DELAY CHARACTERISTICS IN PLACE-RESERVATION QUEUES WITH CLASS-DEPENDENT SERVICE TIMES

SABINE WITTEVRONGEL*, BART Feyaerts and HERWIG BRUNEEL
SMACS Research Group
Department TELIN
Ghent University
Sint-Pietersnieuwstraat 41, 9000 Gent, Belgium

STIJN DE VUYST
Department of Industrial Systems Engineering and Product Design
Ghent University
Technologiepark 903, 9052 Zwijnaarde, Belgium

Abstract. This paper considers a discrete-time single-server infinite-capacity queue with two classes of packet arrivals, either delay-sensitive (class 1) or delay-tolerant (class 2), and a reservation-based priority scheduling mechanism. The objective is to provide a better quality of service to delay-sensitive packets at the cost of allowing higher delays for the best-effort packets. To this end, the scheduling mechanism makes use of an in-queue reserved place intended for future class-1 packet arrivals. A class-1 arrival takes the place of the reservation in the queue, after which a new reservation is created at the tail of the queue. Class-2 arrivals always take place at the tail of the queue. We study the delay characteristics for both packet classes under the assumption of a general independent packet arrival process. The service times of the packets are independent and have a general distribution that depends on the class of the packet. Closed-form expressions are obtained for the probability generating functions of the per-class delays. From this, moments and tail probabilities of the packet delays of both classes are derived. The results are illustrated by some numerical examples.

1. Introduction. Modern heterogeneous packet-based communication networks support a wide spectrum of applications, each bearing their own set of quality-of-service (QoS) requirements. In particular, applications can roughly be divided into two classes depending on their delay sensitivity. In what follows, we refer to packets from delay-sensitive applications (e.g. real-time applications such as telephony, video or media streaming) as class-1 or high-priority packets; their mean delay and delay jitter should be kept minimal. Packets originating from delay-tolerant (best-effort) applications (e.g. email or file transfer) are referred to as class-2 or low-priority packets; their delay requirements are typically less stringent.

Many approaches have been proposed to provide delay differentiation between class-1 and class-2 packets, all varying in complexity and effect. A common method

2010 Mathematics Subject Classification. Primary: 60K25, 90B22; Secondary: 68M20.
Key words and phrases. Discrete-time queue, priority scheduling, place reservation, delay analysis, class-dependent service times.
The reviewing process of this paper was handled by Yutaka Takahashi and Wuyi Yue.
* Corresponding author: Sabine Wittevrongel.
is the use of Absolute Priority (AP) scheduling, where transmission priority is always given to class-1 packets and class-2 packets can only enter the server in the absence of class-1 packets. It has been shown (see e.g. [11, 14]) that AP indeed reduces the class-1 packet delay, but may lead to excessively long class-2 packet delays, a phenomenon also referred to as low-priority packet starvation. To overcome this drawback of AP, several less drastic solutions for priority scheduling have been introduced as well. For instance, in [13], a probabilistic priority mechanism is presented that uses a small probability by which the server may service a class-2 packet, even though class-1 packets are available. Another idea is to limit the priority given to class-1 packets to a given window. Examples include slot-bound priority scheduling [4], where class-1 packets are only given priority over class-2 packets arriving during the same time slot, and time-limited overtake priority scheduling [5], where class-1 packets cannot overtake class-2 packets if the latter arrived at least \( N \) time slots earlier than the former. Packets can also be allowed to promote to a higher-priority class under certain well-specified conditions, or conversely, to degrade to a lower class; this is referred to with the term priority jumps, see e.g. [9, 10] for a comparison of several variants. Finally, on the other extreme of the spectrum, we find the classical first-come-first-served (FCFS) scheduling, where all packets are treated equally and no delay differentiation is obtained.

In this paper, we focus on a reservation-based priority scheduling mechanism, which makes use of an in-queue reserved position for future class-1 packets. As a class-1 packet arrives to the queue, it takes the place of the reservation in the queue, after which a new reservation is created at the tail of the queue. Class-2 arrivals always take place at the tail of the queue in the usual FCFS way. If multiple packets arrive during the same time slot, first the class-1 packets are inserted one by one and then the class-2 packets are inserted, as in the case of slot-bound priority. In addition, it is clear that the first class-1 packet of each slot can jump over some class-2 packets. The reservation-based scheduling will therefore lead to a delay reduction for class-1 packets as compared to FCFS. Since any class-2 packet can, however, be jumped over only once, there is also a limit to the disadvantage experienced by class-2 packets, thus alleviating class-2 packet starvation as compared to AP.

Past work on place reservation queues assumed independent and identically distributed (\( iid \)) transmission (or service) times for both packet classes, either deterministically equal to one slot [6], geometrically distributed [8] or with a general distribution [7]. In contrast, we now consider heterogeneous service requirements with class-dependent service-time distributions and we further complete the analysis of [15]. The key element in the analysis method for class-dependent service times is the use of a new Markovian system state vector consisting of the total amount of work in the queue in front of the reservation and the number of class-2 packets in the queue behind the reservation, at the beginning of a slot. Based on this, closed-form expressions can be obtained for the probability generating functions (pgfs), the moments and the tail probabilities of the packet delays of both classes.

The rest of this paper is structured as follows. The queueing model under study is described in Section 2. In Section 3, we identify an adequate Markovian system state vector that will enable the delay analysis under class-dependent service times and we construct the system equations that govern the time evolution of the system state vector. In Section 4, we derive an expression for the joint pgf of the system state vector in stochastic equilibrium. Based on this result, we then analyze the
class-1 and class-2 packet delays in Section 5. Several numerical examples are presented in Section 6 and conclusions are given in Section 7.

2. Queueing model description. We consider a discrete-time queueing system with a single server and an infinite storage capacity. Time is divided into fixed-length time slots. The system is a late arrival system with delayed access (LAS-DA) [12], such that packets arrive by the end of a slot and their service can start no earlier than at the beginning of the next slot. Packets arrive to the queue according to a two-class arrival process, described by the joint pgf

$$A(z_1, z_2) = \mathbb{E}[z_1^{a_{1,k}} z_2^{a_{2,k}}],$$

with $a_{j,k}$ being the number of packet arrivals of class $j$ during slot $k$. Note that the numbers of arrivals are assumed to be iid from slot to slot, but $a_{1,k}$ and $a_{2,k}$ can be dependent during any slot $k$. For convenience we introduce $A_1(z) = A(z, 1)$, $\lambda_1 = A_1'(1)$ and similarly we define $A_2(z) = A(1, z)$, $\lambda_2 = A_2'(1)$.

The queueing system operates under a reservation-based priority scheduling discipline. First, this means that packets arriving during the same time slot are inserted in the queue in order of their priority. More specifically, before actual insertion, all the $a_{1,k} + a_{2,k}$ arrivals during any slot $k$ are reordered such that all $a_{1,k}$ class-1 packets are inserted (one by one) in the queue before the $a_{2,k}$ class-2 arrivals are inserted. This is referred to as slot-bound priority in e.g. [4]. Secondly, the queue contains one reserved place or reservation $R$, which serves as a placeholder for future class-1 packet arrivals. Whenever a class-1 packet is inserted in the queue, it replaces the reservation, after which a new reservation is immediately inserted at the queue’s tail, such that the queue always contains exactly one reservation. Class-2 packets are always inserted at the tail of the queue. This way, the first class-1 packet of each slot can jump over some class-2 packets, while any class-2 packet can be outpaced by at most one class-1 packet. The packet insertion in the queue under reservation-based scheduling is further illustrated by an example in Figure 1. Specifically, the figure shows the system content just before and after the insertion of 4 packets arriving during a given slot.

![Figure 1](image-url)

Figure 1. Insertion of 4 packets arriving during the same slot under reservation-based scheduling

Once packets are actually inserted, they move through the queue in the traditional FCFS way. If the reservation reaches the server, it remains in the queue and is jumped over by the first class-2 packet, if any.

The distribution of the service times is class-dependent, i.e., class-1 service times have pgf $S_1(z)$ with mean $\mu_1$ and class-2 service times have pgf $S_2(z)$ with mean
\( \mu_2 \). All service times are independent from packet to packet. The system load can then be found as

\[ \rho = \lambda_1 \mu_1 + \lambda_2 \mu_2. \tag{2} \]

As another illustration, Figure 2 shows a sample path of the queueing model. Specifically, above the time axis in Figure 2, the content of the queue at the beginning of 12 consecutive slots is shown for a particular sequence of packet arrivals, constant class-1 service times of 2 slots and constant class-2 service times of 1 slot. Below the time axis, arrows indicate the service duration and the departure instant for each of the packets in the server. The variables \( v_k \) and \( w_k \) in the figure are Markovian system state variables and are introduced in the next section.

![Figure 2. Sample path of the queueing model](image)

3. Markovian state description and system equations. As Figures 1 and 2 illustrate, due to the reservation-based scheduling discipline, there can never be class-1 packets positioned behind the reservation. In contrast, the converse does not apply, i.e., in front of the reservation, packets can be of either class. Moreover, once a packet is in front of the reservation, its remaining delay is only dependent on the service process. A (class-2) packet positioned behind the reservation can, however, be jumped over by (at most) one class-1 arrival, as long as the service of that class-2 packet has not yet begun (which is the case for a packet behind the reservation).

These observations suggest the following system state variables at the beginning of slot \( k \):

- \( w_k \): the total amount of work (expressed in slots) in front of the reservation (including the server) at the beginning of slot \( k \). This is the sum of the service times of all packets in the queue in front of the reservation at the beginning of slot \( k \) and the remaining service time of the packet in service – if any – again at the beginning of slot \( k \);
- \( v_k \): the number of class-2 packets behind the reservation at the beginning of slot \( k \).

For illustrative purposes, the values of these system state variables at the beginning of the consecutive slots are indicated in Figure 2. Note that due to the work
conserving nature of the system, if \( w_k = 0 \) then \( v_k = 0 \) as well, implying that the system is empty at the beginning of slot \( k \).

In the rest of this section, we derive a set of system equations that govern the evolution of the system state variables from slot to slot. For this purpose, we first look at a random slot \( k \) that does not feature any class-1 arrival (\( a_{1,k} = 0 \)), implying that the reservation does not get seized during slot \( k \). Several possibilities regarding the system state and arrival process have then to be taken into account. Firstly, consider the case where the system is empty or where there is only one slot’s worth of work in front of the reservation at the beginning of slot \( k \) (i.e., \( w_k \leq 1 \)). Then, if \( a_{1,k} = 0 \) and \( w_k \leq 1 \), no more work will be in front of the reservation \( R \) by the end of slot \( k \). This yields an empty system at the beginning of slot \( k + 1 \) (as e.g. for \( k = 4 \) or \( k = 5 \) in Figure 2), unless there were still some class-2 packets behind the reservation (\( v_k > 0 \)) (as e.g. for \( k = 3 \) in Figure 2) or unless there were class-2 arrivals during slot \( k \) (\( a_{2,k} > 0 \)) (as e.g. for \( k = 6 \) in Figure 2). As can be seen from Figure 2 (for slots \( k = 3 \) or \( k = 6 \)), in the latter cases where \( v_k > 0 \) and/or \( a_{2,k} > 0 \), a class-2 packet jumps over the reservation \( R \) and starts its service at the beginning of slot \( k + 1 \), such that the work \( w_{k+1} \) in front of \( R \) corresponds to the full service time of this class-2 packet; the other \( a_{2,k} + v_k - 1 \) class-2 packets in the system remain positioned behind \( R \) and, hence, correspond to \( v_{k+1} \). Therefore, for \( a_{1,k} = 0 \), \( w_k \leq 1 \) and \( a_{2,k} + v_k = 0 \), we have

\[
\begin{align*}
v_{k+1} &= 0, \quad (3) \\
w_{k+1} &= 0, \quad (4)
\end{align*}
\]

whereas for \( a_{1,k} = 0 \), \( w_k \leq 1 \) and \( a_{2,k} + v_k > 0 \), we find

\[
\begin{align*}
v_{k+1} &= a_{2,k} + v_k - 1, \quad (5) \\
w_{k+1} &= s_2, \quad (6)
\end{align*}
\]

where \( s_2 \) denotes the service time of the class-2 packet that is taken into service at the beginning of slot \( k + 1 \).

Secondly, still for \( a_{1,k} = 0 \), if there is more than one slot’s worth of work in front of the reservation at the beginning of slot \( k \) (i.e., \( w_k > 1 \)) (as e.g. for \( k - 2 \) or \( k = 9 \) in Figure 2), all \( a_{2,k} \) class-2 arrivals are simply appended to the queue, such that the number of class-2 packets behind \( R \) increases by \( a_{2,k} \), whereas the work in front of \( R \) decreases by one. For \( a_{1,k} = 0 \) and \( w_k > 1 \), we thus find

\[
\begin{align*}
v_{k+1} &= a_{2,k} + v_k, \quad (7) \\
w_{k+1} &= w_k - 1. \quad (8)
\end{align*}
\]

In case there is at least one class-1 arrival during slot \( k \) (\( a_{1,k} > 0 \)) (as is the case for \( k = 1, k = 7, k = 8 \) or \( k = 10 \) in Figure 2), then at the beginning of slot \( k + 1 \), the reservation will be near the queue’s tail, with only the \( a_{2,k} \) new class-2 arrivals behind it. Thus, the \( v_k \) class-2 packets that were behind the reservation at the beginning of slot \( k \), will now be in front of \( R \), and the same goes for the \( a_{1,k} \) new class-1 arrivals: the service times of all these \( v_k + a_{1,k} \) packets are added to the work in front of \( R \). During slot \( k \) the server can reduce the remaining work (if any) in front of the reservation at the beginning of slot \( k \). Since at least one class-1 packet arrives during slot \( k \), we know that no additional jump of a class-2 packet over the reservation will take place. These observations yield the following
set of system equations for \( a_{1,k} > 0 \):

\[
v_{k+1} = a_{2,k}, \quad (9)
\]

\[
w_{k+1} = (w_k - 1)^+ + \sum_{i=1}^{a_{1,k}} s_{1,i} + \sum_{i=1}^{v_k} s_{2,i}, \quad (10)
\]

where \((\cdot)^+ = \max(\cdot, 0)\) and the \( s_{j,i} \)s correspond to individual service times of class-\( j \) packets.

The system equations (3)–(10) clearly show that the set of vectors \( \{v_k, w_k\} \) indeed constitutes a two-dimensional first-order discrete-time Markov chain. In what follows, we derive the equilibrium distribution of the Markovian system state, as an intermediate step towards the delay analysis of Section 5.

4. Equilibrium analysis of the system state. We define the joint pgf \( P_k(x, y) \) of the system state variables \( v_k \) and \( w_k \) at the beginning of slot \( k \) as

\[
P_k(x, y) \triangleq \mathbb{E}[x^{v_k} y^{w_k}]. \quad (11)
\]

By means of the system equations (3)–(10), the system state pgf \( P_{k+1}(x, y) \) at the beginning of slot \( k + 1 \) can then be expressed in terms of its slot \( k \) counterpart. Specifically, we start from the definition of \( P_{k+1}(x, y) \), use the law of total expectation to split up the expectation into four terms corresponding to the four cases distinguished above in (3)–(10), and introduce the appropriate system equations, as follows:

\[
P_{k+1}(x, y) = \mathbb{E}[x^{v_{k+1}} y^{w_{k+1}}]
\]

\[
= \mathbb{E}[x^{v_{k+1}} y^{w_{k+1}} \mid a_{1,k} = 0, w_k \leq 1, a_{2,k} + v_k = 0] + \mathbb{E}[x^{v_{k+1}} y^{w_{k+1}} \mid a_{1,k} = 0, w_k \leq 1, a_{2,k} + v_k > 0] + \mathbb{E}[x^{v_{k+1}} y^{w_{k+1}} \mid a_{1,k} = 0, w_k > 1] + \mathbb{E}[x^{v_{k+1}} y^{w_{k+1}} \mid a_{1,k} > 0] \tag{12}
\]

where \( \mathbb{E}[X \mid \{Y\}] \triangleq \mathbb{P}[Y \mid \mathbb{E}[X \mid Y]] \) and where in the last term we also used the independence of the service times from packet to packet. The four terms in the above expression for \( P_{k+1}(x, y) \) can now be determined separately.

Note first that due to the independent nature of the packet arrival process from slot to slot, the random variables \( a_{1,k} \) and \( a_{2,k} \) are independent of the system state variables \( v_k \) and \( w_k \) at the beginning of slot \( k \). The first term in (12) is then immediately obtained as

\[
\mathbb{P}[a_{1,k} = 0, w_k \leq 1, a_{2,k} + v_k = 0] = \mathbb{P}[a_{1,k} = 0, a_{2,k} = 0] \mathbb{P}[v_k = 0, w_k \leq 1] = A(0, 0) \{\mathbb{P}[v_k = 0, w_k = 1] + \mathbb{P}[v_k = 0, w_k = 0]\} = A(0, 0) (R_k(0) + p_{0,k}), \tag{13}
\]
where we introduced the probability \( p_{0,k} \triangleq \text{Prob}[v_k = 0, w_k = 0] \) of an empty system and the partial pgf

\[
R_k(x) \triangleq \mathbb{E}[x^{v_k} \{ w_k = 1 \}],
\]

such that \( \text{Prob}[v_k = 0, w_k = 1] = R_k(0) \). Note also that \( P_k(x,0) = p_{0,k}, \forall x \). Similarly, with the law of total expectation and the property that \( w_k = 0 \) implies \( v_k = 0 \), the second term in equation (12) can be further rewritten as

\[
\begin{align*}
&= \frac{1}{x} S_2(y) \left\{ \mathbb{E}[x^{a_{1,k} + v_k} \{ a_{1,k} = 0, w_k = 0, a_{2,k} + v_k > 0 \}] + \mathbb{E}[x^{a_{2,k} + v_k} \{ a_{1,k} = 0, w_k = 1 \}] - \mathbb{E}[x^{a_{2,k} + v_k} \{ a_{1,k} = 0, w_k = 1, a_{2,k} + v_k = 0 \}] \right\} \\
&= \frac{1}{x} S_2(y) \left\{ \mathbb{E}[x^{a_{1,k}} | a_{1,k} = 0, w_k = 0] \mathbb{E}[x^{v_k} \{ w_k = 1 \}] - \mathbb{E}[x^{a_{2,k}} | a_{1,k} = 0, w_k = 0] \mathbb{E}[x^{v_k} \{ w_k = 1 \}] \right\}
\end{align*}
\]

Again using the property that the variables \( a_{1,k} \) and \( a_{2,k} \) are independent of the system state variables \( v_k \) and \( w_k \), together with the law of total expectation, we get the third term in (12) as

\[
\begin{align*}
&= \frac{1}{y} A(0,x) \left\{ E[x^{v_k} y^{w_k} | w_k = 1] - E[x^{v_k} y^{w_k} \{ w_k = 0 \}] \right\}
\end{align*}
\]

Analogously, the fourth term in (12) is determined as

\[
\begin{align*}
&= \frac{1}{y} A(0,x) \left\{ A(S_1(y), x) - A(0,x) \right\} [P_k(S_2(y), y) + (y - 1)p_{0,k}].
\end{align*}
\]

Finally, substitution of the expressions (13)–(17) into (12) together with some rearrangements of the terms on the right-hand side, leads to the following relationship between the system state pgfs at the beginning of the slots \( k \) and \( k + 1 \):

\[
P_{k+1}(x, y) = \frac{1}{x} (x - S_2(y)) \left\{ A(0,0) \left( R_k(0) + p_{0,k} \right) - A(0,x) \left( R_k(x) + p_{0,k} \right) \right\}
\]
We now assume that a steady state exists for the queueing system. This assumption holds if the system is stable, i.e., if $\rho < 1$. In steady state (for $k \to \infty$) the pgfs $P_k(x,y)$ and $P_{k+1}(x,y)$ both converge to the same limiting function $P(x,y)$. Therefore, by removing the time indices $k$ and $k+1$ from (18), we obtain the following linear equation for the steady-state pgf $P(x,y)$ of the system state:

\[
P(x,y) = \frac{1}{x} (x - S_2(y)) \left[ A(0,0) (R(0) + p_0) - A(0,x) (R(x) + p_0) \right] \\
+ \frac{1}{y} [(y-1) A(S_1(y),x)p_0 + A(0,x)P(x,y)] \\
+ \frac{1}{y} [A(S_1(y),x) - A(0,x)] P(S_2(y),y),
\]

where we also introduced $p_0$ and $R(x)$ as the steady-state counterparts of $p_{0,k}$ and $R_k(x)$. From (19), the pgf $P(x,y)$ is then found as

\[
P(x,y) = \frac{1}{y - A(0,x)} \left\{ (y-1) A(S_1(y),x)p_0 \\
+ \frac{y}{x} (x - S_2(y)) \left[ A(0,0) (R(0) + p_0) - A(0,x) (R(x) + p_0) \right] \\
+ [A(S_1(y),x) - A(0,x)] P(S_2(y),y) \right\}.
\]

Expression (20) still contains a number of unknowns: the unknown probabilities $R(0)$ and $p_0$ and the unknown functions $R(x)$ and $P(S_2(y),y)$. We will now tackle these unknowns one by one.

First, substituting $x = S_2(y)$ in (20), we find

\[
P(S_2(y),y) = \frac{(y-1) A(S_1(y),S_2(y))}{y - A(S_1(y),S_2(y))} p_0.
\]

Note that $P(S_2(y),y)$ corresponds to the pgf of the unfinished work in the system at the beginning of an arbitrary slot, which could also be found from e.g. [2] as the pgf of the system content in a discrete-time single-server infinite-capacity queueing system with constant service times of one slot and a general independent arrival process with pgf of the number of arrivals per slot given by $A(S_1(y),S_2(y))$. With (21), $P(x,y)$ can be rewritten as

\[
P(x,y) = \frac{1}{y - A(0,x)} \left\{ (y-1) \frac{y A(S_1(y),x) - A(0,x)A(S_1(y),S_2(y))}{y - A(S_1(y),S_2(y))} p_0 \\
+ \frac{y}{x} (x - S_2(y)) \left[ A(0,0) (R(0) + p_0) - A(0,x) (R(x) + p_0) \right] \right\}.
\]

Secondly, we make use of the fact that all pgfs are bounded when their arguments are on the closed unit disk. Specifically, substitution of $y = y(x) \triangleq A(0,x)$ in (22) would cause the denominator to become 0, presumably causing a singularity. However, since it can be shown that $|A(0,x)| \leq 1$ for $|x| \leq 1$, there must not be a singularity for $y = y(x)$ with $|x| \leq 1$. The only way such singularity can be avoided is by expressing that the corresponding numerator must become 0 as well. Solving
the resulting linear equation for $R(x)$, we get
\[ R(x) = \frac{A(0,0)}{y(x)} (R(0) + p_0) - p_0 \]
\[ + \frac{x(y(x) - 1)}{y(x)(x - S_2(y(x)))} \frac{A(S_1(y(x)), x) - A(S_1(y(x)), S_2(y(x)))}{y(x) - A(S_1(y(x)), S_2(y(x)))} p_0. \]

In order to determine the unknown $R(0)$, we take a look at the empty system probability $p_0$. Note that only two scenarios can lead to an empty system at the beginning of a random steady-state slot: either the system was already empty at the beginning of the previous slot (which occurs with probability $p_0$), or there was exactly one packet in the system at the beginning of the previous slot, with a remaining service time of one slot. Note that in that case, the single packet in the system will have been in front of the reservation at the beginning of that previous slot, a situation which occurs with probability $\text{Prob}[v = 0, w = 1] = R(0)$. In both scenarios, there must not have been any arrivals during that previous slot, such that we have that
\[ p_0 = A(0,0) (p_0 + R(0)). \]

This then allows us to rewrite (23) as
\[ R(x) = -p_0 + \frac{1}{y(x)} p_0 \left\{ 1 + x \frac{y(x) - 1}{x - S_2(y(x))} \frac{A(S_1(y(x)), x) - A(S_1(y(x)), S_2(y(x)))}{y(x) - A(S_1(y(x)), S_2(y(x)))} \right\}. \]

Substitution of (24) and (25) into (22) yields
\[ P(x, y) = \frac{p_0}{y - A(0, x)} \left\{ Q(x, y) - \frac{y}{A(0, x)} \frac{x - S_2(y)}{x - S_2(A(0, x))} Q(x, A(0, x)) \right\}, \]
where we introduced the shorthand function
\[ Q(x, y) \triangleq (y - 1) \frac{yA(S_1(y), x) - A(0, x)A(S_1(y), S_2(y))}{y - A(S_1(y), S_2(y))}. \]

Finally, we still need to determine the empty system probability $p_0$. We do so by expressing the normalization condition $P(1, 1) = 1$. In particular, from (26) and with $A(0, 1) = A_1(0)$, we have
\[ P(1, 1) = P(1, y)|_{y=1} \]
\[ = \frac{p_0}{y - A_1(0)} \left\{ Q(1, y) - \frac{y}{A_1(0)} \frac{1 - S_2(y)}{1 - S_2(A_1(0))} Q(1, A_1(0)) \right\}|_{y=1} \]
\[ = \frac{Q(1, 1)}{1 - A_1(0)} p_0, \]
where $Q(1, 1)$ can be found from (27) by means of de l'Hôpital’s rule as
\[ Q(1, 1) = \lim_{y \to 1} (y - 1) \frac{yA_1(S_1(y)) - A_1(0)A(S_1(y), S_2(y))}{y - A(S_1(y), S_2(y))} \]
\[ = \frac{1 - A_1(0)}{1 - \lambda_1 \mu_1 - \lambda_2 \mu_2} \]
\[ = \frac{p_0}{1 - \lambda_1 \mu_1 - \lambda_2 \mu_2}, \]
such that
\[ P(1, 1) = \frac{p_0}{1 - \lambda_1 \mu_1 - \lambda_2 \mu_2}. \]
The normalization condition then leads to
\[ p_0 = 1 - \lambda_1 \mu_1 - \lambda_2 \mu_2. \]  
(31)

Note that we obtain \( p_0 = 1 - \rho \), as expected from Little’s law. Note also that by breaking up the calculation of \( P(1,1) \), we have at the same time obtained a closed-form expression for \( P(1,y) \), which will be useful in the remainder of this paper.

5. Analysis of the packet delay. The packet delay analysis is performed for the two packet classes separately. In either case, we select a random steady-state packet \( P_j \) of class \( j \) \((j = 1, 2)\) and investigate its delay \( d_j \). Although the arrival slot \( \mathcal{S}_j \) of \( P_j \) is not a random slot (due to the fact that it is by definition a slot with one or more class-\( j \) arrivals), in view of the independent nature of the packet arrival process from slot to slot, the system state at the beginning of a slot is independent of the arrival process during that slot, and hence the system state distribution at the beginning of slot \( \mathcal{S}_j \) is identical to the system state distribution at the beginning of a random steady-state slot, i.e.,

\[ E[x^{v_{\mathcal{S}_j}}y^{w_{\mathcal{S}_j}}] = P(x,y). \]  
(32)

Since \( P_j \) is an arbitrarily chosen class-\( j \) packet, the distribution of the numbers of arrivals of either class during slot \( \mathcal{S}_j \) can be obtained as

\[ \text{Prob}[a_{1,\mathcal{S}_j} = \alpha_1, a_{2,\mathcal{S}_j} = \alpha_2] = \frac{\alpha_j}{\lambda_j} \text{Prob}[a_1 = \alpha_1, a_2 = \alpha_2], \]  
(33)

proportional to both \( \alpha_j \) and the distribution of the numbers of arrivals during a random slot.

5.1. Pgf of the class-1 delay. The position at which \( P_1 \) is inserted in the queue highly depends on the order of all \( a_{1,\mathcal{S}_1} \) class-1 arrivals during slot \( \mathcal{S}_1 \). If \( P_1 \) is the first of the \( a_{1,\mathcal{S}_1} \) class-1 packets to be inserted, it will replace the reservation at its position at the beginning of slot \( \mathcal{S}_1 \). If \( P_1 \) is not the first one to be inserted, it will be positioned somewhere behind the \( v_{\mathcal{S}_1} \) class-2 packets that were behind the reservation at the beginning of slot \( \mathcal{S}_1 \).

Let us therefore introduce \( f_{P_1} \) as the number of class-1 packets that have arrived during slot \( \mathcal{S}_1 \) and that are inserted in the queue before \( P_1 \) is. Given that the order of arrivals during the same slot is completely random, \( f_{P_1} \) will be uniformly distributed between 0 and \( a_{1,\mathcal{S}_1} - 1 \), such that with (33) the pgf \( F_{P_1}(z) \) of \( f_{P_1} \) can be found as

\[ F_{P_1}(z) = E[z^{f_{P_1}}] = E \left[ \sum_{i=0}^{a_{1,\mathcal{S}_1}-1} \frac{1}{a_{1,\mathcal{S}_1}} z^i \right] = E \left[ \frac{1 - z^{a_{1,\mathcal{S}_1}}}{a_{1,\mathcal{S}_1}(1-z)} \right] = \frac{A_1(z) - 1}{\lambda_1(z-1)}, \]  
(34)

such that

\[ F_{P_1}(0) = \frac{1 - A_1(0)}{\lambda_1}, \quad \text{and} \quad F_{P_1}'(1) = \frac{\lambda_1'}{2A_1}, \]  
(35)

with \( \lambda_j' \equiv A_j''(1) \) \((j = 1, 2)\).

The delay \( d_1 \) of \( P_1 \) can then be expressed as

\[ d_1 = \begin{cases} (w_{\mathcal{S}_1} - 1)^+ + s_1, & f_{P_1} = 0, \\ (w_{\mathcal{S}_1} - 1)^+ + \sum_{i=1}^{f_{P_1}+1} s_{1,i} + \sum_{i=1}^{v_{\mathcal{S}_1}} s_{2,i}, & f_{P_1} > 0. \end{cases} \]  
(36)
From the above equation, together with the independence of the service times from packet to packet, expressions (32) and (34), and the observation that \( f_{P_1} \) is independent of both \( v_{S_1} \) and \( w_{S_1} \), the pgf \( D_1(z) \) of \( d_1 \) follows as

\[
D_1(z) = E[z^{d_1}]
= S_1(z) E \left[ z^{(w_{S_1}-1)^+} \{ f_{P_1} = 0 \} \right]
+ E \left[ z^{(w_{S_1}-1)^+} S_1(z)^{f_{P_1}+1} S_2(z)^{w_{S_1}} \{ f_{P_1} > 0 \} \right]
= S_1(z) \text{Prob}[f_{P_1} = 0] E \left[ z^{(w_{S_1}-1)^+} \right]
+ S_1(z) E \left[ S_1(z)^{f_{P_1}} \{ f_{P_1} > 0 \} \right] E \left[ S_2(z)^{w_{S_1}} z^{(w_{S_1}-1)^+} \right]
\]

(37)

Also note that in the above derivation the \((\cdot)^+\) operator was removed based on the law of total expectation in a similar way as indicated in (17). Finally, with (21), (26), (34) and (35), this is further transformed into

\[
D_1(z) = p_0 S_1(z) \left\{ \frac{F_{P_1}(0)}{z-A(1(0))} \left( z - 1 \right) \frac{A_1(S_1(z)) - A(S_1(z), S_2(z))}{z-A(S_1(z), S_2(z))} - \frac{Q(1,A_1(0))}{A_1(0)} \frac{1-S_2(z)}{1-S_2(A_1(0))} \right\}
\]

(38)

\[
= p_0 S_1(z) \left\{ \frac{z-1}{z-A(S_1(z), S_2(z))} \left( A_1(S_1(z)) - A(S_1(z), S_2(z)) \right) \frac{1-A_1(0)}{z-A_1(0)} + \frac{A_1(S_1(z)) - 1}{S_1(z) - 1} \right\}
\]

5.2. Pgf of the class-2 delay. When a random class-2 packet \( P_2 \) arrives to the system, it is appended to the queue, along with all other \( a_2, S_2 - 1 \) class-2 arrivals during slot \( S_2 \). Although the position at which \( P_2 \) is inserted is more straightforward than for \( P_1 \), we do have to take into account the possibility that \( P_2 \) is passed by a class-1 arrival in a slot later than \( S_2 \). We therefore first focus on the total time needed to serve all packets in front of \( P_2 \) at the moment of its insertion, i.e., any remaining packets that were in the system already at the beginning of slot \( S_2 \), all \( a_1, S_2 \) class-1 arrivals during slot \( S_2 \) and all \( f_{P_2} \) class-2 packets that have arrived during \( S_2 \) and are to be served before \( P_2 \). We can find this time \( \delta \) as

\[
\delta = (w_{S_2} - 1)^+ + \sum_{i=1}^{a_1} s_{1,i} + \sum_{i=1}^{v_{S_2}} s_{2,i} + f_{P_2}.
\]

(39)
The pgf $\Delta(z)$ of $\delta$ is then calculated as follows:

\[
\Delta(z) = E[z^{\delta}]
\]

\[
= E\left[S_1(z)^{a_1,\delta_2} S_2(z)^{v_{\delta_2}} + f_{\delta_2} z (w_{\delta_2} - 1)^+\right]
\]

\[
= E\left[S_1(z)^{a_1,\delta_2} S_2(z)^{f_{\delta_2}}\right] E\left[S_2(z)^{v_{\delta_2} z (w_{\delta_2} - 1)^+}\right]
\]

\[
= E\left[S_1(z)^{a_1,\delta_2} S_2(z)^{f_{\delta_2}}\right] \cdot \{E\left[S_2(z)^{v_{\delta_2} z w_{\delta_2} - 1}\{w_{\delta_2} > 0\}\right] + \text{Prob}[w_{\delta_2} = 0, w_{\delta_2} = 0]\}
\]

\[
= \frac{1}{z} E\left[S_1(z)^{a_1,\delta_2} S_2(z)^{f_{\delta_2}}\right] \left[P(S_2(z), z) + (z - 1) p_0\right]
\]

\[
= E\left[S_1(z)^{a_1,\delta_2} S_2(z)^{f_{\delta_2}}\right] \frac{(z - 1) p_0}{z - A(S_1(z), S_2(z))},
\]

where we have subsequently used the independence of the packet service times, the independent nature of the packet arrival process from slot to slot, the law of total expectation and the expression (21), in a similar way as before. Note that $a_1, S_2$ and $f_{\delta_2}$ are not necessarily independent: $f_{\delta_2}$ is uniformly distributed between 0 and $a_2, S_2$, so $f_{\delta_2}$ is dependent on $a_2, S_2$ which can be correlated to $a_1, S_2$. The joint pgf of $a_1, S_2$ and $f_{\delta_2}$ can be found from (33) as

\[
E[x^{a_1,\delta_2} y^{f_{\delta_2}}] = E \left[ x^{a_1,\delta_2} \sum_{i=0}^{a_2, S_2 - 1} \frac{1}{a_2, S_2} y^i \right] = \frac{A(x, y) - A_1(x)}{\lambda_2 (y - 1)},
\]

such that (40) can be rewritten as

\[
\Delta(z) = \frac{p_0}{\lambda_2} \frac{z - 1}{S_2(z) - 1} \frac{A(S_1(z), S_2(z)) - A_1(S_1(z))}{z - A(S_1(z), S_2(z))}.
\]

The delay of $P_2$ can now be determined as the sum of the $\delta$ slots needed to serve all packets in front of $P_2$ when it was inserted, the service time of $P_2$ itself and, in case the reservation is seized before $P_2$ enters the server, the service time of the class-1 packet that jumped in front of $P_2$. In order for the reservation not to be taken before $P_2$ enters the server, there must not be any class-1 arrival during the first $\delta$ slots following slot $S_2$. The probability for the reservation not to be taken is therefore given by $A_1(0)^\delta$ and the probability of the reservation to be taken is then $1 - A_1(0)^\delta$. The delay $d_2$ of $P_2$ can thus be found as

\[
d_2 = \delta + \gamma_\delta s_1 + s_2,
\]

where $\gamma_\delta$ is a Bernoulli variable which is 0 with probability $A_1(0)^\delta$ and 1 with probability $1 - A_1(0)^\delta$. From (43) and (42), the pgf $D_2(z)$ of $d_2$ follows as

\[
D_2(z) = E[z^{d_2}]
\]

\[
= S_2(z) E[z^{\delta + \gamma_\delta s_1}]
\]

\[
= S_2(z) E[z^{\delta} S_1(z)^{\gamma_\delta}]
\]

\[
= S_2(z) E[z^{\delta} \left(A_1(0)^\delta + (1 - A_1(0)^\delta) S_1(z)\right)]
\]

\[
= S_2(z) \left[S_1(z) \Delta(z) + (1 - S_1(z)) \Delta(z A_1(0))\right]
\]

\[
= \frac{p_0}{\lambda_2} S_2(z) \left\{ \frac{S_1(z) (z - 1) (A(S_1(z), S_2(z)) - A_1(S_1(z)))}{(S_2(z) - 1) (z - A(S_1(z), S_2(z)))} \right\}
\]
Similarly, the mean delay of a random class-2 packet can be found as

\[
A(S_1(zA_1(0)), S_2(zA_1(0)) - A_1(S_1(zA_1(0))))
\]

\[
\frac{zA_1(0) - A(S_1(zA_1(0)), S_2(zA_1(0)))}{zA_1(0) - A(S_1(zA_1(0)), S_2(zA_1(0)))}
\]

5.3. Calculation of moments and tail probabilities. Using the moment generating property of pgfs, we can now determine the moments of the class-\(j\) packet delay from the pgf \(D_j(z)\). In particular, the mean delay of a random class-1 packet can be found in closed form as the first derivative of \(D_1(z)\), evaluated for \(z = 1\), i.e.,

\[
E[d_1] = D'_1(1)
\]

\[
= \mu_1 \left(1 + \frac{\lambda_1}{2\lambda_1} + \frac{1}{2p_0} (\lambda_1 \mu'_1 + \lambda_1 \mu_1^2 + 2\lambda_1 \mu_1 \mu_2 + \lambda_2 \mu'_2 + \lambda_2 \mu_2^2) \right) - \frac{\mu_2}{\lambda_1} \left(\lambda_2 - \frac{p_0 Q(1, A_1(0))}{A_1(0) (1 - S_2(1))}\right)
\]

where we have introduced \(\mu'_j \triangleq S''_j(1)\) and \(\lambda_{12} \triangleq \frac{\partial^2}{\partial t^2} A(1, 1)\) for convenience. Similarly, the mean delay of a random class-2 packet can be found as

\[
E[d_2] = D'_2(1)
\]

\[
= \mu_1 \left(1 + \frac{\lambda_1}{\lambda_2} - \Delta(A_1(0))\right) + \mu_2 \left(1 + \frac{\lambda_2}{2\lambda_2}\right)
\]

\[
+ \frac{1}{2p_0} (\lambda_1 \mu'_1 + \lambda_1^2 \mu_1^2 + 2\lambda_1 \mu_1 \mu_2 + \lambda_2 \mu'_2 + \lambda_2 \mu_2^2).
\]

We can use (45) and (46) to obtain the mean number of packets \(E[u]\) in the system at the beginning of a slot in steady state, using Little’s law:

\[
E[u] = \lambda_1 E[d_1] + \lambda_2 E[d_2].
\]

Expressions (38) and (44) can also be used to calculate higher-order moments. Thus, one can calculate the variance of the class-\(j\) packet delay from

\[
\text{var}[d_j] = D''_j(1) + D'_j(1) - D'_j(1)^2.
\]

Here, the second factorial moment of the type-1 delay is obtained from (38) as

\[
D''_1(1) = 2\mu'_1 + \frac{B''(1)}{3p_0} + \frac{(B''(1))^2}{2p_0^2} + \left(\frac{\lambda_1^2 \mu_1}{2p_0} + \frac{\lambda_1 \mu_1 - \lambda_2 \mu_2}{p_0} - 1\right)\frac{B''(1)}{\lambda_1}
\]

\[
+ \frac{2\lambda_2 \mu_2}{(1 - A_1(0)) \lambda_1} - \frac{p_0}{\lambda_1 A_1(0) (1 - A_1(0)) (1 - S_2(1))}\left[2\mu_2 - (1 - A_1(0)) (2\mu_1 \mu_2 - \mu'_2)\right]
\]

\[
+ \frac{2\lambda_1 \mu_1^2 + 3\lambda_1 (4\mu'_1 + \mu'_2) - 12\mu_1 \mu_2 \lambda_2}{6\lambda_1},
\]

with \(B(z) = A(S_1(z), S_2(z))\) the pgf of the arriving amount of work per slot, so that

\[
B'(1) = \lambda_1 \mu_1 + \lambda_2 \mu_2 = 1 - p_0,
\]

\[
B''(1) = \lambda'_1 \mu_1^2 + 2\lambda_1 \mu_1 \mu_2 + \lambda'_2 \mu_2^2 + \lambda_1 \mu'_1 + \lambda_2 \mu'_2.
\]
\[ B''(1) = \lambda'' \mu'' + 3 \lambda' \mu' + \frac{\partial^3}{\partial z_1 \partial z_2} A(z_1, z_2) \bigg|_{z_1=z_2=1} + 3 \mu_1^2 \frac{\partial^3}{\partial z_1 \partial z_2} A(z_1, z_2) \bigg|_{z_1=z_2=1} + \lambda'' \mu'' + 3 \lambda' \mu' + \lambda'' \mu'' + 3 \lambda_2^2 \mu_2^2 + \lambda_1' \mu_1' + \lambda_2' \mu_2', \]

and where \( \lambda'' \mu'' \) is \( S''(1) \) and \( \mu'' \) is \( S''(1) \). The second factorial moment of the type-2 delay follows from (44) as

\[ D''(1) \equiv \mu' + 2 \mu_1 + \mu_2 + 2(\mu_1 + \mu_2) \Delta'(1) + \Delta''(1) \]

\[ - (\mu' + 2 \mu_1 + \mu_2) \Delta(A_1(0)) - 2 \mu_1 A_1(0) \Delta'(A_1(0)), \tag{51} \]

where \( \Delta(z) \) is given by (42), so that

\[ \Delta'(1) = \frac{1 - \lambda_1 + \mu_1}{2 \lambda_2 \mu_2 - \lambda_2 \mu_2} - \frac{\lambda' \mu'}{2 \lambda_2 \mu_2} \]

\[ \Delta''(1) = \frac{\mu_2^2(\lambda_2 \mu_2 + \lambda_1 \mu_2 + \lambda_1 \mu_1) - \lambda'' \mu'' + 3 \lambda_1' \mu_1' + \lambda_1' \mu_1' + \lambda_2' \mu_2'}{3 \lambda_2 \mu_2} \]

\[ - \frac{\mu_2(1 - \lambda_1 \mu_1 + \lambda_1 \mu_1^2 + \lambda_1 \mu_2 + \lambda_1 \mu_2) B''(1)}{2 \lambda_2 \mu_2 - \lambda_2 \mu_2} \]

\[ + \frac{1 - \lambda_1 \mu_1}{\lambda_2 \mu_2} \left( \frac{B''(1)}{2 \mu_2^2 - \lambda_2 \mu_2} + \frac{B''(1)}{3 \mu_2^2 - \lambda_2 \mu_2} \right). \tag{53} \]

From the inversion formula for \( \mathcal{z} \)-transforms, it follows that the probability mass function \( \text{Prob}[d_j = n] \) of the class-\( j \) delay can be written as a weighted sum of negative \( n \)th powers of the poles of \( D_j(z) \). Since all these poles have a modulus larger than 1, \( \text{Prob}[d_j = n] \) is dominated by the contribution of the pole \( z_{d,j} \) with the smallest modulus. To guarantee non-negative values for the probability mass function, this dominant pole \( z_{d,j} \) must be real and positive, see e.g. [3]. It can also be shown (see Appendix) that \( z_{d,j} \) is always a pole of multiplicity 1. As such, the tail distribution of the class-\( j \) packet delay can be approximated by the following geometric form:

\[ \text{Prob}[d_j = n] \approx -\theta_j \left( \frac{1}{z_{d,j}} \right)^{n+1}, \tag{54} \]

for sufficiently large values of \( n \), where \( z_{d,j} \) is the dominant pole of the pgf \( D_j(z) \) and \( \theta_j \) is the residue of \( D_j(z) \) for \( z = z_{d,j} \). From the expressions (38) and (44) for \( D_1(z) \) and \( D_2(z) \), it can be understood that the tail behaviour of the packet delay for both classes is governed by the same dominant pole \( z_d \triangleq z_{d,1} = z_{d,2} \), where \( z_d \) is the smallest positive real root larger than 1 of the equation

\[ z - A(S_1(z), S_2(z)) = 0. \tag{55} \]

The value of \( z_d \) can be calculated numerically, e.g. by means of the Newton-Raphson method. The complex residue \( \theta_j \) of the pgf \( D_j(z) \) can then be found as \( \theta_j \triangleq \lim_{z \to z_d} (z - z_d) D_j(z) \). This yields

\[ \theta_1 = \frac{p_0}{\lambda_1} (z_d - 1) S_1(z_d) \left( \lim_{z \to z_d} \frac{z - z_d}{z - A(S_1(z), S_2(z))} \right) \]

\[ \cdot \left[ \left( A_1(S_1(z_d)) - z_d \right) \frac{1 - A_1(0)}{z_d - A_1(0)} + \frac{A_1(S_1(z_d)) - 1}{S_1(z_d) - 1} \right]. \tag{56} \]
and
\[ \theta_2 = \frac{p_0}{\lambda_2} (z_d - 1) S_1(z_d) S_2(z_d) \left( \frac{z_d - A_1(S_1(z_d))}{S_2(z_d) - 1} \right) \left( \lim_{z \to z_d} \frac{z - z_d}{z - A(S_1(z), S_2(z))} \right) \],

where the limit in (56) and (57) is calculated with use of de l’Hôpital’s rule as
\[ \lim_{z \to z_d} \frac{z - z_d}{z - A(S_1(z), S_2(z))} = \left( 1 - S_1'(z_d) \frac{\partial}{\partial 1} A(S_1(z_d), S_2(z_d)) - S_2'(z_d) \frac{\partial}{\partial 2} A(S_1(z_d), S_2(z_d)) \right)^{-1}. \]

6. **Discussion of numerical results.** In this section, we illustrate the results with some numerical examples. We consider an \( M \times M \) switch with output buffers, where each output buffer has one outlet, as shown in Figure 3. Packets arrive on one of the inlets of the switch according to a Bernoulli process and are then routed to the output buffer corresponding to their destination in a uniform and independent way. The joint pgf of the numbers of packet arrivals of both classes in an output buffer during a slot is then given by
\[ A(z_1, z_2) = \left( 1 - \frac{\lambda_1}{M} (1 - z_1) - \frac{\lambda_2}{M} (1 - z_2) \right)^M, \]

where \( \lambda_j \) is the probability of a class-\( j \) arrival on an arbitrary switch inlet during a slot. At most \( M \) packets can arrive in the switch per slot, which implies that the numbers of class-1 and class-2 packet arrivals during a slot are correlated. The load \( \rho \) in an output buffer is given by \( \rho = \lambda_1 \mu_1 + \lambda_2 \mu_2 \). We also define the traffic mix \( \alpha \) as \( \alpha = \frac{\lambda_1 \mu_1}{\rho} \). In the sequel, we choose the number of inlets/outlets \( M = 16 \).

![Figure 3. An \( M \times M \) switch with output buffers](image-url)

In a first set of examples, we consider deterministic service times for both classes, i.e.,
\[ S_j(z) = z^{\mu_j}. \]

The mean delays \( E[d_1] \) and \( E[d_2] \) experienced by class-1 and class-2 packets, respectively, in an output buffer of the switch can be calculated from the exact closed-form expressions (45) and (46). Figures 4 and 5 show the mean packet delays of both classes as a function of the load \( \rho \) for various values of the traffic mix \( \alpha \). In Figure 4 we consider identical service times of 3 slots for both classes,
i.e., $\mu_1 = \mu_2 = 3$, whereas in Figure 5 class-dependent service times are considered where $\mu_1 = 3$ and $\mu_2 = 20$. Figure 4 clearly illustrates that the mean delay of class-1 packets is significantly reduced due to the reservation-based scheduling; the price to pay is a somewhat larger mean class-2 delay. Note that due to the reserved position for class-1 arrivals, the first class-1 packet of each slot can jump over a number of class-2 packets, which explains the delay reduction. The first class-1 packet in a slot can in fact jump over those class-2 packets that were inserted in the queue since the previous class-1 packet, except if the service of those class-2 packets has already started. For $\mu_1 = \mu_2 = 3$, we observe that the smaller the portion $\alpha$ of class-1 packets in the total load, the lower the mean packet delays of both classes will be. Figure 5 shows that the latter is not the case for $\mu_1 = 3$ and $\mu_2 = 20$. Note that in this case the mean class-2 delay is at least 20 slots. In this case the influence of $\alpha$ is opposite. This can be explained intuitively based on the long class-2 service times; it is expected that their impact on the packet delays will be higher when the portion of class-2 traffic in the total load is higher (lower values of $\alpha$). The traffic mix $\alpha$ also has a similar effect on the standard deviations of the packet delays, as can be observed from Figure 6, where the standard deviations of the packet delays of both classes are plotted as a function of the load $\rho$ for the same set of system parameters as considered in Figure 5.

Approximations for the tail probabilities of the delays of class-1 and class-2 packets can be calculated from expressions (54)–(58). Figure 7 shows the tail probabilities of the packet delays of both classes on a logarithmic scale, again for deterministic service times ($\mu_1 = 3$, $\mu_2 = 20$), for a load $\rho = 0.8$ and various values of $\alpha$. Note that for a given value of $\alpha$ the curves for both classes are parallel, which could be expected due to the fact that the packet delay pgfs $D_1(z)$ and $D_2(z)$ have the same dominant pole $z_d$. The results are also in full agreement with the conclusions from Figure 5. We observe that for each value of $\alpha$ the tail probabilities of the class-1 packet delay are lower than those of the class-2 delay. The tail probabilities of both classes increase as $\alpha$ increases, in agreement with Figure 5. For $\alpha = 0.8$ we have compared our approximations for the tail probabilities with the exact probability mass functions of the packet delays. The latter probabilities were obtained by numerical inversion of the pgfs $D_1(z)$ and $D_2(z)$ using the algorithm presented in [1].
For a random variable with pgf $F(z)$, the inversion formula for $z$-transforms is

$$\text{Prob}[d_j = n] = \frac{1}{2\pi j} \int_{C_r} D_j(z)z^{-n-1}dz,$$

where $j$ is the complex imaginary unit $\sqrt{-1}$ and $C_r$ is a circular contour around the origin with radius $0 < r < 1$. In [1], this integral is approximated by sampling the integrand on $2n$ points of the contour $C_r$. Using the discrete Poisson summation formula, the error bound for this approximation is shown to be $r^{2n}$ for large $n$ and small $r$, such that any desired accuracy is guaranteed by choosing $r$ sufficiently small. Figure 7 shows that the approximations for the tail probabilities of the packet delays of both classes are very good.

To illustrate the influence of the mean service times of the class-1 and class-2 packets, Figures 8 and 9 show the tail probabilities of the packet delays of both classes, for $\rho = 0.8$, $\alpha = 0.15$ and various values of $\mu_1$ and $\mu_2$. Specifically, Figure 8 considers $\mu_2 = 4$ and several values of $\mu_1$ ($\mu_1 = 2, 4, 6, 8, 10$), whereas Figure 9 considers $\mu_1 = 4$ and several values of $\mu_2$ ($\mu_2 = 2, 4, 6, 8, 10$). The class-1 service
times are geometrically distributed and the class-2 service times have a uniform distribution between 1 and $m$ in both figures, i.e.,

$$S_1(z) = \frac{1}{\mu_1 - (\mu_1 - 1)z},$$

$$S_2(z) = \frac{z(1 - z^m)}{m(1 - z)}, \quad \mu_2 = \frac{m + 1}{2}.$$  \hspace{1cm} (63)

The figures clearly show that the mean lengths of the service times can have a big effect on the tail probabilities of both the class-1 and the class-2 delay. Note again that for each set of curves, the tail probabilities for class 1 and class 2 have the same decay rate.

In order to study the impact of the variance of the service times, in a next set of examples, we will consider deterministic service times for only one of the packet classes and for the other class we will assume the service times to be equal to 1 with probability $p$ and equal to $m$ with probability $1 - p$; for such class-$j$ packets, the
The pgf of the service times is then given by
\[ S_j(z) = pz + (1-p)z^m. \]  
(64)

The corresponding mean value and variance equal
\[ \mu_j = p + (1-p)m, \]  
(65)
\[ \sigma^2_j = p(1-p)(m-1)^2, \]  
(66)
respectively. To study the influence of \( \sigma^2_j \), the parameters \( p \) and \( m \) will be varied so that \( \sigma^2_j \) is varied, but \( \mu_j \) remains constant.

Figure 10 shows the mean packet delays of both classes versus the standard deviation \( \sigma_1 \) of the class-1 service times, for deterministic class-2 service times and variable class-1 service times (with pgf given by (64)), for \( \rho = 0.8, \mu_1 = \mu_2 = 3 \) and various values of the traffic mix \( \alpha \). On the contrary, in Figure 11, the mean packet delays of both classes are plotted versus the standard deviation \( \sigma_2 \) of the class-2 service times, for deterministic class-1 service times and variable class-2 service times, again for \( \rho = 0.8, \mu_1 = \mu_2 = 3 \) and various values of \( \alpha \). We observe that in case of variable class-1 service times the mean packet delays of both classes decrease with decreasing values of the fraction \( \alpha \) of class-1 packets in the total load. In case of more variable class-2 service times, however, the influence of \( \alpha \) is opposite, as expected intuitively.

Finally, in Figure 12, the mean values and the standard deviations of the packet delays of both classes are plotted as a function of the traffic mix \( \alpha \), for deterministic class-2 service times, variable class-1 service times (with pgf given by (64) with \( m = 20 \)), for \( \rho = 0.8 \) and \( \mu_1 = \mu_2 = 3 \). We see that for both packet classes the standard deviation keeps pace with the mean value, which points to delay distributions being close to exponential. The latter is also confirmed by the probability mass functions obtained by numerical inversion, as plotted in Figure 12.

7. Conclusions. We have analyzed the delay characteristics of a discrete-time single-server infinite-capacity queueing system with a two-class packet arrival process and class-dependent service times under a reservation-based priority scheduling mechanism. We have obtained exact closed-form expressions for the pgfs, the mean values and the tail probabilities of the packet delays of both classes. The analysis

\[ \text{Figure 9. Tail probabilities of the packet delays, for } M = 16, \rho = 0.8, \alpha = 0.15, \mu_2 = 4 \text{ and various values of } \mu_2. \]
\[E[d_1] \quad E[d_2] \quad \alpha = 0.8\]
\[E[d_1] \quad E[d_2] \quad \alpha = 0.5\]
\[E[d_1] \quad E[d_2] \quad \alpha = 0.2\]

**Figure 10.** Mean packet delays versus the standard deviation of the class-1 service times \(\sigma_1\), for \(M = 16, \rho = 0.8, \mu_1 = \mu_2 = 3\) and various values of \(\alpha\)

\[S_1(z) = z\mu_1, \quad \mu_1 = 3\]
\[S_2(z) = z\mu_2, \quad \mu_2 = 3\]
\[\sigma^2_1 = p(1-p)(m-1)^2\]

\[S_1(z) = pz + (1-p)zm, \quad \mu_1 = p + (1-p)m = 3, \quad \mu_2 = p + (1-p)m = 3\]

\[\sigma^2_2 = p(1-p)(m-1)^2\]

**Figure 11.** Mean packet delays versus the standard deviation of the class-2 service times \(\sigma_2\), for \(M = 16, \rho = 0.8, \mu_1 = \mu_2 = 3\) and various values of \(\alpha\)

\[S_1(z) = z^{\mu_1}, \quad \mu_1 = 3\]
\[S_1(z) = z^{\mu_1}, \quad \mu_1 = 3\]
\[S_1(z) = z^{\mu_1}, \quad \mu_1 = 3\]
\[S_1(z) = z^{\mu_1}, \quad \mu_1 = 3\]

**Figure 12.** Mean values and standard deviations of packet delays versus traffic mix \(\alpha\), for \(M = 16, \rho = 0.8, \) and \(\mu_1 = \mu_2 = 3\)
shows that the use of an in-queue reservation for class 1 allows to reduce class-1 delays at the expense of somewhat larger class-2 delays.

Since the studied reservation-based scheduling mechanism makes use of only one reserved position for class-1 arrivals, only the first class-1 packet in a slot can be inserted in the queue at a more favourable position and in fact, such packet can jump over at most the class-2 packets that were inserted in the queue since the previous class-1 packet, thus limiting the delay reduction for class 1. A possible extension of the considered reservation-based scheduling mechanism is to introduce multiple reservations for class-1 packets in the queue, which is expected to lead to a larger delay differentiation between the two types of packets. In such case, it is expected that the number of reservations will be an important parameter to tune the degree of delay differentiation between class 1 and class 2.

Acknowledgments. This research has been partially funded by the Interuniversity Attraction Poles Programme initiated by the Belgian Science Policy.

Appendix. In this Appendix, we show that the dominant pole \( z_d \) of the pgfs \( D_1(z) \) and \( D_2(z) \) is always a pole of multiplicity 1. First, note that in \( D_1(z) \) and \( D_2(z) \), the pole \( z_d \) is determined by the factor \( (z-A(S_1(z),S_2(z)))^{-1} \), in the sense that all other singularities of the delay pgfs have a modulus larger than that of \( z_d \). This can be verified by inspecting (38) and (44). Secondly, recall that \( S_1(z) \), \( S_2(z) \) and \( A(z_1,z_2) \) are analytic functions inside a disc around the origin with a modulus strictly larger than 1. That is, we exclude distributions for which not all moments exist. On the positive real axis, these pgfs are therefore smooth functions at least in \( z \in [1,L[ \), where \( L \) is the radius of convergence of \( U(z) = A(S_1(z),S_2(z)) \), the pgf of the amount of work \( U \) arriving in an arbitrary slot. In that region, \( U(z) = \sum_{n=0}^{\infty} u_n z^n \) with \( u_n = \text{Prob}[U = n] \).

We can prove that \( U(z) \) is strictly increasing in \([1,L[ \) if and only if \( u_0 < 1 \) and \( U'(z) \) is strictly increasing if and only if \( u_0 + u_1 < 1 \). Indeed, if \( u_0 < 1 \) then the power series \( U(z) \) contains terms in \( z^n \) with \( n > 0 \), so that for \( 0 < \varepsilon < L - z \),

\[
U(z + \varepsilon) = \sum_{n=0}^{\infty} u_n (z + \varepsilon)^n \\
\geq \sum_{n=0}^{\infty} u_n (z^n + \varepsilon) = U(z) + \varepsilon \sum_{n=0}^{\infty} u_n = U(z) + \varepsilon,
\]

and thus \( U(z + \varepsilon) > U(z) \), which means that \( U(z) \) is strictly increasing. Likewise, if \( u_0 + u_1 < 1 \) the power series \( U(z) \) contains terms in \( z^n \) with \( n > 1 \), so that now

\[
U'(z + \varepsilon) = \sum_{n=0}^{\infty} n u_n (z + \varepsilon)^{n-1} \\
\geq \sum_{n=0}^{\infty} n u_n (z^{n-1} + \varepsilon) = U'(z) + \varepsilon \sum_{n=0}^{\infty} n u_n = U'(z) + \varepsilon E[U],
\]

and thus \( U'(z + \varepsilon) > U'(z) \). Both conditions \( u_0 < 1 \) and \( u_0 + u_1 < 1 \) are fulfilled for our system because otherwise there would never be packets waiting in the queue at the beginning of a slot and so there would not by an actual queuing problem to be studied.

If \( z_d \) had a multiplicity larger than 1, then (58) would not exist and neither would \( \theta_1 \) or \( \theta_2 \). In other words, the denominator \( 1 - U'(z_d) \) of (58) would be 0. This,
however, never happens when the equilibrium condition is satisfied, in which case $U'(z_d) > 1$. This follows immediately from $U'(1) = 1, U(z_d) = z_d > 1$, and the mean value theorem which says there must be a $1 < x < z_d$ so that $U'(x) = 1$. Since $U'(z)$ is strictly increasing and $x < z_d$ we have indeed that $U'(z_d) > U'(x) = 1$.

REFERENCES

[1] J. Abate and W. Whitt, Numerical inversion of probability generating functions, *Operations Research Letters*, 12 (1992), 245–251.
[2] H. Bruneel, Performance of discrete-time queueing systems, *Computers & Operations Research*, 20 (1993), 303–320.
[3] H. Bruneel and B. G. Kim, *Discrete-Time Models for Communication Systems Including ATM*, Kluwer Academic Publishers, Boston, 1993.
[4] S. De Clercq, B. Steyaert and H. Bruneel, Delay analysis of a discrete-time multiclass slot-bound priority system, *4OR: A Quarterly Journal of Operations Research*, 10 (2012), 67–79.
[5] S. De Clercq, B. Steyaert, S. Wittevrongel and H. Bruneel, Analysis of a discrete-time queue with time-limited overtake priority, *Annals of Operations Research*, 238 (2016), 69–97.
[6] S. De Vuyst, S. Wittevrongel and H. Bruneel, Place reservation: Delay analysis of a novel scheduling mechanism, *Computers & Operations Research*, 35 (2008), 2447–2462.
[7] B. Feyaerts, S. De Vuyst, H. Bruneel and S. Wittevrongel, Delay analysis of a discrete-time GI-GI-1 queue with reservation-based priority scheduling, *Stochastic Models*, 32 (2016), 179–205.
[8] B. Feyaerts, S. De Vuyst, S. Wittevrongel and H. Bruneel, Analysis of a discrete-time priority queue with place reservations and geometric service times, in *Proceedings of the Summer Computer Simulation Conference, SCSC 2008/DASD* (Edinburgh, June 16-18, 2008), SCS, (2008), 140–147.
[9] T. Maertens, J. Walraevens and H. Bruneel, Performance comparison of several priority schemes with priority jumps, *Annals of Operations Research*, 162 (2008), 109–125.
[10] A. Melikov, L. Ponomarenko and C. Kim, Approximate method for analysis of queuing models with jump priorities, *Automation and Remote Control*, 74 (2013), 62–75.
[11] S. Ndreca and B. Scoppola, Discrete time GI/Geom/1 queueing system with priority, *European Journal of Operational Research*, 189 (2008), 1403–1408.
[12] H. Takagi, *Queueing Analysis: A Foundation of Performance Evaluation, Volume 3: Discrete-Time Systems*, North-Holland, Amsterdam, 1993.
[13] C.-K. Tham, Q. Yao and Y. Jiang, A multi-class probabilistic priority scheduling discipline for differentiated services networks, *Computer Communications*, 25 (2002), 1487–1496.
[14] J. Walraevens, B. Steyaert and H. Bruneel, Delay characteristics in discrete-time GI-GI-1 queues with non-preemptive priority queueing discipline, *Performance Evaluation*, 50 (2002), 53–75.
[15] S. Wittevrongel, B. Feyaerts, H. Bruneel and S. De Vuyst, Delay analysis of a queue with reservation-based scheduling and class-dependent service times, *Stoch. Models*, 32 (2016), 179–205.

Received March 2017; revised July 2017.

E-mail address: sabine.wittevrongel@UGent.be
E-mail address: bart.feyaerts@UGent.be
E-mail address: herwig.bruneel@UGent.be
E-mail address: stijn.devuyst@UGent.be