Noname manuscript No.
(will be inserted by the editor)

Performance analysis of an unreliable $M/G/1$ retrial queue with two-way communication

Author 1 · **Author 2** · **Author 3**

Received: date / Accepted: date

Abstract Efficient use of call center operators through technological innovations more often come at the expense of added operation management issues. In this paper, the stationary characteristics of an $M/G/1$ retrial queue is investigated where the single server, subject to active failures, primarily attends incoming calls and directs outgoing calls only when idle. The incoming calls arriving at the server follow a Poisson arrival process, while outgoing calls are made in an exponentially distributed time. On finding the server unavailable (either busy or temporarily broken down), incoming calls intrinsically join the virtual orbit from which they re-attempt for service at exponentially distributed time intervals. The system stability condition along with probability generating functions for the joint queue length distribution of the number of calls in the orbit and the state of the server are derived and evaluated numerically in the context of mean system size, server availability, failure frequency and orbit waiting time.

Keywords Retrial queueing system · Server breakdown · Coupled switching · Performance evaluation · Steady-state distribution

Mathematics Subject Classification (2000) 60K25 · 62N05

1 Introduction

Blended call centers have recently evolved as an effective and profitable communication asset in bridging companies and their customers. Unlike conventional call centers, such modern communication systems are capable of managing a mixture of both, inbound and outbound call operations that require instant service (Bhulai and Koole 2003; Aksin et al. 2007). An outgoing call is initiated by the server only when no incoming call is in the system. This feature, commonly referred to as *coupled switching* or *two-way communication*, yields higher productivity by reducing the idle time experienced by the serving operator (Artalejo and Phung-Duc 2012; Legros et al. 2017). Moreover, incoming calls that find the server busy enter a virtual orbit and tend to retry for service after some random time (Artalejo and Gomez-Corral 2008). As a result, in-depth analysis of the influence of retrying customer calls on the dynamics of coupled switching in call centers is of great significance

Address(es) of author(s) should be given

not only to the research community, but also serves as guidelines to statistical practitioners, network managers and system administrators.

Current advancement in scale and scope of call centers as socio-technical systems has initiated the need for formulation and analysis of more refined queueing models. The range of seminal works dedicated to coupled switching in the retrial queueing literature is relatively diverse (see Artalejo (2010) for a comprehensive overview). In the study conducted by Choi et al. (1995), some expected performance measures for an $M/G/1/K$ priority retrial queue with coupled switching were derived under the assumption that incoming and outgoing calls follow the same service time distribution. Nevertheless, such an assumption limits the practicality of the model as customers may have different service needs. Although Artalejo and Resing (2010) derived the first partial moments for an $M/G/1/1$ retrial queue model with general service time distributions using mean value analysis, it cannot be used to obtain the stationary distribution and factorial moments. Comprehensive analysis of the $M/M/1/1$ retrial queue with coupled switching and different service time distributions for single and multiple server cases have been reported by Artalejo and Phung-Duc (2012). The work was further extended to incorporate multiple types of outgoing calls by Sakurai and Phung-Duc (2015), for which the joint stationary distribution of the number of calls in the orbit and the state of the server were obtained both, asymptotically and recursively. Furthermore, Artalejo and Phung-Duc (2013) proposed an embedded Markov chain approach to study the steady-state behavior of a couple-switched $M/G/1$ retrial queue with tailed asymptotic analysis of number of customers residing in the orbit. Nonetheless, the server may undergo multiple failures and resume service upon repair (Krishnamoorthy et al. (2014)). Despite its prevalence in practice, research efforts to address server failure in blended call center system models are scarce.

The reliability of an $M/G/1$ retrial queue with only inbound calls and server breakdowns was investigated by Wang et. al. (2001). In another related work by Martin and Artalejo (1995), an $M/G/1$ service system model with two types of impatient units was exhaustively analyzed. The equilibrium balking strategies of customers in the $M/M/1$ queue with set-up times, breakdowns and reparis were scrutinized by Chen and Zhou (2015). Moreover, Ouazine and Abbas (2016) reported a functional approximation of the stationary performance of the $M1,M2/G1,G2/1$ retrial queue with two-way communication and finite orbit capacity. Perfect and imperfect repair of a single server $M/M/1/1$ retrial queue with only incoming customers were modeled by Chang and Wang (2017). A more recent work by Phung-Duc (2017) successfully identified higher order moments for single server retrial queues with set-up time using Taylor series method. To our best knowledge, however, explicit reliability indices for an unreliable $M/G/1$ retrial queue with two-way communication have not yet been analytically derived in modeling service systems. Hence, continuous-time analytical characterization of an unreliable single server retrial queue with coupled switching is imperative from the viewpoint of both, queueing as well as reliability analysis.

The foremost goal of this letter is to study the impact of server failure on the steady-state performance of the $M/G/1$ queue with two-way communication having an orbit with infinite capacity and generally distributed server repair time. In particular, we obtain the system stability condition using the embedded Markov chain technique, followed by the supplementary variable approach to obtain in closed-form the probability generating functions (pgfs) for incoming calls in orbit and in the system, followed by their second order moments. The numerical simulations conducted for various performance metrics of interest corroborate the theoretical findings of the proposed system model.

The rest of the paper is organized as follows: Sect. 2 is dedicated to the description and mathematical formulation of the unreliable $M/G/1$ retrial queue with two-way commu-

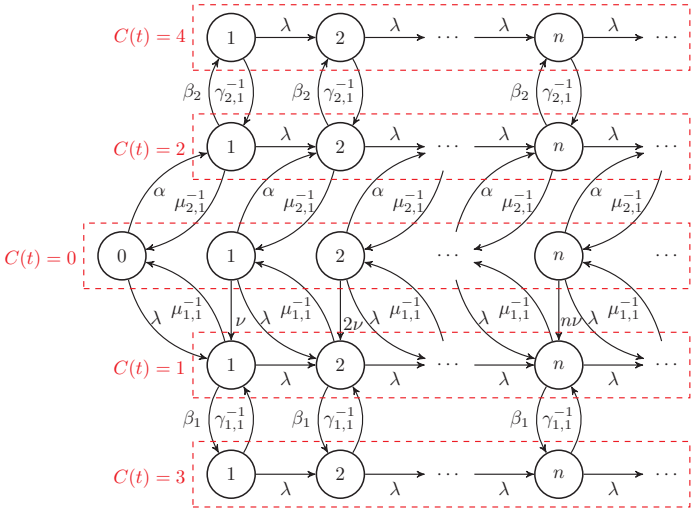


Fig. 1 State transitions of the proposed system model with each row highlighting the state of the server.

nication. In Sect. 3, the necessary and sufficient condition for system stability is presented, followed by derivation of the corresponding steady-state distribution. Various system performance measures as well as reliability indices are derived in Sect. 4. Numerical simulation results are discussed in Sect. 5. Finally, Sect. 6 concludes the paper with suggestions for potential future research directions.

2 System model formulation

We consider a single server retrial queue in which primary inbound calls follow a Poisson arrival process with rate λ. If the server is idle, an outgoing call is initiated after an exponentially distributed time with rate α. As in reality, the time taken to serve incoming and outgoing calls is assumed to be different. If an incoming call finds the server busy, it then enters the orbit and re-attempts to seek service after an exponentially distributed time with rate ν. Otherwise, the incoming call commences service immediately. Since the server may breakdown while serving calls, without loss of generality, we assume that the lifetime of the server follows an exponential distribution with rates β_1 and β_2 during the service of inbound and outbound calls, respectively. On failure, the server is instantly sent for repair which has a generally distributed time. The corresponding state transition diagram of the proposed model is given in Fig. 1, where $n \geq 0$ is the number of incoming calls waiting in the orbit.

For the sake of consistency, we define $i \in \{1,2\}$ to differentiate between incoming and outgoing calls. Henceforth, $i = 1$ refers to incoming calls, while $i = 2$ indicates outgoing calls. Let $S_i(x)$ and $R_i(x)$ be the cumulative distributions of service and repair times of i-type calls, respectively. Similarly, let $s_i(x)$ and $r_i(x)$ denote respectively, the probability density functions of service and repair times of i-type calls. The Laplace transforms of the service and repair times for each type of call are denoted as $\tilde{S}_i(\theta)$ and $\tilde{R}_i(\theta)$, respectively. We also define $S_i^o(x)$ and $R_i^o(x)$ as the remaining service and repair times, respectively. Moreover, let $\mu_{i,k}$ and $\gamma_{i,k}$ denote the k^{th} moment of service and repair times, respectively. In what follows, the arrival flows of incoming calls, outgoing calls, service time, repair time,

and intervals between successive re-attempts are all assumed to be mutually independent. Finally, let $N(t)$ be the number of incoming customer calls in orbit and $M(t)$ be the total number of customers in the system at time t. We now define the state of the server, denoted by $C(t)$, to be as follows:

$$C(t) = \begin{cases} 0, & \text{if the server is } idle, \\ 1, & \text{if the server is } busy \text{ serving an } incoming\ call, \\ 2, & \text{if the server is } busy \text{ making an } outgoing\ call, \\ 3, & \text{if the server } fails \text{ while serving an } incoming\ call, \\ 4, & \text{if the server } fails \text{ while making an } outgoing\ call. \end{cases}$$

For service and repair times that are exponentially distributed, the state transitions for $\{(C(t),N(t));t \geq 0\}$ on the state space $S=\{0,1,2,3,4\}\times\mathbb{Z}_+$ are as shown in Fig. 1, where $\mathbb{Z}_+$ denotes the set of non-negative integers. In the case of generally distributed service and repair times, the Markov process $\{(C(t),N(t),S_1^o(t),S_2^o(t),R_1^o(t),R_2^o(t));t \geq 0\}$ can be used to describe the state of the system. Based on this generalized definition, the state probabilities are given by:

$$P_{0,n}(t) = \Pr[C(t)=0,\ N(t)=n], \tag{1}$$

$$P_{1,n}(x,t)\,dx = \Pr[C(t)=1,\ N(t)=n,\ x < S_1^o(t) \leq x+dx], \tag{2}$$

$$P_{2,n}(x,t)\,dx = \Pr[C(t)=2,\ N(t)=n,\ x < S_2^o(t) \leq x+dx], \tag{3}$$

$$P_{3,n}(x,y,t)\,dy = \Pr[C(t)=3,\ N(t)=n,\ S_1^o(t)=x,\ y < R_1^o(t) \leq y+dy], \tag{4}$$

$$P_{4,n}(x,y,t)dy = \Pr[C(t)=4,\ N(t)=n,\ S_2^o(t)=x,\ y < R_2^o(t) \leq y+dy], \tag{5}$$

where $x,y \geq 0$ are time epochs. In (1), $P_{0,n}(t)$ is the probability of the server being idle while having n calls in the orbit at time t. For $i \in \{1,2\}$ in (2) and (3), $P_{i,n}(x,t)\,dx$ denotes the joint probability that the server is busy with an i-type call during the remaining service time $(x,x+dx)$ and there are n calls residing in the orbit at time t. Likewise, for $j \in \{3,4\}$ in (4) and (5), $P_{j,n}(x,y,t)\,dy$ refers to the joint probability that at time t there are n calls residing in the orbit, the remaining service time is x, and the failed server is fixed within the remaining repair time $(y,y+dy)$ while serving an inbound $(j=3)$ or an outbound $(j=4)$ call.

3 Steady-state distribution

In this section, we identify the pgfs of orbit size and number of incoming calls in the system. To do so, we first determine the system stability condition using the following theorem.

Theorem 1 *The necessary and sufficient condition for system stability is given by the inequality* $\lambda\mu_{1,1}(1+\beta_1\gamma_{1,1}) < 1$.

Proof Let $\hat{X}_n$ be the service completion time of the n^{th} call which includes possible down times (due to server failure) while providing service. For the sufficient condition, we need to prove the ergodicity of $\{L_n; n \geq 1\}$, where $\{L_n\}$ is an irreducible and aperiodic discrete-time Markov chain of $\{(C(t),N(t),S_1^o(t),S_2^o(t),R_1^o(t),R_2^o(t));t \geq 0\}$ and is defined as $L_n = N(\hat{X}_n^+)$. Using Foster's criterion and undertaking the same approach as in Artalejo and Phung-Duc (2013), $\{L_n\}$ is positive recurrent if $|\eta_k| < \infty$ and $\lim_{k\to\infty} \sup\{\eta_k\} < 0$ for all k, where $\eta_k = E[(L_{n+1}-L_n)|L_n=k]$. By conditioning on the identity of the n^{th} call, we arrive at:

$$\eta_k = \frac{kv[\lambda\mu_{1,1}(1+\beta_1\gamma_{1,1})-1]}{\lambda+kv+\alpha} + \frac{\lambda[\lambda\mu_{1,1}(1+\beta_1\gamma_{1,1})]}{\lambda+kv+\alpha} + \frac{\alpha[\lambda\mu_{2,1}(1+\beta_2\gamma_{2,1})]}{\lambda+kv+\alpha}. \tag{6}$$

It is straightforward to observe that for all k values, $\eta_k < \infty$ and $\lim_{k\to\infty} \sup\{\eta_k\} < 0$ if $\lambda\mu_{1,1}(1+\beta_1\gamma_{1,1}) < 1$, which proves the sufficiency criteria.

As pointed out in Sennott et al. (1983), the non-ergodicity of $\{L_n\}$ can be guaranteed if Kaplan's condition is satisfied, i.e. there exists some $k_0 \in Z_+$ such that $\eta_k \geq 0$ for $k \geq k_0$ and $\eta_k < \infty$ for all $k \geq 0$. In our setting, this condition is satisfied as $r_{i,j} = 0$ for $j < i-1$, where $P = [r_{i,j}]$ is the one-step transition probability matrix. Hence, $\lambda\mu_{1,1}(1+\beta_1\gamma_{1,1}) \geq 1$ implies the non-ergodicity of $\{L_n; n \geq 1\}$, which completes the proof. □

Adopting the supplementary variable technique, the system of balance equations for (1)-(5) is obtained in terms of the limiting state probabilities as follows:

$$(\lambda + n\nu + \alpha)P_{0,n} = P_{1,n}(0) + P_{2,n}(0)\,, \tag{7}$$

$$P'_{1,n}(x) = (\lambda+\beta_1)P_{1,n}(x) - \lambda P_{0,n}s_1(x) - \lambda P_{1,n-1}(x) - (j+1)\nu P_{0,n+1}s_1(x) - P_{3,n}(x,0)\,, \tag{8}$$

$$P'_{2,n}(x) = (\lambda+\beta_2)P_{2,n}(x) - \lambda P_{2,n-1}(x) - \alpha P_{0,n}s_2(x) - P_{4,n}(x,0)\,, \tag{9}$$

$$\frac{\partial}{\partial y}P_{3,n}(x,y) = -\lambda[P_{3,n-1}(x,y) - P_{3,n}(x,y)] - \beta_1 P_{1,n}(x)r_1(y)\,, \tag{10}$$

$$\frac{\partial}{\partial y}P_{4,n}(x,y) = -\lambda[P_{4,n-1}(x,y) - P_{4,n}(x,y)] - \beta_2 P_{2,n}(x)r_2(y)\,, \tag{11}$$

with the following normalizing condition, where the terms $W(x)$ and $Z(x,y)$ respectively, stand for $P_{1,n}(x) + P_{2,n}(x)$ and $P_{3,n}(x,y) + P_{4,n}(x,y)$:

$$\sum_{n=1}^{\infty}\left[P_{0,n} + \int_0^{\infty} W(x)dx + \int_0^{\infty}\int_0^{\infty} Z(x,y)dxdy\right] = 1, \tag{12}$$

Taking the Laplace transforms $\mathscr{L}\{\cdot\}$ of (7)-(11) results in the following marginal generating functions, where notations $\tilde{P}_{i,n}(\theta)$ and $\widetilde{\tilde{P}}_{j,n}(\theta,s)$ are used to denote $\mathscr{L}\{P_{i,n}(x)\}$ and $\mathscr{L}\{\mathscr{L}\{P_{j,n}(x,y)\}\}$, respectively:

$$P_0(z) = \sum_{n=0}^{\infty} P_{0,n}\,z^n\,, \tag{13}$$

$$\tilde{P}_i(z,\theta) = \sum_{n=0}^{\infty} \tilde{P}_{i,n}(\theta)\,z^n\,, \qquad i \in \{1,2\}\,, \tag{14}$$

$$P_i(z,0) = \sum_{n=0}^{\infty} P_{i,n}(0)\,z^n\,, \qquad i \in \{1,2\}\,, \tag{15}$$

$$\widetilde{\tilde{P}}_j(z,\theta,s) = \sum_{n=0}^{\infty} \widetilde{\tilde{P}}_{j,n}(\theta,s)\,z^n\,, \qquad j \in \{3,4\}\,, \tag{16}$$

$$\tilde{P}_j(z,\theta,0) = \sum_{n=0}^{\infty} \tilde{P}_{j,n}(\theta,0)\,z^n\,, \qquad j \in \{3,4\}\,. \tag{17}$$

Subsequently, the following pgfs are obtained through some algebraic manipulations, with function definitions $\phi(z) \triangleq \exp\big(-\int_z^1 \frac{[\lambda(1-\delta_1(u))+\alpha(1-\delta_2(u))]}{\nu[\delta_1(u)-u]}du\big)$, $\delta_i(\cdot) \triangleq \tilde{S}_i(h_i(\cdot))$ and

$h_i(z) \triangleq \lambda + \beta_i - \lambda z - \beta_i \tilde{R}_i(\lambda - \lambda z)$ for $i \in \{1,2\}$:

$$P_0(z) = \frac{1-\lambda\mu_{1,1}(1+\beta_1\gamma_{1,1})}{1+\alpha\mu_{2,1}(1+\beta_2\gamma_{2,1})}\phi(z), \tag{18}$$

$$\tilde{P}_1(z,0) = \frac{[\lambda(1-z)+\alpha(1-\delta_2(z))][1-\delta_1(z)]}{[\delta_1(z)-z]h_1(z)}P_0(z), \tag{19}$$

$$\tilde{P}_2(z,0) = \frac{\alpha[1-\delta_2(z)]}{h_2(z)}P_0(z), \tag{20}$$

$$\widetilde{\tilde{P}}_3(z,0,0) = \frac{\beta_1[1-\tilde{R}_1(\lambda-\lambda z)]}{\lambda-\lambda z}\tilde{P}_1(z,0), \tag{21}$$

$$\widetilde{\tilde{P}}_4(z,0,0) = \frac{\beta_2[1-\tilde{R}_2(\lambda-\lambda z)]}{\lambda-\lambda z}\tilde{P}_2(z,0). \tag{22}$$

In steady-state, the pgfs of orbit occupancy size, $P(z)$, and system size, $R(z)$, at an arbitrary epoch can now be expressed in terms of the equations derived in (18)-(22). By ignoring the non-zero probability of server failure, it is straightforward to show that the following results are in complete agreement with Artalejo and Phung-Duc (2013):

$$\begin{aligned} P(z) &= P_0(z) + \tilde{P}_1(z,0) + \tilde{P}_2(z,0) + \widetilde{\tilde{P}}_3(z,0,0) + \widetilde{\tilde{P}}_4(z,0,0) \\ &= \frac{\lambda(1-z)+\alpha[1-\delta_2(z)]}{\lambda[\delta_1(z)-z]}P_0(z), \end{aligned} \tag{23}$$

$$\begin{aligned} R(z) &= P_0(z) + z[\tilde{P}_1(z,0) + \tilde{P}_2(z,0) + \widetilde{\tilde{P}}_3(z,0,0) + \widetilde{\tilde{P}}_4(z,0,0)] \\ &= \frac{(\lambda-\lambda z)\delta_1(z)+\alpha[1-\delta_2(z)]}{\lambda[\delta_1(z)-z]}P_0(z). \end{aligned} \tag{24}$$

4 Performance and reliability analysis

In this section, some performance and reliability metrics for the queueing system under study are discussed. Specifically, our performance analysis involves finding the expected number of calls in the system and expected waiting time in the orbit, while server availability and server failure frequency characterize the reliability indices. To improve readability, we introduce and use the notations $\rho_1 = (1+\beta_1\gamma_{1,1})$, $\rho_2 = (1+\beta_2\gamma_{2,1})$, $\sigma_1 = \lambda\rho_1\mu_{1,1}$, and $\sigma_2 = \alpha\rho_2\mu_{2,1}$ throughout this section.

4.1 Expected number of customer calls in the system

This measure accounts for the mean number of incoming calls retrying for service, either due to server failure or it being busy, as well as those being served by the server. This is readily obtained by differentiating the pgfs in (23) and (24) before evaluating them at $z=1$. As a result, the equation given in (23) yields the first moment of the orbit size as follows:

$$\begin{aligned} E[N] &= P'(1) \\ &= \frac{\lambda^2(\beta_1\mu_{1,1}\gamma_{1,2}+\rho_1^2\mu_{1,2})}{2(1-\sigma_1)} + \frac{\lambda\alpha(\beta_2\mu_{2,1}\gamma_{2,2}+\rho_2^2\mu_{2,2})}{2(1+\sigma_2)} + \frac{\lambda(\sigma_1+\sigma_2)}{\nu(1-\sigma_1)}, \end{aligned} \tag{25}$$

Similarly, the mean system size resulting from (24) is given as:

$$E[M] = R'(1) = P'(1) + \frac{\sigma_1 + \sigma_2}{1+\sigma_2}. \tag{26}$$

By differentiating (23) and (24) twice with respect to z and evaluating them at $z=1$, we also obtain the second order moment of the orbit size as follows:

$$E[N^2] = P''(1) = M_1 + M_2 + M_3, \tag{27}$$

where M_1, M_2, and M_3 are derived to be:

$$\begin{aligned} M_1 =& \frac{\lambda^2(\sigma_1+\sigma_2)^2}{\nu^2(1-\sigma_1)(1+\sigma_2)} + \frac{\lambda^3(1+\sigma_2)(\beta_1\mu_{1,1}\gamma_{1,2}+\rho_1\mu_{1,2})}{2\nu(1-\sigma_1)(1+\sigma_2)} \\ &+ \frac{\lambda^2\alpha(1-\sigma_1)(\beta_2\mu_{2,1}\gamma_{2,2}+\rho_2\mu_{2,2})}{2\nu(1-\sigma_1)(1+\sigma_2)}, \end{aligned} \tag{28}$$

$$\begin{aligned} M_2 =& \frac{\lambda^3(\beta_1\gamma_{1,3}\mu_{1,1}+2\rho_1\beta_1\gamma_{1,2}\mu_{1,2}+\rho_1^3\mu_{1,3})}{3(1-\sigma_1)} + \frac{\lambda^4(\beta_1\mu_{1,1}\gamma_{1,2}+\rho_1\mu_{1,2})^2}{2(1-\sigma_1)^2} \\ &+ \frac{\lambda^3(\beta_1\mu_{1,1}\gamma_{1,2}+\rho_1\mu_{1,2})}{2(1-\sigma_1)^2}\left(\frac{\sigma_1+\sigma_2}{\nu} + \frac{\alpha\big(1-\sigma_1(\beta_2\mu_{2,1}\gamma_{2,2}+\rho_2\mu_{2,2})\big)}{2(1+\sigma_2)}\right) \\ &+ \frac{\lambda^2\sigma_1}{1-\sigma_1}\left(\frac{\alpha\rho_2\big(1-\sigma_1(\beta_2\gamma_{2,3}\mu_{2,1}+2\rho_2\beta_2\gamma_{2,2}\mu_{2,2}+\rho_2^3\mu_{2,3})\big)}{3(1+\sigma_2)}\right) \\ &+ \frac{\lambda^2\sigma_1(\sigma_1+\sigma_2)(\beta_2\mu_{2,1}\gamma_{2,2}+\rho_2\mu_{2,2})}{\nu(1-\sigma_1)} + \frac{M_4}{\nu(1-\sigma_1)}, \end{aligned} \tag{29}$$

$$\begin{aligned} M_3 =& \frac{\lambda^2\alpha(1-\sigma_1)(\beta_1\gamma_{1,3}\mu_{1,1}+2\rho_1\beta_1\gamma_{1,2}\mu_{1,2}+\rho_1^3\mu_{1,3})}{3(1+\sigma_2)} \\ &+ \frac{\lambda^2\alpha(\sigma_1+\sigma_2)(\beta_2\mu_{2,1}\gamma_{2,2}+\rho_1\mu_{2,2})}{\nu(1+\sigma_2)} + \frac{\lambda^2\alpha\sigma_2 M_4}{\nu(1-\sigma_1)(1+\sigma_2)}, \end{aligned} \tag{30}$$

with M_4 given as below:

$$M_4 = \frac{(\sigma_1+\sigma_2)^2}{\nu} + \frac{\lambda(1+\sigma_2)(\beta_1\mu_{1,1}\gamma_{1,2}+\rho_1\mu_{1,2})}{2} + \frac{\alpha(1-\sigma_1)(\beta_2\mu_{2,1}\gamma_{2,2}+\rho_2\mu_{2,2})}{2}. \tag{31}$$

4.2 Expected waiting time in orbit

Denoted by W, the steady-state delay experienced by an incoming customer call in obit depends on the total idle time of the server not serving an incoming call (W_0), the total service time (including the server failure time) of the server providing service to an incoming call (W_1), and the total service time (including the server failure time) of the server busy with an outgoing call (W_2). The probability of an inbound call entering the orbit (P_w) is thus, calculated as follows:

$$P_w = \lim_{z\to 1}\{\tilde{P}_1(z,0) + \tilde{P}_2(z,0) + \widetilde{\tilde{P}}_3(z,0,0) + \widetilde{\tilde{P}}_4(z,0,0)\} = \frac{(\sigma_1+\sigma_2)}{(1+\sigma_2)}. \tag{32}$$

Using the above equation and the first moments of the pgfs in (18)-(22), we derive the mean waiting time in the orbit to be (see Choi et al. (1995)):

$$\mathrm{E}[W] = \mathrm{E}[W_0] + \mathrm{E}[W_1] + \mathrm{E}[W_2] = \mathrm{E}[N]/\lambda\,, \tag{33}$$

where $\mathrm{E}[W_0] = P_w/\nu$, $\mathrm{E}[W_1] = \mathrm{E}[N]\mathrm{E}[B_1] + \sigma_1\mathrm{E}[R_1]$, and $\mathrm{E}[W_2] = \sigma_2(1-\sigma_1)\mathrm{E}[R_2]/(1+\sigma_2) + \sigma_2\mathrm{E}[W_0]$. Note that the notations $\mathrm{E}[B_i]$ and $\mathrm{E}[R_i]$ represent the mean service time (including failure time) and the mean remaining service time (including failure time) while serving i-type calls, which are given as $\mu_{i,1}(1+\beta_i\gamma_{i,1})$ and $\mathrm{E}[B_i^2]/(2\mathrm{E}[B_i])$, respectively.

4.3 Server availability

Availability is a system characteristic that measures how often the server is available for use, even though it may not be functioning properly. The probability that the server is operational at a given time instant t is defined as its point-wise availability, $A(t)$, and its steady state availability (i.e. $\lim_{t\to\infty} A(t) = P_a$) is given as:

$$P_a = \lim_{z\to 1}\{P_0(z) + \tilde{P}_1(z,0) + \tilde{P}_2(z,0)\} = \frac{(1+\alpha\mu_{2,1})(1-\sigma_1) + \lambda\mu_{1,1}(1+\sigma_2)}{(1+\sigma_2)}\,. \tag{34}$$

4.4 Server failure frequency

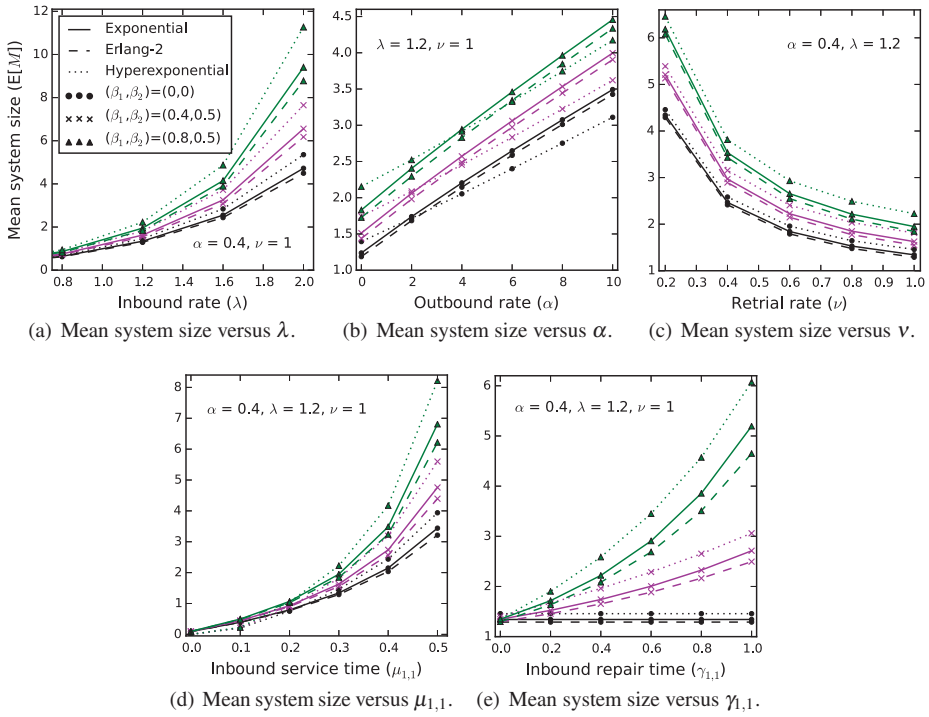
This measure corresponds to the probability that the server fails at time $t > 0$ given that it was operating at $t = 0$ (see SenthilKumar and Arumuganathan (2010)). It can be easily shown that the following closed-form expression results from (23):

$$P_f = \lim_{z\to 1}\{\beta_1\tilde{P}_1(z,0) + \beta_2\tilde{P}_2(z,0)\} = \lambda\mu_{1,1}\beta_1 + \alpha\mu_{2,1}\beta_2\frac{(1-\sigma_1)}{(1+\sigma_2)}\,. \tag{35}$$

5 Numerical Examples and Discussions

To illustrate the impact of system parameters on the performance primitives, we present numerical examples for service and repair times with three arbitrary distributions namely, exponential with density function $c_1 e^{-c_1 x}$, Erlangian of order two with density function $c_1^2 x e^{-c_1 x}$ and hyperexponential given as $ac_1 e^{-c_1 x} + (1-a)c_2 e^{-c_2 x}$, where $c_1, c_2 > 0$ and $0 \le a \le 1$. Throughout this section, we assume $\lambda = 1.2$, $\alpha = 0.4$, $\nu = 1$, $\mu_{2,1} = 0.1$, and $\gamma_{2,1} = 0.2$ to satisfy the ergodic condition of the analytical system. We also consider an $M/G/1$ retrial queue without server failure, i.e. $(\beta_1, \beta_2) = (0,0)$, as in Artalejo and Phung-Duc (2013) to serve as the baseline scenario for our comparison.

Fig. 2 shows the variation in mean system size ($E[M]$) as functions of the inbound arrival rate (λ), outbound rate (α), retrial rate (ν), and inbound service ($\mu_{1,1}$) and repair ($\gamma_{1,1}$) times. As evident in Fig.2(a), increase in the number of arriving calls reduces the chance of finding the server active and idle. Consequently, these unattended incoming calls enter the orbit to retry for service thus, increasing the average system size as shown in the figure. In comparison to the failure-free baseline scenario, we note that the system size of our model increases with the failure rate β_1 as λ increases. A similar relationship can be observed in Fig.2(b) between $E[M]$ and α as well. On the other hand, $E[M]$ steeply decreases initially



(a) Mean system size versus λ. (b) Mean system size versus α. (c) Mean system size versus ν.

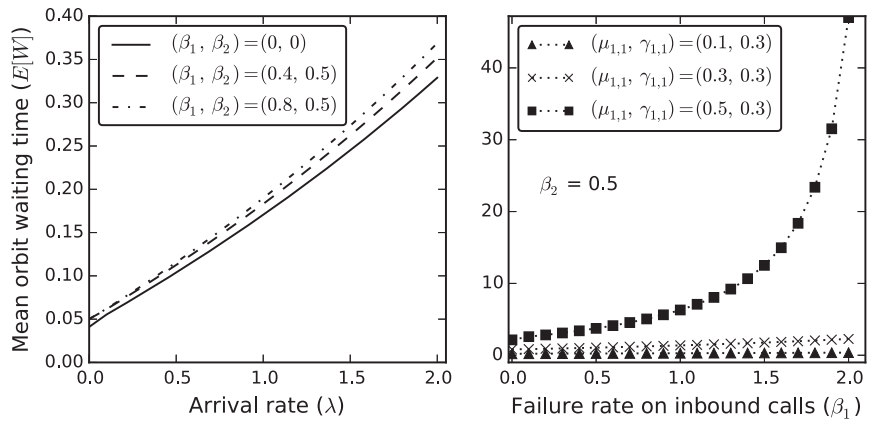
(d) Mean system size versus $\mu_{1,1}$. (e) Mean system size versus $\gamma_{1,1}$.

Fig. 2 Mean system size ($E[M]$) against incoming arrival rate (λ), outgoing rate (α), retrial rate (ν), inbound service time ($\mu_{1,1}$), and inbound repair time ($\gamma_{1,1}$).

and gradually stabilizes to some constant value with increase in the rate for service retrial as depicted in Fig.2(c). The observation justifies the fact that an incoming call residing in orbit has a higher chance of being served and thus, leaving the orbit if it re-attempts for service more frequently. For lower values of β_1, a primary incoming call is more probable to find the server available, resulting in a reduced orbit size. The influence of the service and repair times of primary calls on the mean system size are also illustrated in Fig.2(d) and Fig.2(e), respectively. As the average time to serve incoming calls increases, $E[M]$ grows steeper as depicted in Fig. 2(d). In other words, longer service times increases the number of incoming calls waiting in the orbit. Likewise, the shorter the server repair time, the more active it would be thus, reducing the system size which is mainly dominated by the orbit length. As seen in Fig.2(e), depending on the distribution type, the difference in $E[M]$ increases substantially with increase in the server repair time while serving an inbound call.

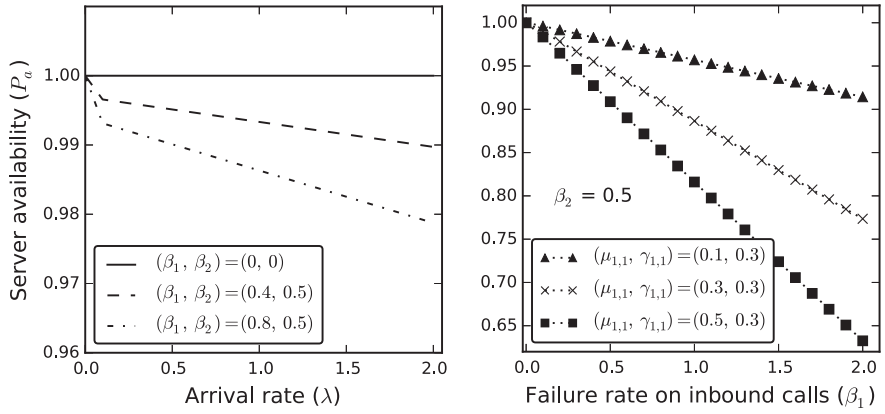
The mean orbit waiting time of incoming calls is given in Fig. 3 as functions of parameters λ and β_1. With respect to the benchmark, we observe the impact of server failure while serving incoming calls on the gradual increase in $E[W]$ in Fig.3(a). As the value of λ rises from 1 to 2, $E[W]$ shows a percentage increase of almost 133% for $(\beta_1,\beta_2) = (0.4,0.5)$ and nearly 111% for $(\beta_1,\beta_2) = (0.8,0.5)$. Moreover, as the value of β_1 increases to 2 in Fig.3(b), we observe that the average orbit delay grows exponentially with increase in $\mu_{1,1}$.

Fig. 4 shows the impact of λ and β_1 on server availability. It is noteworthy that all three distributions exhibit the same results for different values of β_1 and β_2. Therefore, they have



(a) Mean orbit waiting time versus λ. (b) Mean orbit waiting time versus β_1.

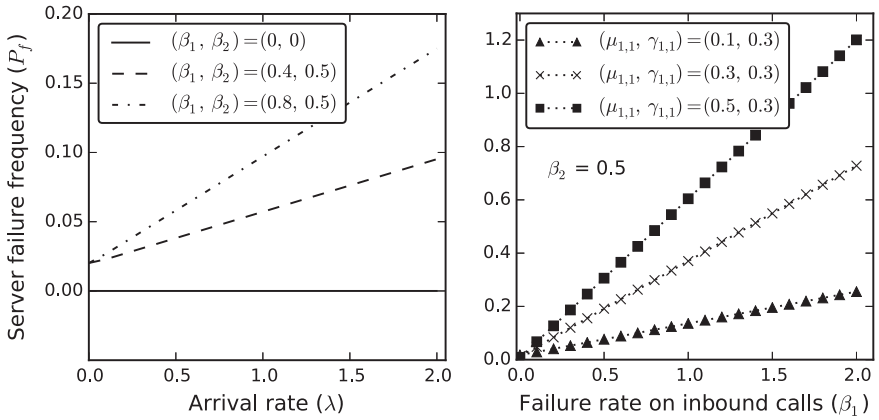
Fig. 3 Expected waiting time in orbit ($E[W]$) versus inbound arrival rate (λ) and server failure rate while serving such calls (β_1).



(a) Server availability versus λ. (b) Server availability versus β_1.

Fig. 4 Server availability (P_a) as a function of the inbound arrival rate (λ) and the server failure rate while serving such calls (β_1).

been demonstrated using a single plot. In absence of server failure, P_a is obviously always equal to 1. However, as β_1 increases, the availability of the server to incoming calls reduces with rise in λ. For instance, at $\lambda = 1.6$ in Fig. 4(a), P_a falls by slightly less than 1% as β_1 goes from 0.4 to 0.8. Fig.4(b) further portrays the prominent effect of β_1 on server availability under varying first moment of service and repair times. Note that there is a steeper fall in P_a as the service time of incoming calls increases. For instance, given $\beta_1 = 1$, P_a drops by approximately 15% as $\mu_{1,1}$ increases by a factor of 5. This measure further deteriorates by around 31.8% as the value of β_1 rises to 2. The figure reveals that the probability of finding the server available is higher when $\mu_{1,1} < \gamma_{1,1}$ and is more likely to reduce with increase in the inbound service time $\mu_{1,1}$.



(a) Server failure frequency versus λ. (b) Server failure frequency versus β_1.

Fig. 5 Server failure frequency (P_f) as a function of the inbound arrival rate (λ) and the server failure rate while serving such calls (β_1).

Similarly, Fig. 5(a) depicts the server failure frequency as functions of parameters λ and β_1. Apparent from (35), we observe that P_f monotonically increases with the number of incoming calls in our model. Additionally, at $\lambda = 2$, as β_1 increases from 0.4 to 0.8, the value of P_f rises drastically by over 74%. For various values of $(\mu_{1,1}, \gamma_{1,1})$, Fig.5(b) shows that P_f constantly increases with β_1 and is higher when $\mu_{1,1}$ is more than $\gamma_{1,1}$. In comparison to Fig.4, such behavior is not far from expectation as the measures P_f and P_a are inversely related.

6 Conclusion

In this paper, we have conducted an exhaustive steady-state analysis of the $M/G/1$ retrial queue incorporated with two-way communication and the possibility of server failure. Having immediate applications in blended call centers, our study of the stationary characteristics provided explicit expressions for the joint distribution of the server state and the expected number of incoming customer calls in the orbit. Results from extensive numerical simulations were provided for various performance measures to validate and compare our findings with that of a baseline system with no server breakdown. A promising follow-up on this work would be an extended analysis involving an unreliable multi-server retrial queueing model with prioritized classes of incoming service requests. The consideration of customer impatience in conjunction with service prioritization is yet another interesting direction for further exploration.

References

Aksin Z, Armony M, Mehrotra V (2007) The modern call center: a multi-disciplinary perspective on operations management research. Prod Oper Manag 16(6):665–688.

Artalejo JR (2010) Accessible bibliography on retrial queues: progress in 2000-2009. Math Comput Model 51(9-10):1071–1081.

Artalejo JR, Gomez-Corral A (2008) Retrial queueing systems: a compuational approach, Springer-Verlag Berlin Heidelberg.

Artalejo JR, Resing JAC (2010) Mean value analysis of single server retrial queues. Asia Pac J Oper Res 27(3):335–345.

Artalejo JR, Phung-Duc T (2012) Markovian retrial queues with two way communication. J Ind Manag Optim 8(4):781–806.

Artalejo JR, Phung-Duc T (2013) Single server retrial queues with two way communication. Appl Math Model 37(4):1811–1822.

Bhulai S, Koole G (2003) A queueing model for call blending in call centers. IEEE Trans Autom Control 48(8):1434–1438.

Chang J, Wang J (2017) Unreliable $M/M/1/1$ retrial queues with set-up time. Qual Technol Quant M https://doi.org/10.1080/16843703.2017.1320459.

Chen P, Zhou Y (2015) Equilibrium balking strategies in the single server queue with setup times and breakdowns. Oper Res Int J 15(2):213–231.

Choi BD, Choi K, Lee YW (1995) $M/G/1$ retrial queueing systems with two types of calls and finite capacity. Queueing Syst 19(1-2):215–229.

Legros B, Jouini O, Koole G (2017) Blended call center with idling times during the call service. IISE Transactions https://doi.org/10.1080/24725854.2017.1387318

Krishnamoorthy A, Pramod PK, Chakravarthy SR (2014) Queues with interruptions: a survey TOP 22(1):290–320.

Martin M, Artalejo JR (1995) Analysis of an M/G/1 queue with two types of impatient units. Adv Appl Probab 27(3):840–861.

Ouazine S, Abbas K (2016) A functional approximation for retrial queues with two way communication. Ann Oper Res 247(1):211–227.

Phung-Duc T (2017) Single server retrial queues with setup time. J Ind Manag Optim 13(3):1329–1345.

Sakurai H, Phung-Duc T (2015) Two-way communication retrial queues with multiple types of outgoing calls. TOP 23(2):466–492.

SenthilKumar M, Arumuganathan R (2010) An $M^X/G/1$ retrial queue with two phase service subject to active server breakdown and two types of repair. Int J Oper Res 8(3):261-291.

Sennott LI, Humblet PA, Tweedie RL (1983) Mean drifts and the non-ergodicity of Markov chains. Oper Res 31(4):783–789.

Sherman NP, Kharoufeh JP (2006) An $M/M/1$ retrial queue with unreliable server. Oper Res Lett 34(6):697–705.

Wang J, Cao J, Li Q (2001) Reliability analysis of the retrial queue with server breakdowns and repairs. Queueing Syst 38(4):363–380.