Two Way Communication Retrial Queues with Balanced Call Blending

Tuan Phung-Duc[1] and Wouter Rogiest[2],[⋆]

[1] Graduate School of Informatics, Kyoto University
Yoshida-Honmachi, Sakyo-ku, Kyoto 606-8501, Japan
`tuan@sys.i.kyoto-u.ac.jp`
[2] Department of Telecommunications and Information Processing,
Ghent University
St.-Pietersnieuwstraat 41, B-9000 Ghent, Belgium
`wouter.rogiest@telin.ugent.be`

Abstract. In call centers, call blending consists in the mixing of incoming and outgoing call activity. Artalejo and Phung-Duc recently provided an apt model for such a setting, with a two way communication retrial queue. However, by assuming a classical (proportional) retrial rate for the incoming calls, the outgoing call activity is largely blocked when many incoming calls are in orbit, which may be unwanted, especially when outgoing calls are vital to the service offered.

In this paper, we assume a balanced way of call blending, through a retrial queue with constant retrial rate for incoming calls. For the single server case (one operator), a generating functions approach enables deriving explicit formulas for the joint stationary distribution of the number of incoming calls and the system state, and also for the factorial moments. This is complemented with a stability analysis, expressions for performance measures, and also recursive formulas, allowing reliable numerical calculation. For the multiserver case (multiple operators), we provide a quasi-birth-and-death process formulation, enabling deriving a sufficient and necessary condition for stability in this case, as well as a numerical recipe to obtain the stationary distribution.

Keywords: Markov chain, retrial queues, single server, multiserver, call centers, call blending.

1 Introduction

Retrial queues have received considerable attention over recent years, providing an apt model for the performance evaluation of call centers, computer networks, and communications systems. An overview is given in [1,2]. Characteristic of retrial queues is the fact that calls (or, in general, customers) that cannot be served upon arrival enter an orbit and request for a retrial after some random time. Due to this, analysis of a retrial queue is more difficult than that of its

---

[⋆] Corresponding author.

counterpart model without retrial, and explicit results can only be obtained in a few special cases [3].

Here, we consider a specific retrial queue model with application to call centers. As explained in [4], a central characteristic of a call center is whether it handles inbound traffic, with incoming calls, or outbound traffic, with outgoing calls. Correspondingly, they are referred to as inbound and outbound call centers. Most retrial queue models in literature assume such a system, with one way communication. However, often, call centers are not strictly inbound or outbound, and typically handle a mixture of incoming and outgoing calls. Typically, incoming calls are assigned to operators by an Automatic Call Distributor (ACD). For outgoing calls, calls are either initiated by the ACD (automatically), or by the operators (manually).

The principle of mixing is referred to as call blending, with two way communication, serving several purposes.

- Firstly, it may be added to the regular tasks, as in the case of an inbound call center in which operators utilize their idle time to perform secondary, non-urgent outgoing calls. Then, call blending is primarily a way to increase overall productivity, by increasing operator utilization, potentially through a control policy. A mathematical analysis and optimization of such a policy is presented in [5].
- Secondly, it may occur as an integral part of the tasks performed at the call center. In this case, incoming as well as outgoing calls are vital elements of the service delivered, and should both be performed. This occurs when tasks necessitate several calls in both directions.

Both cases can be modeled with a retrial queue supporting two way communication. More precisely, [6] assumes a model with classical retrial rate for incoming calls (see Section 2.1 for definition of retrial rate). Such a choice results in an apt model for the first case, since the outgoing call activity is indeed largest when few incoming calls are in orbit, and smallest when many are in orbit. However, in the second case, such behavior is undesirable, since the outgoing call activity should also continue regularly while many incoming calls are awaiting service. By assuming a constant retrial rate, outgoing calls are still initiated regularly (either by the ACD or by the operators), even if the number of incoming calls in orbit is high.

Further, note that many types of call blending can be identified; Koole and Mandelbaum [4] provide an excellent overview. A high-level discussion and basic performance analysis is provided in [7]. The paper [8] presents a collection of Markov chain models for call centers, including a discussion of model fidelity and efficacy, in a simulation context. Although different in several ways, Model M1 in [8], with "all blend agents and no mismatches" shares many of the assumptions of the two way communication retrial queue model presented in [6] and here (see also Section 2.1): inbound calls arriving according to a Poisson process, with independent and identically distributed (i.i.d.) service times drawn from an exponential distribution for inbound as well as outbound calls. Further, in [8] and here, multiple identical blend agents (or operators) are assumed (only

1 in [6]), as well as a First-Come-First-Served (FCFS) order for the queue of incoming calls. More precisely, although the model presented here does not make assumptions on the service order for the queue, constant retrial rate is commonly associated with FCFS ordering in the orbit queue (see e.g. [9,10], and [11,12] for discussion), with only the customer at the head of the queue able to request for service. In this regard, FCFS ordering of incoming call requests may be viewed as a more natural (but not only) way to realize a constant retrial rate.

As mentioned earlier, [6] shares many of the assumptions of the single server part of this work. However, assuming constant instead of classical retrial rate leads to completely different expressions for the variables analyzed, with no simple (mathematical) way to relate the obtained results to those of [6]. In terms of analysis, more closely related to this contribution is [13]. In this paper, a service system is analyzed in which a processor must serve two types of impatient units, with either infinitely impatient or infinitely patient customers. Assuming general service times, a variant of Takács' equation is derived which also holds for the system considered in this work, with exponentially distributed service times. In this regard, this paper [13] provides an interesting reference, but does not contain any of the derivations and expressions reported here.

The contribution of this paper is twofold. First, we carry out an extensive analysis for the single server retrial queue with two way communication and constant retrial rate in which we derive explicit expressions for the joint stationary distribution and their partial generating functions as well as recursive formulae. Second, we formulate the multiserver case by a quasi-birth-and-death process for which the stability condition and a numerical recipe are presented.

The remainder of this paper is structured as follows. In Section 2, we set out the model and assumptions of the current work, as well as the balance equations governing the system's behavior. Section 3 presents an exhaustive analysis of the single server case (one operator), including a study of stability, as well as closed-form expressions for the joint stationary distribution of the number of incoming calls and the system state, and several other measures of interest. In Section 4, we consider the multiserver model (multiple operators), through a formulation using a quasi-birth-and-death process. As explained, this allows to apply standard numerical recipes to obtain the stationary distribution as well as the stability condition. Finally, conclusions are drawn in Section 5.

2   Model

In this section, we first list the assumptions made in this work, introducing notation for the parameters involved. This allows to formulate a set of balance equations, which will provide the starting point for the analysis in the next section.

2.1   Assumptions

A single server retrial queue with two way communication is considered. Primary incoming call requests arrive at the server (or operator) according to a Poisson

process with rate λ. Incoming calls finding an idle server receive service instantly. In case of a busy server, the incoming call enters an orbit. Within the orbit, a constant retrial policy is applied, i.e., the arrival rate of customers from the orbit is $\mu(1 - \delta_{0,n})$ provided that there are n customers in the orbit, where $\delta_{0,n}$ denotes the Kronecker delta. This is opposed to the case analyzed in [6], with a classical retrial rate $n\mu$, which depends on the number of customers in orbit, n. As mentioned, constant retrial rate occurs when customers form a FCFS queue in the orbit, and only the customer at the head of the queue can request service. In addition, when the server turns idle, it makes an outgoing call after an exponentially distributed time with rate α. The service times of the incoming and outgoing calls are i.i.d., exponentially distributed with rate ν_1 and ν_2 respectively.

2.2   Markov Chain and Balance Equations

Let $S(t)$ denote the state of the server at time t,

$$
S(t) = \begin{cases} 0 & \text{if the server is idle,} \\ 1 & \text{if the server is providing an incoming service,} \\ 2 & \text{if the server is providing an outgoing service,} \end{cases}
$$

and let $N(t)$ denote the number of calls in orbit at time t. Here, the couple $\{(S(t), N(t)); t \geq 0\}$ forms a Markov chain on the state space $\{0, 1, 2\} \times \mathbb{Z}_+$, with $\mathbb{Z}_+ = \{0, 1, 2, \ldots\}$. Given that the Markov chain is aperiodic and irreducible, under the condition that the system is stable, the probability distributions associated with the variables involved converge to a unique stochastic equilibrium for $t \to \infty$, to

$$
\pi_{i,j} = \lim_{t \to \infty} \Pr[S(t) = i, \, N(t) = j], \qquad (i, j) \in \{0, 1, 2\} \times \mathbb{Z}_+ \, .
$$

The condition for stability will be derived in Section 3. Now, the probabilities $\pi_{i,j}, (i, j) \in \{0, 1, 2\} \times \mathbb{Z}_+$, are characterized by following set of balance equations,

$$
(\lambda + \alpha + \mu(1 - \delta_{0,j}))\pi_{0,j} = v_1 \pi_{1,j} + v_2 \pi_{2,j}, \tag{1}
$$
$$
(\lambda + v_1)\pi_{1,j} = \lambda \pi_{0,j} + \mu \pi_{0,j+1} + \lambda \pi_{1,j-1}, \tag{2}
$$
$$
(\lambda + \nu_2)\pi_{2,j} = \alpha \pi_{0,j} + \lambda \pi_{2,j-1}, \tag{3}
$$

for $j \in \mathbb{Z}_+$, with $\delta_{0,j} = 1$ for $j = 0$ and zero elsewhere, and $\pi_{i,-1} = 0$ for $i \in \{1, 2\}$.

Let $\Pi_i(z), i \in \{0, 1, 2\}$, denote the partial generating functions

$$
\Pi_i(z) = \sum_{j=0}^{\infty} \pi_{i,j} z^j, \qquad i \in \{0, 1, 2\} \, ,
$$

with z a complex number, $z \in \mathbb{C}$. Multiplying the balance equations (1-3) by z^j and taking the sum over $j \in \mathbb{Z}_+$, the balance equations are transformed to the z-domain, yielding

$$(\lambda + \alpha + \mu)\Pi_0(z) - \mu\pi_{0,0} = v_1\Pi_1(z) + v_2\Pi_2(z), \tag{4}$$

$$(\lambda + v_1)\Pi_1(z) = (\lambda + \mu z^{-1})\Pi_0(z) - \mu\pi_{0,0}z^{-1} + \lambda z\Pi_1(z), \tag{5}$$

$$(\lambda + \nu_2)\Pi_2(z) = \alpha\Pi_0(z) + \lambda z\Pi_2(z). \tag{6}$$

Summing (4-6), multiplying with z and dividing by $(z - 1)$ leads to an orbit balance equation,

$$\mu(\Pi_0(z) - \pi_{0,0}) = \lambda z(\Pi_1(z) + \Pi_2(z)), \tag{7}$$

which will prove useful at several points in the analysis below.

3   Analysis

This section provides the analysis of the single-server case. We first derive explicit expressions for the three partial probability generating functions $\Pi_i(z)$, $i \in \{0, 1, 2\}$, associated with the stationary distribution probabilities $\pi_{i,j}$, $(i, j) \in \{0, 1, 2\} \times \mathbb{Z}_+$. From these, a stability condition is derived, and inversion of the generating functions to the probability domain yields closed-form expressions for the stationary distribution. Further, we treat the factorial moments, recursive formulas, first moments and a cost model.

3.1   Generating Functions

Looking for explicit expressions for the $\Pi_i(z)$, $i \in \{0, 1, 2\}$, we first remark that $\Pi_1(z)$ can be expressed in terms of $\Pi_0(z)$ through (4), leading to

$$\Pi_1(z) = \frac{1}{\nu_1}(\lambda + \alpha + \mu - \nu_2\frac{\alpha}{\lambda + \nu_2 - \lambda z})\Pi_0(z) - \frac{\mu}{\nu_1}\pi_{0,0}, \tag{8}$$

while (6) yields that

$$\Pi_2(z) = \frac{\alpha}{\lambda + \nu_2 - \lambda z}\Pi_0(z). \tag{9}$$

Substituting (9) and (8) in (7), we obtain

$$\Pi_0(z) = \pi_{0,0}(1 - \frac{\lambda z}{\nu_1})\left(1 - \frac{\lambda z}{\mu\nu_1}\left(\lambda + \mu + \alpha\frac{\lambda + \nu_1 - \lambda z}{\lambda + \nu_2 - \lambda z}\right)\right)^{-1}. \tag{10}$$

Now, only $\pi_{0,0}$ needs to be determined to make (10) explicit. To obtain $\pi_{0,0}$, we evaluate the partial generating functions in $z = 1$, and then verify the normalization condition. We obtain

$$\Pi_0(1) = \pi_{0,0}\left(1 - \frac{\lambda}{\nu_1}\right)\left(1 - \frac{\lambda}{\mu\nu_1}\left(\lambda + \mu + \alpha\frac{\nu_1}{\nu_2}\right)\right)^{-1},$$

$$\Pi_1(1) = \frac{\lambda + \mu}{\nu_1}\Pi_0(1) - \frac{\mu}{\nu_1}\pi_{0,0}, \tag{11}$$

$$\Pi_2(1) = \frac{\alpha}{\nu_2}\Pi_0(1).$$

Introducing some additional notation,

$$\rho = \frac{\lambda}{\nu_1}, \qquad \sigma = \frac{\alpha}{\nu_2} \,,$$

and requiring that $\sum_{i=0}^{2} \Pi_i(1) = 1$, some calculations yield

$$\pi_{0,0} = \frac{1 - \frac{\lambda}{\mu}(\rho + \sigma + \frac{\mu}{\nu_1})}{1 + \sigma} \,. \tag{12}$$

Expression (10) and (12) together provide an expression for the partial generating function $\Pi_0(z)$, which is a function of only the model parameters, and thus explicit, as wanted. Explicit expressions for $\Pi_1(z)$ and $\Pi_2(z)$ are readily obtained, by substituting $\Pi_0(z)$ in (8) and (9), respectively.

Finally, with (12), we can simplify (11), to obtain

$$\Pi_0(1) = \frac{1 - \rho}{1 + \sigma} \,, \ \Pi_1(1) = \rho \,, \ \Pi_2(1) = \sigma \frac{1 - \rho}{1 + \sigma} \,. \tag{13}$$

These steady-state probabilities have also been obtained in [6] for the model with classical retrial rate. This is somewhat surprising, since assumptions are different, and all three generating functions ($\Pi_0(z)$, $\Pi_1(z)$ and $\Pi_2(z)$) reported here differ significantly from those in [6]. The fact that the values of (13) match can be understood from the fact that no incoming calls are ever lost (and thus, $\Pi_1(1)$ should amount to the traffic load).

3.2 Stability Condition

With $\pi_{0,0}$ obtained by (12), a characterization of stability is now straightforward. More precisely, requesting $\pi_{0,0}$ to be larger than zero leads to the stability condition for the single-server system,

$$-\mu + (\lambda + \mu)\frac{\lambda}{\nu_1} + \alpha\frac{\lambda}{\nu_2} < 0 \,. \tag{14}$$

3.3 Stationary Distribution

At this point, we derive explicit formulae for $\pi_{0,j}$, $\pi_{1,j}$ and $\pi_{2,j}$, $j \in \mathbb{Z}_+$. We already have $\pi_{0,0}$ from (12), and start by deriving $\pi_{0,j}$, $j \geq 1$, from (10). To this end, we transform $\Pi_0(z)$ from (10) as follows

$$\Pi_0(z) = \frac{\pi_{0,0}(1 - \rho z)(1 - \theta z)}{\frac{1}{b}z^2 - \frac{a}{b}z + 1} \,, \tag{15}$$

where

$$\theta = \frac{\lambda}{\lambda + \nu_2} \,, \ a = \frac{(\lambda + \mu)(\lambda + \nu_2) + \alpha(\lambda + \nu_1) + \mu\nu_1}{\lambda(\lambda + \alpha + \mu)} \,, \ b = \frac{\mu\nu_1(\lambda + \nu_2)}{\lambda^2(\lambda + \alpha + \mu)} \,.$$

Through a partial fraction expansion, this can be rewritten as

$$\Pi_0(z) = \pi_{0,0}\left(1 + \frac{z}{z_1}\frac{C}{1 - \frac{z}{z_1}} + \frac{z}{z_2}\frac{D}{1 - \frac{z}{z_2}}\right), \tag{16}$$

with

$$C = \frac{(1-\rho z)(1-\theta z)}{1 - \frac{z}{z_2}}\Big|_{z=z_1} = \frac{z_2(1-\rho z_1)(1-\theta z_1)}{z_2 - z_1},$$

$$D = \frac{(1-\rho z)(1-\theta z)}{1 - \frac{z}{z_1}}\Big|_{z=z_2} = \frac{z_1(1-\rho z_2)(1-\theta z_2)}{z_1 - z_2},$$

where z_1 and z_2 denote the real and positive poles of $\Pi_0(z)$, namely

$$z_1 = \frac{a + \sqrt{a^2 - 4b}}{2}, \qquad z_2 = \frac{a - \sqrt{a^2 - 4b}}{2},$$

with $z_1 + z_2 = a$, and $z_1 z_2 = b$. If stability condition (14) holds, $z_1 > z_2 > 1$, so enabling inversion of (16) to the probability domain, as

$$\pi_{0,j} = \pi_{0,0}\left[C\left(\frac{1}{z_1}\right)^j + D\left(\frac{1}{z_2}\right)^j\right], \qquad j \geq 1. \tag{17}$$

To obtain $\pi_{1,j}$ and $\pi_{2,j}$, we expand the partial generating functions into simpler fractions, which can easily be inverted from the z-domain to the probability domain. A useful expression in the calculation is obtained from (15), by performing a partial fraction expansion, leading to

$$\frac{1}{1-\theta z}\Pi_0(z) = \pi_{0,0}(1-\rho z)\left(\frac{1}{(1-\frac{z_1}{z_2})(1-\frac{z}{z_1})} + \frac{1}{(1-\frac{z_2}{z_1})(1-\frac{z}{z_2})}\right).$$

Using this, and (8) and (17), we obtain

$$\pi_{1,0} = \frac{1}{\nu_1}\pi_{0,0}(\lambda + \alpha - \frac{\alpha\nu_2}{\lambda + \nu_2}), \tag{18}$$

$$\pi_{1,j} = \frac{1}{\nu_1}\pi_{0,0}\left[C\left(\lambda + \alpha + \mu - \frac{\alpha\nu_2}{(\lambda + \nu_2)(1-\theta z_1)}\right)\left(\frac{1}{z_1}\right)^j\right.$$

$$\left. + D\left(\lambda + \alpha + \mu - \frac{\alpha\nu_2}{(\lambda + \nu_2)(1-\theta z_2)}\right)\left(\frac{1}{z_2}\right)^j\right], \quad j \geq 1. \tag{19}$$

Similarly, with (9) and (17), we find

$$\pi_{2,0} = \frac{\alpha}{\lambda + \nu_2}\pi_{0,0}, \tag{20}$$

$$\pi_{2,j} = \frac{\alpha}{\lambda + \nu_2}\pi_{0,0}\left[\frac{C}{1-\theta z_1}\left(\frac{1}{z_1}\right)^j + \frac{D}{1-\theta z_2}\left(\frac{1}{z_2}\right)^j\right], \quad j \geq 1. \tag{21}$$

As such, for $j \in \mathbb{Z}_+$, $\pi_{0,j}$ is given by (12) and (17), $\pi_{1,j}$ by (18) and (19), and $\pi_{2,j}$ by (20) and (21).

Remark 1. In the derivations presented above, we implicitly assumed that $z_1 \notin \{z_2, \rho^{-1}, \theta^{-1}\}$, and $z_2 \notin \{\rho^{-1}, \theta^{-1}\}$. This assumption excludes some minor special cases where the inversion from z-domain to the probability domain requires small modifications.

3.4   Factorial Moments

Next, we derive explicit expressions for the partial factorial moments M_k^i, $(i, k) \in \{0, 1, 2\} \times \mathbb{Z}_+$, which relate to the coefficients of z^k in the series $\Pi_i(1 + z)$ as follows,

$$\Pi_i(1 + z) = \sum_{k=0}^{\infty} \frac{M_k^i}{k!} z^k, \qquad i \in \{0, 1, 2\}. \tag{22}$$

For $k = 0$, (13) already provides the answer, since $M_0^i = \Pi_i(1)$. For $k \geq 1$, expressing $\Pi_0(1 + z)$ using (16), we obtain

$$M_k^0 = k! \pi_{0,0} \left(\frac{C z_1}{(z_1 - 1)^{k+1}} + \frac{D z_2}{(z_2 - 1)^{k+1}} \right), \quad k \geq 1,$$

$$M_k^1 = \frac{k! \pi_{0,0}}{\nu_1} \left[\left(\lambda + \alpha + \mu - \frac{\alpha \nu_2}{(\lambda + \nu_2)(1 - \theta z_1)} \right) \frac{C z_1}{(z_1 - 1)^{k+1}} \right.$$
$$\left. + \left(\lambda + \alpha + \mu - \frac{\alpha \nu_2}{(\lambda + \nu_2)(1 - \theta z_2)} \right) \frac{C z_2}{(z_2 - 1)^{k+1}} \right], \quad k \geq 1,$$

$$M_k^2 = \frac{\alpha k! \pi_{0,0}}{\lambda + \nu_2} \left[\frac{1}{1 - \theta z_1} \frac{C z_1}{(z_1 - 1)^{k+1}} + \frac{1}{1 - \theta z_2} \frac{C z_2}{(z_2 - 1)^{k+1}} \right], \quad k \geq 1.$$

Together with (13) (with $\Pi_i(1) = M_0^i$), this provides explicit expressions for all M_k^i, $k \in \mathbb{Z}_+$, $i \in \{0, 1, 2\}$.

3.5   Recursive Formulae

In Sections 3.3 and 3.4, explicit expressions are given for the stationary distribution and the partial factorial moments. However, since the coefficients involved may be either positive or negative, numerical computation may be unreliable. Opposed to this, a recursive computation with only positive terms provides a numerically stable alternative.

The stationary probabilities can be expressed recursively as follows,

$$\pi_{0,j} = \frac{\lambda(\pi_{1,j-1} + \pi_{2,j-1})}{\mu}, \qquad j \geq 1, \tag{23}$$

$$\pi_{2,j} = \frac{\alpha \pi_{0,j} + \lambda \pi_{2,j-1}}{\lambda + \nu_2}, \qquad j \geq 1, \tag{24}$$

$$\pi_{1,j} = \frac{\lambda(\pi_{0,j} + \pi_{2,j} + \pi_{1,j-1})}{\nu_1}, \qquad j \geq 1, \tag{25}$$

where we recall that $\pi_{0,0}$ is given by (12), $\pi_{1,0}$ by (18), and $\pi_{2,0}$ by (20). Expression (23) can be obtained from (7), whereas (24) is derived from (3), and (25) from the combination of (2) and (24).

For the partial factorial moments, starting point is the expression (7), substituting z with $(1 + z)$. Appealing to (22), we find that

$$M_k^0 = \frac{\lambda(M_k^1 + M_k^2) + k\lambda(M_{k-1}^1 + M_{k-1}^2)}{\mu}, \qquad k \geq 1. \tag{26}$$

Similarly, from (6), substituting z with $(1 + z)$, we obtain

$$M_k^2 = \frac{\alpha M_k^0 + k\lambda M_{k-1}^2}{\nu_2}, \qquad k \geq 1. \tag{27}$$

Further, adding (7) to the product of (5) and z allows to derive that

$$M_k^1 = \frac{\lambda(M_k^0 + M_k^2 + kM_{k-1}^1)}{\nu_1 - \lambda}, \qquad k \geq 1. \tag{28}$$

Substituting (27) and (28) in (26) yields

$$M_k^0 = k\lambda \cdot \frac{\nu_1\nu_2 M_{k-1}^1 + [\lambda\nu_1 + \nu_2(\nu_1 - \lambda)]M_{k-1}^2}{\mu\nu_2(\nu_1 - \lambda) - \lambda(\alpha\nu_1 + \lambda\nu_2)}, \qquad k \geq 1. \tag{29}$$

Expressions (27), (28) and (29), and (13) (with $\Pi_i(1) = M_0^i$) together provide the recursive formulation for the partial factorial moments. It should be noted that the denominator of (29) is positive due to the stability condition (14).

3.6   First Moments and Cost Model

In this section, deriving first moments allows formulating a cost model. From (10), we find that

$$M_1^0 = \Pi_0'(1) = \pi_{0,0}\left(\frac{Cz_1}{(1 - z_1)^2} + \frac{Dz_2}{(1 - z_2)^2}\right),$$

which, after some calculations using $z_1 + z_2 = a$ and $z_1 z_2 = b$, is simplified as

$$M_1^0 = \pi_{0,0} b\frac{a - 2 + (\rho + \theta)(1 - b) + \rho\theta(2b - a)}{(1 - a + b)^2}. \tag{30}$$

Let $E[N]$ denote the average number of customers in the orbit, i.e.,

$$E[N] = M_1^0 + M_1^1 + M_1^2.$$

It follows from (26) that

$$M_1^1 + M_1^2 = \frac{\mu}{\lambda}M_1^0 - (M_0^1 + M_0^2),$$
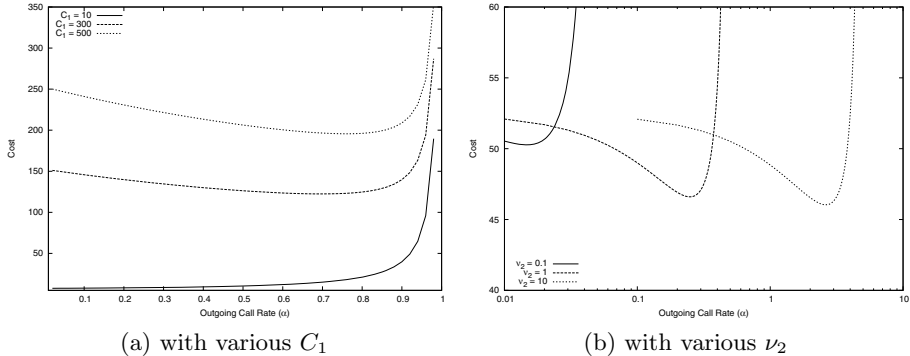
(a) with various C_1 (b) with various ν_2

Fig. 1. Cost as a function of varying α, for the parameter settings specified in Table 1

leading to

$$\mathrm{E}[N] = \frac{\lambda + \mu}{\lambda} M_1^0 - \frac{\rho + \sigma}{1 + \sigma},$$

where M_1^0 is given by (30).

Let U denote the utilization of the server, i.e., $U = M_0^1 + M_0^2$. From a management point of view, we need to minimize $1 - U$. At the same time, we also need to minimize the number of customers in the orbit $\mathrm{E}[N]$. These considerations motivate the following cost model,

$$\min \quad f(\alpha) = C_1(1 - U) + C_2 \mathrm{E}[N],$$

$$s.t. \quad -\mu + (\lambda + \mu)\frac{\lambda}{\nu_1} + \alpha\frac{\lambda}{\nu_2} < 0, \qquad \alpha \geq 0,$$

where the inequality comes from the stability condition and C_1 and C_2 are the cost of idle server and of a retrial customer. The cost model formulation boils down to finding the optimal α while keeping all other parameters constant.

Remark 2. In our model, there are a number of free parameters such as $\lambda, \mu, \nu_1, \nu_2$ and α. Thus, the optimization formulation presented above is only one of several options. However, aiming for the optimization of α is natural in the call center context, as it can be controlled by the operator (directly), or by the ACD (automatically).

In Fig. 1, the cost model is evaluated for varying α, under the parameter setting specified in Table 1. For the setting considered in Fig. 1a, cost evaluation yields a non-trivial optimal value for α when $C_1 \in \{300, 500\}$. This corresponds to the case where the cost of idle server (C_1) is (much) larger than the cost of a retrial customer (C_2, fixed to 1). Opposed to this, when the cost of idle server is small, $C_1 = 10$, the cost function is monotonically increasing, and the optimum is trivially found for $\alpha = 0$, with no outgoing call activity, which is intuitive. For various ν_2, illustrated in Fig. 1b, cost evaluation yields clear optima for α,

Table 1. Parameter setting for the numerical examples considered in Fig. 1

Figure	λ	ν_1	ν_2	μ	C_1	C_2
(a)	0.5	1	2	1	various	1
(b)	0.5	1	various	1	100	1

increasing with increasing ν_2, rate parameter of the outgoing call service time distribution. In the optimum, apparently, reduced duration of outgoing calls is matched by an increased rate for outgoing call activity, which is also quite intuitive.

4 Multiple Operators

While the analysis of the previous section assumed a single operator (single server), we now shift focus to the case of multiple operators (multiserver). In particular, we consider an M/M/c/c ($c \geq 1$) retrial queue with constant retrial rate and two way communication, where the notations λ, μ, ν_1, ν_2 and α have the same definitions as above. The behavior of each server in this multiserver model is the same as that of the single server case, i.e., an idle server makes an outgoing call after an exponentially distributed time with rate α. We first provide a Quasi-Birth-and-Death (QBD) process formulation, identifying all the components of the involved infinitesimal generator and block matrices. Next, we highlight stability, and also examine the numerical recipe for the calculation of the stationary probabilities.

4.1 Infinitesimal Generator and Matrices

Let $S_1(t)$, $S_2(t)$ and $N(t)$ denote the numbers of incoming calls and outgoing calls in the servers and the number of customers in the orbit at time t, respectively. It is easy to see that $\{X(t) = (S_1(t), S_2(t), N(t)); t \geq 0\}$ forms a level-independent QBD process in the state space

$$\mathcal{S} = \{(i, j, k); i = 0, 1, \ldots, c, j = 0, 1, \ldots, c - i, j \in \mathbb{Z}_+\}.$$

Let O denote a matrix with an appropriate size with all zero entries. It is easy to see that the infinitesimal generator of $\{X(t); t \geq 0\}$ is given by

$$Q = \begin{pmatrix} A^0 & A^+ & O & O & \cdots \\ A^- & A & A^+ & O & \cdots \\ O & A^- & A & A^+ & \cdots \\ O & O & A^- & A & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}.$$

The block matrices A^-, A, A^+ and A^+ are explicitly written as follows,

$$A^- = \begin{pmatrix} O & A_0^- & O & \cdots & O \\ O & O & A_1^- & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & O \\ \vdots & & & O & A_{c-1}^- \\ O & \cdots & \cdots & O & O \end{pmatrix}, \qquad A = \begin{pmatrix} A_{0,1} & A_{0,0} & O & \cdots & \cdots & O \\ A_{1,2} & A_{1,1} & A_{1,0} & \ddots & & \vdots \\ O & A_{2,2} & A_{2,1} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & O \\ \vdots & & & \ddots & A_{c-1,1} & A_{c-1,0} \\ O & \cdots & \cdots & O & A_{c,2} & A_{c,1} \end{pmatrix},$$

$$A^+ = \begin{pmatrix} A_0^+ & O & O & \cdots & O \\ O & A_1^+ & O & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & O \\ \vdots & & & A_{c-1}^+ & O \\ O & \cdots & \cdots & O & A_c^+ \end{pmatrix}, \qquad A^0 = \begin{pmatrix} A_{0,1}^0 & A_{0,0} & O & \cdots & \cdots & O \\ A_{1,2} & A_{1,1}^0 & A_{1,0} & \ddots & & \vdots \\ O & A_{2,2} & A_{2,1}^0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & O \\ \vdots & & & \ddots & A_{c-1,1}^0 & A_{c-1,0} \\ O & \cdots & \cdots & O & A_{c,2} & A_{c,1}^0 \end{pmatrix},$$

where A_i^- $(i = 0, 1, \ldots, c-1)$, A_i^+ $(i = 0, 1, \ldots, c)$, $A_{i,2}$ $(i = 1, 2, \ldots, c)$, $A_{i,1}$, $A_{i,1}^0$ $(i = 0, 1, \ldots, c)$ and $A_{i,0}$ $(i = 0, 1, \ldots, c-1)$ are $(c-i+1) \times (c-i)$, $(c-i+1) \times (c-i+1)$, $(c-i+1) \times (c-i+2)$, $(c-i+1) \times (c-i+1)$, $(c-i+1) \times (c-i+1)$ and $(c-i+1) \times (c-i)$ matrices respectively, with entries given by

$$A_i^-(j, j') = \begin{cases} \mu, & j' = j \ (j = 0, 1, \ldots, c-i-1), \\ 0, & \text{otherwise}, \end{cases}$$

$$A_i^+(j, j') = \begin{cases} \lambda, & j' = j = c-i, \\ 0, & \text{otherwise}, \end{cases}$$

$$A_{i,2}(j, j') = \begin{cases} i\nu_1, & j' = j \ (j = 0, 1, \ldots, c-i), \\ 0, & \text{otherwise}, \end{cases}$$

$$A_{i,1}(j, j') = \begin{cases} (c-i-j)\alpha, & j' = j+1 \ (j = 0, 1, \ldots, c-i-1), \\ j\nu_2, & j' = j-1 \ (j = 1, 2, \ldots, c-i), \\ -\gamma_{i,j}, & j' = j \ (j = 0, 1, \ldots, c-i), \\ 0, & \text{otherwise}, \end{cases}$$

$$A_{i,1}^0(j, j') = \begin{cases} (c-i-j)\alpha, & j' = j+1 \ (j = 0, 1, \ldots, c-i-1), \\ j\nu_2, & j' = j-1 \ (j = 1, 2, \ldots, c-i), \\ -\gamma_{i,j}^0, & j' = j \ (j = 0, 1, \ldots, c-i), \\ 0, & \text{otherwise}, \end{cases}$$

$$A_{i,0}(j, j') = \begin{cases} \lambda, & j' = j \ (j = 0, 1, \ldots, c-i-1), \\ 0, & \text{otherwise}, \end{cases}$$

where $\gamma_{i,j} = \lambda + \mu + i\nu_1 + j\nu_2 + (c-i-j)\alpha$, and $\gamma_{i,j}^0 = \lambda + i\nu_1 + j\nu_2 + (c-i-j)\alpha$.

4.2   Stability Condition

We consider the matrix $P = A^- + A + A^+$, which is the infinitesimal generator of the irreducible Markov chain $\{C(t) = (S_1(t), S_2(t)); t \geq 0\}$ on the state space $\mathcal{V} = \{(i, j); i = 0, 1, \ldots, c, j = 0, 1, \ldots, c-i\}$. It should be noted that this Markov chain represents the behavior of the servers regardless of the number of customers in the orbit when it is large enough. Let $p_{i,j} = \lim_{t\to\infty} \Pr(S_1(t) = i, S_2(t) = j)$ for $(i, j) \in \mathcal{V}$. Furthermore, let $\boldsymbol{p}_i = (p_{i,0}, p_{i,1}, \ldots, p_{i,c-i})$ $(i = 0, 1, \ldots, c)$. In addition, let $\boldsymbol{p} = (\boldsymbol{p}_0, \boldsymbol{p}_1, \ldots, \boldsymbol{p}_c)$ denote the stationary distribution of $\{C(t); t \geq 0\}$, which is the unique solution of the following system of equations.

$$\boldsymbol{p}P = \boldsymbol{0}, \qquad \boldsymbol{p}\boldsymbol{e} = 1,$$

where $\boldsymbol{0}$ and $\boldsymbol{e}$ denote a row and a column vector with an appropriate size with all zero and all one entries, respectively. The necessary and sufficient condition for the stability of $\{X(t); t \geq 0\}$ is given by

$$\boldsymbol{p}A^+\boldsymbol{e} < \boldsymbol{p}A^-\boldsymbol{e}, \tag{31}$$

according to [14]. Because the number of states of $\{C(t); t \geq 0\}$ is finite the stability condition presented by (31) itself is explicit. However, it seems that a simple scalar form in terms of given parameters is not easily obtainable.

Special Case. As a way to verify consistency, we apply the multiserver stability condition to the single-server case. For the matrices, we obtain

$$A^+ = \begin{pmatrix} 0 & 0 & 0 \\ 0 & \lambda & 0 \\ 0 & 0 & \lambda \end{pmatrix}, \qquad A^- = \begin{pmatrix} 0 & 0 & \mu \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix},$$

$$A = \begin{pmatrix} -(\lambda + \alpha + \mu) & \alpha & \lambda \\ \nu_2 & -(\lambda + \nu_2) & 0 \\ \nu_1 & 0 & -\nu_1 \end{pmatrix},$$

$$P = \begin{pmatrix} -(\lambda + \alpha + \mu) & \alpha & \lambda + \mu \\ \nu_2 & -\nu_2 & 0 \\ \nu_1 & 0 & -\nu_1 \end{pmatrix}.$$

We have

$$p_{0,1} = \frac{\alpha}{\nu_2} p_{0,0}, \qquad p_{1,0} = \frac{\lambda + \mu}{\nu_1} p_{0,0}.$$

Thus, the stability condition (31) yields,

$$\lambda \frac{\alpha}{\nu_2} + \lambda \frac{\lambda + \mu}{\nu_1} < \mu,$$

which is consistent with (14), as should.

4.3   Stationary Distribution

In this section, we derive the stationary distribution for $\{X(t); t \geq 0\}$. Under the stability condition derived in the previous section, the stationary distribution exists. Let

$$\pi_{i,j,k} = \lim_{t \to \infty} \Pr(S_1(t) = i, S_2(t) = j, N(t) = k), \qquad (i, j, k) \in \mathcal{S},$$

$$\boldsymbol{\pi}_{i,k} = (\pi_{i,0,k}, \pi_{i,1,k}, \dots, \pi_{i,c-i,k}),$$

$$\boldsymbol{\pi}_k = (\boldsymbol{\pi}_{0,k}, \boldsymbol{\pi}_{1,k}, \dots, \boldsymbol{\pi}_{c,k}).$$

According to the matrix-analytic method, we have

$$\boldsymbol{\pi}_{k+1} = \boldsymbol{\pi}_k R, \qquad k \in \mathbb{Z}_+,$$

where the matrix R is the minimal nonnegative solution of

$$A^+ + AR + A^- R^2 = O,$$

for which several efficient numerical algorithms are available [14]. For example, R can be obtained as $\lim_{n \to \infty} R_n$, where $\{R_n; n \in \mathbb{Z}_+\}$ is defined by

$$R_0 = O, \qquad R_{n+1} = A^{-1} A^+ + A^{-1} A^- R_n^2, \qquad n \in \mathbb{Z}_+.$$

Furthermore, R can be also obtained by the matrix continued fraction approach presented in [15]. Finally, the boundary vector $\boldsymbol{\pi}_0$ is determined by

$$\boldsymbol{\pi}_0 A^0 + \boldsymbol{\pi}_1 A^- = \mathbf{0}, \qquad \sum_{k=0}^{\infty} \boldsymbol{\pi}_k e = 1,$$

which is equivalent to

$$\boldsymbol{\pi}_0 (A^0 + R A^-) = \mathbf{0}, \qquad \boldsymbol{\pi}_0 (I - R)^{-1} e = 1.$$

4.4   First Moments and Cost Model

We define the generating function for $\{\boldsymbol{\pi}_k; k \in \mathbb{Z}_+\}$ as

$$\boldsymbol{\pi}(z) = \sum_{k=0}^{\infty} \boldsymbol{\pi}_k z^k = \boldsymbol{\pi}_0 (I - zR)^{-1}.$$

Let $\boldsymbol{M}_n$, $n \in \mathbb{Z}_+$ denote the nth factorial moment vector of partial factorial moments $M_n^{(i,j)}$, $(i, j) \in \mathcal{V}$. We then have

$$\boldsymbol{M}_n = \frac{d}{dz} \boldsymbol{\pi}(z) \Big|_{z=1} = \boldsymbol{\pi}_0 n! (I - R)^{-(n+1)} R^n.$$

Let $\pi_{i,j}^S = \lim_{t \to \infty} \Pr(S_1(t) = i, S_2(t) = j)$, $(i, j) \in \mathcal{V}$. We also define

$$\boldsymbol{\pi}_i^S = (\pi_{i,0}^S, \pi_{i,1}^S, \dots, \pi_{i,c-i}^S), \qquad \boldsymbol{\pi}^S = (\boldsymbol{\pi}_0, \boldsymbol{\pi}_1, \dots, \boldsymbol{\pi}_c).$$

We then have

$$\boldsymbol{\pi}^S = \sum_{k=0}^{\infty} \boldsymbol{\pi}_k = \boldsymbol{\pi}_0 (I - R)^{-1}.$$

Let $\mathrm{E}[S_1]$ and $\mathrm{E}[S_2]$ denote the average number of incoming and outgoing calls in the servers, respectively. We have

$$\mathrm{E}[S_1] = \sum_{i=0}^{c} i \sum_{j=0}^{c-i} \pi_{i,j}^S, \qquad \mathrm{E}[S_2] = \sum_{i=0}^{c} \sum_{j=0}^{c-i} \pi_{i,j}^S j.$$

On the other hand, Little's formula yields

$$\mathrm{E}[S_1] = \frac{\lambda}{\nu_1}.$$

Let U denote the utilization of a server at the steady state, i.e.,

$$U = \frac{\mathrm{E}[S_1] + \mathrm{E}[S_2]}{c} = \frac{\lambda}{c\nu_1} + \frac{\mathrm{E}[S_2]}{c}.$$

From a management point of view, we need to minimize $1-U$, i.e., the fraction of time where the server is idle. At the same time, from a service point of view, we also need to minimize the average number of customers in the orbit $\mathrm{E}[N] = \boldsymbol{M}_1 \boldsymbol{e}$. These needs motivate us to consider an optimization problem finding the optimal value of the rate of outgoing calls.

$$\begin{aligned} \min \quad & f(\alpha) = C_1(1 - U) + C_2 \mathrm{E}[N], \\ s.t. \quad & \boldsymbol{p} A^+ \boldsymbol{e} < \boldsymbol{p} A^- \boldsymbol{e}, \qquad \alpha \geq 0, \end{aligned}$$

where the inequality is the stability condition, and C_1 and C_2 reflect the cost of idle server and of a retrial customer. Similar to the single-server case, the optimization consists in finding the optimal α while keeping all other parameters constant.

5   Conclusion

This paper presents the analysis of a two way communication retrial queue model applicable to a call center with balanced call blending. By assuming a constant retrial rate for the incoming calls, outgoing call activity is still possible when many incoming calls are in orbit, corresponding to balanced blending.

For the single server case, we derived the partial generating functions associated with the joint stationary distribution of the number of incoming calls and the system state. From this, we extracted explicit (closed-form) expressions for the involved probabilities, and also for the partial factorial moments. Both were also characterized with a recursive formulation. Further, the system's stability condition was derived, and a cost model was proposed. For the multiserver case, a formulation by a quasi-birth-and-death process was assumed. The involved matrices were derived, as well as an expression for the multiserver stability condition. Finally, also a numerical recipe for the stationary distribution was presented, and an associated cost model was proposed.

References

1. Falin, G.I., Templeton, J.G.C.: Retrial Queues. Chapman and Hall, Boca Raton (1997)
2. Artalejo, J.R., Gomez-Córral, A.: Retrial Queueing Systems: A Computational Approach. Springer, Berlin (2008)
3. Phung-Duc, T., Masuyama, H., Kasahara, S., Takahashi, Y.: State-dependent M/M/c/c+r retrial queues with bernoulli abandonment. Journal of Industrial and Management Optimization 6(3), 517–540 (2010)
4. Koole, G., Mandelbaum, A.: Queueing models of call centers: an introduction. Annals of Operations Research 113(1), 41–59 (2002)
5. Bhulai, S., Koole, G.: A queueing model for call blending in call centers. IEEE Transactions on Automatic Control 48, 1434–1438 (2003)
6. Artalejo, J.R., Phung-Duc, T.: Markovian single server retrial queues with two way communication. In: Proceedings of the 6th International Conference on Queueing Theory and Network Applications, QTNA 2011, pp. 1–7. ACM, New York (2011)
7. Bernett, H.G., Fischer, M.J., Masi, D.M.B.: Blended call center performance analysis. IT Professional 4, 33–38 (2002)
8. Deslauriers, A., L'Ecuyer, P., Pichitlamken, J., Ingolfsson, A., Avramidis, A.N.: Markov chain models of a telephone call center with call blending. Computers & Operations Research 34, 1616–1645 (2007)
9. Atencia, I., Moreno, P.: A single-server retrial queue with general retrial times and Bernoulli schedule. Applied Mathematics and Computation 162(2), 855–880 (2005)
10. Atencia, I., Fortes, I., Moreno, P., Sánchez, S.: An M/G/1 retrial queue with active breakdowns and Bernoulli schedule in the server. Information and Management Sciences 17(1), 1–17 (2006)
11. Fayolle, G.: A simple telephone exchange with delayed feedbacks. In: Proc. of the International Seminar on Teletraffic Analysis and Computer Performance Evaluation, pp. 245–253. North-Holland Publishing Co., Amsterdam (1986)
12. Farahmand, K.: Single line queue with repeated demands. Queueing Systems 6, 223–228 (1990)
13. Martin, M., Artalejo, J.R.: Analysis of an M/G/1 queue with two types of impatient units. Advances in Applied Probability 27(3), 840–861 (1995)
14. Latouche, G., Ramaswami, V.: Introduction to Matrix Analytic Methods in Stochastic Modelling. SIAM, Philadelphia (1999)
15. Phung-Duc, T., Masuyama, H., Kasahara, S., Takahashi, Y.: A simple algorithm for the rate matrices of level-dependent QBD processes. In: Proceedings of the 5th International Conference on Queueing Theory and Network Applications, QTNA 2010, pp. 46–52. ACM, New York (2010)