JOURNAL OF INDUSTRIAL AND
MANAGEMENT OPTIMIZATION
Volume **7**, Number **3**, August **2011**

doi:10.3934/jimo.2011.7.735

pp. **735–751**

PARTIALLY SHARED BUFFERS WITH FULL OR MIXED PRIORITY

Thomas Demoor, Dieter Fiems, Joris Walraevens
and Herwig Bruneel

Department of Telecommunications and Information Processing, Ghent University
St-Pietersnieuwstraat 41, 9000 Gent, Belgium

Abstract. This paper studies a finite-sized discrete-time two-class priority queue. Packets of both classes arrive according to a two-class discrete batch Markovian arrival process (2-DBMAP), taking into account the correlated nature of arrivals in heterogeneous telecommunication networks. The model incorporates time and space priority to provide different types of service to each class. One of both classes receives absolute time priority in order to minimize its delay. Space priority is implemented by the partial buffer sharing acceptance policy and can be provided to the class receiving time priority or to the other class. This choice gives rise to two different queueing models and this paper analyses both these models in a unified manner. Furthermore, the buffer finiteness and the use of space priority raise some issues on the order of arrivals in a slot. This paper does not assume that all arrivals from one class enter the queue before those of the other class. Instead, a string representation for sequences of arriving packets and a probability measure on the set of such strings are introduced. This naturally gives rise to the notion of intra-slot space priority. Performance of these queueing systems is then determined using matrix-analytic techniques. The numerical examples explore the range of service differentiation covered by both models.

1. **Introduction.** Packet based (IP) networks endure an ever-increasing tension as numerous applications, each requiring different Quality of Service (QoS) standards, concurrently utilize the network. In general, two requirements can be distinguished and packets requiring minimal mean delay and delay jitter are said to request time priority whereas space priority requests minimal packet loss. The most popular approach to fulfilling different requirements for different applications distributes packets into traffic classes according to their type of application. QoS differentiation is then obtained by providing different types of service to each class.

This contribution studies a single-server queueing model with two traffic classes and a single finitely-sized buffer. Time is discretizised into fixed-length intervals (slots). Arrivals are modelled as a two-class discrete batch Markovian arrival process (2-DBMAP), which can take into account burstiness of network traffic as well as

2000 *Mathematics Subject Classification.* Primary: 60J10, 60K25; Secondary: 68M12.

Key words and phrases. Queueing theory, time priority, space priority, partial buffer sharing, matrix analytic method.

The second and third authors are Postdoctoral Fellows with the Research Foundation Flanders (FWO-Vlaanderen), Belgium. This paper was presented at the QTNA2010 conference which was held in Beijing, China, during July 24-26, 2010. An earlier and brief version of this paper was published in the Conference Proceedings. The reviewing process of the paper was handled by Wuyi Yue and Yutaka Takahashi as Guest Editors.

correlation between both classes. The system can differentiate the service it delivers by giving a class time priority and/or space priority. The class receiving time priority has absolute transmission priority. Furthermore, space priority is provided by adopting the partial buffer sharing (PBS) acceptance policy. As there are two classes, say class 1 and class 2, there are four possible combinations of the two priority types. However, we only need to consider two as the two others then follow directly by swapping class 1 and 2 around. Therefore, we can choose class 1 to have time priority. The model wherein this class also receives space priority is called Full Priority (FP), whereas the term Mixed Priority (MP) is used when class 2 receives space priority. We study both models in a unified manner.

First, we survey literature on providing time and/or space priority and point out differences with the current contribution. Time priority has been studied extensively in discrete-time absolute priority queues with infinite queue size, e.g. [13, 8, 15, 10, 16]. However, infinite queue sizes imply no packet loss and thus space priority cannot be considered. Finite-sized priority queueing systems [2, 14] were discussed as well but here space priority is not studied intensively. Here, each class has its dedicated queue whereas the current paper a single queue is shared by both classes and space priority is also provided. PBS is easily implemented [7] and has been widely studied in, [17, 6, 18]. However, these papers do not include time priority as packets of all classes are served in a First-In-First-Out manner. The current contribution encapsulates [3] and extends [5]. The former paper studies networks where packets are categorized into two classes: real-time packets (telephony, multimedia, gaming,...) requiring time priority and data packets (file transfer, email,...) requiring space priority and consequently the MP-model is perfect for providing QoS. In contrast, [5] studies the FP-model in order to provide QoS in scalable video coding (SVC) (see e.g. [11]), which uses two types of packets: base layer and enhancement layer packets. The former are required to decode and playback the video, although at poor quality, whereas enhancement packets only increase quality. Here, the FP-model is clearly appropriate. However, this paper assumed that, at a slot boundary, all base-layer packets arrive before enhancement packets. The current contribution does not make any assumptions on the order of arrivals and provides unified formulas for both models.

Note that the used terminology stems from the IP network context. However, the results can be applied in any domain by thinking of packets as customers, time priority as fast service and space priority as guaranteed service. The queueing models under consideration are solved using matrix analytic methods, see e.g. [9]. The remainder of this contribution is organized as follows. The queueing model is described in the next section and is subsequently analysed in section 3. The next section determines how to obtain several performance measures and is followed by a section elaborating on intra-slot space priority. Section 6 illustrates the obtained results by means of some numerical examples. Finally, this paper is concluded in section 7.

2. **Queueing model.**

We consider a discrete-time single-server priority queueing system. Time is divided into fixed-length intervals (slots) and arrivals and departures are synchronized with respect to slot boundaries. There are two classes of packets, say class 1 and 2. Transmission times of packets of both classes are assumed to be fixed and equal to the slot length. Each slot, a packet enters the server for transmission if any packets are present in the queue. In the remainder,

we thus distinguish between the queue and the server. The queue capacity of the system under investigation is finite as the queue can only store up to N packets simultaneously.

Actually, this paper studies two queueing models. In both models, time priority is granted to class 1 in the form of absolute transmission priority over class 2. In the first model, class 1 gets time and space priority over class 2 and it is called the FP-model. In contrast, class 2 receives space priority in the MP-model. This paper presents unified formulas for both models. Space priority is provided by adopting the PBS buffer acceptance policy with threshold T ($0 \leq T \leq N$). A packet of the class with space priority can enter the buffer containing less than N packets upon arrival of the packet whereas a packet of the other class is only allowed into a buffer containing no more than T packets upon arrival of the packet. Thereby, packet loss is minimized for the prioritized class. Here, the packets that are present "upon arrival" of a certain packet include the packets that arrived at the same slot boundary but that entered the queue before this packet.

Queue finiteness and PBS cause packet loss and the order in which packets arrive at a slot boundary determines which packets are lost. In the literature, this peculiarity is avoided by assuming that all packets of a certain class arrive concurrently. However, this may not hold in practice. Furthermore, when rearrangement is possible, it often can be exploited to improve performance. We consider a more formal arrival process making no assumptions on the order of arrivals at a slot boundary by using a string representation leading to the notion of intra-slot space priority (ISP), which is discussed in section 5. The arrival sequence at a slot boundary is embodied by a vector $\boldsymbol{x}$ with i-th element $x_i \in \{1, 2\}$ denoting the class-type of the ith packet. The total number of arrivals obviously equals the total number of elements of $\boldsymbol{x}$, given by $\dim(\boldsymbol{x})$. For instance, a class-2 arrival followed by a class-1 arrival and another class-2 arrival is depicted by the vector $\boldsymbol{x} = [2\ 1\ 2]$ whereas a slot with no arrivals corresponds to $\boldsymbol{x} = [\ ]$. For each $n \in \mathbb{N}$, there are 2^n vectors representing a possible arrival sequence. Let the set of all vectors representing an arrival sequence be denoted by Ω. The arrival process is then specified by defining appropriate probability measures on Ω.

Remark: the length of such sequences is, in general, unbounded. Furthermore, considering only a finite number of elements of a sequence is not sufficient as an infinite amount of packets without space priority can be dropped due to the threshold while the following packets with space priority are accepted. However, assigning probability to such sequences is merely a cruiserweight challenge as "run length encoding" immediately yields that only a finite number of "transitions" between consecutive arrivals of the same class (and thus only a finite number of elements) need to be considered to determine which packets enter the buffer.

Let the vector $\boldsymbol{a}_k$ represent the arrival sequence at the kth slot boundary. In this paper, class-1 and class-2 arrivals are modelled by means of a 2-class discrete-time batch Markovian arrival process (2-DBMAP). As we need to keep track of the entire sequence of arrivals, the definition of this process is more general than the standard one [19]. In the current contribution, a 2-DBMAP is completely characterized by the $Q \times Q$ matrices $\boldsymbol{A}(\boldsymbol{x})$ governing the transitions from slot to slot of the underlying discrete-time Markov chain when arrivals occur according to the sequence $\boldsymbol{x} \in \Omega$. Here, Q denotes the size of the state space of the underlying chain. We have

$$\boldsymbol{A}(\boldsymbol{x}) = \Big[\Pr[\boldsymbol{a}_k = \boldsymbol{x}, s_{k+1} = j | s_k = i] \Big]_{i,j=1,\ldots,Q}, \tag{1}$$

with s_k the state of the underlying Markov chain during slot k. As $\boldsymbol{A}(\cdot)$ is to be a proper (probability) measure, for any set $\Phi \subseteq \Omega$ we have

$$\boldsymbol{A}(\Phi) = \sum_{\boldsymbol{x} \in \Phi} \boldsymbol{A}(\boldsymbol{x}) \,. \tag{2}$$

The number of class-m packets amongst the first n packets in an arrival sequence $\boldsymbol{x}$ is given by

$$c_m^n(\boldsymbol{x}) = \sum_{i=1}^{min(n,dim(\boldsymbol{x}))} 1\{x_i = m\} \,, \tag{3}$$

$m = 1, 2$, with $1\{.\}$, the indicator function, evaluating to 1 if its argument is true and to 0 if it is false. Obviously, the vector dimension cannot be exceeded. The total number of arrivals at the kth boundary, $a_{T,k}$, equals $\dim(\boldsymbol{a}_k)$. Also note that the number of class-i packets arriving at the kth boundary, $a_{i,k}$, is easily found to be equal to $c_i^\infty(\boldsymbol{a}_k)$.

For further use, let ρ_i denote the mean number of packets of class i ($i = 1, 2$) that arrive at a slot boundary and be defined as

$$\rho_1 = \sum_{\boldsymbol{x} \in \Omega} c_1^\infty(\boldsymbol{x}) \boldsymbol{\psi} \boldsymbol{A}(\boldsymbol{x}) \boldsymbol{e} \;, \quad \rho_2 = \sum_{\boldsymbol{x} \in \Omega} c_2^\infty(\boldsymbol{x}) \boldsymbol{\psi} \boldsymbol{A}(\boldsymbol{x}) \boldsymbol{e} \,. \tag{4}$$

Here $\boldsymbol{e}$ is a column vector of ones and $\boldsymbol{\psi}$ is the steady-state probability row vector of the underlying Markov chain, i.e., it is the unique non-negative solution of

$$\boldsymbol{\psi} = \boldsymbol{\psi} \boldsymbol{A}(\Omega) \,, \quad \boldsymbol{\psi} \boldsymbol{e} = 1 \,. \tag{5}$$

Furthermore, let $\rho = \rho_1 + \rho_2$ denote the total arrival load.

Due to the possible simultaneity of arrivals of both classes and departures at slot boundaries, one needs to specify the order in which these arrivals and departures are processed at a boundary. We here assume that the departure, if any, occurs before any arrivals. In the remainder, observation of the queue "at slot boundaries" means after possible departures but before arrivals.

3. Queueing analysis.

We first relate the total number of packets and the number of class-2 packets in the queue at consecutive slot boundaries. These relations contain the notion of effective arrivals and these are subsequently derived in the second subsection. Finally, a set of balance equations can be established and solved numerically.

3.1. System equations.

Consider slot boundary k and let u_k and v_k denote the total queue content and the class-2 queue content — i.e., the total number of packets and the number of class-2 packets in the queue — at this slot boundary. Possibly, some arriving packets are not accepted into the queue giving rise to packet loss. Therefore, let $\tilde{a}_{1,k}$ and $\tilde{a}_{2,k}$ denote the number of class-1 and class-2 packets arriving at the kth slot boundary that the system accommodates (called effective arrivals in the remainder).

The system equations relate the total queue content and the class-2 queue content at consecutive slot boundaries. As a packet leaves the queue at the $(k+1)$th boundary if there are any packets present, the total queue content evolves according to

$$u_{k+1} = (u_k + \tilde{a}_{1,k} + \tilde{a}_{2,k} - 1)^+ \,. \tag{6}$$

Here $(\cdot)^+$ is the usual shorthand notation for $\max(\cdot, 0)$.

The evolution of the class-2 queue content is more intricate. If a class-1 packet enters the server at the $(k+1)$st slot boundary, this is if $u_k - v_k + \tilde{a}_{1,k} > 0$, class-2 packets obviously have no access to the server yielding

$$v_{k+1} = v_k + \tilde{a}_{2,k}\,. \tag{7}$$

On the other hand, if there are no class-1 packets present, this is if $u_k - v_k + \tilde{a}_{1,k} = 0$, a class-2 packet enters the server, if any is present. This produces

$$v_{k+1} = (v_k + \tilde{a}_{2,k} - 1)^+\,. \tag{8}$$

3.2. **Effective arrivals.** Before constructing the balance equations from the system equations, we introduce some auxiliary functions which will allow us to describe both models in unified formulas in the remainder of this paper. The number of effective arrivals when the queue content equals n and packets arrive according to the vector $\boldsymbol{x}$ are given by

$$\begin{aligned} \tilde{a}_1^n(\boldsymbol{x}) &= \begin{cases} \min(c_1^\infty(\boldsymbol{x}), N - n - \tilde{a}_2^n(\boldsymbol{x}))\,, & \text{Full Priority} \\ c_1^{T-n}(\boldsymbol{x})\,, & \text{Mixed Priority} \end{cases} \\ \tilde{a}_2^n(\boldsymbol{x}) &= \begin{cases} c_2^{T-n}(\boldsymbol{x})\,, & \text{Full Priority} \\ \min(c_2^\infty(\boldsymbol{x}), N - n - \tilde{a}_1^n(\boldsymbol{x}))\,, & \text{Mixed Priority} \end{cases} \end{aligned}. \tag{9}$$

Consequently, $\tilde{a}_{i,k} = \tilde{a}_i^{u_k}(\boldsymbol{a}_k)$. Note that the queue accommodates arriving packets of the class receiving space priority until there are N packets in the queue and packets of the other class until there are T packets in the queue. Especially note that the number of effective arrivals of the class without space priority is obtained first as it appears in the expression for that of the other class. This stems from the fact that, obviously, the threshold is reached before the entire buffer is full (concurrently if $T = N$).

The maximum number of effective class-i arrivals in a slot given the queue content equals n is denoted by $\tilde{a}_i^{\max}$(n) yielding

$$\begin{aligned} \tilde{a}_1^{\max}(n) &= \begin{cases} N - n\,, & \text{Full Priority} \\ (T-n)^+\,, & \text{Mixed Priority} \end{cases} \\ \tilde{a}_2^{\max}(n) &= \begin{cases} (T-n)^+\,, & \text{Full Priority} \\ N - n\,, & \text{Mixed Priority} \end{cases} \end{aligned}. \tag{10}$$

Notice that, evidently, these functions exactly oppose each other for both models. Next, the class-2 queue content given that the total queue content equals n ranges from $v_{\min}(n)$ to $v^{\max}(n)$ with

$$\begin{aligned} v_{\min}(n) &= \begin{cases} 0\,, & \text{Full Priority} \\ (n+1-T)^+\,, & \text{Mixed Priority} \end{cases} \\ v^{\max}(n) &= \begin{cases} \min(T, n)\,, & \text{Full Priority} \\ n\,, & \text{Mixed Priority} \end{cases} \end{aligned}. \tag{11}$$

Especially note that, in the FP-model, there are at most T class-2 packets present in the buffer, whereas in the MP-model, there must be class-2 packets present if the total content exceeds $T-1$ as the buffer can only contain up to $T-1$ class-1 packets immediately following a departure. Furthermore, in the MP-model, $T = 0$ is an exceptional case as then $v_{\min}(n)$ should equal n and not $n+1$. Here, the system

behaves as a FIFO queue with a single class of (class-2) packets as all class-1 packets are dropped.

Let $\tilde{\boldsymbol{A}}_u(m,n)$ denote the matrix governing the transitions of the underlying Markov chain at a slot boundary when there are m effective class-1 arrivals and n effective class-2 arrivals, given that there are u packets in the queue at that slot boundary. That is

$$\begin{aligned}\tilde{\boldsymbol{A}}_u(m,n) &= \Big[\Pr[\tilde{a}_{1,k}=m,\tilde{a}_{2,k}=n,s_{k+1}=j|s_k=i,u_k=u]\Big]_{i,j=1,\dots,Q}\\ &= \sum_{x\in\Omega}\boldsymbol{A}(\boldsymbol{x})1\{m=\tilde{a}_1^{u_k}(\boldsymbol{x}),\ n=\tilde{a}_2^{u_k}(\boldsymbol{x})\}\,,\end{aligned} \tag{12}$$

for $u=0,\dots,N-1$, $m\geq 0$ and $n\geq 0$.

3.3. **Balance equations.** Clearly, the triple (u_k,v_k,s_k) describes the state of the queueing system at the kth slot boundary in the Markovian sense. Therefore, let $\boldsymbol{\pi}_k(m,n)$ denote the row vector whose ith entry is the probability to have $n-m$ class-1 and m class-2 packets in the queue at the kth slot boundary while the arrival process is in state i, i.e.,

$$\boldsymbol{\pi}_k(m,n)=\Big[\Pr\left[v_k=m,u_k=n,s_k=i\right]\Big]_{i=1,\dots,Q}\,, \tag{13}$$

for $n=0,\dots,N-1$ and $m=v_{\min}(n),\dots,v^{\max}(n)$. In view of (6), (7) and (8), relating slots k and $k+1$ and conditioning on the state of the server yields

$$\begin{aligned}\boldsymbol{\pi}_{k+1}(m,n) = &\sum_{j=0}^{\min(n+1,N-1)}\sum_{i=v_{\min}(j)}^{v^{\max}(j)}\boldsymbol{\pi}_k(i,j)\tilde{\boldsymbol{A}}_j(n-m+i+1-j,m-i)\\ &+1\{n=m\}\sum_{j=0}^{\min(n+1,N-1)}\boldsymbol{\pi}_k(j,j)\tilde{\boldsymbol{A}}_j(0,n-j+1)\\ &+1\{n=m=0\}\,\boldsymbol{\pi}_k(0,0)\tilde{\boldsymbol{A}}_0(0,0)\,,\end{aligned} \tag{14}$$

for $n=0,\dots,N-1$ and $m=v_{\min}(n),\dots,v^{\max}(n)$.

Grouping the vectors $\boldsymbol{\pi}_k(m,n)$ by total total queue content defines the row vectors

$$\boldsymbol{\pi}_k(n)=[\boldsymbol{\pi}_k(v_{\min}(n),n),\dots,\boldsymbol{\pi}_k(v^{\max}(n),n)]\,, \tag{15}$$

for $n=0,\dots,N-1$. The set of equations (14) then has block matrix representation

$$\boldsymbol{\pi}_{k+1}(n)=\sum_{j=0}^{\min(n+1,N-1)}\boldsymbol{\pi}_k(j)\boldsymbol{C}(j,n)\,, \tag{16}$$

where the block elements (of size $Q\times Q$) of $\boldsymbol{C}(j,n)$ are given by

$$\begin{aligned}c_{i+1,m+1}(j,n) = &\tilde{\boldsymbol{A}}_j(n+1-j-m+i,m-i)\\ &+1\{m=n,\,i=j\}\tilde{\boldsymbol{A}}_j(0,n-j+1)\\ &+1\{n=j=0\}\tilde{\boldsymbol{A}}_0(0,0)\,,\end{aligned} \tag{17}$$

for $i=v_{\min}(j),\dots,v^{\max}(j)$ and $m=v_{\min}(n),\dots,v^{\max}(n)$. Note that $c_{i+1,m+1}(j,n)$ corresponds to the evolution of $\boldsymbol{\pi}_k(i,j)$ to $\boldsymbol{\pi}_{k+1}(m,n)$.

Under mild assumptions, the Markov chain under consideration has only one ergodic class, see [12] for details. Consequently, there exists a unique stationary

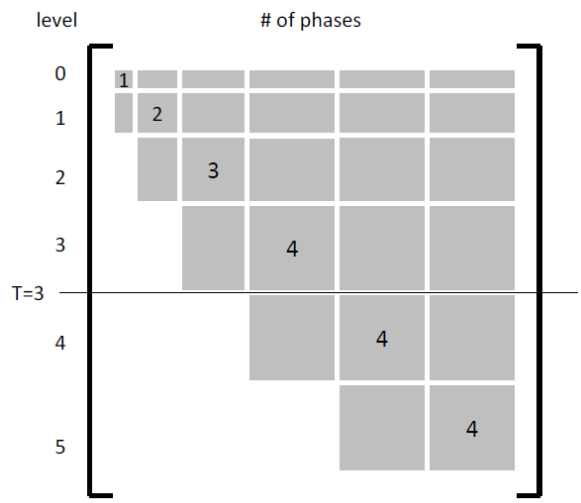

FIGURE 1. Transition matrix block structure for $N = 6$, $T = 3$.

distribution (a non-negative normalized vector), say $\boldsymbol{\pi} = [\boldsymbol{\pi}(0), \ldots, \boldsymbol{\pi}(N-1)]$ — with $\boldsymbol{\pi}(n) = [\boldsymbol{\pi}(v_{\min}(n), n), \ldots, \boldsymbol{\pi}(v^{\max}(n), n)]$ — satisfying the balance equations

$$\boldsymbol{\pi}(n) = \sum_{j=0}^{\min(n+1,N-1)} \boldsymbol{\pi}(j)\boldsymbol{C}(j,n)\,, \tag{18}$$

for $n = 0, \ldots, N-1$. Consequently, the transition matrix of the priority queueing system under consideration has an upper-Hessenberg block-structure with varying block sizes which is efficiently solved by means of a linear level reduction algorithm [1]. In the block matrix, the level (block-row number) indicates the total queue content while the phase (size of a block element) indicates the class-2 queue content and the state of the arrival process. In general, the number of phases equals $(v^{\max}(n) - v_{\min}(n)) \times Q$ at level n Consequently, for $n \leq T$, the number of phases equals $(n+1) \times Q$ as, out of n packets in total, from 0 up to n packets can be of class 2. For levels $n > T$, the block size remains constant at $(T+1) \times Q$ and $T \times Q$ for the FP- and MP-model respectively as the class-2 queue content can vary from 0 to T and from $n+1-T$ to n respectively. Figure 1 demonstrates the block structure of the FP-model for a small example ($N = 6$, $T = 3$).

4. **Performance analysis.** Once $\boldsymbol{\pi}(n)$ has been obtained, various performance measures can be derived. This section describes how to calculate supported load, packet loss, queue content at a random slot boundary and at a random point in time and mean packet waiting time.

The supported class-i load $\tilde{\rho}_i$ is defined as the average number of class-i packets arriving at a slot boundary that are accommodated by the queue. They are determined by

$$\tilde{\rho}_1 = \sum_{i=0}^{N-1} \sum_{m=0}^{\tilde{a}_1^{\max}} \sum_{n=0}^{\tilde{a}_2^{\max}} m\boldsymbol{\pi}(i)\boldsymbol{e}\tilde{\boldsymbol{A}}_i(m,n)\boldsymbol{e}\,, \quad \tilde{\rho}_2 = \sum_{i=0}^{N-1} \sum_{m=0}^{\tilde{a}_1^{\max}} \sum_{n=0}^{\tilde{a}_2^{\max}} n\boldsymbol{\pi}(i)\boldsymbol{e}\tilde{\boldsymbol{A}}_i(m,n)\boldsymbol{e}\,. \tag{19}$$

Note that $\boldsymbol{\pi}(i)\boldsymbol{e}$ is a row vector of size Q with ith element denoting the probability that the queue contains i packets in total and that the underlying chain of the arrival process is in state j, $(1 \leq j \leq Q)$. Furthermore, the total supported load is given by $\tilde{\rho} = \tilde{\rho}_1 + \tilde{\rho}_2$.

Alternatively, the supported load can also be retrieved by observing the departure process. As the system is stationary, the total supported load has to equal the probability that a packet leaves the queue at a random slot boundary. As a packet departs at each slot boundary except when the queue is empty, this produces

$$\tilde{\rho} = 1 - \boldsymbol{\pi}(0,0)\tilde{\boldsymbol{A}}_0(0,0)\boldsymbol{e}\,. \tag{20}$$

Similarly, the class-1 supported load equals the probability of a class-1 departure at a random slot boundary. A class-1 packet leaves the queue if there are class-1 packets present in the queue. Thus

$$\tilde{\rho}_1 = 1 - \sum_{m=0}^{v^{\max}(N-1)} \sum_{n=0}^{\tilde{a}_2^{\max}(m)} \boldsymbol{\pi}(m,m)\tilde{\boldsymbol{A}}_0(0,n)\boldsymbol{e}\,. \tag{21}$$

Note that the appearance of $v^{\max}(N-1)$ indicates that, at a slot boundary, the system can contain up to T and up to $N-1$ class-2 packets for the FP- and MP-model respectively. Also, the class-2 supported load is easily determined as $\tilde{\rho}_2 = \tilde{\rho} - \tilde{\rho}_1$.

The packet loss ratio is the fraction of packets that cannot be accommodated by the queue. In view of the definitions of supported load and packet loss ratio, one easily derives the packet loss ratio of class-1 packets (plr_1), of class-2 packets (plr_2) and of all packets (plr) to be

$$\mathrm{plr}_1 = 1 - \frac{\tilde{\rho}_1}{\rho_1}\,, \quad \mathrm{plr}_2 = 1 - \frac{\tilde{\rho}_2}{\rho_2}\,, \quad \mathrm{plr} = 1 - \frac{\tilde{\rho}}{\rho}\,. \tag{22}$$

Let u_1 and u_2 denote the class-1 and class-2 queue content at a random slot boundary. Since $\boldsymbol{\pi}(m,n)$ is the joint distribution of the queue content of both classes, all moments (mean, variance, etc.) of the random variables u_1 and u_2 are easily obtained. For instance, the i-th moment of the class-j queue content at random slot boundaries $\overline{u}_j^{(i)}$ is given by

$$\overline{u}_1^{(i)} = \sum_{n=0}^{N-1} \sum_{m=v_{\min}(n)}^{v^{\max}(n)} (n-m)^i \boldsymbol{\pi}(m,n)\boldsymbol{e}\,, \quad \overline{u}_2^{(i)} = \sum_{n=0}^{N-1} \sum_{m=v_{\min}(n)}^{v^{\max}(n)} m^i \boldsymbol{\pi}(m,n)\boldsymbol{e}\,. \tag{23}$$

The mean total queue content is given by $\overline{u}^{(1)} = \overline{u}_1^{(1)} + \overline{u}_2^{(1)}$. Similar expressions can be established for joint moments. For instance, the covariance between the queue content of both classes at a random slot boundary is given by

$$\mathrm{Cov}(u_1,u_2) = \sum_{n=0}^{N-1} \sum_{m=v_{\min}(n)}^{v^{\max}(n)} (n-m)m\boldsymbol{\pi}(m,n)\boldsymbol{e} - \overline{u}_1^{(1)}\overline{u}_2^{(1)}\,. \tag{24}$$

When the queue is observed during a random slot (or equivalently at random points in time), this is after all departures and arrivals occurred at the preceding slot boundary, let $\theta(m,n)$ denote the probability that it contains $n-m$ class-1 and m class-2 packets. These packets either were already present at the preceding slot boundary or have arrived at that slot boundary. Consequently, the queue content

at random slots is easily obtained from the one at random slot boundaries yielding

$$\theta(m,n) = \sum_{i=0}^{m}\sum_{j=0}^{n} \boldsymbol{\pi}(i,j)\tilde{\boldsymbol{A}}_j(n-m-j+i, m-i)\boldsymbol{e}\,, \tag{25}$$

for $n = 0, \ldots, N$ and for $m = v_{\min}(n), \ldots, v^{\max}(n)$. Notice that the queue can now contain up to N packets as we no longer observe the system immediately following a departure. Again the i-th moment of the class-j queue content at random points in time $\overline{y}_i$ is given by

$$\overline{y}_1^{(i)} = \sum_{n=0}^{N}\sum_{m=v_{\min}(n)}^{v^{\max}(n)} (n-m)^i\theta(m,n)\boldsymbol{e}\,, \quad \overline{y}_2^{(i)} = \sum_{n=0}^{N}\sum_{m=v_{\min}(n)}^{v^{\max}(n)} m^i\theta(m,n)\boldsymbol{e}\,. \tag{26}$$

Alternatively, $y_i^{(1)}$ can also be obtained by noting that there are $\tilde{\rho}_i$ class-i arrivals at a slot boundary on average yielding

$$\overline{y}_i^{(1)} = \overline{u}_i^{(1)} + \tilde{\rho}_i\,. \tag{27}$$

Consequently, calculating $\theta(m,n)$ is superfluous when one is only interested in the mean values $y_i^{(1)}$.

Packet waiting time is defined as the number of slots a packet spends in the queueing system. Applying Little's law, the mean class-i ($i = 1, 2$) waiting time is found as

$$\overline{w}_i^{(1)} = \frac{1}{\tilde{\rho}_i}\overline{y}_i^{(1)} = \frac{1}{\tilde{\rho}_i}\overline{u}_i^{(1)} + 1\,. \tag{28}$$

Notice that here Little's result does not relate the mean waiting time to the mean queue content at random slot boundaries but to the mean queue content at random slots. This is caused by the chosen order of arrival, observation and departure epochs in our queueing model as illustrated in [4].

5. **Intra-slot space priority.** The order in which packets arrive at a slot boundary can be seen as means of providing intra-slot space priority (ISP), as it partially determines which of these packets, if any, are dropped. Obviously, ISP will have a larger effect when a large number of packets arrive at a slot boundary. The literature generally assumes that all class-1/class-2 packets arrive before packets of the other class (class-1/2 ISP). In some applications, reordering the arrivals at a slot boundary is feasible. This can consequently be exploited to improve performance. For instance, as the FP-model provides time- and space priority to class-1 packets, it is beneficial to use class-1 ISP as well. In contrast, the MP-model gives space priority to class-2 packets so it seems natural to give these packets ISP as well. Furthermore, in lots of real-life applications rearranging is infeasible and packets often arrive in a completely random order (no ISP).

Theoretically, ISP is achieved by only allowing certain forms of arrival sequences $\boldsymbol{x} \in \Omega$ to correspond with non-zero entries in the matrix $\boldsymbol{A}(\boldsymbol{x})$. Let us call the set of vectors of this form Ψ, making this formally equal to $\boldsymbol{A}(\Psi) = \boldsymbol{A}(\Omega)$. When Ψ contains a reasonably small number of vectors, determining $A_u(m,n)$ is straightforward by combining (3), (9) and (12). However, this becomes increasingly tedious as Ψ contains more elements. This can be avoided by giving up some generality on the order of arrivals. Here, information about the order of arrivals is removed from

the arrival process but assumed to be generally known. This enables writing the arrival process as a standard 2-DBMAP [19] given by

$$\boldsymbol{A}(m,n) = \Big[\Pr[a_{1,k} = m, a_{2,k} = n, s_{k+1} = j | s_k = i]\Big]_{i,j=1..Q}, \tag{29}$$

that only keeps track of the number of arrivals of each class at a slot boundary. Several cases where the order of arrivals can be assumed to be generally known were mentioned above: class-1/2 ISP and no ISP. The remainder of this section will elaborate on this matter.

5.1. **Class-1 intra-slot space priority.** Here, all class-1 packets are assumed to arrive before class-2 packets. Consequently, the set Ψ fulfilling $\boldsymbol{A}(\Psi) = \boldsymbol{A}(\Omega)$ is the set of all arrival sequences $\boldsymbol{x}$ of the form

$$\boldsymbol{x} = \big[\underbrace{1\ldots1}_{m}\underbrace{2\ldots2}_{n}\big], \tag{30}$$

for $m, n \geq 0$, representing a slot boundary with m class-1 and n class-2 arrivals. If class-1 ISP is assumed, the only information held by such a vector are the values of m and n. Consequently, (12) simplifies to

$$\tilde{\boldsymbol{A}}_u(m,n) = \sum_{i=m}^{\infty}\sum_{j=n}^{\infty} \boldsymbol{A}(i,j) 1\{m = \min(i, \tilde{a}_1^{\max}(u)), n = \min(j, \tilde{a}_2^{\max}(u+m))\}. \tag{31}$$

5.2. **Class-2 intra-slot space priority.** In this case, each non-zero probability arrival sequence $x \in \Psi$ is of the form

$$\boldsymbol{x} = \big[\underbrace{2\ldots2}_{m}\underbrace{1\ldots1}_{n}\big]. \tag{32}$$

Again, $\boldsymbol{A}(\Psi) = \boldsymbol{A}(\Omega)$ and only m and n need to be accounted for and thus (29) holds again . Here, (12) simplifies to

$$\tilde{\boldsymbol{A}}_u(m,n) = \sum_{i=m}^{\infty}\sum_{j=n}^{\infty} A(i,j) 1\{m = \min(i, \tilde{a}_1^{\max}(u+n)), n = \min(j, \tilde{a}_2^{\max}(u))\}. \tag{33}$$

5.3. **No intra-slot space priority.** This situation is more intricate. When i class-1 and j class-2 packets arrive, these $i+j$ packets are assumed to have a completely random order. We have

Full Priority:

$$\begin{aligned}\tilde{\boldsymbol{A}}_u(m,n) = \sum_{i=m}^{\infty}\sum_{j=n}^{\infty} \boldsymbol{A}(i,j)\Bigg(&1\{i+j < (T-u)^+)\}1\{m=i, n=j\} \\ &+ 1\{i+j \geq (T-u)^+)\}\frac{\binom{(T-u)^+}{n}\binom{(i+j-(T-u)^+)^+}{j-n}}{\binom{i+j}{i}}1\{m = \min(i, N-u-n)\}\Bigg)\end{aligned} \tag{34}$$

Mixed Priority:

$$\tilde{\boldsymbol{A}}_u(m,n) = \sum_{i=m}^{\infty}\sum_{j=n}^{\infty} \boldsymbol{A}(i,j)\Bigg(1\{i+j<(T-u)^+)\}1\{m=i,n=j\}$$
$$+\,1\{i+j\geq (T-u)^+)\}\frac{\binom{(T-u)^+}{m}\binom{(i+j-(T-u)^+)^+}{i-m}}{\binom{i+j}{i}}1\{n=\min(j,N-u-m)\}\Bigg) \tag{35}$$

with $\binom{n}{k} = n!/(k!(n-k)!)$ denoting the binomial coefficient. In this case, a unified formula for both models (FP and MP) cannot be established as the class receiving space priority governs this equation. This can be seen as follows: when i class-1 and j class-2 packets arrive at a slot boundary, choosing i (out of $i+j$) positions for class-1 completely determines the arrival vector. The queue can accommodate $(T-u)^+$ packets until the threshold is reached and packets of the class without priority are no longer accepted. Consequently, in order to accept $m\,(n)$ of these packets, they have to be among the first $(T-u)^+$ arriving packets. The remaining $i-m\,(j-n)$ non-prioritized packets are lost, but all possible combinations among these vectors evidently have to be taken into consideration as well. Once the number of unprioritised effective arrivals is known, it is straightforward that prioritized packets are accepted as long as the queue is not entirely full.

6. **Numerical examples.** In this section, we investigate the impact of time priority, PBS and ISP on the performance measures of both classes in both the FP- and MP-model. Obviously, the impact of ISP increases as multiple packets arrive at the same slot boundary while it has no impact when only a single packet arrives. Therefore, a bursty arrival process where multiple packets arrive at the same slot boundary is considered in this section. Furthermore, ISP only has effect in slots where the threshold is crossed. If the threshold is not reached, all arriving packets are accepted, whereas, if the queue content already exceeds the threshold, only packets with space priority may enter the queue. Consequently, one would expect ISP to have a minor impact but in the following we demonstrate that ISP can considerably influence system performance. Furthermore, time priority and PBS have a large impact as expected.

The arrival process is as follows. Packets are generated by M on/off sources and given that a source is on (off) at a slot boundary, it remains on (off) at the following slot boundary with probability α (β). This is demonstrated in Fig. 2. Consequently, consecutive on-periods (off-periods) constitute a series of geometrically distributed random variables with mean $1/(1-\alpha)$ ($1/(1-\beta)$). When a source is on at a slot boundary, it generates b_1 class-1 packets and b_2 class-2 packets. A source does not generate packets when it is off at a slot boundary. The aggregated DBMAP of these sources is easily established. The arrival process at the buffer is completely characterized by the quintuple $(M, b_1, b_2, \alpha, \beta)$. However, it is equivalent and often more convenient to use the quintuple (M, b_1, b_2, σ, K), where

$$\sigma = \frac{1-\beta}{2-\alpha-\beta}\,, \quad K = \frac{1}{2-\alpha-\beta}\,. \tag{36}$$

The parameter σ denotes the fraction of time a source is on and K is a measure for the absolute lengths of the on- and off-periods. The parameter K takes values between $\max(\sigma, 1-\sigma)$ and ∞. For $K<1$, $K=1$ and $K>1$ the arrivals in

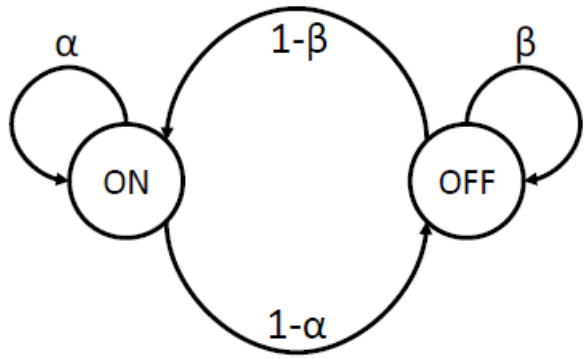

FIGURE 2. Source transition diagram.

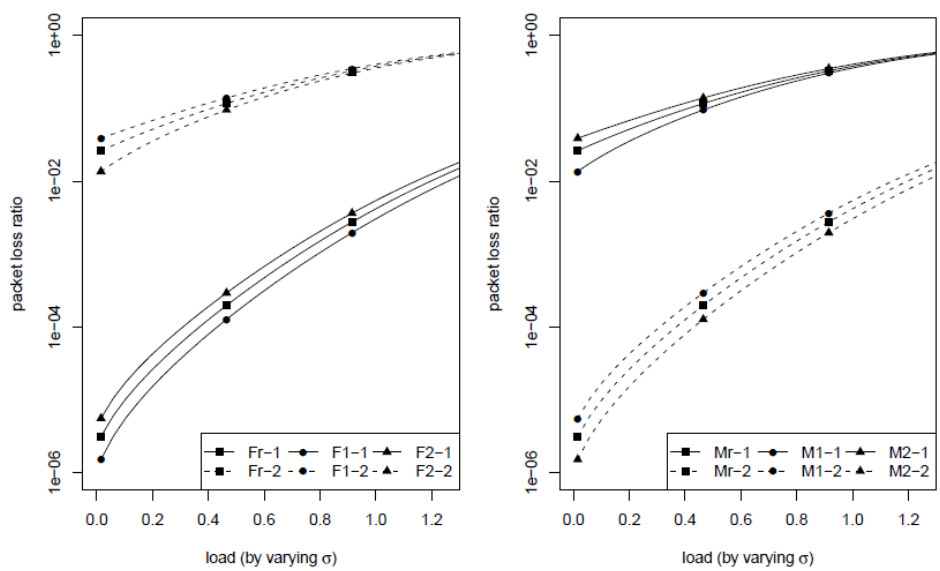

FIGURE 3. Loss vs. load with 3 ISP types for the FP-model (on the left) and the MP-model (on the right).

consecutive slots are negatively correlated, not correlated and positively correlated respectively. Consequently, the class-i arrival load is given by

$$\rho_i = M\sigma b_i \tag{37}$$

We now study the queueing system described in this paper when packets arrive according to the arrival process as described above. The legends use following 3 character notation. The first character denotes the model: F for FP and M for MP, the second denotes the ISP: 1 and 2 for class-1 and class-2 ISP respectively and r for random (no ISP). This is followed by a hyphen and the class number (1 or 2). For instance, when the load is depicted for Fr-2 it denotes the class-2 load for the FP-model with no ISP. Each figure has two graphs. The left one depicts the results

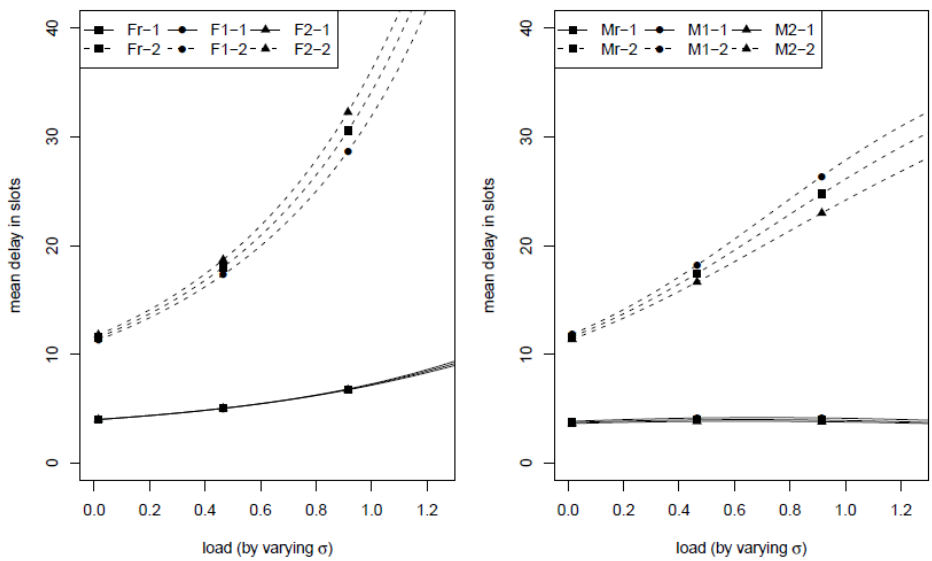

FIGURE 4. Delay vs. threshold with 3 ISP types for the FP-model (on the left) and the MP-model (on the right).

for the FP-model and the right one for the MP-model. Obviously, the results for no ISP will always lie between the values for class-1 and class-2-ISP. Furthermore, in order to make the graphs clearer, curves are full for class-1 and dashed for class-2 and each type of ISP has a symbol: a circle for class-1 ISP, a triangle for class-2 ISP and a square for no ISP.

First, consider a buffer that can hold $N = 50$ packets and has threshold $T = 25$. Packets are generated by $M = 2$ sources with $K = 1.5$ and when a source is on it generates 4 packets of each class. The fraction of time a source is on σ is varied causing the load ρ to vary from 0 to 1.3 (note that the system is finite and thus always stable). We investigate the impact hereof on the packet loss ratio in Fig. 3 and on the mean delay in Fig. 4.

We first study the packet loss ratio. Obviously, it increases when the load increases. The QoS differentiation provided by the model is immediately apparent. The loss is much lower for the class receiving space priority (class-1 on the left and class-2 on the right). Furthermore, the effect of ISP is easily observed as loss is up to three times higher (for light loads) between the different ISP's. For the class without space priority, all packets are discarded once the threshold T is exceeded and thus the ISP only plays a role in the slots where T is crossed. As the load increases the queue content surmounts the threshold more frequently and the packet loss becomes less dependent on the type of ISP and the three lines converge. Also note that, as time priority does not influence packet loss ratio, but only the order in which packets are served, the results are symmetric for the FP- and MP-model (swapping classes and ISP).

The mean delay of class 1 is lower than that of class 2 for both models as time priority is always provided to class 1. ISP affects mean class-2 delay considerably (5-20% difference) whereas class-1 packets are hardly influenced. This can be seen

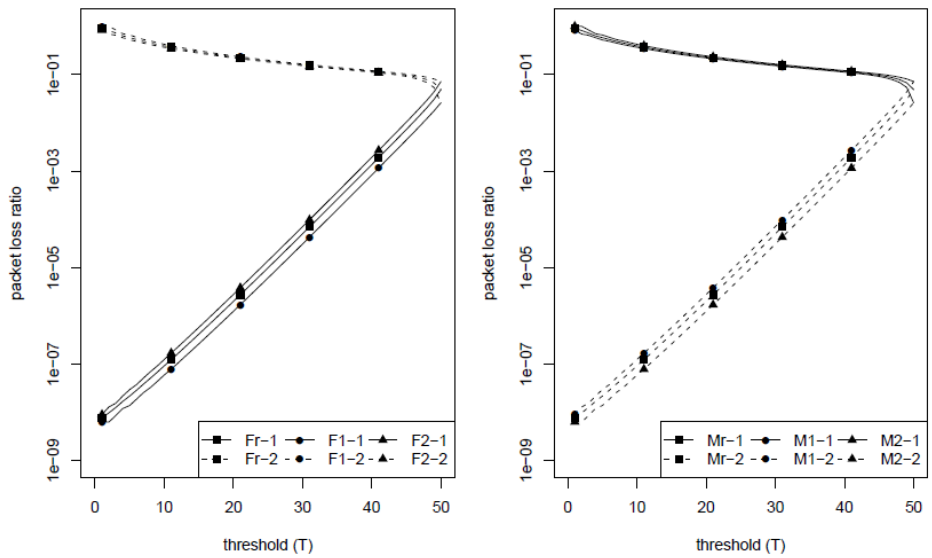

FIGURE 5. Packet loss ratio vs. threshold with 3 ISP types for the FP-model (on the left) and the MP-model (on the right).

by noting that class-1 packets are not affected by other packets arriving while they wait in the queue whereas class-2 packets have to give priority to any arriving class-1 packets and are consequently more reactive to packet drops and thus also to different ISP. For the FP-model, the mean class-1 delay increases with the load as more and more class-1 packets are allowed into the system. The mean class-2 delay increases as more and more packets enter the system. Note that the ISP resulting in the highest packet loss ratio also yields the lowest delay as more and more packets are dropped. In contrast, for the MP-model, the class-1 delay first increases slightly when the load increases and then starts decreasing as more and more class-1 packets are dropped as they do not have space priority and consequently the packets that do get accepted have shorter delay. Furthermore, both the mean class-1 and class-2 delay are lower than for the FP-model because, opposed to that model, the MP-model drops more class-1 packets than class-2 packets and the former have an impact on the delay of both classes whereas the latter only have an impact on the delay of other class-2 packets. This also explains why the ISP resulting in the highest class-2 packet loss ratio also yields the highest class-2 delay for the MP-model.

Next, we will investigate the effect of the threshold (T) as it controls how the available space (N) is distributed between both classes. Consider, $N = 50$, $M = 2$, $\sigma = 0.12$, $K = 1.5$ and $b_1 = b_2 = 2$ yielding a load $\rho = 0.96$. We let T vary from 0 to N and again depict the packet loss ratio (Fig. 5) and the mean delay (Fig. 6).

For $T = 0$, the system behaves as a system with only one traffic class (those with space priority). The differentiation in packet loss ratio between both classes decreases as T increases as more and more packets are allowed into the system (packets of both classes can utilize the spaces up to T). For $T = N$ there is no space priority and thus no difference between both classes. Furthermore, as explained for Fig. 3, ISP has only a limited effect on class-1 packet loss for high load (recall that

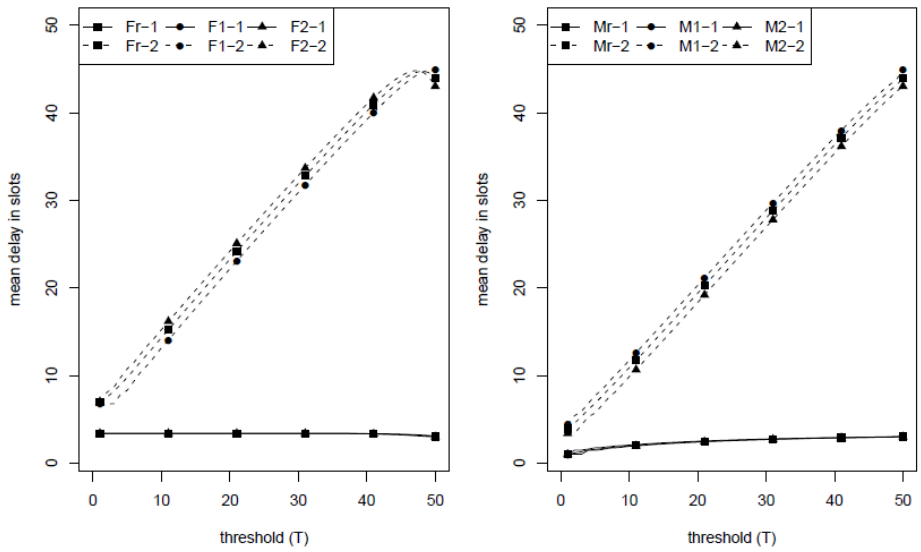

FIGURE 6. Delay vs. threshold with 3 ISP types for the FP-model (on the left) and the MP-model (on the right).

$\rho = 0.96$) whereas its impact on class-2 is bigger. Obviously, the ISP equivalent to the class receiving space priority corresponds to the smallest amount of packet loss. Again, it is apparent that both models are symmetric concerning packet loss.

The mean class-1 delay is hardly affected by varying the threshold for the FP-model and for larger N it even decreases slightly as the system even starts to drop space prioritized (class-1) packets resulting in a shorter delay for packets of this class that are accepted. This also explains the decrease in class-2 delay when T approaches N. Furthermore, class-2 delay increases as the threshold increases as more and more class-2 packets are allowed into the system causing a longer delay for other packets of this class (recall that they do not affect the delay of class-1 packets). For the MP-model, when the threshold increases, more class-1 packets are allowed into the system at the cost of class-2 packets. But, as stated before, class-1 packets affect the mean delay of both classes which thus get longer as T increases. Concerning ISP, similar arguments as above lead to the same conclusions. It is clear that choosing the threshold T appropriately (with respect to the required QoS) is of paramount importance

7. **Conclusions.** This paper studies a finite-sized discrete-time two-class priority queue where packets arrive according to a two-class discrete batch Markovian arrival process (2-DBMAP). Time and space priority are incorporated in the queueing model to provide different types of service to each class. One of both classes receives absolute time priority in order to minimize its delay. Space priority is implemented by the partial buffer sharing acceptance policy and can be provided to the class receiving time priority or to the other class. This choice gives rise to two different queueing models (Full and Mixed Priority) and this paper analyses both these models in a unified manner. Furthermore, the buffer finiteness and the use of space priority make it interesting to consider a general order of arrivals at a slot

boundary. This paper introduces a string representation for sequences of arriving packets. This naturally gives rise to intra-slot space priority (ISP) governing space priority between the packets arriving at a slot boundary. Performance of these queueing systems is then determined using matrix-analytic techniques. One can conclude that the range of service differentiation covered by these models is large and that ISP has a major impact for certain parameter settings and can thus not be neglected for bursty arrival processes. Determining an appropriate value for the threshold (space priority) is of paramount importance as it not only affects packet loss but also the queue content (and thus delay/time priority performance) of packets of both classes, especially for Mixed Priority.

REFERENCES

[1] C. Blondia and O. Casals, *Statistical multiplexing of VBR sources: A matrix-analytic approach,* Performance Evaluation, **16** (1992), 5–20.

[2] T. Demoor, J. Walraevens, D. Fiems, S. De Vuyst and H. Bruneel, *Influence of real-time queue capacity on system contents in DiffServ's expedited forwarding per-hop-behavior,* Journal of Industrial and Management Optimization, **6** (2010), 587–602.

[3] T. Demoor, D. Fiems, J. Walraevens and H. Bruneel, *Time and space priority in a partially shared priority queue,* in "Proceedings of the 5th International Conference on Queueing Theory and Network Applications," ACM, (2010), pp. 120.

[4] D. Fiems and H. Bruneel, *A note on the discretization of Little's result,* Operations Research Letters, **30** (2002), 17–18.

[5] D. Fiems, J. Walraevens and H. Bruneel, *Performance of a partially shared priority buffer with correlated arrivals,* Lecture Notes in Computer Science, **4516** (2007), 582–593.

[6] G. Hwang and B. Choi, *Performance analysis of the $DAR(1)/D/c$ priority queue under partial buffer sharing policy,* Computers & Operations Research, **31** (2004), 2231–2247.

[7] H. Kröner, G. Hébuterne, P. Boyer and A. Gravey, *Priority management in ATM switching nodes,* IEEE Journal on Selected Areas in Communications, **9** (1991), 418–427.

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

[9] G. Latouche and V. Ramaswami, "Introduction to Matrix Analytic Methods in Stochastic Modeling," Series on Statistics and Applied Probability. ASA-SIAM, Philadelphia, PA, American Statistical Association, Alexandria, VA, 1999.

[10] M. Mehmet Ali and X. Song, *A performance analysis of a discrete-time priority queueing system with correlated arrivals,* Performance Evaluation, **57** (2004), 307–339.

[11] H. Radha, Y.Chen, K. Parthasarathy and R. Cohen, *Scalable internet video using MPEG-4,* Signal Processing: Image Communication, **15** (1999), 95–126.

[12] K. Spaey, "Superposition of Markovian Traffic Sources and Frame Aware Buffer Acceptance," Ph.D. thesis, Universitaire Instelling Antwerpen, 2002.

[13] T. Takine, B. Sengupta and T. Hasegawa, *An analysis of a discrete-time queue for broadband ISDN with priorities among traffic classes,* IEEE Transactions on Communications, **42** (1994), 1837–1845.

[14] J. Van Velthoven, B. Van Houdt and C. Blondia, *The impact of buffer finiteness on the loss rate in a priority queueing system,* Lecture Notes in Computer Science, **4054** (2006), 211–225.

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

[16] J. Walraevens, S. Wittevrongel and H. Bruneel, *A discrete-time priority queue with train arrivals,* Stochastic Models, **23** (2007), 489–512.

[17] Y. Wang, C. Liu and C. Lu, *Loss behavior in space priority queue with batch Markovian arrival process - discrete-time case,* Performance Evaluation, **41** (2000), 269–293.

[18] Y. C. Wang, J. S. Wang and F. H. Tsai, *Analysis of Discrete-Time Space Priority Queue with Fuzzy Threshold,* Journal of Industrial and Management Optimization, **5** (2009), 467–479.

[19] J. Zhao, B. Li, X. Cao and I. Ahmad, *A matrix-analytic solution for the $DBMAP/PH/1$ priority queue,* Queueing Systems, **53** (2006), 127–145.

Received September 2010; revised January 2011.

E-mail address: thdemoor@telin.ugent.be
E-mail address: df@telin.ugent.be
E-mail address: jw@telin.ugent.be
E-mail address: hb@telin.ugent.be