A dynamic prioritization policy for the callback option in a call center

Balcioglu, B.; Kanavetas, O.

Flexible Services and Manufacturing Journal
https://doi.org/10.1007/s10696-021-09413-y

A dynamic prioritization policy for the callback option in a call center

Barış Balcıoğlu[1] · **Odysseas Kanavetas**[2]

Accepted: 19 March 2021

Abstract

In this paper, we study the $M_n/M_n/c/(K_1 + K_2) + M_n$ system with two finite-size queues where underlying exponential random variables have state-dependent rates. When all servers are busy, upon arrival customers may join the online or the offline/callback queue or simply balk. Customers waiting in the online queue are impatient and if their patience expires, they may choose to join the callback queue instead of abandoning the system for good. Customers in the callback queue are assumed to be patient. Customers are served following a threshold policy: when the number of customers in the callback queue surpasses a threshold level, the next customer to serve is picked from here. Otherwise, only after a predetermined number of agents are reserved for future arrivals, customers remaining in the callback queue can be served. We conduct an exact analysis of this system and obtain its steady-state performance measures. The times spent in both queues are expressed as Phase-type distributions. With numerical examples, we present how the policy responds when shorter callback times are promised or customer characteristics vary.

Keywords The $M_n/M_n/c/(K_1 + K_2) + M_n$ queue · Impatient customers · Callback option · Phase-type distribution · Call centers

✉ Barış Balcıoğlu
baris.balcioglu@sabanciuniv.edu

Odysseas Kanavetas
o.kanavetas@math.leidenuniv.nl

[1] Faculty of Engineering and Natural Sciences, Sabancı University, Orhanlı-Tuzla, 34956 Istanbul, Turkey

[2] Mathematical Institute, Leiden University, Niels Bohrweg 1, 2333 CA Leiden, The Netherlands

Published online: 31 March 2021

1 Introduction

Providing high service levels with limited workforce has been the major challenge in customer contact/call centers. The stress is felt more when callers are impatient, who can hang up and be lost if their queueing time turns out to be longer than their patience time. As a remedy to this problem, call centers may offer a callback option to customers. Such an offer appears more practical and appealing nowadays because most customers use smart (or cellular) phones instead of landline phones. With smart phones, unless the caller is shortly going to join an environment where she will not be able to pick up her phone, waiting to be called back imposes less inconvenience when compared to a caller that had to wait sitting next to a landline phone to receive the same callback service. There is also empirical evidence such as the one from Software Advice[1] showing that nearly two thirds of the callers prefer the callback option instead of waiting on the line.

Obviously, a successfully implemented callback policy leads to fewer customer loss and this, in return, increases the proportion of time customer representatives work. For such a policy to be successful, customers must be announced a reasonably long (or maybe short) time window during which they would be called back, possibly coupled with some delay information regarding the online queue. A callback during the announced time window can be realized if the callback customers can gain priority over those waiting on the line from time to time. Such a dynamic prioritization is possible following a threshold policy. Therefore,—if feasible—the call center management identifies the threshold level and a customer representative completing a service calls the longest awaiting callback customer if the number of customers waiting in this queue is more than or equal to this threshold level. This would help a desired proportion (e.g., 90%) of those customers be called back within the announced time period. Otherwise, i.e., if the number of the callback customers is fewer than the threshold level, the call of the longest awaiting customer on the line—if there is any—is picked up. While the first-come first-served (FCFS) policy arises naturally for the online queue implementing the FCFS policy for the callback queue is also easier, in practice, than scheduled call returns. It is also what 67% of callers choosing the callback option expect from the call centers according to a survey conducted by ContactBabel (2016).

To analyze this policy, in this paper, we model the setting as a queueing system with two queues, one representing the online and the other the callback customers, where customers in either queue can be dynamically prioritized using the threshold policy described above. The idea of using the threshold policy to do such a dynamic prioritization is due to Armony and Maglaras (2004a, 2004b). However, there are differences in the problem settings we study. Armony and Maglaras assume that the callback option is offered only to a newly arriving call. They also assume that those callers not opting for it and choosing to wait on the line are patient to be eventually served. We, on the other hand, assume that online customers are impatient. In

[1] https://www.softwareadvice.com/resources/3-ways-to-offer-callback/

addition to the newly arriving customers, online customers, upon expiry of their patience, can also request to be called back. Regarding the customer characteristics as summarized here, our setting is more similar to that modeled by Brandt and Brandt (1999). Yet, Brandt and Brandt, like many of the callback models to be summarized below, stipulate that the online customers are always the high-priority customers. In their model, the callback customers can be served only when there are no more online customers waiting and after a certain number of agents/customer representatives become idle and are reserved for future callers. Thus, they design a server reservation policy.

We include the server reservation idea from Brandt and Brandt in our study as well. That is to say,—if feasible—the call center management can determine how many agents to reserve for future callers: when the online queue is empty and the number of awaiting callback customers is below the threshold level, only after the determined number of agents start idling while waiting for future calls, any remaining callback customers can be served by the unreserved agents.

Our contribution in this paper is to combine the threshold and server reservation policies from the literature in a queueing system with impatient online and patient offline/callback customers where the latter over the former can be dynamically prioritized for service from time to time. We assume that underlying exponential interarrival, service, and patience times have state-dependent rates. This makes the model available to approximately analyze the cases with general service and impatience times if, as pointed out by Kanvetas and Balcıoğlu (2018), future research can successfully map the characteristics of the original general distributions on the state-dependent rates of the approximating exponential random variables.

We obtain the steady-state performance measures of this queueing system such as the proportion of customers lost (those hanging up right away or abandoning from the online queue for good) and the queueing time distributions in both queues in Sects. 2 and 3 , respectively. This enables us to design a problem in a numerical study in Sect. 4.1 where the proportion of lost customers is minimized while at least a given proportion of the callback customers—if feasible—are called back within the announced time window. With narrower time windows, we see that server reservation policy is usually not possible and the threshold level to prioritize the callback queue decreases. Yet, with shorter time windows announced, if more customers can be induced to opt for the callback, the system starts losing fewer customers without degrading the online customer experience seriously. In Sect. 4.2, via another numerical example, we show that the proposed policy can be robust even if arrival and abandonment rates can change significantly.

The potential benefits of offering the callback option have led many call centers to provide their customers with this service. The 2016 ContactBabel states that 39% of the surveyed companies offer the callback option. This has also ignited a recent academic research interest on the topic as well. Kim et al. (2012) consider impatient online customers who can also request to be called back when their patience expires. They assume Markovian arrival process to capture the bursty nature of the call arrivals. However, they suggest that the callback customers can be served only if there are no online customers left. Dudin et al. (2013) consider any lost customer (directly hanging up or abandoning) a callback customer. The customer representatives are

grouped into two. Those whose primary customers are online/callback customers can serve callback/online customers only if there are no more customers in the queue they are assigned to. When to offer the callback option is also investigated by some researchers. Legros et al. (2016) suggest offering this option only to new customers when the number of online customers is above a threshold level. However, if such a customer does not choose to be called back, she cannot use this service if her patience expires later on. The authors also examine the server reservation policy: Depending on the number of customers in the offline queue, the number of busy customer representatives, and the type of job in service, the number of representatives to be reserved for the online customers is determined. In this paper and also in Legros et al. (2017), the online customers have non-preemptive priority over those in the offline queue. Legros et al. design a different postponed callback offer scheme. Only the first customer in the online queue, after having waited for a certain amount of time, hears the callback option. They obtain the performance measures of interest assuming finite size queues when they consider impatience for the online customers. Ata and Peng (2017) employ a look-ahead policy to determine which customer to offer a callback option. Due to the cost structure, they give preemptive-resume priority to the online customers. Yom-Tov and Zeitler (2018) point out the importance of the delay guarantee to make the callback offer appealing to callers and they employ a simulation-based approach to determine the appropriate delay guarantee depending on the workload over the day. They assume that the callback decision is taken only by new arrivals. In their system, the callback customer for whom the delay guarantee is about to be violated gains priority and is taken into service.

Another setting where customer representatives serve two queues is when there is call blending. That is, in addition to inbound calls, from time to time, agents try to make outbound calls. The major difference in this setting, in comparison to ours, is that the population of the outbound customers is infinite and there is no service guarantee considered for them. The goal could be maximizing the outbound call rate as in Bhulai and Koole (2003) and Gans and Zhou (2003) while there could be service levels for the inbould calls. Then, the agent reservation policy can be implemented, that is, when the number of agents reserved for the inbound calls reaches a threshold, an agent can make an outbound call or respond to an email request. We refer the interested reader for further reading on call blending to Deslauriers et al. (2007), Pang and Perry (2014), Legros et al. (2015, 2020).

In comparison to the settings in the papers considering the callback option, the call center we model in this study provides the callback option not only to new callers but also to the impatient online customers. The service policy can reserve agents for future arrivals but the fundamental aspect is the threshold policy that dynamically prioritizes the callback customers in order to call them back within the announced time window.

We obtain the steady-state system size distribution of the underlying queueing system by constructing an appropriate continuous-time Markov chain (CTMC). With this distribution in hand, we are able to express the proportions of customers who hang up, abandon, get served, or request to be called back. Then, we revert our attention to obtaining the queueing time distributions and their moments. For both customer types in the online and callback queues, after representing a customer's

queueing position via three dimensional CTMC's, we express the queueing time distributions as Phase-type distributions (PHD). With these, the call center management can decide on what to announce to customers upon their arrival and the lengths of time they may need to wait if they choose the callback option. These decisions help the management to minimize the proportion of customers lost/to maximize the throughput of the call center.

The paper is organized as follows. In Sect. 2, we conduct the exact analysis of the underlying queueing system to obtain its steady-state performance measures. In Sect. 3, we obtain the queueing time distributions for both the online and callback customers. The numerical examples presented in Sect. 4 demonstrate how the callback option improves the throughput of customers reached through two mediums, the online and the callback queues. Sect. 5 presents conclusions and possible directions for future research.

2 The exact analysis of the $M_n/M_n/c/(K_1+K_2)+M_n$ queueing system

In this section, we analyze the $M_n/M_n/c/(K_1+K_2)+M_n$ queueing system to obtain its steady-state system size distribution alongside some frequently employed performance measures. This system comprises c parallel and identical servers and two finite-size queues, that is, the *online/primary* and the *offline* queues that can accommodate at most K_1 and K_2 customers, respectively. As a subsystem, the *online system* refers to the servers (or customers being served) and (customers waiting in) the online queue. Customers arrive according to a Poisson process with rate λ_0, and join one of the queues or balk right away depending on the number of customers in the online system and offline queue, denoted by n and m, respectively. If customers, when they call, are informed about the number of customers waiting in the online (as we recommend in Sect. 4.1) and offline queues, finite K_1 and K_2 arise as a result of their equilibrium joining strategy. The longer the queue(s) they see upon arrival, the less likely the customers will join the related queue(s), and thus, one can have finite queue sizes. We also need them to be finite to obtain the queueing time distributions in Sect. 3.

It is reasonable to assume that $\lambda_{n,m}=\lambda_0$ when there are $n<c$ customers being served. When all servers are busy and there are $j=0,1,\ldots,K_1$ queued customers in the online queue and $m=0,1,\ldots,K_2$ customers in the offline queue, with rate $\lambda_{c+j,m}/\lambda'_{c+j,m}$, customers join the online/offline queue or balk right away from the system with rate $\lambda_0-(\lambda_{c+j,m}+\lambda'_{c+j,m})$. Obviously, $\lambda_{c+K_1,m}=0$ for all m.

Remark 1 When the offline queue is full with K_2 customers waiting in it, for all n, customers balk with rate $\lambda_0-\lambda_{n,K_2}$, which includes λ'_{n,K_2}.

In the context of a call center, the online queue corresponds to callers waiting on the line for an agent whereas the offline queue to those customers that have requested to be called back. While customers in the offline queue are patient, each

customer waiting in the online queue has a random patience time. If its queueing time exceeds this random variable (r.v.), the ith queued customer among $k(= i, \ldots, K_1)$ customers in the online queue can either renege and leave the system without receiving service (which we simply refer to as abandonment/abandoning the system) with rate $\delta_{k,i,m}$ or join the offline queue with rate $\delta'_{k,i,m}$.

Remark 2 If the offline queue is full, a customer whose patience for waiting expires reneges without receiving service, that is, abandons, with rate $\delta_{k,i,K_2} + \delta'_{k,i,K_2}$.

"Successful" customers are those who reach one of the servers either upon their arrival or before their patience time expires while waiting in the online queue. Once their service starts, customers stay in the system until their service finishes. Upon completion of a service, the server picks up the longest awaiting customer in the offline queue if there are at least $T(= 1, \ldots, K_2 + 1)$ customers here. If the number of customers in the offline queue is fewer than T and there are queued customers in the online queue, the longest awaiting customer in the latter is picked for service. Although the service priority between the two queues changes depending on the size of the offline queue being less than T or not, within each queue customers are served according to the FCFS policy. If the online queue is empty, even though there may be (fewer than T) customers waiting in the offline queue, the server becomes idle unless this would reduce the total number of busy servers to $a - 1$, $(0 \le a - 1 \le c - 1)$. If the offline queue is empty, obviously the server status changes to idle. When the online queue is empty and the number of busy servers is to be $a - 1$, if there are (fewer than T) customers in the offline queue, one of the $c - a + 1$ idle servers picks the longest awaiting customer here keeping the total number of busy servers at a.

Obviously, this dynamic priority policy for choosing the next customer to serve is quite flexible and by choosing the parameters appropriately, it can turn into a static priority policy. For instance setting $T > K_2$, customers in the offline queue will have to wait until the online queue is empty and $c - a$ idle servers are reserved for future arrivals. With this change, one obtains a server reservation policy similar to that designed by Brandt and Brandt. If $T = 1$, then the customers in the offline queue gain a non-preemptive priority over those in the online queue.

Service and patience time r.v.s are assumed to be exponentially distributed. Moreover, the total rate of service ($\mu_{n,m}$), the total rate of abandonment ($\delta_{n-c,m}$), and the total rate of moving into the offline queue ($\delta'_{n-c,m}$) are,

$$\mu_{n,m} = \begin{cases} \sum_{i=1}^{n} \mu_{n,i,m}, & 1 \le n \le c, \ \ 0 \le m \le T-1, \\ \sum_{i=1}^{c} \mu_{n,i,m}, & c+1 \le n \le c+K_1, \ \ 0 \le m \le K_2, \end{cases}$$

$$\delta_{k,m} = \sum_{i=1}^{k} \delta_{k,i,m}, \quad 1 \le k = (n-c) \le K_1, \quad 0 \le m \le K_2,$$

$$\delta'_{k,m} = \sum_{i=1}^{k} \delta'_{k,i,m}, \quad 1 \le k = (n-c) \le K_1, \quad 0 \le m \le K_2,$$

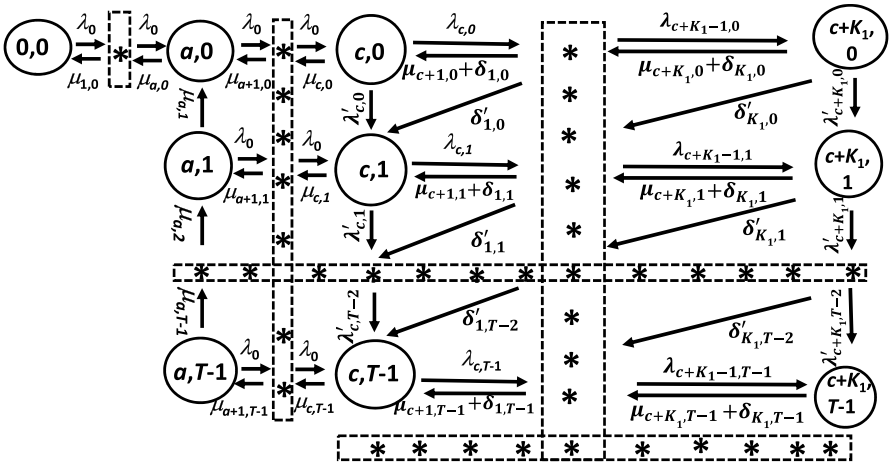

Fig. 1 Part of the two-dimensional CTMC with states when the offline queue has fewer than T customers

where $\mu_{n,i,m}$ is the service rate of the server working on the ith oldest customer taken into service when there are n/m customers in the online system/offline queue. Similarly, $\delta_{k,i,m}/\delta'_{k,i,m}$ is the abandonment rate/the rate of moving into the offline queue of the ith customer in the online queue when there are k and m customers in the online and offline queues, respectively. Thus, information updates on the system size upon departures (either in the form of a service completion or abandonment or moving into the offline queue) or a new customer arrival, each server (queued customer in the online queue) may change its rate of service (rate of abandonment or rate of flow into the offline queue). Another reason to allow these rates to change when the system state changes is the following: if more accurate methods can be devised to map general service and impatience distributions into state-dependent rates for exponential random variables as discussed by Kanavetas and Balcıoğlu, the $M_n/M_n/c/(K_1+K_2)+M_n$ queue with these rates can be employed to approximately analyze the $M_n/GI/c/(K_1+K_2)+GI$ queue.

Let $(N_P(t), N_C(t))$ denote the number of customers in the online/primary system and the offline/callback queue at time t, respectively, which form a two-dimensional CTMC as depicted in Figs. 1 and 2 . Let

$$\pi_{n,m} = \lim_{t\to\infty} P(N_P(t) = n, N_C(t) = m),\ n = 0,\ldots,c+K_1, m = 0,\ldots,K_2,$$

denote the steady-state probability distribution of this CTMC which can be computed following standard techniques. Using them, we can have the steady-state probability of having n customers in the online system (P_n) and m customers in the offline queue (P_m^C) as

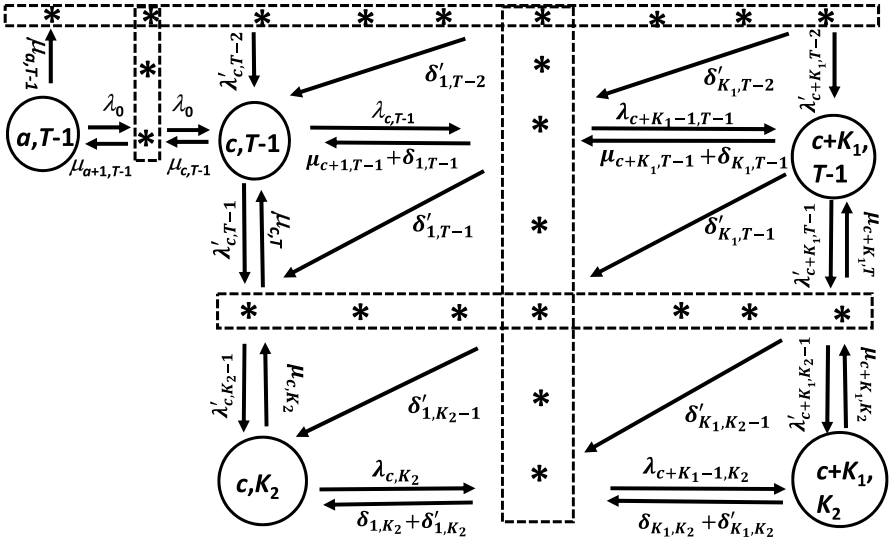

Fig. 2 Part of the two-dimensional CTMC with states when the offline queue has more than $T-2$ customers

$$P_n = \begin{cases} \pi_{n,0}, & 0 \le n \le a-1, \\ \sum_{m=0}^{T-1} \pi_{n,m}, & a \le n < c, \\ \sum_{m=0}^{K_2} \pi_{n,m}, & c \le n \le c+K_1, \end{cases}$$
$$P_m^C = \begin{cases} \sum_{n=0}^{c+K_1} \pi_{n,0}, & m = 0, \\ \sum_{n=a}^{c+K_1} \pi_{n,m}, & 1 \le m \le T-1, \\ \sum_{n=c}^{c+K_1} \pi_{n,m}, & T \le m \le K_2. \end{cases} \tag{1}$$

Then, the mean server utilization can be computed as

$$U = \frac{\sum_{n=1}^{c} nP_n + c\sum_{n=c+1}^{c+K_1} P_n}{c}. \tag{2}$$

Let S be the event that a newly arriving customer receives service without moving into the offline queue, A the event that she eventually abandons, and C the event that she is served after waiting in the offline queue (where she may have joined possibly after spending some time in the online queue). Then, recalling Remarks 1 and 2 as well, the probability that a customer balks instead of joining either one of the queues when all servers are busy is

$$P_B = \frac{\sum_{n=c}^{c+K_1}\sum_{m=0}^{K_2-1}(\lambda_0 - \lambda_{n,m} - \lambda'_{n,m})\pi_{n,m} + \sum_{n=c}^{c+K_1}(\lambda_0 - \lambda_{n,K_2})\pi_{n,K_2}}{\lambda_0}, \tag{3}$$

that she eventually abandons is

$$P_A = \frac{\sum_{n=c+1}^{c+K_1} \sum_{m=0}^{K_2-1} \delta_{n-c,m}\pi_{n,m} + \sum_{n=c+1}^{c+K_1} (\delta_{n-c,K_2} + \delta'_{n-c,K_2})\pi_{n,_{K_2}}}{\lambda_0}. \tag{4}$$

Therefore, we compute the probability that a customer that cannot join a full offline queue when she attempts to join it as follows:

$$P_{RCB} = \frac{\sum_{n=c}^{c+K_1} \lambda'_{n,K_2}\pi_{n,_{K_2}} + \sum_{n=c+1}^{c+K_1} \delta'_{n-c,K_2}\pi_{n,_{K_2}}}{\lambda_0}. \tag{5}$$

Observe that one part of the P_{RCB} due to λ'_{n,K_2} is included in P_B and the remaining part is included in P_A. The probability that a customer ever waits in the offline queue until service can be computed as

$$P_C = \frac{\overline{\lambda}_C}{\lambda_0},$$

where

$$\overline{\lambda}_C = \sum_{m=0}^{K_2-1} \left[\sum_{n=c}^{c+K_1} \lambda'_{n,m}\pi_{n,m} + \sum_{n=c+1}^{c+K_1} \delta'_{n-c,m}\pi_{n,m} \right], \tag{6}$$

is the mean rate of customers joining the offline queue. With these probabilities, the remaining probability that a customer is successful can be computed as

$$P_S = 1 - P_A - P_B - P_C. \tag{7}$$

Letting W_S denote the queueing time r.v. for a successful customer, the probability that such a customer does not wait before service is

$$P(W_S = 0) = \frac{\sum_{n=0}^{c-1} \lambda_0 P_n}{\lambda_0 P_S} = \frac{\sum_{n=0}^{c-1} P_n}{P_S}. \tag{8}$$

Before we look into the queueing time distributions more in detail in Sect. 3, observe that with P_m^C as provided in Eq. (1), we can compute $E[N_C^r]$, that is the rth moment of the number of customers in the offline queue. Invoking the Little's law, we can compute the mean waiting time spent here as

$$E[W_C] = \frac{E[N_C]}{\overline{\lambda}_C}. \tag{9}$$

Waiting time distribution and its moments in the offline queue can be alternatively computed with the analysis presented in Sect. 3.2.

3 Queueing time distributions in the online and offline queues

We first characterize the distribution and moments of the time spent in the online queue in Sect. 3.1. In addition to W_S introduced in Sect. 2, we introduce W and W_{AC} denoting queueing time r.v.s of an arbitrary customer in the online queue and a customer leaving here (because of abandonment or moving into the offline queue), respectively and we obtain the distributions for all three. Then, we obtain the distribution and moments of the time spent in the offline queue, denoted by W_C in Sect. 3.2.

3.1 Queueing time distribution in the online queue

Since an arbitrary customer in the online queue (with the online queue time r.v W) is either a successful customer (with the online queue time r.v. W_S) or a customer who leaves the online queue by eventually abandoning or becoming a callback customer (with the online queue time r.v. W_{AC}), employing the law of total probability, one has the following relationships for their complementary queueing time distributions and their rth moments, respectively:

$$\begin{aligned} P(W \geq w) =& Q_S P(W_S \geq w) + (1 - Q_S) P(W_{AC} \geq w), \\ E[W^r] =& Q_S E[W_S^r] + (1 - Q_S) E[W_{AC}^r], \end{aligned} \tag{10}$$

where Q_S is the probability that a customer, out of those who join the online system, is successful, and is computed as

$$Q_S = \frac{P_S}{1 - P_B - P_{CD}},$$

with P_B given in Eq. (3) and P_{CD}, i.e., the probability that a customer upon arrival moves into the offline queue without ever entering the online queue, being

$$P_{CD} = \frac{\sum_{n=c}^{c+K_1} \sum_{m=0}^{K_2-1} \lambda'_{n,m} \pi_{n,m}}{\lambda_0}.$$

We first obtain the statistics of W for which we employ the r.v.s N_P^α and N_C^α denoting, respectively, the number of customers found in the online system and offline queues upon arrival by a customer joining the online system with the following steady-state distribution:

$$P(N_P^\alpha = n, N_C^\alpha = m) = \frac{\lambda_{n,m} \pi_{n,m}}{\overline{\lambda}_P}, n = c, \ldots, c + K_1 - 1, m = 0, \ldots, K_2,$$

where

$$\overline{\lambda}_P = \sum_{n=0}^{c+K_1} \sum_{m=0}^{K_2} \lambda_{n,m} \pi_{n,m} = \lambda_0 (1 - P_B - P_{CD}),$$

is the mean arrival rate at the online system. Considering all the possible numbers of customers in the online system (n) and in the offline queue (m) that an arrival to join the online queue can find and their corresponding probabilities, we invoke the law of total probability to write

$$\begin{aligned} P(W \geq w) &= \sum_{n=c}^{c+K_1} \sum_{m=0}^{K_2} P(W \geq w | N_P^{\alpha} = n, N_C^{\alpha} = m) \frac{\lambda_{n,m} \pi_{n,m}}{\overline{\lambda}_P}, \\ E[W^r] &= \sum_{n=c}^{c+K_1} \sum_{m=0}^{K_2} E[W^r | N_P^{\alpha} = n, N_C^{\alpha} = m] \frac{\lambda_{n,m} \pi_{n,m}}{\overline{\lambda}_P}. \end{aligned} \tag{11}$$

Consider W_m^{n-c+1}, which is the queueing time r.v. of a "tagged" customer joining the online queue when there are $n - c(\geq 0)/m$ customers in the online/offline queue (in other words, $W_m^{n-c+1} \equiv W | N_P^{\alpha} = n, N_C^{\alpha} = m$). This r.v. is the time spent until absorption in a CTMC (to be referred to as CTMC_P) where absorption occurs if this customer abandons, moves into the offline queue or eventually reaches one of the c parallel servers. This means that W_m^{n-c+1} follows a PHD. The CTMC_P consists of three dimensional states of the form (i, j, k) denoting that the tagged customer is the ith queued customer from the head of the online queue when there are j/k customers in the online system/offline queue. Observe that the rate of leaving the state (i, j, k) is

$$\nu_{j,k} = \mu_{j,k} + \lambda_{j,k} + \lambda'_{j,k} + \delta_{j-c,k} + \delta'_{j-c,k}, \tag{12}$$

and does not depend on i.

W_m^{n-c+1} for the tagged customer starts in state $(n - c + 1, n + 1, m)$ of the CTMC_P. We denote reaching one of the c parallel servers by a state of the form $(0,j,k)$, $c \leq j$, $k = \min\{m, T - 1\}, \ldots, T - 1$, if $T > 1$. If $T = 1$, then $k = 0$ when the customer reaches a server. To characterize the underlying PHD (see, e.g. Altıok 1997, Section 2.8), we need

- the $L \times L$ $\mathbf{F}$ matrix that lists the transition rates among the transient states where L is the number of transient states
- the $1 \times L$ initial probability vector $\boldsymbol{\alpha}$ which has 1 for the entry corresponding to the starting state $(n - c + 1, n + 1, m)$ and 0's for other transient states.

We next discuss how we determine L, and characterize $\mathbf{F}$ and $\boldsymbol{\alpha}$ under three cases.

When $T = 1$: Recall that this leads to a policy giving non-preemptive priority to the customers in the offline queue. Starting from the state $(n - c + 1, n + 1, m)$, until the tagged customer leaves the online queue (abandoning/moving into offline queue) or moves one step closer to the head of the queue (or she is picked by one of the servers if $n = c$ and $k = 0$), the following transient states $(n - c + 1, j, k)$ can be visited in the CTMC_P, $j = n + 1, \ldots, c + K_1$ and $k = 0, \ldots, K_2$. That is, there are $(K_2 + 1) \times (c + K_1 - n)$ transient states the CTMC_P that can be visited while the customer stays as the $n - c + 1$st queued customer. Other transient states when the customer is the ith queued customer can be identified in the same way and with some algebra, one can determine

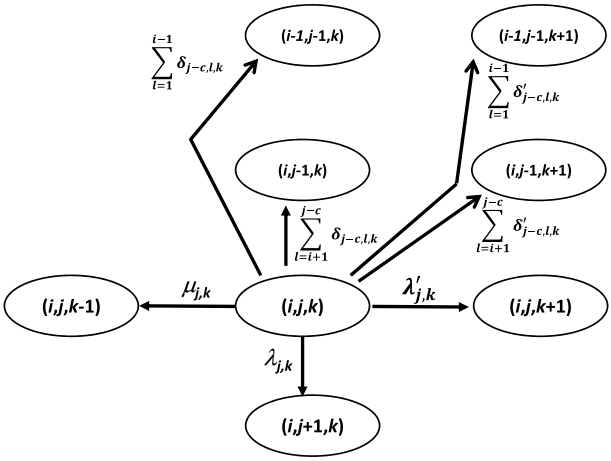

Fig. 3 The transition rate diagram of moving from the transient state (i, j, k) to neighboring transient states when $K_1 > j - c > i > 1, 0 < k < K_2$

$$L = [(n - c + 1)K_1 - (n - c + 1)(n - c)/2] \times (K_2 + 1) = A \times B.$$

Indexing states $(1, c + 1, 0), \ldots, (1, c + 1, K_2), (1, c + 2, 0), \ldots$ as the 1st, $\cdots$ $K_2 + 1$st, $K_2 + 2$nd $\cdots$, states, the initial probability vector $\boldsymbol{\alpha}$ has 1 in its $[A - (K_1 - n + c)] \times B + (m + 1)$st entry.

The **F** matrix can be constructed using Fig. 3, which is an example demonstrating how one can move to the neighboring transient states of the transient state (i, j, k) in the CTMC$_P$. Otherwise, with rate $\delta_{j-c,i,k}/\delta'_{j-c,i,k}$, the process can end in the absorbing states of abandonment/moving into the offline queue or with rate $\mu_{c+1,0}$ (when $i = 1$ and $k = 0$) in the absorbing state of reaching a server. When $k = 0$, $\mu_{j,0}$ is added to $\sum_{l=1}^{i-1} \delta_{j-c,l,0}$ yielding the rate with which the CTMC$_P$ in Fig. 3 moves from state $(i, j, 0)$ to state $(i - 1, j - 1, 0)$.

When $T > 1$, $m < T$**:** Starting from the state $(n - c + 1, n + 1, m)$, until the tagged customer moves into an absorbing state or one step ahead in the queue, there are $(K_2 - m + 1) \times (c + K_1 - n)$ transient states as $(n - c + 1, j, k)$, $j = n + 1, \ldots, c + K_1$ and $k = m, \ldots, K_2$ that can be visited. The transient states that can be visited when the customer occupies the ith position in the online queue can be identified similarly. Then, one can calculate

$$L = [(n - c + 1)K_1 - (n - c + 1)(n - c)/2] \times (K_2 - m + 1) = A \times B'.$$

Indexing states $(1, c + 1, m), \ldots, (1, c + 1, K_2), (1, c + 2, m), \ldots$ as the 1st, $\cdots$ $K_2 - m + 1$st, $K_2 - m + 2$nd $\cdots$, states, the $\boldsymbol{\alpha}$ vector has 1 in its $[A - (K_1 - n + c)] \times B' + 1$st entry.

The **F** matrix can be constructed as explained in Fig. 3 with the following change: When $k < T$, $\mu_{j,k}$ is added to $\sum_{l=1}^{i-1} \delta_{j-c,l,k}$ and this gives the rate that one

moves from state (i, j, k) to state $(i-1, j-1, k)$ in the CTMC$_P$ in Fig. 3. Otherwise (when $k \geq T$), with rate $\mu_{j,k}$, one moves from state (i, j, k) to state $(i, j, k-1)$.

When $T > 1$, $m \geq T$: Starting from the state $(n-c+1, n+1, m)$, until the tagged customer moves into an absorbing state or one step ahead in the queue, there are $(K_2 - T + 2) \times (c + K_1 - n)$ transient states as $(n-c+1, j, k)$, $j = n+1, \ldots, c+K_1$ and $k = T-1, \ldots, K_2$ that can be visited. The transient states of $CTMC_P$ when the customer is in the ith position in the online queue can be identified similarly and one can calculate

$$L = [(n-c+1)K_1 - (n-c+1)(n-c)/2] \times (K_2 - T + 2) = A \times B''.$$

Indexing states $(1, c+1, T-1), \ldots, (1, c+1, K_2), (1, c+2, T-1), \ldots$ as the 1st, $\cdots$ $K_2 - T + 2$nd, $K_2 - T + 3$rd $\cdots$, states, the $\boldsymbol{\alpha}$ vector has 1 in its $[A - (K_1 - n + c)] \times B'' + (m - T + 2)$nd entry.

The $\mathbf{F}$ matrix can be constructed as explained in the previous case.

With the $L \times L$ identity matrix $\mathbf{I}$, and the $L \times 1$ vector $\mathbf{1}$ of 1's, for W_m^{n-c+1}, we have the moments

$$E[(W_m^{n-c+1})^r] = E[W^r | N^\alpha = n, N_C^\alpha = m] = (-1)^r r! \boldsymbol{\alpha} \mathbf{F}^{-r} \mathbf{1}, \tag{13}$$

and the complementary distribution

$$P(W_m^{n-c+1} \geq w) = P(W \geq w | N^\alpha = n, N_C^\alpha = m) = \boldsymbol{\alpha} \exp(\mathbf{F}w)\mathbf{F}\mathbf{1}, \tag{14}$$

with which, via Eq. (11), we can obtain $P(W \geq w)$ and $E[W^r]$.

When it comes to W_S, i.e., the queueing time of a successful customer, with parallel reasoning to the derivation of Eq. (11), we can write

$$P(W_S \geq w) = \sum_{n=c}^{c+K_1} \sum_{m=0}^{K_2} P(W_S \geq w | N^\alpha = n, N_C^\alpha = m, S) P(N^\alpha = n, N_C^\alpha = m | S),$$

where

$$P(N^\alpha = n, N_C^\alpha = m | S) = \frac{P(S | N^\alpha = n, N_C^\alpha = m) P(N^\alpha = n, N_C^\alpha = m)}{Q_S},$$

for which we need to compute $P(S | N^\alpha = n, N_C^\alpha = m)$.

Let $q_{i,j,k}$ denote the probability that the ith customer from the head of the online queue will eventually be served (without flowing into the offline queue) when there are a total of j/k customers in the online system/offline queue. Then, with $q_{0,j,k} = 1$, and considering the possible one-step transitions (e.g., Fig. 3) for the ith customer in the online queue when there are j/k customers in the online system/offline queue, we have

$$P(S | N^\alpha = n, N_C^\alpha = m) = q_{n-c+1,n+1,m},$$

and for $i \geq 1, j \geq c, k < T, T > 1$, with $\nu_{j,k}$ as given in Eq. (12)

$$
\begin{aligned}
q_{i,j,k} =& \frac{\mu_{j,k} + \sum_{l=1}^{i-1} \delta_{j-c,l,k}}{\nu_{j,k}} q_{i-1,j-1,k} + \frac{\sum_{l=1}^{i-1} \delta'_{j-c,l,k}}{\nu_{j,k}} q_{i-1,j-1,k+1} + \frac{\lambda_{j,k}}{\nu_{j,k}} q_{i,j+1,k} \\
&+ \frac{\lambda'_{j,k}}{\nu_{j,k}} q_{i,j,k+1} + \frac{\sum_{l=i+1}^{j-c} \delta_{j-c,l,k}}{\nu_{j,k}} q_{i,j-1,k} + \frac{\sum_{l=i+1}^{j-c} \delta'_{j-c,l,k}}{\nu_{j,k}} q_{i,j-1,k+1}.
\end{aligned} \tag{15}
$$

When $T = 1$ and $k > 1$ or when $k \geq T > 1$, Eq. (15) slightly changes and instead of the first expression on the RHS, the following needs be substituted:

$$
\frac{\mu_{j,k}}{\nu_{j,k}} q_{i,j,k-1} + \frac{\sum_{l=1}^{i-1} \delta_{j-c,l,k}}{\nu_{j,k}} q_{i-1,j-1,k}.
$$

Combining all these we have

$$
P(N^{\alpha} = n, N_C^{\alpha} = m|S) = \frac{\lambda_{n,m}\pi_{n,m}q_{n-c+1,n+1,m}}{\overline{\lambda}_P Q_S} = \frac{\lambda_{n,m}\pi_{n,m}q_{n-c+1,n+1,m}}{\lambda_0 P_S}. \tag{16}
$$

Now let us consider $W_S^{n-c+1,m}$, which is the queueing time r.v. of a tagged "successful" customer joining the online queue when there are $n - c(\geq 0)/m$ customers in the online/offline queue. This r.v. has also a PHD since it is the time spent until absorption in one of the states $(0,j,k)$, $c \leq j$, $(k = \min\{m, T-1\}, \ldots, T-1$, if $T > 1$ and $k = 0$ if $T = 1)$ in which the customer reaches a server in a three-dimensional CTMC that has the same transient states as those in CTMC_P. The difference is that this customer neither abandons nor moves into the offline queue. Thus, in addition to $\mathbf{F}$ and $\boldsymbol{\alpha}$ introduced earlier, we need another $L \times 1$ vector $\boldsymbol{\alpha}_S$ that has 0's as entries except for $\mu_{j,k}$ for entries corresponding to states when $i = 1$ and $k < T$. In other words, this is the vector of transition into the absorbing state (i.e., reaching one of the servers) for a successful customer.

Then, one can show that

$$
E[(W_S^{n-c+1,m})^r] \equiv E[W_S^r|N^{\alpha} = n, N_C^{\alpha} = m, S] = \frac{r!\boldsymbol{\alpha}\mathbf{F}^{-(r+1)}\boldsymbol{\alpha}_S}{\boldsymbol{\alpha}(-\mathbf{F})^{-1}\boldsymbol{\alpha}_S}, \tag{17}
$$

which is multiplied by $P(N^{\alpha} = n, N_C^{\alpha} = m|S)$ given in Eq. (16) and summed over for all possible n and m values to compute the rth moment and variance of a successful customer's queueing time as follows:

$$
\begin{aligned}
E[W_S^r] =& \sum_{n=c}^{c+K_1} \sum_{m=0}^{K_2} E[W_S^r|N^{\alpha} = n, N_C^{\alpha} = m, S] \frac{\lambda_{n,m}\pi_{n,m}q_{n-c+1,n+1,m}}{\lambda_0 P_S}, \\
Var[W_S] =& E[W_S^2] - E[W_S]^2.
\end{aligned} \tag{18}
$$

Similarly, after the complementary distribution

$$
P(W_S^{n-c+1,m} \geq w) \equiv P(W_S \geq w|N^{\alpha} = n, N_C^{\alpha} = m, S) = \frac{-\boldsymbol{\alpha}\exp(\mathbf{F}w)\mathbf{F}^{-1}\boldsymbol{\alpha}_S}{\boldsymbol{\alpha}(-\mathbf{F})^{-1}\boldsymbol{\alpha}_S},
$$

is multiplied by $P(N^{\alpha}=n, N_C^{\alpha}=m|S)$ given in Eq. (16) and summed over for all possible n and m values, we can compute the probability that a successful customer's queueing time exceeds w as

$$P(W_S \geq w) = \sum_{n=c}^{c+K_1} \sum_{m=0}^{K_2} P(W_S \geq w | N^{\alpha}=n, N_C^{\alpha}=m, S) \frac{\lambda_{n,m} \pi_{n,m} q_{n-c+1,n+1,m}}{\lambda_0 P_S}. \quad (19)$$

With the distributions and the moments of the queueing time r.v.s W and W_S in hand, those of W_{AC} can be found from Eq. (10).

3.2 Queueing time distribution in the offline queue

The analysis to obtain the distribution of W_C follows similar patterns to that we have presented for W in Sect. 3.1. Consider a tagged customer joining the offline queue as the mth ($m = 1, \ldots, K_2$) customer here while there are n ($n = c, \ldots, c + K_1$) customers in the online system. Let us denote this r.v. with $W_C^{n,m}$ and its rth moment with $E[(W_C^{n,m})^r]$. There are two types of customers who can enter the offline queue to have $W_C^{n,m}$ as its offline queue time r.v.:

(i) a customer finding $n/m-1$ customers in the online system/offline queue upon arrival who immediately moves into the offline queue with rate $\lambda'_{n,m-1}$,
(ii) a customer in the online queue when there are $n+1/m-1$ customers in the online system/offline queue whose patience expires and moves into the offline queue with rate $\delta'_{n+1-c,m-1}$.

Thus,

$$\begin{aligned} P(W_C \geq w) &= \sum_{m=1}^{K_2} \left[\sum_{n=c}^{c+K_1} P(W_C^{n,m} \geq w) \frac{\lambda'_{n,m-1} \pi_{n,m-1}}{\overline{\lambda}_C} \right. \\ &\quad \left. + \sum_{n=c}^{c+K_1-1} P(W_C^{n,m} \geq w) \frac{\delta'_{n+1-c,m-1} \pi_{n+1,m-1}}{\overline{\lambda}_C} \right], \\ E[W_C^r] &= \sum_{m=1}^{K_2} \left[\sum_{n=c}^{c+K_1} E[(W_C^{n,m})^r] \frac{\lambda'_{n,m-1} \pi_{n,m-1}}{\overline{\lambda}_C} \right. \\ &\quad \left. + \sum_{n=c}^{c+K_1-1} E[(W_C^{n,m})^r] \frac{\delta'_{n+1-c,m-1} \pi_{n+1,m-1}}{\overline{\lambda}_C} \right], \end{aligned} \quad (20)$$

where $\overline{\lambda}_C$ is given in Eq. (6)

$W_C^{n,m}$ follows a PHD because it can be represented as the time until absorbtion (i.e., until the tagged customer reaches one of the servers) in a three-dimensional CTMC (to be referred to as CTMC$_C$) where the state (i, k, j) denotes that the tagged customer is the ith queued customer from the head of the offline queue when there are k/j customers in the offline queue/online system. Observe that the

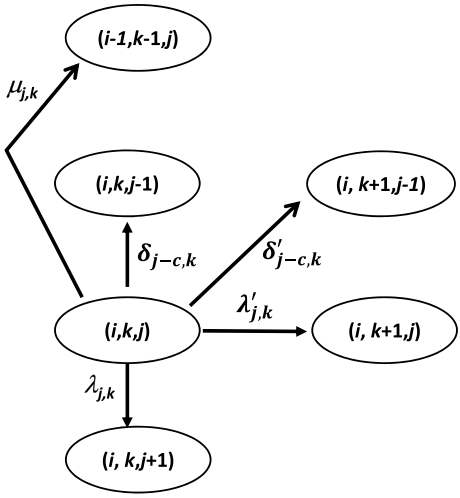

Fig. 4 The transition rate diagram of moving from the transient state (i, k, j) to neighboring transient states when $i > 1, c \leq j < c + K_1$

rate of leaving the transient state (i, k, j) is $\nu_{j,k}$ as given in Eq. (12). The absorption state is represented as $(0,k,j)$, $a \leq j$, $k = 0, \ldots, K_2 - 1$. To characterize the underlying PHD fully, one needs

- the $L_C \times L_C$ $\mathbf{F_C}$ matrix that lists the transition rates among the transient states where L_C is the number of transient states
- the $1 \times L_C$ initial probability vector $\boldsymbol{\alpha_C}$ which has 1 for the entry corresponding to the starting state (m, m, n) and 0's for other transient states.

We next discuss how we determine L_C, and characterize $\mathbf{F_C}$ and $\boldsymbol{\alpha_C}$ under two cases.

When $T = 1$: In this case, customers in the offline queue have non-preemptive priority over those in the online queue. Starting from the state (m, m, n), until the tagged customer moves one step closer to the head of the queue (or she is picked by one of the servers), one can visit states (m, k, j), $j = c, \ldots, c + K_1$ and $k = m, \ldots, K_2$ in the CTMC$_C$. That is, there are $(K_2 - m + 1) \times (K_1 + 1)$ transient states in the CTMC$_C$ that can be visited while the customer stays as the mth queued customer. Other transient states when the customer is the ith queued customer can be identified in the same way and with some algebra, one can determine

$$L_C = [K_2(K_2 + 1) - (K_2 - m)(K_2 - m + 1)] \times (K_1 + 1)/2.$$

Indexing states $(1, 1, c), \ldots, (1, K_2, c), (1, 1, c + 1), \ldots$ as the 1st, $\cdots$ K_2 th, $K_2 + 1$st $\cdots$, states, the initial probability vector $\boldsymbol{\alpha_C}$ has 1 in its $[K_2(K_2 + 1) - (K_2 - m + 1)(K_2 - m + 2)] \times [(K_1 + 1)/2] + (n - c)(K_2 - m + 1) +$ 1st entry.

Figure 4 is an example demonstrating how one can move to the neighboring transient states of the transient state (i, k, j) in the CTMC$_C$. We order the neighboring states in accordance with the way we index the transient states as explained above. That is, if (i, k, j) is the Mth transient state, $(i, k + 1, j)$ is the $M + 1$st state, $(i, k, j - 1)$

and $(i,k,j+1)$ are the $M-(K_2-k)+1$st and $M+(K_2-k)+1$st transient states, respectively.

When $T>1$**:** Starting with state $(m,\ m,\ n)$, one can visit states $(m,\ k,\ j)$ in the CTMC_C for $j=c,\ldots,c+K_1$ and $k=m,\ldots,K_2$ or for $j=a,\ldots,c-1$ and $k=m,\ldots,T-1$. This means there are $(K_2-m+1)\times(K_1+1)+(c-a)\times(T-m)$ transient states that can be visited in the CTMC_C while the customer stays as the mth queued customer. Other transient states when the customer is the ith queued customer can be identified in the same way and with some algebra, one can determine

$$\begin{aligned}L_C =&[K_2(K_2+1)-(K_2-m)(K_2-m+1)]\times(K_1+1)/2\\&+(c-a)\times[(T-1)T-A'''],/2,\end{aligned}$$

where $A'''=(T-m-1)(T-m)/2$ for $m\le T$ and 0 otherwise.

Indexing states $(1,1,a),\ldots,(1,T-1,a),(1,1,a+1),\ldots,(1,1,c)$ as the 1st, $\cdots$ $T-1$st, Tth $\cdots$, $(c-a)(T-1)+1$st states, the initial probability vector $\boldsymbol{\alpha_C}$ has 1 in its $[K_2(K_2+1)-(K_2-m+1)(K_2-m+2)]\times[(K_1+1)/2]+(n-c)(K_2-m+1)+(c-a)\times[(T-1)T-A''']/2+A^*+1$st entry where

$$A^*=\begin{cases}(n-c)(K_2-m+1), & m\ge T,\\(c-a)(T-m)+(n-c)(K_2-m+1), & m<T,n\ge c,\\(n-a)(T-m), & m<T,n<c.\end{cases}$$

Figure 4 holds true showing how one can move to the neighboring transient states of the transient state $(i,\ k,\ j)$ if $k\ge T$ or $j=a$. Otherwise, with rate $\mu_{j,k}+\delta_{j-c,k}$, the CTMC_C moves from state (i,k,j) to state $(i,k,j-1)$.

Having obtained $\mathbf{F_C}$ and $\boldsymbol{\alpha_C}$, with the $L_C\times L_C$ identity matrix $\mathbf{I}$, and the $L_C\times 1$ vector $\mathbf{1}$ of 1's, Eqs. (13) and (14) can be employed to compute $E[(W_C^{n,m})^r]$ and $P(W_C^{n,m}\ge w)$. Finally, Eq. (20) yields the moments (in specific, an alternative to Eq. 9 for $E[W_C]$) and tail distribution of W_C.

4 The numerical examples

In this section, we first consider a small size call center example in Sect. 4.1, to demonstrate how dynamic prioritization of customers in the offline queue under the threshold policy helps the management improve the system performance while trying to meet the standards announced to callers. We also explore how the policy responds when customers are announced shorter time windows to call back. Later, in Sect. 4.2, in a larger call center, we observe what may happen to the threshold level T and the number of agents $c-a$ reserved for future customers (when the call-back queue has fewer than T customers) if the customer arrival rate increases and/or the customers become more impatient.

In both examples, customers call according to a Poisson process with rate λ_0. A customer reaching a server receives service for an exponential amount of time with rate $\mu=1$. Those who need to wait in the online queue have exponentially

distributed patience times with rate δ. We assume that service and reneging rates are constants since we prefer focusing on the policy without complicating the notation.

Estimating/constructing the arrival rate function depending on the state of a system, which is going to offer the callback option is challenging if prior data does not exist. For instance, if a call center providing no callback option is considering to announce something like "If you request to be called back, we will reach you in 10 minutes" to allure some customers to opt for the callback option, they may be unable to foresee how such a promise can change the behavior of callers. Similarly, announcements can alter the distribution of their patience times as well (see Akşin et al. (2017) for an empirical study on the customer behavior depending on delay announcements). Yet, exploring this type of behavioral change is beyond the scope of our study and deserves a multidisciplinary research effort involving psychology and marketing as well. We refer the reader to Hathaway et al. (2019) who attempt to fill this gap via an empirical study.

In this setting, the first thing the call center has to decide is what to announce new callers that need to wait in the online queue. In Sect. 4.1, we suggest that announcing in which position customers are about to join the online queue be sufficient since it is more difficult to predict other delay statistics and this could also be misleading as to be discussed in that section. In Sect. 4.2, on the other hand, since we would like to focus on the impact of higher λ_0 and δ, customers are solely informed if all agents are busy when they call. In both sections, we assume that announcing callback related delay statistics will not change λ_n and δ that we have when no callback is offered but only $\lambda_n^{'}$ and $\delta^{'}$. Yet, one can use the model, envisioning a wide range of changes in customer behavior, to test whether the resources could serve the goal set (e.g., minimizing the proportion of customers who hang up or abandon) while meeting the service levels announced (to call a customer back in a certain period of time). To this end, we present a small example in Sect. 4.1 where arrival rates may change when the callback option is offered.

4.1 The small call center example

In this example, designed to demonstrate how the policy responds when shorter time windows to call back are announced, five agents ($c = 5$) serve customers and we set $\lambda_0 = 5$ and $\delta = 0.5$. We assume that a customer finding n in the online system joins the online queue with rate

$$\lambda_n = \begin{cases} 5, & 0 \leq n \leq 5, \\ 5\frac{15-n}{10}, & n \geq 6, \end{cases}$$

implying that $K_1 = 10$.

In Table 1, the performance measures of interest are listed in the Performance Measures column. The values these measures take when no callback option is offered are presented in the No Callback column. We see that the proportion of customers lost, $P(Loss) = P_B + P_A = 17\%$, is quite high where P_B and P_A are given in Eqs. (3), (4), respectively. Thus, offering a callback option operating under the proposed threshold policy can be used to minimize $P(Loss)$. Notice that

Table 1 Comparison of the different policies with or without the callback option

Performance Measures	No Callback	$D_c = 10$ $T > K_2$ $a = 4$	$D_c = 8$ $T > K_2$ $a = 5$	$D_c = 6$ $T = 9$ $a = 5$	$D_c = 4$ $T = 7$ $a = 5$	$D_c = 2$ $T = 1$ $a = 5$	$D_c = 2$ and m $T = 1$ $a = 5$
$P(Loss)$	17%	15%	14%	12%	8%	8%	9%
P_C		3%	5%	10%	24%	44%	46%
P_S	83%	82%	81%	78%	68%	48%	45%
U	83%	85%	86%	88%	92%	92%	91%
$P(W_S = 0)$	52%	49%	45%	38%	30%	44%	51%
$P(W_S \geq 1 \mid W_S > 0)$	4.5%	4.5%	4.5%	4.9%	9.2%	25%	26%
$E[W_S \mid W_S > 0]$	0.36	0.36	0.36	0.37	0.45	0.77	0.8
$Var[W_S \mid W_S > 0]$	0.10	0.10	0.10	0.11	0.29	0.89	1
$P(W_C \geq D_c)$		4.6%	2.5%	8.8%	9.8%	11%	10%
$E[W_C]$		3.43	2.2	2.7	2.4	0.9	0.8
$Var[W_C]$		9.9	4.6	4.7	1.5	0.7	0.7

losing customers makes other performance measures to appear somewhat misleadingly good. In the No Callback column, we see that the average server utilization $U = 83\%$ (given in Eq. 2) is low. If we also compute the following statistics for successful customers that wait in the online queue before their service starts,

$$P(W_S \geq w \mid W_S > 0) = \frac{P(W_S \geq w)}{1 - P(W_S = 0)},$$

$$E[W_S^r \mid W_S > 0] = \frac{E[W_S^r]}{1 - P(W_S = 0)},$$

together with Eqs. (8), (19), and (18), we see that for only 4.5% of them the queueing time exceeds the mean service time of 1. Yet, is it reasonable to announce $P(W_S \geq 1 \mid W_S > 0) = 4.5\%$ (which is due to a large proportion of customers lost) as "Those who choose to wait can reach an agent in around 1 minute," (assuming that the unit time is a minute)? Sharing such an information may radically change the customer behavior: all customers may choose to join the online queue and wait patiently anticipating a wait of 1 time unit converting the system to an $M/M/5$ queue with $\rho = 1$. Thus, unless more agents are hired, this announced statistics cannot be realized.

On the other hand, if the management aims at diverting callers to the callback queue, some sort of anticipated delay information has to be shared stating that such requests will be handled within D_c time units. In a probabilistic environment such goals cannot be met 100% of the time unless – practically useless – $D_c = \infty$ is announced. Therefore, for this example, we assume that as long as $P(W_C \geq D_c) \leq 10\%$, the management can consider the promise made to callers to have been met.

Setting $K_2 = 15$ in cases with the callback option results in negligibly small P_{RCB} given in Eq. (5). Since a single D_c is to be announced, we assume that the following rates do not depend on m, namely, the offline queue length at the instant when a customer may decide to join the offline queue,

$$\lambda_n^{'} = \theta(5 - \lambda_n), \quad n = 5, \ldots, 15,$$
$$\delta_{k,i}^{'} = (\theta + 0.1)0.5, \quad k = 1, \ldots, 10, i = 1, \ldots, k.$$

When shorter D_c is announced, we increase θ and through this linear model, we capture the rise in the proportion of customers choosing the callback option. We assume that when their patience expires customers from the online queue are more likely (by 10%) to choose the callback option than those who join the callback queue right away instead of balking.

What happens if D_c is chosen to be 10 time units? D_c =10 may appear as a long time and may not be too convincing for many callers to opt for the callback, thus, we assume that $\theta = 0.1$. That is, 10% of the callers that would hang up and 20% of online customers who would abandon if a callback option were not offered request to be called back if D_c =10. The D_c =10 column in the table lists the results: The policy of Brandt and Brandt, giving non-preemptive priority to the online customers (since $T > K_2$) satisfies the constraint $P(W_C \geq 10) = 4.6\% < 10\%$. The policy reserves one of the servers for future arrivals and callback customers are served by whoever gets idle from the remaining $a = 4$ servers. One can use the alternative policy with $T = 11$ and $a = 4$ (reserving one agent for future calls but from time to time allowing dynamic prioritization of callback customers) that gives the same $P(Loss)$ and even slightly reduces $E[W_C]$ and $Var[W_C]$.

Among all the cases with a callback option to be discussed, D_c =10 case is the only one where the system can reserve one server for future arrivals while never dynamically prioritizing the callback customers. For instance if $D_c = 8$, assuming a higher $\theta = 0.2$, the system can still give non-preemptive priority to the online customers but with $a = 5$, that is reserving no agents. Here, $P(Loss)$ gets lower to 14%. Loosing fewer customers than in the previous cases, as expected, makes servers busier. Fewer customers lost together with a higher proportion of callback customers served, through Eq. (7), decreases P_S as well. One can employ alternative policies yielding the same $P(Loss)$. The policy with $T = 12$ and $a = 5$ slightly reduces $E[W_C]$ and $Var[W_C]$, while the policy with $T = 6$ and $a = 4$ increases them.

When D_c is lowered to 6, we set $\theta = 0.4$, the callback customers need be prioritized when their number is at least (T =)9 and the statistics $P(W_S \geq 1 | W_S > 0)$, $E[W_S | W_S > 0]$, and $Var[W_S | W_S > 0]$ start getting slightly higher than those in the no callback case as listed in $D_c = 6$ column. Note that for D_c =10 and $D_c = 8$, these statistics have stayed the same as in the no callback case while only $P(W_S = 0)$ has got lower. In these two cases, callback customers have lower priority, thus, their impact is negligible for successful customers that need to wait. Their service times shorten the periods over which successful customers could have otherwise found idle servers. In cases with $T < 15$, on the other hand, more callback customers start delaying the successful customers waiting in the online queue and the related statistics start getting higher than those in the no callback case.

For $D_c = 6$, if we attempt to reserve one server for future arrivals by setting $a = 4$, T should get lower to 5. Although this set up also gives 12% as the proportion of lost customers, the statistics for successful customers get worse, e.g., $E[W_S|W_S > 0]$ increases to 0.4.

When $D_c = 4$, we increase θ to 0.7 for which the results are listed in $D_c = 4$ column. Now that 24% of the customers are served from the callback queue, we are not surprised to see that with $T = 7$, prioritization of this queue starts when its length is shorter than for the case with $D_c = 6$. As seen, although the mean server utilization increases, since fewer customers are lost, the statistics for the successful customers get worse.

Offering much shorter callback time may lead any delayed customer to choose the callback option which would again make the mean server utilization 100%. Obviously, in this case, the queue length will explode and the announcement will become a false announcement. However, the system can stop improving even before such a "perfect" announcement is made. Consider $D_c = 2$ for which $\theta = 0.8$ and the results are listed in the second last column. The system can respond to this only by giving non-preemptive priority to the callback customers by setting $T = 1$. Observe that even in this case at its best, $P(W_C \geq 2) = 11\%$ is larger than 10% and this is a violation of the announced delay statistics. Even if this were acceptable, we do not see a further reduction in $P(Loss)$. More people opt for the callback option and for them the statistics get better (e.g., the mean callback time is 0.9) but those for the delayed successful customers get worse when compared to their statistics discussed in the previous columns.

Although, as stated in the introduction of this section, we are not focusing on customer behavioral changes that may affect the arrival rates once the callback option is offered, suppose that in addition to $D_c = 2$, customers are not only informed in which order they would join the online queue but that in the offline queue as well. Let us assume that the arrival rate changes as follows when the callback is offered

$$\lambda_{n,m} = \begin{cases} 5, & 0 \leq n \leq 5, \quad \forall m, \\ 0.8^{(4-m)} 5 \frac{15-n}{10}, & n \geq 6, \quad m = 0, 1, \ldots, 4, \\ \frac{15-n}{10}, & \text{otherwise}, \end{cases}$$

with $\lambda'_{n,m} = 0.8\lambda_{n,m}$. That is, when the customers perceive the offline queue length as short (with less than five callback customers) and considering the $D_c = 2$ information, too, now it is less appealing for them to join the online queue in the second or a later position. For instance, while we had $\lambda_{6,1} = \lambda_6 = 4.5$ with $\lambda'_{6,1} = \lambda'_6 = 0.4$ when only $D_c = 2$ were announced, now we have $\lambda_{6,1} = 2.304$ with $\lambda'_{6,1} = 1.8432$. The results are listed in the last column in Table 1. These results slightly differ with respect to the corresponding ones listed in the column on their left. Although $P(Loss)$ at 9% is higher by 1%, now the constraint $P(W_C \geq 2) = 10\%$ is met.

In summary, the results indicate that due to the flexibility of the dynamic prioritization of the callback customers the proportion of lost customers can reduce from 17% to 8% while increasing the mean server utilization.

4.2 The large call center example

In the previous section, we observe how the proposed policy changes when stricter service levels are promised to customers. In this section, we vary the arrival rate λ_0 and the rate of abandonment δ to observe their impact on the callback policy parameters. To do this, we consider a larger call center example with $c = 100$ agents where callers are only informed if they need to wait on the line and are offered to be called back in $E[W_C] \leq 0.1$, namely, in one tenth of the mean service time. We set $K_1 = 40$ and $K_2 = 100$, which are effectively setting both queue capacities to infinity as verified by the corresponding simulation runs in which these capacities were infinite. However, as the queue capacities get larger, the Matlab code, in which the analytical model is implemented, slows down. We run the code on a Windows-based computer with Intel(R) Core(TM) i5-7200U CPU at 2.50 Hz, and in the example to be presented here, it fails to obtain the queue time distributions derived in Sect. 3 because the PHD matrices needed get too large. Luckily, we are still able to obtain $\pi_{n,m}$ and, thus, $P(Loss) = P_B + P_A$ via Eqs. (3), (4), the server utilization U from Eq. (2), and $E[W_C]$ from Eq. (9) in a couple of seconds. We assume that a customer joins the online queue with rate

$$\lambda_n = \begin{cases} \lambda_0, & 0 \leq n \leq 99, \\ 0.6 \times \lambda_0, & 100 \leq n < 140. \end{cases}$$

In other words, 40% of the customers finding all agents busy hang up if no callback option is offered. If the callback option is offered,

$$\begin{aligned} \lambda_n^{'} &= 0.8(\lambda_0 - \lambda_n), \quad n = 100, \ldots, 140, \\ \delta_{k,i}^{'} &= 0.9 \times \delta, \quad k = 1, \ldots, 40, i = 1, \ldots, k. \end{aligned}$$

In other words, 80%/90% of customers that would hang up/abandon without a callback option would choose this service upon arrival/while quitting the online queue. Note that the choice of these rates is not particularly important. What we aim is to have some probability of loss when no callback option is offered and to prevent the agent utilization from approaching the inhumane 100%, especially when the callback option is available. With this simple setup for determining the rates, we vary λ_0 from 95 to 105 and δ from 0.5 to 2. We do not present the results for the intermediate values because those for the lowest and highest values of these two parameters, as we do in Table 2, would be sufficient for discussion. Here, under each λ_0 value, we list three columns, the first one describing the policy, the second and third ones giving the average agent utilization, U, the probability of lost customers, $P(Loss)$, respectively. The first three rows where the results in percentages are displayed are when customers are more patient with $\delta = 0.5$, whereas the last three rows are for more impatient customers with $\delta = 2$. For each case, that is for each (δ, λ_0) pair, we present the results for three policies. The No Callback policy, as the name implies, shows us the server utilization and the proportion of customers lost when no callback is offered. We see that as λ_0 increases both statistics increase: server utilization from 91% to 95%, and $P(Loss)$ from 4.3% to 9.7%/9.8% when $\delta = 0.5/2$. $P(Loss)$ is much smaller when compared to the same statistics in Table 1. This seems to be the

Table 2 Comparison of the change of callback policy parameters

λ_0	95			105		
δ	Policy	$U\%$	$P(Loss)\%$	Policy	$U\%$	$P(Loss)\%$
0.5	No Callback	91	4.3	No Callback	95	9.7
	$a = 98, T = 4$	93	2.2	$a = 97, T = 4$	98	6.3
	$a = 100, T = 6$	93	2.4	$a = 100, T = 5$	98	6.5
2	No Callback	91	4.3	No Callback	95	9.8
	$a = 97, T = 4$	93	2.2	$a = 97, T = 4$	98	6.7
	$a = 100, T = 7$	93	2.4	$a = 100, T = 5$	98	6.8

case when the number of agents is high as we can see in the numerical examples of Brandt and Brandt, Yom-Tov and Zeitler, etc.

Still, in all the cases, offering a callback option reduces $P(Loss)$ by 2.1% to 3.4%, which is significant. For each case, the optimal policy results are listed in their respective second row. For instance, for ($\delta = 0.5$, $\lambda_0 = 95$), the optimal policy reserves $c - a = 100 - 98 = 2$ agents for future arrivals as long as there are fewer than $T = 4$ customers in the offline queue. When 4 or more customers wait to be called back, an agent finishing a service picks up the next customer from the offline queue. When this policy is implemented, $P(Loss)$ decreases by 2.1% from what it is for the no callback policy (4.3%) to 2.2%. For each case, in their respective third row, we also present the optimal policy when $a = 100$, i.e., when no agent reservation is allowed. In this policy, therefore, T is optimized. Using this sub-optimal policy increases $P(Loss)$ by at most 0.2%, as in, e.g., the ($\delta = 0.5$, $\lambda_0 = 105$) case.

What do we see in the results of Table 2? For the policies with $a = 100$, when δ is fixed, we see that T gets smaller as λ_0 increases. Although the impatience of customers makes the queue dynamics complicated (since an abandoning customer may help the next arrival be served in the online system), in these cases more customers choosing the callback option must have caused a strain on the constraint of $E[W_C] \leq 0.1$ so that the policy responds by lowering T. When $\lambda_0 = 95$, T increases when δ increases. However, for the ($\delta = 2$, $\lambda_0 = 95$) case, even if we used $a = 100$, $T = 6$ $P(Loss)$ would increase only by 0.01%.

When optimal policies are compared, we see that when λ_0 or δ increases, one more agent is reserved for future arrivals by setting $a = 97$. However, if we used $a = 98$, $T = 4$, which is the optimal policy of the ($\delta = 2$, $\lambda_0 = 95$) case, as a sub-optimal policy for the other three cases, $P(Loss)$ would increase, at most, by 0.03% for the ($\delta = 2$, $\lambda_0 = 95$) case and by 0.01% for the cases with $\lambda_0 = 105$. In summary, an optimal policy appears to sustain its good performance even if arrival and abandonment rates can change significantly. This may be an advantage if these parameters change over time or we do not have reliable estimates for them.

5 Conclusion

In this paper, we study the $M_n/M_n/c/(K_1+K_2)+M_n$ queue in which the rates of the exponential interarrival, service and reneging times can change when the system size changes. Our numerical examples show that customer loss can be reduced by dynamically prioritizing the callback customers according to a threshold policy. This demonstrates the power of the callback option as a tool to mitigate the customer loss that arises especially if many customers are impatient to wait on the line. Currently, when the finite queue sizes increase, we have computational difficulties and this drawback can be overcome with more computing power with the advent of technology in the future or by designing accurate and fast approximations that can be worked on as future research. Our model can be considered in a setting with non-homogenous Poisson call arrivals, too. Thus, depending on the time of the day, different time windows to call back can be announced to callers to balance the traffic load more uniformly over the day. One other aspect to incorporate can be when to make the callback offer. In our study, a customer is able to exercise this option at any time she wants. Another threshold policy to determine the timing of the callback offer appears as the most straightforward policy: make the callback offer to customers only when the online queue length surpasses a level. Yet, this may cause some problems if, for modeling purposes, we have to keep track of those to whom the announcement has been made.

Acknowledgements We would like to thank the two anonymous referees and the editors for their invaluable suggestions to improve the manuscript. We also thank Miklós Telek who helped us with the derivation of Eqs. (17) and (19). This work was supported in part by TÜBİTAK, The Scientific and Technological Research Council of Turkey, under the Grant No. 213M428.

Funding Information This work was supported in part by TÜBİTAK, The Scientific and Technological Research Council of Turkey, under the Grant No. 213M428.

Declarations

Code availability Codes are available in Matlab and simulation in Arena

References

Akşin Z, Ata B, Emadi SM, Su C (2017) Impact of delay announcements in call centers: an empirical approach. Oper Res 65(1):242–265

Altıok T (1997) Performance analysis of manufacturing systems. Springer, New York

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

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

Ata B, Peng X (2017) An optimal callback policy for general arrival processes: a pathwise analysis. working paper, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2947368

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

Brandt A, Brandt M (1999) On a two-queue priority system with impatience and its application to a call center. Methodol Comput Appl Probab 1:191–210

ContactBabel (2016) US contact center decision makers' guide, 9th edition. http://www.contactbabel.com/pdfs/july16/The-2016-US-Contact-Center-Decision-Makers-Guide.pdf

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

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 Probl Eng. https://doi.org/10.1155/2013/983723

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

Hathaway B, Emadi SM, Deshpande V (2019) Don't call us, we'll call you: an empirical study of caller behavior under a callback option. working paper, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3349486

Kanvetas O, Balcıoğlu B (2018) The "sensitive" Markovian queueing system and its application for a call center problem. Annals Oper Res. https://doi.org/10.1007/s10479-018-2802-6

Kim C, Dudina O, Dudin A, Dudin S (2012) Queueing system MAP/M/N as a model of call center with call-back option. In: Al-Begain K, Fiems D, Vincent JM (eds) Analytical and stochastic modeling techniques and applications. ASMTA 2012. Lecture Notes in Computer Science, vol 7314. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-30782-9_1

Legros B, Jouini O, Koole G (2015) Adaptive threshold policies for multi-channel call center. IIE Trans 47:414–430

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

Legros B, Ding S, van der Mei R, Jouini O (2017) Call centers with a postponed callback offer. OR Spectr 39(4):1097–1125

Legros B, Jouini O, Akçin OZ, Koole G (2020) Front-office multitasking between service encounters and back-office tasks. Eur J Oper Res 287(3):946–963

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

Yom-Tov GB, Zeitler T (2018) Delay guarantee planning of call-back options in time-varying service systems. Proceedings of the 2018 Winter Simulation Conference, 2084–2094

Barış Balcıoğlu is a professor of industrial engineering at the Faculty of Engineering and Natural Sciences of Sabancı University, Istanbul, Turkey. He received his PhD in industrial and systems engineering in 2003 at Rutgers, the State University of New Jersey. Between 2003–2011, he worked first as an assistant and then as an associate professor at the Mechanical and Industrial Engineering Department of the University of Toronto. His research interests are modeling and analyzing production/inventory systems and call centers employing queueing theory and discrete-event simulation.

Odysseas Kanavetas is an assistant professor of mathematical institute at the Faculty of Science of Leiden University, Leiden, The Netherlands. He received his PhD in mathematics and operations research in December, 2014 at National and Kapodistrian University of Athens, Athens, Greece. Between 2015–2018, he was postdoctoral researcher at Sabancı University and Koç University, Istanbul, Turkey. In 2018 he worked as a visiting assistant professor at the Management Science and Information Systems Department of Rutgers Business School, USA. His research interests are machine learning, and modeling, analysis, estimation, simulation, optimization in inventory management, supply chain management, health care management, and queueing systems.