Performance Evaluation 95 (2016) 1–40

Contents lists available at ScienceDirect

Performance Evaluation

journal homepage: www.elsevier.com/locate/peva

Optimal scheduling in call centers with a callback option

Benjamin Legros [a], Oualid Jouini [a,*], Ger Koole [b]

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

[b] VU University Amsterdam, Department of Mathematics, De Boelelaan 1081a, 1081 HV Amsterdam, The Netherlands

ARTICLE INFO

Article history:
Received 1 August 2014
Received in revised form 29 June 2015
Accepted 23 September 2015
Available online 1 November 2015

Keywords:
Call centers
Callback option
Routing optimization
Queueing systems
Markov chains
Markov decision processes

ABSTRACT

We consider a call center model with a callback option, which allows to transform an inbound call into an outbound one. A delayed call, with a long anticipated waiting time, receives the option to be called back. We assume a probabilistic customer reaction to the callback offer (option). The objective of the system manager is to characterize the optimal call scheduling that minimizes the expected waiting and abandonment costs. For the single-server case, we prove that non-idling is optimal. Using a Markov decision process approach, we prove for the two-server case that a threshold policy on the number of queued outbound calls is optimal. For the multi-server case, we numerically characterize a switching curve of the number of agents reserved for inbound calls. It is a function of the number of queued outbound calls, the number of busy agents and the identity of jobs in service. We also develop a Markov chain method to evaluate the system performance measures under the optimal policy.

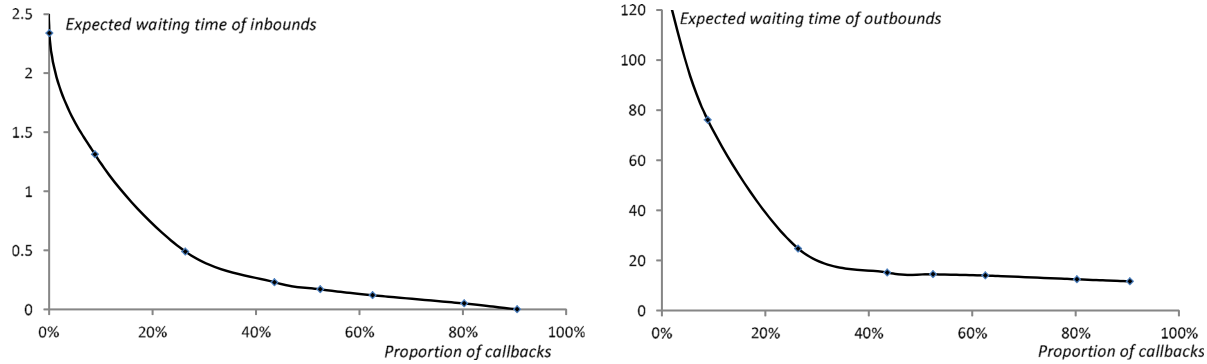
We next conduct a numerical study to examine the impact of the policy parameters on the system performance. We observe that the value of the callback offer is especially important for congested situations. It also appears that the benefits of a reservation policy are more apparent in large call centers, while they almost disappear in the extreme situations of light or heavy workloads. We moreover observe in most cases that the callback offer should be given upon arrival to any delayed call. However, if balking and abandonment are very high (which helps to reduce the workload) or if the overall treatment time spent to serve an outbound call is too large compared to that of an inbound one, there is a value in delaying the proposition of the callback offer.

1. Introduction

Context and motivation. Call centers serve as the public face in various areas and industries: insurance companies, emergency centers, banks, information centers, help-desks, tele-marketing, just to name a few. The success of call centers is due to the technological advances in information and communication systems. The most used form of communication is the telephone. However, in the context of highly congested call centers, the use of alternative service channels can be proposed to customers so as to better match demand and capacity. Alternative channels could be email, chat, blog, or postponed callback service. We focus on this last alternative. The idea is that customers, who are expected to experience long waiting times, receive the option to be called back later. This leads to a contact center with two channels, one for inbound calls (inbounds), and another for outbound calls (outbounds). The recent study of ICMI [1], based on the analysis of 361 large contact centers, reports that 76% of them use the outbound channel.

* Corresponding author.
E-mail addresses: benjamin.legros@centraliens.net (B. Legros), oualid.jouini@centralesupelec.fr (O. Jouini), ger.koole@vu.nl (G. Koole).

http://dx.doi.org/10.1016/j.peva.2015.09.002



(a) Inbounds. (b) Outbounds.

Fig. 1. Effect of the callback option on performance (arrival rate = 5.5, service rate = 0.2, number of agents = 28).

The flexibility of the callback option comes from the willingness of some customers to accept future processing. The call center can then make use of this opportunity to better manage arrival uncertainty, which in turn would improve the system performance. An illustration of callback option benefits is provided in Fig. 1. The figure gives simulated performance measures of a call center example with various levels for the use of the callback option. We consider a non-idling system where inbounds have a non-preemptive higher priority over outbounds. We observe that the expected waiting times of inbound and outbound calls are considerably improved by using the callback option. For instance, the expected waiting time of inbounds could be divided by around 20 (it decreases from 8 min and 55 s to 23 s) while only 10% of arriving calls choose to be called back.

The unpredicted and flexible call center environment offers the potential for a routing optimization that would lead to a significant operational improvement. It is a non-expensive approach compared to staffing optimization [2,3]. One important question for managers in our context is how should be the routing rule of jobs that would ensure non-excessive waiting times for both job types, i.e., upon a service completion, should the agent handle an inbound or an outbound call? when should be proposed the callback offer? We address these questions under a queueing modeling framework and a probabilistic customer reaction to the callback option.

A call center where agents simultaneously handle inbound and outbound calls is commonly referred to as *call blending*. The key distinction of call center problems with blending comes from the fact that outbound tasks have less urgency relative to inbound calls. Blended operations problems have led to research on performance evaluation [4–6], staffing [7] and analysis of blending policies [8–13]. Because of the lack of service level requirement on outbounds, it is best to give higher priority to inbounds. Moreover, to reduce the number of inbounds who may experience long waiting before service, one has to guarantee that there is sufficient idleness in the system. In the patent of Dumas et al. [14], based on extensive simulation experiments, it is shown that blending inbound and outbound calls and employing a threshold policy, ensure that the outbound throughput rate is met while waiting times of inbounds are very short. It is also shown that blending the two types of calls in one pool requires less agents than employing two distinct pools. Bhulai and Koole [9] and Gans and Zhou [2], prove this optimal control, which is of threshold type, when the service rates of the two types of jobs are equal. More precisely, they show that it is optimal to schedule outbound tasks only when no outbounds are in the queue and the number of idle agents exceeds a certain threshold.

In the case of a callback option, this policy cannot be directly applied. The reason is that the above literature considers an infinite amount of non-priority jobs. In a call center with a callback option, the number of customers waiting to be called back has to be finite in order to avoid infinite waiting. The routing policy should then account for the length of the callback queue. Another difference, compared to cases with classical infinite amount of outbound tasks, is that inbound and outbound arrivals are negatively correlated. This requires further analysis, and may lead to different managerial recommendations.

Contributions. We consider a call center with a single customer type. A delayed call, with a long anticipated waiting time, receives the option to be called back. We develop a modeling that accounts for balking, abandonment, probabilistic customer reaction to a state-dependent delay information, unequal service requirements for job types, and the eventual non-availability of a called back customer. The objective of the system manager is to find the optimal call scheduling policy that minimizes the expected operating costs of inbounds and outbounds. The control actions concern the number of agents reserved for inbounds and the system state situations at which the callback offer should be proposed.

We distinguish three main contributions. The first contribution is related to the agent reservation policy. We prove for the single-server case that non-idling is optimal. Using a Markov decision process (MDP) approach, we prove for the two-server case with equal service requirements that a threshold policy on the number of queued outbounds is optimal. Based on the two-server result, we conjecture for the multi-server case that the optimal policy is of switch type. The number of agents to reserve for inbounds depends on the number of queued outbounds, the number of busy agents and the identity of jobs in service. Moreover, we examine the impact of the system exogenous parameters on the agent reservation policy. We observe, for example, that a reservation policy is not likely to be used under light or heavily loaded situations.