Performance Evaluation 92 (2015) 24–39

Contents lists available at ScienceDirect

Performance Evaluation

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

Fluid approximation of a call center model with redials and reconnects

S. Ding [a,*], M. Remerova [a], R.D. van der Mei [a,b], B. Zwart [a]

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

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

HIGHLIGHTS

- We model the customer redial and reconnect behaviors in call centers.
- We approximate the service levels and abandonment percentages of such a model.
- A fluid model is proposed, and the corresponding fluid limit is derived.
- The performance of our approximation is evaluated numerically.

ARTICLE INFO

Article history:
Received 4 September 2014
Received in revised form 2 July 2015
Accepted 10 July 2015
Available online 20 July 2015

Keywords:
Call centers
Fluid model
Redial
Reconnect
Erlang A

ABSTRACT

In many call centers, callers may call multiple times. Some of the calls are re-attempts after abandonments (redials), and some are re-attempts after connected calls (reconnects). The combination of redials and reconnects has not been considered when making staffing decisions, while not distinguishing them from the total calls will inevitably lead to under- or overestimation of call volumes, which results in improper and hence costly staffing decisions.

Motivated by this, in this paper we study call centers where customers can abandon, and abandoned customers may redial, and when a customer finishes his conversation with an agent, he may reconnect. We use a fluid model to derive first order approximations for the number of customers in the redial and reconnect orbits in the heavy traffic. We show that the fluid limit of such a model is the unique solution to a system of three differential equations. Furthermore, we use the fluid limit to calculate the expected total arrival rate, which is then given as an input to the Erlang A formula for the purpose of calculating the service levels and abandonment probabilities. The performance of such a procedure is validated numerically in the case of both single intervals with constant parameters and multiple intervals with time-dependent parameters. The results demonstrate that this approximation method leads to accurate estimations for the service levels and the abandonment probabilities.

1. Introduction

Nowadays, call centers are important means of communication with customers. Therefore, the response-time performance of call centers is crucial for the customer satisfaction. It is essential to the costs and the performances of call centers

* Corresponding author.
E-mail addresses: s.ding@cwi.nl (S. Ding), maria.frolkova@gmail.com (M. Remerova), mei@cwi.nl (R.D. van der Mei), bert.zwart@cwi.nl (B. Zwart).

http://dx.doi.org/10.1016/j.peva.2015.07.003
0166-5316/

that managers make the right staffing decisions (i.e., determine the right number of agents). Various models have been developed in order to support such decision processes. One of the most widely used models is the Erlang C model and there is a lot of literature on it (see [1] and the references therein). The square-root staffing rule is a simplified and approximated staffing rule for the Erlang C model, which is proposed by Halfin and Whitt [2]. However, the Erlang C model does not include customer abandonments, while the Erlang A model does. Garnett et al. [3] show that the square-root staffing rule remains valid for the Erlang A model. However, both the Erlang C and the Erlang A model ignore customer redial (a re-attempt after an abandoned call) behaviors in call centers, while this behavior can be quite significant (see [1] and reference therein). Aguir et al. [4] discover that ignoring redials can lead to under-staffing or over-staffing, depending on the forecasting assumption being made. This model with reneging is also studied in [5], and later extended by Phung-Duc and Kawanishi [6] and Phung-Duc and Kawanishi [7] with an extra feature of after-call work. Sze [8] studies a queueing model where abandonments and redials are included, focusing on the heavily loaded systems. We refer to Falin and Templeton [9] for more references in retrial queues.

Besides redials, there also exists another important feature, which is called reconnect (a re-attempt after a connected call). The reconnect customer behavior is first mentioned in [1] as revisit. Motivated by the application in healthcare staffing with reentrant patients, Liu and Whitt [10]; Yom-Tov and Mandelbaum [11] develop methods to set staffing levels for models with and without Markovian routing. Such methods remain valid for time-varying demand. In [12], the authors use real call center data to show that an inbound call can either be a fresh call (an initial attempt), a redial or a reconnect. Also, as argued in [12], redials and reconnects should be considered and modeled, since without distinguishing them from the fresh calls can lead to significantly over- or underestimation of the total inbound volume. As a consequence, neglecting the impact of redials and reconnects will lead to either overstaffing or understaffing. In case of overstaffing the performance of the call center will be good, but at unnecessarily high costs. In case of understaffing, the performance of the call center will be degraded, which may lead to customer dissatisfaction and possibly customer churn. Despite the economic relevance of including both features in staffing models, to the best of the authors' knowledge no papers have appeared on staffing of call centers where *both* redials and reconnects are included. This paper aims to fill this gap, that is, we investigate the staffing problem in call centers with the features of both redials and reconnects. We focus on the case of large call centers that operate under heavy load.

In the Erlang C model, if the system is heavily loaded, the expected queueing length will go to infinity in stationarity, and arriving customers will on average experience infinity long waiting. However, for large call centers with customer abandonments, especially during the busy hours when the inbound volume is quite large such that the system operates under heavy load, it is possible that most customers will experience relatively short waiting times while having only a small customer abandonment percentage. Further discussions of this effect can be found in [3].

In this paper, we aim to answer the following question: "In large call centers, for given number of agents, what are the service level (SL) and the abandonment percentage (AP) if both redialing and reconnection of customers are taken into account?" In this paper, the SL is defined as the probability that customers get served and wait less than certain given acceptable waiting time, and the AP is defined as the probability that customers abandon. To answer this question, one must first estimate the total number of arrivals into the call center. This is not trivial, since the number of total arrivals depends on the number of agents (see [12]). This dependency becomes more complicated in real life, due to the fact that the rate of fresh calls arriving and the number of agents are often time-dependent. If the number of arrivals cannot be determined, it is impossible to calculate the SL. Therefore, in this paper, we take a two-step approach to calculate the SL and AP. First, we numerically calculate the expected total arrival rate at any instant time by using a fluid limit approximation. We also show that the fluid limit of this model is a unique solution to a system of three deterministic differential equations. In the second step, under the assumption of the total arrival process being Poisson, we apply the Erlang A formula to obtain the SL and the AP. This approximation turns out to be quite accurate. In this paper, we consider only the expected SL and AP, for discussions about the SL variability, we refer to the work by Roubos et al. [13].

Fluid models for call centers have been extensively studied. Whitt [14] develops a deterministic fluid limit which they use to provide first-order performance descriptions for the $G/GI/s + GI$ queueing model under heavy traffic, where the second GI stands for the i.i.d. patience distribution. In [14], the redial behavior is not considered, though. The existence and uniqueness of the fluid limit are given as conjectures. Mandelbaum et al. [15] use the fluid and diffusion approximation for the multi-server system with abandonments and redials. He obtains first order approximations of queue length and expected waiting time as well as their confidence bounds. In [16], the authors use a fluid and a diffusion approximation for the time varying multiserver queue with abandonments and retrials. They show that both approximations can be obtained by solving sets of non-linear differential equations, where the diffusion process can provide confidence bounds for the fluid approximation. The work by Mandelbaum et al. [17] gives more general theoretical results for fluid and diffusion approximations for Markovian service networks. Aguir et al. [18] extend the model by allowing customer balking behavior, but no formal proof of the fluid limit is given. Besides the applications in staffing call centers, fluid models have also been applied in delay announcement of customers in call centers (see [19,20]). Besides the fluid or the diffusion limits, there are other methods that can be used to approximate queueing models, such as the Gaussian Variance Approximation (GVA) method developed by Massey and Pender [21]. Such a GVA approach is generalized by Pender and Massey [22] to Jackson networks with abandonment, which leads to better approximations comparing to approximation results obtained by the corresponding fluid and diffusion limits.

The rest of the paper is structured as follows. In Section 2, we describe the queueing model with the features of the redial and reconnect. In Section 3, we propose a fluid model, which is a deterministic analogue of the stochastic model. We prove