Annals of Operations Research
https://doi.org/10.1007/s10479-023-05263-z

ORIGINAL RESEARCH

A single server retrial queue with event-dependent arrival rates

Ioannis Dimitriou[1]

Accepted: 24 February 2023

Abstract
In this work, we consider a novel single-server retrial queue with event-dependent arrival rates. Contrary to other related works, the primary customers' arrival rates depend on the last realized event, which refers either to a departure, or to an arrival of either type, or to when a customer arrives during a busy period, compared with others. Our motivation stems from the modeling of service systems, in which the customers express their willingness to join the system based on the last realized event. We investigate the stability conditions, and derive the stationary distribution both at service completion epochs, and at an arbitrary epoch. We also study the asymptotic behaviour under high rate of retrials. Performance measures are explicitly derived, and extensive numerical examples are performed to investigate the impact of event-dependency. Moreover, constrained optimisation problems are formulated and solved with ultimate goal to derive optimal joining probabilities.

Keywords Queueing · Event-dependent arrival rates · General retrials · Linear control policy · Performance · Variable arrival rate

1 Introduction

In this work, we introduce a novel retrial queueing system, by incorporating a special feature for the customers' behaviour, called *event-dependency*. Retrial queues are used to model service systems, in which arriving customers choose to get served remotely. Thus, in case they find an idle server they begin their service immediately; otherwise, they leave the service area and enter a pool of blocked customers, called the orbit queue. Three major policies are employed to model the access from the pool: (i) In the classical retrial policy, any blocked customer retries independently of each other to access the server after exponentially distributed random time. (ii) In some situations the time intervals between successive repeated attempts are independent of the number of blocked customers (constant retrial policy). In such a case, it is assumed that the server when becomes available start seeking for a blocked customer, while the seeking time intervals are either exponentially or arbitrarily distributed

✉ Ioannis Dimitriou
idimit@uoi.gr

[1] Department of Mathematics, University of Ioannina, 45110 Ioannina, Greece

Published online: 21 March 2023

(i.e., general retrials). (iii) In the linear control policy (assuming exponentially distributed repeated attempts), the classical and the constant retrial policy are combined.

A particular example of such a situation arises in modern call centers, where the call-back option (i.e., the seeking/retrieval time) allows to essentially improve its overall performance; see e.g., Armony and Maglaras (2004a, b); Dudin et al. (2004); Phung-Duc and Kawanishi (2014). Motivated by the fact that often a quick observation of the system (i.e., the nature of the last realized event) may influence customers' decision about the utility of joining a busy system (see e.g., Gencer et al. (2014)), and thus, using the call-back option, our aim is to study a versatile model for the representation of such service systems, by introducing a queueing model with repeated attempts and *event-dependent* arrival rates.

This phenomenon arises when customers either cannot estimate their expected wait from the observation of the queue length, or they are not aware at all about the queue length. The latter case is common in retrial systems, since the arriving customers is most likely to not be aware of the number of the already blocked customers (i.e., the orbit queue length). Thus, in the call center example, assume that a *potential* customer is only aware of last realized event. If he/she receives a busy signal along with the information of the type of customer in service, as well as whether he/she is the first that arrives during the current service or not (see below for more details), he/she analogously adapts his/her arrival rate, leaves his/her contact details and wait to get called back later. In case the last event is a service completion, the customer also adapts its arrival rate to occupy the idle server, but now there is also a competitive stream due to the seeking process.

The main contributions of the paper are summarized as follows.

- On the modelling side, we introduce the concept of *event-dependent* arrival rates in the retrial setting. We pay particular emphasis to the case of general retrial times, but we also provided results under the linear retrial policy, thus, considering all the retrial policies. In particular, we employ a *multi-level event dependency* framework, where the customer's willingness to join the system, which is reflected on the choice of the arrival rates depend (i) on the last realized event, i.e., an arrival or a departure, (ii) in the former case, on the type of the customer that has occupied the idle server, (i.e., whether it is a primary or a retrial customer), (iii) on whether an arriving customer is the first after the arrival that has occupied the idle server, or they have already joined the system other primary customers before the corresponding arriving customer.
- On the technical side, we investigate the stationary behaviour at service completion epoch, as well as at an arbitrary epoch, and provide explicit expressions for various performance metrics. Under the event-dependency framework, PASTA does not hold, and the arrival process is no longer a standard, but a modified Poisson process. The effect of event-dependency on system's performance, is extensively investigated through numerical experiments. Constraint optimisation problems are solved and provide insights on how event-dependency affects the way an arriving customer join or not the system. Moreover, we investigate the stability condition, and the asymptotic behaviour of our system under high rate of retrials. A basic framework on how we can investigate the optimal admission policy based on a Markov decision process (MDP) is also discussed.

1.1 Literature review

Our work is classified into the intersection of the literature of retrial queues and of queues with state dependent parameters.

For a detailed treatment on the development of retrial queues see the books in Falin and Templeton (1997); Artalejo and Gómez-Corral (2008), and references therein; see also Phung-Duc (2017). The vast majority of works on retrial queues assume the classical retrial policy; see e.g., Langaris and Dimitriou (2010). However, in specific service settings, the time intervals between successive attempts are independent of the number of attempting customers (i.e., the constant retrial policy); see e.g., Farahmand (1990); Fayolle (1986); Dimitriou (2018). In Artalejo and Gomez-Corral (1997), the authors introduced the linear control policy, which combined the classical and the constant retrial policies. In Gómez-Corral (1999), the author presented an exhaustive analysis of the single-server retrial queue with general service and seeking times; see also Choi et al. (1993). We further mention the very recent works in Baron et al. (2018, 2022), where the authors considered the state-dependent version of the model in Gómez-Corral (1999) based on the number of customers in orbit by using a probabilistic and an efficient computational method.

The performance analysis of the standard (i.e., without retrials) M/G/1 queue with event-dependent arrival rates was recently introduced in Legros (2018). In Legros and Sezer (2018), the authors studied queueing models where arrivals depend on the remaining service time. Recently, in Legros (2021), the authors investigated the admission control problem with state-dependent arrivals, and provided an algorithm for dimensioning the system. In Legros (2022), the author studied a G/M/1 queue with event-dependent service rate. For other works on standard (i.e., without retrials) queues with workload-dependent or waiting time-dependent arrival and/or service rates see e.g., Bekker et al. (2004); Boxma et al. (2005); Kerner (2008); Boxma and Vlasiou (2007); D'Auria et al. (2022).

To our best knowledge, the concept of event-dependency has never been treated in the retrial queueing literature so far. The main objective of our work is to fill this gap. Our work generalizes the seminal work in Legros (2018) in the retrial setting, and in a multi-level framework due to the the presence of primary and orbiting customers; for initial results on this work see Dimitriou (2022). Our work also differs from Baron et al. (2018), since the arrival-dependency is based on the last realized event instead of the observed number of orbiting customers, and we have also considered all the well known retrial policies.

The rest of the paper is summarized as follows. In Sect. 2, we describe in detail the mathematical model with general retrials. The stability condition and the stationary analysis at service completion epochs is given in Sect. 3. The stationary analysis at an arbitrary epoch is presented in Sect. 4 (we used both the supplementary variable method and the Markov renewal theory). Explicit expressions for various performance metrics along with an asymptotic result are also given. The stationary analysis for the model with the linear control policy is given in Sect. 5. In Sect. 6, we present extensive numerical results that reveal the effect of event-dependency on the system's performance. Moreover, we also solve some interesting constrained optimisation problems in the presence of event-dependency. A conclusion along with some future research plans are given in Sect. 7, where we also discuss in detail how one can investigate the optimal admission policy using a constrained MDP framework.

2 Model description

We consider a single-server queueing system with no waiting space. The service times are iid random variables with cumulative distribution function (cdf) $B(\cdot)$, density $b(\cdot)$, Laplace-Stieltjes (LST) $\beta^*(.)$ and firsts moments $\bar{b}^{(k)} = (-1)^k \frac{d^k}{ds^k}\beta^*(s) \mid_{s=0}, k = 1, 2$. The customers that find the server busy upon arrival, they abandon the system, but leave their contact details

Table 1 Summary of event description and the corresponding arrival rates

Description of last event	Next arrival at rate
Service completion	λ^-
An external arrival has occupied the idle server	λ^e
At least one external customer has arrived after the occupation of the idle server by an external customer	λ_+^e
A retrial customer has occupied the idle server	λ^r
At least one external customer has arrived after the occupation of the idle server by a retrial customer	λ_+^r

so they are called back by the server in a later instant; hence, we assume that they join an infinite capacity orbit queue, waiting to be retrieved by the server. After finishing service, a customer leaves the system and the server seeks for a customer from the orbit. The seeking/retrieving times are iid random variables with cdf $A(.)$, density $a(.)$ and LST $\alpha^*(.)$. However, a new/primary customer may arrive during the seeking process, and in such a case, the server interrupts the seeking process, and starts serving the newly arriving customer. We assume that the interarrival, service and seeking times are mutually independent.

Recall that after a service completion, there is a competition between external/primary arrivals and retrials. The type of the customer that will occupy the server influences the arrival rates of the subsequent customers. More precisely, based on the last realized event, the next customer arrives according to a Poisson process, as follows (see also Table 1):

- If the last realized event is a service completion, the next primary customer will arrive at a rate λ^-.
- In case a primary customer has occupied the idle server, then, the first primary customer that arrive during the corresponding busy period will arrive at a rate λ^e. Moreover, the subsequent primary customers (i.e., the second, third, etc arriving customers during the corresponding busy period) will arrive at a rate λ_+^e.
- In case a retrial customer has occupied the idle server, then, the first primary customer that arrive during the corresponding busy period will arrive at a rate λ^r. Moreover, the subsequent primary customers (i.e., the second, third, etc arriving customers during the corresponding busy period) will arrive at a rate λ_+^r.

Remark 1 From the customer's perspective one might expect that $\lambda^- > \lambda^e \geq \lambda_+^e$, and $\lambda^r \geq \lambda_+^r$. This is justified as follows: if a customer knows that the last realised event is an arrival that has occupied the idle server, she knows that if she decides to join the system, she will be routed to the orbit queue. So she has to wait to be called back by the server in a later instant (i.e., $\lambda^- > \lambda^e$). Normally, the subsequent arrivals that already know that other customers have arrived previously, they might be even more doubted to join the orbit queue (thus, $\lambda^e \geq \lambda_+^e$). Similar arguments may hold for the case $\lambda^r \geq \lambda_+^r$.

3 The embedded Markov chain at service completion epochs

Let τ_i be the time of the i-th departure and $X_i = X(\tau_i^+)$ be the number of customers left in orbit just after the departure of the i-th customer. Then,

$$X_i = \begin{cases} X_{i-1} - B_i + A_i(B_i), & X_{i-1} > 0, \\ A_i(0), & X_{i-1} = 0, \end{cases} \tag{1}$$

where $B_i \in \{0, 1\}$ is the number of orbiting customers, which enter service at time the i-th service starts (i.e. $B_i = 1$ if the i-th customer is an orbiting customer, and $B_i = 0$ if the i-th customer is an external customer), and $A_i(B_i)$ is the number of external arrivals during the time the i-th served customer stays in the service station ($A_i(0)$ (resp. $A_i(1)$) is the number of arriving customers during the service of a primary (resp. a retrial) customer). The random variable B_i depends on the history of the system before the time τ_{i-1} only through the variable X_{i-1}, and its conditional distribution is given by

$$\begin{aligned} \mathbb{P}(B_i = 0 \mid X_{i-1} = n) &= 1 - (1 - \delta_{0,n})\alpha^*(\lambda^-), \\ \mathbb{P}(B_i = 1 \mid X_{i-1} = n) &= (1 - \delta_{0,n})\alpha^*(\lambda^-), \end{aligned}$$

where n is the orbit size and $\delta_{0,n}$ denotes the Kronecker's delta function.

The service time of the i-th customer is independent of previous service times and the number of orbiting customers. Denote by S the corresponding service time. We now focus on the distribution of A_i. Note that since we consider event-dependent arrival rates, we must take into account all the possible events mentioned at the end of the previous section. More precisely,

Case 1: If $X_i > 0$, the last event is a service completion that leaves the server idle. Thus, the next customer that occupies the server is either an external customer (at a Poisson rate λ^-), or a registered (i.e., a retrial) customer. Therefore, the last event for the first customer who arrives during the service of $(i + 1)$-th customer is either an arrival or a successful retrial. In case the server was occupied by an external customer, the first customer will arrive at rate λ^e, and all subsequent customers at rate λ_+^e. In case the server was occupied by a retrial customer, the first customer will arrive at rate λ^r, and all subsequent customers at rate λ_+^r. With such a framework, the next arrival depends both on the last event (i.e., arrival or service completion), and on the type of the customer that have occupied the server in the last (arrival) event.

Case 2: If $X_i = 0$, the last event is a service completion that leaves the system empty. The next customer that occupies the server is an external customer (at rate λ^-). Therefore, the last event for the first customer that will arrive after the server's occupation is an arrival, thus will arrive at rate λ^e, and all the subsequent customers will arrive at rate λ_+^e.

Thus, due to the event-dependency we need to obtain the distribution of the number of arrivals in a service of length t given the type of the customer that occupied the server. Denote by $N(t)$ the number of arriving customers during a service of length t and let

$$\begin{aligned} \mathbb{P}_e(N(t) = n) &= \mathbb{P}(N(t) = n \mid \text{the server is occupied by a primary customer}), \\ \mathbb{P}_r(N(t) = n) &= \mathbb{P}(N(t) = n \mid \text{the server is occupied by a retrial customer}). \end{aligned}$$

Then, the distribution of $N(t)$, is given by the following set of differential equations:

$$\begin{aligned} \mathbb{P}_e(N(0) = 0) &= 1, \\ \tfrac{d}{dt}\mathbb{P}_e(N(t) = 0) &= -\lambda^e \mathbb{P}_e(N(t) = 0), \\ \tfrac{d}{dt}\mathbb{P}_e(N(t) = 1) &= -\lambda_e^+ \mathbb{P}_e(N(t) = 1) + \lambda^e \mathbb{P}_e(N(t) = 0), \\ \tfrac{d}{dt}\mathbb{P}_e(N(t) = n) &= -\lambda_e^+ \mathbb{P}_e(N(t) = n) + \lambda_e^+ \mathbb{P}_e(N(t) = n - 1), \ n \geq 2. \end{aligned} \tag{2}$$

Similarly,

$$\begin{aligned}
&\mathbb{P}_r(N(0)=0)=1,\\
&\tfrac{d}{dt}\mathbb{P}_r(N(t)=0)=-\lambda^r\mathbb{P}_r(N(t)=0),\\
&\tfrac{d}{dt}\mathbb{P}_r(N(t)=1)=-\lambda_r^+\mathbb{P}_r(N(t)=1)+\lambda^r\mathbb{P}_r(N(t)=0),\\
&\tfrac{d}{dt}\mathbb{P}_r(N(t)=n)=-\lambda_r^+\mathbb{P}_r(N(t)=n)+\lambda_r^+\mathbb{P}_r(N(t)=n-1),\ n\geq 2.
\end{aligned} \tag{3}$$

The solutions of systems (2), (3) are respectively

$$\begin{aligned}
&\mathbb{P}_e(N(t)=0)=e^{-\lambda^e t},\\
&\mathbb{P}_e(N(t)=n)=\tfrac{\lambda^e}{\lambda_e^+-\lambda^e}\left(\tfrac{\lambda_e^+}{\lambda_e^+-\lambda^e}\right)^{n-1}[e^{-\lambda^e t}-e^{-\lambda_e^+ t}\textstyle\sum_{k=0}^{n-1}\tfrac{((\lambda_e^+-\lambda^e)t)^k}{k!}],\ n\geq 1.
\end{aligned}$$

and

$$\begin{aligned}
&\mathbb{P}_r(N(t)=0)=e^{-\lambda^r t},\\
&\mathbb{P}_r(N(t)=n)=\tfrac{\lambda^r}{\lambda_r^+-\lambda^r}\left(\tfrac{\lambda_r^+}{\lambda_r^+-\lambda^r}\right)^{n-1}[e^{-\lambda^r t}-e^{-\lambda_r^+ t}\textstyle\sum_{k=0}^{n-1}\tfrac{((\lambda_r^+-\lambda^r)t)^k}{k!}],\ n\geq 1.
\end{aligned}$$

Note that the arrival processes are modified Poisson processes (see also Legros (2018)) where the first interarrival time follows a different distribution than the other interarrival times, and at the same time (for the first time in our work) depend also on the type of the customer that occupy the server (i.e., a primary or a retrial customer). Then, for $i, n \geq 0$,

$$\mathbb{P}(A_i(0)=n \mid X_i \geq 0)=\textstyle\int_0^\infty \mathbb{P}_e(N(t)=n)b(t)dt=b_n^e.$$

Similarly, for $i, n \geq 0$

$$\mathbb{P}(A_i(1)=n \mid X_i > 0)=\textstyle\int_0^\infty \mathbb{P}_r(N(t)=n)b(t)dt=b_n^r.$$

Let $A_k(z)=\sum_{n=0}^\infty b_n^k z^n=\sum_{n=0}^\infty \int_0^\infty \mathbb{P}_k(N(t)=n)z^n b(t)dt$, $|z|\leq 1$, $k=e,\ r$, i.e., $A_k(z)$ is the probability generating function (pgf) of the number of customers that arrive at the system during the service time of a customer of type k, $k=e,r$. Then, extensive computations lead to

$$A_k(z)=\tfrac{\beta^*(\lambda^k)(\lambda_+^k-\lambda^k)(1-z)-\lambda^k z\beta^*(\lambda_+^k(1-z))}{\lambda_+^k(1-z)-\lambda^k},\ k=e,r.$$

Remark 2 Note that in case of *no event-dependency*, i.e., for $k=e,r$, $\lambda_+^k=\lambda^k=\lambda$, then, $A_k(z)=\beta^*(\lambda(1-z))$.

The one-step transition probabilities $p_{m,n}=\mathbb{P}(X_i=n \mid X_{i-1}=m)$ are given by the formulae:

$$\begin{aligned}
p_{m,n}&=(1-\alpha^*(\lambda^-))b_{n-m}^e+\alpha^*(\lambda^-)b_{n-m+1}^r,\ m=1,2,\ldots,n,\\
p_{0,n}&=b_n^e,\ n\geq 0,\\
p_{m+1,m}&=\alpha^*(\lambda^-)b_0^r,\ m\geq 0.
\end{aligned}$$

Theorem 1 *Let X_i be the orbit length at the time of the ith departure, $i \geq 1$. Then, $\{X_i, i \geq 1\}$ is ergodic if and only if*

$$\tilde{b}^{(1)}<\tfrac{\alpha^*(\lambda^-)[\lambda_+^r+(\lambda^r-\lambda_+^r)\beta^*(\lambda^r)]}{\lambda^r[\lambda_+^e+(\lambda_+^r-\lambda_+^e)\alpha^*(\lambda^-)]}-\tfrac{(1-\alpha^*(\lambda^-))(\lambda^e-\lambda_+^e)\beta^*(\lambda^e)}{\lambda^e[\lambda_+^e+(\lambda_+^r-\lambda_+^e)\alpha^*(\lambda^-)]}. \tag{4}$$

Proof See Appendix A. □

Let π_n, $n \geq 0$, be the stationary probability that n customers are in the orbit at a service completion epoch. Then, the Kolmogorov equations reads:

$$\begin{aligned}\pi_n = \pi_0 b_n^e + (1-\delta_{0,n})(1-\alpha^*(\lambda^-)) \textstyle\sum_{j=1}^{n} \pi_j b_{n-j}^e \\ +\alpha^*(\lambda^-) \textstyle\sum_{j=1}^{n+1} \pi_j b_{n+1-j}^r,\ n \geq 0.\end{aligned} \tag{5}$$

Let $\Pi(z) = \sum_{n=0}^{\infty} \pi_n z^n$, $|z| \leq 1$. In Theorem 2, we obtain $\Pi(z)$, π_0, and give the condition of the existence of $\Pi(z)$, which also ensures the stationary regime.

Theorem 2 *Under the stability condition (4), we have*

$$\Pi(z) = \pi_0 \alpha^*(\lambda^-) \frac{zA_e(z) - A_r(z)}{\alpha^*(\lambda^-)(zA_e(z) - A_r(z)) + z(1 - A_e(z))}, \tag{6}$$

where

$$\pi_0 = \frac{\lambda^e \alpha^*(\lambda^-)[\lambda_+^r(1-\lambda^r \bar{b}^{(1)}) + \beta^*(\lambda^r)(\lambda^r - \lambda_+^r)] - \lambda^r(1-\alpha^*(\lambda^-))[\lambda_+^e \lambda^e \bar{b}^{(1)} + (1-\beta^*(\lambda^e))(\lambda^e - \lambda_+^e)]}{\alpha^*(\lambda^-)\left[\lambda^e[\lambda_+^r(1-\lambda^r \bar{b}^{(1)}) + \beta^*(\lambda^r)(\lambda^r - \lambda_+^r)] + \lambda^r[\lambda_+^e \lambda^e \bar{b}^{(1)} + (1-\beta^*(\lambda^e))(\lambda^e - \lambda_+^e)]\right]}.$$

Asking $\pi_0 > 0$ we have that (4) is also necessary for the ergodicity of the chain.

Proof The proof is straightforward by using (5) and applying the generating function approach. The normalization condition implies the expression for π_0. □

Remark 3 Note that our model exhibits a behaviour closely related to the stochastic decomposition behaviour, which normally arise in the standard (i.e., no event-dependency) retrial systems. In particular, when $\lambda^r = \lambda^-$, $\lambda_+^r = \lambda^+$,

$$\Pi(z) = \Pi_{M/G/1}^{(event)}(z) \chi_1(z) \chi_2(z),$$

where $\Pi_{M/G/1}^{(event)}(z)$ is the pgf of the number of customers at service completion epochs in the standard M/G/1 queue with event-dependent arrivals in Legros (2018), and

$$\begin{aligned}\chi_1(z) &= \frac{z - A_r(z)}{\alpha^*(\lambda^-)(zA_e(z) - A_r(z)) + z(1 - A_e(z))} \times \frac{\alpha^*(\lambda^-)(1 + A_e^{(1)}(1) - A_r^{(1)}(1))}{1 - A_r^{(1)}(1)}, \\ \chi_2(z) &= \frac{zA_e(z) - A_r(z)}{zB(z) - A_r(z)} \times \frac{(1 + \lambda^+ \bar{b}^{(1)} - A_r^{(1)}(1))}{1 + A_e^{(1)}(1) - A_r^{(1)}(1)},\end{aligned}$$

where $A_k^{(1)}(1)$, $k = e, r$, are given in Corollary 2. Moreover, $\chi_1(z)$ is the pgf of the number of orbiting customers given the system is idle. However, although $\chi_2(1) = 1$, $\chi_2(z)$, is not obvious that constitutes a pgf.

4 Performance analysis at arbitrary instants

Let $X(t)$ be the number of orbiting customers, $C(t)$ be the state of the server, and $I(t)$ be the last realized event at time t, with values as described in Table 2.

Let also $Z(t)$ be the remaining time until the next service (when $C(t) = 1$), or seeking time completion (when $C(t) = 0$) at time t. Then, $\{(C(t), X(t), I(t), Z(t)); t \geq 0\}$ is an irreducible continuous time Markov chain describing the system model with state space $\{(0, 0, E_1)\} \cup \{(0, j, E_1, r) : j \geq 1, r \geq 0\} \cup \{(1, j, E_k, r) : j \geq 0, r \geq 0, k = 2, 3\} \cup \{(1, j, E_k, r) : j \geq 1, r \geq 0, k = 4, 6\} \cup \{(1, j, E_k, r) : j \geq 2, r \geq 0, k = 5, 7\}$.

Table 2 Description of the states of $I(t)$

Symbol	Last realized event
E_1	Service completion
E_2	An external arrival occupied the server
E_3	A retrial customer occupied the server
E_4	The 1st external customer during The busy period initiated in E_2, has arrived
E_5	At least one customer has arrived after E_4
E_6	The 1st external customer during The busy period initiated in E_3, has arrived
E_7	At least one customer has arrived after E_6

Let also,

$$
\begin{aligned}
p_{0,0}(t) &= P(C(t)=0, X(t)=0, I(t)=E_1),\\
p_{0,j}(r,t) &= P(C(t)=0, X(t)=j, I(t)=E_1, Z(t)\in(r,r+dr]),\ j\geq 1,\\
p^{(k)}_{1,j}(r,t) &= P(C(t)=1, X(t)=j, I(t)=E_k, Z(t)\in(r,r+dr]),\ j\geq 0, k=2,3,\\
p^{(k)}_{1,j}(r,t) &= P(C(t)=1, X(t)=j, I(t)=E_k, Z(t)\in(r,r+dr]),\ j\geq 1, k=4,6,\\
p^{(k)}_{1,j}(r,t) &= P(C(t)=1, X(t)=j, I(t)=E_k, Z(t)\in(r,r+dr]),\ j\geq 2, k=5,7.
\end{aligned}
$$

We are interested in the steady-state counterparts (as $t\to\infty$) of these probabilities.

Lemma 1 *Let* $p_{0,0}=\lim_{t\to\infty}p_{0,0}(t)$, $p_{0,j}(r)=\lim_{t\to\infty}p_{0,j}(r,t)$, *and* $p^{(k)}_{1,j}(r)=\lim_{t\to\infty}p^{(k)}_{1,j}(r,t)$, $k=2,\ldots,7$. *Then:*

$$
\begin{aligned}
\lambda^- p_{0,0} &= \textstyle\sum_{k=2}^{3} p^{(k)}_{1,0}(0),\\
-\tfrac{d}{dr}p_{0,j}(r) &= -\lambda^- p_{0,j}(r) + a(r)\textstyle\sum_{k=2}^{7} p^{(k)}_{1,j}(0), && j\geq 1,\\
-\tfrac{d}{dr}p^{(2)}_{1,j}(r) &= -\lambda^e p^{(2)}_{1,j}(r) + \lambda^- p_{0,j}b(r), && j\geq 0,\\
-\tfrac{d}{dr}p^{(3)}_{1,j}(r) &= -\lambda^r p^{(3)}_{1,j}(r) + p_{0,j+1}(0)b(r), && j\geq 0,\\
-\tfrac{d}{dr}p^{(4)}_{1,j}(r) &= -\lambda^e_+ p^{(4)}_{1,j}(r) + \lambda^e p^{(2)}_{1,j-1}(r), && j\geq 1,\\
-\tfrac{d}{dr}p^{(5)}_{1,j}(r) &= -\lambda^e_+ p^{(5)}_{1,j}(r) + \lambda^e_+(p^{(5)}_{1,j-1}(r)+p^{(4)}_{1,j-1}(r)), && j\geq 2,\\
-\tfrac{d}{dr}p^{(6)}_{1,j}(r) &= -\lambda^r_+ p^{(6)}_{1,j}(r) + \lambda^r p^{(3)}_{1,j-1}(r), && j\geq 1,\\
-\tfrac{d}{dr}p^{(7)}_{1,j}(r) &= -\lambda^r_+ p^{(7)}_{1,j}(r) + \lambda^r_+(p^{(7)}_{1,j-1}(r)+p^{(6)}_{1,j-1}(r)), && j\geq 2.
\end{aligned}
\tag{7}
$$

Proof See Appendix B. □

Let for $Re(s)\geq 0$, $|z|\leq 1$,

$$P^*_0(s,z)=\textstyle\sum_{j=1}^{\infty}\int_0^{\infty}e^{-sr}p_{0,j}(r)dr z^j,$$

$$
\begin{aligned}
P^*_{1,k}(s,z) &= \textstyle\sum_{j=0}^{\infty} \int_0^{\infty} e^{-sr} p^{(k)}_{1,j}(r)dr z^j,\ k=2,3,\\
P^*_{1,k}(s,z) &= \textstyle\sum_{j=1}^{\infty} \int_0^{\infty} e^{-sr} p^{(k)}_{1,j}(r)dr z^j,\ k=4,6,\\
P^*_{1,k}(s,z) &= \textstyle\sum_{j=2}^{\infty} \int_0^{\infty} e^{-sr} p^{(k)}_{1,j}(r)dr z^j,\ k=5,7.
\end{aligned} \tag{8}
$$

Theorem 3 *The stationary distribution of* (C, X, I) *has the following pgfs*

$$
\begin{aligned}
P^*_0(0,z) &= \tfrac{p_{0,0}z(1-\alpha^*(\lambda^-))(A_e(z)-1)}{\alpha^*(\lambda^-)(zA_e(z)-A_r(z))+z(1-A_e(z))},\\
P^*_{1,2}(0,z) &= \lambda^- \tfrac{1-\beta^*(\lambda^e)}{\lambda^e}(p_{0,0}+P^*_0(0,z)),\\
P^*_{1,3}(0,z) &= \tfrac{1-\beta^*(\lambda^r)}{\lambda^r}K(z),\\
P^*_{1,4}(0,z) &= \lambda^- z[\tfrac{\lambda^e_+(1-\beta^*(\lambda^e))-\lambda^e(1-\beta^*(\lambda^e_+))}{\lambda^e_+(\lambda^e_+-\lambda^e)}](p_{0,0}+P^*_0(0,z)),\\
P^*_{1,5}(0,z) &= \lambda^e\lambda^- z[\tfrac{z}{\lambda^e\lambda^e_+(1-z)}+\tfrac{\beta^*(\lambda^e_+(1-z))}{(\lambda^e_+(1-z)-\lambda^e)\lambda^e_+(1-z)}\\
&\quad -\tfrac{\beta^*(\lambda^e_+)}{\lambda^e_+(\lambda^e_+-\lambda^e)}-\tfrac{\lambda^e_+z\beta^*(\lambda^e)}{\lambda^e(\lambda^e_+-\lambda^e)(\lambda^e_+(1-z)-\lambda^e)}](p_{0,0}+P^*_0(0,z)),\\
P^*_{1,6}(0,z) &= zK(z)[\tfrac{\lambda^r_+(1-\beta^*(\lambda^r))-\lambda^r(1-\beta^*(\lambda^r_+))}{\lambda^r_+(\lambda^r_+-\lambda^r)}],\\
P^*_{1,7}(0,z) &= \lambda^r zK(z)[\tfrac{z}{\lambda^r\lambda^r_+(1-z)}+\tfrac{\beta^*(\lambda^r_+(1-z))}{(\lambda^r_+(1-z)-\lambda^r)\lambda^r_+(1-z)}\\
&\quad -\tfrac{\beta^*(\lambda^r_+)}{\lambda^r_+(\lambda^r_+-\lambda^r)}-\tfrac{\lambda^r_+z\beta^*(\lambda^r)}{\lambda^r(\lambda^r_+-\lambda^r)(\lambda^r_+(1-z)-\lambda^r)}],
\end{aligned} \tag{9}
$$

where

$$
\begin{aligned}
K(z) &= \tfrac{\lambda^-\alpha^*(\lambda^-)[p_{0,0}(A_e(z)-1)+A_e(z)P^*_0(0,z)]}{z-\alpha^*(\lambda^-)A_r(z)},\\
p_{0,0} &= \tfrac{\alpha^*(\lambda^-)(\lambda^e t_r+\lambda^r t_e)-\lambda^r t_e}{\alpha^*(\lambda^-)[(1+\lambda^-\bar{b})\lambda^e t_r+\lambda^-\bar{b}\lambda^r t_e]},
\end{aligned}
$$

where

$$
\begin{aligned}
t_e &= (\lambda^e-\lambda^e_+)(1-\beta^*(\lambda^e))+\lambda^e\lambda^e_+\bar{b}^{(1)},\\
t_r &= (\lambda^r-\lambda^r_+)\beta^*(\lambda^r)+\lambda^r_+(1-\lambda^r\bar{b}^{(1)}).
\end{aligned}
$$

Proof See Appendix C. □

Note that asking $p_{0,0} > 0$, we obtain the necessary stability condition, which is the same as the one in (4).

4.1 The event-independent case

We now consider the event-independent case, where $\lambda^k = \lambda^k_+ = \lambda$, $k = e, r$. Then, using the previous results, our model reduces to the one in Gómez-Corral (1999), where the event-independent case was treated. Indeed, it easy to see that when $\lambda^k = \lambda^k_+ = \lambda$, $k = e, r$, then $A_k(z) = \beta^*(\lambda - \lambda z)$, and $\pi_0 = p_{0,0} = 1 - \frac{\lambda\bar{b}^{(1)}}{\alpha^*(\lambda)}$ with $\lambda\bar{b}^{(1)} < \alpha^*(\lambda)$ being the stability condition; see Theorems 1, 2 in Gómez-Corral (1999).

4.2 Performance metrics

Having obtained explicitly the pgfs, we can have in closed form the basic performance metrics.

Corollary 1 *The probabilities of server's state are:*

$$\begin{aligned}
P(C=0) &= p_{0,0} + P_0^*(0,1) = \tfrac{\lambda^e t_r}{(1+\lambda^-\bar{b})\lambda^e t_r + \lambda^-\bar{b}^{(1)}\lambda^r t_e},\\
P(C=1, I=E_2) &= P_{1,2}^*(0,1) = \tfrac{\lambda^- t_r(1-\beta^*(\lambda^e))}{(1+\lambda^-\bar{b}^{(1)})\lambda^e t_r + \lambda^-\bar{b}^{(1)}\lambda^r t_e},\\
P(C=1, I=E_3) &= P_{1,3}^*(0,1) = \tfrac{\lambda^- t_e(1-\beta^*(\lambda^r))}{(1+\lambda^-\bar{b}^{(1)})\lambda^e t_r + \lambda^-\bar{b}^{(1)}\lambda^e t_r},\\
P(C=1, I=E_4) + P(C=1, I=E_5) &= P_{1,4}^*(0,1) + P_{1,5}^*(0,1)\\
&= \tfrac{\lambda^-\lambda^r t_r}{(1+\lambda^-\bar{b}^{(1)})\lambda^e t_r + \lambda^-\bar{b}^{(1)}\lambda^r t_e}(\bar{b}^{(1)} - \tfrac{1-\beta^*(\lambda^e)}{\lambda^e}),\\
P(C=1, I=E_6) + P(C=1, I=E_7) &= P_{1,6}^*(0,1) + P_{1,7}^*(0,1)\\
&= \tfrac{\lambda^-\lambda^r t_e}{(1+\lambda^-\bar{b}^{(1)})\lambda^e t_r + \lambda^-\bar{b}^{(1)}\lambda^r t_e}(\bar{b}^{(1)} - \tfrac{1-\beta^*(\lambda^r)}{\lambda^r}).
\end{aligned}$$

Proof From the results obtained in Theorem 3 and using the normalization condition, the Corollary 1 is proved after heavy but straightforward computations. □

The following corollary provides the throughput (TH_S) generated by the system, the expected orbit queue length ($E(X)$), and the expected sojourn time ($E(W)$).

Corollary 2 *We have,*

$$\begin{aligned}
E(X) &= p_{0,0}[(1-\alpha^*(\lambda^-))(1+\lambda^-\bar{b}^{(1)})G - \tfrac{\lambda^-\alpha^*(\lambda^-)}{1-\alpha^*(\lambda^-)}S_r]\\
&\quad + \bar{b}^{(1)}F + \lambda^- P(C=0)[S_e + \tfrac{\alpha^*(\lambda^-)}{1-\alpha^*(\lambda^-)}S_r],\\
TH_S &= \tfrac{\lambda^-(\lambda^e t_r + \lambda^r t_e)}{\lambda^e t_r + \lambda^-\bar{b}^{(1)}(\lambda^e t_r + \lambda^r t_e)},\\
E(W) &= \tfrac{E(X)}{TH_S},
\end{aligned} \tag{10}$$

where $p_{0,0}$, t_r, t_e, $P(C=0)$ *are given in Theorem 3 and Corollary 1, and*

$$\begin{aligned}
S_k &:= \tfrac{\lambda^k - \lambda_+^k}{\lambda^k}(\bar{b}^{(1)} - \tfrac{1-\beta^*(\lambda^k)}{\lambda^k}) + \tfrac{\lambda_+^k \bar{b}^{(2)}}{2},\ k = e, r,\\
G &:= \tfrac{\alpha^*(\lambda^-)[(2A_e^{(1)}(1) + A_e^{(2)}(1))(1-A_r^{(1)}(1)) + A_e^{(1)}(1)A_r^{(2)}(1)]}{2[A_e^{(1)}(1)(\alpha^*(\lambda^-)-1) + \alpha^*(\lambda^-)(1-A_r^{(1)}(1))]^2},\\
F &:= \tfrac{[p_{0,0}(1-\alpha^*(\lambda^-))G + A_e^{(1)}(1)P(C=0)] - P_0^*(0,1)(1-\alpha^*(\lambda^- A_r^{(1)}(1)))}{(1-\alpha^*(\lambda^-))^2},
\end{aligned}$$

with

$$\begin{aligned}
A_k^{(1)}(1) &= \tfrac{(\lambda^k - \lambda_+^k)(1-\beta^*(\lambda^k)) + \lambda^k\lambda_+^k\bar{b}^{(1)}}{\lambda^k},\\
A_k^{(2)}(1) &= \lambda_+^k(2\bar{b}^{(1)} + \lambda_+^k\bar{b}^{(2)}) - 2\tfrac{\lambda_+^k}{\lambda^k}A_k^{(1)}(1),
\end{aligned}$$

for $k = e, r$.

Proof See Appendix D. □

4.3 Asymptotic behaviour under high rate of retrials

Let $P(z) = p_{0,0} + P_0^*(0,z) + z\sum_{k=2}^{7} P_{1,k}^*(0,z)$. Then,

$$\lim_{a^*(\lambda^-)\to 1} P(z) = P^{(\infty)}(z),$$

where $P^{(\infty)}(z)$ is the pgf of the number of customers in the system obtained in the seminal paper Legros (2018), by assuming that $\lambda^r = \lambda^-$, and $\lambda^e = \lambda_+^e = \lambda_+^r = \lambda^+$. Note that in

such a case, the customers who find the server busy repeat their calls almost immediately. Note also that if we further assume $\lambda^- = \lambda_+ = \lambda$, $P^{(\infty)}(z)$ coincides with the pgf of the number of customers in the standard M/G/1 queue.

Theorem 4 *As* $\alpha^*(\lambda^-) \to 1$,

$$\frac{2\lambda^r t_e(1-\alpha^*(\lambda^-))}{\alpha^*(\lambda^-)[(1+\lambda^-\tilde{b}^{(1)})\lambda^e t_r+\lambda^-\tilde{b}^{(1)}\lambda^r t_e]} \leq \sum_{n=0}^{\infty}[P(X=n)-P^{(\infty)}(X=n)] \leq \frac{2\lambda^r t_e(1-\alpha^*(\lambda^-))}{\alpha^*(\lambda^-)\lambda^e t_r}.$$

Proof The proof follows the steps given in Artalejo and Falin (1994), and further details are omitted. □

Theorem 4 provides a measure of the proximity between the steady state distributions for the standard M/G/1 queueing system with event dependent arrivals in Legros (2018) and our queueing system. The importance of these bounds is to provide upper and lower estimates for the distance between both distributions.

4.4 Explicit expressions for the Markovian case

In the following, we consider the purely Markovian case, where we assume that service, and retrieval times are exponentially distributed with rates μ, and α, respectively. To simplify further the analysis, we also assume that $\lambda^r_+ = \lambda^r$, $\lambda^e_+ = \lambda^e$. Note that under such a setting there are three events that affect the arrival rates:

1. a service completion, after which an external arrival occurs with rate λ^-,
2. a service initiation by an external arrival, after which the next arrivals occur at rate λ^e,
3. a service initiation by a retrial customer, after which the next arrivals occur at rate λ^r.

Let $p_{0,j} = \lim_{t\to\infty} P(X(t) = j, C(t) = 0, I(t) = E_1)$, $p^{(k)}_{1,j} = \lim_{t\to\infty} P(X(t) = j, C(t) = 1, I(t) = E_k)$, $j = 0, 1, \ldots, k = 2, 3$. The following theorem states the main result.

Theorem 5 *Under the stability condition* $\alpha\lambda^r + \lambda^-\lambda^e < \alpha\mu$, *the stationary probabilities of the system state are as follows:*

$$\begin{aligned}
p_{0,0} &= \frac{\mu(\alpha\mu-\alpha\lambda^r-\lambda^-\lambda^e)}{a((\lambda^-+\mu)(\mu-\lambda^r)+\lambda^e\lambda^-)},\\
p_{0,j} &= \left(\frac{\lambda^e\lambda^r(\lambda^-+\alpha)}{\alpha\mu(\lambda^e+\mu)}\right)^j \frac{\lambda^r p_{0,0}}{\lambda^-+\alpha}\left[\sum_{i=0}^{j} z_1^{j-i}z_2^i - \frac{\alpha\mu}{\lambda^e(\lambda^-+\alpha)}\sum_{i=0}^{j-1} z_1^{j-i+1}z_2^i\right],\ j \geq 1,
\end{aligned} \tag{11}$$

$$p^{(2)}_{1,j} = \frac{\lambda^- p_{0,0}}{\mu}\left(\frac{\lambda^e\lambda^r(\lambda^-+\alpha)}{\alpha\mu}\right)^j\left[\mu\sum_{i=0}^{j} z_1^{j-i}z_2^i - \frac{\alpha\mu(\lambda^e+\mu)}{\lambda^e(\lambda^-+\alpha)}\sum_{i=0}^{j-1} z_1^{j-i+1}z_2^i\right],\ j \geq 0, \tag{12}$$

$$p^{(3)}_{1,j} = \frac{\lambda^-\lambda^- p_{0,0}}{\mu}\left(\frac{\lambda^e\lambda^r(\lambda^-+\alpha)}{\alpha\mu}\right)^j \sum_{i=0}^{j} z_1^{j-i}z_2^i,\ j \geq 0, \tag{13}$$

where z_1, z_2, *with* $z_i > 1$, $i = 1, 2$, *the two zeros of the polynomial*

$$f(z) := z(\alpha\mu\lambda^r + \lambda^e(\lambda^- + \alpha)(\lambda^r + \mu)) - \alpha\mu(\lambda^e + \mu) - \lambda^e\lambda^r(\lambda^- + \alpha)z^2.$$

Moreover, the probabilities of server's state are:

$$\begin{aligned}
P(C=0) &= \frac{\alpha(\mu-\lambda^r)p_{0,0}}{\alpha\mu-\alpha\lambda^r-\lambda^-\lambda^e},\\
P(C=1) &= \frac{\lambda^-\alpha(\mu+\lambda^e-\lambda^r)p_{0,0}}{\mu(\alpha\mu-\alpha\lambda^r-\lambda^-\lambda^e)}.
\end{aligned} \tag{14}$$

Proof Appendix E. □

4.5 The Markov regenerative approach

Our aim in this subsection is to provide expressions for the joint distribution of the state of the server, the number of jobs in orbit, and the last realized event by following the method of regenerative processes. This method provides quicker the results on the stationary behaviour than the method of supplementary variables applied in Theorem 3, since the probability distribution of the embedded Markov chain in service completion epochs is known; see Theorem 2. For ease of computations, we assume that the seeking/retrieving times are iid exponentially distributed random variables with rate α (i.e., we adopt the constant retrial policy).

It can be easily verified that $\{(C(t), X(t), I(t); t \geq 0)\}$ is a Markov regenerative process with the embedded Markov renewal process $\{X_m; m \in \mathbb{N}\}$. Thus, using the classical limiting theorems in Cinlar (1975), we have that under the ergodicity conditions established in Theorem 1

$$\begin{aligned} q_{i,j}^{(k)} &:= \lim_{t\to\infty} P((C(t), X(t), I(t) = (i, j, E_k)) \\ &= \frac{\sum_{n=0}^{\infty} \pi_n \tau_n(i,j,k)}{\sum_{n=0}^{\infty} \pi_n \tau_n}, \ (i, j, k) \in S, \end{aligned}$$

where $S = \{(0, j, E_1); j \geq 0\} \cup \{(1, j, E_k); j \geq 0, k = 2, 3\} \cup \{(1, j, E_k); j \geq 1, k = 4, 6\} \cup \{(1, j, E_k); j \geq 2, k = 5, 7\}$, $\tau_n(i, j, k)$ denotes the expected amount of time spent by the process $\{(C(t), X(t), I(t); t \geq 0)\}$ in the state (i, j, E_k) during an interval between two successive service completion epochs, given that at the beginning of this interval there were n jobs in orbit, τ_n denotes the expectation of the time interval between two successive service completion epoch given that at the beginning of this interval there were n jobs in orbit, and $\{\pi_n; n \geq 0\}$ are the limiting probabilities of $\{X_m; m \in \mathbb{N}\}$, as given through the pgf in (6) for $\alpha^*(\lambda^-) = \frac{\alpha}{\alpha+\lambda^-}$.

For the model at hand,

$$\tau_n = \frac{1}{\lambda^- + \alpha(1 - \delta_{0,n})} + \bar{b}^{(1)}, \ n \geq 0, \tag{15}$$

so that $U := \sum_{n=0}^{\infty} \pi_n \tau_n := \frac{\pi_0\alpha+\lambda^-(1+\bar{b}^{(1)}(\lambda^-+\alpha))}{\lambda^-(\lambda^-+\alpha)}$, and π_0 as given in Theorem 2 for $\alpha^*(\lambda^-) = \frac{\alpha}{\alpha+\lambda^-}$. The following theorem states our main result.

Theorem 6 *Under the stability conditions given in Theorem 1, we have*

1. *The limiting probabilities are given by*

$$
\begin{aligned}
q_{0,j}^{(1)} &= \tfrac{1}{U}\tfrac{\pi_j}{\lambda^- + \alpha(1-\delta_{0,j})},\ j \geq 0 \\
q_{1,j}^{(2)} &= \tfrac{1}{U}\tfrac{\pi_j \lambda^-}{\lambda^- + \alpha(1-\delta_{0,j})} \int_0^\infty e^{-\lambda^e t}(1-B(t))dt \\
&= \lambda^-\left(\tfrac{1-\beta^*(\lambda^e)}{\lambda^e}\right)\tfrac{1}{U}\tfrac{\pi_j}{\lambda^- + \alpha(1-\delta_{0,j})},\ j \geq 0, \\
q_{1,j}^{(3)} &= \tfrac{1}{U}\tfrac{\pi_{j+1}\alpha}{\lambda^- + \alpha} \int_0^\infty e^{-\lambda^r t}(1-B(t))dt \\
&= \alpha\left(\tfrac{1-\beta^*(\lambda^r)}{\lambda^r}\right)\tfrac{1}{U}\tfrac{\pi_{j+1}}{\lambda^- + \alpha},\ j \geq 0, \\
q_{1,j}^{(4)} &= \tfrac{1}{U}\tfrac{\pi_{j-1}\lambda^-}{\lambda^- + \alpha(1-\delta_{0,j-1})} \int_0^\infty (1-B(t)) \int_{u=0}^t \lambda^e e^{-\lambda^e u} e^{-\lambda_+^e (t-u)} du dt \\
&= \lambda^-\left(\tfrac{\lambda^e(1-\beta^*(\lambda_+^e)) - \lambda_+^e(1-\beta^*(\lambda^e)}{\lambda_+^e(\lambda_+^e - \lambda^e)}\right)\tfrac{1}{U}\tfrac{\pi_{j-1}}{\lambda^- + \alpha(1-\delta_{0,j-1})},\ j \geq 1, \\
q_{1,j}^{(5)} &= \tfrac{1}{U}\textstyle\sum_{n=0}^{j-2} \tfrac{\pi_n \lambda^-}{\lambda^- + \alpha(1-\delta_{0,n})} \\
&\quad \times \int_0^\infty (1-B(t)) \int_{u=0}^t \lambda^e e^{-\lambda^e u} e^{-\lambda_+^e (t-u)} \tfrac{(\lambda_+^e (t-u))^{j-n-1}}{(j-n-1)!} du dt,\ j \geq 2, \\
q_{1,j}^{(6)} &= \tfrac{1}{U}\tfrac{\pi_j \lambda^-}{\lambda^- + \alpha} \int_0^\infty (1-B(t)) \int_{u=0}^t \lambda^r e^{-\lambda^r u} e^{-\lambda_+^r (t-u)} du dt \\
&= \alpha\left(\tfrac{\lambda^r(1-\beta^*(\lambda_+^r)) - \lambda_+^r(1-\beta^*(\lambda^r)}{\lambda_+^r(\lambda_+^r - \lambda^r)}\right)\tfrac{1}{U}\tfrac{\pi_j}{\lambda^- + \alpha},\ j \geq 1, \\
q_{1,j}^{(7)} &= \tfrac{1}{U}\textstyle\sum_{n=1}^{j-1} \tfrac{\pi_n \alpha}{\lambda^- + \alpha} \\
&\quad \times \int_0^\infty (1-B(t)) \int_{u=0}^t \lambda^r e^{-\lambda^r u} e^{-\lambda_+^r (t-u)} \frac{(\lambda_+^r (t-u))^{j-n}}{(j-n)!} du dt,\ j \geq 2.
\end{aligned}
\tag{16}
$$

2. *The partial generating functions* $Q_{i,k}(z) = \sum_{j=0}^\infty q_{i,j}^{(k)} z^j$, $i = 0, 1$, $k = 1, \ldots, 7$, *are given by*

$$
\begin{aligned}
Q_{0,1}(z) &= \tfrac{\alpha\pi_0(\lambda^- + \alpha)}{\alpha\pi_0 + \lambda^-(1+\bar{b}^{(1)}(\lambda^- + \alpha))}\left(\tfrac{z - A^r(z)}{\lambda^- z(1-A^e(z)) + \alpha(z - A^r(z))}\right), \\
Q_{1,2}(z) &= \lambda^- \tfrac{1-\beta^*(\lambda^e)}{\lambda^e} Q_{0,1}(z), \\
Q_{1,3}(z) &= \alpha \tfrac{1-\beta^*(\lambda^r)}{\lambda^r}(Q_{0,1}(z) - q_{0,0}^{(1)}), \\
Q_{1,4}(z) &= \lambda^- z\left(\tfrac{\lambda^e(1-\beta^*(\lambda_+^e)) - \lambda_+^e(1-\beta^*(\lambda^e)}{\lambda_+^e(\lambda_+^e - \lambda^e)}\right) Q_{0,1}(z), \\
Q_{1,5}(z) &= \lambda^e \lambda^- z\left[\tfrac{z}{\lambda^e \lambda_+^e (1-z)} + \tfrac{\beta^*(\lambda_+^e(1-z))}{(\lambda_+^e(1-z) - \lambda^e)\lambda_+^e(1-z)}\right. \\
&\quad \left. - \tfrac{\beta^*(\lambda_+^e)}{\lambda_+^e(\lambda_+^e - \lambda^e)} - \tfrac{\lambda_+^e z \beta^*(\lambda^e)}{\lambda^e(\lambda_+^e - \lambda^e)(\lambda_+^e(1-z) - \lambda^e)}\right] Q_{0,1}(z), \\
Q_{1,6}(z) &= \alpha\left[\tfrac{\lambda_+^r(1-\beta^*(\lambda^r)) - \lambda^r(1-\beta^*(\lambda_+^r))}{\lambda_+^r(\lambda_+^r - \lambda^r)}\right](Q_{0,1}(z) - q_{0,0}^{(1)}), \\
Q_{1,7}(z) &= \alpha\lambda^r\left[\tfrac{z}{\lambda^r \lambda_+^r (1-z)} + \tfrac{\beta^*(\lambda_+^r(1-z))}{(\lambda_+^r(1-z) - \lambda^r)\lambda_+^r(1-z)}\right. \\
&\quad \left. - \tfrac{\beta^*(\lambda_+^r)}{\lambda_+^r(\lambda_+^r - \lambda^r)} - \tfrac{\lambda_+^r z \beta^*(\lambda^r)}{\lambda^r(\lambda_+^r - \lambda^r)(\lambda_+^r(1-z) - \lambda^r)}\right](Q_{0,1}(z) - q_{0,0}^{(1)}),
\end{aligned}
\tag{17}
$$

where $q_{0,0}^{(1)} = \frac{1}{U}\frac{\pi_0}{\lambda^-} = \frac{\pi_0(\lambda^- + \alpha)}{\pi_0 + \lambda^-(1+\bar{b}^{(1)}(\lambda^- + \alpha))}$, *and* π_0 *as given in Theorem 2 when* $\alpha^*(\lambda^-) = \frac{\alpha}{\alpha + \lambda^-}$.

Proof See Appendix F. □

Remark 4 With simple but tedious calculations, it is easy to realize that the expressions of pgfs in Theorem 6 coincide with those in Theorem 3 when $\alpha^*(\lambda^-) = \frac{\alpha}{\alpha + \lambda^-}$ (i.e., when the seeking/retrieval times are iid exponentially distributed random variables with rate α).

More precisely, $Q_{0,1}(z) = p_{0,0} + P_0^*(0,z)$, $Q_{1,k}(z) = P_{1,k}^*(0,z)$, $k = 2,\ldots,7$, with $K(z) = \alpha(\frac{Q_{0,1}(z) - q_{0,0}^{(1)}}{z})$.

Remark 5 Note that in case of *no-event dependency*, i.e., $\lambda^- = \lambda^e = \lambda_+^e = \lambda^r = \lambda_+^r = \lambda$, $U = 1/\lambda$, as expected, since U is the expected time between two successive departures.

5 The model under the linear control policy

In this section, we provide results on the model with multi-level event-dependent arrival rates, under the linear retrial policy Artalejo and Gomez-Corral (1997). Note that such a policy incorporates two types of retrial requests, i.e., the classical retrial policy and the constant retrial policy. More precisely, in the former policy, any orbiting customer retries independently to access the server, while in the latter one the time between two successive repeated attempts is independent of the number of customers applying for service. This section deals with the M/G/1 retrial queue with event-dependent arrival rates, allowing the simultaneous presence of both types of repeat requests.

Thus, for the model as described in Sect. 2, we assume now that the control policy to access from the orbit queue to the server is governed by an exponential law with linear intensity $n\alpha + \beta(1 - \delta_{0,n})$ when the orbit size is $n \geq 0$, where $\delta_{0,n}$ denotes Kronecker's delta. Due to the linear retrial policy, the state probabilities defined in section 4 are the same except the one that refer to the case where the server is idle. Thus, we have $p_{0,j}(t) := P(C(t) = 0, X(t) = j, I(t) = E_1)$, $j \geq 0$, and $p_{0,j} = \lim_{t\to\infty} p_{0,j}(t)$. The pgfs in (8) remain valid, and let $P_0(z) = \sum_{j=0}^{\infty} p_{0,j} z^j$, $|z| \leq 1$.

Theorem 7 *The embedded Markov chain* $\{X_i; i \geq 0\}$ *at service completion epochs is ergodic if and only*

$$A_r^{(1)}(1) < 1 - \frac{\lambda^-}{\beta} A_e^{(1)}(1)\delta_{0,a}, \tag{18}$$

where $A_k^{(1)}(1) = \frac{d^j}{dz^j} A_k(z)|_{z=1}$, $k = e, r$.

Proof To prove ergodicity, we use the Foster's criterion as in the proof of Theorem 1. By considering the function $f(n) = n$, the mean drifts $x_n := E(f(X_{i+1}) - f(X_i) \mid f(X_i) = n)$, $n \geq 0$ are:

$$x_n = \begin{cases} A_e^{(1)}(1), & n = 0, \\ \frac{\lambda^-}{\lambda^- + \beta + n\alpha} A_e^{(1)}(1) + \frac{\beta + n\alpha}{\lambda^- + \beta + n\alpha} A_r^{(1)}(1) - \frac{\beta + n\alpha}{\lambda^- + \beta + n\alpha}, & n \geq 1. \end{cases}$$

Clearly, if $\beta > 0$ and

$$A_r^{(1)}(1) < 1 - \frac{\lambda^-}{\beta} A_e^{(1)}(1)\delta_{0,\alpha}, \tag{19}$$

then, $\lim_{n\to\infty} x_n < 0$ and (19) is a sufficient condition for ergodicity. The proof of necessity follows Theorem 1 in Sennott et al. (1983) and further details are omitted. □

Remark 6 In case $\alpha = 0$, and $A_e(.) = A_r(.) := A(.)$, i.e., no event dependency, (19) reduces to the ergodicity condition for the retrial model under the constant retrial policy; see (Farahmand (1990), eq. (2.14)). In case $\alpha > 0$, $\beta \geq 0$, the ergodicity condition reduces to $A_r^{(1)}(1) < 1$.

The main result is as follows.

Theorem 8 *If* $\alpha > 0$, $\beta \geq 0$ *and* $A_r^{(1)}(1) < 1$, *then*

$$P_0(z) = z^{-\beta/\alpha} H(z) \left(1 - (1 - \delta_{0,\alpha}) \frac{\int_z^1 x^{\frac{\beta}{\alpha}-1} H^{-1}(x) dx}{\int_0^1 x^{\frac{\beta}{\alpha}-1} H^{-1}(x) dx} \right), \tag{20}$$

where

$$H(z) = P_0(1) exp\{ \frac{\lambda^-}{\alpha} \int_z^1 \frac{A_e(x) - 1}{A_r(x) - 1} dx \}, \tag{21}$$

$$P_0(1) = \frac{1 - A_r^{(1)}(1)}{(1 + \lambda^- \bar{b}^{(1)})(1 - A_r^{(1)}(1) + \lambda^- \bar{b}^{(1)} A_e^{(1)}(1)}, \tag{22}$$

and

$$\begin{aligned}
P_{1,2}^*(s,z) &= \tfrac{\lambda^-(\beta^*(\lambda^e) - \beta^*(s))}{s - \lambda^e} P_0(z), \\
P_{1,3}^*(s,z) &= \tfrac{(\alpha z P_0'(z) + \beta(P_0(z) - p_{0,0}))(\beta^*(\lambda^r) - \beta^*(s))}{z(s - \lambda^r)}, \\
\textstyle\sum_{k=4}^5 P_{1,k}^*(s,z) &= \tfrac{\lambda^e \lambda^- z}{s - \lambda_+^e(1-z)} [\tfrac{\beta^*(\lambda^e) - \beta^*(\lambda_+^e(1-z))}{\lambda_+^e(1-z) - \lambda^e} - \tfrac{\beta^*(s) - \beta^*(\lambda^e)}{s - \lambda^e}] P_0(z), \\
\textstyle\sum_{k=6}^7 P_{1,k}^*(s,z) &= \tfrac{\lambda^r(\alpha z P_0'(z) + \beta(P_0(z) - p_{0,0}))}{s - \lambda_+^r(1-z)} [\tfrac{\beta^*(\lambda^r) - \beta^*(\lambda_+^r(1-z))}{\lambda_+^r(1-z) - \lambda^r} - \tfrac{\beta^*(s) - \beta^*(\lambda^r)}{s - \lambda^r}]
\end{aligned} \tag{23}$$

Proof See Appendix G. □

Remark 7 Note that in the special case $\alpha = 0$, our model refers to the retrial queue with constant retrial policy and event dependent arrival rates. In such a case, (G32) reduces to a simple expression for obtaining $P_0(z)$,

$$\begin{aligned}
P_0(z) &= \tfrac{\beta(z - A_r(z))}{\beta(z - A_r(z)) + \lambda^- z(1 - A_e(z))} p_{0,0}, \\
p_{0,0} &= \tfrac{\beta(1 - A_r^{(1)}(1)) - \lambda^- A_e^{(1)}(1)}{\beta[(1 + \lambda^- \bar{b})(1 - A_r^{(1)}(1)) + \lambda^- \bar{b} A_r^{(1)}(1))]}.
\end{aligned}$$

The rest expressions are derived analogously by using (23) and $\alpha = 0$. When we further assume that $\lambda^- = \lambda^r = \lambda_+^r = \lambda^e = \lambda_+^e = \lambda$ (i.e., no event-dependency), we derive the expressions in Farahmand (1990).

In case $\beta = 0$, $\alpha > 0$ we have the model under the classical retrial policy, and $P_0(z) = H(z)$. The rest expressions are derived by (23) for $\beta = 0$.

6 Numerical results

Our aim here is twofold. First, to focus on the effect of event-dependency on system performance: In particular, we aim to investigate how the event-dependency, i.e., the customers' behaviour based on the last realized event, affects the major performance metrics of our system; see Sect. 6.1. In this direction, we model the customer's behaviour by a vector $q := (q_1, q_2, q_3, q_4)$ that refers to the joining probabilities of arriving customers after an arrival. In particular, the vector q states the willingness of an arriving customer to join the system when the server is busy, by taking into account the type of the customer that has occupied the server (i.e., either a primary, or a retrial customer), and whether he/she is the first after the arrival that has occupied the idle server, or they have already joined the system other primary customers before him/her. More precisely, let q_1 be the joining probability when the

last event is an arrival of an external customer that occupies the idle server, q_2 be the joining probability when at least one customer has arrived after the occupation of the server by an external/primary customer, q_3 be the joining probability when the last event is an arrival of a retrial customer that occupies the idle server, and q_4 be the joining probability when at least one customer has arrived after the occupation of the server by an orbiting customer. Thus, we assume that $(\lambda^e, \lambda_+^e, \lambda^r, \lambda_+^r) = \lambda^+ q = \lambda^+(q_1, q_2, q_3, q_4)$.

Second, we aim to obtain optimal joining probabilities that helps a system manager to determine whether an arriving customer will be allowed to enter the system or will be rejected. These probabilities are selected with ultimate goal to maximize the system generated throughput subject to certain constraints on the expected number of orbiting customers; see Sect. 6.2. Note that throughput optimisation problems with delay constraints have been recently investigated in several service systems with shared resources as in Chen et al. (2018); Pappas et al. (2018); Ploumidis et al. (2017).

Assume hereon that the service and the retrieval times are such that $B \sim Erlang(M, \mu)$, $A \sim Erlang(N, \alpha)$, respectively. All the results are based on the derivations in Theorem 3, and in Corollary 2.

6.1 The effect of event-dependency on system performance

Example 1a: Set $\lambda^+ = 0.3$, $\mu = 2.5$ and $q = (q_1, q_2, q_3, q_4) = (0.5, 0.4, 0.6, 0.4)$. Our aim is to investigate the effect of the number of phases of service times and retrial times on $E(X)$ for increasing values of λ^-. Moreover, we aim to investigate how the relation among λ^- and λ^+ affects $E(X)$.

Note that by increasing the number of phases of service times, $E(X)$ is also increasing as expected (see Fig. 1). Moreover, as $\lambda^- < \lambda^+$, $E(X)$ decreases, and as $\lambda^- \geq \lambda^+$, $E(X)$ increases. Thus, under such a setting, by keeping the arrival rates after a service time lower than the arrival rates after an arrival, we ensure a better performance. This is due to the fact that keeping lower as possible λ^- with respect to λ^+, we give better chances for the blocked (i.e., retrial customers) to connect with the server. Thus, under the current setting, the customers feel more comfortable to arrive after an arrival, and connect with the server as retrial customers (as long as $\lambda^- < \lambda^+$).

Figure 2 indicate that by decreasing α from 3.5 to 1.5, we cannot have the advantage of the previous setting. Thus, even if $\lambda^- < \lambda^+$, by increasing λ^-, $E(X)$ increases as expected.

Furthermore, by increasing the number of phases of retrial times, we also fail to retain the advantage of the previous setting (see Fig. 3). This is because in such a case, there is a longer delay for the retrial customers to connect with the server, which in turn results in increasing the expected number of orbiting customers.

Example 1b: Let know $\lambda^+ = 0.5$, $\alpha = 3.5$, $N = 2$, and $q = (0.1, 0.4, 0.5, 0.2)$. In Fig. 4 we observe that by increasing the number of phases of service we effectively reduce the expected waiting time when $q_1 < q_2$, as $\lambda^- \to \lambda^+$ and beyond. This result is surprising, since as shown in Fig. 5, it is in contradiction with the improvement in the expected waiting time which results from the increase in the variability in the service process when the number of phases decreases and $q_1 > q_2$.

Example 1c: In Fig. 6 we can observe the effect of $\lambda^e = \lambda^+ q_1$ on the expected number of orbiting customers for $\lambda^- = 0.1$, $(q_2, q_3, q_4) = (0.5, 0.6, 0.1)$. It is seen that there is a critical value of λ^+ (close to 0.5 in that example), where we can achieve a substantial decrease on $E(X)$. Moreover, we can see that a small increase on q_1 (from 0.2 to 0.4) will

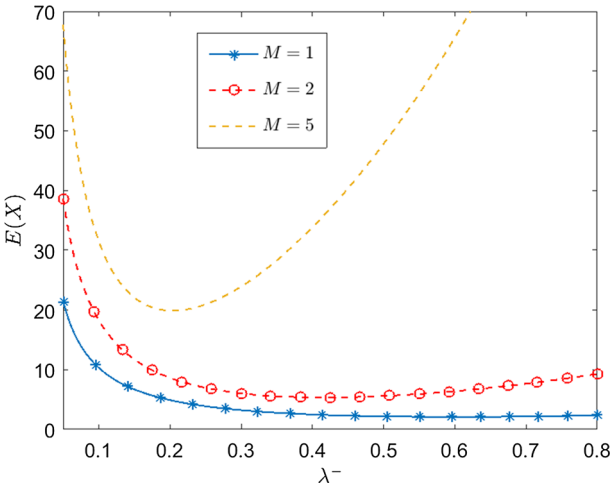

Fig. 1 Effect of service phases when $N = 2, \alpha = 3.5$

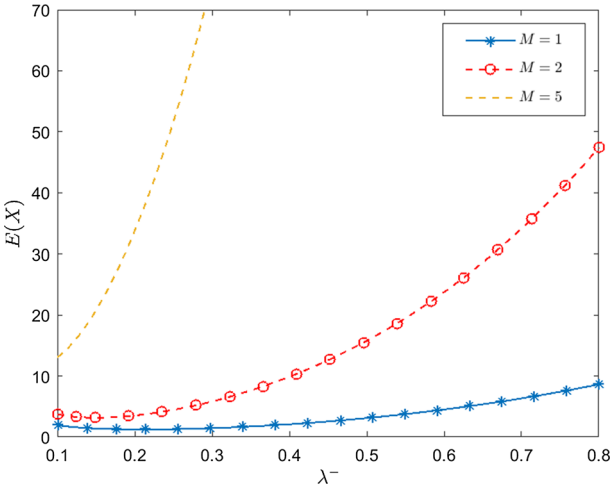

Fig. 2 Effect of decreasing α from 3.5 to 1.5 ($N = 2$)

cause an increase on $E(X)$, which is expected. However, by further increasing q_1, we observe that $E(X)$ start decreasing. It seems that such a behaviour arises as soon as $q_1 > q_2$. Thus, it seems that it is better for the system manager to allow with high probability the arriving customers to join the system when the last event is the occupation of the server by the arrival of an external customer and then, to keep lower the joining probability after that event.

A similar behaviour is observed (even clearer) in Fig. 7, where now we focused the effect of q_3, which is the joining probability when the last event was an arrival from the orbit that occupied the idle server. It seems that the more we increase q_3 (for fixed values of $q_1 = 0.6 > q_2 = 0.5, q_4 = 0.4$), the better performance we achieve, i.e., $E(X)$ decreases. In Fig. 8, we assume that $q_1 = 0.2 < q_2 = 0.5$, and observe a similar behaviour as in Fig. 7. However, for small values of λ^+, $E(X)$ is smaller when $q_3 < q_4$. As λ^+ increases, $E(X)$ remains smaller as long as $q_3 > q_4$.

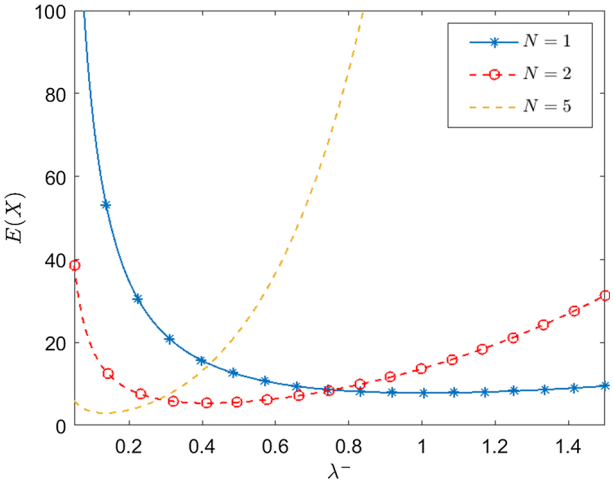

Fig. 3 Effect of phases of retrial times ($\alpha = 3.5$, $M = 2$)

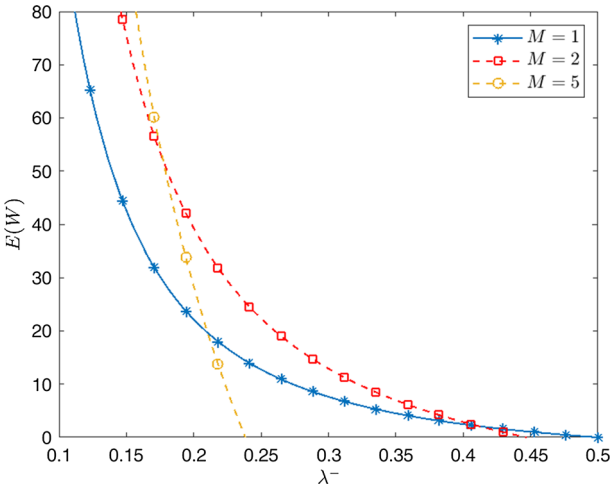

Fig. 4 Effect of M when $N = 2$, $\alpha = 3.5$, $\lambda^+ = 0.5$, $q = (0.1, 0.4, 0.5, 0.2)$

6.2 Optimisation problems

Our goal is to determine the optimal joining probabilities $q = (q_1, q_2, q_3, q_4)$ that maximize the throughput (TH_S) generated by the system, subject to constraints on the maximum attained service level on $E(X)$, and the stability.

These probabilities will dictate the way the arriving customers join the system when the last event is known. Therefore, they can serve as a guide for the system manager to see what would ideally be the arriving customer's behaviour when the last event is known. Equivalently, based on these values, he/she can adequately accept or reject arriving jobs subject to the last realized event, so that to maximize the system throughput. Throughput optimisation problems with delay constraints have been recently investigated in wireless systems with shared resources;

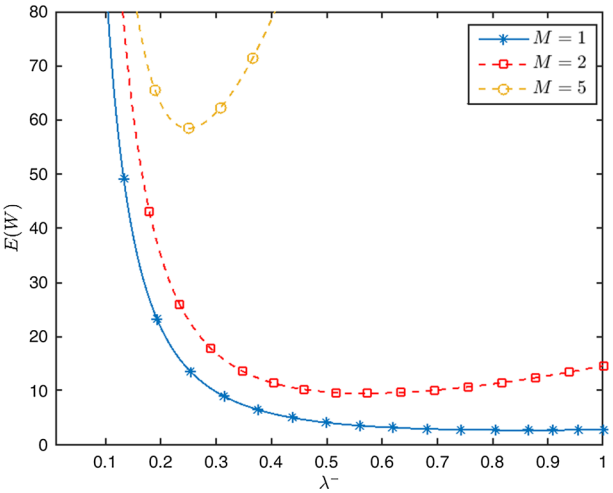

Fig. 5 Effect of M when $N = 2, \alpha = 3.5, \lambda^+ = 0.5, q = (0.6, 0.4, 0.5, 0.2)$

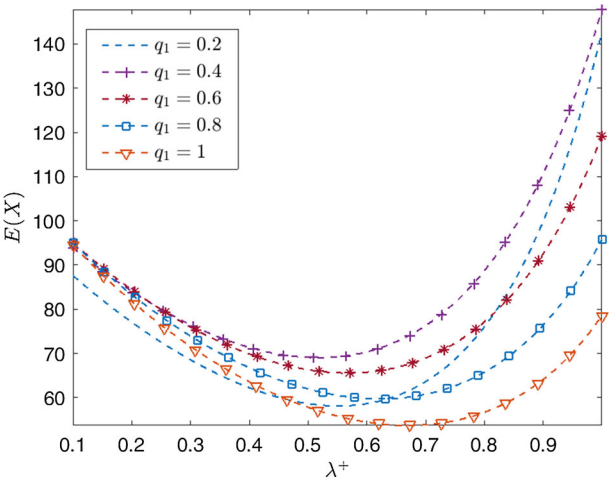

Fig. 6 Effect of q_1 when $N = 2, M = 5\ \alpha = 3.5, \lambda^- = 0.1$

see e.g., Mehmeti et al. (2023); Chen et al. (2018); Pappas et al. (2018); Ploumidis et al. (2017).

Thus, we focus on the following problem:

$$\begin{cases} \text{Maximize } TH_S, \\ \text{subject to} \\ E(X) \le \overline{E(X)}, \\ \bar{b} < \frac{\alpha^*(\lambda^-)[\lambda_+^r + (\lambda^r - \lambda_+^r)\beta^*(\lambda^r)]}{\lambda^r[\lambda_+^e + (\lambda_+^r - \lambda_+^e)\alpha^*(\lambda^-)]} - \frac{(1-\alpha^*(\lambda^-))(\lambda^e - \lambda_+^e)\beta^*(\lambda^e)}{\lambda^e[\lambda_+^e + (\lambda_+^r - \lambda_+^e)\alpha^*(\lambda^-)]}, \\ q_2 < q_1,\ q_4 < q_3, \\ 0 \le q_i \le 1,\ i = 1, 2, 3, 4, \end{cases} \tag{24}$$

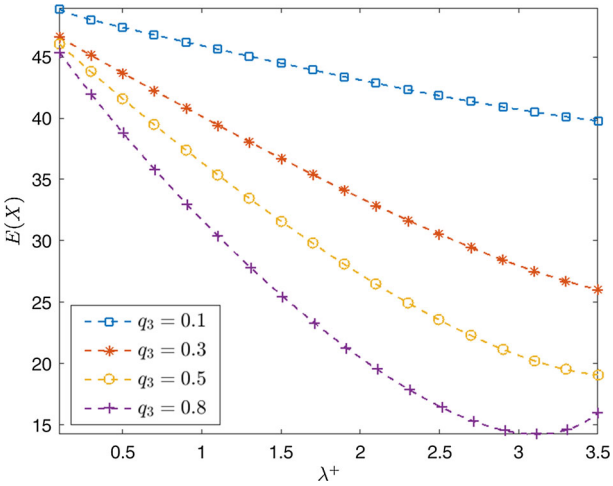

Fig. 7 Effect of q_3 when $N = 2$, $M = 5$ $\alpha = 3.5$, $\lambda^- = 0.1$, $q_1 = 0.6 > 0.5 = q_2$

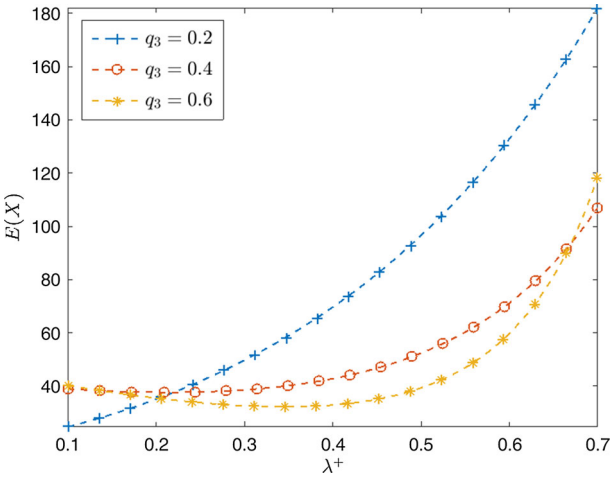

Fig. 8 Effect of q_3 when $N = 2$, $M = 5$ $\alpha = 3.5$, $\lambda^- = 0.1$, $q_1 = 0.2 < 0.5 = q_2$

where $\bar{b} = \frac{M}{\mu}, \beta^*(s) = (\frac{\mu}{\mu+s})^M, \alpha^*(s) = (\frac{\alpha}{\alpha+s})^N$, and $(\lambda^e, \lambda^e_+, \lambda^r, \lambda^r_+) = \lambda^+(q_1, q_2, q_3, q_4)$. In the optimisation problem (24), the first constraint refers to the maximum number of orbiting jobs, and the second one on the stability condition. The last two conditions are related to the ordering of the joining probabilities (q_1, q_2, q_3, q_4). In particular, we assume $q_2 < q_1$, $q_4 < q_3$, by assuming that if a customer knows that another one has already arrived, it is less likely to join. The optimal joining probabilities are in Tables 3, 4, 5.

Example 2a: Set $\lambda^+ = 2, \mu = 1.5, \alpha = 3, M = 4, N = 3, \overline{E(X)} = 20$. The optimal joining probabilities as functions of λ^- that maximize TH_S are given in Table 3. We observe that by increasing λ^-, TH_S increases too. It seems that when $\lambda^- < \lambda^+$, it is better to reject newly arriving customers after the occupation of the server by a primary customer (i.e., small q_1,

Table 3 Optimal values of joining probabilities as functions of λ^-

λ^-	$q_{opt} = (q_1, q_2, q_3, q_4)$	Optimal TH_S
0.1	(0.0001, 0.0001, 0.3148, 0.1431)	0.0788
1	(0.0547, 0.0287, 0.1719, 0.033)	0.3082
2	(0.0727, 0.027, 0, 0)	0.3289
4	(0.0094, 0.0048, 0.0962, 0.0271)	0.3452
6.1	(0.0006, 0.0003, 0.2856, 0.0687)	0.354

Table 4 Optimal values of joining probabilities as functions of N

N	$q_{opt} = (q_1, q_2, q_3, q_4)$	Optimal TH_S
1	(1, 0, 0.7199, 0)	0.328
2	(0.4614, 0.1678, 0.6339, 0.1532)	0.3221
5	(0.2669, 0.0383, 0.0309, 0.0157)	0.2916
15	(0.0062, 0.0031, 0.3373, 0.1446)	0.2737
30	(0.0001, 0, 0.3269, 0.1533)	0.2727

Table 5 Optimal values of joining probabilities as functions of M

M	$q_{opt} = (q_1, q_2, q_3, q_4)$	Optimal TH_S
1	(0.4755, 0.4755, 1, 0)	0.6702
2	(0.4259, 0.0974, 0.7478, 0.6007)	0.4958
5	(0.2721, 0.0415, 0.1334, 0.0512)	0.2484
15	(0.0104, 0.0053, 0.4011, 0.0143)	0.0936

q_2 compared with q_3, q_4). When $\lambda^- = \lambda^+$, TH_S is maximized by rejecting all the newly arriving customers that arrive after the occupation of the server by a retrial customer.

Example 2b: Set now $\lambda^- = 1$, $\lambda^+ = 0.5$, $\mu = 1.5$, $\alpha = 3$, $M = 4$, $\overline{E(X)} = 20$. In Table 4 we observe that the more we increase number of phases of retrial times, the less number of arriving customers we allow to enter the system. More precisely, we observe that when $N = 30$, TH_S is maximized when we reject customers that arrive after the arrival of a primary customer, and it is better to accept customers that arrive after a successful retrial that occupies the idle server. This make sense since by increasing the number of phases of retrial times we also increase the time to retrieve an orbiting customer. Thus, if we aim to maximize the system throughput we need to somehow prioritize the orbiting customers, and give more chances to arriving customers to join the system when an orbiting customer is in service.

Under the same setting, but now fixing $N = 4$, and by varying M, we observe in Table 5 that the optimal joining probabilities are more sensitive on the number of service phases compared with the number of retrieval phases. In particular, if the number of service phases increases, then the maximum throughput decreases very fast.

Example 2c: We now consider the optimisation problem in (24) with $\lambda^- = 1$, $\lambda^+ = 0.5$, $\mu = 1.5, \alpha = 3, M = 4, \overline{E(X)} = 20$, but now by excluding the constraints $q_2 < q_1, q_4 < q_3$. Table 6 contains the corresponding optimal joining probabilities. We can now observe that contrary to the case in Table 4, there is no specific trend on the values of the optimal joining probabilities. Moreover, in most of the cases $q_2^{opt} > q_1^{opt}$ and $q_4^{opt} > q_3^{opt}$. Thus, it seems that

Table 6 Optimal values of joining probabilities

N	$q_{opt} = (q_1, q_2, q_3, q_4)$	Optimal TH_S
1	(1, 0, 0.7199, 0)	0.328
2	(0.0933, 0.7073, 0.2362, 0.5123)	0.2908
5	(0.0744, 0.714, 0.2016, 0.5106)	0.287
15	(0.179, 0.9938, 0.0001, 0.7527)	0.295
30	(0.0293, 0.0374, 0.1065, 0.0439)	0.276

Table 7 Optimal values of joining probabilities as functions of λ^+, λ^-

λ^-	λ^+	$q_{opt} = (q_1, q_2, q_3, q_4)$	Optimal TH_S
0.1	0.5	(1, 1, 1, 0.2129)	0.5654
	1.5	(0.3813, 0.2076, 0.3001, 0.1480)	0.4742
	2.5	(0.1316, 0.0610, 0.2394, 0.1141)	0.4459
	3.5	(0.1170, 0.0359, 0.1898, 0.0742)	0.4435
0.5	0.5	(1, 0, 0.6694, 0)	0.5161
	1.5	(0.4142, 0.0682, 0.2474, 0.0543)	0.4726
	2.5	(0.0614, 0.0239, 0.3470, 0.0076)	0.4375
	3	(0.1497, 0.0017, 0.2509, 0.1190)	0.4476
1.5	0.5	(0.2422, 0.1515, 0.6712, 0.1349)	0.4676
	1.5	(0.0584, 0.0252, 0.3116, 0.0361)	0.4369
	2.5	(0.1795, 0.0153, 0.0205, 0.0088)	0.4479
	3	(0.0138, 0.0068, 0.4500, 0.0272)	0.4308

it is better to accept customers with small probability when the last event is the occupation of the idle server by a primary (resp. orbiting) customer, and then to increase the accepting probability for the subsequent customers. Similarly to the case in Table 4, the number of phases of the retrieval times heavily affects the optimal values of the joining probabilities.

Example 2d: Table 7 contains optimal joining probabilities as functions of λ^+, λ^-. Remind that these rates characterize the arrival rates when the last event is an arrival and a departure, respectively. We can observe that when both λ^+, λ^- are small, the maximum throughput is achieved by allowing all the customers that arrive after the occupation of the idle server by either a primary or a retrial customer, but keeping lower joining probability for subsequent arriving customers after the server occupation by an orbiting customer.

Example 2e: We now investigate the effect of service rate μ of a phase of the service time, on the optimal joining probabilities when the server is busy (we assume no specific ordering of joining probabilities). In particular, we assume that $\alpha = 3$, $M = 4$, $N = 2$. Table 8 contains the optimal joining probabilities when $\lambda^- = \lambda^+ = 0.5$, i.e., we assume that arrivals occur with rate λ^+, and if a service completion is the last event, the will join with certainty the system, and seek for optimal joining probabilities when an arrival has occurred (i.e., $\lambda^+ q_i$, $i = 1, 2, 3, 4$). Clearly, the more we increase μ, the less the expected service time, and the maximum throughput is achieved by allowing with high probability the arriving jobs to enter the system. For example, when $\mu > 4$, arriving jobs after the occupation of the idle server by a primary customer are accepted with certainty, and the more we increase μ, the more

Table 8 Optimal values of joining probabilities as functions of μ

μ	$q_{opt} = (q_1, q_2, q_3, q_4)$	Optimal TH_S
0.5	(0.1274, 0.0282, 0.0369, 0.1675)	0.4432
1.5	(0.2497, 0.1249, 0.2694, 0.0902)	0.4598
2.5	(0.2448, 0.3541, 0.4338, 0.3820)	0.4672
3.5	(0.1221, 0.7456, 0.2277, 0.4977)	0.4503
4.5	(1, 0.1736, 0.8286, 0)	0.5274
5.5	(1, 0.7668, 0.8736, 0)	0.5471
7.5	(1, 1, 1, 1)	0.6

Table 9 Optimal values of joining probabilities as functions of λ^+, μ

λ^+	μ	$\tilde{q}_{opt} = (q_1, q_2, q_3, q_4, q_5)$	Optimal TH_S
0.1	0.5	(1, 0, 1, 0, 0.4704)	0.5275
	1.5	(0.0084, 0.7301, 0.0046, 0.4843, 0.8284)	0.43
	2.5	(0.0399, 0.8051, 0.867, 0.4714, 0.7932)	0.4414
	3.5	(0.1928, 0.8553, 0.8635, 0.7085, 0.956)	0.4844
0.3	0.5	(1, 0, 1, 0, 0.151)	0.5275
	1.5	(1, 1, 1, 1, 0.1788)	0.6
	2.5	(1, 1, 1, 1, 0.5323)	0.6
	3.5	(1, 1, 1, 1, 0.572)	0.6
0.6	0.5	(0.1773, 0.2851, 0.299, 0.1755, 0.0298)	0.454
	1.5	(1, 1, 1, 0.3026, 0)	0.5687
	2.5	(1, 1, 1, 1, 0.0386)	0.6
	3.5	(1, 1, 1, 1, 0.3052)	0.6

we increase the joining probability for subsequent arrivals, as well as for arriving customers when the last event is the occupation of the server by an orbiting customer.

Example 2f: In all the preceding examples we have assumed that arrivals at an idle server join the system with certainty. Assume now that $\lambda^- = \lambda^+ q_5, q_5 \in [0, 1]$. In the following we observe that the TH_s is heavily affected by the arrivals that see an idle server; see Table 9. More precisely, we can observe that in order to maximize the system throughput, it is better to keep low the probabilities of entering the system just after a service completion. This seems to be crucial as λ^+ increases and μ remains as small as possible. However, as μ increases the situation becomes smoother. Interestingly, in all cases, the maximum throughput is achieved when the joining probability after a departure is strictly less than 1.

7 Conclusion and future work

In this work, we introduced the concept of *event-dependent* arrivals in the retrial setting. We studied the stability condition, and investigated the stationary behaviour both at service completion, and at an arbitrary epoch (by applying both the supplementary variable method and the Markov renewal theory). We provided results for all the well known retrial policies, i.e., with general retrials, and under the linear retrial policy (which combines the classical and

the constant retrial policy). Explicit expressions for various performance metrics are derived, and used to numerically investigate the effect of event-dependency on system's performance. We also formulated and solved constrained optimisation problems that shown insights into the effect of event dependency on the optimal joining probabilities.

It seems that event-dependency is a result of customer's strategic behaviour, although we postpone the formal game-theoretical investigation as a future work. Indeed, consider an initial unobservable system with a potential arrival rate λ. At an arrival epoch, a customer decides either to join or not system. Although the system is unobservable, an arriving customer is informed about the last realized event; an arrival or a service completion. Let the net benefit for a customer who joins by the value of service, B, minus the cost of waiting proportional to a waiting cost per time unit, C. Given that the expected waiting time of an arriving customer is different whether the last event was an arrival or a service completion, a strategy after an arrival and after a service is described by five different probabilities of joining $\bar{p}^+ := (q_1, q_2, q_3, q_4)$ and p^-, by taking also into account the type of a customer that occupies the server (i.e., a retrial or a primary). More precisely, let $\bar{\lambda}^+ := \lambda \bar{p}^+$, $\lambda^- := \lambda p^-$. The values of p^-, $\bar{p}^+$ are such that the stability condition is satisfied, and the values of $\bar{p}^+$ are such that: q_1 refers to the case the last event is an arrival of an external customer to an idle server, q_2 refers to the case the last event is a subsequent arrival to a busy server, which was initially occupied by an external customer, q_3 refers to the case the last event is an arrival of a retrial customer to an idle server, and q_4 refers to the case the last event is a subsequent arrival to a busy server, which was initially occupied by a retrial customer. Having that description in mind one can consider the net benefit for a customer who arrives after and arrival and service and seeking for optimal joining probabilities and identify the equilibrium, social and profit maximizing strategies.

Several other questions are open for future research. It would be worth investigating the effect of event-dependency on service times, i.e., to allow the service rates to depend on the last realized event. Our modeling framework allows also to consider several characteristics, such as priorities, vacations, breakdowns and repairs. Another option is to investigate the possibility of obtaining the performance metrics through the QMCD (Queueing & Markov Chain Decomposition) method Baron et al. (2018).

7.1 On the admission control problem

We now discuss how one can investigate the admission control problem Koole (2007) by using an MDP framework. In the admission control problem, the system manager must determine upon the the arrival of a customer whether to allow him/her to enter the system or to reject him/her.

With ultimate goal to maximize the throughput of served customers with a service level constraint on the expected number of orbiting customers, such a decision can be based on the number of customers present in orbit, the distribution of the remaining service time and the last realized event. Due to the fact that a deterministic admission policy is easier (than the randomized) to implement in real system one can focus on determining a threshold policy, depending on the last realized event and on the remaining service time of the customer in service. More precisely, given that there will be k customers in orbit, and the remaining service equal s, we seek for thresholds, say $k_s^{e,1}$, $k_s^{e,2}$, $k_s^{r,1}$, $k_s^{r,2}$, k^- such that,

– If a primary arrival occurs after an arrival of a primary customer that occupies an idle server, then, this customer is rejected (and leave the system without service) if $k > k_s^{e,1}$.

- If a primary arrival occurs after at least one primary arrival after the occupation of an idle server by a primary customer, then, this customer is rejected (and leave the system without service) if $k > k_s^{e,2}$.
- If a primary arrival occurs after an arrival of an orbiting customer that occupies an idle server, then, this customer is rejected if $k > k_s^{r,1}$.
- If a primary arrival occurs after at least one primary arrival after the occupation of an idle server by an orbiting customer, then, this customer is rejected (and leave the system without service) if $k > k_s^{r,2}$.
- If a primary arrival occurs after a service completion, then, this customer is rejected if $k > k^-$ (i.e., we give priority to the orbiting customers to occupy the idle server, if there are more than k^-).

To proceed with the set up of the problem, we assume that the service time is Coxian distributed (note that this is a proper choice to approximate the service time distribution since it is dense in the class of all non-negative distributions. However, one still needs to set the number of phases to appropriately approximate the considered service time distribution Horváth and Telek (2002)). Coxian distribution is defined by the parameters $\mu_j > 0$, and $r_j \in [0, 1]$, $1 \leq j \leq N$, with $r_1 = 0$, where r_j (resp. $\bar{r}_j = 1 - r_j$) denotes the probability of entering the remaining phase $j-1$ after leaving the remaining phase j (resp. completing the service time after leaving phase j), while and the parameter μ_j is the rate of the exponentially distributed random duration of the remaining phase j. For simplicity, we assume that the retrial times can be exponentially distributed with rate α. We seek for necessary conditions under which a deterministic threshold-type policy can be optimal. This can be done by using the value iteration technique, where structural properties of the value function can be proven by induction.

Our problem can be considered as a constrained Markov decision process (MDP) Altman (1999) (i.e., maximize the throughput of served customers with a constraint on the expected number of customers in the system), and can be investigated using various techniques, such that the one that introduces the constraint into the objective function by using a Lagrange multiplier. Then, one can realize that the optimal policy for a certain Lagrange multiplier is optimal for the constrained problem when the value of the constraint under this policy attains exactly $E(X)$. Thus, it follows that this policy is stationary and randomizes in at most one state. Then, the optimisation problem can be rewritten as $min(E(X) - c \times TH_S)$, where the non-negative coefficient c is the Lagrange multiplier which refers to the relative importance, given by the system manager, of the throughput of served customers (TH_S) compared to the expected number of customers in the system ($E(X)$).

Denote by (x, y) the state of the system, where $x \geq 0$ is the number of remaining phases of work, and y denotes the nature of the last event, i.e., $y \in \{+_e, +_{e,1}, +_{e,2}, +_r, +_{r,1}, +_{r,2}, -\}$, where $y = +_e$ (resp. $y = +_r$), when the last event is the occupation of the idle server by an external (resp. orbiting) customer, $y = +_{e,1}$ (resp. $y = +_{r,1}$) when the last event is the first arrival after the occupation of the idle server by an external (resp. orbiting) customer, $y = +_{e,2}$ (resp. $y = +_{r,2}$) when at least one arrival has occurred after the occupation of the idle server by an external (resp. orbiting) customer, and $y = -$ when the last event is a service completion. Denote the transition rate from state (x, y) to state (x', y') by $q_{(x,y),(x',y')}$. Then,

$$
q_{(x,y),(x',y')} = \begin{cases} \alpha, & x > 0, y = -, x' = x, y' = +_r, \\ \lambda^-, & x \geq 0, y = -, x' = x + N, y' = +_e, \\ \lambda^e, & x \geq 0, y = +_e, x' = x + N, y' = +_{e,1}, \\ \lambda^e_+, & x \geq 0, y = \in \{+_{e,1}, +_{e,2}\}, x' = x + N, y' = +_{e,2}, \\ \lambda^r, & x \geq 0, y = +_r, x' = x + N, y' = +_{r,1}, \\ \lambda^r_+, & x \geq 0, y \in \{+_{r,1}, +_{r,2}\}, x' = x + N, y' = +_{r,2}, \\ r_j\mu_j, & x = kN + j, y \in \{+_e, +_{e,1}, +_{e,2}, +_r, +_{r,1}, +_{r,2}\}, x' = x - 1, y' = y, \\ \bar{r}_j\mu_j, & x = kN + j, y \in \{+_e, +_{e,1}, +_{e,2}, +_r, +_{r,1}, +_{r,2}\}, x' = kN, y' = -, \\ 0, & \text{otherwise.} \end{cases}
$$

It would be better to proceed by applying a uniformization method (Puterman (1994), Sect. 11.5.2) and consider the discrete time version of the model asking $\lambda^- + \lambda^e + \lambda^e_+ + \lambda^r + \lambda^r_+ + \alpha + \sum_{j=1}^N \mu_j = 1$. Then, we formulate a two-step value function so that to separate transitions and actions, and then, we define the dynamic programming value functions $V_n^y(x)$, $y \in \{+_e, +_{e,1}, +_{e,2}, +_r, +_{r,1}, +_{r,2}, -\}$, $W_n^y(x)$, $y \in \{+_{e,1}, +_{e,2}, +_{r,1}, +_{r,2}, -\}$, with $V_0^y(x) = W_0^y(x) = 0$, $x \geq 0$. In particular, for $1 \leq j \leq N$, $k \geq 0$,

$$
\begin{aligned}
V_{n+1}^{+_e}(kN + j) &= k + 1 + \lambda^e W_n^{+_{e,1}}(kN + j) + r_j\mu_j V_n^{+_e}(kN + j - 1) \\
&\quad + \bar{r}_j\mu_j V_n^-(kN) + (1 - \lambda^e - \mu_j) V_n^{+_e}(kN + j), \\
V_{n+1}^{+_{e,1}}(kN + j) &= k + 1 + \lambda^e_+ W_n^{+_{e,2}}(kN + j) + r_j\mu_j V_n^{+_{e,1}}(kN + j - 1) \\
&\quad + \bar{r}_j\mu_j V_n^-(kN) + (1 - \lambda^e_+ - \mu_j) V_n^{+_{e,1}}(kN + j), \\
V_{n+1}^{+_{e,2}}(kN + j) &= V_{n+1}^{+_{e,1}}(kN + j), \\
V_{n+1}^{+_r}(kN + j) &= k + 1 + \lambda^r W_n^{+_{r,1}}(kN + j) + r_j\mu_j V_n^{+_r}(kN + j - 1) \\
&\quad + \bar{r}_j\mu_j V_n^-(kN) + (1 - \lambda^r - \mu_j) V_n^{+_r}(kN + j), \\
V_{n+1}^{+_{r,1}}(kN + j) &= k + 1 + \lambda^r_+ W_n^{+_{r,2}}(kN + j) + r_j\mu_j V_n^{+_{r,1}}(kN + j - 1) \\
&\quad + \bar{r}_j\mu_j V_n^-(kN) + (1 - \lambda^e_+ - \mu_j) V_n^{+_{r,1}}(kN + j), \\
V_{n+1}^{+_{r,2}}(kN + j) &= V_{n+1}^{+_{r,1}}(kN + j), \\
V_{n+1}^-(kN) &= k + \lambda^- W_n^-((k + 1)N) + \alpha V_n^{+_r}(kN) \\
&\quad + (1 - \lambda^- - \alpha) V_n^-(kN),
\end{aligned}
\tag{25}
$$

where the operator W_n denotes the decision to accept or to reject a newly arriving customer. In particular,

$$
\begin{aligned}
W_n^{+_{e,1}}(kN + j) &= min\{V_n^{+_e}(kN + j) - c, V_n^{+_e}(kN + j)\}, \\
W_n^{+_{e,2}}(kN + j) &= min\{V_n^{+_{e,1}}(kN + j) - c, V_n^{+_{e,1}}(kN + j)\}, \\
W_n^{+_{r,1}}(kN + j) &= min\{V_n^{+_r}(kN + j) - c, V_n^{+_r}(kN + j)\}, \\
W_n^{+_{r,2}}(kN + j) &= min\{V_n^{+_{r,1}}(kN + j) - c, V_n^{+_{r,1}}(kN + j)\}, \\
W_n^-(kN + j) &= min\{V_n^{+_e}(kN + j) - c, V_n^-(kN + j)\}.
\end{aligned}
\tag{26}
$$

For each $n > 0$, $x \geq 0$, we have a minimizing action upon a customer's arrival, i.e., to allow this customer to enter the system or reject him/her. In order to obtain the long-run average optimal actions one can use the value iteration method developed in Howard (1960), by recursively evaluating V_n using equations (25), (26), for $n \geq 0$. The form of the optimal policy is closely related to the structural properties of the value function, and it would be interesting to derive conditions for which one can have optimal threshold policies based on the number of orbiting customers for a given number of remaining phases of service. We believe that by using a standard MDP approach where structural properties of the value

function are proved by induction, one can derive such optimal threshold-type policy. The formal and detailed investigation of this task is postponed as a future research work.

Due to the complicated nature of this problem, it seems to be a difficult task to find, and further to implement the optimal policy in a real system. With that in mind, one can consider a simplified version, where the thresholds on the orbit queue length would depend only on the number of orbiting customers, and further to assume that $k^y = k^+$, $y \in \{+_e, +_{e,1}, +_{e,2}, +_r, +_{r,1}, +_{r,2}\}$, so that we would have two thresholds, k^+, k^-: If an arrival occur after an arrival (resp. a service), this customer is rejected if $k > k^+$ (resp. $k > k^-$). Although the proposed policy would be not optimal, it is simpler to implement in practice. Moreover, compared to the Coxian approximation, which leads to the application of the matrix-geometric method to derive the performance measures, the simplified approach we propose here leads to the derivation of explicit expressions for the performance metrics for any service/retrieval time distribution.

Assuming that $k^+ > k^-$, the stationary probabilities at departure instants, say $\tilde{p} := (\tilde{\pi}_0, \ldots, \tilde{\pi}_{k^+})$, can be derived using a similar approach as in Sect. 3, where now the one step transition probability matrix, say $\tilde{P} := (\tilde{p}_{i,j})$, $i, j = 0, \ldots, k^+$, is given by

$$
\begin{aligned}
\tilde{p}_{0,j} &= b^e_j,\ j = 0, 1, \ldots, k^+ - 1,\\
\tilde{p}_{0,k^+} &= 1 - \textstyle\sum_{j=0}^{k^+-1} b^e_j,\\
\tilde{p}_{i,i-1} &= \alpha^*(\lambda^-) b^r_0,\\
\tilde{p}_{i,j} &= (1 - \alpha^*(\lambda^-)) b^e_{j-i} + \alpha^*(\lambda^-) b^r_{j-i+1},\ i = 1, 2, \ldots, k^-,\ j = i, i+1, \ldots, k^+ - 1,\\
\tilde{p}_{i,k^+} &= 1 - \textstyle\sum_{j=i-1}^{k^+-1} \tilde{p}_{i,j},\ i = 1, 2, \ldots, k^-,\\
\tilde{p}_{i,i-1} &= 1,\ \tilde{p}_{i,j} = 0,\ j \neq i-1,\ i = k^- + 1, \ldots, k^+,
\end{aligned}
$$

Note that from $i = k^- + 1$ to $i = k^+$, there cannot be any arrival during a service since the first arrival after the service completion cannot occur. Thus, we can only have an orbiting customer that will enter the service with probability 1, and the orbiting customers will be reduced by one at the next service completion.

One can then solve the system $\tilde{\pi} = \tilde{\pi}\tilde{P}$, with $\sum_{i=0}^{k^+} \tilde{\pi}_i = 1$, to obtain the stationary distribution of the orbit size at service completion epochs. The stationary probabilities at an arbitrary epoch can be also derived in a similar fashion as in Sect. 4, so further details are omitted. With these results, one can solve optimisation problems (as for the case of the infinite thresholds in the optimisation problem in (24)) to find optimal thresholds, say k^+_{opt}, k^-_{opt}, to maximize the generating system throughput. Again, we do not claim that this can be an optimal two-threshold policy.

Acknowledgements I would like to thank the Editor and the Reviewers for the insightful remarks, which helped to improve the original exposition.

Funding Open access funding provided by HEAL-Link Greece.

Declarations

Conflict of interest We declare that we have no financial and personal relationships with other people or organizations that can inappropriately influence our work. There is no professional or other personal interest of any nature or kind in any product, service, and/or company that could be construed as influencing the position presented in, or the review of, the manuscript entitled.

Appendix A: Proof of Theorem 1

(Sufficiency) We use standard Foster-Lyapunov arguments. Let $A_k^{(j)}(1) = \frac{d^j}{dz^j} A_k(z)|_{z=1}$, $k = e, r$, $j = 1, 2$. Then, the mean drifts are given by:

$$\begin{aligned} x_n &= E(X_{i+1} - X_i \mid X_i = n) = E(A_{i+1}(B_{i+1}) \mid X_i = n) - E(B_{i+1} \mid X_i = n) \\ &= (1 - \alpha^*(\lambda^-)(1 - \delta_{0,n}))A_e^{(1)}(1) + \alpha^*(\lambda^-)(1 - \delta_{0,n})(A_r^{(1)}(1) - 1), \end{aligned}$$

where $\delta_{0,n}$ denotes the Kronecker's delta. Suppose that

$$(1 - \alpha^*(\lambda^-))A_e^{(1)}(1) + \alpha^*(\lambda^-)A_r^{(1)}(1) < \alpha^*(\lambda^-).$$

Then,

$$\epsilon = \tfrac{1}{2}[\alpha^*(\lambda^-) - (1 - \alpha^*(\lambda^-))A_e^{(1)}(1) - \alpha^*(\lambda^-)A_r^{(1)}(1)] > 0,$$

and there exists

$$\begin{aligned} \lim_{n\to\infty} x_n &= (1 - \alpha^*(\lambda^-))A_e^{(1)}(1) + \alpha^*(\lambda^-)A_r^{(1)}(1) - \alpha^*(\lambda^-) \\ &= -2\epsilon < -\epsilon. \end{aligned}$$

Hence, $x_n < -\epsilon$ for all the states except for a finite number. Therefore,

$$(1 - \alpha^*(\lambda^-))A_e^{(1)}(1) + \alpha^*(\lambda^-)A_r^{(1)}(1) < \alpha^*(\lambda^-), \tag{A1}$$

is a sufficient condition for the ergodicity of the embedded Markov chain. After straightforward computations (A1) reduces to (4). **(Necessity)** The necessity part can be proved using the Kaplan's condition and further details are omitted (The necessity can also proved using the generating function approach; see Theorem 3).

Appendix B: Proof of Lemma 1

Considering the evolution of $\{(C(t), X(t), I(t), Z(t)); t \geq 0\}$ in the interval $[0, t + dt]$ and conditioning on its value at time t we have for $dt \to 0^+$, the equations

$$\frac{d}{dt} p_{0,0}(t) = -\lambda^- p_{0,0}(t) + \sum_{k=2}^{3} p_{1,0}^{(k)}(0, t), \tag{B2}$$

$$\left(\frac{\partial}{\partial t} - \frac{\partial}{\partial r}\right) p_{0,j}(r, t) = -\lambda^- p_{0,j}(r, t) + a(r) \sum_{k=2}^{7} p_{1,j}^{(k)}(0, t), \ j \geq 1, \tag{B3}$$

$$\left(\frac{\partial}{\partial t} - \frac{\partial}{\partial r}\right) p_{1,j}^{(2)}(r, t) = -\lambda^e p_{1,j}^{(2)}(r, t) + \lambda^- p_{0,j}(t)b(r), \ j \geq 0, \tag{B4}$$

$$\left(\frac{\partial}{\partial t} - \frac{\partial}{\partial r}\right) p_{1,j}^{(3)}(r, t) = -\lambda^r p_{1,j}^{(3)}(r, t) + p_{0,j+1}(0, t)b(r), \ j \geq 0, \tag{B5}$$

$$\left(\frac{\partial}{\partial t}-\frac{\partial}{\partial r}\right)p_{1,j}^{(4)}(r,t)=-\lambda_{+}^{e}p_{1,j}^{(4)}(r,t)+\lambda^{e}p_{1,j-1}^{(2)}(r,t),\ j\geq 1, \tag{B6}$$

$$\left(\frac{\partial}{\partial t}-\frac{\partial}{\partial r}\right)p_{1,j}^{(5)}(r,t)=-\lambda_{+}^{e}p_{1,j}^{(5)}(r,t)+\lambda_{+}^{e}(p_{1,j-1}^{(5)}(r,t)+p_{1,j-1}^{(4)}(r,t)),\ j\geq 2, \tag{B7}$$

$$\left(\frac{\partial}{\partial t}-\frac{\partial}{\partial r}\right)p_{1,j}^{(6)}(r,t)=-\lambda_{+}^{r}p_{1,j}^{(6)}(r,t)+\lambda^{r}p_{1,j-1}^{(3)}(r,t),\ j\geq 1, \tag{B8}$$

$$\left(\frac{\partial}{\partial t}-\frac{\partial}{\partial r}\right)p_{1,j}^{(7)}(r,t)=-\lambda_{+}^{r}p_{1,j}^{(7)}(r,t)+\lambda_{+}^{r}(p_{1,j-1}^{(7)}(r,t)+p_{1,j-1}^{(6)}(r,t)),\ j\geq 2. \tag{B9}$$

Letting $t\rightarrow\infty$, Eqs. (B2)-(B9) reduce to those given in (7).

Appendix C: Proof of Theorem 3

Multiplying the second in (7) with e^{-sr}, integrating with respect to $s\in[0,\infty)$, and having in mind that $p_{0,j}^{*}(s)=\int_{0}^{\infty}e^{-sr}p_{0,j}(r)dr$, yields

$$(s-\lambda^{-})p_{0,j}^{*}(s)=p_{0,j}(0)-\alpha^{*}(s)\sum_{k=2}^{7}p_{1,j}^{(k)}(0).$$

Multiplying with z^{j}, and summing for all $j\geq 0$ yields

$$(s-\lambda^{-})P_{0}^{*}(s,z)=P_{0}(0,z)-\alpha^{*}(s)\left[\sum_{k=2}^{7}P_{1,k}(0,z)-\lambda^{-}p_{0,0}\right]. \tag{C10}$$

For $s=\lambda^{-}$ we obtain,

$$P_{0}(0,z)=\alpha^{*}(\lambda^{-})\left[\sum_{k=2}^{7}P_{1,k}(0,z)-\lambda^{-}p_{0,0}\right], \tag{C11}$$

and substituting back in (C10),

$$P_{0}^{*}(s,z)=\frac{\alpha^{*}(\lambda^{-})-\alpha^{*}(s)}{s-\lambda^{-}}\left[\sum_{k=2}^{7}P_{1,k}(0,z)-\lambda^{-}p_{0,0}\right]. \tag{C12}$$

By repeating the same procedure for the third in (7), we obtain

$$(s-\lambda^{e})P_{1,2}^{*}(s,z)=P_{1,2}(0,z)-\lambda^{-}\beta^{*}(s)[p_{0,0}+P_{0}^{*}(0,z)], \tag{C13}$$

and setting $s=\lambda^{e}$ yields

$$P_{1,2}(0,z)=\lambda^{-}\beta^{*}(\lambda^{e})[p_{0,0}+P_{0}^{*}(0,z)].$$

Substituting back in (C13) yields

$$P_{1,2}^{*}(s,z)=\frac{\lambda^{-}}{s-\lambda^{e}}(\beta^{*}(\lambda^{e})-\beta^{*}(s))[p_{0,0}+P_{0}^{*}(0,z)]. \tag{C14}$$

So the second in (9) has been proved. By applying the same procedure for the forth in (7) we obtain

$$\begin{aligned}P_{1,3}(0,z)&=\frac{\beta^{*}(\lambda^{r})}{z}P_{0}(0,z),\\ P_{1,3}^{*}(s,z)&=\frac{\beta^{*}(\lambda^{r})-\beta^{*}(s)}{z(s-\lambda^{r})}P_{0}(0,z).\end{aligned} \tag{C15}$$

Now repeat the same procedure for the fifth and sixth in (7) we obtain

$$(s-\lambda_{+}^{e})P_{1,4}^{*}(s,z)=P_{1,4}(0,z)-\lambda^{e}zP_{1,2}^{*}(s,z), \tag{C16}$$

$$(s-\lambda_+^e)P_{1,5}^*(s,z) = P_{1,5}(0,z) - \lambda_+^e z(P_{1,5}^*(s,z) + P_{1,4}^*(s,z)). \tag{C17}$$

Substituting (C14) in (C16), and setting $s = \lambda_+^e$ yields an expressions for $P_{1,4}(0,z)$. Substituting back in (C16) yields,

$$P_{1,4}^*(s,z) = \frac{\lambda^e\lambda^- z}{s-\lambda_+^e}\left[\frac{\beta^*(\lambda^e)-\beta^*(\lambda_+^e)}{\lambda_+^e-\lambda^e} - \frac{\beta^*(\lambda^e)-\beta^*(s)}{s-\lambda^e}\right](p_{0,0}+P_0^*(0,z)). \tag{C18}$$

Using (C18) in (C17), setting now $s = \lambda_+^e(1-z)$, and following the same procedure as above, we obtain

$$\begin{aligned} P_{1,5}^*(s,z) = {} & \frac{\lambda^e\lambda^- z}{s-\lambda_+^e(1-z)}\left[\frac{\beta^*(\lambda^e)-\beta^*(\lambda_+^e(1-z))}{\lambda_+^e(1-z)-\lambda^e} - \frac{\beta^*(\lambda^e)-\beta^*(\lambda_+^e)}{\lambda_+^e-\lambda^e}\left(\frac{s-\lambda_+^e(1-z)}{s-\lambda_+^e}\right)\right. \\ & \left. + \frac{\lambda_+^e z}{s-\lambda_+^e}\left(\frac{\beta^*(\lambda^e)-\beta^*(s)}{s-\lambda^e}\right)\right](p_{0,0}+P_0^*(0,z)). \end{aligned} \tag{C19}$$

Summing the above equations and using (C14), we obtain

$$\begin{aligned} \textstyle\sum_{k=4}^5 P_{1,k}(0,z) &= \frac{\lambda^-\lambda^e z[\beta^*(\lambda^e)-\beta^*(\lambda_+^e(1-z))]}{\lambda_+^e(1-z)-\lambda^e}[p_{0,0}+P_0^*(0,z)], \\ \textstyle\sum_{k=4}^5 P_{1,k}^*(s,z) &= \frac{\lambda^-\lambda^e z}{s-\lambda_+^e(1-z)}\left(\frac{\beta^*(\lambda^e)-\beta^*(\lambda_+^e(1-z))}{\lambda_+^e(1-z)-\lambda^e} - \frac{\beta^*(\lambda^e)-\beta^*(s)}{s-\lambda^e}\right) \\ &\quad \times[p_{0,0}+P_0^*(0,z)]. \end{aligned} \tag{C20}$$

Similar operations for the last two equations in (7) yield

$$\begin{aligned} P_{1,6}^*(s,z) &= \frac{\lambda^r}{s-\lambda_+^r}\left[\frac{\beta^*(\lambda^r)-\beta^*(\lambda_+^r)}{\lambda_+^r-\lambda^r} - \frac{\beta^*(\lambda^r)-\beta^*(s)}{s-\lambda^r}\right]P_0(0,z), \\ P_{1,7}^*(s,z) &= \frac{\lambda^r}{s-\lambda_+^r(1-z)}\left[\frac{\beta^*(\lambda^r)-\beta^*(\lambda_+^r(1-z))}{\lambda_+^r(1-z)-\lambda^r} - \frac{\beta^*(\lambda^r)-\beta^*(\lambda_+^r)}{\lambda_+^r-\lambda^r}\left(\frac{s-\lambda_+^r(1-z)}{s-\lambda_+^r}\right)\right. \\ &\quad \left. + \frac{\lambda_+^r z}{s-\lambda_+^r}\left(\frac{\beta^*(\lambda^r)-\beta^*(s)}{s-\lambda^r}\right)\right]P_0(0,z). \end{aligned} \tag{C21}$$

and

$$\begin{aligned} \textstyle\sum_{k=6}^7 P_{1,k}(0,z) &= \frac{\lambda^r[\beta^*(\lambda^r)-\beta^*(\lambda_+^r(1-z))]}{\lambda_+^r(1-z)-\lambda^r}P_0(0,z), \\ \textstyle\sum_{k=6}^7 P_{1,k}^*(s,z) &= \frac{\lambda^r P_0(0,z)}{s-\lambda_+^r(1-z)}\left(\frac{\beta^*(\lambda^r)-\beta^*(\lambda_+^r(1-z))}{\lambda_+^r(1-z)-\lambda^r} - \frac{\beta^*(\lambda^r)-\beta^*(s)}{s-\lambda^r}\right). \end{aligned} \tag{C22}$$

Now from (C12),

$$\sum_{k=2}^{7} P_{1,k}(0,z) = \lambda^- p_{0,0} + \frac{P_0(0,z)}{\alpha^*(\lambda^-)}. \tag{C23}$$

Using (C14)-(C22), and after lengthy algebraic computations, we obtain

$$\sum_{k=2}^{7} P_{1,k}(0,z) = \lambda^- A_e(z)(p_{0,0}+P_0^*(0,z)) + \frac{P_0(0,z)}{z}A_r(z). \tag{C24}$$

Equating (C23), (C24) we obtain

$$P_0(0,z) = \frac{\lambda^- z\alpha^*(\lambda^-)(p_{0,0}(A_e(z)-1)+A_e(z)P_0^*(0,z))}{z-\alpha^*(\lambda^-)A_r(z)}. \tag{C25}$$

Using (C11), (C25), we obtain

$$P_0^*(s,z) = \frac{\alpha^*(\lambda^-)-\alpha^*(s)}{s-\lambda^-}\frac{P_0(0,z)}{\alpha^*(\lambda^-)}. \tag{C26}$$

Setting $s = 0$ in (C26), and using (C25) we obtain the first in (9). Setting $s = 0$ in (C15), and letting $K(z) := \frac{P_0(0,z)}{z}$, we obtain the third in (9). Similarly we can obtain the rest expressions in (9).

Having obtain the expressions in (9), and having in mind that $1 = p_{0,0} + P_0^*(0,1) + \sum_{k=2}^{7} P_{1,k}^*(0,1)$ we derive after lengthy but straightforward computations the probability of an empty system $p_{0,0}$.

Appendix D: Proof of Corollary 2

Let $A_k^{(j)}(1) = \frac{d^j}{dz^j} A_k(z)|_{z=1}$, $k = e, r$, $j = 1, 2$. Then, for $k = e, r$ we have

$$A_k^{(1)}(1) = \frac{(\lambda^k - \lambda_+^k)(1-\beta^*(\lambda^k)) + \lambda^k \lambda_+^k \bar{b}^{(1)}}{\lambda^k},$$
$$A_k^{(2)}(1) = \lambda_+^k (2\bar{b}^{(1)} + \lambda_+^k \bar{b}^{(2)}) - 2\frac{\lambda_+^k}{\lambda^k} A_k^{(1)}(1).$$

Let also,

$$G := \frac{\alpha^*(\lambda^-)[(2A_e^{(1)}(1) + A_e^{(2)}(1))(1 - A_r^{(1)}(1)) + A_e^{(1)}(1) A_r^{(2)}(1)]}{2[A_e^{(1)}(1)(\alpha^*(\lambda^-) - 1) + \alpha^*(\lambda^-)(1 - A_r^{(1)}(1))]^2},$$
$$F := \frac{[p_{0,0}(1-\alpha^*(\lambda^-))G + A_e^{(1)}(1) P(C=0)] - P_0^*(0,1)(1-\alpha^*(\lambda^-) A_r^{(1)}(1))}{(1-\alpha^*(\lambda^-))^2}.$$

Then, using the results in Theorem 3, and differentiating with respect to z, at $z = 1$, we obtain after heavy computations

$$\begin{aligned}
\tfrac{d}{dz} P_0^*(0,z)|_{z=1} &= p_{0,0}(1 - \alpha^*(\lambda^-))G, \\
\tfrac{d}{dz} P_{1,2}^*(0,z)|_{z=1} &= p_{0,0}(1 - \alpha^*(\lambda^-))\lambda^- \tfrac{1-\beta^*(\lambda^e)}{\lambda^e} G, \\
\tfrac{d}{dz} P_{1,3}^*(0,z)|_{z=1} &= \lambda^- \alpha^*(\lambda^-) \tfrac{1-\beta^*(\lambda^r)}{\lambda^r} F, \\
\tfrac{d}{dz} \textstyle\sum_{k=4}^{5} P_{1,k}^*(0,z)|_{z=1} &= \lambda^- (\bar{b}^{(1)} - \tfrac{1-\beta^*(\lambda^e)}{\lambda^e})[\tfrac{d}{dz} P_0^*(0,z)|_{z=1} + \tfrac{\lambda^e - \lambda_+^e}{\lambda^e} P(C=0)] \\
&\quad + \tfrac{\lambda^- \lambda_+^e \bar{b}^{(2)}}{2} P(C=0), \\
\tfrac{d}{dz} \textstyle\sum_{k=6}^{7} P_{1,k}^*(0,z)|_{z=1} &= (\bar{b}^{(1)} - \tfrac{1-\beta^*(\lambda^r)}{\lambda^r})(\tfrac{\lambda^r - \lambda_+^r}{\lambda^e} P_0(0,1) + F) + \tfrac{\lambda_+^r \bar{b}^{(2)}}{2} P_0(0,1).
\end{aligned} \tag{D27}$$

Sum the terms in (D27) to obtain $E(X)$ in (10). Let TH_S is the throughput generated by the system. Then,

$$\begin{aligned}
TH_S = \lambda^- P_0^*(0,1) + \lambda^e P_{1,2}^*(0,1) + \lambda_+^e \textstyle\sum_{k=4}^{5} P_{1,k}^*(0,1) \\
+ \lambda^r P_{1,3}^*(0,1) + \lambda_+^r \textstyle\sum_{k=6}^{7} P_{1,k}^*(0,1).
\end{aligned}$$

After tedious computations we can have the expression given in (10).

Appendix E: Proof of Theorem 5

Let $P_0(z) = \sum_{j=0}^{\infty} p_{0,j} z^j$, $P_{1,k}(z) = \sum_{j=0}^{\infty} p_{1,j}^{(k)} z^j$, $|z| \leq 1$. Writing down the global balance equations and forming the generating functions we come up with the following

system:

$$\begin{aligned} P_0(z)(\lambda^- + \alpha) - \alpha p_{0,0} &= \mu(P_{1,1}(z) + P_{1,2}(z)), \\ P_{1,1}(z) &= \tfrac{\lambda^-}{\mu+\lambda^e(1-z)} P_0(z), \\ P_{1,2}(z) &= \tfrac{\alpha(P_0(z)-p_{0,0})}{z(\mu+\lambda^r(1-z))}. \end{aligned}$$

Simple calculation leads to

$$P_0(z) = \frac{\alpha(\mu + \lambda^e(1-z))(\lambda^r z - \mu)p_{0,0}}{\alpha(\mu + \lambda^e(1-z))(\lambda^r z - \mu) + \lambda^-\lambda^e z(\mu + \lambda^e(1-z))}.$$

Setting $z = 1$, and asking $P_0(1)+P_{1,1}(1)+P_{1,2}(1) = 1$, we obtain $p_{0,0}$ as given in the first in (11), so that $\alpha\lambda^r + \lambda^-\lambda^e < \alpha\mu$ is a necessary condition for ergodicity. Having obtained $p_{0,0}$, simple algebraic calculation provide the stationary probabilities of server's state as given in (14).

Note that $P_0(z)$ can be written as follows:

$$P_0(z) = \frac{\alpha p_{0,0}}{\lambda^- + \alpha}[1 + \mu\lambda^- \frac{\lambda^r z - \lambda^e - \mu}{f(z)}],$$

where $f(z)$ as given in Theorem 5. Note that $f(0) = -\alpha\mu(\lambda^e+\mu)$, $f(1) = \mu(\alpha\lambda^r+\lambda^e\lambda^- - \alpha\mu) < 0$, due to the ergodicity condition, and $\frac{d^2 f(z)}{dz^2} = -2\lambda^e\lambda^r(\lambda^- + \alpha) < 0$, so that $f(z)$ is a concave function for all z, and $\frac{df(z)}{dz} = 0 \Rightarrow z = \frac{\alpha\mu\lambda^r+\lambda^e(\lambda^-+\alpha)(\lambda^r+\mu)}{2\lambda^e\lambda^r(\lambda^-+\alpha)} > 1$. Thus, $f(z) = 0$ has two positive zeros, say z_1, z_2, such that $z_i > 1$. Thus, $f(z) = -\lambda^e\lambda^r(\lambda^- + \alpha)(z - z_1)(z - z_2)$, with $z_1 z_2 = \frac{\alpha\mu(\lambda^e+\mu)}{\lambda^e\lambda^r(\lambda^-+\alpha)}$, $z_1 + z_2 = \frac{\alpha\mu\lambda^r+\lambda^e(\lambda^-+\alpha)(\lambda^r+\mu)}{\lambda^e\lambda^r(\lambda^-+\alpha)}$. Then, the expression of $P_0(z)$ can be rewritten as

$$P_0(z) = \frac{\alpha p_{0,0}}{\lambda^- + \alpha}[1 - \frac{\mu\lambda^-}{\lambda^e\lambda^r(\lambda^- + \alpha)} \sum_{j=0}^{\infty} \left(\frac{A z_2^{j+1} + B z_1^{j+1}}{(z_1 z_2)^{j+1}} \right) z^j],$$

where

$$A = \frac{\mu + \lambda^e - \lambda^r z_1}{z_1 - z_2}, \; B = -\lambda^r - A.$$

Tedious but simple calculation leads to the expression in (11). Similar work can be done by using the expressions of $P_{1,k}(z)$, $k = 1, 2$ and $P_0(z)$ in order to obtain the (12), (13), so further details are omitted.

Appendix F: Proof of Theorem 6

We start by obtaining the expressions for $\tau_n(i, j, k)$, $(i, j, k) \in S$, $n \geq 0$. Remind that τ_n, $n \geq 0$ is given in (15). Since,

$$\tau_n(0, j, 1) = \frac{1}{\lambda^- + \alpha(1 - \delta_{0,n})}\delta_{j,n}, \; j, n \geq 0,$$

it is readily seen that $q_{0,j}^{(1)} = \frac{1}{U}\frac{\pi_j}{\lambda^- + \alpha(1-\delta_{0,j})}$, $j \geq 0$, and forming the pgf we obtain after some algebra the expression of $Q_{0,1}(z)$ in (17). Similarly,

$$\tau_n(1, j, 2) = \frac{\lambda^-}{\lambda^- + \alpha(1 - \delta_{0,n})}\delta_{j,n} \int_{t=0}^{\infty} e^{-\lambda^e t}(1 - B(t))dt, \; j \geq 0,$$

$$\tau_n(1,j,3) = \frac{\alpha\delta_{j+1,n}}{\lambda^- + \alpha\delta_{j+1,n}} \int_{t=0}^{\infty} e^{-\lambda^r t}(1-B(t))dt,\ j \geq 0,$$

$$\tau_n(1,j,4) = \frac{\lambda^-\delta_{j-1,n}}{\lambda^- + \alpha(1-\delta_{0,n})} \int_{t=0}^{\infty}(1-B(t)) \int_{u=0}^{t} \lambda^e e^{-\lambda^e u} e^{-\lambda_+^e(t-u)} dudt,\ j \geq 1,$$

$$\tau_n(1,j,5) = \frac{\lambda^-}{\lambda^- + \alpha(1-\delta_{0,n})} \int_{t=0}^{\infty}(1-B(t)) \times \int_{u=0}^{t} \lambda^e e^{-\lambda^e u} e^{-\lambda_+^e(t-u)} \frac{(\lambda_+^e(t-u))^{j-n-1}}{(j-n-1)!} dudt,\ j \geq 2, 0 \leq n \leq j-2,$$

$$\tau_n(1,j,6) = \frac{\alpha\delta_{j,n}}{\lambda^- + \alpha\delta_{j,n}} \int_{t=0}^{\infty}(1-B(t)) \int_{u=0}^{t} \lambda^r e^{-\lambda^r u} e^{-\lambda_+^r(t-u)} dudt,\ j \geq 1,$$

$$\tau_n(1,j,7) = \frac{\alpha}{\lambda^- + \alpha} \int_{t=0}^{\infty}(1-B(t)) \times \int_{u=0}^{t} \lambda^r e^{-\lambda^r u} e^{-\lambda_+^r(t-u)} \frac{(\lambda_+^r(t-u))^{j-n}}{(j-n)!} dudt,\ j \geq 2, 1 \leq n \leq j-1. \quad \text{(F28)}$$

To have a clearer view on how we obtain the above expressions, let us briefly explain $\tau_n(1,j,3)$. So, we need to have $j+1$ jobs in orbit upon a service completion (i.e., $n = j+1$), and an orbiting job joins the server. Then, its service time begins at time, say $t = 0$, and the time interval $(t, t+dt)$ contributes to $\tau_n(1,j,3)$ if (i) service time has not been completed before time t (with probability $1 - B(t)$), and (ii) No primary jobs arrive during $(0,t]$ (with probability $e^{-\lambda^r t}$). The other expressions are derived in a similar reasoning.

Now, by using (F28), simple computations result in the expressions given in (16). After routine algebraic calculations of (16), we can derive the rest partial generating functions $Q_{1,k}(z)$, $k = 2,\ldots,7$, as given in (17).

Appendix G: Proof of Theorem 8

Using the state probabilities and usual arguments lead to the differential difference equations:

$$\begin{aligned}
(\lambda^- + j\alpha + \beta(1-\delta_{0,j}))p_{0,j} &= \textstyle\sum_{k=2}^{3} p_{1,j}^{(k)}(0) + \sum_{k=4,6} p_{1,j}^{(k)}(0)1_{\{j\geq 1\}} \\
&\quad + \textstyle\sum_{k=5,7} p_{1,j}^{(k)}(0)1_{\{j\geq 2\}}, \\
-\tfrac{d}{dr}p_{1,j}^{(2)}(r) &= -\lambda^e p_{1,j}^{(2)}(r) + \lambda^- p_{0,j} b(r),\ j \geq 0, \\
-\tfrac{d}{dr}p_{1,j}^{(3)}(r) &= -\lambda^r p_{1,j}^{(3)}(r) + p_{0,j+1}((j+1)\alpha+\beta)b(r),\ j \geq 0, \\
-\tfrac{d}{dr}p_{1,j}^{(4)}(r) &= -\lambda_+^e p_{1,j}^{(4)}(r) + \lambda^e p_{1,j-1}^{(2)}(r),\ j \geq 1, \\
-\tfrac{d}{dr}p_{1,j}^{(5)}(r) &= -\lambda_+^e p_{1,j}^{(5)}(r) + \lambda_+^e(p_{1,j-1}^{(5)}(r) + p_{1,j-1}^{(4)}(r)),\ j \geq 2, \\
-\tfrac{d}{dr}p_{1,j}^{(6)}(r) &= -\lambda_+^r p_{1,j}^{(6)}(r) + \lambda^r p_{1,j-1}^{(3)}(r),\ j \geq 1, \\
-\tfrac{d}{dr}p_{1,j}^{(7)}(r) &= -\lambda_+^r p_{1,j}^{(7)}(r) + \lambda_+^r(p_{1,j-1}^{(7)}(r) + p_{1,j-1}^{(6)}(r)),\ j \geq 2.
\end{aligned} \quad \text{(G29)}$$

By taking Laplace transforms in (G29), and applying the generating function approach (following similar arguments and notation as in the proof of Theorem 3) we obtain after some algebra

$$\alpha z P_0'(z) + P_0(z)(\lambda^- + \beta) = \beta p_{0,0} + \textstyle\sum_{k=2}^{7} P_{1,k}(0,z), \quad \text{(G30)}$$

and equation (23). Moreover,

$$\begin{aligned} P_{1,2}(0,z) &= \lambda^{-}\beta^{*}(\lambda^{e})P_0(z), \\ P_{1,3}(0,z) &= \frac{\beta^{*}(\lambda^{r})}{z}[\alpha z P_0'(z) + \beta(P_0(z) - p_{0,0})], \\ \sum_{k=4}^{5} P_{1,k}(0,z) &= \frac{\lambda^{-}\lambda^{e} z[\beta^{*}(\lambda^{e}) - \beta^{*}(\lambda_{+}^{e}(1-z))]}{\lambda_{+}^{e}(1-z) - \lambda^{e}} P_0(z), \\ \sum_{k=6}^{7} P_{1,k}(0,z) &= \frac{\lambda^{r}(\alpha z P_0'(z) + \beta(P_0(z) - p_{0,0}))}{\lambda_{+}^{r}(1-z) - \lambda^{r}} (\beta^{*}(\lambda^{r}) - \beta^{*}(\lambda_{+}^{r}(1-z))). \end{aligned} \tag{G31}$$

Then, simple computations yield

$$\sum_{k=2}^{7} P_{1,k}(0,z) = \alpha A_r(z) P_0'(z) + P_0(z)(\lambda^{-} A_e(z) + \frac{\beta}{z} A_r(z)) - \beta p_{0,0} \frac{A_r(z)}{z}.$$

Substituting back in (G30) yields

$$\alpha z P_0'(z) + P_0(z)(\beta + \frac{\lambda^{-} z(1 - A_e(z))}{z - A_r(z)}) = \beta p_{0,0}. \tag{G32}$$

Setting $z = 1$ in (G32) yields

$$\alpha P_0'(1) + \beta(P_0(1) - p_{0,0}) = \lambda^{-} \frac{A_e^{(1)}(1)}{1 - A_r^{(1)}(1)} P_0(1). \tag{G33}$$

Using (G33) and setting $s \to 0, z \to 1$ in (23) we obtain after simple calculations Eq. (22).

The solution of the (G32) is

$$\begin{aligned} P_0(z) = & \, exp\{\tfrac{1}{\alpha} \int_z^1 u^{-1}(\beta + \tfrac{\lambda^{-} u(1 - A_e(u))}{u - A_r(u)}) du\} \\ & \times \left(P_0(1) - \tfrac{\beta p_{0,0}}{\alpha} \int_z^1 x^{-1} exp\{\tfrac{1}{\alpha} \int_x^1 u^{-1}(\beta + \tfrac{\lambda^{-} u(1 - A_e(u))}{u - A_r(u)}) du\} dx \right). \end{aligned} \tag{G34}$$

Setting $z = 0$ in (G34), and using (22) we obtain after manipulations

$$p_{0,0} = \delta_{0,\beta} H(0) + (1 - \delta_{0,\beta}) \frac{\alpha}{\beta} \left(\int_0^1 x^{\beta/\alpha - 1} H^{-1}(x) dx \right)^{-1}.$$

Substituting $p_{0,0}$ and (22) into (G34) leads after some algebra to the expression (20).

References

Armony, M., & Maglaras, C. (2004). Contact centers with a call-back option and real-time delay information. *Operations Research, 52*(4), 527–545. https://doi.org/10.1287/opre.1040.0123

Armony, M., & Maglaras, C. (2004). On customer contact centers with a call-back option: Customer decisions, routing rules, and system design. *Operations Research, 52*(2), 271–292. https://doi.org/10.1287/opre.1030.0088

Dudin, A. N., Krishnamoorthy, A., Joshua, V., & Tsarenkov, G. V. (2004). Analysis of the BMAP/G/1 retrial system with search of customers from the orbit. *European Journal of Operational Research, 157*(1), 169–179. https://doi.org/10.1016/S0377-2217(03)00245-5

Phung-Duc, T., & Kawanishi, K. (2014). Performance analysis of call centers with abandonment, retrial and after-call work. *Performance Evaluation, 80*, 43–62. https://doi.org/10.1016/j.peva.2014.03.001

Gencer, B., Karaesmen, Z., Gunes, E., & Pala, O. (2014). The impact of queue features on customers' queue joining and reneging behavior: A laboratory experiment. In *2014 Proceedings of 20th Conference of the International Federation of Operational Research Societies, IFORS* , Barcelona, Spain .

Falin, G., & Templeton, J. G. (1997). *Retrial queues*. Chapman & Hall.

Artalejo, J., Gómez-Corral, A. (2008). Retrial queueing systems. A computational approach. Springer. https://doi.org/10.1007/978-3-540-78725-9.

Phung-Duc, T. (2017). Retrial queueing models: a survey on theory and applications. In *Stochastic operations research in business and industry* (pp. 1–31). World Scientific.

Langaris, C., & Dimitriou, I. (2010). A queueing system with n-phases of service and (n-1)-types of retrial customers. *European Journal of Operational Research, 205*(3), 638–649. https://doi.org/10.1016/j.ejor.2010.01.034

Farahmand, K. (1990). Single line queue with repeated demands. *Queueing Systems, 6*(1), 223–228. https://doi.org/10.1007/BF02411475

Fayolle, G. (1986). 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.

Dimitriou, I. (2018). A two-class queueing system with constant retrial policy and general class dependent service times. *European Journal of Operational Research, 270*(3), 1063–1073. https://doi.org/10.1016/j.ejor.2018.03.002

Artalejo, J. R., & Gomez-Corral, A. (1997). Steady state solution of a single-server queue with linear repeated requests. *Journal of Applied Probability, 34*(1), 223–233. https://doi.org/10.2307/3215189

Gómez-Corral, A. (1999). Stochastic analysis of a single server retrial queue with general retrial times. *Naval Research Logistics (NRL), 46*(5), 561–581. https://doi.org/10.1002/(SICI)1520-6750(199908)46:5<561::AID-NAV7>3.0.CO;2-G

Choi, B. D., Park, K. K., & Pearce, C. E. M. (1993). An M/M/1 retrial queue with control policy and general retrial times. *Queueing Systems, 14*(3–4), 275–292. https://doi.org/10.1007/BF01158869

Cinlar, E. (1975). Introduction to stochastic processes Prentice-Hall. Englewood Cliffs, New Jersey (420p).

Baron, O., Economou, A., & Manou, A. (2018). The state-dependent M/G/1 queue with orbit. *Queueing Systems, 90*(1), 89–123. https://doi.org/10.1007/s11134-018-9582-1

Baron, O., Economou, A., & Manou, A. (2022). Increasing social welfare with delays: Strategic customers in the M/G/1 orbit queue. *Production and Operations Management, 31*(7), 2907–2924. https://doi.org/10.1111/poms.13728

Legros, B. (2018). M/G/1 queue with event-dependent arrival rates. *Queueing Systems, 89*(3), 269–301. https://doi.org/10.1007/s11134-017-9557-7

Legros, B., & Sezer, A. D. (2018). Stationary analysis of a single queue with remaining service time-dependent arrivals. *Queueing Systems, 88*(1), 139–165. https://doi.org/10.1007/s11134-017-9552-z

Dimitriou, I. (2022). The M/G/1 retrial queue with event-dependent arrivals. arXiv preprint arXiv:2203.02757.

Legros, B. (2021). Dimensioning a queue with state-dependent arrival rates. *Computers and Operations Research, 128*, 105179. https://doi.org/10.1016/j.cor.2020.105179

Legros, B. (2022). The principal-agent problem for service rate event-dependency. *European Journal of Operational Research, 297*(3), 949–963. https://doi.org/10.1016/j.ejor.2021.09.020

Bekker, R., Borst, S., Boxma, O. J., & Kella, O. (2004). Queues with workload-dependent arrival and service rates. *Queueing Systems, 46*, 537–556. https://doi.org/10.1023/B:QUES.0000027998.95375.ee

Boxma, O., Kaspi, H., Kella, O., & Perry, D. (2005). On/off storage systems with state-dependent input, output, and switching rates. *Probability in the Engineering and Informational Sciences, 19*(1), 1–14. https://doi.org/10.1017/S0269964805050011

Kerner, Y. (2008). The conditional distribution of the residual service time in the Mn/G/1 queue. *Stochastic Models, 24*(3), 364–375. https://doi.org/10.1080/15326340802232210

Boxma, O. J., & Vlasiou, M. (2007). On queues with service and interarrival times depending on waiting times. *Queueing Systems, 56*(3), 121–132. https://doi.org/10.1007/s11134-007-9011-3

D'Auria, B., Adan, I. J. B. F., Bekker, R., & Kulkarni, V. (2022). An M/M/c queue with queueing-time dependent service rates. *European Journal of Operational Research, 299*(2), 566–579. https://doi.org/10.1016/j.ejor.2021.12.023

Artalejo, J., & Falin, J. (1994). Stochastic decomposition for retrial queues. *Top, 2*(2), 329–342. https://doi.org/10.1007/BF02574813

Sennott, L. I., Humblet, P. A., & Tweedie, R. L. (1983). Technical noteúmean drifts and the non-ergodicity of Markov chains. *Operations Research, 31*(4), 783–789. https://doi.org/10.1287/opre.31.4.783

Chen, Z., Pappas, N., Kountouris, M., & Angelakis, V. (2018). Throughput with delay constraints in a shared access network with priorities. *IEEE Transactions on Wireless Communications, 17*(9), 5885–5899. https://doi.org/10.1109/TWC.2018.2851213

Pappas, N., Chen, Z., & Dimitriou, I. (2018). Throughput and delay analysis of wireless caching helper systems with random availability. *IEEE Access, 6*, 9667–9678. https://doi.org/10.1109/ACCESS.2018.2801246

Ploumidis, M., Pappas, N., & Traganitis, A. (2017). Flow allocation for maximum throughput and bounded delay on multiple disjoint paths for random access wireless multihop networks. *IEEE Transactions on Vehicular Technology, 66*(1), 720–733. https://doi.org/10.1109/TVT.2016.2547181

Mehmeti, F., Papa, A., Kellerer, W. (2023). Maximizing network throughput using SD-RAN. In *IEEE Consumer communications and networking conference (IEEE CCNC 2023)At: Las Vegas, NV, USA* (pp. 1–7).

Koole, G. (2007). *Monotonicity in Markov reward and decision chains: Theory and applications* (Vol. 1). Now Publishers Inc. https://doi.org/10.1561/0900000002

Horváth, A., & Telek, M. (2002). Phfit: A general phase-type fitting tool. In T. Field, P. G. Harrison, J. Bradley, & U. Harder (Eds.), *Computer performance evaluation: Modelling techniques and tools* (pp. 82–91). Springer. https://doi.org/10.1007/3-540-46029-2_5

Altman, E. (1999). *Constrained Markov decision processes: Stochastic modeling*. Routledge.

Puterman, M. L. (1994). *Markov decision processes: Discrete stochastic dynamic programming*. Wiley Series in Probability and Statistics. Wiley. https://doi.org/10.1002/9780470316887.

Howard, R. A. (1960). *Dynamic programming and Markov processes*. John Wiley.