}[d] = \alpha\mathrm{E}[d_1] + (1-\alpha)\mathrm{E}[d_2]$ (with $\mathrm{E}[d_j]$ the mean type-j packet delay, $j = 1, 2$), because the probabilities that a random arriving packet is of type-1 and type-2 equal α and $1-\alpha$ respectively. Assuming α fixed then, a lowered $\mathrm{E}[d_2]$ for a scheme with priority jumps thus implies an increased $\mathrm{E}[d_1]$. Note also that when $\alpha \to 0$, $\mathrm{E}[d_2]$ converges for all priority schemes and for the FIFO scheme, since basically only type-2 packets arrive at the system. When $\alpha \approx 1$ (i.e., when the overall traffic mix only exists of type-1 traffic) on the other hand, $\mathrm{E}[d_1]$ is the same for all schemes.

4.2 Schemes with a jumping parameter

We first briefly compare the (simulated) performance of the HOL-PJ scheme with the (calculated) performance of the HOL-MBP scheme. To make a valid comparison between both schemes, it is necessary to define a proper relation between the jumping parameters of both schemes, i.e., the jumping probability β of the HOL-MBP scheme and the delay limit L of the HOL-PJ scheme. It is easily verified that the number of slots until a type-2 packet jumps is deterministically equal to L in the latter scheme, while it is geometrically distributed with parameter $(1-\beta)$ in the first. Choosing L equal to $(1-\beta)/\beta$ seems a natural choice since this equalises their respective means. Note that the variances of the two distributions then equal 0 and $(L+1)L$ respectively.

In [5], we have already illustrated the influence of the total arrival rate on the mean packet delays of both types of traffic. Here, we examine the influence of the traffic mix on $\mathrm{E}[d_1]$ and $\mathrm{E}[d_2]$ for the two jumping schemes. Figures 3a. and 3b. show the mean packet delays of both types of traffic when $L = 3$ (and thus $\beta = 0.25$), as functions of α, for $\lambda_T = 0.7$ and $\lambda_T = 0.9$ respectively. We notice that the HOL-PJ scheme leads to a lower $\mathrm{E}[d_1]$, while the HOL-MBP scheme performs better for $\mathrm{E}[d_2]$. The difference, altough small, is firstly caused by the different variances of the number of slots until a type-2 packet jumps for both schemes. This variance is namely larger in the HOL-MBP scheme than in the HOL-PJ scheme. A larger variance means that type-2 packets are subject to more varying waiting times in the low-priority queue. It is then possible that the low-priority queue builds up, and that a lot of packets are transferred to the high-priority queue. This effect is further increased by the fact that the *total* content of the low-priority queue jumps in the HOL-MBP scheme. Arriving type-1 packets thus suffer from larger delays due to the 'burstiness' of the number of jumping packets, resulting in a higher $\mathrm{E}[d_1]$ for the HOL-MBP scheme. This phenomenen especially appears when α is low (i.e., when a lot of type-2 packets enter the system). E.g., when $\alpha \approx 0$ in the HOL-MBP scheme, a rare type-1 arrival can be preceded by a merge of the high-priority queue and a "big" low-priority queue. In the HOL-PJ scheme, jumps are more spread over time. The exceptionally arriving type-1 packet is thus expected to be delayed longer in the HOL-MBP scheme than in the HOL-PJ scheme. When α is high, the low-priority queue cannot build up as much since there are fewer type-2 arrivals.

In general, we can state that the HOL-PJ scheme and the HOL-MBP scheme have a similar performance (provided the right choices of the jumping parameters β and L) with regard to the



a. for $\lambda_T = 0.7$ b. for $\lambda_T = 0.9$

Figure 4: Mean value of packet delays versus α

mean packet delays of both types of traffic; HOL-MBP scheduling results in a slightly lower mean type-2 packet delay. The figures also illustrate that the mean type-2 packet delay can be decreased significantly by introducing a jumping mechanism, especially when α is high.

4.3 Schemes without a jumping parameter

We furthermore compare the three jumping schemes that do not have a jumping parameter. First note that the curves of E$[d_1]$ (i.e., the mean type-1 packet delay) and E$[d_2]$ (i.e., the mean type-2 packet delay) for the HOL-JIA schemes will always lie between those for the HOL scheme and the HOL-JOS scheme. This is due to the different jumping mechanisms in the stated schemes. Indeed, in the HOL scheme, there are no jumps at all, while in the HOL-JOS scheme, the HOL-packet of the low-priority queue jumps *in every slot* in which a packet of the high-priority queue is transmitted. In the HOL-JIA schemes, jumps occur but are restricted to those slots where type-2 packets arrive, and the number of jumps is thus more controlled than in the HOL-JOS scheme. In the remainder of this subsection, we perform a thorough comparison of the schemes and examine their impact on the mean packet delays of both types of traffic, as functions of the traffic mix, the total arrival rate, the arrival rate of type-2 traffic, and the variance of the number of type-2 arrivals in a slot respectively. We thereby pay special attention to the newly introduced HOL-JIA^2 scheme.

4.3.1 Impact of the traffic mix

In Figures 4a. and 4b., we show the mean packet delays of both types of traffic for $\lambda_T = 0.7$ and $\lambda_T = 0.9$ respectively, as functions of α. We first notice that when $\alpha \to 0$ (i.e., when the overall

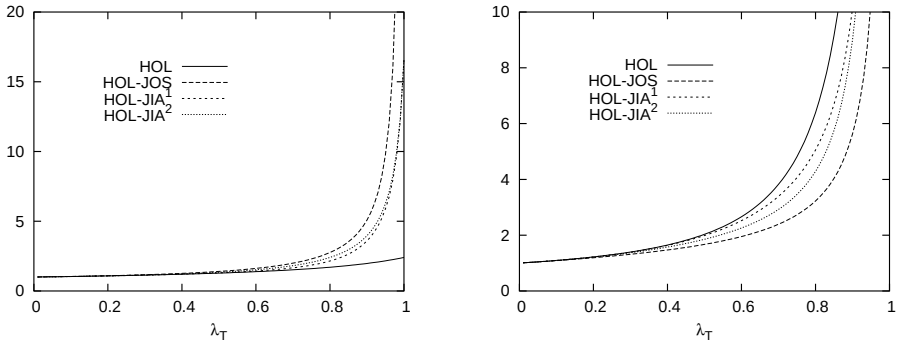
traffic mix basically only exists of type-2 traffic), $\mathrm{E}[d_1] \to 1$ for all considered priority schemes. When $\alpha \approx 0$ in these schemes, type-2 packets are immediately transmitted out of the low-priority queue, and the probability that an exceptionally arriving type-1 packet enters an empty high-priority queue thus approximately equals 1. The delay of the type-1 packet is then not influenced by type-2 packets. As a consequence, $\mathrm{E}[d_1]$ for these priority schemes equals the transmission time of one packet, i.e., one slot.

Secondly, it is seen from Figures 4a. and 4b. that when α is low (i.e., when few type-1 packets arrive at the system), the curves for the HOL-JIA schemes, of both $\mathrm{E}[d_1]$ and $\mathrm{E}[d_2]$, lie near the curves for the HOL scheme. When α is low in the HOL-JIA schemes, the high-priority queue is often empty and few type-2 packets jump to the high-priority queue. Hence, both type-1 and type-2 packets behave similarly as in the HOL scheme. In the HOL-JOS scheme, a type-2 packet jumps to the high-priority queue every slot where both queues are non-empty. As a result, a considerably higher $\mathrm{E}[d_1]$ and lower $\mathrm{E}[d_2]$ than for the HOL(-JIA) schemes is observed.

Furthermore, when α increases, more type-1 packets arrive at the system, and more type-2 packets thus suffer from larger delays. As a consequence, both $\mathrm{E}[d_1]$ and $\mathrm{E}[d_2]$ increase. In these jumping schemes, increasing α also means that the probability of having an empty high-priority queue decreases and that more occasions arise for type-2 packets to jump. A higher increase of $\mathrm{E}[d_1]$ and a repressed increase of $\mathrm{E}[d_2]$ compared to the HOL scheme is the logic consequence. For the HOL-JOS scheme, this further implies that the curves of $\mathrm{E}[d_1]$ and $\mathrm{E}[d_2]$ lie close to the curve for the FIFO scheme, especially when the total arrival rate is high (see Figure 4b.).

When α is high, one can see that the HOL-JIA schemes perform quite similar with respect to $\mathrm{E}[d_1]$. The corresponding curves lie somewhere in the middle between the curve for the HOL-JOS scheme and the curve for the HOL scheme. The HOL-JIA schemes however show a large performance difference in the mean type-2 packet delays. When α is high in the HOL-JIA^2 scheme, a large portion of the limited number of arriving type-2 packets can immediately jump to the high-priority queue upon arrival. This results in a lower $\mathrm{E}[d_2]$ than for the HOL-JIA^1 scheme. The price to pay, i.e., a higher $\mathrm{E}[d_1]$ is limited. E.g., when $\lambda_T = 0.9$ and $\alpha = 0.9$ (see Figure 4b.), $\mathrm{E}[d_2]$ decreases from about 12.3 for HOL-JIA^1 to 8 for HOL-JIA^2, with only a small increase for $\mathrm{E}[d_1]$ (from about 4.4 to about 4.9).

Finally, when $\alpha \to 1$, we observe a totally different behaviour for $\mathrm{E}[d_2]$ for both HOL-JIA



a. E[d_1] b. E[d_2]

Figure 5: Mean value of packet delays versus λ_T for $\alpha = 0.75$

schemes. When $\alpha \approx 1$ in the HOL-JIA schemes, an exceptionally arriving type-2 packet has a large probability of entering an empty low-priority queue. In the HOL-JIA^1 scheme, this type-2 packet is only allowed to jump if another type-2 packet would arrive, which is not possible. To be transmitted, the packet thus basically has to wait in the low-priority queue until the high-priority queue becomes empty. So, when $\alpha \approx 1$, no jumps occur and the HOL-JIA^1 scheme performs as the HOL scheme. Hence, E[d_2] converges for the HOL-JIA^1 scheme and the HOL scheme for $\alpha \to 1$. In the HOL-JIA^2 scheme on the other hand, an exceptionally arriving type-2 packet jumps immediately upon arrival and can thus be transmitted within a relatively short time. Here, a similar behaviour is noticed as in the HOL-JOS scheme. Dropping the restriction that type-2 packets are not allowed to jump at the end of their arrival slot thus prevents a type-2 packet from wasting time in the low-priority queue.

Note that the mean type-2 packet delay for the HOL-JIA^2 scheme reaches a maximum as a function of α. This is due to two counteracting mechanisms. First, increasing α means more type-1 packets and longer waiting times for type-2 packets (because of the priority scheduling). Increasing α also means that less type-2 packets actually have to wait in the low-priority queue with a decrease of the delay as a consequence. The last effect is however only dominant when α is large.

4.3.2 Impact of the total arrival rate

Figure 5a. shows the mean type-1 packet delay for $\alpha = 0.75$, as function of λ_T. Obviously, E[d_1] increases when λ_T increases. Furthermore, when $\lambda_T \to 1$, E[d_1] $\to \infty$ for the HOL-JOS scheme. This is not the case for the other schemes. Thus for high λ_T, E[d_1] can still be limited for the HOL-JIA schemes, though it is increased compared to HOL. We also see that the two HOL-JIA schemes

perform quite similar with regard to the mean type-1 packet delay. In fact, $\mathrm{E}[d_1]$ is equal for both HOL-JIA schemes for $\lambda_T \to 1$. Indeed, the content of the low-priority queue then approaches infinity due to the system becoming unstable, and the jumping type-2 packets are thus always packets that were already in the queue at that time.

In Figure 5b., we have depicted the mean type-2 packet delay for $\alpha = 0.75$, as function of λ_T. When λ_T is high, one can clearly see the effect of the different jumping mechanisms. As expected, the HOL scheme performs the worst and the HOL-JOS scheme the best with regard to $\mathrm{E}[d_2]$. The performance of the HOL-JIA schemes lies somewhere in the middle, with HOL-JIA^2 better performing than HOL-JIA^1. For intuitive explanations of these observations, we refer to subsection 4.3.1.

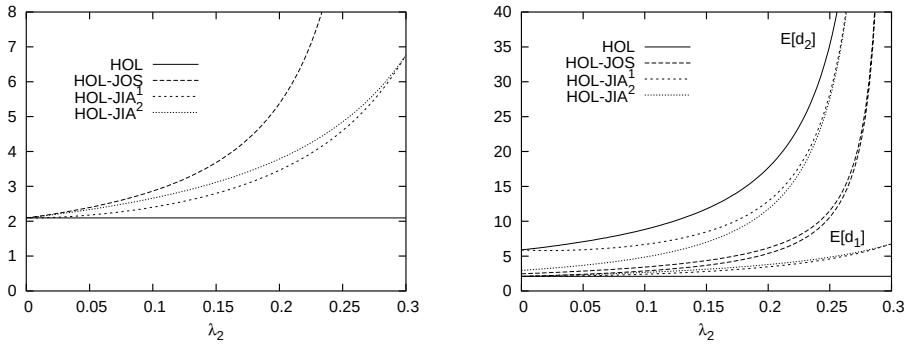
4.3.3 Impact the arrival rate of type-2 traffic

Up to now, we have considered a *two-dimensional* binomial arrival process, fully determined by expression (14). It is clear that for this specific arrival process the number of arrivals of both types of traffic during a time slot are correlated. This arrival process is thus not entirely suitable for studying the influence of the arrival characteristics of one single type of traffic on the performance of the various priority schemes. For this purpose, we use an arrival process in which the numbers of arrivals of both types of traffic are uncorrelated in a slot, i.e., $A(z_1, z_2) = A_1(z_1)A_2(z_2)$. We consider

$$\begin{cases} A_1(z) = \left(1 - \dfrac{\lambda_1}{N}(1-z)\right)^N \\ A_2(z) = \dfrac{1}{1 + \lambda_2 - \lambda_2 z} \end{cases}, \tag{15}$$

with $N = 16$. The number of type-1 arrivals during a slot is thus binomially distributed, while we assume a geometric distribution for the number of type-2 arrivals. The arrival rate of type-j traffic is again given by λ_j $(j = 1, 2)$, and the fraction α of type-1 traffic in the overall traffic mix is still defined as λ_1/λ_T (with $\lambda_T = \lambda_1 + \lambda_2$).

In Figure 6a., we show the mean packet delay of type-1 traffic for $\lambda_1 = 0.7$, as function of λ_2. Obviously, the number of arriving type-2 packets has no impact on $\mathrm{E}[d_1]$ for the HOL scheme (since there are no jumps in this scheme). In the jumping schemes, the number of jumps increases with the number of type-2 arrivals. As a consequence, $\mathrm{E}[d_1]$ increases when λ_2 increases. Note that



a. E[d_1] b. E[d_1] and E[d_2]

Figure 6: Mean value of packet delays versus λ_2 for $\lambda_1 = 0.7$

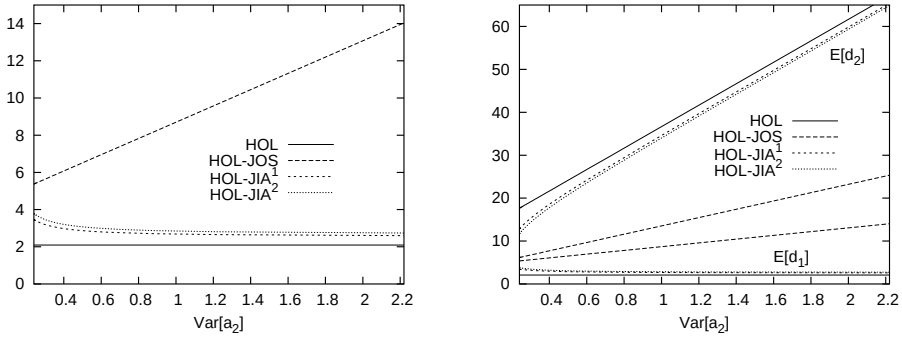
one would not immediately expect an influence of λ_2 on E[d_1] for the HOL-JIA^2 scheme: λ_2 does not appear in expression (12). However, λ_2 has an influence on E[d_1] via $A_2(0)$. The probability of no type-2 arrivals in a slot namely decreases when λ_2 increases. In Figure 6b., we have also included the mean type-2 packet delay. We see that E[d_2] depends more on the type-2 arrival rate than E[d_1], which is quite logic. Note also the small difference between E[d_1] and E[d_2] for the HOL-JOS scheme, while the HOL-JIA schemes achieve more differentiation in the delay of both types of traffic.

When we compare both HOL-JIA schemes, we can conclude the following: for low λ_2, the HOL-JIA^2 scheme smartly performs better than the HOL-JIA^1 scheme with respect to E[d_2], while both HOL-JIA schemes behave rather similarly for E[d_1]. This is again because a reasonable portion of the arriving type-2 packets directly jump to the high-priority queue. When λ_2 is high, both schemes perform identically due to the low-priority queue being non-empty with high probability.

4.3.4 Impact of the variance of the number of type-2 arrivals

The arrival process of (15) is sufficient for studying the effect of the mean number of type-2 arrivals on the behaviour of the various jumping schemes. To study the impact of the variance of the number of type-2 arrivals in a slot however, we assume the following arrival process:

$$\begin{cases} A_1(z) = \left(1 - \dfrac{\lambda_1}{N}(1 - z_1)\right)^N \\ A_2(z) = p\dfrac{1}{1 + \lambda_{2,1} - \lambda_{2,1}z} + (1-p)\dfrac{1}{1 + \lambda_{2,2} - \lambda_{2,2}z} \end{cases} . \tag{16}$$



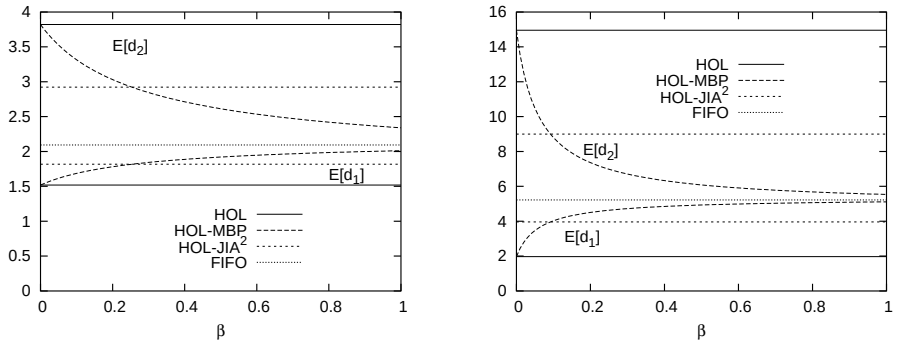
a. E[d_1] b. E[d_1] and E[d_2]

Figure 7: Mean value of packet delays versus Var[a_2] for $\lambda_1 = 0.7$, $\lambda_2 = 0.2$, and $\lambda_{2,1} = 0.1$

N equals 16 in the figures. The number of type-2 arrivals are now assumed to be distributed according to a weighted sum of two geometrics. We choose the weight such that the arrival rate of type-2 traffic remains constant, i.e., $\lambda_2 = p\lambda_{2,1} + (1-p)\lambda_{2,2} = 0.2$. We set $\lambda_{2,1} = 0.1$, and vary $\lambda_{2,2}$ between 0.2 and ∞. The variance of the number of type-2 arrivals then varies from 0.24 to ∞.

Figure 7a. illustrates the mean type-1 packet delay for $\lambda_1 = 0.7$ and $\lambda_2 = 0.2$, as function of Var[a_2]. We see that the effect of Var[a_2] on E[d_1] for the HOL-JIA schemes is only visible when Var[a_2] is low; this effect is a decrease of E[d_1] when Var[a_2] increases and is again due to a varying $A_2(0)$. For a low Var[a_2], the type-2 arrivals are nicely spread over time. This however means that type-2 packets arrive in a lot of slots, causing numerous jumps, and thus leading to a slightly higher mean type-1 packet delay when Var[a_2] is low. In the HOL-JOS scheme, E[d_1] is linearly increasing with Var[a_2] (see [6]). Hence, the variability of the mean number of type-2 arrivals has a much larger impact on E[d_1] for this scheme.

In Figure 7b., we have also depicted the mean type-2 packet delay. In most queueing systems, the mean packet delay is linearly dependent on the variance of the number of corresponding arrivals. Here, E[d_2] increases when Var[d_2] increases, for all schemes. For the HOL-JIA schemes, we furthermore see the added effect of $A_2(0)$ for low Var[a_2]. We also observe a big influence of Var[a_2] on E[d_2] for the HOL scheme and the HOL-JIA schemes, while the impact for the HOL-JOS scheme is smaller. It is however noticed that the curves of E[d_1] and E[d_2] for the latter scheme diverge for increasing Var[a_2]. HOL-JOS scheduling thus provides more differentiation in the delays when the traffic becomes burstier. This is a fortiori the case in the other schemes.



a. *for* $\lambda_T = 0.7$ b. *for* $\lambda_T = 0.9$

Figure 8: Mean value of packet delays versus β when $\alpha = 0.75$

4.3.5 Summary

In summary, the HOL-JIA schemes perform similarly with regard to the mean type-1 packet delay. As to the mean type-2 packet delay, both schemes also achieve a similar performance when α is low (i.e., when the overall traffic mix mainly consists of type-2 traffic). When α is high on the other hand, the HOL-JIA^2 scheme outperforms the HOL-JIA^1 scheme. The JIA^2 mechanism thus shows promising results with regard to the mean type-2 packet delay. Furthermore, the mean type-1 packet delays only depend on the arrival characteristics of type-2 traffic through a single parameter, namely the probability of no type-2 arrivals during a slot. Therefore, the influence of type-2 packets on the performance of type-1 traffic is kept small for the HOL-JIA schemes. The HOL-JOS scheme further achieves a limited delay differentiation, and is thus only practicable if there is little difference in the delay requirements of both types of traffic. This subsection finally shows that subtle differences between jumping schemes can yield large differences between their performance.

4.4 The (dis)advantages of a jumping parameter

In this final subsection, we briefly illustrate the (dis)advantages of introducing a jumping parameter. Figures 8a. and 8b. show the mean packet delays of both types of traffic when $\alpha = 0.75$, as functions of the jumping parameter β of the HOL-MBP scheme, for $\lambda_T = 0.7$ and $\lambda_T = 0.9$ respectively. We only depict the HOL-MBP scheme and the HOL-JIA^2 scheme, since the HOL-PJ scheme performs similarly as the first (see subsection 4.2), and since the HOL-JIA^2 scheme is the best performing one of the jumping schemes without a jumping parameter (see subsection 4.3). The

advantage of a jumping parameter is obvious from these figures: β can be chosen by an operator depending on the delay requirements of both types of traffic. A low β e.g., will highly favour the type-1 traffic, while chosing a higher β will give the type-1 traffic only a small reduction (compared to the FIFO scheme). The HOL-JIA^2 scheme does not have such a parameter, and the performance of this scheme is thus completely determined by the arrival process. A possible disadavantage of a jumping parameter however is that it is necessary to anticipate on varying arrival characteristics to still achieve the required performance. Indeed, when for example the total arrival rate increases, β has to be lowered to still meet the same delay requirements. The HOL-MBP scheme thus seems practically difficult to implement for a system with fast-varying incoming traffic.

5 Conclusions and future work

In this paper, we have first given an overview of jumping schemes in the literature. We have further introduced a new self-adaptive jumping scheme: the HOL-JIA^2 scheme. We have derived the probability generating functions of the system contents and the delay of type-1 packet. Moments can be easily calculated from these pgfs. We have furthermore provided a method to determine the mean delay of a type-2 packet, although its pgf seems hard to calculate. Then, we have extensively compared the performance of the various priority schemes with priority jumps. Special attention is given to the new HOL-JIA^2 scheme. Specifically, the self-adaptiveness of this scheme produces promising results. We also show that subtle differences between jumping schemes can yield considerable differences between their performance. Finally, we note that there does not exist something like "the ultimate jumping scheme". Indeed, depending on the applications, the required differentiation, the arrival characteristics (e.g., the arrival rates), and the variability of these characteristics, one can opt for either the HOL-JOS scheme (when there is little difference in the delay requirements of both types of traffic), the HOL-MBP scheme (when the characteristics do not change a lot in time), or the HOL-JIA^2 scheme (when the characteristics constantly vary).

Letting jumps depend on the arrival characteristics of the type-2 traffic introduces the notion of self-adaptiveness. In the future, we plan to further investigate this self-adaptiveness of jumping schemes. Instead of the jumping condition in this paper, we could incorporate other conditions. The HOL-packet of the low-priority queue could for example jump to the high-priority queue only when

a certain number of type-2 packets have arrived in a predetermined time period (of more than one slot). Jumps can also be conditioned on the arrival characteristics of type-1 traffic. The ultimate goal is to find a self-adaptive jumping scheme that performs well for every traffic scenario. The fact that the analysis in this paper is quite straightforward gives us the hope that this ideal jumping scheme is still analytically tractable. We have further shown in the paper that it is complicated to derive an explicit expression for the pgf of the delay of a type-2 packet for the HOL-JIA schemes, due to correlations between the involved quantities. Finding an exact solution seems extremely difficult, so we are looking for good approximate solutions. One possible approximation may for example be obtained by ignoring the correlations.

Further possible future work includes the study of the effect of the various jumping schemes on systems with more general mathematical models. We for example think of time correlation in the arrival process, of geometrically distributed transmission times, and of a general number of priorities. Note also that in the schemes studied so far only the HOL-packet of the low-priority queue can jump to the high-priority queue. It may be interesting to analyse a jumping scheme in which a random packet of the low-priority queue can jump to the high-priority queue (to incorporate 'impatient' packets). We note that most of the extensions will complicate the analysis considerably. However, this makes them in turn interesting research topics.

Acknowledgement

The authors would like to thank the anonymous referees and the editor for their constructive suggestions, which led 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 on Networking*, 2(5):508–519, 1994.

[2] 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.

[3] 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.

[4] 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.

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

[6] T. Maertens, J. Walraevens, and H. Bruneel. A modified HOL priority scheduling discipline: performance analysis. *European Journal of Operational Research*, 180(3):1168–1185, 2007.

[7] T. Maertens, J. Walraevens, M. Moeneclaey, and H. Bruneel. A new dynamic priority scheme: performance analysis. In *Proceedings of the 13th International Conference on Analytical and Stochastic Modelling Techniques and Applications (ASMTA2006)*, pages 74–84, 2006.

[8] 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.