OR Spectrum
DOI 10.1007/s00291-017-0487-x

REGULAR ARTICLE

Call centers with a postponed callback offer

Benjamin Legros[1] · **Sihan Ding**[2] ·
Rob van der Mei[2] · **Oualid Jouini**[3]

Received: 29 June 2016 / Accepted: 16 August 2017

Abstract We study a call center model with a postponed callback option. A customer at the head of the queue whose elapsed waiting time achieves a given threshold receives a voice message mentioning the option to be called back later. This callback option differs from the traditional ones found in the literature where the callback offer is given at customer's arrival. We approximate this system by a two-dimensional Markov chain, with one dimension being a unit of a discretization of the waiting time. We next show that this approximation model converges to the exact one. This allows us to obtain explicitly the performance measures without abandonment and to compute them numerically otherwise. From the performance analysis, we derive a series of practical insights and recommendations for a clever use of the callback offer. In particular, we show that this time-based offer outperforms traditional ones when considering the waiting time of inbound calls.

✉ Sihan Ding
dingsihan@hotmail.com

Benjamin Legros
benjamin.legros@centraliens.net

Rob van der Mei
R.D.van.der.Mei@cwi.nl

Oualid Jouini
oualid.jouini@centralesupelec.fr

[1] Laboratoire Métis, EM Normandie, 64 Rue du Ranelagh, 75016 Paris, France

[2] Stochastics Group, Center for Mathematics and Computer Science (CWI),
Science Park 123, 1098 XG Amsterdam, The Netherlands

[3] Laboratoire Genie Industriel, CentraleSupélec, Université Paris-Saclay,
Grande Voie des Vignes, 92290 Chatenay-Malabry, France

Published online: 31 August 2017

Keywords Queueing systems · Markov chains · Performance evaluation · Waiting time · Callback

1 Introduction

Call centers serve as the public face in various areas and industries: insurance companies, emergency centers, banks, information centers, help desks, telemarketing, just to name a few. The success of call centers is due to the technological advances in information and communications systems. The most used form of communication is the direct telephone contact. However, in the context of highly congested call centers, the use of alternative options can be proposed to customers so as to better match demand and capacity. Alternative options could be email, chat, blog, callback service, etc.

The callback offer allows the call center to change the nature of the channel from an inbound call to an outbound one. For the call center manager, this change is valuable because it reduces the congestion in the inbound queue. Another important aspect in call centers is customers' abandonment (e.g., see Mandelbaum and Zeltyn 2004; Dai and He 2012). While waiting in the inbound queue, a customer may decide to leave the system without being served. This customer is then lost for the call center without possibilities to be recontacted. Instead, an outbound customer can be reached later. Even with a long delay before being called back, this customer is potentially not lost. From customers' perspective, the willingness to accept future processing depends on the urge to get an answer and the waiting cost. If waiting is painful and getting an answer is not urgent, then a customer may accept the callback offer.

In practice, several types of callback offers are developed with the same purpose of changing inbound calls into outbound ones. A large number of patents reflect this wide variety and the technological challenges to implement this option in the Automatic Call Distributor (ACD) (Livanos 1994; Metcalf 2006; Rafter et al. 2010; Blaesi 2015). Nevertheless, from our discussion with our partner INTERACTIV GROUP, the effects of the callback option are not well understood by managers and the implementation still needs to be improved to achieve some service level objectives.

In call centers, a percentile of the waiting time is the usually chosen as a service level objective. This metric is often preferred to the average speed of answer because the former was perceived to be more informative; see Bailey and Sweeney (2003). It is therefore important for managers to develop a callback offer which can be adjusted to this type of service level agreement. At the same time, the callback offer should be carefully used. Even when the callback offer is accepted by a customer, most customers would prefer being served directly. So, the callback offer should not be automatically proposed, but should be proposed in a way which allows the call center to control the proportion of outbound calls. As mentioned above, the other aspect is abandonment. In case of a too important use of the callback offer, the proportion of non-abandoning customers may get too important which in turn may lead to the impossibility to ensure a sufficiently short delay for callback customers. In summary, an efficient callback offer should:

- Help the manager to achieve a service level objective for inbound calls;
- Control the proportion of outbound calls;

- Be easy to implement in the ACD;
- Be sufficiently simple to develop staffing solutions and predict performance.

In the literature on operations research, different callback options have already been studied and optimized (Armony and Maglaras 2004a, b; Kim et al. 2012; Dudin et al. 2013; Legros et al. 2016). These callback models will be discussed in detail below. A common element in these models is that the decision to propose a callback offer is based on the system size. For instance, above a threshold on the queue length, a callback option is proposed to all arriving customers. Unlike these models, we propose a new callback option given to the first customer in line when its experienced waiting time reaches a given waiting time threshold, the service level objective. We call this callback option the *postponed* call back offer.

This makes sense both from theoretical and practical points of view, especially for objectives that are functions of the waiting time such as the percentage of calls that have waited shorter than a specific threshold. One can imagine, and it is indeed shown in this paper, that a policy that uses actual waiting time information performs well for this type of objective.

The motivation to let customers wait before the callback offer in our model is to avoid giving a callback offer to a customer who could have been served in a reasonable time. If a callback offer is given at arrival based eventually on the queue size, it may be possible due to the variability in the service times to encounter a series of small service times which would have enable to serve this customer in a reasonable time. By letting the customer wait before the callback offer, the call center gives a chance to serve the customer without using the callback option. Recall that most customers prefer being directly served than being called back later.

In addition, we assume that customers have a probabilistic reaction to the callback offer and that a non-preemptive priority is given to inbound calls since these ones are more urgent. A precise definition of the queueing model is given in Sect. 2. Another value of this callback model is that it is completely tractable. Without abandonment, closed-form expressions of the performance measures can be obtained. This allows for workforce management solutions and a simple implementation of the callback offer.

In Sect. 3, we determine the proportion of customers who have waited less than the waiting time objective and the proportion of callback customers. In order to differentiate between inbound and outbound customers, we are also interested in their respected expected waiting times. Closed-form expressions of these performance measures are derived without abandonment, and a numerical method is developed with abandonment. The difficulty to compute these metrics is that the decision to change a high priority customer into a low priority one does not depend on a classical state definition like the number of high priority customers, but on the experienced waiting time of a given customer. To overcome this difficulty, we propose the following approach:

1. We develop an approximating model, in which the waiting time of the first customer in line is modeled by a succession of exponential phases. The number of waiting phases and the elapsing of time rate per phase are the control parameters of the approximation.
2. Since this new model is a Markov chain, the transitions rate can be obtained and the stationary probabilities can be derived.

3. Finally, as the control parameters of the approximating model tend to infinity, we show that this model converges to the exact one which in turn leads to the exact performance measures.

The key operational findings derived in Sect. 4 are that (1) the callback offer can be used as a tool to reduce a waiting time percentile, (2) the value of a callback option is more apparent under intermediate loaded situations, with abandonment, for small call center, or when customers react mostly positively to the callback option, (3) two rational strategies are possible for customers; either they all accept or they all reject the callback offer, (4) the time at which the callback offer is proposed should be sufficiently postponed, especially when the abandonment is significant or when customers do not have a rational reaction to the callback offer, and (5) compared to a non-postponed callback option, a postponed offer improves the waiting time of inbound calls and the proportion of abandonment, especially in highly loaded situations.

In what follows, we discuss the related literature.

Literature review There is an extensive and growing literature on call centers. We refer the reader to Gans et al. (2003) and Akşin et al. (2007) for an overview. The main topics encountered in call center studies are routing decisions (e.g., see Helber and Henken 2010; Robbins and Harrison 2010; Legros 2016), staffing (e.g., see Cezik and L'Ecuyer 2008; Liao et al. 2012), or performance evaluation (e.g., see Koole and Mandelbaum 2002; Stolletz and Helber 2004; Shumsky 2004). Our article focuses on performance evaluation based on a particular routing mechanism defined through a callback offer.

There are a few papers on different callback options in call centers. Armony and Maglaras (2004a) consider a model in which customers are given a choice of whether to wait online for their call to be answered or to leave a number and be called back within a specified time or to immediately balk. Upon arrival, customers are informed (or know from prior experience) of the expected waiting time if they choose to wait and the delay guarantee for the callback option. Their decision is probabilistic and based on this information. Under the heavy traffic regime, Armony and Maglaras (2004a) develop an estimation scheme for the anticipated real-time delay that is asymptotically correct. They also propose an asymptotically optimal routing policy that minimizes real-time delay subject to a deadline on the postponed service mode. Armony and Maglaras (2004b) develop an asymptotically optimal routing rule, characterize the unique equilibrium regime of the system, and propose a staffing rule that picks the minimum number of agents that satisfies a set of operational constraints on the performance of the system.

There are two recent papers by Kim et al. (2012) and Dudin et al. (2013). Kim et al. (2012) consider a call center model with a callback option where the capacity of the queue for the inbound calls is finite. Customer balking and abandonment are allowed. They provide an efficient algorithm for calculating the stationary probabilities of the system. Moreover, they derive the Laplace–Stieltjes transform of the sojourn time distribution of virtual customers. Dudin et al. (2013) consider a slightly different model, where agents make outbound calls to those lost customers. There are two agent teams: one that handles in priority inbound calls and another that handles in priority outbound calls. They compute the stationary probabilities and deduce from that some performance measures. They also numerically address the staffing issue of the two teams.

Finally, Legros et al. (2016) consider in their callback model, a probabilistic customer reaction to the callback offer. They show using a Markov decision process approach that the optimal reservation policy for inbound calls is of switch type. Thereafter, the system performance measures are computed under the optimal policy. It appears from this study that the value of the callback offer is apparent for congested situations and that the benefits of a reservation policy are more apparent in large call centers, while they almost disappear in the extreme situations of light or heavy workloads. Moreover, if balking and abandonment are very high or if the overall treatment time spent to serve an outbound call is very large compared to that of an inbound one, there is a value in delaying the proposition of the callback offer.

Another stream of literature less closely related to our article deals with the analysis of queueing multi-channel call center models with blending. This can be related to callback models by assuming an infinite amount of customers to callback at the next working period. Some papers focus on performance evaluation, and others address the analysis of blending policies or staffing decisions. Deslauriers et al. (2007) develop various continuous Markov chain models for a call center with inbound and outbound calls. The authors consider a threshold policy and characterize the rate of outbounds and the waiting time distribution of inbounds. Other call center papers address the analysis of blending policies. Gans and Zhou (2003) and Bhulai and Koole (2003) prove that a threshold policy on the number of idle agents is optimal to maximize the outbound throughput under a service level constraint on the inbound waiting time. Similar results are also found in Legros et al. (2015), for a non-stationary model where inbound calls arrive according to a non-homogeneous Poisson process. Pang and Perry (2014) consider a large call blending model and propose a logarithmic safety staffing rule, combined with a threshold control policy to ensure that agents' utilization is always close to one with always idle agents present.

2 Setting

In this section, we define the queueing model and present an approximation model which can be studied through a Markov chain analysis.

2.1 Queueing model

We consider a multi-server single queue with s identical, parallel servers. The arrival process of customers is Poisson with rate λ. Service times are independent and exponentially distributed with rate μ. When a customer calls, if at least one agent is available, then this customer is directly served; otherwise, he/she is routed to a first-come first-served queue called Queue 1. After having waited K time units, the first customer in line waiting in Queue 1 hears a voice message, proposing to be called back later. We assume that a proportion r of customers accepts the callback offer and becomes then outbound calls. These calls are routed to another queue called Queue

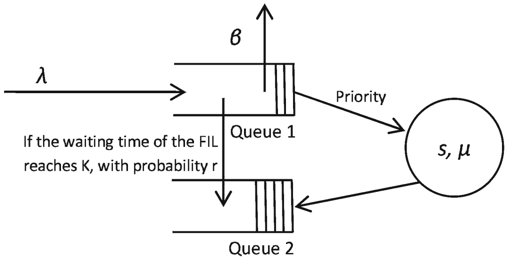


Fig. 1 Queueing model

2. Since inbound calls are more urgent, a non-preemptive priority is given to Queue 1. Another reason for the priority of inbound calls is the cost of waiting. In many call centers, inbound customers pay per waiting time unit, whereas an outbound customer would not pay. A priority for inbound calls would then help to reduce their waiting cost.

Moreover, customers' patience is limited. We assume that the patience of a customer in Queue 1 is exponentially distributed with rate β. Customers in Queue 2 are infinitely patient since they are outbound calls. Our queueing model is equivalent to a particular V-queueing model with two queues: Queue 1 and Queue 2, where customers in Queue 1 have a non-preemptive priority over customers in Queue 2. The arrival process in Queue 1 is Poisson with parameter λ, and the arrival process in Queue 2 is generated by customers in Queue 1 who have waited exactly K time units without being served and accept the callback offer. This equivalent queueing model is depicted in Fig. 1. For this queueing model, we are interested in the proportion of callback customers, P_c, the proportion of abandonment, P_a, the expected waiting time of customers served from Queue 1, $E(W_1)$, the expected waiting time of callback customers, $E(W_2)$ (it includes the time also spent in Queue 1), and the probability of waiting less than the instant at which the callback option is proposed, $P(W < K)$, where W is the waiting time of an arbitrary customer. Note that without abandonment, this queueing model can be seen as an M/M/s queue where the queue discipline has been modified.

2.2 An approximating model

In order to have a Markov chain, one may only have exponential durations between two successive events. Yet, the time at which the callback offer is given is deterministic. To overcome this difficulty, we develop here an approximating model in which all durations are exponential. The resulting Markov chain will be studied in Sect. 3 to obtain the performance measures of the exact model.

The approximation is based on a Markov chain where the states constitute a discrete representation of the waiting time of the first customer in line (FIL) in Queue 1 when one or more customers are waiting. The waiting time of the FIL in Queue 1 is modeled by a succession of exponential phases with rate γ per phase as proposed in Koole et al. (2012). Instead, Queue 2 is modeled as in most queueing models by its number of customers. The number of waiting phases in Queue 1 after which the callback offer is proposed to the FIL is denoted by n. After leaving this waiting phase, a customer—if not served—is routed to Queue 2 with probability r or stays in Queue 1 with probability $1-r$. The queue discipline in both queues is still FCFS.

After giving a state definition and the transition rates, we will explain how this approximation converges to the real model.

State definition The system is modeled using a two-dimensional continuous-time Markov chain. We denote by (x, y) a state of the system for $x \geq -s$ and $y \geq 0$, where x represents the servers state or the waiting time in Queue 1 and y represents the number of customers in Queue 2. More precisely, states with $-s \leq x \leq 0$ correspond to an empty Queue 1 and $s+x$ busy agents. States with $x > 0$ correspond to the phase at which the FIL in Queue 1 is waiting and all agents are busy.

Transitions We next describe the seven possible transitions in the Markov chain. When the FIL changes, because of a service completion or an abandonment (see transition Type 5), or because of the current FIL moving to Queue 2 (see transition Type 8), the waiting time phase changes from $x > 0$ to $x-h$ with probability $q_{x,x-h}$. This means that either the new first in line is in waiting phase $x-h > 0$ or that Queue 1 is empty if $x-h = 0$, for $0 \leq h < x$. The probabilities $q_{x,x-h}$ are given in Theorem 2 of Legros et al. (2017) by

$$q_{x,x-h} = \left(1 - \left[1 + \frac{\lambda}{\gamma}\left(\frac{\gamma}{\gamma+\beta}\right)^{x-h}\right]^{-1}\right) \cdot \prod_{k=x-h+1}^{x} \left[1 + \frac{\lambda}{\gamma}\left(\frac{\gamma}{\gamma+\beta}\right)^{k}\right]^{-1}$$

for $0 \leq h < x$ and

$$q_{x,0} = \prod_{k=1}^{x} \left[1 + \frac{\lambda}{\gamma}\left(\frac{\gamma}{\gamma+\beta}\right)^{k}\right]^{-1}.$$

Moreover, the probability of abandonment after a given waiting phase is $\frac{\beta}{\gamma+\beta}$ (see Table 1, Line 3 in Legros et al. 2017)

1. An arrival with rate λ while Queue 1 is empty ($-s \leq x \leq 0, y = 0$), which changes the state to $(x+1, 0)$. If $x < 0$, then the number of busy servers is increased by 1. Otherwise, if $x = 0$, then the FIL entity is created.
2. A service completion with rate $(s+x)\mu$ while Queues 1 and 2 are empty ($-s < x \leq 0, y = 0$), which changes the state to $(x-1, y)$. The number of busy servers is reduced by 1.
3. A service completion with rate $s\mu$ while Queue 1 is empty, Queue 2 is not empty and all servers are busy ($x = 0, y \geq 1$), which changes the state to $(0, y-1)$. The number of customers in Queue 2 is reduced by 1.
4. A service completion with rate $s\mu q_{x,x-h}$ or an abandonment with rate $\gamma \frac{\beta}{\gamma+\beta}$ while Queue 1 is not empty ($x > 0, y \geq 0$), which changes the state to $(x-h, y)$, that is, the new FIL is in waiting phase $x-h$.

5. A phase increase without abandonment with rate $\gamma \frac{\gamma}{\gamma+\beta}$ while Queue 1 is not empty and the FIL is not in waiting phase n $(0 < x < n, y \geq 0)$, which changes the state to $(x + 1, y)$. The waiting phase of the FIL is increased by 1.
6. A phase increase with rate $(1 - r)\gamma$ while the FIL is in waiting phase n $(y \geq 0)$, which changes the state to $(n + 1, y)$. The waiting phase of the FIL is increased by 1.
7. A phase increase with rate $r\gamma q_{x,x-h}$ while the FIL in Queue 1 is in waiting phase n $(x = n, y \geq 0)$, which changes the state to $(x - h, y + 1)$, that is, the new FIL is in waiting phase $x - h$ and the number of customers in Queue 2 is increased by 1.

Convergence to the real system We approximate the deterministic duration before giving the callback offer by an Erlang random variable with n phases and rate γ per phase. We choose n and γ such that $\frac{n}{\gamma} \overset{\Delta}{=} K$. The Laplace transform of the Erlang distribution with parameters n and γ is $\left(\frac{\gamma}{\gamma+s}\right)^n$. We have

$$\left(\frac{\gamma}{\gamma + s}\right)^n = e^{n \ln((1+s/\gamma)^{-1})} \underset{\gamma\to\infty}{\sim} e^{n \ln(1-s/\gamma)} \underset{\gamma\to\infty}{\sim} e^{-ns/\gamma} = e^{-sK},$$

where we write $f(a) \underset{a\to a_0}{\sim} g(a)$ to express that $\lim_{a\to a_0} \frac{f(a)}{g(a)} = 1$, for $a_0 \in \mathbb{R}$. Applying the Levy continuity theorem for Laplace transforms, this result ensures that as n and γ go to infinity, the considered Erlang random variable converges in distribution to the deterministic duration K.

The other approximation is the transition from Queue 1 to Queue 2. It is assumed in our modeling that after one γ-transition from state $x = n$, only one customer is routed to Queue 2. However, more than one customer could be in phase n (as in any other phase). More precisely (with no abandonment), given that one customer is in phase n, this customer is the only one with probability $\frac{\gamma}{\lambda+\gamma}$, or two customers or more are in phase n with probability $\frac{\lambda}{\lambda+\gamma}$. Again, as γ tends to infinity, the probability that only one customer is in one phase is equal to one.

3 Performance analysis

In Sect. 3.1, we derive explicitly the performance measures without abandonment. The method developed here is adapted numerically in Sect. 3.1.2 to include abandonment.

3.1 Explicit performance measures without abandonment

In Sect. 3.1, we give the stationary probabilities of the discretized system. Next, in Sect. 3.1.2, we let the elapsing of time rate tends to infinity in order to obtain the exact performance measures.

3.1.1 Stationary probabilities

Recall that in the case with no abandonment ($\beta = 0$), we simply have

$$q_{x,x-h} = \left(\frac{\lambda}{\lambda+\gamma}\right)\left(\frac{\gamma}{\lambda+\gamma}\right)^h$$

for $0 \leq h < x$ and

$$q_{x,0} = \left(\frac{\gamma}{\lambda+\gamma}\right)^x$$

as in Theorem 2.1 of Koole et al. (2012). Let us introduce the notations $a = \frac{\lambda}{\mu}$ and $a_\gamma = s \cdot \frac{a+\gamma/\mu}{s+\gamma/\mu}$. The ratio a represents the traffic intensity of the system and a_γ is a modified version of the traffic intensity. The parameter a_γ is an increasing function of γ which is equal to a for $\gamma = 0$ and equal to s for $\gamma = \infty$. Proposition 1 gives the stationary probability $p_{x,y}$ to be in state (x, y) for $x \geq -s$ and $y \geq 0$.

Proposition 1 *Under the stability condition $\lambda < s\mu$, we have*

$$p_{-s,0} = \left[\sum_{x=0}^{s-1}\frac{a^x}{x!} + \frac{a^s}{s!}\frac{\left(1 + \frac{a}{s}\frac{\lambda}{\gamma} - r\frac{a}{s}\left(1+\frac{\lambda}{\gamma}\right)\left(\frac{a_\gamma}{s}\right)^n\right)}{(1-a/s)\left(1 - r\frac{a}{s}\left(\frac{a_\gamma}{s}\right)^n\right)}\right]^{-1},$$

$$p_{x-s,0} = \frac{a^x}{x!} \cdot p_{-s,0},\ \textit{for}\ 0 \leq x \leq s,$$

$$p_{x,0} = p_{0,0}\frac{\lambda}{\gamma}\frac{\left(\frac{a_\gamma}{s}\right)^x(s\mu - \lambda(1-r)) - r\lambda\left(\frac{a_\gamma}{s}\right)^n}{s\mu - \lambda(1-r) - r\lambda\left(\frac{a_\gamma}{s}\right)^n},\ \textit{for}\ 1 \leq x \leq n,$$

$$p_{x,0} = p_{0,0}(1-r)\frac{\lambda}{\gamma}\frac{(s\mu-\lambda)\left(\frac{a_\gamma}{s}\right)^{x-n}}{s\mu - \lambda(1-r) - r\lambda\left(\frac{a_\gamma}{s}\right)^n},\ \textit{for}\ x > n,$$

$$p_{x,y} = \frac{\lambda}{\gamma}p_{0,0}\frac{\left(\frac{a_\gamma}{s}\right)^x(s\mu - \lambda(1-r)) - r\lambda\left(\frac{a_\gamma}{s}\right)^n}{s\mu - \lambda\left(\frac{a_\gamma}{s}\right)^n}\frac{s\mu - \lambda(1-r)\left(\frac{a_\gamma}{s}\right)^x - r\lambda\left(\frac{a_\gamma}{s}\right)^n}{s\mu - \lambda(1-r) - r\lambda\left(\frac{a_\gamma}{s}\right)^n}$$
$$\times\left(\frac{r\lambda}{s\mu}\frac{s\mu - \lambda\left(\frac{a_\gamma}{s}\right)^n}{s\mu - \lambda(1-r) - r\lambda\left(\frac{a_\gamma}{s}\right)^n}\right)^y,\ \textit{for}\ 1 \leq x \leq n, y \geq 1,$$

$$p_{x,y} = (1-r)\left(\frac{a_\gamma}{s}\right)^{x-n}p_{n,y},\ \textit{for}\ x > n, y \geq 1.$$

Proof We adopt the following approach to derive the stationary probabilities. First, we determine a set of equilibrium equations. Next, using these equilibrium equations we derive a simple explicit expression of the probability that the FIL in Queue 1 is in waiting phase x; $p_x = \sum_{y=0}^{\infty} p_{x,y}$ for $x \geq 0$. Considering this probability leads to a one-dimensional problem which in turn allows us to compute the probability of an empty system using the normalizing condition. Finally, we derive the other stationary probabilities.

Equilibrium equations Let S be the state space. Consider the cut between $A_1 = \{(-s,0),\ldots,(x,0)\}$ and $S\backslash A_1$, where $x \geq -s$. Observing that $\left(\frac{\gamma}{\lambda+\gamma}\right)^x + \sum_{l=h}^{x-1}\left(\frac{\lambda}{\lambda+\gamma}\right)\left(\frac{\gamma}{\lambda+\gamma}\right)^l = \left(\frac{\gamma}{\lambda+\gamma}\right)^h$, we deduce that the cumulative transition rate from state (x,y) to states $(0,y),(1,y)\cdots(x-h,y)$ is $s\mu\left(\frac{\gamma}{\lambda+\gamma}\right)^h$, for $0 \leq h < x < n$ and $y \geq 0$. Therefore, by equating flows across the cut, one may write

$$\lambda p_{x,0} = (s+x+1)\mu p_{x+1,0}, \text{ for } -s \leq x < 0, \tag{1}$$

$$\lambda p_{0,0} = s\mu p_{0,1} + s\mu\sum_{i=1}^{\infty} p_{i,0}\left(\frac{\gamma}{\lambda+\gamma}\right)^i, \tag{2}$$

$$\gamma p_{x,0} = s\mu p_{0,1} + s\mu\sum_{i=x+1}^{\infty} p_{i,0}\left(\frac{\gamma}{\lambda+\gamma}\right)^{i-x}, \text{ for } 0 < x \leq n, \tag{3}$$

$$\gamma p_{x,0} + r\gamma p_{n,0} = s\mu p_{0,1} + s\mu\sum_{i=x+1}^{\infty} p_{i,0}\left(\frac{\gamma}{\lambda+\gamma}\right)^{i-x}, \text{ for } x > n. \tag{4}$$

Consider now the cut between $A_2 = \{(x,y') : y' \leq y\}$ and $S\backslash A_2$, where $y \geq 0$. This leads to

$$r\gamma p_{n,y} = s\mu p_{0,y+1}, \text{ for } y \geq 0. \tag{5}$$

Finally, from the cut between $A_3 = \{(0,y),(1,y),\cdots(x,y)\}$ and $S\backslash A_3$, where $x \geq 0$ and $y \geq 1$, we get

$$(s\mu+\lambda)p_{0,y} = s\mu p_{0,y+1} + s\mu\sum_{i=1}^{\infty} p_{i,y}\left(\frac{\gamma}{\lambda+\gamma}\right)^i + r\gamma\left(\frac{\gamma}{\lambda+\gamma}\right)^n p_{n,y-1}, \text{ for } y \geq 1, \tag{6}$$

$$\gamma p_{x,y} + s\mu p_{0,y} = s\mu p_{0,y+1} + s\mu\sum_{i=x+1}^{\infty} p_{i,y}\left(\frac{\gamma}{\lambda+\gamma}\right)^{i-x} + r\gamma\left(\frac{\gamma}{\lambda+\gamma}\right)^{n-x} p_{n,y-1}, \text{ for } 0 < x \leq n \text{ and } y \geq 1, \tag{7}$$

$$\gamma p_{x,y} + s\mu p_{0,y} = s\mu p_{0,y+1} + s\mu\sum_{i=x+1}^{\infty} p_{i,y}\left(\frac{\gamma}{\lambda+\gamma}\right)^{i-x} + r\gamma p_{n,y-1}, \text{ for } x > n \text{ and } y \geq 1. \tag{8}$$

Probability of an empty system Summing up Eqs. (4) and (8) for $y \geq 1$ yields

$$\gamma p_x = s\mu\sum_{k=1}^{\infty}\left(\frac{\gamma}{\lambda+\gamma}\right)^k p_{x+k},$$

for $x > n$. Let us denote by z, a root of the related homogeneous equation. We then have

$$\gamma = s\mu \sum_{k=1}^{\infty} \left(\frac{\gamma}{\lambda+\gamma}\right)^k z^k,$$

which leads to $\gamma(\lambda + \gamma(1-z)) = s\mu\gamma z$. This equation has a unique solution; $z = \frac{\lambda+\gamma}{s\mu+\gamma} = \frac{a_\gamma}{s}$. Therefore, we have $p_{x+n+1} = \left(\frac{a_\gamma}{s}\right)^x p_{n+1}$, for $x \geq 0$. Summing up now Eqs. (3) and (7) for $y \geq 1$ and $x = n$ yields

$$(1-r)\gamma p_n = s\mu \sum_{k=1}^{\infty} \left(\frac{\gamma}{\lambda+\gamma}\right)^k p_{n+k},$$

so we deduce that $p_{x+n} = (1-r)\left(\frac{a_\gamma}{s}\right)^x p_n$ for $x \geq 0$. We now prove by induction on x that $p_{n-x} = \left(\frac{s}{a_\gamma}\right)^x p_n$, for $0 \leq x < n$. This relation is clearly true for $x = 0$. Assume now that this relation holds for $p_n, p_{n-1}, \ldots, p_{n-x}$. Summing up now Eqs. (3) and (7) for $y \geq 1$ yields

$$\begin{aligned}
\gamma p_{n-(x+1)} &= s\mu \left(\frac{\gamma}{\lambda+\gamma}\right)\left(\frac{s}{a_\gamma}\right)^x p_n + s\mu \left(\frac{\gamma}{\lambda+\gamma}\right)^2 \left(\frac{s}{a_\gamma}\right)^{x-1} p_n + \cdots \\
&\quad + s\mu \left(\frac{\gamma}{\lambda+\gamma}\right)^x \left(\frac{s}{a_\gamma}\right) p_n + (r\gamma + s\mu)\left(\frac{\gamma}{\lambda+\gamma}\right)^{x+1} p_n \\
&\quad + s\mu(1-r)\sum_{k=1}^{\infty} \left(\frac{\gamma}{\lambda+\gamma}\right)^{x+1+k} \left(\frac{a_\gamma}{s}\right)^k p_n \\
&= s\mu \sum_{i=1}^{x+1} \left(\frac{\gamma}{\lambda+\gamma}\right)^i \left(\frac{s}{a_\gamma}\right)^{x+1-i} p_n + \gamma r \left(\frac{\gamma}{\lambda+\gamma}\right)^{x+1} p_n \\
&\quad + \gamma(1-r)\left(\frac{\gamma}{\lambda+\gamma}\right)^{x+1} p_n.
\end{aligned}$$

Using $\left(\frac{\gamma}{\lambda+\gamma}\right)\left(\frac{s}{a_\gamma}\right)^{-1} = \frac{\gamma}{s\mu+\gamma}$, we may write

$$\begin{aligned}
\gamma p_{n-(x+1)} &= s\mu \left(\frac{s}{a_\gamma}\right)^{x+1} \sum_{i=1}^{x+1} \left(\frac{\gamma}{s\mu+\gamma}\right)^i p_n + \gamma \left(\frac{\gamma}{\lambda+\gamma}\right)^{x+1} p_n \\
&= s\mu \left(\frac{s}{a_\gamma}\right)^{x+1} \frac{\gamma}{s\mu+\gamma} \frac{1-\left(\frac{\gamma}{s\mu+\gamma}\right)^{x+1}}{1-\frac{\gamma}{s\mu+\gamma}} p_n + \gamma \left(\frac{\gamma}{\lambda+\gamma}\right)^{x+1} p_n \\
&= \gamma \left(\frac{s}{a_\gamma}\right)^{x+1} p_n,
\end{aligned}$$

which proves the induction step. Using Eq. (6), with the same approach we also obtain $p_0 = \frac{\gamma}{\lambda}\left(\frac{s}{a_\gamma}\right)^n p_n$; therefore, $p_x = \frac{\lambda}{\gamma}\left(\frac{a_\gamma}{s}\right)^x p_0$ for $1 \le x \le n$ and $p_x = (1-r)\frac{\lambda}{\gamma}\left(\frac{a_\gamma}{s}\right)^x p_0$ for $x > n$. From the last expression, the stability condition is $\frac{a_\gamma}{s} < 1$. This is equivalent to $\lambda < s\mu$ as for a simple M/M/s queue. Moreover, summing up Eq. (5) for $y \ge 0$ leads to $s\mu(p_0 - p_{0,0}) = r\gamma p_n$. So, $p_0 = \frac{p_{0,0}}{1-r\frac{a}{s}\left(\frac{a_\gamma}{s}\right)^n}$. Using now Eq. (1), we finally deduce that $p_0 = \frac{\frac{a^s}{s!}p_{-s,0}}{1-r\frac{a}{s}\left(\frac{a_\gamma}{s}\right)^n}$. Using the fact that the overall sum of the stationary probabilities is equal to one, we obtain the probability of an empty system as in Proposition 1.

Other stationary probabilities We can show that $p_{n+x,0} = (1-r)\left(\frac{\alpha_\gamma}{s}\right)^x p_{n,0}$ for $x > 0$. The proof is identical to the proof for p_{n+x} above.
We now show by induction on x that

$$p_{n-x,0} = p_{n,0}\left\{\left(\frac{s}{a_\gamma}\right)^x + \frac{r\lambda}{s\mu-\lambda}\left(\left(\frac{s}{a_\gamma}\right)^x - 1\right)\right\}, \tag{9}$$

for $0 \le x < n$. This relation is clearly true for $x = 0$. Assume now that this relation holds for $p_{n,0}$, $p_{n-1,0}$, $p_{n-x,0}$. One may write using Eq. (3) that

$$\begin{aligned}\gamma p_{n-(x+1),0} = s\mu p_{0,1} + s\mu\sum_{k=0}^{x}\left(\frac{\gamma}{\lambda+\gamma}\right)^{x+1-k} p_{n-k,0}\\ + s\mu(1-r)\sum_{k=1}^{\infty}\left(\frac{\gamma}{\lambda+\gamma}\right)^{x+1+k}\left(\frac{\alpha_\gamma}{s}\right)^k p_{n,0}.\end{aligned}$$

We now replace $p_{n,0}, p_{n-1,0}, \ldots, p_{n-x,0}$ by their expressions as a function of $p_{n,0}$ and $s\mu p_{0,1}$ by $r\gamma p_{n,0}$ (Eq. 5). We obtain

$$\begin{aligned}\gamma p_{n-(x+1),0} = r\gamma p_{n,0} + s\mu(1-r)\sum_{k=1}^{\infty}\left(\frac{\gamma}{\lambda+\gamma}\right)^{x+1+k}\left(\frac{\alpha_\gamma}{s}\right)^k p_{n,0}\\ + s\mu p_{n,0}\sum_{k=0}^{x}\left(\frac{\gamma}{\lambda+\gamma}\right)^{x+1-k}\left\{\left(\frac{s}{a_\gamma}\right)^k + \frac{r\lambda}{s\mu-\lambda}\left(\left(\frac{s}{a_\gamma}\right)^k - 1\right)\right\}.\end{aligned}$$

Using now $\sum_{k=0}^{x}\left(\frac{\gamma}{\lambda+\gamma}\right)^{x+1-k} = \frac{\gamma}{\lambda}\left(1-\left(\frac{\gamma}{\lambda+\gamma}\right)^{x+1}\right)$, $\sum_{k=0}^{x}\left(\frac{\gamma}{\lambda+\gamma}\right)^{x+1-k}\left(\frac{s}{a_\gamma}\right)^k = \frac{\gamma}{s\mu}\left(\left(\frac{s}{a_\gamma}\right)^{x+1} - \left(\frac{\gamma}{\lambda+\gamma}\right)^{x+1}\right)$, and $\sum_{k=1}^{\infty}\left(\frac{\gamma}{\lambda+\gamma}\right)^{x+1+k}\left(\frac{\alpha_\gamma}{s}\right)^k = \frac{\lambda+\gamma}{s\mu}\left(\frac{\gamma}{\lambda+\gamma}\right)^{x+2}$, we prove the induction step. Observe that Eq. (2) is almost identical to Eq. (3) in which we would replace x by 0. The only difference is the multiplicative coefficient on the left hand side of Eq. (2). This one is λ instead of γ. Therefore, using the corrective coefficient $\frac{\gamma}{\lambda}$, we deduce the explicit expression of $p_{0,0}$;

$$p_{0,0} = \frac{\gamma}{\lambda} p_{n,0} \left\{ \left(\frac{s}{a_\gamma}\right)^n + \frac{r\lambda}{s\mu - \lambda} \left(\left(\frac{s}{a_\gamma}\right)^n - 1 \right) \right\}.$$

This last equation relates $p_{0,0}$ and $p_{n,0}$. By substituting the expression of $p_{n,0}$ as a function of $p_{0,0}$ into Eq. (9), we get

$$\begin{aligned} p_{x,0} &= p_{0,0} \frac{\lambda}{\gamma} \frac{\left(\frac{s}{a_\gamma}\right)^{n-x} + \frac{r\lambda}{s\mu-\lambda} \left(\left(\frac{s}{a_\gamma}\right)^{n-x} - 1 \right)}{\left(\frac{s}{a_\gamma}\right)^{n} + \frac{r\lambda}{s\mu-\lambda} \left(\left(\frac{s}{a_\gamma}\right)^{n} - 1 \right)} \\ &= p_{0,0} \frac{\lambda}{\gamma} \frac{\left(\frac{a_\gamma}{s}\right)^x (s\mu - \lambda(1-r)) - r\lambda \left(\frac{a_\gamma}{s}\right)^n}{s\mu - \lambda(1-r) - r\lambda \left(\frac{a_\gamma}{s}\right)^n}, \end{aligned}$$

for $1 \leq x \leq n$, and

$$p_{x,0} = p_{0,0}(1-r) \frac{\lambda}{\gamma} \frac{(s\mu - \lambda) \left(\frac{a_\gamma}{s}\right)^{x-n}}{s\mu - \lambda(1-r) - r\lambda \left(\frac{a_\gamma}{s}\right)^n},$$

for $x > n$.

With the same approach, one can show by induction that $p_{n+x,y} = p_{n,y}(1-r) \left(\frac{a_\gamma}{s}\right)^x$, for $x > 0$ and

$$\begin{aligned} p_{n-x,y} = p_{n,y} &\left\{ \left(\frac{s}{a_\gamma}\right)^x + \frac{r\lambda}{s\mu - \lambda} \left(\left(\frac{s}{a_\gamma}\right)^x - 1 \right) \right\} \\ &+ \frac{r\lambda}{s\mu - \lambda} p_{n,y-1} \left[1 - \left(\frac{s}{a_\gamma}\right)^x \right], \end{aligned} \tag{10}$$

for $0 \leq x < n$. Combining now Eq. (6) with Eq. (10), we get

$$\begin{aligned} p_{0,y} = p_{n,y} \frac{\gamma}{\lambda} &\left\{ \left(\frac{s}{a_\gamma}\right)^n + \frac{r\lambda}{s\mu - \lambda} \left(\left(\frac{s}{a_\gamma}\right)^n - 1 \right) \right\} \\ &+ \frac{r\lambda}{s\mu - \lambda} p_{n,y-1} \frac{\gamma}{\lambda} \left[1 - \left(\frac{s}{a_\gamma}\right)^n \right]. \end{aligned}$$

This last equation relates $p_{0,y}$, $p_{n,y}$ and $p_{n,y-1}$. Since $s\mu p_{0,y} = r\gamma p_{n,y-1}$ for $y \geq 1$ (Eq. 5), we obtain a relation between $p_{0,y}$ and $p_{n,y}$;

$$\begin{aligned} p_{0,y} = p_{n,y} \frac{\gamma}{\lambda} &\left\{ \left(\frac{s}{a_\gamma}\right)^n + \frac{r\lambda}{s\mu - \lambda} \left(\left(\frac{s}{a_\gamma}\right)^n - 1 \right) \right\} \\ &+ \frac{\lambda}{s\mu - \lambda} p_{0,y} \frac{s\mu}{\lambda} \left[1 - \left(\frac{s}{a_\gamma}\right)^n \right]. \end{aligned}$$

This last equation can be finally simplified into

$$p_{n,y} = \frac{\lambda}{\gamma} p_{0,y} \frac{s\mu - \lambda \left(\frac{a_\gamma}{s}\right)^n}{s\mu - \lambda(1-r) - r\lambda \left(\frac{a_\gamma}{s}\right)^n},$$

for $y \geq 1$.

Equation (5) gives an expression of $p_{n,y-1}$ as a function of $p_{0,y}$. Inserting these two results into Eq. (10) leads to an expression of $p_{x,y}$ as a function of $p_{0,y}$;

$$p_{x,y} = \frac{\lambda}{\gamma} p_{0,y} \frac{s\mu - \lambda(1-r)\left(\frac{a_\gamma}{s}\right)^x - r\lambda \left(\frac{a_\gamma}{s}\right)^n}{s\mu - \lambda(1-r) - r\lambda \left(\frac{a_\gamma}{s}\right)^n},$$

for $0 < x \leq n$ and $y \geq 1$. Finally, from Eq. (5) we get

$$p_{n,y} = \left(\frac{r\lambda}{s\mu} \frac{s\mu - \lambda \left(\frac{a_\gamma}{s}\right)^n}{s\mu - \lambda(1-r) - r\lambda \left(\frac{a_\gamma}{s}\right)^n}\right)^y p_{n,0}.$$

This finishes the proof of the proposition. □

3.1.2 Performance measures

In Theorem 1, we derive the performance measures. In order to relate the performance measures to those of an M/M/s queue, we introduce the notation $C(s,a) = P(W > 0)$ (i.e., probability of queueing in an M/M/s queue). Recall from Kleinrock (1975, p. 103) that $C(s,a) = \frac{\frac{a^s}{s!}}{\sum\limits_{x=0}^{s-1} \frac{a^x}{x!} + \frac{a^s}{s!}\frac{1}{1-a/s}} \cdot \frac{1}{1-a/s}$.

Theorem 1 *We have*

$$P_c = r \cdot C(s,a) \cdot \frac{(1-a/s)e^{-s\mu(1-a/s)\cdot K}}{1 - r\frac{a}{s}e^{-s\mu(1-a/s)\cdot K}},$$

$$P(W > K) = C(s,a)\frac{(1-r\frac{a}{s})e^{-s\mu(1-a/s)\cdot K}}{1 - r\frac{a}{s}e^{-s\mu(1-a/s)\cdot K}},$$

$$E(W_1) = \frac{\frac{a^s}{s!}}{s\mu} \cdot \frac{1 - re^{-s\mu(1-a/s)\cdot K}(1 + s\mu(1-a/s)\cdot K)}{(1-a/s)^2\left(\left(1 - r\frac{a}{s}e^{-s\mu(1-a/s)\cdot K}\right)\sum\limits_{x=0}^{s-1}\frac{a^x}{x!} + \frac{a^s}{s!}\frac{1-re^{-s\mu(1-a/s)\cdot K}}{1-a/s}\right)},$$

$$E(W_2) = \frac{1 + s\mu \cdot K}{s\mu(1-a/s)}.$$

Proof The approach to derive the performance measures first consists of defining the embedded Markov chain at specific instants chosen in order to reach the performance measures at arbitrary instants. Next, by letting γ and n tend to infinity we obtain the results.

The embedded Markov chain Arriving customers either enter service upon arrival, enter service from Queue 1 after some wait, or are routed to Queue 2. Call the instants when one of these three events occurs Q-instants. Since the events at Q-instants all occur one at a time, in the long run the system is identical at arrival instants and Q-instants. Since the Poisson arrival process of customers is independent of the system state, the system is identical at arrival instants and arbitrary instants. So, the system is also identical at arbitrary instants and Q-instants. We therefore choose to consider the system at Q-instants to obtain the performance measures (the arrival instants cannot be seen in our Markov chain).

The Q-instants are determined by λ-transitions from state with a vacant server, $s\mu$-transitions from the other states except in states $(0, y)$ and γ-transitions from states (n, y), for $y \geq 0$. The overall customer flow at Q-instants is identical to the customer flow at arrival instants and has a rate λ. Therefore, the probability at Q-instants that x servers are busy for $0 \leq x < s$ is $\frac{\lambda}{\lambda} p_{-s+x,0} = p_{-s+x,0}$. The probability that the FIL is in waiting phase x and y customers are in Queue 2 is $\frac{s\mu}{\lambda} p_{x,y}$ for $0 < x < n$ or $x > n$, 0 for $x = 0$ and $\frac{s\mu + r\gamma}{\lambda} p_{n,y}$ for $x = n$. The stationary probabilities at Q-instants are then completely known. This allows us to derive the performance measures.

Performance measures The approach to obtain the performance measures is to let γ and n tend to infinity with respect to $\frac{n}{\gamma} = K$. First, we have

$$\lim_{n,\gamma\to\infty} \left(\frac{a_\gamma}{s}\right)^n = e^{-s\mu(1-a/s)\cdot K}.$$

We now derive the proportion of customers who are routed to Queue 2, P_c. A customer moves from Queue 1 to Queue 2 due to a γ-transition from states (n, y), $y \geq 0$. The proportion of customers which are moved from Queue 1 to Queue 2 is therefore

$$P_c = \lim_{n,\gamma\to\infty} r\frac{\gamma}{\lambda} p_n.$$

Recall from the proof of Proposition 1 that $p_n = \frac{\lambda}{\gamma}\left(\frac{a_\gamma}{s}\right)^n p_0$ and $p_0 = \frac{\frac{a^s}{s!} p_{-s,0}}{1 - r\frac{a}{s}\left(\frac{a_\gamma}{s}\right)^n}$. Therefore,

$$r\frac{\gamma}{\lambda} p_n = r\left(\frac{a_\gamma}{s}\right)^n \frac{\frac{a^s}{s!} p_{-s,0}}{1 - r\frac{a}{s}\left(\frac{a_\gamma}{s}\right)^n}. \tag{11}$$

From the expression of $p_{-s,0}$ in Proposition 1, we get the probability of an empty system in an M/M/s queue:

$$\lim_{n,\gamma\to\infty} p_{-s,0} = \left[\sum_{x=0}^{s-1} \frac{a^x}{x!} + \frac{a^s}{s!}\frac{1}{1 - a/s}\right]^{-1}. \tag{12}$$

By applying the last result in Eq. (11), we obtain the explicit expression of P_c.

We now derive the proportion of customers who waits less than K, $P(W < K)$. A customer is served from Queue 1 due to a $s\mu$-transition from states (x, y), $y \geq 0$. Therefore,

$$P(W < K) = \lim_{n,\gamma\to\infty} p_{-s,0} + p_{-s+1,0} + \cdots + p_{-1,0} + \frac{s\mu}{\lambda}(p_1 + p_2 + \cdots + p_n).$$

Therefore, we get

$$P(W < K) = \lim_{n,\gamma\to\infty} p_{-s,0}\left(\sum_{x=0}^{s-1}\frac{a^x}{x!} + \frac{\frac{a^s}{s!}}{1-a/s}\frac{\lambda+\gamma}{\gamma}\frac{1-\left(\frac{\lambda+\gamma}{s\mu+\gamma}\right)^n}{1-r\frac{a}{s}\left(\frac{\lambda+\gamma}{s\mu+\gamma}\right)^n}\right);$$

this in turn leads to the result of the theorem.

Consider now the served customers from Queue 1. A served customer from Queue 1 waits x γ-phases with probability $\frac{s\mu}{\lambda}p_x$ for $x > 0$, and each phase has an expected duration of $1/\gamma$. Therefore,

$$\begin{aligned}(1-P_c)E(W_1) &= \lim_{n,\gamma\to\infty}\frac{s\mu}{\lambda}\sum_{x=1}^{\infty}\frac{x}{\gamma}p_x\\ &= \lim_{n,\gamma\to\infty} p_0\frac{s\mu}{\gamma^2}\frac{a_\gamma}{s}\frac{-r(n+1)\left(1-\frac{a_\gamma}{s}\right)\left(\frac{a_\gamma}{s}\right)^n + 1 - r\left(\frac{a_\gamma}{s}\right)^n}{\left(1-\frac{a_\gamma}{s}\right)^2}.\end{aligned}$$

In order to compute this limit, we separate the last expression in three parts. First, we may write

$$\lim_{n,\gamma\to\infty} p_0 = \lim_{n,\gamma\to\infty}\frac{\frac{a^s}{s!}p_{-s,0}}{1-r\frac{a}{s}\left(\frac{a_\gamma}{s}\right)^n} = \frac{\frac{a^s}{s!}\left[\sum_{x=0}^{s-1}\frac{a^x}{x!}+\frac{a^s}{s!}\frac{1}{1-a/s}\right]^{-1}}{1-r\frac{a}{s}e^{-s\mu(1-a/s)\cdot K}}. \tag{13}$$

Second, we have

$$\begin{aligned}\lim_{n,\gamma\to\infty}\frac{s\mu}{\gamma^2}\frac{a_\gamma}{s}\frac{1}{\left(1-\frac{a_\gamma}{s}\right)^2} &= \lim_{n,\gamma\to\infty}\frac{s\mu}{(s-a)^2}\frac{\left(a+\frac{\gamma}{\mu}\right)\left(s+\frac{\gamma}{\mu}\right)}{\gamma^2}\\ &= \frac{1}{s\mu(1-a/s)^2}.\end{aligned} \tag{14}$$

Finally, one may write

$$\begin{aligned}&-r(n+1)\left(1-\frac{a_\gamma}{s}\right)\left(\frac{a_\gamma}{s}\right)^n + 1 - r\left(\frac{a_\gamma}{s}\right)^n\\ &= 1 - r\left(\frac{a_\gamma}{s}\right)^n - r\frac{(n+1)(s-a)}{s+\gamma/\mu}\left(\frac{a_\gamma}{s}\right)^n\end{aligned}$$

Applying the assumption $\frac{n}{\gamma} = K$ yields

$$\lim_{n,\gamma \to \infty} -r(n+1)\left(1-\frac{a_\gamma}{s}\right)\left(\frac{a_\gamma}{s}\right)^n + 1 - r\left(\frac{a_\gamma}{s}\right)^n$$
$$= 1 - re^{-s\mu(1-a/s)\cdot K}(1 + s\mu(1-a/s)\cdot K). \tag{15}$$

Combining Eqs. (13), (14) and (15) leads to the expression of $E(W_1)$.

We now consider the expected waiting time of customers who are routed to Queue 2. The probability of having y customers in Queue 2 at Q-instants ($y \geq 0$) is $\sum_{x=1}^{\infty} \frac{s\mu}{\lambda} p_{x,y} + \frac{r\gamma}{\lambda} p_{n,y}$. Using the results of Proposition 1, we can compute explicitly this expression by letting n and γ tends to infinity.

3.2 Numerical analysis with abandonment

The complexity of the transition structure does not allow us to obtain explicit expressions for the performance measures with abandonment. However, since the transition structure is completely known, using space state truncation with a bound, D_1, for the number of waiting phases in Queue 1 and a bound, D_2, for the number of customers in Queue 2, we can derive the performance measures including the proportion of abandonment.

Let S be the state space. Consider the cut between $A_1 = \{(-s, 0), \ldots, (x, 0)\}$ and $S\backslash A_1$, where $-s \leq x \leq D_1$. By equating flows across the cut, one may write

$$\lambda p_{x,0} = (s + x + 1)\mu p_{x+1,0}, \text{ for } -s \leq x < 0, \tag{16}$$

$$\lambda p_{0,0} = s\mu p_{0,1} + \left(s\mu + \gamma\frac{\beta}{\gamma+\beta}\right)\sum_{i=1}^{D_1} p_{i,0}q_{i,0}, \tag{17}$$

$$\gamma p_{x,0} = s\mu p_{0,1} + \left(s\mu + \gamma\frac{\beta}{\gamma+\beta}\right)\sum_{i=x+1}^{D_1} p_{i,0}\sum_{k=0}^{x} q_{i,k}, \text{ for } 0 < x \leq n, \tag{18}$$

$$\gamma p_{x,0} + r\gamma p_{n,0} = s\mu p_{0,1} + \left(s\mu + \gamma\frac{\beta}{\gamma+\beta}\right)\sum_{i=x+1}^{D_1} p_{i,0}\sum_{k=0}^{x} q_{i,k}, \text{ for } n < x < D_1. \tag{19}$$

Consider now the cut between $A_2 = \{(x, y') : y' \leq y\}$ and $S\backslash A_2$, where $0 \leq y \leq D_2$. This leads to

$$r\gamma p_{n,y} = s\mu p_{0,y+1}, \text{ for } 0 \leq y < D_2. \tag{20}$$

Finally, from the cut between $A_3 = \{(0, y), (1, y), \cdots (x, y)\}$ and $S\backslash A_3$, where $-s \leq x \leq D_1$ and $1 \leq y \leq D_2$, we get

$$(s\mu + \lambda)p_{0,y} = s\mu p_{0,y+1} + \left(s\mu + \gamma\frac{\beta}{\gamma+\beta}\right)\sum_{i=1}^{D_1} p_{i,y}q_{i,0} + r\gamma q_{n,0}p_{n,y-1}, \quad (21)$$

for $1 \leq y < D_2$,

$$\gamma p_{x,y} + s\mu p_{0,y} = s\mu p_{0,y+1} + \left(s\mu + \gamma\frac{\beta}{\gamma+\beta}\right)\sum_{i=x+1}^{D_1} p_{i,y}\sum_{k=0}^{x} q_{i,k} + r\gamma\sum_{k=0}^{x} q_{n,k}p_{n,y-1}, \quad (22)$$

for $0 < x \leq n,$ and $1 \leq y < D_2$,

$$\gamma p_{x,y} + s\mu p_{0,y} = s\mu p_{0,y+1} + \left(s\mu + \gamma\frac{\beta}{\gamma+\beta}\right)\sum_{i=x+1}^{D_1} p_{i,y}\sum_{k=0}^{x} q_{i,k} + r\gamma p_{n,y-1} \quad (23)$$

for $n < x < D_1$ and $1 \leq y < D_2$.

We then get a finite number of equations due to the state space truncation. In addition to the normalizing condition (i.e., the sum of the overall probabilities is equal to one), on may obtain numerically all stationary probabilities.

Arriving customers either enter service upon arrival, enter service from Queue 1 or Queue 2 after some wait, abandon from Queue 1 after experiencing some wait, or move from Queue 1 to Queue 2 after waiting n phases. The proportion of customers which accepts the callback offer, P_c, is then given by

$$P_c = r\frac{\gamma}{\lambda}\sum_{y=0}^{D_2} p_{n,y}.$$

The proportion of customers who have waited less than K time units, $P(W < K)$, is

$$P(W < K) = \sum_{x=-s}^{-1} p_{x,0} + \sum_{y=0}^{D_2}\sum_{x=1}^{n}\frac{s\mu + \gamma\frac{\beta}{\gamma+\beta}}{\lambda}p_{x,y}.$$

The proportion of abandonment, P_a, is

$$P_a = \sum_{y=0}^{D_2}\sum_{x=1}^{D_1}\frac{\gamma\frac{\beta}{\gamma+\beta}}{\lambda}p_{x,y}.$$

The expected waiting time in Queue 1, $E(W_1)$, is given by

$$(1 - P_c)E(W_1) = \sum_{y=0}^{D_2}\sum_{x=1}^{D_1}\frac{s\mu + \gamma\frac{\beta}{\gamma+\beta}}{\lambda}\frac{x}{\gamma}p_{x,y}.$$

We now consider the expected waiting time of customers who are routed to Queue 2. The probability of having y customers in Queue 2 ($y \geq 0$) is $\sum_{x=1}^{D_1} \frac{s\mu+\gamma\frac{\beta}{\gamma+\beta}}{\lambda} p_{x,y} + \frac{r\gamma}{\lambda} p_{n,y}$. This leads to the expected number in Queue 2. Next, applying Little's Law leads to $E(W_2)$.

One difficulty in the computation is the choice for the two parameters γ and D_1. The truncation parameter D_1 introduces the risk of having a large probability mass in the truncated state, particularly for large values of γ. The value of γ has an important influence on the approximation. Increasing γ means that more states are required for the truncation. At the same time, γ should be sufficiently large to represent the continuous elapsing of time.

4 Operational findings, discussions and insights

We investigate the issues related to a postponed callback offer. We derive a series of insights which can be proved in the case without abandonment. The proven results are next discussed with abandonment. More precisely, in Sect. 4.1, we show how a postponed callback offer can improve a waiting time percentile. In Sect. 4.2, we analyze how the customer's behavior may impact the system performance and what may be a customer rational strategy. In Sect. 4.3, we investigate the impact of the control parameter K on the performance measures to obtain recommendations to better control the system performance. Finally, in Sect. 4.4, we conduct a comparison between our postponed callback option and a callback option given at customer's arrival as developed in the literature (e.g., see Armony and Maglaras 2004a; Legros et al. 2016).

4.1 The callback offer, a tool to improve a waiting time percentile

We evaluate the impact of the callback offer on $P(W < K)$.

Analysis without abandonment We denote by R the ratio between $P(W > K)$ with the callback offer and $P(W > K)$ without the callback offer. Without the callback offer, we have $P(W > K) = C(s, a) \cdot e^{-s\mu(1-a/s)\cdot K}$. Therefore, using the expression of $P(W > K)$ in Theorem 1, we get

$$R = \frac{1 - r\frac{a}{s}}{1 - r\frac{a}{s}e^{-s\mu(1-a/s)\cdot K}} \leq 1.$$

So, as a first insight, we obtain

Insight 1 *The callback offer allows the manager to reduce a waiting time percentile.*

In Fig. 2, we represent $P(W > K)$ and R as a function of the workload for three different values for the callback acceptance parameter r. We observe that the higher is r, the smaller are $P(W > K)$ and R. This can be proved by

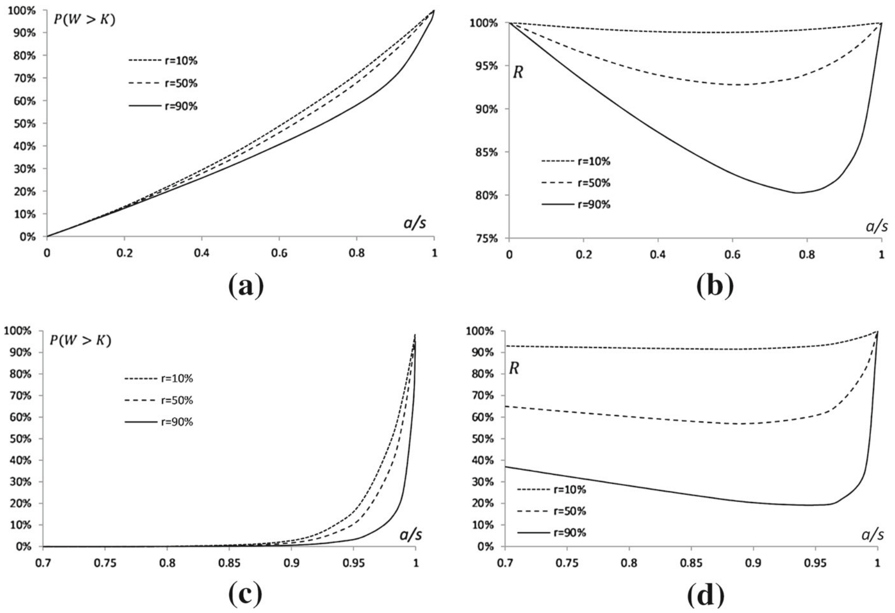


Fig. 2 $P(W > K)$ ($\mu = 1$, $K = 0.5$, $\beta = 0$). **a** $s = 1$, **b** $s = 1$, **c** $s = 50$ and **d** $s = 50$

$$\frac{\partial P(W > K)}{\partial r} = -C(s,a)e^{-s\mu(1-a/s)\cdot K}\,\frac{\frac{a}{s}(1-e^{-s\mu(1-a/s)\cdot K})}{\left(1-r\frac{a}{s}e^{-s\mu(1-a/s)\cdot K}\right)^2} < 0, \text{ and}$$

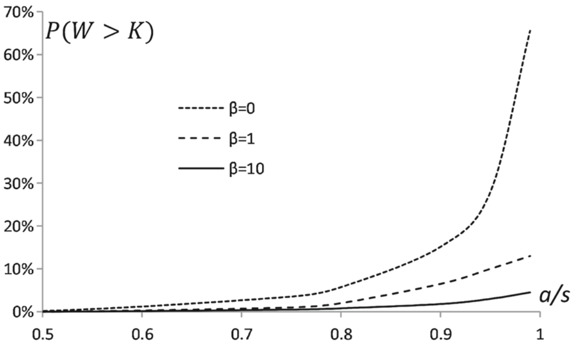
$$\frac{\partial R}{\partial r} = -\frac{\frac{a}{s}(1-e^{-s\mu(1-a/s)\cdot K})}{\left(1-r\frac{a}{s}e^{-s\mu(1-a/s)\cdot K}\right)^2} < 0.$$

One would expect that the impact of accepting the callback offer is stronger under high workload situations. Yet, the highest improvement is for intermediate workload situations as shown in Fig. 2b, d. This can be explained as follows. For low workload situations, the probability of waiting less that the threshold K is high. Therefore, most customers do not hear the callback offer. Under high workload situations, most customers hear the callback offer, but whether they accept it or not, they will wait more than K. The comparison between Fig. 2a and c illustrates that the absolute improvement is stronger in small call centers. The reason is related to the pooling effect. It is well established that the pooling effect in large call centers reduces the improvement that a good routing strategy could bring (Bassamboo et al. 2010; Legros et al. 2015). In summary, our observations lead to a second insight:

Insight 2 *The more customers are likely to accept the callback offer, the more strongly $P(W > K)$ can be improved. The maximal improvement is for intermediate workload situations and for small call center size.*

Impact of the abandonment The callback offer can be used to prevent some customers with too long waiting time to leave the system. It is then interesting to observe how

Fig. 3 $P(W > K)$ ($\mu = 1$, $s = 10$, $K = 0.5$, $r = 90\%$)



abandonment may impact $P(W > K)$. In Fig. 3, we give $P(W > K)$ as a function of the ratio a/s for different values of the abandonment rate. An interesting observation is that the abandonment feature strongly helps to reduce $P(W > K)$. This is particularly apparent in high workload situations. Callback customers then benefit from the abandonment of customers in Queue 1 because the abandonment participates in the departure flow from Queue 1.

4.2 Customer's behavior

We investigate here the customer's reaction to the callback offer.

Impact of r *on* $E(W_2)$ The parameter r is assumed to capture the customer's behavior. An interesting observation is that this parameter r is not part of the expression of $E(W_2)$ without abandonment. This means that the delay for callback customers is insensitive to the willingness of customers to accept the callback offer. Hence, we get the following insight:

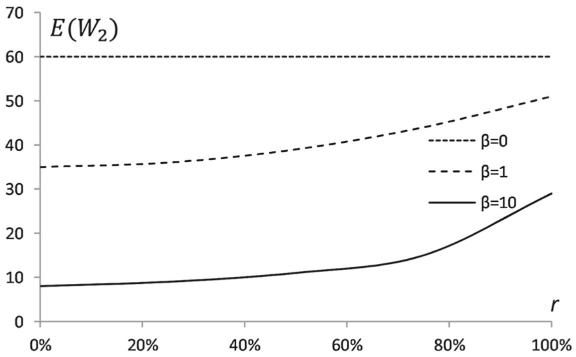
Insight 3 *Without abandonment, the delay for callback customers is insensitive to the parameter* r.

However, Fig. 4 reveals that with abandonment, the parameter r influences the delay for callback customers. More precisely, as r increases, $E(W_2)$ increases. This observation is intuitive. As r increases, the proportion of callback customers also increases. These customers do not abandon which in turn leads to a higher congestion of the system.

Rational customers We study here customers' rational behavior. First, with rational customers, one may neglect the exponential patience. As shown in Mandelbaum and Shimkin (2000), rational abandonments can occur only upon arrival (zero or infinite patience for each customer).

We then investigate the willingness to accept the callback offer without abandonment. The choice for a customer to accept the callback offer or not can be seen as the result of a rational decision. When hearing the callback offer at time K, a customer has the choice to stay in Queue 1 with a remaining expected waiting time of $\frac{1}{s\mu}$ (because

Fig. 4 $E(W_2)$ ($\mu = 1, s = 10, K = 0.5, \lambda = 9.9$)



the callback offer is given to the first customer in line) or can choose to be called back later with an expected delay of $E(W_2) - K$. Of course, accepting the callback offer leads to higher waiting time, but waiting to be called back is less costly/annoying than continuing to wait for an agent to be available. We capture by c_1 and c_2 the cost per time unit of waiting in the initial queue (Queue 1) or in the callback queue (Queue 2), respectively.

The parameter r should therefore be

$$r = \arg\min\left((1-r)c_1\frac{1}{s\mu} + rc_2(E(W_2) - K)\right),$$

with $c_1 \geq c_2$. Since $E(W_2)$ is insensitive to r, the optimal value for r is either 0 or 1. More precisely, we get:

Insight 4 *Only two rational customer strategies are possible. Either all customers who hear the callback offer accept this offer if* $c_2\frac{1+\lambda\cdot K}{1-a/s} < c_1$*; otherwise, they all reject the offer.*

The condition $c_2\frac{1+\lambda\cdot K}{1-a/s} < c_1$ induces that the higher the workload is, the more likely customers will refuse the callback offer. Intuitively, this can be explained by the long delays for callback customers in case of high workload situations due to their low priority. The second consequence is that the smaller is K, the more likely a customer will accept the callback offer. The reason is related to the proportion of callback customers. When K is small, a high proportion of customers will hear the offer. Therefore, if they all accept the offer, the proportion of those who are in Queue 1 is small and the effect of the low prioritization is reduced which in turn makes the callback offer attractive.

4.3 The control parameter K

The control parameter for the call center is the time at which the callback offer is proposed, K.

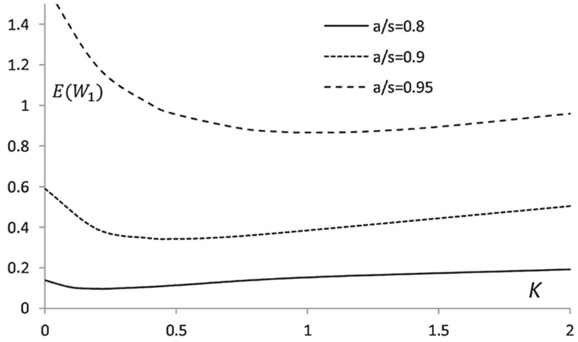


Fig. 5 $E(W_1)$ ($s = 10$, $\mu = 1$, $r = 0.8$)

With rational customers As mentioned in Sect. 4.2, by choosing a too high value for K, a call center with rational customers will induce a rejection of the callback offer ($r = 0$). In this case, the value of K is irrelevant. Under a waiting time threshold for the callback offer, all customers accept the offer ($r = 1$). In the case $r = 1$, both $E(W_1)$ and $E(W_2)$ are strictly increasing in K. This argue for a value of $K = 0$. However, in that case with $r = 1$ and $K = 0$, the call center manager may loose the control of the proportion of callback customers and the inbound queue will always be empty. This might be unwanted because inbound calls can be a source of revenue for the call center; contrary to outbound calls they may pay a waiting cost per waiting time unit. So, the choice of K also depends on the wanted proportion of callback customers. This proportion, P_c, is strictly decreasing in K. This can be seen by

$$\frac{\partial P_c}{\partial K} = -s\mu r \cdot C(s,a) \cdot \frac{(1-a/s)^2 e^{-s\mu(1-a/s)\cdot K}}{(1-r\frac{a}{s}e^{-s\mu(1-a/s)\cdot K})^2} < 0.$$

With irrational customers In the case $r < 1$, the elements mentioned above still hold except the monotonicity of $E(W_1)$. In Fig. 5, we present $E(W_1)$ as a function of K for different workload situations.

Proposition 2 *If $0 < r < 1$, there exists a unique value for K which minimizes $E(W_1)$. It is the unique solution in K of*

$$xA + re^{-x} = 1, \tag{24}$$

with $x = s\mu K(1 - a/s)$ and $A = \frac{\sum_{x=0}^{s-1}\frac{a^x}{x!} + \frac{a^s}{s!(1-a/s)}}{\frac{a}{s}\sum_{x=0}^{s-1}\frac{a^x}{x!} + \frac{a^s}{s!(1-a/s)}}$.

Note that in the case $r = 0$, $E(W_1)$ is insensitive to K.

Proof We obtain Eq. (24) from $\frac{\partial E(W_1)}{\partial K} = 0$. Consider the function $f(x) = xA + re^{-x} - 1$. We want to show that $f(x) = 0$ has a unique solution. We have $f'(x) = A - re^{-x}$. Since $x > 0$, $r < 1$ and $A > 1$, we have $f'(x) > 0$ for $x \geq 0$. So,

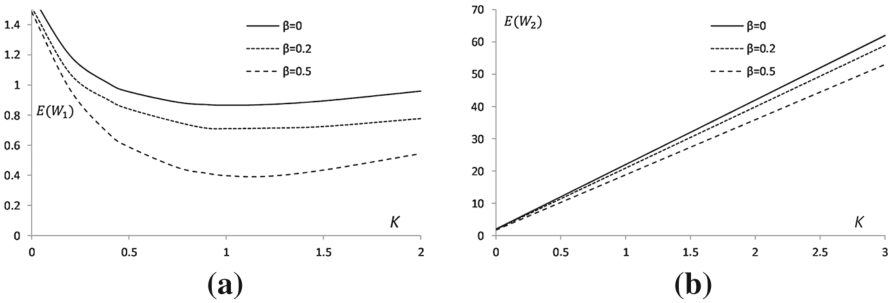


Fig. 6 Impact of the abandonment ($s = 10$, $r = 0.8$, $a/s = 0.95$, $\mu = 1$). **a** $E(W_1)$ and **b** $E(w_2)$

the function f is increasing in x for $x \geq 0$. Moreover, $f(0) = r - 1 < 0$ and $\lim_{x \to +\infty} f(x) = +\infty$. This proves that there exists a unique solution of Eq. (24). □

One way to obtain the unique solution of Eq. (24) is to apply the Newton algorithm by defining recursively x_k by $x_0 = 0$ and $x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}$ for $k \geq 0$ and f defined as in the proof of Proposition 2. Note that since $f'(x) > 0$ for $x \geq 0$, the recursion is well defined.

The reason which explains why $E(W_1)$ is not increasing in K is the definition of the callback offer. Increasing K does not necessarily mean that less customers have the callback proposition. Recall that only the first customer in line can hear the callback offer. In case of high workload situations and low value for K, the probability to be the FIL at waiting time K is low (except if $r = 1$). Most likely, at waiting time K a customer will have other customers in front of him and will not have the callback offer. Increasing K in this case leads to a higher chance to be the FIL at waiting time K. Therefore, increasing K leads to a higher chance to leave Queue 1. This explains how $E(W_1)$ can be decreasing in K. In case of low workload situations, increasing K reduces the proportion of callback customers and therefore increases $E(W_1)$.

With abandonment Figure 6a, b illustrates the impact of K on $E(W_1)$ and $E(W_2)$, for different values of the abandonment rate β. We observe that with abandonment, the value of K which minimizes $E(W_1)$ is higher than the one obtained without abandonment. With abandonment, the increasing of the number of customers in Queue 1 increases also the departure rate (after abandonment or service) of inbounds from the system, which makes the system more efficient and may decrease $E(W_1)$. Therefore, higher values for K may lead to a better performance for inbound calls. We observe that $E(W_2)$ is still increasing in K (Fig. 6b) although the abandonment in Queue 1 also reduces the waiting time in Queue 2.

The abandonment plays a important role in the choice of K. Since by definition outbound calls do not abandon, reducing K reduces abandonment, which is positive. Yet, this may also increase the workload and lead to higher waiting time. This leads to another insight.

Insight 5 *The callback offer may help to reduce the proportion of abandonment. However, the time at which the callback offer is proposed should be carefully chosen in order to avoid congestion.*

4.4 Comparison with a non-postponed callback offer

The callback offer studied in this article differs from the one in the literature by the instant at which it is proposed. In most callback models, the callback offer is given at arrival of a new call if the expected waiting time is too high (e.g., see Armony and Maglaras 2004a; Legros et al. 2016). Instead, we consider in this article a callback offer given after experimenting some wait. We propose to conduct a comparison between these two strategies.

We call Model A our postponed callback offer and by Model B a callback offer proposed at arrival of a new call. For Model B, we assume that at and above a given number of customers in Queue 1 (or equivalently at and above a given expected waiting time for an arriving customer) the callback offer is proposed to all arriving customers. Hence, in Model B, Queue 1 has a limited capacity n. All arriving customers are routed to Queue 2 if Queue 1 size is equal to n. Therefore, n is the control parameter of Model B. The performance measures in Model B can be obtained through a Markov chain analysis or can be deduced from Proposition 3 of Legros et al. (2016). We obtain the following performance measures for Model B:

$$P_c = C(s,a) \cdot \frac{(1-a/s)\left(\frac{a}{s}\right)^n}{1-\left(\frac{a}{s}\right)^{n+1}},$$

$$E(W_1) = \frac{\frac{a^s}{s!}}{s\mu} \cdot \frac{1-\left(\frac{a}{s}\right)^n(1+n(1-a/s))}{(1-a/s)^2\left(\left(1-\left(\frac{a}{s}\right)^{n+1}\right)\sum\limits_{x=0}^{s-1}\frac{a^x}{x!}+\frac{a^s}{s!}\frac{1-\left(\frac{a}{s}\right)^n}{1-a/s}\right)},$$

$$E(W_2) = \frac{1+n}{s\mu(1-a/s)}.$$

The difficulty in the comparison is the customer's reaction to the offer. It may differ whether the callback offer is given at arrival or later. To avoid this complexity, we assume that all customers accept the callback offer in both models. This corresponds to a rational behavior in Model A.

Comparison without abandonment In Theorem 2, we consider a context for which the call center manager wants to maintain the proportion of callback customers at a given level. Under this constraint which forces the two models to have the same proportion of callback customers, we prove that our postponed callback offer leads to a better expected waiting time for inbound calls and a worse expected waiting time for outbound ones.

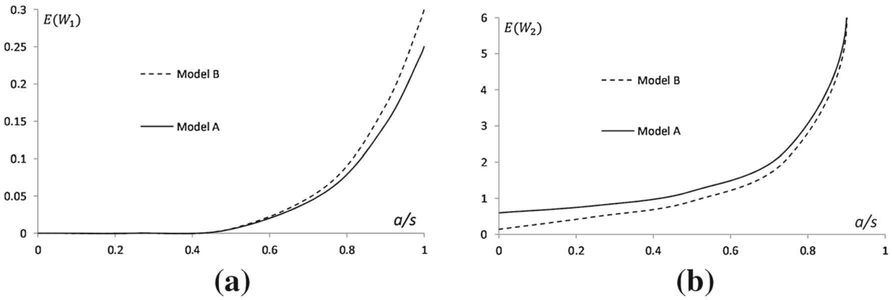


Fig. 7 Comparison between the callback offers ($s = 10$, $r = 1$, $\mu = 1$, $\beta = 0$, $K = 0.5$, $n \ln(a/s) = -s\mu(1 - a/s)K$). **a** $E(W_1)$ and **b** $E(W_2)$

Theorem 2 *Given that the control parameters K (Model A) and n (Model B) are chosen such that the proportion of callback customers in identical in both models, $E(W_1)$ is lower in Model A and $E(W_2)$ is lower in Model B.*

Proof To obtain the same proportion of callback customers in both models, the control parameters n and K should be related by $\left(\frac{a}{s}\right)^n = e^{-s\mu(1-a/s)K}$. This equation is equivalent to $n \ln(a/s) = -s\mu(1 - a/s)K$. Let us denote by $E(W_1)_A$ and $E(W_1)_B$, the expected waiting time of inbound calls in Model A and B. We have

$$E(W_1)_A - E(W_1)_B$$
$$= \frac{\frac{a^s}{s!}}{s\mu} \cdot \frac{e^{-s\mu(1-a/s)\cdot K}(n - s\mu K)}{(1 - a/s)\left(\left(1 - \frac{a}{s}e^{-s\mu(1-a/s)\cdot K}\right)\sum_{x=0}^{s-1}\frac{a^x}{x!} + \frac{a^s}{s!}\frac{1-e^{-s\mu(1-a/s)\cdot K}}{1-a/s}\right)}.$$

Thus, the sign of this difference depends on the sigh of $n - s\mu K$. One may write,

$$n - s\mu K = -\frac{s\mu K}{\ln(a/s)}(\ln(a/s) + 1 - a/s).$$

Since $a/s < 1$, $-\frac{s\mu K}{\ln(a/s)} > 0$. Thus, the sign of the expression depends on the sign of $\ln(a/s) + 1 - a/s$. Consider the function in x, $f(x) = \ln(x) + 1 - x$ for $x > 0$. We have $f'(x) = \frac{1}{x} - 1$. So $f'(x) > 0$ for $0 < x \leq 1$. Since $f(1) = 0$, $\ln(a/s) + 1 - a/s < 0$. This proves that $E(W_1)_A - E(W_1)_B < 0$. With the same approach, we can prove that the expected waiting time for outbound calls is higher with Model A. □

In Fig. 7a, b, we represent $E(W_1)$ and $E(W_2)$ as a function of the workload for Model A and Model B assuming a fixed value of $K = 0.5$ for Model A and n is adjusted in Model B with the relation $n \ln(a/s) = -s\mu(1 - a/s)K$ such that the two models achieve the same proportion of callback customers. An interesting observation is that the improvement for $E(W_1)$ with Model A is higher under high workload situations, whereas the improvement for $E(W_2)$ with Model B is higher under low workload situations. This leads to a last insight.

Insight 6 *A postponed callback offer is preferred under high workload situations.*

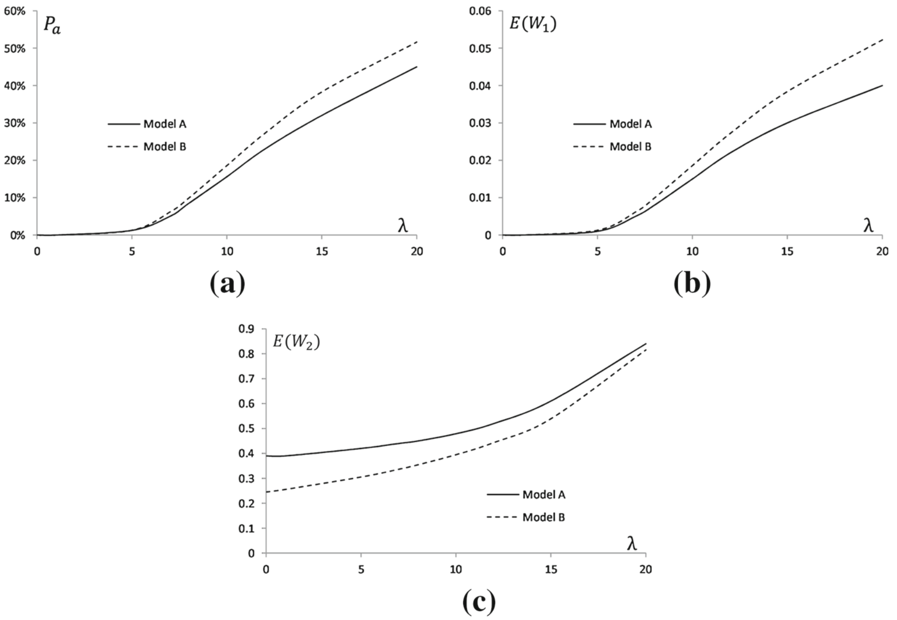


Fig. 8 Comparison between the callback offers ($s = 10$, $r = 1$, $\mu = 1$, $\beta = 10$, $n = 5$). **a** P_a, **b** $E(W_1)$ and **c** $E(W_2)$

Comparison with abandonment In Fig. 8a–c, we represent P_a, $E(W_1)$ and $E(W_2)$ as a function of the arrival rate for Model A and Model B assuming a fixed value of $n = 5$ for Model B and K is adjusted in Model A such that the two models achieve the same proportion of callback customers. We obtain the same qualitative observations as shown in Fig. 7. As mentioned in Insight 6, with abandonment the postponed callback offer is preferred under high workload situation. In addition, Fig. 8a reveals that for a given proportion of callback customers, the postponed callback offer achieves a lower proportion of abandonment. This is an essential value of the postponed callback offer; it allows the call center to reduce the proportion of lost customers.

5 Conclusion

In this article, we propose a new callback model. After experimenting some wait, the first customer in line receives a callback proposition and chooses to accept it or not. This simple model differs from the one in the literature where the callback offer is given at customers' arrival. We first develop a Markov chain analysis to derive the performance measures without abandonment. The same approach is also applied to compute numerically the performance measures with abandonment. This allows us to better understand the effect of the callback offer on the call center performance. We find that our callback offer succeeds in reducing a percentile of the waiting time. In particular, the realized improvement can be significant in intermediate workload situations, with abandonment and small call center size. One surprising result is that the delay for callback customers is insensitive to the willingness of customers to accept the callback offer without abandonment. This result is, however, no longer valid with abandonment. This leads to only two rational customer behaviors: either they all accept or they all reject the callback offer. Next, we evaluate how to derive the optimal value of K without abandonment and show how this parameter can be efficiently used to reduce the proportion of abandonment. Finally, we show that our postponed callback offer outperforms the existing ones in reducing the proportion of abandonment and the expected waiting time of inbound calls.

Several avenues are open for future research. It would be interesting to develop a callback offer with a state-dependent starting time. This may give a trade-off between the benefits of the postponed and non-postponed callback offer. In addition, more complexity could be included in the model like retrials and reconnections, time-dependent parameters or other type of service time or patience distributions.

Acknowledgements The authors want to express their gratitude to the Reviewers for their useful comments that significantly improved this paper. The authors would like also to thank Sébastien Thorel from INTERACTIV GROUP for their helpful discussions.

References

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

Armony M, Maglaras C (2004) Contact centers with a call-back option and real-time delay information. Oper Res 52(4):527–545

Armony M, Maglaras C (2004) On customer contact centers with a call-back option: customer decisions, routing rules, and system design. Oper Res 52(2):271–292

Bailey ED, Sweeney T (2003) Considerations in establishing emergency medical services response time goals. Prehosp Emerg Care 7(3):397–399

Bassamboo A, Randhawa RS, Van Mieghem JA (2010) Optimal flexibility configurations in newsvendor networks: going beyond chaining and pairing. Manag Sci 56(8):1285–1303

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

Blaesi D (2015) Customer callback reward system and method, February 17. US Patent 8,958,538

Cezik MT, L'Ecuyer P (2008) Staffing multiskill call centers via linear programming and simulation. Manag Sci 54(2):310–323

Dai JG, He S (2012) Many-server queues with customer abandonment: a survey of diffusion and fluid approximations. J Syst Sci Syst Eng 21(1):1–36

Deslauriers A, L'Ecuyer P, Pichitlamken J, Ingolfsson A, Avramidis AN (2007) Markov chain models of a telephone call center with call blending. Comput Oper Res 34(6):1616–1645

Dudin S, Kim C, Dudina O, Baek J (2013) Queueing system with heterogeneous customers as a model of a call center with a call-back for lost customers. Math Prob Eng 2013:1–13

Gans N, Koole G, Mandelbaum A (2003) Telephone call centers: tutorial, review, and research prospects. Manuf Serv Oper Manag 5(2):79–141

Gans N, Zhou Y-P (2003) A call-routing problem with service-level constraints. Oper Res 51:255–271

Helber S, Henken K (2010) Profit-oriented shift scheduling of inbound contact centers with skills-based routing, impatient customers, and retrials. OR Spectr 32(1):109–134

Kim C, Dudina O, Dudin A, Dudin S (2012) Queueing system map/m/n as a model of call center with call-back option. Anal Stoch Model Tech Appl 7314:1–15

Kleinrock L (1975) Queueing systems, theory, vol I. A Wiley-Interscience Publication, Hoboken

Koole G, Mandelbaum A (2002) Queueing models of call centers: an introduction. Ann Oper Res 113(1):41–59

Koole G, Nielson BF, Nielson TB (2012) First in line waiting times as a tool for analysing queueing systems. Oper Res 60(5):1258–1266

Legros B, Jouini O, Dallery Y (2015) A flexible architecture for call centers with skill-based routing. Int J Prod Econ 159:192–207

Legros B, Jouini O, Koole G (2016) Optimal scheduling in call centers with a callback option. Perform Eval 95:1–40

Legros B, Jouini O, Koole G (2017) A uniformization approach for the dynamic control of multi-server queueing systems with abandonments. Oper Res (To appear)

Legros B (2016) Unintended consequences of optimizing a queue discipline for a service level defined by a percentile of the waiting time. Oper Res Lett 44(6):839–845

Liao S, Koole G, Van Delft C, Jouini O (2012) Staffing a call center with uncertain non-stationary arrival rate and flexibility. OR Spectr 34(3):691–721

Livanos K (1994) Automatic customer call back for automatic call distribution systems, May 10. US Patent 5,311,574

Mandelbaum A, Shimkin N (2000) A model for rational abandonments from invisible queues. Queueing Syst 36(1):141–173

Mandelbaum A, Zeltyn S (2004) The impact of customers patience on delay and abandonment: some empirically-driven experiments with the M/M/n+G queue. OR Spectr 26(3):377–411

Metcalf M (2006) Callback service, February 16. US Patent App. 11/307,677

Pang G, Perry O (2014) A logarithmic safety staffing rule for contact centers with call blending. Manag Sci 61(1):73–91

Rafter J, Lewis DC, Rawle JD, Irwin M, Artemieff S (2010) Method and system for scheduling a customer service callback, July 20. US Patent 7,761,323

Robbins TR, Harrison TP (2010) A stochastic programming model for scheduling call centers with global service level agreements. Eur J Oper Res 207(3):1608–1619

Shumsky RA (2004) Approximation and analysis of a call center with flexible and specialized servers. OR Spectr 26(3):307–330

Stolletz R, Helber S (2004) Performance analysis of an inbound call center with skills-based routing. OR Spectr 26(3):331–352