A Mixed Discrete-Time Delay/Retrial Queueing Model for Handover Calls and New Calls Competing for a Target Channel

Rein Nobel

Abstract To study the performance of handover calls approaching a target cell in combination with arrivals of new calls competing for the same cell, a mixed discrete-time delay/retrial model with one server and with priorities for the delayed customers is discussed. The handover calls are modeled as high-priority customers and the new calls as low-priority customers. The priority is non-preemptive. Upon arrival high-priority customers are put in a queue which is served on a first come first served basis. The behavior of the low-priority customers is modeled as in a retrial queue. Arrivals are in batches and all customers are served individually according to generally distributed and independent service times. The joint steady-state distribution of the queue length of the high priority customers and the orbit size of the low-priority customers is studied using probability generating functions. Several performance measures will be calculated, such as the mean queue length of the handover calls and the orbit size of the new calls. Also the covariance between the queue length and the orbit size will be studied, among others.

Keywords Handover calls · Discrete-time retrial queue · Priority customers · Generating functions

1 Introduction

In mobile telephony the problem how to handle handover calls is a important topic. When a mobile phone user is moving from one cell [the source] to another cell [the target] then his ongoing call has to be switched from the channel of the source cell to a channel of the target cell. Because neighboring cells cover overlapping regions, usually a so-called soft handover protocol is used, i.e. the ongoing call joins a queue,

R. Nobel (✉)
Department of Econometrics, Vrije University, Amsterdam, The Netherlands
e-mail: r.d.nobel@vu.nl

© Springer International Publishing Switzerland 2016

T.V. Do et al. (eds.), *Queueing Theory and Network Applications*,
Advances in Intelligent Systems and Computing 383,
DOI: 10.1007/978-3-319-22267-7_16

waiting for a free channel at the target cell, but the call continues to use the channel of the source cell until a channel at the target cell is available. Meanwhile also new calls will try to get a free channel at the target cell. To avoid unnecessary interruptions of the ongoing calls waiting for a free channel at the target cell, priority is given to the handover calls over the new calls. When all channels are busy new calls are temporarily rejected and the new calls have to be initiated anew some time later.

To model this protocol of soft handover calls at a target cell in combination with the arrival of new calls at this cell we study a mixed delay/retrial model in discrete time with one server [channel]. More specifically, we consider a one-server queueing model in discrete time with two types of customers. Time is divided in slots, and all events [arrivals, start of a service and departures] are considered to occur at the slot boundaries only. The high-priority customers [handover calls] arrive in batches following a general probability distribution. Upon arrival a batch of high-priority customers is put in a waiting line from which the customers are served one by one on a first come first served basis. The low-priority customers [new calls] also arrive in batches (primary arrivals), possibly following a different probability distribution, and when upon arrival a batch of low-priority customers sees the server busy, all incoming low-priority customers are sent into orbit, a virtual waiting space from which they will try to reenter the system individually some random time later (secondary arrivals). The service times of the high-priority and the low-priority customers are all independent and follow [possibly] a different general distribution. To resolve the conflict of simultaneous arrivals and departures we have chosen for the *late arrival set up with delayed access*, i.e. arrivals have precedence over departures and a service of newly arrived customers can only start at the time slot following the slot of the arrival at the earliest. Also the modeling assumption is made that the time slot after *any* departure the server always stays idle, even when high-priority customers are waiting in line. A new high-priority customer will start service the next slot when the queue of high-priority customers is not empty or a batch of high-priority customers will have arrived during the idle slot. In that case all possibly arrived low-priority customers are sent (back) into orbit. Otherwise, i.e. no high-priority customers present at the end of the idle slot, the server starts the service of a low-priority customer, randomly chosen from the mixed batch of primary and secondary low-priority arrivals. When neither high-priority customers are present at the end of the idle slot, nor low-priority customers will have arrived during the idle slot, the server stays idle also the following slot. All customers are served one by one, and in case a low-priority customer is taken into service all other primary and secondary low-priority customers having arrived in the same slot are sent (back) into orbit.

As is well-known *retrial models* have received much less attention in the literature than the well-known queueing models such as delay-models and loss-models, mainly because the arrival stream of the customers consists of two types, the primary arrivals who enter for the first time, and the secondary arrivals from the orbit, making the 'arrival intensity' dependent of the number of customers in the orbit. Also overtaking takes place, i.e. customers are not served according to a specific queueing discipline, which severely complicates the study of the waiting-time distribution of a customer, here defined as the total time that the customer spends in the orbit. It is probably fair

to say that the unpopularity of the research on retrial models is partly due to their intractability, because from a practical point of view retrial models often describe a more realistic picture of many queueing situations than any of the other type of models. Notwithstanding the mathematical difficulties encountered in the study of retrial systems some models, with the $M/G/1$ retrial queue in a prominent position, have been analyzed thoroughly, and we refer to the monographs of FALIN AND TEMPLETON [4] and ARTALEJO AND GÓMEZ-CORRAL [1] for an overview of the main results.

Although most papers on retrial queues discuss models in continuous time, as a consequence of the revolutionary developments in the computer and telecommunication technology, at the end of the past century people started to study also retrial models in discrete time. LI AND YANG [5], [6] and [9] made a start. NOBEL AND MORENO [8] were the first to study a discrete-time classical retrial queueing model with the so-called *late-arrival* setup, i.e. precedence is given to arrivals over departures. We recall that in a classical retrial model an idle server accepts exactly one customer for service from the batch of all the incoming customers [a mixture of primary customers and customers arriving from the orbit] and sends all the other newly arrived customers (back) to the orbit. As a consequence of the late-arrival setup, after a departure the server always stays idle for at least one time slot, due to the fact that the most recently arrived customers have seen the server still busy and therefore they have been sent into the orbit.

In this paper we will extend the classical discrete-time one-server retrial model of NOBEL AND MORENO [8] by adding a second type of customers [the handover calls] who will be put in a queue and are served one by one on a first come first served basis. These customers are given non-preemptive priority over the original customers [the new calls] who continue to act as retrial customers. In a previous paper (NOBEL AND MORENO [7]) the high-priority customers were lost when upon arrival they found the server busy. A model similar to our delay/retrial model has been studied in CHOI AND KIM [2], but they discuss only single arrivals and all customers follow the same service-time distribution. Further, they have chosen the early arrival setup. A continuous-time retrial model with priority customers has been studied by FALIN, ARTALEJO AND MARTIN [3], but in that paper only single arrivals have been considered. The model discussed in this paper can be seen both as an extension and as the discrete-time counterpart of that model.

We will study the joint steady-state distribution of the length of the queue of high-priority customers and the size of the orbit with low-priority customers. Not surprisingly, the mathematical analysis of our mixed delay/retrial model differs greatly from the analysis of the models discussed in the papers [2], [7] and [8].

Firstly, we will derive the generating function of the joint steady-state distribution of the number of low-priority customers in orbit, the number of high-priority customers in the queue and the residual service time of the customer in service [either a high-priority customer, or a low-priority customer]. This generating function will be used to calculate several performance measures, e.g. the mean queue length, the mean orbit size and the covariance of the queue length and the orbit size. In Section 2 we describe the model in detail. Section 3 discusses the steady-state distributions of

the orbit size and the queue length, among others. In Section 4 we derive an expression for the mean busy period. Numerical results will be presented in a forthcoming extended version of this paper.

2 Description of the Model

For a detailed description of the discrete-time setup with late arrivals and delayed access [LAS/DA] we refer to NOBEL AND MORENO [8]. Recall that in the classical retrial model the time slot after a departure the server always stays idle for at least one slot, due to the late-arrival setup with delayed access. For the mixed delay/retrial model to be discussed in this paper we make the technical assumption that the slot following a departure the server always stays idle, *also in case high-priority customers are waiting in the queue*. We can interpret this idle slot as a preparation time for the next service, but we admit that the main reason to include this idle slot following a departure is to enable tractability: a small price to pay for a deeper insight into this mixed delay/retrial model with priorities for the delayed customers.

We will now give the precise description of our discrete-time mixed delay/retrial queueing model with one server and priorities. During each time slot high-priority customers arrive in batches. The batch sizes are mutually independent and follow a general probability distribution $\left\{a_i^{(H)}\right\}_{i=0}^{\infty}$ with probability generating function (p.g.f.)

$$\mathcal{A}_H(y) = \sum_{i=0}^{\infty} a_i^{(H)} y^i.$$

In every time slot also low-priority customers arrive in batches. These batch sizes follow a general probability distribution $\left\{a_k^{(L)}\right\}_{k=0}^{\infty}$ with p.g.f.

$$\mathcal{A}_L(z) = \sum_{k=0}^{\infty} a_k^{(L)} z^k.$$

These batch sizes are again mutually independent and they are also independent of the batch sizes of the high-priority customers. We call these arrivals primary arrivals. Each individual high-priority customer requires a service time, measured as a number of time slots, which follows the discrete probability distribution $\left\{b_j^{(H)}\right\}_{j=1}^{\infty}$ with p.g.f.

$$\mathcal{B}_H(w) = \sum_{j=1}^{\infty} b_j^{(H)} w^j.$$

Similarly, every low-priority customer requires a generally distributed service time with distribution $\left\{b_j^{(L)}\right\}_{j=1}^{\infty}$ and p.g.f.

$$\mathcal{B}_L(w) = \sum_{j=1}^{\infty} b_j^{(L)} w^j.$$

All service times are mutually independent and they are also independent of the batch sizes of the arriving customers. A service time requires at least one time slot, so $b_0^{(H)} = b_0^{(L)} = 0$. As said before, the high-priority customers are placed in a queue and the high-priority customers are served individually on a first come first served basis [within a batch in random order]. Low-priority customers behave as the customers in the classical retrial queue, with the only difference that all incoming low-priority customers [primary and secondary arrivals] are *also* sent into orbit when high-priority customers are present in the queue or arrive simultaneously, i.e. in the same slot, with the low-priority customers. In each time slot low-priority customers try to reenter the system individually and independently with the so-called retrial probability r $[0 < r < 1]$.

We are interested in the steady-state behavior of the number of high-priority customers in the queue, the number of low-priority customers in orbit and the residual service time of the customer currently in service. To analyze the mixed delay/retrial queueing model, we define a discrete-time Markov chain (DTMC) by observing the system at the epochs $k-$, that is at the start of the time slots k just after, possibly, a service of a (low- or high-priority) customer has started, but before the arrivals during time slot k have occurred. We define the following random variables,

H_k = the residual service time of the [high- or low-priority] customer in service at time $k-$,

L_k = the number of high-priority customers present in the queue at time $k-$,

Q_k = the number of low-priority customers in orbit at time $k-$.

We define $H_k = 0$ when at epoch $k-$ the server is idle. Then, due to the independencies stated in the description of the model, the stochastic process $\{(H_k, L_k, Q_k) : k = 0, 1, 2, \ldots\}$ is an irreducible aperiodic DTMC and under the stability condition that

$$\mathcal{A}'_H(1)[\mathcal{B}'_H(1) + 1] + \mathcal{A}'_L(1)[\mathcal{B}'_L(1) + 1] < 1$$

it is positive recurrent. A formal proof of this stability condition can be given using Foster's criterion [see NOBEL AND MORENO [8] for the details]. Notice the '+1' added to the mean service times $\mathcal{B}'_H(1)$ and $\mathcal{B}'_L(1)$, due to our technical assumption that after *every* departure the server stays idle for at least one time slot.

3 The Joint Distribution of Queue Length and Orbit Size

In this section we will derive the joint generating function of the steady-state distribution of the DTMC $\{(H_k, L_k, Q_k) : k = 0, 1, 2, \ldots\}$. Under the stability condition

we can define the following limiting joint distribution of this DTMC

$$\pi(j,m,n) = \lim_{k\to\infty} \mathrm{P}(H_k = j; L_k = m; Q_k = n), \qquad j,m,n = 0,1,2,\ldots,$$

with its associated three-dimensional generating function

$$\Pi(w,y,z) = \sum_{j=0}^{\infty}\sum_{m=0}^{\infty}\sum_{n=0}^{\infty} \pi(j,m,n) w^j y^m z^n.$$

In the following it is convenient to introduce also the partial generating functions,

$$\Pi_{jm}(z) = \sum_{n=0}^{\infty} \pi(j,m,n) z^n \quad \text{and}$$

$$\Pi_j(y,z) = \sum_{m=0}^{\infty}\sum_{n=0}^{\infty} \pi(j,m,n) y^m z^n = \sum_{m=0}^{\infty} \Pi_{jm}(z) y^m.$$

To find the p.g.f. $\Pi(w,y,z)$ we write down the system of balance equations,

$$\begin{aligned}
\pi(0,m,n) &= \mathbf{I}_{\{m=0\}} a_0^{(H)} a_0^{(L)} (1-r)^n \pi(0,0,n) + \\
&\quad \sum_{i=0}^{m} a_i^{(H)} \sum_{k=0}^{n} a_k^{(L)} \pi(1,m-i,n-k), \qquad (1)\\
&\quad m,n = 0,1,2,\ldots,\\
\pi(j,m,n) &= \sum_{i=0}^{m} a_i^{(H)} \sum_{k=0}^{n} a_k^{(L)} \pi(j+1,m-i,n-k)\\
&\quad + b_j^{(H)} \sum_{i=0}^{m+1} a_i^{(H)} \sum_{k=0}^{n} a_k^{(L)} \pi(0,m+1-i,n-k)\\
&\quad + \mathbf{I}_{\{m=0\}} b_j^{(L)} a_0^{(H)} \left\{ \sum_{k=1}^{n+1} a_k^{(L)} \pi(0,0,n+1-k) \right.\\
&\quad \left. + a_0^{(L)} \left(1-(1-r)^{n+1}\right) \pi(0,0,n+1) \right\}. \qquad (2)\\
&\quad j = 1,2,\ldots; \quad m,n = 0,1,2,\ldots.
\end{aligned}$$

Notice how our technical assumption that after *any* departure the server stays idle for at least one time slot plays its role in these balance equations. This assumption enforces more parallelism between the [services of] low-priority customers and high-priority customers. Below we will see that only due to this enforced parallelism our analysis can be pursued successfully.

From equations (1) and (2) we get by multiplying both sides with z^n and summing over $n = 0, 1, \ldots$, and subsequently multiplying both sides of the result by y^m and summing over $m = 0, 1, \ldots$,

$$\Pi_0(y,z) = a_0^{(H)} a_0^{(L)} \Pi_{00}((1-r)z) + \mathcal{A}_H(y)\mathcal{A}_L(z)\Pi_1(y,z), \tag{3}$$

$$\Pi_j(y,z) = \mathcal{A}_H(y)\mathcal{A}_L(z)\Pi_{j+1}(y,z) + \frac{b_j^{(H)}}{y}\mathcal{A}_L(z)\left[\mathcal{A}_H(y)\Pi_0(y,z) - a_0^{(H)}\Pi_{00}(z)\right] + \frac{b_j^{(L)} a_0^{(H)}}{z}\left[\mathcal{A}_L(z)\Pi_{00}(z) - a_0^{(L)}\Pi_{00}((1-r)z)\right]. \tag{4}$$

Next, multiplying equation (4) by w^j and summing over $j = 1, 2, \ldots$ gives after some simple algebra, using equation (3) to get rid of $\Pi_1(y,z)$,

$$\begin{aligned} yz\left[w - \mathcal{A}_H(y)\mathcal{A}_L(z)\right]\Pi(w,y,z) &= \mathcal{A}_H(y)\mathcal{A}_L(z)z\left[w\mathcal{B}_H(w) - y\right]\Pi_0(y,z) \\ &\quad + a_0^{(H)}\mathcal{A}_L(z)w\left[y\mathcal{B}_L(w) - z\mathcal{B}_H(w)\right]\Pi_{00}(z) \\ &\quad + a_0^{(H)}a_0^{(L)}wy\left[z - \mathcal{B}_L(w)\right]\Pi_{00}((1-r)z). \end{aligned} \tag{5}$$

So, the problem is to find the unknown partial generating functions $\Pi_0(y,z)$ and $\Pi_{00}(z)$. Firstly, take $w = \mathcal{A}_H(y)\mathcal{A}_L(z)$ in (5) to make the left-hand side zero. This gives

$$\begin{aligned} \Pi_0(y,z) &= a_0^{(H)}\frac{\mathcal{A}_L(z)\left[y\mathcal{B}_L(\omega(y,z)) - z\mathcal{B}_H(\omega(y,z))\right]\Pi_{00}(z)}{z\left[y - \omega(y,z)\mathcal{B}_H(\omega(y,z))\right]} \\ &\quad + a_0^{(H)}\frac{a_0^{(L)}y\left[z - \mathcal{B}_L(\omega(y,z))\right]\Pi_{00}((1-r)z)}{z\left[y - \omega(y,z)\mathcal{B}_H(\omega(y,z))\right]} \end{aligned} \tag{6}$$

where $\omega(y,z) := \mathcal{A}_H(y)\mathcal{A}_L(z)$. Now for any z with $|z| \le 1$ let $w = \phi(z)$ be a solution of the system of equations

$$\begin{cases} w = \mathcal{A}_H(y)\mathcal{A}_L(z) \\ y = w\mathcal{B}_H(w) \end{cases} \iff \begin{cases} w = \mathcal{A}_H(w\mathcal{B}_H(w))\mathcal{A}_L(z) \\ y = w\mathcal{B}_H(w). \end{cases}$$

For real z with $0 < z < 1$ it is easy to see that there is a unique real solution $w = \phi(z) \in (0,1)$ and further that $\phi(1) = 1$. So we have for z with $|z| \le 1$

$$\phi(z) = \mathcal{A}_H(\phi(z)\mathcal{B}_H(\phi(z)))\mathcal{A}_L(z) \tag{7}$$

from which we can calculate the derivative $\phi'(z)$ by implicit differentiation. For future use we give the result

$$\phi'(z) = \frac{\mathcal{A}_H(\phi(z)\mathcal{B}_H(\phi(z)))\mathcal{A}_L'(z)}{1 - \mathcal{A}_H'(\phi(z)\mathcal{B}_H(\phi(z)))[\mathcal{B}_H(\phi(z)) + \phi(z)\mathcal{B}_H'(\phi(z))]\mathcal{A}_L(z)}. \tag{8}$$

From equation (6) we get [notice that now $y = \phi(z)\mathcal{B}_H(\phi(z))$ and $\omega(y, z) = \phi(z)$]

$$\Pi_{00}(z) = a_0^{(L)} \frac{\phi(z)\,[z - \mathcal{B}_L(\phi(z))]}{\mathcal{A}_L(z)\,[z - \phi(z)\mathcal{B}_L(\phi(z))]} \Pi_{00}((1-r)z). \tag{9}$$

Introduce (see also NOBEL AND MORENO [8]) the *retrial function*

$$\mathcal{R}(z) := a_0^{(L)} \frac{\phi(z)\,[z - \mathcal{B}_L(\phi(z))]}{\mathcal{A}_L(z)\,[z - \phi(z)\mathcal{B}_L(\phi(z))]}.$$

We see that $\mathcal{R}(0) = 1$ and after some calculation, using L'Hôpital and result (8) we find

$$\mathcal{R}(1) = a_0^{(L)} \frac{1 - \mathcal{A}'_H(1)[\mathcal{B}'_H(1) + 1] - \mathcal{A}'_L(1)\mathcal{B}'_L(1)}{1 - \mathcal{A}'_H(1)[\mathcal{B}'_H(1) + 1] - \mathcal{A}'_L(1)[\mathcal{B}'_L(1) + 1]}.$$

Notice that in the denominator the stability condition shows up. Rewriting equation (9) gives via iteration

$$\begin{aligned}\Pi_{00}(z) = \mathcal{R}(z)\Pi_{00}((1-r)z) = \mathcal{R}(z)\mathcal{R}((1-r)z)\Pi_{00}\left((1-r)^2 z\right) = \cdots \\ = \prod_{i=0}^{n-1} \mathcal{R}((1-r)^i z)\Pi_{00}\left((1-r)^n z\right), \end{aligned} \tag{10}$$

and now, sending n to infinity, we get

$$\Pi_{00}(z) = \prod_{i=0}^{\infty} \mathcal{R}((1-r)^i z)\Pi_{00}(0) \tag{11}$$

For the technique to prove the convergence of the infinite product $\prod_{i=0}^{\infty} \mathcal{R}((1-r)^i z)$ we refer to [8]. So, our next problem is to calculate $\Pi_{00}(0)$. From equation (11) we see that is it sufficient to calculate $\Pi_{00}(1-r)$. We plug the result (9) in equation (6). This gives

$$\begin{aligned}\Pi_0(y, z) = a_0^{(H)} \frac{\mathcal{A}_L(z)\,[y\mathcal{B}_L(\omega(y,z)) - z\mathcal{B}_H(\omega(y,z))]\,\mathcal{R}(z) + a_0^{(L)} y\,[z - \mathcal{B}_L(\omega(y,z))]}{z\,[y - \omega(y,z)\mathcal{B}_H(\omega(y,z))]} \\ \times \Pi_{00}((1-r)z). \end{aligned} \tag{12}$$

Because $\Pi_0(1, 1)$ is the long-run fraction of time slots that the server is idle and we can conclude from Little's Law that

$$\Pi_0(1, 1) = 1 - \mathcal{A}'_H(1)\mathcal{B}'_H(1) - \mathcal{A}'_L(1)\mathcal{B}'_L(1)$$

we can find an expression for $\Pi_{00}(1-r)$ using equation (12). Notice that $\omega(y, 1) = \mathcal{A}_H(y)$.

$$\Pi_0(1,1) = \lim_{y\to 1} a_0^{(H)} \frac{[y\mathcal{B}_L(\mathcal{A}_H(y)) - \mathcal{B}_H(\mathcal{A}_H(y))]\,\mathcal{R}(1) + a_0^{(L)} y\,[1 - \mathcal{B}_L(\mathcal{A}_H(y))]}{y - \mathcal{A}_H(y)\mathcal{B}_H(\mathcal{A}_H(y))}$$

$$\times \Pi_{00}(1-r) =$$

$$a_0^{(H)} a_0^{(L)} \frac{\begin{array}{l}\left(1 - \mathcal{A}'_H(1)[\mathcal{B}'_H(1) - \mathcal{B}'_L(1)]\right)\left(1 - \mathcal{A}'_H(1)[\mathcal{B}'_H(1)+1] - \mathcal{A}'_L(1)\mathcal{B}'_L(1)\right) + \\ -\mathcal{A}'_H(1)\mathcal{B}'_L(1)\left(1 - \mathcal{A}'_H(1)[\mathcal{B}'_H(1)+1] - \mathcal{A}'_L(1)[\mathcal{B}'_L(1)+1]\right)\end{array}}{\left(1 - \mathcal{A}'_H(1)[\mathcal{B}'_H(1)+1]\right)\left(1 - \mathcal{A}'_H(1)[\mathcal{B}'_H(1)+1] - \mathcal{A}'_L(1)[\mathcal{B}'_L(1)+1]\right)}$$

$$\times \Pi_{00}(1-r) =$$

$$\text{[after some algebra!]} \quad = a_0^{(H)} a_0^{(L)} \frac{1 - \mathcal{A}'_H(1)\mathcal{B}'_H(1) - \mathcal{A}'_L(1)\mathcal{B}'_L(1)}{1 - \mathcal{A}'_H(1)[\mathcal{B}'_H(1)+1] - \mathcal{A}'_L(1)[\mathcal{B}'_L(1)+1]} \Pi_{00}(1-r),$$

from which we find

$$\Pi_{00}(1-r) = \frac{1 - \mathcal{A}'_H(1)[\mathcal{B}'_H(1)+1] - \mathcal{A}'_L(1)[\mathcal{B}'_L(1)+1]}{a_0^{(H)} a_0^{(L)}}.$$

We remark here that interchanging the limits, i.e. considering $\Pi_0(1,1) = \lim_{z\to 1} \Pi_0(1,z)$, leads to the same result, because $\Pi_0(y,z)$ is continuous at the point $(1,1)$, although at first sight the expression looks very different. To double-check our result we give the details. Notice that $\omega(1,z) = \mathcal{A}_L(z)$ and we get

$$\Pi_0(1,1) = \lim_{z\to 1} a_0^{(H)} \frac{\mathcal{A}_L(z)\,[\mathcal{B}_L(\mathcal{A}_L(z)) - z\mathcal{B}_H(\mathcal{A}_L(z))]\,\mathcal{R}(z) + a_0^{(L)}\,[z - \mathcal{B}_L(\mathcal{A}_L(z))]}{z\,[1 - \mathcal{A}_L(z)\mathcal{B}_H(\mathcal{A}_L z))]}$$

$$\times \Pi_{00}((1-r)z) =$$

$$a_0^{(H)} a_0^{(L)} \frac{\begin{array}{l}\left(1 - \mathcal{A}'_L(1)[\mathcal{B}'_L(1) - \mathcal{B}'_H(1)]\right)\left(1 - \mathcal{A}'_H(1)[\mathcal{B}'_H(1)+1] - \mathcal{A}'_L(1)\mathcal{B}'_L(1)\right) + \\ -[1 - \mathcal{A}'_L(1)\mathcal{B}'_L(1)]\left(1 - \mathcal{A}'_H(1)[\mathcal{B}'_H(1)+1] - \mathcal{A}'_L(1)[\mathcal{B}'_L(1)+1]\right)\end{array}}{\mathcal{A}'_L(1)[\mathcal{B}'_H(1)+1](1 - \mathcal{A}'_H(1)[\mathcal{B}'_H(1)+1] - \mathcal{A}'_L(1)[\mathcal{B}'_L(1)+1])}$$

$$\times \Pi_{00}(1-r) =$$

[again after some algebra!]

$$= a_0^{(H)} a_0^{(L)} \frac{1 - \mathcal{A}'_H(1)\mathcal{B}'_H(1) - \mathcal{A}'_L(1)\mathcal{B}'_L(1)}{1 - \mathcal{A}'_H(1)[\mathcal{B}'_H(1)+1] - \mathcal{A}'_L(1)[\mathcal{B}'_L(1)+1]} \Pi_{00}(1-r).$$

So, slightly rewriting equation (11), we get an explicit expression for the partial p.g.f. $\Pi_{00}(z)$,

$$\Pi_{00}(z) = \prod_{i=0}^{\infty} \frac{\mathcal{R}((1-r)^i z)}{\mathcal{R}((1-r)^i)} \mathcal{R}(1)\Pi_{00}(1-r)$$
$$= \frac{1 - \mathcal{A}'_H(1)[\mathcal{B}'_H(1)+1] - \mathcal{A}'_L(1)\mathcal{B}'_L(1)}{a_0^{(H)}} \prod_{i=0}^{\infty} \frac{\mathcal{R}((1-r)^i z)}{\mathcal{R}((1-r)^i)}. \quad (13)$$

Next, using this expression for $\Pi_{00}(z)$ and the expression for $\mathcal{R}(z)$ we also get an expression for $\Pi_0(y, z)$ from equation (12). After canceling out common factors we find

$$\Pi_0(y,z) = a_0^{(H)} \frac{\mathcal{A}_L(z)\,[y\mathcal{B}_L(\omega(y,z)) - z\mathcal{B}_H(\omega(y,z))]\,\mathcal{R}(z) + a_0^{(L)} y\,[z - \mathcal{B}_L(\omega(y,z))]}{z\,[y - \omega(y,z)\mathcal{B}_H(\omega(y,z))]}$$
$$\times \Pi_{00}((1-r)z) =$$
$$(1 - \mathcal{A}'_H(1)[\mathcal{B}'_H(1)+1] - \mathcal{A}'_L(1)[\mathcal{B}'_L(1)+1]) \left(\prod_{i=1}^{\infty} \frac{\mathcal{R}((1-r)^i z)}{\mathcal{R}((1-r)^i)} \right)$$
$$\times \frac{\phi(z)\,[y\mathcal{B}_L(\omega(y,z)) - z\mathcal{B}_H(\omega(y,z))]\,[z - \mathcal{B}_L(\phi(z))] + y\,[z - \mathcal{B}_L(\omega(y,z))]\,[z - \phi(z)\mathcal{B}_L(\phi(z))]}{z\,[y - \omega(y,z)\mathcal{B}_H(\omega(y,z))]\,[z - \phi(z)\mathcal{B}_L(\phi(z))]}. \quad (14)$$

Finally, we approach our main goal, an expression for the three-dimensional p.g.f. $\Pi(w, y, z)$. From equation (5) we have

$$\Pi(w,y,z) = \frac{\begin{array}{c}\mathcal{A}_H(y)\mathcal{A}_L(z)z\,[w\mathcal{B}_H(w) - y]\,\Pi_0(y,z) + a_0^{(H)}\mathcal{A}_L(z)w\,[y\mathcal{B}_L(w) - z\mathcal{B}_H(w)]\,\Pi_{00}(z) \\ + a_0^{(H)} a_0^{(L)} wy\,[z - \mathcal{B}_L(w)]\,\Pi_{00}((1-r)z)\end{array}}{yz\,[w - \mathcal{A}_H(y)\mathcal{A}_L(z)]}. \quad (15)$$

For future use it is worthwhile to factorize out the common factor $\Pi_{00}((1-r)z)$ in the numerator. This gives after some manipulations and writing throughout $\omega(y, z)$ for $\mathcal{A}_H(y)\mathcal{A}_L(z)$,

$$\Pi(w,y,z) = \frac{1 - \mathcal{A}'_H(1)[\mathcal{B}'_H(1)+1] - \mathcal{A}'_L(1)[\mathcal{B}'_L(1)+1]}{a_0^{(L)}} \left(\prod_{i=1}^{\infty} \frac{\mathcal{R}((1-r)^i z)}{\mathcal{R}((1-r)^i)} \right)$$
$$\times \frac{\left[\begin{array}{c}\omega(y,z)\,[w\mathcal{B}_H(w) - y]\left(\begin{array}{c}\mathcal{A}_L(z)\,[y\mathcal{B}_L(\omega(y,z)) - z\mathcal{B}_H(\omega(y,z))]\,\mathcal{R}(z) \\ + a_0^{(L)} y\,[z - \mathcal{B}_L(\omega(y,z))]\end{array}\right) \\ + \mathcal{A}_L(z)w\,[y\mathcal{B}_L(w) - z\mathcal{B}_H(w)]\,\mathcal{R}(z)\,[y - \omega(y,z)\mathcal{B}_H(\omega(y,z))] \\ + a_0^{(L)} wy\,[z - \mathcal{B}_L(w)]\,[y - \omega(y,z)\mathcal{B}_H(\omega(y,z))]\end{array}\right]}{yz\,[w - \omega(y,z)]\,[y - \omega(y,z)\mathcal{B}_H(\omega(y,z))]}. \quad (16)$$

Notice that in the denominator still the factor $a_0^{(L)}$ is present because we did not spell out the retrial function $\mathcal{R}(z)$ in the numerator. Doing that would also cancel out the factor $a_0^{(L)}$.

From expression (16) we find the marginal p.g.f.'s $\mathcal{L}(y) := \Pi(1, y, 1)$ and $\mathcal{Q}(z) := \Pi(1, 1, z)$ of the limiting distribution of the queue length and the orbit size, respectively. To get rid of the factor $a_0^{(L)}$ introduce $\mathcal{R}^*(z) = \mathcal{R}(z)/a_0^{(L)}$. Then we find

$$\mathcal{L}(y) = (1 - \mathcal{A}'_H(1)[\mathcal{B}'_H(1) + 1] - \mathcal{A}'_L(1)[\mathcal{B}'_L(1) + 1]) \frac{1 - y}{y[1 - \mathcal{A}_H(y)]} \times \frac{\mathcal{A}_H(y) \begin{pmatrix} [y\mathcal{B}_L(\mathcal{A}_H(y)) - \mathcal{B}_H(\mathcal{A}_H(y))]\mathcal{R}^*(1) \\ +y\,[1 - \mathcal{B}_L(\mathcal{A}_H(y))] \end{pmatrix} - \mathcal{R}^*(1)\,[y - \mathcal{A}_H(y)\mathcal{B}_H(\mathcal{A}_H(y))]}{y - \mathcal{A}_H(y)\mathcal{B}_H(\mathcal{A}_H(y))}, \tag{17}$$

and, using the definition of $\mathcal{R}^*(z)$ and some further simplification,

$$\mathcal{Q}(z) = (1 - \mathcal{A}'_H(1)[\mathcal{B}'_H(1) + 1] - \mathcal{A}'_L(1)[\mathcal{B}'_L(1) + 1]) \left(\prod_{i=1}^{\infty} \frac{\mathcal{R}((1-r)^i z)}{\mathcal{R}((1-r)^i)} \right) \times \left(\frac{1 - z}{1 - \mathcal{A}_L(z)} \right) \left(\frac{\phi(z) - 1}{z - \phi(z)\mathcal{B}_L(\phi(z))} \right). \tag{18}$$

Notice that from the expressions (17) and (18) we can check that $\mathcal{L}(1) = 1$ and $\mathcal{Q}(1) = 1$. Of course we can also write down the two-dimensional p.g.f. $\mathcal{T}(y, z) := \Pi(1, y, z)$ of the joint limiting distribution of the queue length and the orbit size,

$$\mathcal{T}(y, z) = (1 - \mathcal{A}'_H(1)[\mathcal{B}'_H(1) + 1] - \mathcal{A}'_L(1)[\mathcal{B}'_L(1) + 1]) \left(\prod_{i=1}^{\infty} \frac{\mathcal{R}((1-r)^i z)}{\mathcal{R}((1-r)^i)} \right) \times \frac{\left[\begin{array}{c} \omega(y,z)\,[1-y] \begin{pmatrix} \mathcal{A}_L(z)\,[y\mathcal{B}_L(\omega(y,z)) - z\mathcal{B}_H(\omega(y,z))]\mathcal{R}^*(z) \\ +y\,[z - \mathcal{B}_L(\omega(y,z))] \end{pmatrix} \\ +\mathcal{A}_L(z)\,[y - z]\,\mathcal{R}^*(z)\,[y - \omega(y,z)\mathcal{B}_H(\omega(y,z))] \\ +y\,[z-1]\,[y - \omega(y,z)\mathcal{B}_H(\omega(y,z))] \end{array} \right]}{yz\,[1 - \omega(y,z)]\,[y - \omega(y,z)\mathcal{B}_H(\omega(y,z))]}. \tag{19}$$

Because $\mathcal{T}(y, z) \neq \mathcal{L}(y)\mathcal{Q}(z)$ we see immediately that the queue length and the orbit size are dependent. Our next step is to calculate the mean queue length $\overline{\mathcal{L}}$ and the mean orbit size $\overline{\mathcal{Q}}$. Of course we have

$$\overline{\mathcal{L}} = \mathcal{L}'(1) \quad \text{and} \quad \overline{\mathcal{Q}} = \mathcal{Q}'(1).$$

After tedious calculations we find

$$
\begin{aligned}
\overline{\mathcal{L}} = &- \left(\frac{\mathcal{A}''_H(1)}{2\mathcal{A}'_H(1)} + \mathcal{A}'_H(1)\mathcal{B}'_H(1) + \mathcal{A}'_L(1)\mathcal{B}'_L(1) \right) \\
&+ \frac{1 - \mathcal{A}'_H(1)[\mathcal{B}'_H(1)+1] - \mathcal{A}'_L(1)[\mathcal{B}'_L(1)+1]}{1 - \mathcal{A}'_H(1)[\mathcal{B}'_H(1)+1]} \\
&\times \left[\frac{\mathcal{A}''_H(1)}{2\mathcal{A}'_H(1)} + \left(\mathcal{A}'_H(1) - \frac{\mathcal{A}''_H(1)}{2\mathcal{A}'_H(1)} - 1 \right) \mathcal{B}'_H(1) + \left(1 + \frac{\mathcal{A}''_H(1)}{2\mathcal{A}'_H(1)} \right) \mathcal{B}'_L(1) + \right. \\
&\left. \frac{1}{2}\mathcal{A}'_H(1)\left(\mathcal{B}''_L(1) - \mathcal{B}''_H(1)\right) \right] \\
&+ \frac{\mathcal{A}''_H(1)[\mathcal{B}'_H(1)+1] + \left[\mathcal{A}'_H(1)\right]^2 [\mathcal{B}''_H(1) + 2\mathcal{B}'_H(1)]}{2(1 - \mathcal{A}'_H(1)[\mathcal{B}'_H(1)+1])} \\
&+ \frac{\mathcal{A}'_L(1)}{1 - \mathcal{A}'_H(1)[\mathcal{B}'_H(1)+1]} \left[\frac{\mathcal{A}''_H(1)}{2\mathcal{A}'_H(1)} + \left(1 + \frac{\mathcal{A}''_H(1)}{2\mathcal{A}'_H(1)} \right) \mathcal{B}'_L(1) + \mathcal{A}'_H(1)\mathcal{B}'_H(1) + \right. \\
&\left. \frac{\mathcal{A}''_H(1)}{2\mathcal{A}'_H(1)} \mathcal{B}''_L(1) \right]
\end{aligned}
$$

and

$$
\begin{aligned}
\overline{\mathcal{Q}} = &(1 - \mathcal{A}'_H(1)[\mathcal{B}'_H(1)+1] - \mathcal{A}'_L(1)[\mathcal{B}'_L(1)+1]) \\
&\left\{ \frac{\phi''(1)[1 - \phi'(1)\mathcal{B}''_L(1)] + \left[\phi'(1)\right]^3 \left[2\mathcal{B}'_L(1) + \mathcal{B}''_L(1)\right]}{2\mathcal{A}'_L(1)\left(1 - \phi'(1)[\mathcal{B}'_L(1)+1]\right)^2} \right. \\
&\left. + \left(\frac{\phi'(1)}{1 - \phi'(1)[\mathcal{B}'_L(1)+1]} \right) \left[\frac{\mathcal{A}''_L(1)}{2\left[\mathcal{A}'_L(1)\right]^2} + \frac{1}{\mathcal{A}'_L(1)} \sum_{i=1}^{\infty} \frac{(1-r)^i \mathcal{R}'((1-r)^i)}{\mathcal{R}((1-r)^i)} \right] \right\}.
\end{aligned}
$$

Using equation (8) we can easily evaluate $\phi'(1)$ and $\phi''(1)$ in terms of the p.g.f.'s $\mathcal{A}_L(\cdot)$, $\mathcal{A}_H(\cdot)$, $\mathcal{B}_L(\cdot)$ and $\mathcal{B}_H(\cdot)$. It is more cumbersome to evaluate the terms of the series because for every argument $(1-r)^i$ the calculation of $\mathcal{R}((1-r)^i)$ and $\mathcal{R}'((1-r)^i)$ requires that the values $\phi((1-r^i)$ and $\phi'((1-r)^i)$ are determined as the solution of the two equations (7) and (8) with $z = (1-r)^i$. This solution must be found numerically. We skip further details.

To find the covariance of the queue length and the orbit size we first calculate $\overline{\mathcal{LQ}} := \sum_{i=1}^{\infty}\sum_{n=1}^{\infty} in\pi(1,i,n)$. Using the two-dimensional p.g.f $\mathcal{T}(y,z)$ we have $\overline{\mathcal{LQ}} = \left[\frac{\partial^2}{\partial y \partial z}\mathcal{T}(y,z)\right]_{y=1,z=1}$ and then the covariance is $\mathrm{Cov}(L,Q) = \overline{\mathcal{LQ}} - \overline{\mathcal{L}} \cdot \overline{\mathcal{Q}}$, where we used L and Q as artifact random variables denoting the steady-state queue length and the orbit size, respectively. We do not spell out the long expression for $\overline{\mathcal{LQ}}$, the evaluation simply requires a lot of tedious algebra. We end this section to announce that numerical results for $\overline{\mathcal{L}}$, $\overline{\mathcal{Q}}$ and $\mathrm{Cov}(L,Q)$ will be presented in an extended version of this paper. This work is in preparation.

4 The Mean Busy Period

The busy period in the delay/retrial model is defined as the time lapse from the epoch that the server starts a first service after the server has been idle due to the fact that the system was empty, i.e. no waiting high-priority customers in the queue and no low-priority customers in the orbit, until the first departure epoch leaving behind an empty system again. Introduce B for this busy period and I for the time lapse that the system is empty between two successive busy periods. It is clear that the idle period is geometrically distributed with parameter $1 - a_0^{(H)} a_0^{(L)}$. So, from the the Renewal Reward Theorem we get

$$\pi(0,0,0) = \frac{1 \Big/ \left(1 - a_0^{(H)} a_0^{(L)}\right)}{1 \Big/ \left(1 - a_0^{(H)} a_0^{(L)}\right) + \mathrm{E}[B]}.$$

From (13) we have

$$\pi(0,0,0) = \Pi_{00}(0) = \frac{1 - \mathcal{A}'_H(1)[\mathcal{B}'_H(1)+1] - \mathcal{A}'_L(1)\mathcal{B}'_L(1)}{a_0^{(H)}} \prod_{i=0}^{\infty} \frac{1}{\mathcal{R}((1-r)^i)}.$$

So we get

$$E[B] = \frac{1}{1 - a_0^{(H)} a_0^{(L)}} \left[\frac{a_0^{(H)}}{1 - \mathcal{A}'_H(1)[\mathcal{B}'_H(1)+1] - \mathcal{A}'_L(1)\mathcal{B}'_L(1)} \prod_{i=0}^{\infty} \mathcal{R}((1-r)^i) - 1 \right].$$

References

1. Artalejo, J.R., Gómez-Corral, A.: Retrial Queueing Systems. Springer-Verlag, Heidelberg (2008)
2. Choi, B.D., Kim, J.W.: Discrete-time $Geo_1, Geo_2/G/1$ retrial queueing systems with two types of calls. Computers and Mathematics with Applications **33**, 79–88 (1997)
3. Falin, G.I., Artalejo, J.R., Martin, M.: On the single server retrial queue with priority customers. Queueing Systems **14**, 439–455 (1993)
4. Falin, G.I., Templeton, J.G.C.: Retrial Queues. Chapman & Hall, London (1997)
5. Li, H., Yang, T.: $Geo/G/1$ discrete time retrial queue with Bernoulli schedule. European Journal of Operational Research **111**, 629–649 (1998)
6. Li, H., Yang, T.: Steady-State Queue Size Distribution of Discrete-Time $PH/Geo/1$ Retrial Queues. Mathematical and Computer Modelling **30**, 51–63 (1999)
7. Nobel, R.D., Moreno, P.: A discrete-time priority loss/retrial queueing model with two types of traffic. In: Choi, B.D. (ed.) Proceedings of the Korea-Netherlands Joint Conference on Queueing Theory and its Applications to Telecommunication Systems, Seoul, pp. 189–207 (2005)
8. Nobel, R.D., Moreno, P.: A discrete-time retrial queueing model with one server. EJOR **189**(3), 1088–1103 (2008)
9. Yang, T., Li, H.: On the steady-state queue size distribution of the discrete-time $Geo/G/1$ queue with repeated customers. Queueing Systems **25**, 199–215 (1995)