# Performance analysis of a discrete-time queueing system with priority jumps

Lic. Tom Maertens*, Dr. ir. Joris Walraevens, Prof. dr. ir. Marc Moeneclaey,
Prof. dr. ir. Herwig Bruneel
Ghent University – UGent
Department of Telecommunications and Information Processing – IR07
Sint-Pietersnieuwstraat 41, B-9000 Gent, Belgium
Phone: +32-9-2648901
Fax: +32-9-2644295
E-mail: {tmaerten,jw,mm,hb}@telin.UGent.be

## 1 Introduction

Over the years, priority scheduling has become the basis for a large class of queueing disciplines that are designed to support different types of traffic in modern telecommunication systems. In the Head-Of-Line (HOL) priority scheduling discipline, *delay-sensitive* traffic is always given priority over *delay-tolerant* traffic. This HOL priority scheme provides low delays for the delay-sensitive traffic, but it can cause excessive delays for the delay-tolerant traffic, especially when the network is highly loaded (see e.g., [1, 5]).

One way to prevent this *starvation* of delay-tolerant traffic is the introduction of *priority jumps* (see e.g., [1]). In a priority scheme with priority jumps, the priority level of delay-tolerant packets may be increased in the course of time. In [3] for instance, the packet at the HOL-position of *the low-priority queue* jumps at the end of each time unit to *the high-priority queue* if during that time unit a packet of the high-priority queue is transmitted. The *flow* of delay-tolerant traffic into the high-priority queue may then however be too drastic in some cases, with a too negative effect for the delay-sensitive traffic as a result. Therefore, we add an extra jumping condition to this modified HOL (m-HOL) scheme of [3].

Concretely, we present a jumping scheme in which a possible jump in a time unit also depends on the number of delay-tolerant packets arriving in the system during that time unit: the Head-

*Corresponding author

Of-Line Jump-If-Arrival (HOL-JIA) priority scheme. Via a probability generating function (pgf) approach, we derive the pgfs of the system contents of the high- and low-priority queue, and the pgf of the delay of a delay-sensitive packet. From these pgfs, we can easily calculate expressions for some interesting performance measures, such as mean values and variances. A numerical example finally shows the impact of the jumping mechanism.

The contribution of this paper concerns the model that is considered, as well as the solution technique that we have used and the specific results that are efficiently determined by this technique. First, we show that by letting priority jumps depend on the arrival characteristics of the delay-tolerant traffic the flow of this type of traffic into the high-priority queue is restricted. The negative effect for the delay-sensitive traffic is then limited. The effect for the delay-tolerant traffic can however still be considerably positive. The jumping mechanism in the HOL-JIA priority scheme is thus better *adapted* to the amount of traffic that arrives in the system. Secondly, we demonstrate that an analysis based on probability generating functions is very suitable for analysing queueing systems with priority jumps. Specifically, some *boundary functions* need to be determined during the solution process. This is a well-known feature for coupled queues (see e.g., [2, 4] for similar cases). The pgf technique provides an efficient and fast method for the determination of these boundary functions.

The outline of the paper is as follows. Section 2 describes the mathematical model. In Sections 3 and 4, we derive the steady-state system contents and study the delays of both types of packets, respectively. Section 5 presents a small numerical example. Finally, we formulate some conclusions in Section 6.

## 2 Mathematical model

We consider a *discrete-time* queueing system with two queues of infinite capacity and one transmission channel. Two types of packets arrive at the system: packets of type 1, which enter the first queue, and packets of type 2, which enter the second queue. The numbers of both types of packets arriving in slot $k$ are denoted by $a_{1,k}$ and $a_{2,k}$ respectively. The $a_{1,k}$s and $a_{2,k}$s are assumed to be independent and identically distributed (i.i.d.) from slot-to-slot. Within one slot however, $a_{1,k}$ and $a_{2,k}$ can be correlated. This possible correlation is described by the joint probability

generating function (pgf) $A(z_1, z_2) \triangleq \lim_{k\to\infty} \mathrm{E}\left[z_1^{a_{1,k}} z_2^{a_{2,k}}\right]$. We furthermore define the marginal pgfs $A_T(z) \triangleq A(z, z)$, $A_1(z) \triangleq A(z, 1)$ and $A_2(z) \triangleq A(1, z)$ of the total number, the number of type-1 and the number of type-2 arrivals during a slot respectively. The corresponding arrival rates are defined as $\lambda_T \triangleq A_T'(1)$, $\lambda_1 \triangleq A_1'(1)$ and $\lambda_2 \triangleq A_2'(1)$, with $\lambda_T = \lambda_1 + \lambda_2$. Note that since there is only one transmission channel, the stability condition for this system is given by $\lambda_T < 1$.

The transmission times of all the packets are deterministically equal to one slot. Packets in the first queue have a higher priority than those in the second. So, whenever there are packets present in the high-priority queue, they have tranmission priority, and only when the high-priority queue is empty, packets of the low-priority queue are transmitted. Within both queues separately, packets are stored according to a First-In, First-Out (FIFO) rule.

Packets of the low-priority queue can jump to the high-priority queue according to 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 have arrived at the system, the *HOL-packet* of the low-priority queue jumps to the high-priority queue. Since the possible jump occurs at the end of the slot, the jumped type-2 packet enters the high-priority queue after the type-1 packets that arrived during the same slot.

## 3 System content

We derive an expression for the joint pgf of the system contents of both queues at the beginning of a random slot in the *steady state*. Under the assumption that the packet in transmission (if one) is part of the queue that is "served" in that slot, we define $u_{H,k}$ and $u_{L,k}$ as the system contents of the high- and low-priority queue at the beginning of slot $k$, and $u_{T,k}$ as the total system content at the beginning of slot $k$. The joint pgf of $u_{H,k}$ and $u_{L,k}$ is denoted by and defined as $U_k(z_1, z_2) \triangleq \mathrm{E}\left[z_1^{u_{H,k}} z_2^{u_{L,k}}\right]$. The system under consideration satisfies the following *system equations*:

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

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

- if $u_{H,k} > 0, u_{L,k} = 0$:

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

- if $u_{H,k} > 0, u_{L,k} > 0$:

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

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

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

$$\begin{cases} u_{H,k+1} = u_{H,k} + a_{1,k} \\ u_{L,k+1} = u_{L,k} - 1 + a_{2,k} \end{cases}, \tag{4}$$

where $[\cdots]^+$ denotes the maximum of the argument and zero. If one of the queues is empty at the beginning of slot $k$, a packet of the other queue (if non-empty) is transmitted during slot $k$ (see Eqs. (1) and (2)). If both queues are non-empty on the other hand, it depends on the number of type-2 arrivals in slot $k$ whether a packet of the low-priority queue jumps to the high-priority queue (Eqs. (3) and (4)). Introducing pgfs in the system equations, letting $k \to \infty$ to reach the steady state, and isolating $U(z_1, z_2)$, yields

$$U(z_1, z_2) = \frac{\left\{ \begin{array}{l} [z_2(z_1 - 1)A(z_1, z_2) + (z_2 - z_1)A(z_1, 0)]\, U(0,0) \\ + (z_2 - z_1)(A(z_1, z_2) - A(z_1, 0))\, U(z_1, 0) + (z_1 - z_2)A(z_1, 0)\, U(0, z_2) \end{array} \right\}}{z_1(z_2 - A(z_1, z_2) + A(z_1, 0)) - z_2 A(z_1, 0)}. \tag{5}$$

In the right-hand side of (5), there are three quantities yet to be determined: the constant $U(0,0)$ and the boundary functions $U(z_1, 0)$ and $U(0, z_2)$. First, $U(0,0)$ is calculated via the normalization condition $U(1,1) = 1$. We obtain the probability of having an empty system: $U(0,0) = 1 - \lambda_T$. Secondly, the boundary functions $U(z_1, 0)$ and $U(0, z_2)$ can be found from (5), by applying Rouché's theorem and exploiting the analyticity of $U(z_1, z_2)$ inside the unit circle ($|z_1| < 1, |z_2| < 1$).

For $U(z_1,0)$, we first take the limit of (5) for $z_2 \to 0$:

$$U(z_1,0) = A(z_1,0)\frac{(z_1-1)(1-\lambda_T) + z_1 U^{(2)}(0,0)}{z_1 - A(z_1,0)}, \tag{6}$$

where $U^{(2)}(0,0) \triangleq \left.\dfrac{\partial U(z_1,z_2)}{\partial z_2}\right|_{z_1=z_2=0}$ denotes the probability of having an empty high-priority queue and one packet in the low-priority queue. Applying Rouché's theorem on the numerator $z_1 - A(z_1,0)$ of the right-hand side of (6) further implies a unique solution $s \triangleq A(s,0)$ in the unit circle ($|s| < 1$). The analyticity of $U(z_1,0)$ inside the unit circle ($|z_1| < 1$) then leads to an expression for $U^{(2)}(0,0)$ and by (6) for $U(z_1,0)$ itself:

$$U(z_1,0) = \frac{1-\lambda_T}{s}\frac{(z_1-s)A(z_1,0)}{z_1 - A(z_1,0)}. \tag{7}$$

Furthermore, by applying Rouché's theorem on the numerator of (5), we can show that for a given $z_2$ ($|z_2| < 1$), the equation $z_1(z_2 - A(z_1,z_2) + A(z_1,0)) - z_2 A(z_1,0) = 0$ has a unique solution $z_1 = Y(z_2)$ in the unit circle ($|z_1| < 1$):

$$Y(z_2) \triangleq \frac{z_2 A(Y(z_2),0)}{z_2 - A(Y(z_2),z_2) + A(Y(z_2),0)}. \tag{8}$$

$U(0,z_2)$ can again be calculated by using the analyticity of $U(z_1,z_2)$ inside the unit circle. This produces

$$U(0,z_2) = \frac{(1-\lambda_T)\left\{\begin{array}{l}(z_2 - Y(z_2))A(Y(z_2),0)\left[s + A(Y(z_2),z_2) - A(Y(z_2),0)\right] \\ -sz_2(1-Y(z_2))A(Y(z_2),z_2) - s(z_2-1)A(Y(z_2),z_2)A(Y(z_2),0)\end{array}\right\}}{s(z_2 - A(Y(z_2),z_2))(Y(z_2) - A(Y(z_2),0))}, \tag{9}$$

where we have used Eqs. (7) and (8).

We now have derived all unknown quantities in (5), and have thus obtained an expression for the joint pgf $U(z_1,z_2)$. By further substituting $z_1$ and $z_2$ by the appropriate values, we can determine the marginal pgfs $U_T(z) \triangleq U(z,z)$, $U_H(z) \triangleq U(z,1)$ and $U_L(z) \triangleq U(1,z)$ of the total system content, and of the high- and low-priority system contents respectively. It should be noted that the expression for $U_T(z)$ is identical to the pgf of the system content of a queue with a FIFO-discipline and with one type of arrivals, determined by $A_T(z)$. This is logic, because the total system content

is independent of the order the packets are being served in. Furthermore, from the marginal pgfs, expressions for the moments can be easily calculated by invoking the moment generating property of pgfs. For further use, we here give the expression of the *mean* total system content $\mathrm{E}[u_T]$:

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

with $\lambda_T = A_T'(1)$ the total arrival rate and $\lambda_{TT} \triangleq A_T''(1)$.

# 4 Packet delay

Let us first consider the delay of a type-1 packet. We therefore tag a type-1 packet and denote its arrival slot by slot $I$. Since jumps occur at the end of slots, the tagged type-1 packet is queued in front of a type-2 packet that possible jumps at the end of slot $I$. As a consequence, 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, only depends on the system content of the high-priority queue at the beginning of slot $I$ ($u_{H,I}$). Due to the i.i.d. arrivals from slot-to-slot, $u_{H,I}$ in slot $I$ and $u_H$ in an arbitrary slot have the same distribution due to the PASTA property. The pgf $D_1(z)$ of the delay of a random type-1 packet can then be expressed in terms of $U_H(z)$ (see e.g., [5] for a similar procedure) and is thus easily found. By further taking the first derivative of $D_1(z)$ for $z = 1$, we get an expression for the mean delay $\mathrm{E}\left[d_1\right]$ of a random type-1 packet. It is given by

$$\begin{aligned}\mathrm{E}\left[d_1\right] = &\frac{A_2(0)(1-\lambda_T) + s\lambda_T - sA_2(0)^2}{sA_2(0)(1-A_2(0))} + \frac{\lambda_{11}A_2(0)}{2\lambda_1(A_2(0)-\lambda_1)} - \frac{A^{(1)}(1,0)(1-\lambda_2)}{A_2(0)(A_2(0)-\lambda_1)} \\ &+ \frac{\lambda_1^2 + \lambda_1 - A_2(0)}{A_2(0)(A_2(0)-\lambda_1)},\end{aligned} \tag{11}$$

with $\lambda_{11} \triangleq A_1''(1)$ and $A^{(1)}(1,0) \triangleq \left.\dfrac{\partial A(z_1,z_2)}{\partial z_1}\right|_{z_1=1,z_2=0}$. By taking higher order derivatives for $z = 1$, expressions of higher moments can also be obtained. Note also that for $\lambda_T \to 1$, i.e., for the total system going to its stability boundary, $\mathrm{E}[d_1]$ remains finite.

Determining an expression for the pgf $D_2(z)$ of the delay of a random type-2 packet is rather complex, and still an open issue at the moment. It is however possible to calculate the mean delay $\mathrm{E}\left[d_2\right]$ of a random type-2 packet. Indeed, according to Little's law, we first have that $\mathrm{E}\left[u_T\right] = \lambda_T \mathrm{E}\left[d\right]$, with $\mathrm{E}\left[u_T\right]$ being the mean total system content and $\mathrm{E}\left[d\right]$ the mean delay of an arbitrary

packet. The probability that a randomly arriving packet is of type 1 (type 2) equals $\frac{\lambda_1}{\lambda_T}$ ($\frac{\lambda_2}{\lambda_T}$), and we thus further have that $\mathrm{E}[d] = \frac{\lambda_1}{\lambda_T}\mathrm{E}[d_1] + \frac{\lambda_2}{\lambda_T}\mathrm{E}[d_2]$. Combining these two expressions, we find

$$\mathrm{E}[d_2] = \frac{\mathrm{E}[u_T] - \lambda_1 \mathrm{E}[d_1]}{\lambda_2}. \tag{12}$$

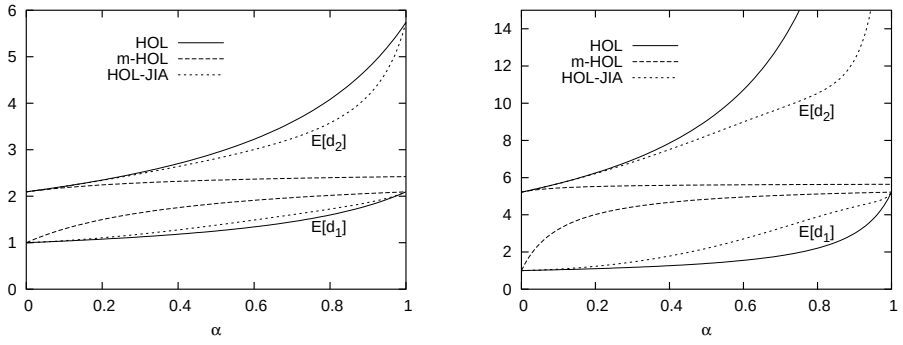
$\mathrm{E}[u_T]$ and $\mathrm{E}[d_1]$ have been calculated (in Eqs. (10) and (11) respectively), and we are thus able to derive an expression for the mean delay of a random type-2 packet.

# 5 Numerical example

In the previous section, we have briefly described the procedures to obtain expressions for the mean packet delays of both types of traffic. In this section, we present a small numerical example to illustrate the impact and the significance of the HOL-JIA priority scheme for varying traffic mixes. We therefore compare the mean delays with the mean delays in a HOL priority queue (see [5]) and in a queue with the m-HOL priority scheme (see [3]). We consider an arrival process with

$$A(z_1, z_2) = \left(1 - \frac{\lambda_1}{16}(1 - z_1) - \frac{\lambda_2}{16}(1 - z_2)\right)^{16}. \tag{13}$$

Here, $\lambda_1$ and $\lambda_2$ are the arrival rates of type-1 (delay-sensitive) and type-2 (delay-tolerant) traffic respectively. This is the arrival process to a queue in an 16x16 output-queueing switch with Bernoulli arrivals at its inlets, and with uniform routing. We furthermore define $\alpha$ as the fraction of type-1 traffic in the overall traffic mix, i.e., $\alpha = \lambda_1/\lambda_T$, with $\lambda_T = \lambda_1 + \lambda_2$.




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

Figure 1: Mean value of packet delays versus $\alpha$

Figures 1a. and 1b. show the mean packet delays of both types of traffic for $\lambda_T = 0.7$ and $\lambda_T =$

0.9 respectively, as functions of $\alpha$, for the HOL, the m-HOL and the HOL-JIA priority schemes. It is first noticed that the curves of $\mathrm{E}[d_1]$ and $\mathrm{E}[d_2]$ for the HOL-JIA scheme lie between those for the HOL scheme and the m-HOL scheme, as expected. Secondly, we observe little performance difference between the HOL scheme and the HOL-JIA scheme when $\alpha$ is low, i.e., when few type-1 packets arrive at the system. When $\alpha$ is low in the HOL-JIA scheme, the high-priority queue is often empty and few type-2 packets jump to the high-priority queue. Both types of traffic thus behave similarly as in the HOL scheme, and the effect of the jumping mechanism is thus limited in the HOL-JIA scheme. In the m-HOL scheme on the other hand, a type-2 packet jumps to the high-priority queue *every* slot where both queues are non-empty. As a result, a higher $\mathrm{E}[d_1]$ and lower $\mathrm{E}[d_2]$ than for the HOL(-JIA) scheme is observed. As already mentioned in the introduction, the effect of the m-HOL scheme may then be too drastic in some cases, and the HOL-JIA scheme mimicing the HOL scheme is satisfactory for low $\alpha$.

When $\alpha$ is high, the m-HOL scheme achieves a small delay differentation. The HOL-JIA scheme on the other hand, can considerably influence $\mathrm{E}[d_2]$ compared to the HOL scheme, while the negative effect on $\mathrm{E}[d_1]$ stays limited (see Figure 1b.). E.g., when $\lambda_T = 0.9$ and $\alpha = 0.8$, $\mathrm{E}[d_2]$ decreases from about 17.3 for HOL to about 10.5 for HOL-JIA, with only a small increase for $\mathrm{E}[d_1]$ (from about 2.21 to about 3.82). We also see that for $\alpha \to 1$, $\mathrm{E}[d_2]$ for the HOL-JIA and HOL schemes converge (see Figure 1a.). When $\alpha \approx 1$ in the HOL-JIA scheme, an exceptionally arriving type-2 packet has a large probability of entering an empty low-priority queue. To be transmitted however, this type-2 packet has to wait in the low-priority queue until the high-priority queue becomes empty (since it is only allowed to jump if another rare type-2 packet arrives). Hence, we find a similar behaviour as in the HOL scheme.

# 6 Conclusions

We first conclude that the HOL-JIA scheme does what it is designed for: the delay-tolerant traffic can be saved from starvation compared to the HOL scheme, while the delay-sensitive traffic is only mildly affected, as opposed to the m-HOL scheme. Secondly, an analysis based on probability generating functions efficiently overcomes mathematical challenges and is moreover useful for the calculation of important performance measures. The pgf of the low-priority delay is however still

an open issue.

## Acknowledgement

The second author is a Postdoctoral Fellow with the Fund for Scientific Research, Flanders (F.W.O.-Vlaanderen), Belgium.

## Author biographies

**TOM MAERTENS** was born in Bruges, Belgium, in 1981. He received the degree of Licentiate in Computer Science from Ghent University, Belgium, in 2003. In September 2003, he joined the SMACS Research Group, Department of Telecommunications and Information Processing, at the same university. His main personal research interests include discrete-time queueing models and performance analysis of communication networks.

**JORIS WALRAEVENS** was born in Zottegem, Belgium, in 1974. He received the M.S. degree in Electrical Engineering and the Ph.D. degree in Engineering in 1997 and 2004 respectively, all from Ghent University, Belgium. In September 1997, he joined the SMACS Research Group, Department for Telecommunications and Information Processing, at the same university. His main research interests include discrete-time queueing models and performance analysis of communication networks. His personal webpage can be found at http://telin.UGent.be/~jw.

**MARC MOENECLAEY** received the diploma of Electrical Engineering and the Ph.D. degree in electrical engineering from the University of Gent, Gent, Belgium, in 1978 and 1983, respectively. He is presently a Professor in the Department of Telecommunications and Information Processing (TELIN), Ghent University. His research interests include statistical communication theory, carrier and symbol synchronization,

bandwidth-efficient modulation and coding, spread-spectrum, and satellite and mobile communication. He is the author of about 300 scientific papers in international journals and conference proceedings. He coauthors the book *Digital Communication Receivers—Synchronization, Channel Estimation, and Signal Processing* (J. Wiley, New York, 1998). He has been active in various international conferences as a Technical Program Committee Member and Session Chairman.

**HERWIG BRUNEEL** was born in Zottegem, Belgium, in 1954. He received the M.S. degree in Electrical Engineering, the degree of Licentiate in Computer Science, and the Ph.D. degree in Computer Science in 1978, 1979 and 1984 respectively, all from Ghent University, Belgium. He is full Professor in the Faculty of Engineering and head of the Department of Telecommunications and Information Processing at the same university. He also leads the SMACS Research Group within this department. His main personal research interests include stochastic modeling and analysis of communication systems, discrete-time queueing theory, and the study of ARQ protocols. He has published more than 250 papers on these subjects and is coauthor of the book *H. Bruneel and B. G. Kim, "Discrete-Time Models for Communication Systems Including ATM"* (Kluwer Academic Publishers, Boston, 1993). From October 2001 to September 2003, he has served as the Academic Director for Research Affairs at Ghent University.

## 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.H.S. Van Leeuwaarden and J.A.C. Resing. A tandem queue with coupled processors: computational issues. *Queueing Systems*, 51(1–2):29–52, 2005.

[3] 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.

[4] M. Sidi and A. Segall. An acknowledgement-based access scheme in a two-node packet-radio network. *IEEE Transactions on Communications*, 32(6):741–744, 1984.

[5] 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.