Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2012, Article ID 425630, 20 pages
doi:10.1155/2012/425630

Research Article

Queue Content Analysis in a 2-Class Discrete-Time Queueing System under the Slot-Bound Priority Service Rule

Sofian De Clercq, Bart Steyaert, and Herwig Bruneel

SMACS Research Group, Ghent University, Sint-Pietersnieuwstraat 4, 9000 Gent, Belgium

Correspondence should be addressed to Sofian De Clercq, sofian.declercq@telin.ugent.be

Received 10 April 2012; Accepted 2 September 2012

Academic Editor: Sri Sridharan

The paper we present here introduces a new priority mechanism in discrete-time queueing systems. It is a milder form of priority when compared to HoL priority, but it favors customers of one type over the other when compared to regular FCFS. It also provides an answer to the starvation problem that occurs in HoL priority systems. In this new priority mechanism, customers of different priority classes entering the system during the same time slot are served in order of their respective priority class—hence the name slot-bound priority. Customers entering during different slots are served on an FCFS basis. We consider two customer classes (pertaining to two levels of priority) such that type-1 customers are served before type-2 customers that enter the system during the same slot. A general independent arrival process and generally distributed service times are assumed. Expressions for the probability generating function (PGF) of the system content (number of type-j customers, $j = 1, 2$) in regime are obtained using a slot-to-slot analysis. The first moments are calculated, as well as an approximation for the probability mass functions associated with the found PGFs. Lastly, some examples allow us some deeper insight into the inner workings of the slot-bound priority mechanism.

1. Introduction

Multiclass queueing systems, or queueing systems buffering multiple types of customers, have been widely adopted in queueing theory, since they enable the modelling of non-identical behaviour of different types of customers that enter the same system. In a multiclass environment, virtually any combination of features with respect to the arrival characteristics, service requirements, buffer management rules that pertain to the individual classes (Fiems et al. [1]) could be considered.

In this paper we study a 2-class discrete-time queueing system with infinite waiting room and one server, under the so-called *slot-bound* priority service rule (SBP) (which is based on the work in De Clercq et al. [2]). That is to say, class-1 customers receive preferential treatment over class-2 customers that have arrived during the same slot. In addition, customers that enter the system during consecutive slots are served on a first-come-first-served (FCFS) basis, regardless of the class they belong to. Slot-bound priority can be used to model any system in which batches of, for example, customers, packets, or tasks, arrive which have to be addressed or serviced in a specific order for whatever reason, while the batches themselves need to be served FCFS. For instance, a batch of customers may be the traffic that accumulates before a traffic light. When the light turns green, the faster drivers (high priority customers) will gain an edge and arrive at the next lights sooner, where they will be "served" once those lights turn green. Moreover SBP can be seen as a polling mechanism in which, during each slot, a gate is placed after each of the two queues (see f.i. Takagi [3], Boxma et al. [4]). Whenever the server encounters a gate it discards this gate, and starts service on the other queue, hence preserving the FCFS order across different arrival slots. The lower priority customer's service times may be seen as server vacations for the high priority customers and as we will see the type-1 (and even type-2) population of the queue has the so-called decomposition property. Fuhrmann and Cooper [5] and Shantikumar [6] show this property in continuous time and Ishizaki [7] shows a decomposition property in discrete time.

The complexity of the analysis of this type of multiclass queueing system is, among others, highly dependent on the service-time distributions. It is often found that deterministic service times are used in order to ease the analysis (e.g., in Fiems et al. [1] or Stavrakakis [8]) while still being widely applicable, since, for instance, the ATM transport paradigm employs fixed-size packages. Another service-time distribution that is frequently adopted is the geometric distribution, which reduces the analysis's complexity due to its memoryless nature (e.g., in Ndreca and Scoppola [9]). Nevertheless, as will be demonstrated in the subsequent sections, the approach that is proposed in this paper allows us to obtain results for generally distributed service times that are independent of one another; however, note that their probability distribution may be dependent on the class of the customer being served. On the other hand, the numbers of arrivals during consecutive slots are assumed to be mutually independent as well, albeit that the numbers of arrivals of customers of different types during the same slot can be correlated in our setting. Discrete-time queueing systems with independent and identically distributed (i.i.d.) customer arrivals have been studied mostly under a general arrival distribution (Bruneel [10], Walraevens et al. [11], Stavrakakis [8], and Ndreca and Scoppola [9]), since the exact nature of the arrival process has, apart from some pathological cases, little or no impact on the complexity of the analysis. The ultimate purpose of this contribution is to analyze the joint probability distribution of the system content at random slot marks as contributed by all types of customers, following a slot-to-slot approach. We refer to De Clercq et al. [12] for the delay analysis for this model.

In the related literature, multiclass systems may be looked upon as having multiple arrival streams with different characteristics, (e.g., Masuyama and Takine [13], Takine [14]) or servicing multiple types of customers/fluids (e.g., He [15], Kulkarni and Glazebrook [16]). Masuyama analysed a system much like the one considered here, in a continuous-time setting with multiple batch Markovian arrival streams and batches consisting of customers of the same type, whereas our discrete time model allows batches containing customers of different types. The references indicated above assume a continuous-time setting. A number of contributions have also been made regarding multiclass systems in discrete-time, mostly in combination with some sort of priority rule, for example, nonpreemptive priority

(e.g., Walraevens et al. [11], Fiems et al. [1], Ndreca and Scoppola [9]), or gated priority (e.g., Stavrakakis [8], Ishizaki et al. [17]). There are some results for discrete-time FCFS-based systems with multiple types of customers as well, as can be seen, for instance, in Van Houdt and Blondia [18], where the customer delay distributions are studied for the specific case of *MMAP[K]/PH[K]/1* and where the arrival process is generalised to include batch arrivals. Note however that the delay analysis in Van Houdt and Blondia [18] is based on the total amount of work in the system at a random slot mark, while we are interested in studying the numbers of customers of both types in the system, which is a much more difficult task. In addition, as far as the service-time distributions are concerned, our model is more general as well.

Interestingly, when studying the SBP policy described previously in a discrete-time multiclass system, one encounters the same intricate problem as Takine [14] (page 349) mentioned having in his analysis of a multiclass FIFO system with class-dependent nonexponential interarrival times: "It is widely recognized that the queue length distribution in a FIFO queue with multiple non-Poissonian arrival streams having different service-time distributions is very hard to analyze, since we have to keep track of the complete order of customers in the queue to describe the queue length dynamics." We will see that under certain assumptions concerning the arrival process, our approach will suffice to deliver a discrete time solution to this problem. A first assumption is the general independent nature of the arrival process (batch Bernoulli process). When we demand that during a slot only customers pertaining to one class can enter the system, a pure multiclass FCFS policy in discrete-time is the result.

The remainder of the paper is organised as follows. In the next section we present the mathematical model of the system. Next, we detect a 4-dimensional Markov chain which we can analyze and derive a steady-state expression for the joint *probability generating function* (PGF) of the system occupancies of both types of customers. From this main result, we subsequently derive expressions for the mean values of these random variables, determine their tail probabilities, and for some specific examples compare them to the mean system occupancies in a system governed by the nonpreemptive head-of-line (np-HoL) priority rule, before concluding this paper.

2. Mathematical Model

In a discrete-time setting, consider a queue with infinite waiting room and a single server serving 2 types of customers in FCFS order. When customers of different types enter the system during the same slot (i.e., simultaneously), the type-1 customers among them are served first. This principle is known as slot-bound priority (De Clercq et al. [2]). The number of type-j customers ($j = 1, 2$) entering the system during slot n is denoted by $a_{j,n}$. We will adopt the notation $A_n(z_1, z_2) \triangleq E[z_1^{a_{1,n}} z_2^{a_{2,n}}]$ for the joint pgf of $a_{1,n}$ and $a_{2,n}$. Hence, our model is not limited to uncorrelated $a_{1,n}$ and $a_{2,n}$. However, we will consider i.i.d. arrivals, meaning $(a_{1,n}, a_{2,n})$ and $(a_{1,m}, a_{2,m})$ are i.i.d. random vectors for $n \neq m$. Therefore, we will omit the index and use $A(z_1, z_2)$ instead. The first-order partial derivatives of this function taken in $(1, 1)$ are the arrival rates of each separate type of customer:

$$\lambda_j \triangleq \left.\frac{\partial A(z_1, z_2)}{\partial z_j}\right|_{(z_1,z_2)=(1,1)} = E\left[a_j\right], \qquad j = 1, 2. \tag{2.1}$$

We consider a single-server system, where all service times are modelled as being independent and generally distributed. Furthermore all service times of type-j customers are i.i.d. discrete random variables (DRVs). The earliest a customer's service time can start is the slot following its arrival slot. Let s_j (with pgf $S_j(z) \triangleq E[z^{s_j}]$) denote the service time of a random type-j customer.

Additionally we define the auxiliary DRVs b_j, a_g, and s_g. b_j will denote the number of type-j customers entering the system during a random slot given that at least one customer of any type enters the system during the slot. Consequently $b_1 + b_2 > 0$ by definition. Secondly, a_g is an indicator which is 1 if at least one customer enters the system during a slot and 0 otherwise. If $a_g = 1$, we aggregate the arriving customers and call this set of customers a *group* (of customers) for short (hence the index g). In such a setting, a_g is the number of groups (of customers) entering the system during a slot. s_g is the service time of such a group, meaning the combined service time of all customers making up such a group. Based on these definitions, we may write

$$B(z_1, z_2) \triangleq E\left[z_1^{b_1} z_2^{b_2}\right] = \frac{A(z_1, z_2) - A(0,0)}{1 - A(0,0)}, \tag{2.2}$$

$$A_g(z) \triangleq E[z^{a_g}] = A(0,0) + (1 - A(0,0))z, \tag{2.3}$$

$$S_g(z) \triangleq E[z^{s_g}] = B(S_1(z), S_2(z)). \tag{2.4}$$

Note that $A_g(B(z_1, z_2)) = A(z_1, z_2)$, a relation that will be fully exploited further in this paper.

In this paper we are interested in the queue content, meaning the number of customers of each type in the system at typical slot marks. With $v_{j,n}$ we denote the number of type-j customers in the queue at the beginning of slot n. $\Pr[v_{j,n} = k]$ is in general a function of n, the slot index. To avoid having to specify the slot index, we assume that the system reaches stochastic equilibrium and as such let n approach infinity. A sufficient condition in our model would be that the arrival rate of customers multiplied by the average service time of each of said customers is less than 1. Formally let $\rho_j \triangleq \lambda_j E[s_j]$. Then ρ_j can be interpreted as being the average number of slots it takes the server to serve all type-j customers that enter the system during a single slot. We can then say that stochastic equilibrium is reached if $\rho \triangleq \rho_1 + \rho_2 < 1$. We hold this inequality for true in the rest of this paper.

In general, $v_{1,n}$ and $v_{2,n}$ are correlated. We will focus our efforts towards finding an expression for

$$V(z_1, z_2) \triangleq \lim_{n\to\infty} V_n(z_1, z_2) \triangleq \lim_{n\to\infty} E\left[z_1^{v_{1,n}} z_2^{v_{2,n}}\right], \tag{2.5}$$

the steady-state joint pgf for v_1 and v_2.

3. Queue Content Analysis

We already introduced the concept of groups. The reason why we group customers that enter the system during the same slot is that we know that customers of different groups are served in FCFS order, while those that are part of the same group are subject to the slot-bound priority rule (type-1 customers are served before type-2 customers). On top of that, we know that each group's content, that is, the number of customers of each type in it, is an i.i.d.

process described by (b_1, b_2). Let w_n denote the number of groups in the queueing system at the beginning of slot n. If there is only a part of a certain group in the system at the beginning of that slot because some customers of that group have already been served, we still include this group in w_n. As such $w_n = 0$ means the system is empty.

If the system is not empty, then the server must be serving a customer. The group that customer belongs to (called the active group) might already have had some customers served and might house more customers than just the one in service. Therefore, let $h_{j,n}$ denote the number of type-j customers in the active group that are still in the system at the beginning of slot n. By convention, we will set $h_{1,n} = h_{2,n} = 0$ if the queueing system is empty.

Lastly, whenever the server is serving a customer, we define r_n to be its remaining service time at the beginning of slot n. In the other case, we set $r_n = 0$, once again implying that the system is empty. Summarizing, we find that

$$w_n = 0 \Longleftrightarrow h_{1,n} = h_{2,n} = 0 \Longleftrightarrow r_n = 0 \Longleftrightarrow v_{1,n} = v_{2,n} = 0. \tag{3.1}$$

The purpose of introducing these new DRVs is twofold. For one, together they constitute a discrete-time Markov chain. Secondly and most importantly, we can determine the number of customers of each type using these drv's. The number of type-j customers in the queue at the beginning of slot n is the sum of those present in the active group, $h_{j,n}$, and those in the $(w_n - 1)^+$ queued groups, that have not yet been served (we adopt the notation $(\cdot)^+ \equiv \max(\cdot, 0)$). The number of type-$j$ customers in these latter groups is all i.i.d. drv's with distribution equal to that of b_j (see Section 2). And so we find

$$v_{j,n} = \sum_{i=1}^{(w_n-1)^+} b_{j,i} + h_{j,n}, \quad j = 1, 2. \tag{3.2}$$

The index i in $b_{j,i}$ was added as an enumeration index for the unserved groups that are queued in the system. Technically you do not need r_n in this equation. However, we do need r_n to form a Markov chain together with w_n and both $h_{j,n}$, given the service-time distributions and arrival process. The renewal period for this Markov chain equals one slot, so we will focus on a slot-to-slot analysis to determine the joint pgf of $h_{1,n}$, $h_{2,n}$, r_n, and w_n in regime.

First notice that $(h_{1,n}, h_{2,n}, r_n, w_n) = (0,0,0,0)$, $(1,0,1,1)$ or $(0,1,1,1)$ are similar cases, in the sense that the corresponding transition probabilities of the system's state at the beginning of slot $n+1$ are same for these three cases. In the first we have an empty system, and in the latter two cases we have a system with only one customer sitting out its last slot of service. The state at the beginning of the next slot is in either case going to be dependent solely on the arrival process during slot n. In view of the SBP service paradigm we find that

$$\begin{gathered} h_{j,n+1} = a_{j,n}, \quad j = 1, 2, \\ r_{n+1} = \begin{cases} s_1, & \text{if } a_{1,n} > 0, \\ s_2, & \text{if } a_{1,n} = 0,\ a_{2,n} > 0, \\ 0, & \text{if } a_{1,n} = a_{2,n} = 0, \end{cases} \\ w_{n+1} = a_g. \end{gathered} \tag{3.3}$$

When $h_{1,n} + h_{2,n} = 1$ and $r_n = 1$, the active group will leave the system at the end of slot n. In the above we saw what the system state evolves to when this group was the only group in the system at the start of slot n. When $w_n > 1$, the next group in line will be served at the beginning of slot $n+1$. It consists of b_j type-j customers and hence our state evolves to

$$\begin{gathered} h_{j,n+1} = b_j, \quad j = 1,2, \\ r_{n+1} = \begin{cases} s_1, & \text{if } b_1 > 0, \\ s_2, & \text{if } b_1 = 0 \ (\text{implying } b_2 > 0), \end{cases} \\ w_{n+1} = w_n - 1 + a_g. \end{gathered} \tag{3.4}$$

If $r_n = 1$, then we know that a customer will finish service at the end of slot n. Moreover when $h_{1,n} + h_{2,n} > 1$, the leaving customer will not be the last of the active group and hence this group will still be in the system at the beginning of slot $n+1$.

(i) Assume $h_{1,n} > 1$. In this case the customer leaving the system was of type 1 and the active group contains at least one additional type-1 customer. Hence the following customer selected for service will be again of type 1, leading to the following set of system equations:

$$\begin{gathered} h_{1,n+1} = h_{1,n} - 1, \\ h_{2,n+1} = h_{2,n}, \\ r_{n+1} = s_1, \\ w_{n+1} = w_n + a_g. \end{gathered} \tag{3.5}$$

(ii) Second, assume $h_{1,n} = 1$ and $h_{2,n} > 0$. This case covers what happens if the customer leaving the system was of type-1, but contrary to the above case, the active group does not contain any additional type-1 customers. It does however contain a type-2 customer. In such a case, our state evolves into

$$\begin{gathered} h_{1,n+1} = 0, \\ h_{2,n+1} = h_{2,n}, \\ r_{n+1} = s_2, \\ w_{n+1} = w_n + a_g. \end{gathered} \tag{3.6}$$

(iii) Third, when $h_{1,n} = 0$ and $h_{2,n} > 1$, we know that the served customer was of type-2 and the active group contains at least one additional type-2 customer. Hence the following customer selected for service will again be of type-2:

$$\begin{gathered} h_{1,n+1} = 0, \\ h_{2,n+1} = h_{2,n} - 1, \end{gathered}$$

$$r_{n+1} = s_2,$$
$$w_{n+1} = w_n + a_g. \tag{3.7}$$

This covers the cases where the active group does not leave the system although a customer does at the end of slot n.

When $r_n > 1$, no customer will leave the system, and hence, the system state evolves to

$$\begin{aligned} h_{j,n+1} &= h_{j,n}, \quad j = 1,2, \\ r_{n+1} &= r_n - 1, \\ w_{n+1} &= w_n + a_g. \end{aligned} \tag{3.8}$$

Equations (3.3)–(3.8) cover the possible evolution of the state description $(h_{1,n}, h_{2,n}, r_n, w_n)$. We define the distribution function $p_n(i_1, i_2, j, k)$ and joint pgf $P_n(x_1, x_2, y, z)$ of these drv's as

$$\begin{aligned} p_n(i_1, i_2, j, k) &= \Pr\left[h_{1,n} = i_1, h_{2,n} = i_2, r_n = j, w_n = k\right], \\ P_n(x_1, x_2, y, z) &\triangleq E\left[x_1^{h_{1,n}} x_2^{h_{2,n}} y^{r_n} z^{w_n}\right] \\ &= \sum_{i_1,i_2,j,k=0}^{\infty} x_1^{i_1} x_2^{i_2} y^j z^k p_n(i_1, i_2, j, k). \end{aligned} \tag{3.9}$$

Using the system equations in (3.3) and (3.4) we can deduce an expression for $P_{n+1}(x_1, x_2, y, z)$, the joint pgf of the system's state at the start of slot $n+1$, as follows:

$$\begin{aligned} P_{n+1}(x_1, x_2, y, z) &= E\left[x_1^{h_{1,n+1}} x_2^{h_{2,n+1}} y^{r_{n+1}} z^{w_{n+1}}\right] \\ &= E\left[x_1^{a_{1,n}} x_2^{a_{2,n}} y^{r_{n+1}} z^{a_g} \{h_{1,n} + h_{2,n} = r_n = w_n \le 1\}\right] \\ &\quad + E\left[x_1^{b_1} x_2^{b_2} y^{r_{n+1}} z^{w_n - 1 + a_g} \{h_{1,n} + h_{2,n} = r_n = 1, w_n > 1\}\right] \\ &\quad + E\left[x_1^{h_{1,n}-1} x_2^{h_{2,n}} y^{s_1} z^{w_n + a_g} \{h_{1,n} > 1, r_n = 1\}\right] \\ &\quad + E\left[x_1^{0} x_2^{h_{2,n}} y^{s_2} z^{w_n + a_g} \{h_{1,n} = 1, h_{2,n} > 0, r_n = 1\}\right] \\ &\quad + E\left[x_1^{0} x_2^{h_{2,n}-1} y^{s_2} z^{w_n + a_g} \{h_{1,n} = 0, h_{2,n} > 1, r_n = 1\}\right] \\ &\quad + E\left[x_1^{h_{1,n}} x_2^{h_{2,n}} y^{r_n - 1} z^{w_n + a_g} \{r_n > 1\}\right], \end{aligned} \tag{3.10}$$

in which we used the notation $E[A\{B\}]$ for $E[A \mid B]\Pr[B]$. Conditioning further on the arrival process to remove r_{n+1} from the first two terms in (3.10) yields the following:

$$\begin{aligned}
P_{n+1}(x_1,x_2,y,z) = {} & A_g\big(z\big(B(x_1,x_2)S_1(y) + B(0,x_2)\big(S_2(y) - S_1(y)\big)\big)\big) \\
& \times \big(P_n(0,0,0,0) + p_n(1,0,1,1) + p_n(0,1,1,1)\big) \\
& + A_g(z)\big(B(x_1,x_2)S_1(y) + B(0,x_2)\big(S_2(y) - S_1(y)\big)\big) \\
& \times \left(\sum_{k=1}^{\infty} z^{k-1}\big(p_n(1,0,1,k) + p_n(0,1,1,k)\big) - p_n(1,0,1,1) - p_n(0,1,1,1)\right) \\
& + \frac{A_g(z)S_1(y)}{x_1}\sum_{i_1=2}^{\infty}\sum_{i_2=0}^{\infty}\sum_{k=1}^{\infty} x_1^{i_1}x_2^{i_2}z^k p_n(i_1,i_2,1,k) \\
& + A_g(z)S_2(y)\sum_{i_2,k=1}^{\infty} x_2^{i_2}z^k p_n(1,i_2,1,k) \\
& + \frac{A_g(z)S_2(y)}{x_2}\sum_{i_2=2}^{\infty}\sum_{k=1}^{\infty} x_2^{i_2}z^k p_n(0,i_2,1,k) \\
& + \frac{A_g(z)}{y}\left(P_n(x_1,x_2,y,z) - P_n(0,0,0,0) - \sum_{i_1,i_2,k=0}^{\infty} x_1^{i_1}x_2^{i_2}z^k p_n(i_1,i_2,1,k)\right).
\end{aligned} \tag{3.11}$$

In order to tackle this elaborate expression, we introduce some short-hand notation

$$\begin{aligned}
R_n(x_1,x_2,z) &\triangleq \sum_{i_1=1}^{\infty}\sum_{i_2=0}^{\infty}\sum_{k=1}^{\infty} x_1^{i_1-1}x_2^{i_2}z^{k-1}p_n(i_1,i_2,1,k), \\
Q_n(x_2,z) &\triangleq \sum_{i_2=1}^{\infty}\sum_{k=1}^{\infty} x_2^{i_2-1}z^{k-1}p_n(0,i_2,1,k).
\end{aligned} \tag{3.12}$$

As we agreed in the previous section, we assume a system in stochastic equilibrium. Concretely we define

$$\begin{aligned}
P(x_1,x_2,y,z) &\triangleq \lim_{n\to\infty} P_n(x_1,x_2,y,z), \\
R(x_1,x_2,z) &\triangleq \lim_{n\to\infty} R_n(x_1,x_2,z), \\
Q(x_1,x_2,z) &\triangleq \lim_{n\to\infty} Q_n(x_1,x_2,z).
\end{aligned} \tag{3.13}$$

We are interested in the steady-state joint pgf $P(x_1,x_2,y,z)$, since $P_n(x_1,x_2,y,z)$ will resemble $P(x_1,x_2,y,z)$ for large enough values of n given any starting conditions. Taking

the limit of both sides of (3.11) while using the previous definitions yields after rearranging the terms

$$\begin{aligned}
P(x_1,x_2,y,z)\left(1-\frac{A_g(z)}{y}\right) = {} & A_g(z)z(S_1(y)-x_1)R(x_1,x_2,z) \\
& + A_g(z)z(S_2(y)-S_1(y))R(0,x_2,z) \\
& + A_g(z)z(S_2(y)-x_2)Q(x_2,z) \\
& - P_0\frac{A_g(z)}{y} - A_g(z)zS_2(y)(R(0,0,z)+Q(0,z)) \\
& + A_g(z)(S_1(y)B(x_1,x_2)+(S_2(y)-S_1(y))B(0,x_2)) \\
& \times (R(0,0,z)+Q(0,z)-R(0,0,0)-Q(0,0)) \\
& + A_g(z(S_1(y)B(x_1,x_2)+(S_2(y)-S_1(y))B(0,x_2))) \\
& \times (P_0+R(0,0,0)+Q(0,0)),
\end{aligned} \tag{3.14}$$

where we wrote $P_0 \triangleq P(0,0,0,0)$ for short. Since $p_n(i_1,i_2,l,0) = 0$ if $(l,i_1,i_2) \neq (0,0,0)$, the equivalence

$$P(x_1,x_2,y,0) \equiv P_0 \tag{3.15}$$

holds by definition. Considering the left-hand side and the right-hand side of (3.14) for $z = 0$, and taking into account that $A_g(0) = A(0,0)$, gives us the following relation between $R(0,0,0)+Q(0,0)$ and P_0

$$P_0 = A(0,0)(P_0+R(0,0,0)+Q(0,0)). \tag{3.16}$$

The equation in (3.14) still contains the unknown functions $R(x_1,x_2,z)$ and $Q(x_2,z)$. To solve for them, notice that (3.14) holds for all values x_1, x_2, y, and z with moduli less than 1 (i.e., in the complex unit disk) since the (partial) probability generating functions that appear in (3.14) are then analytic functions. Because $|A_g(z)| < 1$ when $|z| < 1$, (3.14) holds when we substitute y by $A_g(z)$. For the same reason we can substitute x_1 by 0 therein. To keep the resulting expression and all following expressions a tad bit more tidy, we omit the function parentheses and write $XY(z)$ where we mean $X(Y(z))$, in which X is a function with only one variable. Thus we write

$$\begin{aligned}
A_g(z)zS_2A_g(z)R(0,x_2,z) = {} & A_g(z)z(x_2-S_2A_g(z))Q(x_2,z) \\
& - A_g(z)S_2A_g(z)(B(0,x_2)-z)(R(0,0,z)+Q(0,z)) \\
& - S_2A_g(z)B(0,x_2)(z-1)(1-A(0,0))P_0.
\end{aligned} \tag{3.17}$$

This is a first equation in $R(0,x_2,z)$, $Q(x_2,z)$, and $R(0,0,z)+Q(0,z)$, all three of which are at this point unknown. If we substitute y by $A_g(z)$ and x_1 by $S_1A_g(z)$ in (3.14) we can find a second equation in those same three unknown functions. This second equation reads

$$\begin{aligned}
&A_g(z)z\big(S_2A_g(z)-S_1A_g(z)\big)R(0,x_2,z)\\
&\quad= A_g(z)z\big(x_2-S_2A_g(z)\big)Q(x_2,z)+A_g(z)zS_2A_g(z)(R(0,0,z)+Q(0,z))\\
&\qquad+\big(S_1A_g(z)B\big(S_1A_g(z),x_2\big)+\big(S_2A_g(z)-S_1A_g(z)\big)B(0,x_2)\big)\\
&\qquad\cdot\big(P_0\big(1-A_g(z)\big)-A_g(z)(R(0,0,z)+Q(0,z))\big).
\end{aligned}\tag{3.18}$$

Notice that in both of the above-described substitutions we aimed to remove $P(x_1, x_2, y, z)$ and $R(x_1,x_2,z)$ from the equation. Following this same strategy, we eliminate $Q(x_2,z)$ from the equation by substituting x_2 by $S_2A_g(z)$ in both (3.17) and (3.18), granting us a system of two equations in $R(0,S_2A_g(z),z)$ and $R(0,0,z)+Q(0,z)$. Solving it for the latter function eventually leads to

$$A_g(z)(R(0,0,z)+Q(0,z))=P_0(1-A(0,0))\frac{(z-1)S_gA_g(z)}{z-S_gA_g(z)}.\tag{3.19}$$

Note that (3.19) for $z=0$ implies (3.16). We can now invoke the system of equations in (3.17) and (3.18) to obtain $R(0,x_2,z)$ and $Q(x_2,z)$, now that $R(0,0,z)+Q(0,z)$ is known explicitly. This produces the following expressions:

$$A_g(z)R(0,x_2,z)=P_0(1-A(0,0))(z-1)\frac{B\big(S_1A_g(z),x_2\big)-B(0,x_2)}{z-S_gA_g(z)},\tag{3.20}$$

$$A_g(z)Q(x_2,z)\frac{x_2-S_2A_g(z)}{S_2A_g(z)}=P_0(1-A(0,0))(z-1)\frac{B\big(S_1A_g(z),x_2\big)-S_gA_g(z)}{z-S_gA_g(z)}.\tag{3.21}$$

Notice that (3.19) can be checked using (3.20) and (3.21) setting $x_2=0$. Substituting y by $A_g(z)$ in (3.14) removes only $P(x_1,x_2,y,z)$ from the equation and thus we are able to calculate $R(x_1,x_2,z)$, resulting in

$$\begin{aligned}
&A_g(z)\big(x_1-S_1A_g(z)\big)R(x_1,x_2,z)\\
&\quad=P_0(1-A(0,0))(z-1)\frac{S_1A_g(z)\big(B(x_1,x_2)-B\big(S_1A_g(z),x_2\big)\big)}{z-S_gA_g(z)},
\end{aligned}\tag{3.22}$$

which can again be checked by means of (3.20). The unknown functions now known, we substitute them in (3.14) and rework the resulting formula, in order to obtain the moderately presentable final result hereafter:

$$\begin{aligned}\frac{P(x_1,x_2,y,z)}{P_0} &= 1 + x_1yz\left(\frac{A_g(z)-1}{y-A_g(z)}\right)\left(\frac{S_1(y)-S_1A_g(z)}{x_1-S_1A_g(z)}\right)\left(\frac{B(x_1,x_2)-B(S_1A_g(z),x_2)}{z-S_gA_g(z)}\right)\\ &\quad + x_2yz\left(\frac{A_g(z)-1}{y-A_g(z)}\right)\left(\frac{S_2(y)-S_2A_g(z)}{x_2-S_2A_g(z)}\right)\left(\frac{B(S_1A_g(z),x_2)-S_gA_g(z)}{z-S_gA_g(z)}\right).\end{aligned} \tag{3.23}$$

Using the normalisation property of pgf's we find the constant P_0 to be $1-\rho$. The above formula is pretty symmetric with respect to the last two terms of the right-hand side, the only exception being their last factor. This pgf actually harbors much more information than we intended to obtain in the first place, but for sake of continuity, we show our results for $V(z_1,z_2)$. Substituting both v_j, $j=1,2$ in $V(z_1,z_2)$ by the system equation found in (3.2) and taking the limit $n\to\infty$, we obtain

$$V(z_1,z_2) = (1-\rho)\left(1-\frac{1}{B(z_1,z_2)}\right) + \frac{P(z_1,z_2,1,B(z_1,z_2))}{B(z_1,z_2)}. \tag{3.24}$$

Substituting the expression in (3.23) in this equation gives our main result, which is a closed-form expression for the joint pgf $V(z_1,z_2)$.

3.1. Main Result

The joint pgf of type-1 and type-2 customers in a single-server discrete-time queueing system under the slot-bound priority rule, where the arrival process of the two types of customers is batch Bernoulli with joint pgf $A(z_1,z_2)$ and independent class-specific service times with pgf's $S_1(z)$ and $S_2(z)$, is given by

$$\begin{aligned}V(z_1,z_2) = (1-\rho)\Bigg(1 &+ z_1\,\frac{S_1A(z_1,z_2)-1}{z_1-S_1A(z_1,z_2)}\,\frac{B(z_1,z_2)-B(S_1A(z_1,z_2),z_2)}{B(z_1,z_2)-S_gA(z_1,z_2)}\\ &+ z_2\,\frac{S_2A(z_1,z_2)-1}{z_2-S_2A(z_1,z_2)}\,\frac{B(S_1A(z_1,z_2),z_2)-S_gA(z_1,z_2)}{B(z_1,z_2)-S_gA(z_1,z_2)}\Bigg).\end{aligned} \tag{3.25}$$

From this pgf it is easy to obtain the steady-state pgf of the number of type-1 or type-2 customers or even the total amount of customers in the queue at random slot bounds

($V(z,1)$, $V(1,z)$, and $V(z,z)$, resp.). The first two generating functions—which we will denote by $V_1(z)$ and $V_2(z)$—are given by the following concise expressions:

$$V_1(z) = (1-\rho)\left(1 + z\frac{S_1A_1(z)-1}{z-S_1A_1(z)}\right)\frac{B(z,1)-B(S_1A_1(z),1)}{B(z,1)-S_gA(z,1)}, \tag{3.26}$$

$$V_2(z) = (1-\rho)\left(1 + z\frac{S_2A_2(z)-1}{z-S_2A_2(z)}\right)\frac{B(S_1A_2(z),z)-S_gA_2(z)}{B(1,z)-S_gA_2(z)}, \tag{3.27}$$

where we adopt the notations $A_1(z) \triangleq A(z,1)$ and $A_2(z) \triangleq A(1,z)$.

We can check our results by considering the case where $A(z_1,z_2) = A(z_1,1)$ holds (i.e., no type-2 customer arrivals are generated). $V(z_1,z_2)$ is then equal to the pgf of the number of customers in a simple single-class queueing system at random slot bounds (see f.i. [10]). The same goes for $A(z_1,z_2) = A(1,z_2)$. Furthermore, if $S_1(z) = S_2(z)$, then $V(z,z)$ once again reduces to the result found in [10], as expected.

4. Decomposition Property

The expressions we found for $V_1(z)$ and $V_2(z)$ bear a great resemblance to the results found for the single class system in f.i. Bruneel [10], be it for the last factor which incorporates the effects of SBP. If we were only interested in the pgfs of v_1 and v_2 and not their joint pgf, the above observation suggests a shortcut in the analysis. In this section we will demonstrate that the decomposition property introduced by Fuhrmann and Cooper [5] for a generalized vacation model in continuous time can be used to determine $V_1(z)$ (and $V_2(z)$) in discrete time as well (see, e.g., page 91 in Takagi [3]). In essence, the decomposition property states that for a generalized vacation system, the number of customers present in the system at the beginning of a random slot is distributed as the sum of two independent random variables. The slot-bound priority rule can be seen as such a vacation system; we can consider service times of type-2 customers as vacations for type-1 customers, and vice versa. The two independent random variables then are the stationary number of type-j customers in the system at the beginning of a random slot when no customers pertaining to other types enter the system (v_j^* with pgf $V_j^*(z)$) and the stationary number of type-j customers in the system at the beginning of a random slot during a vacation period (x_j with pgf $X_j(z)$). The vacations then cover all slots during which no type-j customer is being served.

For the remainder of this section we will concentrate on finding an expression for $V_1(z)$ using this method, as $V_2(z)$ is largely obtained in a similar fashion. The first of the two random variables discussed previously is the stationary number of type-1 customers at the beginning of a random slot, given that there are no arrivals of customers of other types. This drv has a known pgf (see, e.g., Bruneel [10]) and is given by

$$V_1^*(z) = (1-\rho_1)\left(1 + z\frac{S_1A_1(z)-1}{z-S_1A_1(z)}\right). \tag{4.1}$$

Note that this expression—up to a normalization constant—equals the first factor in the right-hand side of (3.26).

Since vacations cover all slots during which no type-1 customer is being served, two scenarios may occur for the second drv. A first is that the randomly chosen slot is an idle

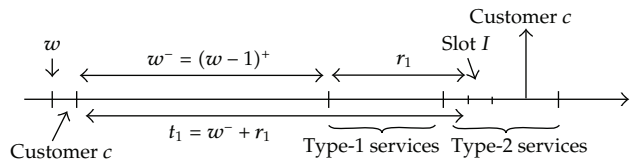

Figure 1: The type-1 customers in the system at the beginning of slot I are accumulated in the queue during the t_1 time slots that it takes customer c to leave the system. Notice that t_1 is stochastically not identical to the delay of a random type-2 customer.

slot, in which case no type-1 customers occupy the system. The second possibility is that the server is serving a type-2 customer, in which case there are x_1^* (with pgf $X_1^*(z)$) type-1 customers occupying the system—that is, x_1^* represents the number of type-1 customers in the system at the beginning of a random slot during which a type-2 customer is being served. Summarizing, the decomposition property leads to the following result:

$$\begin{aligned} V_1(z) &= V_1^*(z)X_1(z), \\ X_1(z) &= \frac{1-\rho}{1-\rho_1} + \frac{\rho_2}{1-\rho_1}X_1^*(z). \end{aligned} \tag{4.2}$$

Note that up until now no features of SBP were used, and hence the remaining unknown pgf $X_1^*(z)$ will characterize SBP. Let slot I be a randomly chosen slot during which a type-2 customer is being served (hereafter called customer c). Then the number of type-1 customers at the beginning of slot I is defined as x_1^* (see Figure 1). Because of the SBP rule, the group being served during slot I does not contain any unserved type-1 customers, and hence x_1^* only contains type-1 customers of groups that have not started their service yet. Since the type-1 customers in the system at the start of slot I entered the system after the arrival of customer c, and none of those customers leaves the system before slot I (since groups are served FCFS), x_1^* can be written as the sum of the number of type-1 customers that arrived during consecutive slots following the arrival slot of customer c (which we know to be a set of i.i.d. drv's). With t_1 representing the time (expressed in slots) ranging from the slot following the arrival of customer c to slot I itself (excluding slot I), and $T_1(z)$ its pgf, we find that

$$x_1^* = \sum_{n=1}^{t_1} a_{1,n} \Longrightarrow X_1^*(z) = T_1 A_1(z). \tag{4.3}$$

The time t_1 is the sum of two drv's, namely, the time it takes for the work pertaining to customers of both types in the system at the beginning of customer c's arrival slot to leave the system (represented by w^- with pgf $W^-(z)$) and the interval starting from the initiation of the service of the group customer c belongs to, until the beginning of slot I (represented by r_1 with pgf $R_1(z)$) (see Figure 1). Since these drv's are mutually independent of one another, $T_1(z)$ is the product of their pgfs (i.e., $T_1(z) = W^-(z)R_1(z)$). First, thanks to the BASTA property (see f.i. Halfin [19]), w^- has the same distribution as the work in the system at

the beginning of a random slot minus one (because we do not count customer c's arrival slot), unless customer c arrives in an empty system. Therefore its pgf is given by

$$W^{-}(z) \triangleq \frac{(1-\rho)(z-1)}{z - A(S_1(z), S_2(z))}. \tag{4.4}$$

Notice that $W^{-}(z)$ is independent of the type of customer c (a consequence of the BASTA property) and that we therefore neglected adding a type subscript to w^{-} and $W^{-}(z)$.

Secondly, obtaining the pgf of r_1 requires a renewal type argument (see f.i. Kleinrock [20]), and it can be checked that the pgf of the number of slots between the service initiation of the group customer c belongs to and slot I is given by

$$R_1(z) \triangleq \frac{A(S_1(z), S_2(z)) - A_1 S_1(z)}{\rho_2(z-1)}. \tag{4.5}$$

From $T_1(z) = W^{-}(z)R_1(z)$, one can find $X_1^*(z)$ using (4.3), which can then be applied to obtain $X_1(z)$ using (4.2). The decomposition result (4.2) then yields expression (3.26) for $V_1(z)$ (in view of the definitions (2.2) and (2.4)).

Notice that $X_1^*(z)$ is a function of $A_1(z)$ and not of z directly. This property however does not hold for $X_2^*(z)$, the pgf of the stationary number of type-2 customers at the beginning of a vacation—where a vacation in this context would be a slot during which no type-2 customer is being served—when searching for $V_2(z)$. Because of the SBP rule, the group customer c (which now represents a random type-1 customer that is being served) belongs to will still have all its type-2 customers. These need to be added to the type-2 customers that entered the system during the t_2 slots following customer c's arrival. Because the former drv is not independent of the latter when correlation exists in the arrival process (i.e., when $A(z_1, z_2) \neq A_1(z_1)A_2(z_2)$), $X_2(z)^*$ cannot be written merely as a function of $A_2(z)$.

Finally we would like to call attention to the fact that this method of analysis could also be applied if instead of two priority classes, customers pertaining to more than two priority classes entered the queue. Naturally only the marginal pgf's could be determined. To calculate the joint pgf of these different types of customers in the queue, the previous slot-to-slot analysis would quickly get cluttered, and other methods should be employed which are not within the scope of this paper (see also De Clercq et al. [2]).

In the next section, among other things that we will observe the tail probabilities will not be dependent on the dominant singularity of $V_1^*(z)$ (or $V_2^*(z)$), but solely on the dominant singularity of $X_1(z)$ and $(X_2(z)$, respectively.

5. Moments and Tail Probabilities

Now that we obtained the joint pgf $V(z_1, z_2)$, we can derive some interesting performance measures concerning v_1 and v_2. First, all moments are derivable from $V(z_1, z_2)$ using

the moment generating property of pgf's. As an illustration we show the first moments below. The second derivatives of the arrival process are defined as $\lambda_{ij} = E[a_i a_j]$:

$$E[v_1] = S_g''(1)\frac{\lambda_1(1-A(0,0))}{2(1-\rho)} + S_1'(1)\frac{\lambda_{11}+\lambda_1}{2},$$
$$E[v_2] = S_g''(1)\frac{\lambda_2(1-A(0,0))}{2(1-\rho)} + S_2'(1)\frac{\lambda_{22}+\lambda_2}{2} + \lambda_{12}S_1'(1). \tag{5.1}$$

The first equation, for example, is comparable to the first moment found for the number of customers in a single-class system (e.g., when $A(z_1, z_2) = A(z_1, 1)$), although the concept of group service times comes across somewhat strange in this setting. Also, the third term in the second equation reflects the effect of SBP on the low-priority customers.

Furthermore, we derive the tail probabilities of v_1, v_2, and even $v_1 + v_2$ using the dominant pole approximation technique (see f.i. Van Mieghem [21]). Therefore, suppose that the dominant singularities of $A(z_1, z_2)$, $S_1(z)$, and $S_2(z)$ are all poles (in contrast to branch points). Then from Vivanti's theorem ([22] page 1254) we know they are real and positive. Furthermore, because pgfs are always analytic inside and well defined on the unit disk, these singularities must have a modulus larger than 1. Since $V(z_1, z_2)$ is a rational function of $A(z_1, z_2)$ and both $S_j(z)$, its dominant singularities z_{v_1}, z_{v_2}, and z_{v_T} (singularities of $V(z, 1)$, $V(1, z)$, and $V(z, z)$, resp.) are poles as well. Hence, for high enough k we can very accurately approximate, for example, $\Pr[v_1 = k]$ calculating only z_{v_1} and its residue as shown in [21]. We start by determining z_{v_1}, dominant pole of $V_1(z)$.

Clearly, z_{v_1} will either be the dominant pole of $B(z, 1)$, $S_1A(z, 1)$, $B(S_1A(z, 1), 1)$, or $S_gA(z, 1) \equiv B(S_1A(z, 1), S_2A(z, 1))$ or be a zero of $z - S_1A(z)$, or $B(z, 1) - S_gA(z, 1)$ with modulus larger than 1—whichever has lowest modulus.

Let R_{A_j} be the radius of convergence of $S_jA(z, 1)$, and $R_A \triangleq \min(R_{A_1}, R_{A_2})$. Since $S_jA(z, 1) > A(z, 1) > z$ for $z \in]1, R_A[$, one can easily deduce that, of the functions $B(z, 1)$, $S_jA(z, 1)$, $B(S_1A(z, 1), 1)$, and $S_gA(z, 1)$, $S_gA(z, 1)$ is the one with the smallest radius of convergence, which will be represented by R_B. In particular, if we denote by R_{a_1} the radius of convergence of $A(z, 1)$ (and of $B(z, 1)$), the inequality $R_B \leq R_{a_1}$ must hold.

As for the zeros of the denominators in (3.26), note that all zeros of $z - S_1A(z)$ are zeros of $B(z, 1) - B(S_1A(z, 1), 1)$ in the numerator as well. Hence we only focus on the zeros of $B(z, 1) - S_gA(z, 1)$. As a result of $R_B \leq R_{a_1}$ and the equilibrium condition we find that this last denominator has exactly one zero in the region $]1, R_B[$ as shown in [23]. Therefore this zero is our dominant pole z_{v_1} we have been searching. Calculating z_{v_1} comes down to finding a solution to the equation $y_0 = A(S_1(y_0), S_2(y_0))$—other than $y_0 = 1$—and solving $y_0 = A(z_{v_1}, 1)$.

The dominant pole approximation then becomes

$$\Pr[v_1 = n] \approx -\theta_1 z_{v_1}^{-n-1}, \tag{5.2}$$

where θ_1 is the residue of $V_1(z)$ at $z = z_{v_1}$ $(= \lim_{z \to z_{v_1}} (z - z_{v_1})V_1(z))$.

Analogous to the previous result we can derive that z_{v_2} is the zero of $B(1, z) - S_gA(1, z)$, and this zero can be calculated from $y_0 = A(1, z_{v_2})$. Finally z_{v_T} is found in a similar way as it satisfies $y_0 = A(z_{v_T}, z_{v_T})$.

When one or more of the dominant singularities in either $A(z_1, z_2)$, $S_1(z)$, or $S_2(z)$ is a branch point, we face a different story. It is no longer certain that $B(z, 1) - S_gA(z, 1)$ has a

zero in $]1, R_C[$ in the case that the smallest singularity of $S_g A(z,1)$ is a branch point. We can use the following criterium (see f.i. Steyaert [23]) to determine whether there is a zero, which can be used in the dominant pole approximation discussed previously:

$$\lim_{z \to R_C^-} \frac{S_g A(z,1)}{B(z,1)} > 1. \tag{5.3}$$

If the above is true, then the dominant pole approximation holds. Otherwise no zero is found in the interval $]1, R_C[$ and hence the dominant singularity is a branch point at R_C, which calls for a case-by-case analysis of the tail behaviour and falls outside the scope of the current paper.

6. Numerical Examples

There are a lot of different parameters incorporated in this model. To get some insight on how v_1 and v_2 will react for various arrival and service-time distributions, we propose an example with a limited number of parameters that appeal to our intuition. We therefore consider

$$\begin{aligned} A(z_1, z_2) &= e^{\alpha(pz_1 + qz_2 + rz_1z_2 - 1)}, \\ S_j(z) &= \frac{z}{z - E[s_j](z-1)}. \end{aligned} \tag{6.1}$$

We choose service times with a geometric distribution and a Poisson arrival process (parameter α), in which each arrival instance generates a type-1 customer with probability p, a type-2 customer with probability q, and customers of both types with probability r. Needless to say $p + q + r = 1$. A useful comparison of our proposed priority rule will include that of total priority or HoL priority (e.g., studied in Walraevens et al. [11]).

In a first graph we choose $p = q = 0$ (and consequently $r = 1$), and $E[s_1] = E[s_2] = 2$. This concretely means that both customer types are indifferentiable concerning their respective service times and always enter the system in pairs. We increase the workload ρ by increasing α and observe its effect on the average buffer content $E[v_1]$ and $E[v_2]$. The resulting graphs can be found in Figure 2 together with those for nonpreemptive HoL priority (abbreviated $E[v_{1,\mathrm{HoL}}]$ and $E[v_{2,\mathrm{HoL}}]$).

Even though we have chosen a symmetric arrival process (i.e., $A(z_1, z_2) = A(z_2, z_1)$) there seems to be a difference between $E[v_1]$ and $E[v_2]$, one that can only be attributed to the presence of priority. By comparing with HoL priority we observe that for low loads, $E[v_j] \approx E[v_{j,\mathrm{HoL}}]$, while for high loads the difference $E[v_2] - E[v_1]$ becomes almost negligible compared to their respective absolute values. The former is a consequence of the fact that for low loads the queue content is largely dominated by the active group's content, since hardly any additional groups get queued up. The latter we can clarify by pointing out that the probability that the server is busy when a random group arrives is ρ. The higher ρ, the more queueing of different groups occurs, and thus the queue content will be dominated by the content of the successive groups that are queued. In case each group counts on average the same amount of type-1 customers as type-2 ones, (due to $\lambda_1 = \lambda_2$) the difference $E[v_2] - E[v_1]$ will be solely the result of the active group's content—which is actually not the same as a

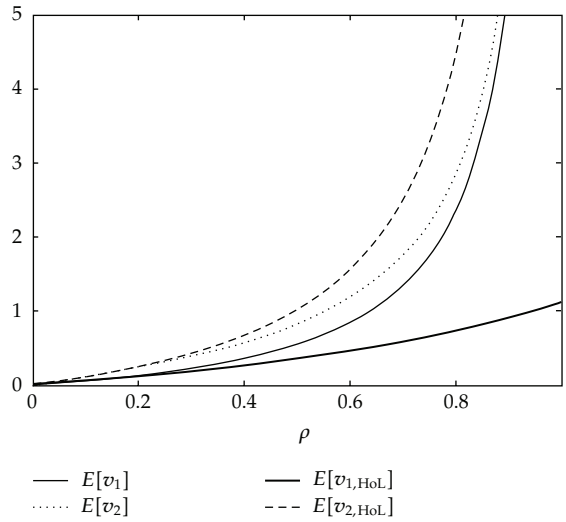

Figure 2: $E[v_j]$ and $E[v_{j,\mathrm{HoL}}]$ versus ρ with $r = 1$, $E[s_j] = 2$ and α variable.

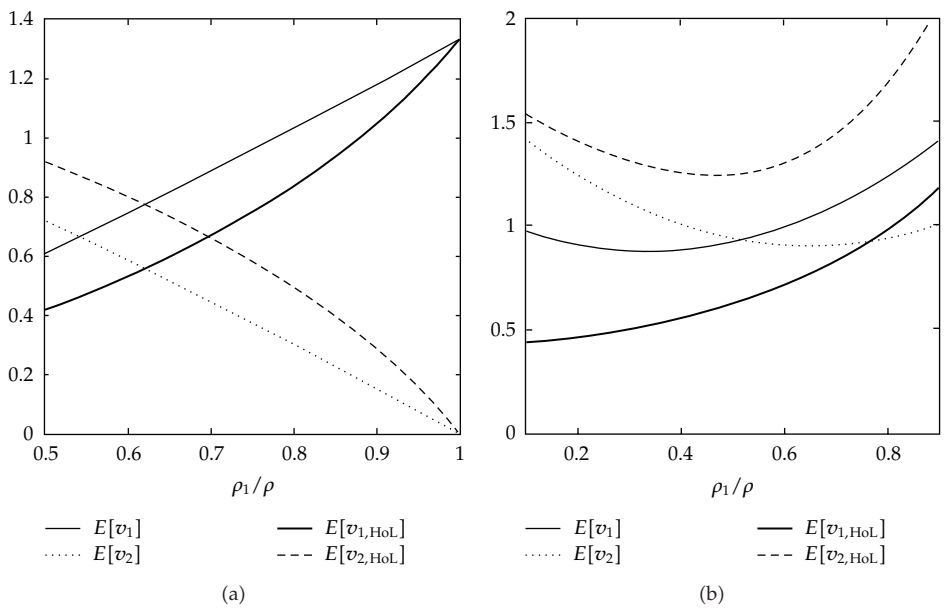

Figure 3: (a) $E[v_j]$ and $E[v_{j,\mathrm{HoL}}]$ versus ρ_1/ρ with $r = 0$, $E[s_j] = 1$ and $\rho = \alpha = 2/3$, p and q being variable. (b) $E[v_j]$ and $E[v_{j,\mathrm{HoL}}]$ versus ρ_1/ρ with $p = q = 0.5$, $E[s_1] + E[s_2] = 10$ and $\rho = 5\alpha = 2/3$, both $E[s_j]$ being variable ($E[s_j] \geq 1$).

random group's content. Lastly, and not surprisingly $\rho = 1$ is an asymptote, in which case the queue content will evolve to infinity in steady state.

Next, we examine what happens to the average type-1 and type-2 population for a fixed load ρ, when we increase ρ_1 (and hence decrease ρ_2 because $\rho_1+\rho_2 = \rho$). We can do this in two ways: by varying the class-specific arrival rates (i.e., by adjusting p and q) or by varying $E[s_1]$ and $E[s_2]$. The parameters used to plot the graphs in Figure 3 are displayed in its

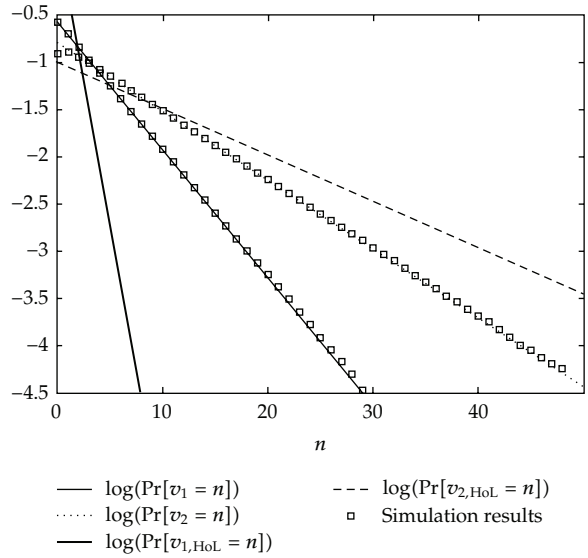

Figure 4: Dominant pole approximation of the probability distribution of both v_j and $v_{j,\mathrm{HoL}}$ on a logarithmic scale with $p = 0.2$, $q = 0.6$ and $r = 0.2$ so that twice as many type-2 customers enter the system as type-1 customers. $E[s_j] = 2$ and $\rho = 9/10$.

caption. Even though the graphs are very different from one another, we point out that in both graphs, when $\rho_1 = \rho_2$, $E[v_1] < E[v_2]$ because of the SBP priority, more so even for absolute priority. In the first graph we vary the arrival rates, and as the customer composition rises in favor of type-1 customers, we observe that $E[v_1] > E[v_2]$ for values of ρ_1 for which still $E[v_{1,\mathrm{HoL}}] < E[v_{2,\mathrm{HoL}}]$. This is an illustration of the more moderate form of priority assignment by SBP. In the second figure in Figure 3(b), the class specific arrival rates are kept constant and equal to one another. The graph plots the average type-1 and type-2 customers versus ρ_1/ρ, by increasing the type-1 service times (and reducing type-2 service times). As $E[s_1]$ increases, the average type-2 customer population increases as well for HoL priority, even though their service time decreases. This is caused by type-1 service times. As such SBP is clearly the better option: type-2 customers do get served before some type-1 customers.

Lastly Figure 4 shows an approximation of $\Pr[v_j = n]$ on a logarithmic scale together with some dots representing simulation results. Approximations that were found for $\Pr[v_{j,\mathrm{HoL}}]$ in Walraevens et al. [24] were used to compare against HoL priority. The number of type-1 and type-2 customers entering the system during the same slot is slightly correlated, and twice as many type-2 customers enter the system as type-1 customers on average. The exact parameters can be found in the figures caption. Clearly, the dominant-pole approximation described in Section 4 constitutes an efficient and accurate method to calculate the queue content distribution of both types of customers.

7. Conclusion

In a dual-class queueing system in discrete time under stochastic equilibrium, we derived expressions for the joint pgf of the number of type-1 and type-2 customers in the queue when the SBP rule is used as a server discipline. We obtained this after a slot-to-slot analysis using

a carefully chosen Markov chain. More concretely we introduced the notion of a "group" of customers, which could be looked upon as classless entities entering and leaving our system, on basis of which we could more easily carry out the analysis. The first moments and tail probabilities were explicitly calculated as well. Moreover, some examples made the effect of SBP clear, comparing it to HoL priority, in which the most important result stated that SBP behaves as FCFS (no difference between the ways customers of different classes are treated) for high workloads while it behaves more as HoL priority for lower loads. Also, our results show that a dominant pole approximation for calculating the queue content distribution of both types of customers is both efficient and accurate.

References

[1] D. Fiems, J. Walraevens, and H. Bruneel, "Performance of a partially shared priority buffer with correlated arrivals," in *Proceedings of the 20th International Teletraffic Congress (ITC '07)*, vol. 4516 of *Lecture Notes in Computer Science*, pp. 582–593, 2007.

[2] S. De Clercq, B. Steyaert, and H. Bruneel, "Analysis of a multi-class discrete-time queueing system under the slot-Bound priority rule," in *Proceedings of the 6th St. Petersburg Workshop on Simulation*, pp. 827–832, 2009.

[3] H. Takagi, *Analysis of Polling Systems*, MIT Press, 1986.

[4] O. Boxma, J. Bruin, and B. Fralix, "Sojourn times in polling systems with various service disciplines," *Performance Evaluation*, vol. 66, no. 11, pp. 621–639, 2009.

[5] S. W. Fuhrmann and R. B. Cooper, "Stochastic decompositions in the $M/G/1$ queue with generalized vacations," *Operations Research*, vol. 33, no. 5, pp. 1117–1129, 1985.

[6] J. G. Shanthikumar, "On stochastic decomposition in $M/G/1$ type queues with generalized server vacations," *Operations Research*, vol. 36, no. 4, pp. 566–569, 1988.

[7] F. Ishizaki, "Decomposition property in a discrete-time queue with multiple input streams and service interruptions," *Journal of Applied Probability*, vol. 41, no. 2, pp. 524–534, 2004.

[8] I. Stavrakakis, "Delay bounds on a queueing system with consistent priorities," *IEEE Transactions on Communications*, vol. 42, no. 2, pp. 615–624, 1994.

[9] S. Ndreca and B. Scoppola, "Discrete time GI/Geom/1 queueing system with priority," *European Journal of Operational Research*, vol. 189, no. 3, pp. 1403–1408, 2008.

[10] H. Bruneel, "Performance of discrete-time queueing systems," *Computers and Operations Research*, vol. 20, no. 3, pp. 303–320, 1993.

[11] J. Walraevens, B. Steyaert, and H. Bruneel, "Performance analysis of the system contents in a discrete-time non-preemptive priority queue with general service times," *Belgian Journal of Operations Research, Statistics and Computer Science*, vol. 40, no. 1-2, pp. 91–103, 2000.

[12] S. De Clercq, B. Steyaert, and H. Bruneel, "Delay analysis of a discrete-time multiclass slot-bound priority system," *A Quarterly Journal of Operations Research*, vol. 10, no. 1, pp. 67–79, 2012.

[13] H. Masuyama and T. Takine, "Analysis and computation of the joint queue length distribution in a FIFO single-server queue with multiple batch Markovian arrival streams," *Stochastic Models*, vol. 19, no. 3, pp. 349–381, 2003.

[14] T. Takine, "Queue length distribution in a FIFO single-server queue with multiple arrival streams having different service time distributions," *Queueing Systems. Theory and Applications*, vol. 39, no. 4, pp. 349–375, 2001.

[15] Q.-M. He, "Queues with marked customers," *Advances in Applied Probability*, vol. 28, no. 2, pp. 567–587, 1996.

[16] V. G. Kulkarni and K. D. Glazebrook, "Output analysis of a single-buffer multiclass queue: FCFS service," *Journal of Applied Probability*, vol. 39, no. 2, pp. 341–358, 2002.

[17] F. Ishizaki, T. Takine, and T. Hasegawa, "Analysis of a discrete-time queue with gated priority," *Performance Evaluation*, vol. 23, no. 2, pp. 121–143, 1995.

[18] B. Van Houdt and C. Blondia, "The delay distribution of A type k customer in a first-come-first-served MMAP[K]/PH[K]/1 queue," *Journal of Applied Probability*, vol. 39, no. 1, pp. 213–223, 2002.

[19] S. Halfin, "Batch delays versus customer delays," *The Bell System Technical Journal*, vol. 62, no. 7, pp. 2011–2015, 1983.

[20] L. Kleinrock, *Queueing Systems, Volume I: Theory*, Wiley, New York, NY, USA, 1975.

[21] P. Van Mieghem, "The asymptotic behavior of queueing systems: large deviations theory and dominant pole approximation," *Queueing Systems: Theory and Applications*, vol. 23, no. 1–4, pp. 27–55, 1996.
[22] K. Ito, *Encyclopedic Dictionary of Mathematics: the Mathematical Society of Japan (2 Vol. Set)*, MIT Press, 2000.
[23] B. Steyaert, *Analysis of generic discrete-time buffer models with irregular packet arrival patterns [Ph.D. thesis]*, 2008.
[24] J. Walraevens, B. Steyaert, and H. Bruneel, "Performance analysis of a single-server ATM queue with a priority scheduling," *Computers and Operations Research*, vol. 30, no. 12, pp. 1807–1829, 2003.