NUMERICAL ALGEBRA,
CONTROL AND OPTIMIZATION
Volume **1**, Number **4**, December **2011**

doi:10.3934/naco.2011.1.657

pp. **657–673**

CONTROLLING DELAY DIFFERENTIATION WITH PRIORITY JUMPS: ANALYTICAL STUDY

Tom Maertens, Joris Walraevens and Herwig Bruneel

SMACS Research Group
Department of Telecommunications and Information Processing (TELIN)
Ghent University (UGent)
Sint-Pietersnieuwstraat 41, B-9000 Gent, Belgium

Abstract. Supporting different services with different Quality of Service (QoS) requirements is not an easy task in modern telecommunication systems: an efficient priority scheduling discipline is of great importance. Fixed or static priority achieves maximal delay differentiation between different types of traffic, but may have a too severe impact on the performance of lower-priority traffic. In this paper, we propose a priority scheduling discipline with priority jumps to control the delay differentiation. In this scheduling discipline, packets can be promoted to a higher priority level in the course of time. We use probability generating functions to study the queueing system analytically. Some interesting mathematical challenges thereby arise. With some numerical examples, we finally show the impact of the priority jumps and of the system parameters.

1. **Introduction.** Modern integrated telecommunication systems are designed to offer a wide variety of services, such as telephony, data transfer, and video conferencing. Different services, however, have extremely diverse *Quality-of-Service* (QoS) requirements. Real-time services, like video conferencing or internet telephony, do not tolerate *delay* but can sustain some *loss*, while non-real-time services, like sending data files, allow for some delay but are quite vulnerable to loss. In this paper, we focus on delay as QoS measure. Regarding their different delay requirements, we categorize real-time traffic as *delay-sensitive* and non-real-time traffic as *delay-tolerant* in the remainder.

To support different types of traffic in modern telecommunication systems, many *priority* scheduling disciplines have been proposed over the years. Priority scheduling can be implemented in two ways, i.e., on queue level or on packet level. In the first case, different types of packets are stored in different queues, and each queue is provided a different priority level. Then either the highest non-empty priority queue is always chosen to be served next (see e.g., [9, 12, 13]), the queues are served in a weighted, fixed order (see e.g., [5, 11]), or the order of service is determined by the contents of the queues (see e.g., [6, 10]). In the second case, all packets are stored in one queue, but each packet is assigned a different priority level. Then priority is always given to the packet with the highest priority level. Individual priority levels

2000 *Mathematics Subject Classification.* Primary: 68M20, 60K25; Secondary: 90B22, 97I80.
Key words and phrases. Queueing theory, performance evaluation, priority scheduling.
The second author is a Postdoctoral Fellow with the Fund for Scientific Research, Flanders (F.W.O.-Vlaanderen), Belgium.

can, for example, be based on the deadlines of packets (see e.g., [4]), on the transmission times of packets (see e.g., [2]), or on a relative weight and the content of the queue (see e.g., [16]). However, this second category of implementations may involve heavy processing as in each time period the priority level of each packet has to be checked and possibly updated. That is the reason why in practice usually (combinations of) priority scheduling disciplines of the first category are used.

In the fixed or static priority scheduling discipline, delay-sensitive packets always have priority over delay-tolerant packets, i.e., delay-tolerant packets can only be transmitted when there are no delay-sensitive packets in the system. Static priority provides low delays for the delay-sensitive traffic, but the performance of the delay-tolerant traffic can be degraded severely. In particular, when the network is highly loaded and a large portion of the network traffic consists of delay-sensitive packets, static priority scheduling may cause excessive delays for the delay-tolerant traffic. Although the delay-tolerant traffic tolerates a certain amount of delay, extreme values obviously 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 being lost, and would consequently decrease its transmission rate. This decreases the throughput, which is detrimental to data transfer. The decrease of the transmission rate, however, is unnecessary, since the delay-tolerant packet is not lost.

In this paper, we consider a queueing system in which delay-tolerant packets can promote or *jump* to the high-priority level in the course of time. Jumped delay-tolerant packets are treated as if they are delay-sensitive, i.e., they have transmission priority over newly arriving delay-sensitive packets. From the transmission channel's point of view, nothing changes in comparison with static priority: the packet at the head of the highest non-empty priority queue is chosen for transmission. Scheduling disciplines with so-called *priority jumps* thus build upon the simplicity and efficiency of static priority, but they prevent delay-tolerant packets from starving out in the system. Scheduling disciplines with priority jumps are the subject of many papers in the recent literature. A nice overview of these priority scheduling disciplines and an in-depth performance comparison between them can be found in [9].

Here, we opt for a model that we can study analytically. Specifically, we introduce a parameter β which is defined as the probability that the full content of the low-priority queue jumps (or, is swapped) to the high-priority queue. This model is based on an earlier model (see [7]), but it eliminates a major limitation of that model as we will describe in the next section. By using probability generating functions (pgfs), we derive the joint distribution of the queue contents and the distributions of the packet delays of both types of traffic. These distributions can lead to some interesting performance measures (such as mean values), which are used to determine the impact of priority jumps and to show the influence of various system parameters.

The contribution of the paper concerns the scheduling discipline that is studied, as well as the solution technique that we have used and the specific analytical results that we find with this technique. First, introducing a jumping parameter has the benefit that the delay differentiation between the different types of traffic can be controlled and adjusted if necessary. The value of β namely can be chosen in such a way that the delay-tolerant traffic stays within its delay requirements: for example, the more stringent the delay requirement, the larger the value of β. Secondly, this paper demonstrates that a queueing analysis based on pgfs is very suitable for studying this type of queueing systems. In particular, some *boundary* functions

need to be determined during the solution process, and it is generally known that this can be a very hard task in priority systems (see e.g., [12, 15]). The pgf technique provides an efficient and fast method for the determination of these functions.

The outline of the paper is as follows. In Section 2, we describe the mathematical model. In Sections 3 and 4, we derive the joint pgf of the system content and study the delays of both types of packets, respectively. Numerical examples are presented in Section 5. Finally, we formulate some conclusions in Section 6.

2. Mathematical model.

We consider a *discrete-time* (i.e., time is assumed to be *slotted*) queueing system with *two queues* of *infinite capacity* and *one transmission channel*. Two types of packets arrive at the system: type-1 packets, representing delay-sensitive traffic, and type-2 packets, which are delay-tolerant. The numbers of arrivals of both types of packets during slot k are denoted by $a_{1,k}$ and $a_{2,k}$, respectively. We assume that the $a_{1,k}$s and $a_{2,k}$s are 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. The joint pgf of $a_{1,k}$ and $a_{2,k}$ is defined as $A(z_1, z_2)$, i.e.,

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

Then the marginal pgfs of the numbers of type-1 and type-2 arrivals per slot are given by $A_1(z) \triangleq A(z, 1)$ and $A_2(z) \triangleq A(1, z)$, respectively. Furthermore, we denote the total number of arrivals during slot k by $a_{T,k}$; so $a_{T,k} \triangleq a_{1,k} + a_{2,k}$. Its pgf is given by $A_T(z) \triangleq A(z, z)$. The corresponding arrival rates, i.e., the mean number of arrivals per slot, are indicated by $\lambda_j \triangleq A_j'(1)$ $(j = 1, 2)$ and $\lambda_T \triangleq A_T'(1)$ $(= \lambda_1 + \lambda_2)$. The transmission times of all the packets are deterministically equal to one slot. Throughout the paper, we assume that $\lambda_T < 1$, so that the system reaches a steady state.

Following their delay requirements, arriving type-1 packets enter the *high-priority* queue, while arriving type-2 packets are originally stored in the *low-priority* queue. The packets of the two queues are transmitted according to a Head-Of-Line (HOL) priority rule: when there are packets present in the high-priority queue at the beginning of a slot, the HOL-packet of this queue is transmitted. Only when the high-priority queue is empty at the beginning of a slot, the HOL-packet of the low-priority queue can be transmitted. Note that within both queues, packets are transmitted according to a First-In-First-Out order.

In the course of time, however, packets of the low-priority queue may jump to the high-priority queue. In [7], we have studied a new priority scheduling discipline with priority jumps. At the end of each slot, specifically, the whole content of the low-priority queue jumps to the high-priority queue with probability β. Or in other words, the contents of both queues may be merged at the end of each slot. That is why we call this the Merge-By-Probability (MBP) mechanism. Numerical examples in [9] show that when the system is highly loaded and a large portion of the system traffic consists of type-1 packets, this mechanism avoids excessive delays for type-2 packets. However, these numerical examples also illustrate that when the traffic mix mainly consists of type-2 traffic, the few type-1 packets may suffer from unnecessary delays. Indeed, since the contents of both queues can be merged even when the high-priority queue is empty, a rare type-1 arrival possibly finds earlier jumped type-2 packets in front of it at entrance.

To avoid this detrimental situation for type-1 packets, we propose a variant of the MBP mechanism in this paper: a merge is only possible when the high-priority queue is non-empty at the beginning of a slot. Packets thus stay in the low-priority queue as long as the high-priority queue is empty. This variant is called the MBP* mechanism. Special attention has to be made to the case in which there are no packets present in the low-priority queue at the beginning of the slot. Two alternative models are being considered here: when type-2 packets arrive in an empty low-priority queue, either they jump with probability β (model A) or they cannot jump (model B). As we will see in the next section, both models are much harder to analyse than the original model of [7].

3. **System content.** In this section, we derive expressions for the joint pgfs of the *system content* at the beginning of a random slot in the *steady state*, for the two models. In the assumption that the packet in transmission (if one) is part of the queue that is "served" in that slot, we denote the contents of the high- and low-priority queue at the beginning of slot k as $u_{H,k}$ and $u_{L,k}$, respectively. Then the system content at the beginning of slot k can be described by the pair $(u_{H,k}, u_{L,k})$. It is easy to see that $(u_{H,k}, u_{L,k})$ is a suitable Markovian description of the system state at the beginning of slot k. First, we express the evolution of the system state from slot to slot (i.e., the so-called *system equations*), thereby clearly indicating the differences between both models. Secondly, these equations are transformed into pgfs, with all its consequences.

3.1. Establishing the system equations.

3.1.1. $u_{H,k} = 0$. 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. The arriving packets are queued according to their priority level and no packets jump from the low-priority queue to the high-priority queue. So we find that

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

where $[\cdots]^+$ denotes the maximum of the argument and zero.

3.1.2. $u_{H,k} > 0$ *and* $u_{L,k} > 0$. On the other hand, when the high-priority queue is non-empty at the beginning of slot k, a packet of the high-priority queue is transmitted during slot k. When, at the beginning of slot k, the low-priority queue is non-empty as well, the whole content of this queue is swapped to the high-priority queue with probability β. This possible swap takes place at the end of slot k, so type-2 packets that have arrived during slot k jump along. This yields the following equations:

- with probability β:

$$\begin{cases} u_{H,k+1} = u_{H,k} - 1 + a_{1,k} + u_{L,k} + a_{2,k} \\ u_{L,k+1} = 0 \end{cases}, \tag{3}$$

- with probability $1 - \beta$:

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

3.1.3. $u_{H,k} > 0$ *and* $u_{L,k} = 0$. As mentioned in the previous section, special attention has to be made to the case in which the high-priority queue is not empty at the beginning of a slot while the low-priority queue is. Two alternatives are being considerd here: the type-2 packets that arrive during the slot also jump with probability β (model A) or they do not jump at all (model B). For model A, we get

- with probability β:

$$\begin{cases} u_{H,k+1} = u_{H,k} - 1 + a_{1,k} + a_{2,k} \\ u_{L,k+1} = 0 \end{cases}, \tag{5}$$

- with probability $1 - \beta$:

$$\begin{cases} u_{H,k+1} = u_{H,k} - 1 + a_{1,k} \\ u_{L,k+1} = a_{2,k} \end{cases}. \tag{6}$$

For model B, this amounts to just (6), with probability 1. Note that the difference between both models is the queue (high- or low-priority) in which arriving type-2 packets may be stored at the end of slot k if they enter an empty low-priority queue.

3.2. **Determining the functional equation.** Next, we introduce pgfs in the system equations. This yields a relationship between $U_{k+1}(z_1, z_2)$ and $U_k(z_1, z_2)$, with

$$U_k(z_1, z_2) \triangleq \mathrm{E}\left[z_1^{u_{H,k}} z_2^{u_{L,k}}\right]. \tag{7}$$

For model A, we find that

$$\begin{aligned} U_{k+1}(z_1, z_2) = {} & A(z_1, z_2)\frac{(z_2 - 1)U_k(0, 0) + U_k(0, z_2)}{z_2} \\ & + \beta A_T(z_1)\frac{U_k(z_1, z_1) - U_k(0, z_1)}{z_1} \\ & + (1 - \beta)A(z_1, z_2)\frac{U_k(z_1, z_2) - U_k(0, z_2)}{z_1}. \end{aligned} \tag{8}$$

This can be arranged as

$$\begin{aligned} U_{k+1}(z_1, z_2) = {} & A(z_1, z_2)\frac{z_1(z_2 - 1)U_k(0, 0) + (1 - \beta)z_2 U_k(z_1, z_2)}{z_1 z_2} \\ & + A(z_1, z_2)\frac{(z_1 - (1 - \beta)z_2)U_k(0, z_2)}{z_1 z_2} \\ & + \beta A_T(z_1)\frac{U_k(z_1, z_1) - U_k(0, z_1)}{z_1}. \end{aligned} \tag{9}$$

Letting $k \to \infty$ to reach the steady state and isolating $U(z_1, z_2)$ afterwards produces the functional equation for the joint pgf of the queue contents:

$$\begin{aligned} U(z_1, z_2) = {} & \frac{z_1(z_2 - 1)A(z_1, z_2)U(0, 0) + (z_1 - (1 - \beta)z_2)A(z_1, z_2)U(0, z_2)}{z_2(z_1 - (1 - \beta)A(z_1, z_2))} \\ & + \frac{\beta z_2 A_T(z_1)[U(z_1, z_1) - U(0, z_1)]}{z_2(z_1 - (1 - \beta)A(z_1, z_2))}. \end{aligned} \tag{10}$$

In a similar way, we obtain the functional equation for model B:

$$U(z_1,z_2) = \frac{(z_1z_2 - z_1 - \beta z_2)A(z_1,z_2)U(0,0) + \beta z_2 A(z_1,z_2)U(z_1,0)}{z_2(z_1-(1-\beta)A(z_1,z_2))} + \frac{(z_1-(1-\beta)z_2)A(z_1,z_2)U(0,z_2)}{z_2(z_1-(1-\beta)A(z_1,z_2))} + \frac{\beta z_2 A_T(z_1)[U(z_1,z_1)-U(z_1,0)-U(0,z_1)+U(0,0)]}{z_2(z_1-(1-\beta)A(z_1,z_2))}. \tag{11}$$

In (10), there are three quantities yet to be determined, namely the constant $U(0,0)$ and the functions $U(0,z)$ and $U(z,z)$; in (11), also the *boundary* function $U(z,0)$ appears. This can be explained as follows. In model A, a possible swap does not depend on the content of the low-priority queue at the beginning of a slot. Indeed, whether the low-priority is empty or not, a swap occurs with probability β. In model B, on the contrary, there have to be packets in the low-priority queue at the beginning of a slot to make a swap possible. This condition on the content of the low-priority queue yields an additional boundary function and thus leads to a more complex solution process.

3.3. **Calculating the function $U(z,z)$ and the constant $U(0,0)$.** Let us first calculate the function $U(z,z)$ and the constant $U(0,0)$. The function $U(z,z)$ is the pgf of the total system content. The transmission times of all packets are equal to one slot and the system is assumed to be *work-conserving*, so the total system content is independent of the chosen scheduling model. By replacing z_1 and z_2 by z in (10) as well as in (11), we find that

$$U(z,z) = U(0,0)\frac{A_T(z)(z-1)}{z-A_T(z)}. \tag{12}$$

Then the constant $U(0,0)$ can be derived by applying the normalisation condition $U(1,1)=1$. By using l'Hôpital's rule, we obtain the probability of having an empty system:

$$U(0,0) = 1-\lambda_T. \tag{13}$$

3.4. **Calculating the boundary functions $U(0,z)$ and/or $U(z,0)$.** Furthermore, we compute the boundary functions $U(0,z)$ and/or $U(z,0)$. This is always the hardest task in this type of two-dimensional queueing systems (see e.g., [12, 15]). We start with model A. With Rouché's theorem, it can be shown that for a given value of z_2 inside the unit circle ($|z_2|<1$), the equation $z_1-(1-\beta)A(z_1,z_2)=0$ has one solution inside the unit circle for z_1 ($|z_1|<1$). This solution is denoted by $Y(z_2)$, with

$$Y(z) \triangleq (1-\beta)A(Y(z),z). \tag{14}$$

Since $Y(z_2)$ is a zero of the denominator in (10), and since $U(z_1,z_2)$ - as a pgf - remains finite inside the unit circle, $Y(z_2)$ must also be a zero of the corresponding numerator. This results, by exploiting the definition of $Y(z)$, in

$$U(0,z) = \frac{(1-\lambda_T)Y(z)A(Y(z),z)(z-1)}{Y(z)(z-A(Y(z),z))} + \frac{\beta z A_T(Y(z))[U(Y(z),Y(z))-U(0,Y(z))]}{Y(z)(z-A(Y(z),z))}. \tag{15}$$

After the use of Expr. (12) to calculate $U(Y(z), Y(z))$, we can arrange this as

$$U(0,z) = a(z) + b(z)U(0, Y(z)), \tag{16}$$

with

$$a(z) = \frac{\beta(1-\lambda_T)zA_T(Y(z))^2(Y(z)-1)}{Y(z)(Y(z)-A_T(Y(z)))(z-A(Y(z),z))} + \frac{(1-\lambda_T)(z-1)A(Y(z),z)}{z-A(Y(z),z)}, \tag{17}$$

$$b(z) = \frac{\beta z A_T(Y(z))}{Y(z)(A(Y(z),z)-z)}. \tag{18}$$

Eq. (16) describes a relation between $U(0,z)$ and $U(0,Y(z))$. We show how this relation can be used in an iterative procedure to compute $U(0,z)$ for any z inside the unit circle. Therefore, we recursively define $Y_i(z)$:

$$Y_i(z) \triangleq Y(Y_{i-1}(z)), \tag{19}$$

with $i \geq 1$ and $Y_0(z) = z$. Based on [1], it can be shown that for $i \to \infty$ and $|z| < 1$, $Y_i(z) \to C$, where $C \triangleq (1-\beta)A_T(C)$. Then, by successively applying Eq. (16), we obtain that

$$\begin{aligned} U(0,z) &= a(z) + b(z)U(0,Y_1(z)) \\ &= a(z) + b(z)a(Y_1(z)) + b(z)b(Y_1(z))U(0,Y_2(z)) \\ &= a(z) + b(z)a(Y_1(z)) + b(z)b(Y_1(z))a(Y_2(z)) \\ &\quad + b(z)b(Y_1(z))b(Y_2(z))U(0,Y_3(z)) \\ &= \cdots \\ &= \sum_{k=0}^{\infty} a(Y_k(z)) \prod_{l=0}^{k-1} b(Y_l(z)) + U(0,C) \prod_{l=0}^{\infty} b(Y_l(z)). \end{aligned} \tag{20}$$

It remains for us to calculate the constant $U(0,C)$. Again with Rouché's theorem, we can prove that the equation $z - A(Y(z),z) = 0$ has one solution inside the unit circle ($|z| < 1$). This solution is denoted by F_1, with $F_1 \triangleq A(Y(F_1),F_1)$, and is a zero of the denominator in (15). Since $|F_1| < 1$ and $U(0,z)$ must be bounded inside the unit circle, the corresponding numerator must also vanish for $z = F_1$. This yields

$$\begin{aligned} U(0,Y(F_1)) = &\frac{(1-\lambda_T)Y(F_1)A(Y(F_1),F_1)(F_1-1)}{\beta F_1 A_T(Y(F_1))} \\ &+ \frac{(1-\lambda_T)A_T(Y(F_1))(Y(F_1)-1)}{Y(F_1)-A_T(Y(F_1))}. \end{aligned} \tag{21}$$

The values of F_1 and $Y(F_1)$, which can be calculated numerically (e.g., via the Newton-Raphson method), lead to a value for $U(0,Y(F_1))$. The replacement of z by $Y(F_1)$ in (20), however, also gives us $U(0,Y(F_1))$ as a function of $U(0,C)$. This equation thus can be used to determine a value for $U(0,C)$. Note that the calculation of $U(0,z)$ here is much more complex than in [7]. In [7], specifically, exploiting the bounded character of $U(z_1,z_2)$ inside the unit circle was sufficient to find an expression for $U(0,z)$ (i.e., we did not need any iterative procedure).

Let us now proceed with model B. As already mentioned, here we need to calculate two boundary functions. By taking the limit of (11) for $z_1 \to z$ and $z_2 \to 0$,

and solving the result for $U(z,0)$, we first find that

$$U(z,0) = \frac{A(z,0)\left[(z-1)U(0,0) + zU^{(2)}(0,0)\right]}{z - A(z,0) + \beta A_T(z)} + \frac{\beta A_T(z)\left[U(z,z) - U(0,z) - U(0,0)\right]}{z - A(z,0) + \beta A_T(z)}, \tag{22}$$

where $U^{(2)}(0,0)$ is defined as $\left.\dfrac{\partial U(z_1,z_2)}{\partial z_2}\right|_{z_1=z_2=0}$ and denotes the probability of having an empty high-priority queue and one packet in the low-priority queue at the beginning of a random slot. It is easy to see that the function $U(z,0)$ is expressed in terms of the other three initial unknowns. However, a new unknown quantity arises, namely the constant $U^{(2)}(0,0)$. We determine this unknown later on. First, we follow the same procedure as we did for model A. $Y(z_2)$ is, as a solution of $z_1-(1-\beta)A(z_1,z_2)=0$ inside the unit circle, a zero of the denominator in (11). Then the bounded character of $U(z_1,z_2)$ inside the unit circle results in

$$U(0,z) = \frac{(Y(z)z - Y(z) - \beta z)A(Y(z),z)U(0,0) + \beta z A(Y(z),z)U(Y(z),0)}{Y(z)(z - A(Y(z),z))} + \frac{\beta z A_T(Y(z))[U(Y(z),Y(z)) - U(Y(z),0) - U(0,Y(z)) + U(0,0)]}{Y(z)(z - A(Y(z),z))}. \tag{23}$$

After using Exprs. (12) and (22) to calculate $U(Y(z),Y(z))$ and $U(Y(z),0)$, respectively, this can be organised as

$$U(0,z) = a(z) + b(z)U^{(2)}(0,0) + c(z)U(0,Y(z)), \tag{24}$$

with

$$a(z) = \frac{\beta(1-\lambda_T)z(1-A(Y(z),0))(A_T(Y(z)) - A(Y(z),z))}{(z - A(Y(z),z))(Y(z) - A(Y(z),0) + \beta A_T(Y(z)))} + \frac{\beta(1-\lambda_T)zY(z)^{-1}(A_T(Y(z)))^2(Y(z)-1)(A(Y(z),z) - A(Y(z),0))}{(z - A(Y(z),z))(Y(z) - A_T(Y(z)))(Y(z) - A(Y(z),0) + \beta A_T(Y(z)))} + \frac{(1-\lambda_T)(z-1)A(Y(z),z)}{z - A(Y(z),z)}, \tag{25}$$

$$b(z) = \frac{\beta z A(Y(z),0)(A(Y(z),z) - A_T(Y(z)))}{(z - A(Y(z),z))(Y(z) - A(Y(z),0) + \beta A_T(Y(z)))}, \tag{26}$$

$$c(z) = \frac{\beta z A_T(Y(z))(A(Y(z),0) - A(Y(z),z))}{Y(z)(z - A(Y(z),z))(Y(z) - A(Y(z),0) + \beta A_T(Y(z)))}. \tag{27}$$

Eq. (24) can be used in an iterative method to compute $U(0,z)$. In a similar way as for model A, we find that

$$U(0,z) = \sum_{k=0}^{\infty} a(Y_k(z)) \prod_{l=0}^{k-1} c(Y_l(z)) + U^{(2)}(0,0) \sum_{k=0}^{\infty} b(Y_k(z)) \prod_{l=0}^{k-1} c(Y_l(z)) + U(0,C) \prod_{k=0}^{\infty} c(Y_k(z)). \tag{28}$$

It remains for us to determine the constants $U^{(2)}(0,0)$ and $U(0,C)$. F_1, previously defined as $A(Y(F_1),F_1)$, is a zero of the denominator of $U(0,z)$. This zero lies inside the unit circle, which implies that the corresponding numerator must be zero for $z =$

F_1 as well. This produces an equation for $U(0, Y(F_1))$ in terms of $U^{(2)}(0,0)$. Then, replacing z by $Y(F_1)$ in (28) leads to a second equation for $U(0, Y(F_1))$, not only in terms of $U^{(2)}(0,0)$ but also in terms of $U(0,C)$. With Rouché's theorem, furthermore, we can show that the factor $Y(z) - A(Y(z),0) + \beta A_T(Y(z)$ of the denominator of $U(0,z)$ also has one zero inside the unit circle. This zero will be denoted by F_2. Similarly as with F_1, we can find two equations for $U(0, Y(F_2))$, one in terms of $U^{(2)}(0,0)$ and one in terms of $U^{(2)}(0,0)$ and $U(0,C)$. In this way, we get a system of four linear equations in four unknown quantities. By repeatedly eliminating the unknowns, we obtain values for $U^{(2)}(0,0)$ and $U(0,C)$ at the end.

3.5. **Bringing everything together.** Now all unknown quantities in (10) and (11) are calculated, so we can derive semi-analytic expressions for the joint pgfs of the system content. These expressions, however, are rather cumbersome, so they are omitted here. As we have noticed during this section, seeming small details between jumping models cause an important shift in the solution process (and sometimes also in the performance, as we will demonstrate later on).

3.6. **Calculating marginal characteristics.** For the sake of convenience, we start from the functional equations to determine the marginal pgfs $U_H(z) \triangleq U(z,1)$ and $U_L(z) \triangleq U(1,z)$ of the contents of the high- and low-priority queue, respectively. For model A (for model B, we can follow the same procedure), we get

$$U_H(z) = \frac{(z-(1-\beta))A_1(z)U(0,1) + \beta A_T(z)(U(z,z) - U(0,z))}{z-(1-\beta)A_1(z)}, \tag{29}$$

$$U_L(z) = \frac{(z-1)A_2(z)(1-\lambda_T) + (1-(1-\beta)z)A_2(z)U(0,z) + \beta z(1-U(0,1))}{z(1-(1-\beta)A_2(z))}. \tag{30}$$

The functions $U(z,z)$ and $U(0,z)$ have been calculated (see Expr. (12) and (20)); the constant $U(0,1)$, which denotes the probability of having an empty high-priority queue at the beginning of a slot, can be computed numerically by making use of (20). Finally, by invoking the moment generating property on $U_H(z)$ and $U_L(z)$, we can find expressions for the moments of the involved quantities. For the mean values, for example, we obtain

$$\begin{aligned} \mathrm{E}[u_H] &= U_H'(1) \\ &= \frac{(1-\beta\lambda_2)(U(0,1)-1) + \lambda_1 + \beta \mathrm{E}[u_T] - \beta U^{(2)}(0,1)}{\beta}, \end{aligned} \tag{31}$$

$$\begin{aligned} \mathrm{E}[u_L] &= U_L'(1) \\ &= \frac{(1-\beta\lambda_2)(1-U(0,1)) - \lambda_1 + \beta U^{(2)}(0,1)}{\beta}, \end{aligned} \tag{32}$$

with $\mathrm{E}[u_T]$ the mean total system content and $U^{(2)}(0,1) \triangleq \left.\dfrac{dU(0,z)}{dz}\right|_{z=1}$. $\mathrm{E}[u_T]$ is easily calculated via (12); $U^{(2)}(0,1)$ can be computed by using (20).

4. **Packet delay.** The second performance characteristic that we study is the *packet delay*, which is defined as the total amount of time that a packet spends in the system (i.e., the number of slots between the end of the packet's arrival slot and the end of its departure slot). Assuming that the system is in the steady state,

we denote the delay of a type-j packet as d_j. In this section, we show how to compute pgfs of d_1 and d_2 for model A. The same procedure can be followed for model B. The computations for model B, however, are omitted because they are not more complicated than for model A. Moreover, both scheduling models yield comparable results.

4.1. **Type-1 packet delay.** Let us first consider a random but "tagged" type-1 packet. We mark the arrival slot of the packet as slot I. Since a possible jump of the content of the low-priority queue to the high-priority queue takes place at the end of a slot, the type-1 packets that arrive during slot I are stored in front of the type-2 packets that possibly jump in slot I. As a consequence, the delay of the tagged type-1 packet only depends on the content of the high-priority queue at the beginning of slot I (i.e., $u_{H,I}$) and the number of type-1 arrivals during slot I. If $f_{1,I}$ represents the number of these arrivals that have to be transmitted before the tagged packet, we find the following equation for d_1:

$$d_1 = [u_{H,I} - 1]^+ + f_{1,I} + 1. \tag{33}$$

Due to the i.i.d. arrivals from slot to slot, $u_{H,I}$ and the content of the high-priority queue at the beginning of an arbitrary slot have the same distribution. For the same reason, $u_{H,I}$ and $f_{1,I}$ are mutually independent. The pgf of $f_{1,I}$, furthermore, can be calculated by observing that an arbitrary packet is more likely to arrive in a larger bulk (see e.g., [3]):

$$F_1(z) = \frac{A_1(z) - 1}{\lambda_1(z-1)}. \tag{34}$$

The pgf of the delay of a random type-1 packet thus can be easily expressed in terms of $U_H(z)$ and $A_1(z)$, i.e.,

$$D_1(z) = \frac{A_1(z) - 1}{\lambda_1(z-1)} \{U_H(z) + (z-1)U_H(0)\}, \tag{35}$$

with $U_H(0) = U(0,1)$. From this expression, we easily obtain that

$$\mathrm{E}[d_1] = \frac{\lambda_{11}}{2\lambda_1} + \mathrm{E}[u_H] + U_H(0). \tag{36}$$

4.2. **Type-2 packet delay.** Secondly, because of the priority scheduling, it is not straightforward to determine an expression for the pgf $D_2(z)$ of the delay of a random type-2 packet (see also e.g., [14]). Moreover, we have to take into account the possibility that type-2 packets may jump to the high-priority queue during their waiting time (see e.g., [7]). Let us tag an arbitrary type-2 packet that enters the system, and again denote its arrival slot by slot I.

The packets that are in the system at the end of slot I and that have to be transmitted before the tagged packet, are referred to as *primary packets*. The tagged type-2 packet can only be transmitted when all primary packets and all type-1 packets that arrive while the tagged packet is waiting in the low-priority queue, are transmitted. Indeed, new type-1 packets can arrive while a primary packet is transmitted. As long as the tagged packet is in the low-priority queue, these type-1 packets get priority over and are scheduled for transmission before the tagged packet. We say that the primary packet adds a so-called *sub-busy period* to the delay of the tagged packet (see e.g., [14]). A sub-busy period *initiated by a packet* starts at the beginning of the slot in which the packet is transmitted, and ends at the beginning of the slot where – for the first time – the number of packets that

have to be transmitted before the tagged packet, is one less than at the beginning of the sub-busy period. When the tagged packet arrives, three possible situations may occur: no packet is in transmission, a packet of the low-priority queue is in transmission, or a packet of the high-priority queue is in transmission. Following equations for d_2, the delay of the tagged type-2 packet, can be derived for the three cases:

- no packet is in transmission during slot I ($u_{H,I} = u_{L,I} = 0$)

$$d_2 = \sum_{m=1}^{f_{1,I}} v_m + \sum_{m=1}^{f_{2,I}} w_m + 1, \tag{37}$$

- a packet of the low-priority queue is in transmission during slot I ($u_{H,I} = 0, u_{L,I} > 0$)

$$d_2 = \sum_{m=1}^{f_{1,I}} v_m + \sum_{m=1}^{u_{L,I}-1+f_{2,I}} w_m + 1, \tag{38}$$

- a packet of the high-priority queue is in transmission during slot I ($u_{H,I} > 0$)

$$d_2 = \sum_{m=1}^{u_{H,I}-1+f_{1,I}} v_m + \sum_{m=1}^{u_{L,I}+f_{2,I}} w_m + 1, \tag{39}$$

where $u_{H,I}$ and $u_{L,I}$ give the contents of the high- and low-priority queue at the beginning of slot I, where $f_{1,I}$ and $f_{2,I}$ represent the type-1 and type-2 packets that arrive during slot I and that have to be transmitted before the tagged packet, and where v_m and w_m denote the lengths of the m-th sub-busy periods initiated by a packet residing in the high- and low-priority queue in slot I, respectively. Since the contents of both queues are not allowed to merge when the high-priority queue is empty, v_m and w_m have different distributions. We determine these distributions later on. The introduction of pgfs in these equations first yields

$$\begin{aligned}
D_2(z) &\triangleq \mathrm{E}\left[z^{d_2}\right] \\
&= \mathrm{E}\left[z^{d_2}\{u_{H,I} = u_{L,I} = 0\}\right] \\
&\quad + \mathrm{E}\left[z^{d_2}\{u_{H,I} = 0, u_{L,I} > 0\}\right] \\
&\quad + \mathrm{E}\left[z^{d_2}\{u_{H,I} > 0\}\right] \\
&= z\mathrm{E}\left[z^{\sum_{m=1}^{f_{1,I}} v_m + \sum_{m=1}^{f_{2,I}} w_m}\{u_{H,I} = u_{L,I} = 0\}\right] \\
&\quad + z\mathrm{E}\left[z^{\sum_{m=1}^{f_{1,I}} v_m + \sum_{m=1}^{u_{L,I}-1+f_{2,I}} w_m}\{u_{H,I} = 0, u_{L,I} > 0\}\right] \\
&\quad + z\mathrm{E}\left[z^{\sum_{m=1}^{u_{H,I}-1+f_{1,I}} v_m + \sum_{m=1}^{u_{L,I}+f_{2,I}} w_m}\{u_{H,I} > 0\}\right],
\end{aligned} \tag{40}$$

with $\mathrm{E}[X\{Y\}] \triangleq \mathrm{E}[X|Y]\,\mathrm{Prob}[Y]$. By conditioning on the possible jumping instants, furthermore, we can consider three cases for a sub-busy period: the tagged packet is still in the low-priority queue at the beginning of the sub-busy period and no jump occurs during the sub-busy period, the tagged packet is still in the low-priority queue at the beginning of the sub-busy period and a jump occurs during the sub-busy period, or there was already a jump before the sub-busy period so that the tagged packet is already in the high-priority queue. In the last case, new

arriving type-1 packets are queued behind the tagged packet and it takes only one slot to decrease the number of packets in front of it by one. This leads to sub-busy periods of length one, i.e., $v_m = w_m = 1$. For the other cases, we define partial pgfs:

$$V_1(z) \triangleq \mathrm{E}\left[z^{v_m}\{\text{t.p. does not jump during } v_m\} \,|\, \text{t.p. does not jump before } v_m\right], \tag{41}$$

$$V_2(z) \triangleq \mathrm{E}\left[z^{v_m}\{\text{t.p. jumps during } v_m\} \,|\, \text{t.p. does not jump before } v_m\right], \tag{42}$$

$$W_1(z) \triangleq \mathrm{E}\left[z^{w_m}\{\text{t.p. does not jump during } w_m\} \,|\, \text{t.p. does not jump before } w_m\right], \tag{43}$$

$$W_2(z) \triangleq \mathrm{E}\left[z^{w_m}\{\text{t.p. jumps during } w_m\} \,|\, \text{t.p. does not jump before } w_m\right], \tag{44}$$

with t.p. an abbreviation for tagged packet. Then, by conditioning on if and when the tagged packet jumps to the high-priority queue and by subsequently using the definitions of these partial pgfs, (40) is transformed into

$$\begin{aligned}
D_2(z) = {} & z\mathrm{E}\Bigg[V_2(z)z^{f_{2,I}}\sum_{i=1}^{a_{1,I}}(V_1(z))^{i-1}z^{a_{1,I}-i} \\
& + (V_1(z))^{a_{1,I}}W_2(z)\sum_{i=1}^{f_{2,I}}(W_1(z))^{i-1}z^{f_{2,I}-i} \\
& + (V_1(z))^{a_{1,I}}(W_1(z))^{f_{2,I}}\Big\{u_{H,I}=0, u_{L,I}=0\Big\}\Bigg] \\
& + z\mathrm{E}\Bigg[V_2(z)z^{u_{L,I}-1+f_{2,I}}\sum_{i=1}^{a_{1,I}}(V_1(z))^{i-1}z^{a_{1,I}-i} \\
& + (V_1(z))^{a_{1,I}}W_2(z)\sum_{i=1}^{u_{L,I}-1+f_{2,I}}(W_1(z))^{i-1}z^{u_{L,I}-1+f_{2,I}-i} \\
& + (V_1(z))^{a_{1,I}}(W_1(z))^{u_{L,I}-1+f_{2,I}}\Big\{u_{H,I}=0, u_{L,I}>0\Big\}\Bigg] \\
& + z\mathrm{E}\Bigg[\beta z^{u_{H,I}-1+a_{1,I}+u_{L,I}+f_{2,I}} \\
& + (1-\beta)V_2(z)z^{u_{L,I}+f_{2,I}}\sum_{i=1}^{u_{H,I}-1+a_{1,I}}(V_1(z))^{i-1}z^{u_{H,I}-1+a_{1,I}-i} \\
& + (1-\beta)(V_1(z))^{u_{H,I}-1+a_{1,I}}W_2(z)\sum_{i=1}^{u_{L,I}+f_{2,I}}(W_1(z))^{i-1}z^{u_{L,I}+f_{2,I}-i} \\
& + (1-\beta)(V_1(z))^{u_{H,I}-1+a_{1,I}}(W_1(z))^{u_{L,I}+f_{2,I}}\Big\{u_{H,I}>0\Big\}\Bigg].
\end{aligned} \tag{45}$$

When the high-priority queue is empty at the beginning of slot I, the tagged packet cannot jump at the end of slot I (first two terms). When there are packets present in the high-priority queue at the beginning of slot I, on the other hand, the tagged packet stays in the low-priority queue with probability $1-\beta$ or jumps to the high-priority queue with probability β (third term). In the latter case, the tagged packet is in the high-priority queue during its entire delay. It has to wait there until all

primary packets - one per slot - are transmitted. When the tagged packet does not jump at the end of slot I, three things can happen during its waiting time: it jumps during one of the sub-busy periods initiated by a packet of the high-priority queue, it jumps during one of the sub-busy periods initiated by a packet of the low-priority queue, or it does not jump at all.

Before we further work out (45), we determine expressions for the partial pgfs of v_m and w_m. During the first slot of all sub-busy periods, i.e., when the initiating packet is transmitted, type-1 packets may arrive at the system. These type-1 packets start sub-busy periods of their own, and these new sub-busy periods are part of the initial sub-busy period. So the length of the initial sub-busy period equals one (the first slot) plus the sum of the lengths of the sub-busy periods initiated by those arriving type-1 packets (possibly none). This yields

$$v_m = 1 + \sum_{i=1}^{a_1} v_{m,i}, \tag{46}$$

and

$$w_m = 1 + \sum_{i=1}^{a_1} v_{m,i}, \tag{47}$$

with a_1 the number of type-1 arrivals during the first slot of the sub-busy period and $v_{m,i}$ the length of the so-called *secondary* sub-busy period initiated by the i-th type-1 arrival in that slot. As one can notice, we have the same equations for v_m and w_m. Yet, they have different distributions as we will now see. Let us first compute $V_1(z)$. The tagged packet does not jump during the complete sub-busy period. This means that it does not jump at the end of the first slot of the sub-busy period (this happens with probability $1-\beta$) and not during the various secondary sub-busy periods. Hence, the $v_{m,i}$s are stochastically indistinguishable from v_m, all having the same pgf $V_1(z)$. Since the numbers of type-1 arrivals are i.i.d. from slot to slot, the $v_{m,i}$s also are independent from a_1. We thus obtain that

$$V_1(z) = (1-\beta)zA_1(V_1(z)). \tag{48}$$

To calculate $V_2(z)$, we again start from Eq. (46). Now, however, we assume that the tagged packet jumps during the sub-busy period. This means that the tagged packet either jumps at the end of the first slot of the sub-busy period (with probability β) or during one of the secondary sub-busy-periods. From the moment that the tagged packet is in the high-priority queue, all $v_{m,i}$s are equal to one. This leads to

$$\begin{aligned} V_2(z) &= \beta z A_1(z) + (1-\beta)zV_2(z)\mathrm{E}\left[\sum_{i=1}^{a_1} V_1(z)^{i-1} z^{a_1-i}\right] \\ &= \beta z A_1(z) + (1-\beta)zV_2(z)\frac{A_1(z) - A_1(V_1(z))}{z - V_1(z)}. \end{aligned} \tag{49}$$

By isolating $V_2(z)$ in the latter and using Expr. (48) afterwards, we find that

$$V_2(z) = \beta \frac{A_1(z)(z - V_1(z))}{1 - (1-\beta)A_1(z)}. \tag{50}$$

In a similar way, we find expressions for $W_1(z)$ and $W_2(z)$. When the high-priority queue is empty at the beginning of a slot, a packet of the low-priority queue is transmitted in that slot. This type-2 packet also starts a sub-busy period, with secondary sub-busy periods, initiated by the type-1 packets arriving during its first

slot, being part of it. However, at the end of that first slot, the content of the low-priority queue cannot jump to the high-priority queue, because of the studied jumping policy (see (2)). For $W_1(z)$, we thus have that

$$W_1(z) = zA_1(V_1(z)). \tag{51}$$

We follow a similar reasoning for $W_2(z)$. The high-priority queue is empty at the beginning of the first slot of the sub-busy period and no jump occurs at the end of this slot. Hence, there is an occurrence of a jump during one of the secondary sub-busy periods. We get that

$$W_2(z) = zV_2(z)\frac{A_1(z) - A_1(V_1(z))}{z - V_1(z)}. \tag{52}$$

Then substituting (50) in Eq. (52) results in

$$W_2(z) = \beta\frac{zA_1(z)(A_1(z) - A_1(V_1(z)))}{1 - (1-\beta)A_1(z)}. \tag{53}$$

So $V_2(z)$, $W_1(z)$ and $W_2(z)$ all can be expressed as a function of $V_1(z)$, which in turn is implicitly defined in (48).

Let us finally go back to (45). By exploiting the uncorrelated nature of the arrival process from slot to slot, by using some standard mathematical (z-transform) techniques, and by making use of the relations of the partial pgfs with $V_1(z)$, we produce

$$\begin{aligned} D_2(z) = {} & \frac{\beta}{1-(1-\beta)A_1(z)}F(z,z)\Big[(z-1)A_1(z)U(0,0) + (A_1(z)-1)U(0,z) \\ & + U(z,z)\Big] \\ & + \frac{\beta(1-\beta)zA_1(z)(A_1(z)-1)}{(1-(1-\beta)A_1(z))((1-\beta)z - V_1(z))}F(V_1(z),z)\Bigg[(z-1)U(0,0) \\ & + \frac{V_1(z)-(1-\beta)z}{V_1(z)}U(0,z) + \frac{(1-\beta)z}{V_1(z)}U(V_1(z),z)\Bigg] \\ & + \frac{(A_1(z)-1)(\beta z(1-(1-\beta)A_1(z)) - z + V_1(z))}{(1-(1-\beta)A_1(z))((1-\beta)z - V_1(z))}F\left(V_1(z), \frac{V_1(z)}{1-\beta}\right) \\ & \times \left[z\frac{V_1(z)-(1-\beta)}{V_1(z)}U(0,0) + \frac{(1-\beta)z}{V_1(z)}U\left(V_1(z), \frac{V_1(z)}{1-\beta}\right)\right]. \end{aligned} \tag{54}$$

The pgf $F(z_1, z_2)$ is found in [8]:

$$F(z_1,z_2) = \frac{A(z_1,z_2) - A_1(z_1)}{\lambda_2(z_2-1)}. \tag{55}$$

The quantities related to the system content are computed in the previous section. From $D_2(z)$, it is possible to obtain expressions for the moments of the delay of a type-2 packet. The mean value, for example, is calculated as $D_2'(1)$, whereas $D_2''(1) + D_2'(1) - (D_2'(1))^2$ would give the variance.

5. **Numerical examples.**

In the previous section, we have described procedures to obtain expressions for the mean packet delays of both types of traffic. These

expressions can be used to illustrate the differences between the merging mechanisms. We consider the following arrival process:

$$A(z_1, z_2) = \left(1 - \frac{\lambda_1}{16}(1 - z_1) - \frac{\lambda_2}{16}(1 - z_2)\right)^{16}, \tag{56}$$

with λ_1 and λ_2 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 independent and uniform routing towards the outlets. Furthermore, we define α as the fraction of type-1 traffic in the overall traffic mix (i.e., $\alpha = \lambda_1/\lambda_T$, with $\lambda_T = \lambda_1 + \lambda_2$).

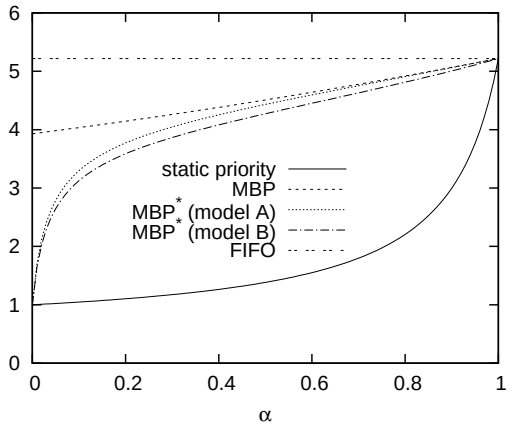

FIGURE 1. Mean type-1 packet delay versus α, with $\beta = 0.4$ and $\lambda_T = 0.9$

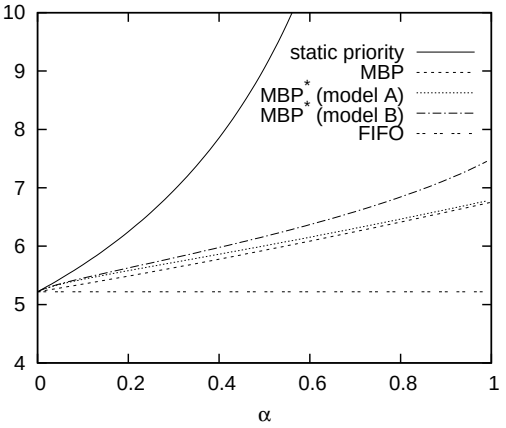

FIGURE 2. Mean type-2 packet delay versus α, with $\beta = 0.4$ and $\lambda_T = 0.9$

Figs. 1 and 2 show the mean packet delays of both types of traffic as functions of α, with $\beta = 0.4$ and the total arrival rate λ_T equal to 0.9, for the original MBP mechanism and the two variants studied in this paper. For the sake of completeness,

we have also depicted the curves for the static priority and FIFO scheduling disciplines. We see that the choice between the merging mechanisms is nearly irrelevant with respect to $\mathrm{E}[d_2]$. For $\mathrm{E}[d_1]$, however, it is easy to notice that MBP* performs better than MBP when α is small, i.e., when the traffic mix mainly consists of type-2 traffic. When α is small, an exceptionally arriving type-1 packet probably enters an empty high-priority queue in the case of MBP*, so the packet is transmitted within a short time period. With MBP, on the other hand, there might be a merge just before the rare type-1 packet arrives. In that case, this packet suffers from an unnecessary delay. When $\alpha = 0.01$, for example, $\mathrm{E}[d_1]$ decreases from about 3.9 for the original MBP mechanism to 1.7 for MBP*. Figs. 1 and 2 finally illustrate that the two MBP* models basically have the same performance. This is confirmed by numerous other examples. In the next figure, we therefore only consider model A.

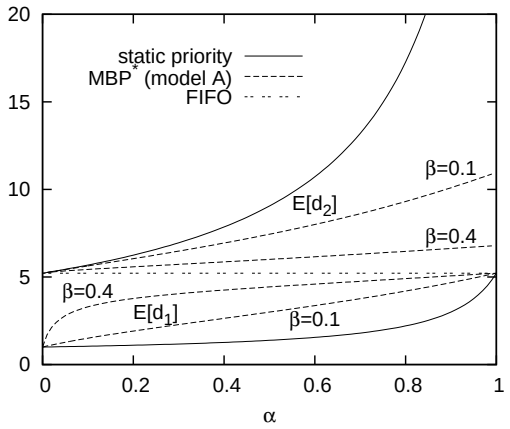

FIGURE 3. Mean packet delays versus α, with $\lambda_T = 0.9$

In Fig. 3, we again show the mean packet delays of both types of traffic as functions of α, but now with $\lambda_T = 0.9$ and for different values of β. A larger value of β implies more jumps and consequently a lower $\mathrm{E}[d_2]$. The price to pay is a higher $\mathrm{E}[d_1]$. The value of β can be chosen according to the delay requirements of both types of traffic. A low β, for example, will highly favour the type-1 (delay-sensitive) traffic, while choosing β higher will achieve a limited delay differentiation between both types of traffic. This is practical if there is little difference in their delay requirements. Most importantly, the parameter β can be fine-tuned to accommodate a required delay differentiation.

6. **Conclusions.** In this paper, we have considered a scheduling discipline with priority jumps. Priority jumps allow for a less drastic delay differentiation between different types of traffic compared to static priority. The introduction of a jumping parameter, moreover, provides a mechanism to control the delay differentiation and to adjust it if necessary (something static priority lacks as well). We have used probability generating functions to analytically study a two-priority queueing system with one server and generally distributed, structured arrivals. We have derived probability generating functions of the system content and of the packet delays. Some mathematical challenges, like the determination of boundary functions and the study of the delay of a low-priority packet, are thereby efficiently

overcome. Probability generating functions, furthermore, are useful in the calculation of important performance measures, such as the mean values of the packet delays. These performance measures have been used to show the impact of the priority scheduling discipline and of some system parameters.

REFERENCES

[1] J. Abate and W. Whitt, *Solving probability transform functional equations for numerical inversion*, Operations Research Letters, **12** (1992), 275–281.

[2] N. Bansal and M. Harchol-Balter, *Analysis of SRPT scheduling: investigating unfairness*, in "Proceedings of the 2001 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems," (2001), 279–290.

[3] H. Bruneel and B. G. Kim, "Discrete-time Models for Communication Systems including ATM," Kluwer Academic Publishers, Boston, 1993.

[4] M. Kargahi and A. Movaghar, *A method for performance analysis of Earliest-Deadline-First scheduling policy*, The Journal of Supercomputing, **37** (2006), 197–222.

[5] M. Katevenis, S. Sidiropoulos and C. Courcoubetis, *Weighted round-robin cell multiplexing in a general-purpose ATM switch chip*, IEEE Journal on Selected Areas in Communications, **9** (1991), 1265–1279.

[6] L.M. Le Ny and B. Tuffin, *Modeling and analysis of multi-class threshold-based queues with hysteresis using Stochastic Petri Nets*, in "Proceedings of the 23rd International Conference on Applications and Theory of Petri Nets; Lecture Notes In Computer Science, Vol. 2360," (2002), 254–272.

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

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

[9] T. Maertens, J. Walraevens and H. Bruneel, *Performance comparison of several priority schemes with prioriy jumps*, Annals of Operations Research, **162** (2008), 109–125.

[10] V. Ramaswami and D. M. Lucantoni, *Algorithmic analysis of a dynamic priority queue*, in "Applied Probability - Computer Science: The Interface, Vol. II" (eds. R. L. Disney and T. J. Ott), (1982), 157–206.

[11] M. Shreedhar and G. Varghese, *Efficient fair queuing using deficit round-robin*, IEEE/ACM Transactions on Networking, **4** (1996), 375–385.

[12] A. Sugahara, T. Takine, Y. Takahashi and T. Hasegawa, *Analysis of a non-preemptive priority queue with SPP arrivals of high class*, Performance Evaluation, **21** (1995), 215–238.

[13] J. Walraevens, "Discrete-time Queueing Models with Priorities," Ph.D thesis, Ghent University, 2004.

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

[15] J. Walraevens, J. S. H. van Leeuwaarden and O. J. Boxma, *Power series approximations for two-class generalized processor sharing systems*, Queueing Systems, **66** (2010), 107–130.

[16] L. Zhang, *Virtual clock: a new traffic control algorithm for packet switching networks*, ACM Transactions on Computer Systems, **9** (1990), 101–124.

Received June 2011; revised August 2011.

E-mail address: tmaerten@telin.UGent.be
E-mail address: jw@telin.UGent.be
E-mail address: hb@telin.UGent.be