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

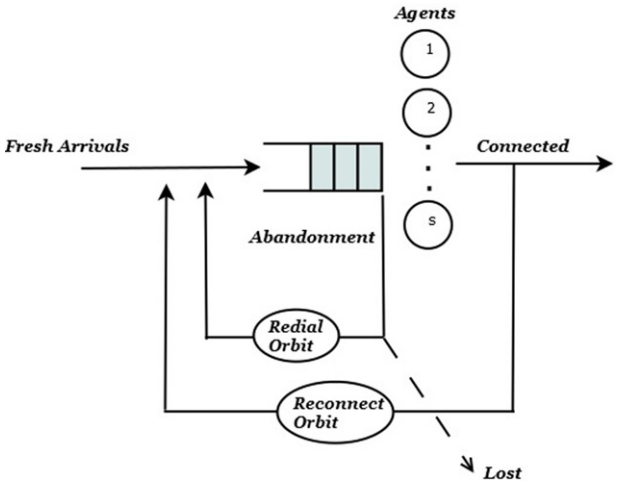


Fig. 1. Call diagram.

that the original stochastic model converges to the fluid model under a proper scaling. We numerically compute the fluid approximations to the number of customers in the queue as well as those in two orbits, and simulate the original model, and compare them in the case of single intervals and multiple intervals, where the parameters are time-dependent but remain piece-wise constants within each interval. The Erlang A formula is then used to approximate the waiting time distributions in Section 5.

2. Model description

Consider the queueing model illustrated in Fig. 1. We assume that calls arrive according to a Poisson process. We refer to these calls as *fresh calls*. There are s agents who handle inbound calls. An arriving call is handled by an available agent, if there is any; otherwise, it waits in an infinite buffer queue. The calls are handled in the order of arrival. After an exponentially distributed amount of time Ψ, a waiting customer who did not get connected to an agent will lose his patience and abandon. We assume $\mathbb{E}\Psi = 1/\theta < \infty$, where θ is the abandonment rate. With probability p, an abandoned customer will enter the redial orbit, and he will redial after an exponentially distributed amount of time Γ_{RD}, with $\mathbb{E}\Gamma_{RD} = 1/\delta_{RD} < \infty$. We refer to these calls as *redials*. With probability $1 - p$, this customer will not call back, and this call is considered as a "lost" call. We assume that the holding time B of a call has an exponential distribution with mean $\mathbb{E}B = 1/\mu < \infty$, regardless whether this call is a fresh call, a redial or a reconnect. After the call has been answered, its corresponding customer will enter the reconnect orbit with probability q, and he will reconnect after an exponentially distributed time Γ_{RC}, with $\mathbb{E}\Gamma_{RC} = 1/\delta_{RC} < \infty$. We refer to such calls as *reconnects*. We assume that p and q do not depend on customers' experiences in the system. These experiences include holding times, waiting times and the numbers of times that customers have already called. We use this queueing model to represent the situation of a single-skill call center. In this paper, we consider independent service times; for the study of dependent service times, please see [23].

The inclusion of reconnect supplemented with the exponential assumption of Γ_{RD} and Γ_{RC} is motivated by the fact that in practice the volume of reconnect is significant and Γ_{RD} and Γ_{RC} are approximately exponential. To demonstrate these, we conduct data analysis of real call center data. The data are collected by a Dutch call center over a half-year period, which is also used in [12]. This data set consists of call records of different types. Each type of calls represents one type of questions of callers. The callers identity information is recorded in the data. Thus, we can compute the reconnect probability by following the call records of each distinct caller. Different types of calls have different reconnect probabilities, and for simplicity, we only select one type of calls, which accounts for nearly 40% of all call records. The redial and reconnect probabilities of this type of calls are 0.40 and 0.15, respectively. Then, in such a case, if all the customers are connected to agents and 15% of the connected customers reconnect exactly once, then more than 13% of the total number of arrivals are reconnects. This further confirms the necessity to include the reconnect customer behavior in call center models. Besides calculating p and q from the data, we also plot the histograms of Γ_{RD} and Γ_{RC} in Figs. 2 and 3. When generating these two figures, if a customer tries to redial or reconnect after one day, we consider that call back as a fresh call. We make this consideration because most of the customers call back in the same day as their corresponding fresh calls [12]. The sample averages of Γ_{RD} and Γ_{RC} are $1/\delta_{RD} = 41.46$ min and $1/\delta_{RC} = 53.49$ min, respectively. As one can see from the shapes in Figs. 2 and 3 that the histograms have longer tails than the exponential distributions. Besides the longer tails, the exponential distributions seem to be good approximations for Γ_{RD} and Γ_{RC}. This can be demonstrated by the two Q–Q plots in Fig. 4(a) and (b). In these two Q–Q plots, we ignore all the samples that are larger than 3 h for Γ_{RD} and that are larger 4 h for Γ_{RC}, which account for less than 8% of total sample size. Therefore, we keep the assumption that Γ_{RD} and Γ_{RC} are exponentially distributed.

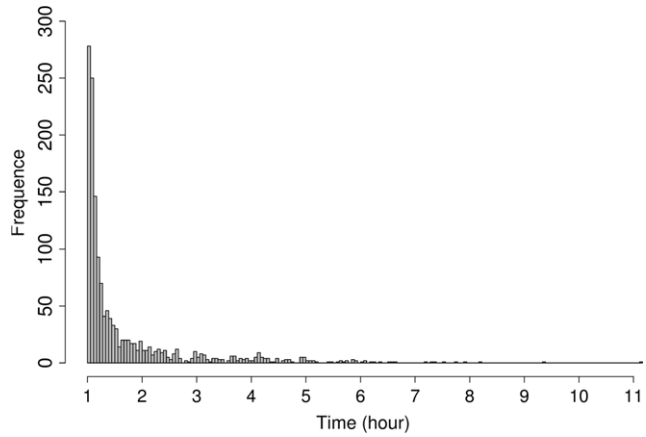


Fig. 2. Histogram of Γ_{RD}.

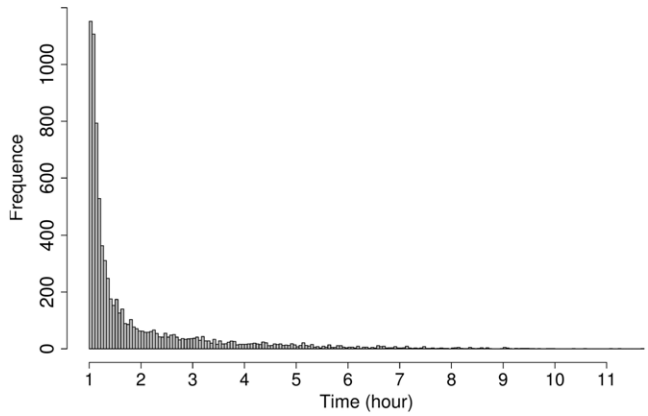
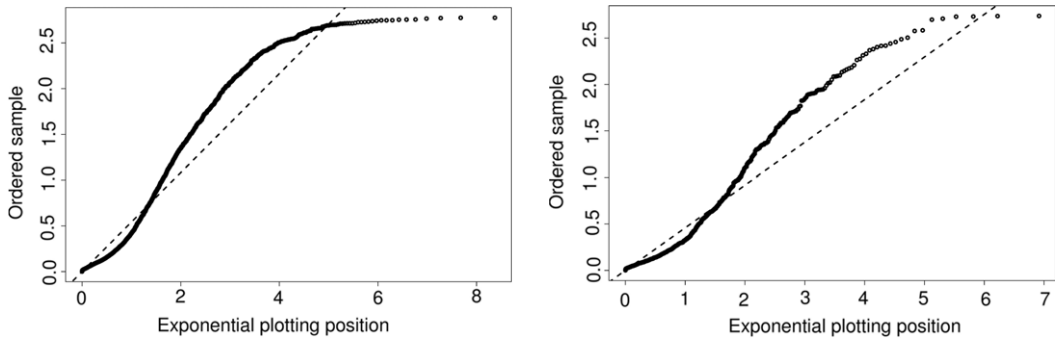


Fig. 3. Histogram of Γ_{RC}.



(a) Q–Q plot of the realizations of Γ_{RD}.

(b) Q–Q plot of the realizations of Γ_{RC}.

Fig. 4. Q–Q plots.

In our model, we assume that there is no difference between the handling times of the fresh calls and those of the reconnects. To validate this assumption, we calculate the average holding times for the fresh calls and the reconnects, e.g., $\mathbb{E}B = 5.14$ min, for the fresh calls, and $\mathbb{E}B = 5.35$ min, for the reconnects. Thus, the fresh calls have slightly lower average handling time. We think that a possible cause for this is that reconnects might represent more difficult questions of customers than the fresh calls, which means it requires longer efforts to handle the reconnects. However, if we differentiate between the handling times of the fresh calls and the reconnects, this would complicate the model significantly, since instead of knowing the total number of customers in the queue, one would need to know both the number of the reconnects and the number of the fresh calls in the queue, as well as their orders in the queue. Therefore, considering the added complexity

and the fact that the difference between them is relatively small, in this paper, we keep our assumption that the fresh calls are statistically the same as the reconnects in terms of service durations.

3. Fluid limit approximations

In this section, we first show that the problem of calculating the expected total arrival rate comes down to the problem of calculating $\mathbb{E}Z_Q(t)$, $\mathbb{E}Z_{RD}(t)$ and $\mathbb{E}Z_{RC}(t)$, where $Z_Q(t)$ is the number of customers in the queue plus the number of customers in service at time t, $Z_{RD}(t)$ is the number of customers in the redial orbit at time t, and $Z_{RC}(t)$ is the number of customers in the reconnect orbit at time t, and $Z_Q(t)$, $Z_{RD}(t)$ and $Z_{RC}(t)$ are random processes. Because an arrival can be a fresh call, a redial or a reconnect, then the following equation holds for any t,

$$\begin{aligned}\mathbb{E}\Lambda(t) &= \lambda(t) + \mathbb{E}\lambda_{RD}(t) + \mathbb{E}\lambda_{RC}(t)\\ &= \lambda(t) + \delta_{RD}\mathbb{E}Z_{RD}(t) + \delta_{RC}\mathbb{E}Z_{RC}(t),\end{aligned} \tag{1}$$

where $\Lambda(t)$ stands for the total arrival rate at time t, which is a stochastic process, $\lambda(t)$ stands for the fresh arrival rate at time t, $\lambda_{RD}(t)$ and $\lambda_{RC}(t)$ stand for the arrival rate due to redials and reconnects at time t, respectively. Therefore, once $\mathbb{E}Z_Q(t)$, $\mathbb{E}Z_{RD}(t)$ and $\mathbb{E}Z_{RC}(t)$ are known, $\mathbb{E}\Lambda(t)$ can be obtained by Eq. (1). Note that $Z_Q(t)$ does not appear in Eq. (1), but we will see later that $Z_{RD}(t)$ and $Z_{RC}(t)$ depend on $Z_Q(t)$.

In fact, the stochastic process $\{\mathbf{Z}(t), t \geq 0\}$, which is defined by

$$\mathbf{Z}(t) := \left(Z_Q(t), Z_{RD}(t), Z_{RC}(t)\right)^T, \tag{2}$$

is a 3-dimensional Markov process, because the inter-arrival time, service duration and other durations are assumed to be exponentially distributed. The state space of this Markov process is $\mathbb{Z}_+^3$. To save space, we will not show the transition diagram here. Since it is a Markov process, we can truncate the system at certain large state, and numerically obtain the steady state distribution of $\mathbf{Z}(t)$ by solving global balance equations. Theoretically, this method offers almost exact results, in the sense that one can control the error by truncating at some sufficiently large state. However, for the model we consider, it is very difficult to formulate and solve the global balance equations, especially for large systems. Therefore, for the convenience of practical usage, we will not consider solving this Markov process, but some approximation methods.

3.1. Fluid limit

In this subsection, we present a fluid model, which we show to arise as the limit under a proper scaling of the stochastic model in Fig. 1.

Often, the fresh arrival rates are time-dependent in real call centers. The operational hours of call centers are divided into several intervals for the convenience of staffing and making schedules, and it is conventional to assume that the fresh arrival rate differs per interval but remains piece-wise constant within each single interval. Thus, we start our analysis by considering the single interval case, where the fresh arrival rate is assumed to be constant for any t (e.g., $\lambda(t) = \lambda,\ t \geq 0$). The cases with time-dependent arrival rates will be discussed later. For the single interval case, the following flow conservation equations hold for this stochastic model:

$$Z_Q(t) = Z_Q(0) + \Pi_\lambda(t) + D_{RD}(t) + D_{RC}(t) - D_s(t) - D_a(t), \tag{3}$$

$$Z_{RD}(t) = Z_{RD}(0) + \sum_{j=1}^{D_a(t)} B_j(p) - D_{RD}(t), \tag{4}$$

$$Z_{RC}(t) = Z_{RC}(0) + \sum_{j=1}^{D_s(t)} B_j(q) - D_{RC}(t), \tag{5}$$

where $\Pi_\lambda(t)$ is the number of fresh arrivals during time interval $[0, t)$, and $\Pi_\lambda(\cdot)$ is a Poisson process of rate λ. In addition, $D_{RD}(t)$, $D_{RC}(t)$, $D_s(t)$, $D_a(t)$ are the number of redials during $[0, t)$, number of reconnects during $[0, t)$, number of served customers during $[0, t)$ and number of abandoned customers during $[0, t)$, respectively. $B_j(p)$ is a Bernoulli random variable with success probability p, $j = 1, 2, \ldots, D_a(t)$. That is, $B_j(p) = 1$, if the jth abandoned customer enters the redial orbit; $B_j(p) = 0$, otherwise. Therefore, for given $D_a(t)$, $\sum_{j=1}^{D_a(t)} B_j(p) \sim \text{Bin}(D_a(t), p)$. By the same argument, we have $\sum_{j=1}^{D_s(t)} B_j(q) \sim \text{Bin}(D_s(t), q)$, for given $D_s(t)$.

Let $\Pi_i(\cdot)$, $i = 1, 2, 3, 4$, be independent Poisson processes of rate 1, then we claim the following:

$$D_s(t) = \Pi_1\left(\int_0^t \mu \min\{s, Z_Q(u)\}du\right),$$

$$D_a(t) = \Pi_2\left(\int_0^t \theta\left(Z_Q(u) - s\right)^+ du\right),$$

$$D_{RD}(t) = \Pi_3\left(\int_0^t \delta_{RD} Z_{RD}(u)\,du\right),$$
$$D_{RC}(t) = \Pi_4\left(\int_0^t \delta_{RC} Z_{RC}(u)\,du\right).$$

Rigorous proof of these four statements can be given along the lines of Pang et al. [24], see Lemma 2.1.

To introduce the fluid limit, we consider a sequence of models as in Fig. 1 such that, in the nth model, the fresh arrival rate is λn and the number of servers is ns. We add the superscript "(n)" to all notations in the nth model. Similarly to (3)–(5), we then have for the nth model:

$$Z_Q^{(n)}(t) = Z_Q^{(n)}(0) + \Pi_{\lambda n}^{(n)}(t) + D_{RD}^{(n)}(t) + D_{RC}^{(n)}(t) - D_s^{(n)}(t) - D_a^{(n)}(t), \tag{6}$$

$$Z_{RD}^{(n)}(t) = Z_{RD}^{(n)}(0) + \sum_{j=1}^{D_a^{(n)}(t)} B_j(p) - D_{RD}^{(n)}(t), \tag{7}$$

$$Z_{RC}^{(n)}(t) = Z_{RC}^{(n)}(0) + \sum_{j=1}^{D_s^{(n)}(t)} B_j(q) - D_{RC}^{(n)}(t). \tag{8}$$

Now we define the fluid scaled process

$$\bar{\mathbf{Z}}^{(n)}(t) := \left(\bar{Z}_Q^{(n)}(t), \bar{Z}_{RD}^{(n)}(t), \bar{Z}_{RC}^{(n)}(t)\right)^T,$$

where

$$\bar{Z}_Q^{(n)}(t) := \frac{Z_Q^{(n)}(t)}{n}, \qquad \bar{Z}_{RD}^{(n)}(t) := \frac{Z_{RD}^{(n)}(t)}{n}, \qquad \bar{Z}_{RC}^{(n)}(t) := \frac{Z_{RC}^{(n)}(t)}{n}.$$

Let $D([0,\infty), \mathbb{R}^3)$ be the space of right continuous functions with left limits in $\mathbb{R}^3$ having the domain $[0,\infty)$. We endow $D([0,\infty), \mathbb{R}^3)$ with the usual Skorokhod J_1 topology. Suppose $\{X^{(n)}\}_{n=1}^{\infty}$ is a sequence of stochastic processes, then notation $X^{(n)} \xrightarrow{d} x$ means that $X^{(n)}$ converges weakly to stochastic process x.

Definition 1. If there exists a limit in distribution for the scaled process $\{\bar{\mathbf{Z}}^{(n)}(\cdot)\}_{n=1}^{\infty}$, i.e. $\bar{\mathbf{Z}}^{(n)}(\cdot) \xrightarrow{d} \mathbf{z}(\cdot)$, then $\mathbf{z}(\cdot)$ is called the fluid limit of the original stochastic model.

3.1.1. Fluid limit for a single interval

To obtain the fluid limit of the system (i.e., a sequence of stochastic processes specified by Eqs. (6)–(8)), we divide both sides of Eqs. (6)–(8) by n, then let $n \to \infty$.

Lemma 1. *The sequence of scaled processes $\{\bar{\mathbf{Z}}^{(n)}(\cdot)\}_{n=1}^{\infty}$ is relatively compact and all weak limits are a.s. continuous.*

Proof. See Appendix B. □

Theorem 1. *If for given deterministic values $\left(z_Q(0), z_{RD}(0), z_{RC}(0)\right)$, we assume $\left(\bar{Z}_Q^{(n)}(0), \bar{Z}_{RD}^{(n)}(0), \bar{Z}_{RC}^{(n)}(0)\right) \xrightarrow{d} \left(z_Q(0), z_{RD}(0), z_{RC}(0)\right)$ as $n \to \infty$, then the fluid limit of the original stochastic model is the unique solution to the following system of equations*

$$z_Q(t) = z_Q(0) + \lambda t + \delta_{RD}\int_0^t z_{RD}(u)du + \delta_{RC}\int_0^t z_{RC}(u)du - \mu\int_0^t \min\{s, z_Q(u)\}du - \theta\int_0^t \left(z_Q(u) - s\right)^+ du, \tag{9}$$

$$z_{RD}(t) = z_{RD}(0) + p\theta\int_0^t \left(z_Q(u) - s\right)^+ du - \delta_{RD}\int_0^t z_{RD}(u)du, \tag{10}$$

$$z_{RC}(t) = z_{RC}(0) + q\mu\int_0^t \min\{s, z_Q(u)\}du - \delta_{RC}\int_0^t z_{RC}(u)du. \tag{11}$$

Proof. See Appendix C. □

Remark. Mandelbaum et al. [17] suggests an alternative proof of this fluid limit result for a more general model. The approach of Mandelbaum et al. [17] is based on a Brownian motion approximation of a Poisson process, thus, a second order approximation, which they simultaneously use to derive both the fluid and diffusion limits. Our derivation of the fluid limit is more straightforward and does not use second order approximations. More precisely, we use the recipe of Ethier and Kurtz [25] which can be considered classic for proving fluid limit results. For us the fluid limit alone is sufficient to obtain an approximation to the waiting time distribution (see the Erlang A approximation in Section 5).

We could not obtain analytic expressions of $z_Q(t)$, $z_{RD}(t)$ and $z_{RC}(t)$ from Eqs. (9)–(11). However, solving them numerically can be done via a standard approach for solving differential equations, and it is relatively fast.

3.1.2. Fluid limit for multiple intervals

We have just shown the fluid limit for a single interval, where the parameters λ and s remain the same within the interval. However, in real call centers, parameters can vary during the day, especially the arrival rate $\lambda(t)$. As shown by Shen and Huang [26] and Ibrahim and L'Ecuyer [27], call volumes normally follow certain intraday patterns. Observing the intraday arrival pattern from the historical data set, managers would schedule different number of agents for each interval to meet the SL requirement or time-stable performance (see for example [28,29]). Therefore, we now show the fluid limit for the case of multiple intervals, where λ and s vary from interval to interval. We assume that other parameters remain constant.

We divide the operational hours of call centers into m intervals. Each interval starts at t_{i-1} and ends at t_i, $i = 1, 2, \ldots, m$. The fresh arrival rate of interval i is denoted by λ_i, and the number of agents in interval i is denoted by s_i, $i = 1, 2, \ldots, m$. In the ith interval, i.e., $t_{i-1} \le t < t_i$, the fluid limit then becomes

$$\begin{aligned} z_Q(t) &= z_Q(t_{i-1}) + \lambda_i (t - t_{i-1}) + \delta_{RD} \int_{t_{i-1}}^{t} z_{RD}(u)du + \delta_{RC} \int_{t_{i-1}}^{t} z_{RC}(u)du \\ &\quad - \mu \int_{t_{i-1}}^{t} \min\{s_i, z_Q(u)\}du - \theta \int_{t_{i-1}}^{t} \left(z_Q(u) - s_i\right)^+ du, \end{aligned} \tag{12}$$

$$z_{RD}(t) = z_{RD}(t_{i-1}) + p\theta \int_{t_{i-1}}^{t} \left(z_Q(u) - s_i\right)^+ du - \delta_{RD} \int_{t_{i-1}}^{t} z_{RD}(u)du, \tag{13}$$

$$z_{RC}(t) = z_{RC}(t_{i-1}) + q\mu \int_{t_{i-1}}^{t} \min\{s_i, z_Q(u)\}du - \delta_{RC} \int_{t_{i-1}}^{t} z_{RC}(u)du. \tag{14}$$

Numerically solving Eqs. (12)–(14) is similar to solving Eqs. (9)–(11), thus, we do not elaborate on the procedure here.

In reality, parameters such as μ, θ, δ_{RD} and δ_{RC} can also be time-dependent, and vary per interval. For example, δ_{RD} may be bigger in the late afternoon than in the morning, since abandoned customers want to have responses by the end of the day. It is possible to extend the model in Eqs. (12)–(14) to adapt such situation by simply replacing the parameters. In this paper, for the simplicity of validation, we will not consider such cases.

3.2. Model under stationarity

We have just shown that one can numerically solve differential equations (9)–(11) to obtain the fluid limit $\mathbf{z}(t)$. We now derive the stationary fluid limit, i.e., we develop conditions under which $\mathbf{z}(t)$ is constant.

By taking the derivative of Eqs. (9)–(11) and assuming that $\frac{d}{dt}\mathbf{z}(t) = 0$ has a constant solution, we can obtain

$$0 = \lambda + \delta_{RD} z_{RD}(\infty) + \delta_{RC} z_{RC}(\infty) - \mu \min\{s, z_Q(\infty)\} - \theta \left(z_Q(\infty) - s\right)^+, \tag{15}$$

$$0 = p\theta \left(z_Q(\infty) - s\right)^+ - \delta_{RD} z_{RD}(\infty), \tag{16}$$

$$0 = q\mu \min\{s, z_Q(\infty)\} - \delta_{RC} z_{RC}(\infty), \tag{17}$$

where $z_Q(\infty) := \lim_{t\to\infty} z_Q(t)$, $z_{RD}(\infty) := \lim_{t\to\infty} z_{RD}(t)$, $z_{RC}(\infty) := \lim_{t\to\infty} z_{RC}(t)$.

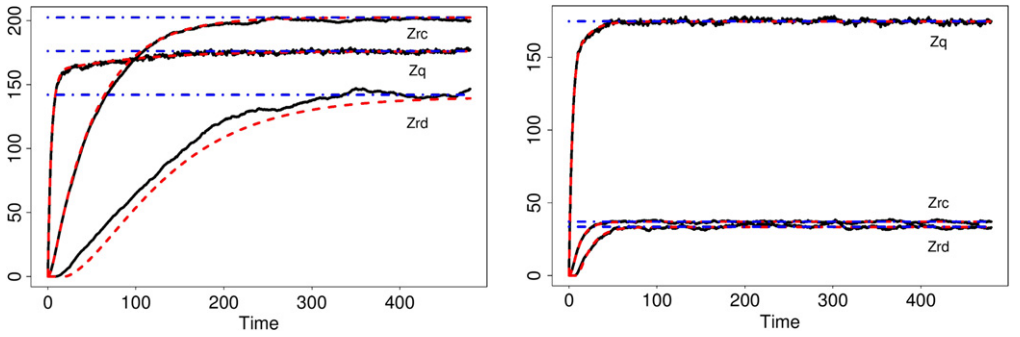
Eqs. (15)–(17) can be easily solved with respect to $z_Q(\infty)$, $z_{RD}(\infty)$ and $z_{RC}(\infty)$, yielding

$$z_Q(\infty) = \begin{cases} \dfrac{\lambda}{(1-q)\mu}, & \text{if } \rho < (1-q) \\ \dfrac{\lambda + q\mu s - \mu s}{\theta(1-p)} + s, & \text{if } \rho \ge (1-q) \end{cases} \tag{18}$$

$$z_{RD}(\infty) = \begin{cases} 0, & \text{if } \rho < (1-q) \\ \dfrac{p\theta\left(z_Q(\infty) - s\right)}{\delta_{RD}}, & \text{if } \rho \ge (1-q) \end{cases} \tag{19}$$

$$z_{RC}(\infty) = \begin{cases} \dfrac{q\mu z_Q(\infty)}{\delta_{RC}}, & \text{if } \rho < (1-q) \\ \dfrac{q\mu s}{\delta_{RC}}, & \text{if } \rho \ge (1-q). \end{cases} \tag{20}$$

The results above would offer some insights. $\rho := \frac{\lambda}{s\mu}$ is the load of the system due to the fresh arrivals. However, the total load into the system is at least $\hat{\rho} := \frac{\lambda}{(1-q)s\mu}$, since $\frac{1}{1-q}$ portion of ρ will reconnect. In the case of $\hat{\rho} < 1$, we have $z_Q(\infty) < s$ and $z_{RD}(\infty) = 0$. This means there is no abandonment at all in the stationary fluid limit when $\hat{\rho} < 1$, and the stationary fluid limit do not depend on δ_{RD} at all. In reality, due to the variabilities in the arrival process, service duration and the patience, abandonments would not be 0 though, but very small numbers. One would expect that the fluid approximation has high



(a) $1/\delta_{RD} = 40$ and $1/\delta_{RC} = 50$. (b) $1/\delta_{RD} = 5$ and $1/\delta_{RC} = 10$.

Fig. 5. Simulation results (black solid curve), fluid approximations (red dashed curve) and stationary fluid limit (blue dot-dashed curve), $\lambda = 40$, $\hat{\rho} = 1.1$. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

approximation errors in such a case. If $\hat{\rho} > 1$, by Eq. (18), $z_Q(\infty) > s$. Therefore, in this case, the stationary fluid limit indicates that there will be $(z_Q(\infty) - s)$ amount of customers waiting in the queue, each abandons the system with rate θ, thus, the total abandonment rate is then $p\theta(z_Q(\infty) - s)$.

As we mentioned before, we could not obtain an analytical expression for the steady state probability of the original system, thus, the stability condition of the original system is then also difficult to derive. Our fluid limit is not for stability but for approximation of a many-arrivals many-servers system, thus, it does not provide the exact stability condition. However, Eqs. (18)–(20) could give some insight. If we consider $z_Q(\infty)$, $z_{RD}(\infty)$ and $z_{RC}(\infty)$ being less than ∞ as the fluid limit being stable, then following conditions are necessary for the stability of the original system: in the case of $\rho/(1-q) < 1$, one requires $\delta_{RC} > 0$; and in the case of $\rho/(1-q) > 1$, one requires $q < 1$, $\theta > 0$, $p < 1$, $\delta_{RD} > 0$ and $\delta_{RC} > 0$.

4. Validation of the fluid limit

In this section, we validate the fluid model via simulation both for a single interval and for multiple intervals. We simulate the system for 480 min of time, i.e., 8 h, which correspond to the busy hours in some call centers. The results obtained via the fluid limit are compared with the simulation results. Since $\mathbf{Z}(t)$ is a stochastic process, it has variability. To remove those variabilities, we do the simulation for 100 times, and then take the average.

4.1. Validation of a single interval

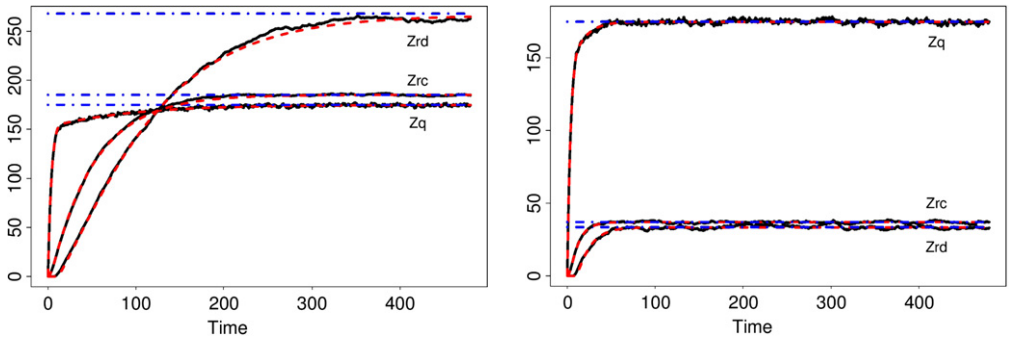
We start with the simple case of a single interval, where $\lambda(t) = \lambda$, for all $t > 0$, and we assume that s, μ as well as other parameters are constants over time. We compare $\mathbf{z}(t)$ (computed via Eqs. (9)–(11)) with $\mathbf{Z}(t)$ (simulation results), and with $\mathbf{z}(\infty) := \left(z_Q(\infty), z_{RD}(\infty), z_{RC}(\infty)\right)^T$ (computed via Eqs. (18)–(20)) for different values of $\hat{\rho}$ and λ. For each value of $\hat{\rho}$, s changes, while $1/\mu = 4$, $p = 0.5$, $q = 0.1$, $\theta = 0.5$ remain the same. We consider two scenarios; in the first scenario, we let $1/\delta_{RD} = 40$ min and $1/\delta_{RC} = 50$ min, which correspond to the real values from the data; in the second scenario, we let $1/\delta_{RD} = 5$ min and $1/\delta_{RC} = 10$ min, which represents the case with "impatient" customers, in the sense that they spend little time in the redial and reconnect orbits. We also consider two different values for λ, i.e., $\lambda = 10$ for relatively small call centers and $\lambda = 40$ for relatively large call centers. Two examples of $\mathbf{z}(t)$ and $\mathbf{Z}(t)$, where $\lambda = 40$, $\hat{\rho} = 1.1$ and $\hat{\rho} = 1.2$, are shown in Figs. 5 and 6, respectively.

One can see from Figs. 5 and 6 that the systems start with zero customers, and as time passes by, $Z_Q(t)$, $Z_{RD}(t)$ and $Z_{RC}(t)$ gradually build up and reach stationarity. These stationarity are well approximated by $\mathbf{z}(\infty)$. Furthermore, in both parameter settings, the fluid limits offer close approximations to the original processes, especially for $Z_Q(t)$ and $Z_{RC}(t)$. The approximation error is larger for $Z_{RD}(t)$, especially when $\hat{\rho} = 1.1$. We now explain why. The fluid limits ignore the variability in the number of customers in the queue; when the queue length is not large, such as the period when $Z_{RD}(t)$ does not reach stationarity, ignoring variability can lead to relatively large errors.

Obtaining an approximation to $\mathbf{Z}(t)$ is the intermediate step for calculating $\lambda_{RD}(t)$ and $\lambda_{RC}(t)$. Therefore, for the purpose of testing the errors of the fluid model in number of redials and reconnects, we introduce the error measurements e_{RD} and e_{RC}, which are defined by

$$e_{RD} := \frac{\int_0^T |\mathbb{E}\lambda_{RD}(u) - \lambda_{RD}^{fl}(u)|du}{\int_0^T \mathbb{E}\lambda_{RD}(u)du} = \frac{\int_0^T |\mathbb{E}Z_{RD}(u) - z_{RD}(u)|du}{\int_0^T \mathbb{E}Z_{RD}(u)du},$$

$$e_{RC} := \frac{\int_0^T |\mathbb{E}\lambda_{RC}(u) - \lambda_{RC}^{fl}(u)|du}{\int_0^T \mathbb{E}\lambda_{RC}(u)du} = \frac{\int_0^T |\mathbb{E}Z_{RC}(u) - z_{RC}(u)|du}{\int_0^T \mathbb{E}Z_{RC}(u)du},$$



(a) $1/\delta_{RD} = 40$ and $1/\delta_{RC} = 50$. (b) $1/\delta_{RD} = 5$ and $1/\delta_{RC} = 10$.

Fig. 6. Simulation results (black solid curve), fluid approximations (red dashed curve) and stationary fluid limit (blue dot-dashed curve), $\lambda = 40$, $\hat{\rho} = 1.2$. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 1
Approximation errors in a single interval, $1/\delta_{RD} = 40$, $1/\delta_{RC} = 50$.

$\hat{\rho}$	$\lambda = 10$		$\lambda = 40$	
	e_{RD}	e_{RC}	e_{RD}	e_{RC}
1.01	91.1%	5.4%	82.9%	2.4%
1.05	45.8%	2.3%	27.0%	1.2%
1.1	19.6%	1.9%	7.8%	0.8%
1.2	7.2%	2.1%	1.3%	0.7%
1.3	1.8%	1.2%	0.8%	0.5%
1.4	1.4%	1.6%	0.4%	0.5%
1.5	1.5%	1.9%	0.6%	0.8%

Table 2
Approximation errors in a single interval, $1/\delta_{RD} = 5$, $1/\delta_{RC} = 10$.

$\hat{\rho}$	$\lambda = 10$		$\lambda = 40$	
	e_{RD}	e_{RC}	e_{RD}	e_{RC}
1.01	85.7%	5.6%	74.3%	1.9%
1.05	39.9%	4.0%	21.2%	1.1%
1.1	16.3%	3.1%	5.4%	1.4%
1.2	5.6%	3.0%	2.8%	1.3%
1.3	4.4%	3.6%	1.8%	1.2%
1.4	2.7%	2.5%	1.6%	1.3%
1.5	2.6%	2.6%	1.3%	1.5%

where $\lambda_{RD}^{fl}(t)$ and $\lambda_{RC}^{fl}(t)$ are the arrival rate due to redial and reconnect in the fluid approximation, respectively, and $T = 480$, as the same length of the simulation time. The parameters and results are shown in Table 1.

One can see from Tables 1 and 2 that for the number of reconnects, the fluid model offers good approximations for both scenarios with all values of $\hat{\rho}$. However, for the number of redials, the fluid model performs badly when $\hat{\rho} < 1.1$. This corresponds to the lingering condition pointed out by Mandelbaum et al. [15], which states that the fluid limit leads to significant inaccuracy when the system stays critically loaded (i.e., $\hat{\rho}$ close to 1 in our case) for a long time. In the next section, we will show that the consequences of these bad performances are not severe in terms of SL and AP. In addition, by comparing two different fresh arrival rates from Tables 1 and 2, we could see that the fluid approximation performs better for bigger call centers.

4.2. Validation of multiple intervals

Similar to the validation procedure in the case of a single interval, now we validate the performance of the fluid model for multiple intervals. We divide 480 min of simulation time into 16 intervals with duration 30 min. The fresh arrival rate λ_i is assumed to be piece-wise constants within each interval, but it varies from interval to interval. The fresh arrival pattern is shown in Fig. 7. This arrival pattern mimics the situation in reality, where there is a morning peak hour and an afternoon peak hour. We validate our approximation for different values of $\hat{\rho}$, and for given value of $\hat{\rho}$, $s_i = \frac{\hat{\rho}\mu(1-q)}{\lambda_i}$, for $i = 1, 2, \ldots, 16$. Other parameters are taken to be same as in the case of a single interval.

We omit the figure for $\mathbf{Z}(t)$, since they are similar to the graph in Fig. 6. The results for e_{RD} and e_{RC} are shown in Table 3.

Similar to Tables 1 and 2, one can see from Table 3 that the fluid model gives close approximations for the number of reconnects for all values of $\hat{\rho}$, and the approximations for the number of redials gets more accurate when $\hat{\rho} > 1.1$.

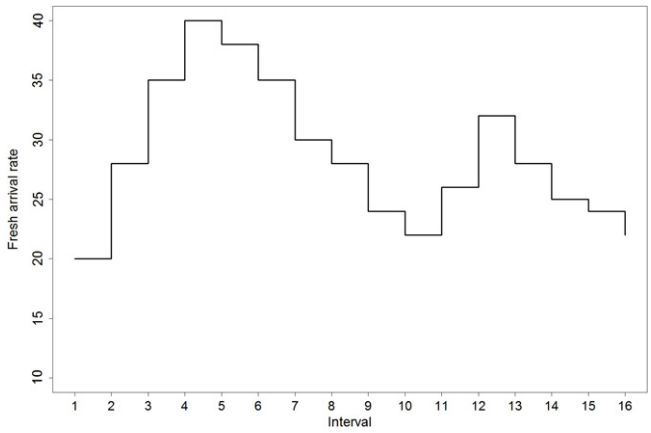


Fig. 7. Fresh arrival rate per interval.

Table 3
Approximation errors in multiple intervals.

$\hat{\rho}$	e_{RD}	e_{RC}	e_{RD}	e_{RC}
1.01	55.5%	1.8%	57.3%	2.1%
1.05	25.8%	1.4%	23.0%	2.0%
1.1	9.8%	1.2%	9.8%	1.2%
1.2	1.3%	0.8%	4.0%	1.6%
1.3	1.1%	0.9%	1.6%	1.5%
1.4	0.7%	0.8%	1.9%	1.7%
1.5	0.7%	0.9%	1.6%	1.6%

5. Erlang A approximation

The fluid model gives first order approximations for $Z_Q(t)$, $Z_{RD}(t)$ and $Z_{RC}(t)$. Based on them, we can approximate the expected total arrival rate and expected number of customers in the queue for any time t, from which the expected waiting time can be obtained. However, this is not the eventual goal, since it gives no information about the waiting time distribution of random customers, which is one of the most used call center performance indicators. Therefore, to this end, we will apply the Erlang A formula to approximate the waiting time distribution. We assume $\Lambda(t)$ to be the arrival rate of the Erlang A model, whose mean can be obtained via Eq. (1).

The reason to use the Erlang A model is intuitively clear, since the redial and reconnect behaviors have only direct influence on the total arrival rate, it has no direct influence on the service, such as the service durations. Therefore, once the total arrival rate $\Lambda(t)$ is given, $Z_{RD}(t)$ and $Z_{RC}(t)$ become irrelevant to what happens in the queue, thus, we can treat the system as an Erlang A system by ignoring the redial and reconnect orbits. Note that this is only an approximation of the Erlang A system, since the arrival process is generally not Poisson.

The analytical expressions for the waiting time distribution and the expected AP of the Erlang A model are known. We refer to Deslauriers et al. [30] and Roubos [31] for the Erlang A formula and the calculation details.

Now, we evaluate the performance of the Erlang A approximation. To save space, we only evaluate the performances in the case of multiple intervals. The arrival pattern is the same as shown in Fig. 7. Given all parameters, we compute $\mathbf{z}(t)$ via Eqs. (12)–(14). After that, $\Lambda(t)$ can be obtained via Eq. (1). $\Lambda(t)$ will be the input as the arrival rate of the Erlang A formula, from which the SL and AP can be obtained. We conduct such a procedure for different values of $\hat{\rho}$, δ_{RD} and δ_{RC}.

We denote SL^{sim} and SL^{a} as the SL from simulation and from Erlang A approximation, respectively. We let the acceptable waiting time to be 0.5 min. The AP from simulation and from the Erlang A approximation are denoted as AP^{sim} and AP^{a}, respectively. Besides the SL and AP, we also compare the probability of waiting from simulation, i.e., P_w^{sim}, with that from the Erlang A approximation, i.e., P_w^{a}. The results are shown in Tables 4 and 5. In Table 4, we let $1/\delta_{RD} = 40$ min and $1/\delta_{RC} = 50$ min, which are taken from a real call center data. In Table 5, we set $\delta_{RD} = 5$ min and $\delta_{RC} = 10$ min, which represents situations where customers spend short times in the redial and reconnect orbits. In both scenarios, we fix $\mu = 1/4$, $\theta = 0.5$, $p = 0.5$ and $q = 0.1$.

Based on the results in Tables 4 and 5, we can see that the Erlang A model offers close approximations both for the SL and AP in all values of $\hat{\rho}$. The approximation errors in probability of waiting is larger, but they are bounded by 5% in all scenarios. The Erlang A approximation performs better when $1/\delta_{RD}$ and $1/\delta_{RC}$ are larger, i.e., with error less than 2% in SL, and 1.2% in the AP in Table 4. However, even for small values of $1/\delta_{RD}$ and $1/\delta_{RC}$, the errors are bounded by 5% in SL and 2% in AP, as shown in Table 5. The approximation results in Table 4 are of special interests, since the parameters are taken to mimic real call centers. One might notice that even though we have large errors in e_{RD} when $\hat{\rho} < 1.1$ in Tables 1 and 2, the errors in SL

Table 4
Approximation errors of the Erlang A approximation, $1/\delta_{RD} = 40$ and $1/\delta_{RC} = 50$.

$\hat{\rho}$	SL^{sim}	SL^{a}	AP^{sim}	AP^{a}	P_w^{sim}	P_w^{a}
1.01	89.2%	92.3%	6.1%	4.9%	50.7%	46.5%
1.05	81.3%	84.6%	9.4%	8.4%	66.3%	65.2%
1.1	67.7%	69.8%	14.2%	13.7%	81.1%	82.9%
1.2	38.1%	37.4%	23.7%	23.8%	93.9%	96.3%
1.3	17.1%	15.2%	32.2%	32.2%	97.3%	99.1%
1.4	7.6%	5.6%	38.9%	39.1%	98.9%	99.7%
1.5	3.8%	2.2%	44.5%	44.6%	99.1%	99.9%

Table 5
Approximation errors of the Erlang A approximation, $1/\delta_{RD} = 5$ and $1/\delta_{RC} = 10$.

$\hat{\rho}$	SL^{sim}	SL^{a}	AP^{sim}	AP^{a}	P_w^{sim}	P_w^{a}
1.01	87.8%	91.7%	6.7%	5.3%	54.5%	50.6%
1.05	78.3%	82.7%	10.6%	9.4%	70.9%	71.3%
1.1	63.4%	65.6%	15.6%	15.3%	84.4%	88.3%
1.2	31.4%	29.6%	25.7%	25.8%	95.7%	95.6%
1.3	11.8%	9.0%	34.3%	34.4%	98.3%	99.9%
1.4	4.5%	2.4%	41.0%	41.3%	99.4%	99.9%
1.5	2.1%	0.7%	45.5%	46.8%	99.3%	99.9%

and AP are small in Tables 4 and 5. This is caused by the fact that when $\hat{\rho} < 1.1$, the number of redials is small compared to the number of reconnects, thus, the errors in number of redials do have big influence in $\Lambda(t)$.

6. Conclusion

In this paper, we investigate staffing of call centers with redials and reconnects. We consider call centers that operate under heavy load. The model can be described as a three-dimensional Markov process $\{\mathbf{Z}(t),\ t > 0\}$, defined in (2). However, to avoid the complexity of solving the Markov process, we use a fluid model to approximate $\mathbf{Z}(t)$. We show that the fluid limit is the unique solution of a set of three differential equations. Under the same fluid scaling, we derive the fluid limit of the queueing system in the non-stationary case to mimic the real situation in call centers, as the parameters can change before the system reaches stationarity. We also performed simulation experiments to assess the accuracy of the approximations. To apply the results to real call center applications, we take a further step by calculating the expected total arrival rate, and use this as an input to the Erlang A formula to calculate the SL and AP. Simulation results show that our approximation to SL is accurate with error less than 2%, and the approximation to AP has errors less than 1.5% when $\hat{\rho} \leq 1.05$ and less than 0.5% when $\hat{\rho} > 1.05$, when the parameters are taken from real data.

The results suggest a number of topics for further research. First, the current paper is focused on the derivation and usage of fluid limits for staffing problems of large call centers featuring both redials and reconnects, with load per server greater than 1. As a next step, it is interesting to supplement the results presented here with the development of staffing methods for the case where the load is strictly less than 1. To this end, the results of the present paper and the results for staffing large call centers without redials/reconnects [32,31,8] will serve as a good starting point. Second, with the presence of the redial and reconnect behaviors, it would be interesting to explicitly quantify the reduction of staffing costs while still meeting the target SL by more efficient planning of call center agents. Third, we use a first order approximation. It would also be interesting to derive the diffusion limit of this model, which suggests a second order approximation. This may lead to a more intuitive and simple staffing formula such as square-root staffing in the spirit of Halfin and Whitt [2]. Moreover, in this paper, we neglect the slight difference between the holding times of the reconnects and those of the fresh calls. As an extension to this, one could relax this assumption and study the correlation between the holding times of the fresh calls and its corresponding reconnects. Last but not the least, besides the influences in call centers staffing, the analysis of reconnect and redial behaviors can also offer insight to call center management. For example, by looking at the reconnect probability of each agent, managers can have some overview information of the quality of service offered by each agent. Furthermore, often the agents have some control on the holding time of each call, and by looking at the correlation between the reconnect probability and the holding time of each call, manager may find the "right" amount of holding time of each call, such that the holding time and the quality of service is well balanced.

Appendix A. Notations

Dividing by n on both sides of Eqs. (6)–(8), we have

$$\bar{Z}_Q^{(n)}(t) = \bar{Z}_Q^{(n)}(0) + G_Q^{(n)}\left(\bar{\mathbf{Z}}^{(n)}\right)(t) + \int_0^t H_Q\left(\bar{\mathbf{Z}}^{(n)}\right)(u)\,du, \tag{A.1}$$

$$\bar{Z}_{RD}^{(n)}(t) = \bar{Z}_{RD}^{(n)}(0) + G_{RD}^{(n)}\left(\bar{\mathbf{Z}}^{(n)}\right)(t) + \int_0^t H_{RD}\left(\bar{\mathbf{Z}}^{(n)}\right)(u)\,du, \tag{A.2}$$

$$\bar{Z}_{RC}^{(n)}(t) = \bar{Z}_{RC}^{(n)}(0) + G_{RC}^{(n)}\left(\bar{\mathbf{Z}}^{(n)}\right)(t) + \int_0^t H_{RC}\left(\bar{\mathbf{Z}}^{(n)}\right)(u)\,du, \tag{A.3}$$

where

$$\begin{aligned} G_Q^{(n)}\left(\bar{\mathbf{Z}}^{(n)}\right)(t) :=& \left(\frac{\Pi_{\lambda n}^{(n)}(t)}{n} - \lambda t\right) - \left(\bar{D}_s^{(n)}(t) - \int_0^t \mu \min\{s, \bar{Z}_Q^{(n)}(u)\}du\right) \\ &- \left(\bar{D}_a^{(n)}(t) - \int_0^t \theta\left(\bar{Z}_Q^{(n)}(u) - s\right)^+ du\right) + \left(\bar{D}_{RD}^{(n)}(t) - \int_0^t \delta_{RD}\bar{Z}_{RD}^{(n)}(u)\,du\right) \\ &+ \left(\bar{D}_{RC}^{(n)}(t) - \int_0^t \delta_{RC}\bar{Z}_{RC}^{(n)}(u)\,du\right), \end{aligned} \tag{A.4}$$

$$G_{RD}^{(n)}\left(\bar{\mathbf{Z}}^{(n)}\right)(t) := \left(\sum_{j=1}^{n\bar{D}_a^{(n)}(t)} B_j(p)/n - \int_0^t p\theta\left(\bar{Z}_Q^{(n)}(u) - s\right)^+ du\right) - \left(\bar{D}_{RD}^{(n)}(t) - \int_0^t \delta_{RD}\bar{Z}_{RD}^{(n)}(u)\,du\right), \tag{A.5}$$

$$G_{RC}^{(n)}\left(\bar{\mathbf{Z}}^{(n)}\right)(t) := \left(\sum_{j=1}^{n\bar{D}_s^{(n)}(t)} B_j(q)/n - \int_0^t q\mu \min\{s, \bar{Z}_Q^{(n)}(u)\}du\right) - \left(\bar{D}_{RC}^{(n)}(t) - \int_0^t \delta_{RC}\bar{Z}_{RC}^{(n)}(u)\,du\right), \tag{A.6}$$

and

$$\bar{D}_s^{(n)}(t) = \Pi_1\left(n\int_0^t \mu \min\{s, \bar{Z}_Q^{(n)}(u)\}du\right)\Big/ n, \tag{A.7}$$

$$\bar{D}_a^{(n)}(t) = \Pi_2\left(n\int_0^t \theta\left(\bar{Z}_Q^{(n)}(u) - s\right)^+ du\right)\Big/ n, \tag{A.8}$$

$$\bar{D}_{RD}^{(n)}(t) = \Pi_3\left(n\int_0^t \delta_{RD}\bar{Z}_{RD}^{(n)}(u)\,du\right)\Big/ n, \tag{A.9}$$

$$\bar{D}_{RC}^{(n)}(t) = \Pi_4\left(n\int_0^t \delta_{RC}\bar{Z}_{RC}^{(n)}(u)\,du\right)\Big/ n, \tag{A.10}$$

and

$$\begin{aligned} \int_0^t H_Q\left(\bar{\mathbf{Z}}^{(n)}\right)(u)\,du &:= \int_0^t \lambda + \delta_{RD}\bar{Z}_{RD}^{(n)}(u) + \delta_{RC}\bar{Z}_{RC}^{(n)}(u) - \mu\min\{s, \bar{Z}_Q^{(n)}(u)\} - \theta\left(\bar{Z}_Q^{(n)}(u) - s\right)^+ du, \\ \int_0^t H_{RD}\left(\bar{\mathbf{Z}}^{(n)}\right)(u)\,du &:= \int_0^t p\theta\left(\bar{Z}_Q^{(n)}(u) - s\right)^+ - \delta_{RD}\bar{Z}_{RD}^{(n)}(u)\,du, \\ \int_0^t H_{RC}\left(\bar{\mathbf{Z}}^{(n)}\right)(u)\,du &:= \int_0^t q\mu\min\{s, \bar{Z}_Q^{(n)}(u)\} - \delta_{RC}\bar{Z}_{RC}^{(n)}(u)\,du. \end{aligned}$$

For the convenience of notation, we rewrite Eqs. (A.1)–(A.3) in the vector form

$$\bar{\mathbf{Z}}^{(n)}(t) = \bar{\mathbf{Z}}^{(n)}(0) + \mathbf{G}^{(n)}\left(\bar{\mathbf{Z}}^{(n)}\right)(t) + \int_0^t \mathbf{H}\left(\bar{\mathbf{Z}}^{(n)}\right)(u)\,du, \tag{A.11}$$

where

$$\begin{aligned} \mathbf{G}^{(n)}\left(\bar{\mathbf{Z}}^{(n)}\right)(t) &:= \left(G_Q^{(n)}\left(\bar{\mathbf{Z}}^{(n)}\right)(t), G_{RD}^{(n)}\left(\bar{\mathbf{Z}}^{(n)}\right)(t), G_{RC}^{(n)}\left(\bar{\mathbf{Z}}^{(n)}\right)(t)\right)^T, \\ \mathbf{H}\left(\bar{\mathbf{Z}}^{(n)}\right)(u) &:= \left(H_Q\left(\bar{\mathbf{Z}}^{(n)}\right)(u), H_{RD}\left(\bar{\mathbf{Z}}^{(n)}\right)(u), H_{RC}\left(\bar{\mathbf{Z}}^{(n)}\right)(u)\right)^T. \end{aligned}$$

Appendix B. Proof of Lemma 1

In order to show that $\{\bar{\mathbf{Z}}^{(n)}(\cdot)\}_{n=1}^{\infty}$ is relatively compact with continuous limits, it is sufficient to show the following two properties (see Corollary 7.4 and Theorem 10.2 of [25]).

1. Compact Containment: for any $T \geq 0, \epsilon > 0$, there exists a compact set $\Gamma_T \subset \mathbb{R}^3$ such that
$$\mathrm{P}\left(\bar{\mathbf{Z}}^{(n)}(t) \in \Gamma_T,\ t \in [0,T]\right) \to 1, \quad \text{as } n \to \infty.$$
2. Oscillation Control: for any $\epsilon > 0$, and $T \geq 0$, there exists a $\delta > 0$, such that
$$\limsup_{n\to\infty} \mathrm{P}\left(\omega\left(\bar{\mathbf{Z}}^{(n)}, \delta, T\right) \geq \epsilon\right) \leq \epsilon, \tag{B.1}$$
where
$$\omega(\mathbf{x}, \delta, T) := \sup_{\substack{\nu,t\in[0,T]\\ |t-\nu|<\delta}} \max_{j\in J} |x_j(t) - x_j(\nu)|,$$
and $J := \{Q, RD, RC\}$.

Proof of Compact Containment property:

The following trivial upper bound holds for the total number of customers in the system (only arrivals are taken into account and no departures): for $t \in [0,T]$,
$$\bar{Z}_Q^{(n)}(t) + \bar{Z}_{RD}^{(n)}(t) + \bar{Z}_{RC}^{(n)}(t) \leq \bar{Z}_Q^{(n)}(0) + \bar{Z}_{RD}^{(n)}(0) + \bar{Z}_{RC}^{(n)}(0) + \Pi_{\lambda n}^{(n)}(T)/n.$$
Since $\Pi_{\lambda n}^{(n)}(\cdot)$ is a Poisson process of rate λn, by the Law of Large Numbers (LLN), we have
$$\Pi_{\lambda n}^{(n)}(T)/n \xrightarrow{d} \lambda T \quad \text{as } n \to \infty.$$
By the assumption of Theorem 1, we have
$$\bar{Z}_Q^{(n)}(0) + \bar{Z}_{RD}^{(n)}(0) + \bar{Z}_{RC}^{(n)}(0) \xrightarrow{d} z_Q(0) + z_{RD}(0) + z_{RC}(0).$$
Hence
$$\mathrm{P}(\bar{\mathbf{Z}}^{(n)}(t) \in \Gamma_T,\ t \in [0,T]) \to 1 \quad \text{as } n \to \infty,$$
where $\Gamma_T = \{(x_1,x_2,x_3) \mid x_1+x_2+x_3 \leq z_Q(0) + z_{RD}(0) + z_{RC}(0) + \lambda T + 1,\ x_1,x_2,x_3 \geq 0\}$, and the compact containment property indeed holds.

Proof of Oscillation Control property:

It follows from Eqs. (6)–(8) that, for all $\nu, t \geq 0$,
$$|\bar{Z}_Q^{(n)}(t) - \bar{Z}_Q^{(n)}(\nu)| \leq |\Pi_{\lambda n}^{(n)}(t) - \Pi_{\lambda n}^{(n)}(\nu)|/n + \sum_{j\in\{s,a,RD,RC\}} |\bar{D}_j^{(n)}(t) - \bar{D}_j^{(n)}(\nu)|,$$
$$|\bar{Z}_{RD}^{(n)}(t) - \bar{Z}_Q^{(n)}(\nu)| \leq \sum_{j\in\{a,RD\}} |\bar{D}_j^{(n)}(t) - \bar{D}_j^{(n)}(\nu)|,$$
$$|\bar{Z}_{RC}^{(n)}(t) - \bar{Z}_Q^{(n)}(\nu)| \leq \sum_{j\in\{s,RC\}} |\bar{D}_j^{(n)}(t) - \bar{D}_j^{(n)}(\nu)|,$$
where the processes $\bar{D}_j^{(n)}(\cdot)$ are defined by (A.8)–(A.10).

Also, from the Compact Containment property, we know that there exists a finite constant V such that
$$\mathrm{P}\left(\underbrace{\bar{Z}_Q^{(n)}(u), \bar{Z}_{RD}^{(n)}(u), \bar{Z}_{RC}^{(n)}(u) \leq V,\ u \in [0,T]}_{=:\Omega_n}\right) \to 1 \quad \text{as } n \to \infty.$$
On the event Ω_n, the following inequalities hold for all $\nu, t \in [0,T]$ such that $|t-\nu| \leq \delta$:
$$\int_\nu^t \mu \min\{s, \bar{Z}_Q^{(n)}(u)\}du \leq c_1\delta, \quad c_1 := \mu s,$$
$$\int_\nu^t \theta\left(\bar{Z}_Q^{(n)}(u) - s\right)^+ du \leq c_2\delta, \quad c_2 := \theta V,$$
$$\int_\nu^t \delta_{RD}\bar{Z}_{RD}^{(n)}(u)du \leq c_3\delta, \quad c_3 := \delta_{RD}V,$$
$$\int_\nu^t \delta_{RC}\bar{Z}_{RC}^{(n)}(u)du \leq c_4\delta, \quad c_4 := \delta_{RC}V.$$

Employing formulas (A.7)–(A.9), we then get

$$\mathrm{P}\left(\omega\left(\bar{\mathbf{Z}}^{(n)}, \delta, T\right) \geq \epsilon\right) \leq \mathrm{P}\left(\Omega_n'\right) + \mathrm{P}\left(\omega\left(\Pi_{\lambda n}^{(n)}(\cdot)/n, \delta, T\right) \geq \epsilon/5\right) + \sum_{j=1}^{4} \mathrm{P}\left(\omega\left(\Pi_j(n\cdot)/n, c_j\delta, c_jT\right) \geq \epsilon/5\right),$$

where

$$\omega\left(\Pi_{\lambda n}^{(n)}(\cdot)/n, \delta, T\right) \xrightarrow{d} \lambda\delta, \qquad \omega\left(\Pi_j(n\cdot)/n, c_j\delta, c_jT\right) \xrightarrow{d} c_j\delta, \qquad 1 \leq j \leq 4,$$

by the LLN for the Poisson processes $\Pi_{\lambda n}^{(n)}(\cdot)/n$, $\Pi_j(n\cdot)/n$ and by the continuity of the moduli of continuity $\omega(x(\cdot), \delta, T)$, $\omega(x(\cdot), c_j\delta, c_jT)$ with respect to $x(\cdot)$.

By the last two displays, the oscillation control property (B.1) indeed holds with any δ such that $\lambda\delta < \epsilon/5$, $c_j\delta < \epsilon/5$, $1 \leq j \leq 4$.

Appendix C. Proof of Theorem 1

In Lemma 1, we have shown that the sequence $\{\bar{\mathbf{Z}}^n(\cdot)\}_{n=1}^{\infty}$ is relatively compact with continuous limits, that is, from any subsequence $\{\bar{\mathbf{Z}}^{n_k}(\cdot)\}_{k=1}^{\infty}$, we can extract another subsequence $\{\bar{\mathbf{Z}}^{n_{k_l}}(\cdot)\}_{l=1}^{\infty}$ that converges weakly in $D([0,\infty), \mathbb{R}^3)$, say to a continuous process $z^*(t)$. We then call $z^*(t)$ a particular limit of the original sequence $\{\bar{\mathbf{Z}}^n(\cdot)\}_{n=1}^{\infty}$.

Consider an arbitrary particular limit $\mathbf{z}^*(\cdot)$ along a subsequence $\{\bar{\mathbf{Z}}^{n_k}(\cdot)\}_{k=1}^{\infty}$. If we can show that $\mathbf{z}^*(\cdot)$ satisfies Eqs. (9)–(11), and Eqs. (9)–(11) have a unique solution, then, due to the arbitrariness of $\mathbf{z}^*(\cdot)$, there must be a unique fluid limit defined by Eqs. (9)–(11).

We have

$$\bar{\mathbf{Z}}^{n_k}(\cdot) - \bar{\mathbf{Z}}^{n_k}(0) - \int_0^{\cdot} \mathbf{H}\left(\bar{\mathbf{Z}}^{n_k}\right) du = \mathbf{G}^{n_k}(\cdot). \tag{C.1}$$

On one hand, since $\bar{\mathbf{Z}}^{n_k}(\cdot) \xrightarrow{d} \mathbf{z}^*(\cdot)$ as $k \to \infty$ and the limit $\mathbf{z}^*(\cdot)$ is continuous,

$$\bar{\mathbf{Z}}^{n_k}(\cdot) - \bar{\mathbf{Z}}^{n_k}(0) - \int_0^{\cdot} \mathbf{H}\left(\bar{\mathbf{Z}}^{n_k}\right) du \xrightarrow{d} \mathbf{z}^*(\cdot) - \mathbf{z}(0) - \int_0^{\cdot} \mathbf{H}\left(\mathbf{z}^*\right) du$$

by the continuous mapping theorem.

On the other hand, below we show that $\mathbf{G}^{n_k}(\cdot) \xrightarrow{d} 0$, and then (C.1) implies that

$$\bar{\mathbf{Z}}^{n_k}(\cdot) - \bar{\mathbf{Z}}^{n_k}(0) - \int_0^{\cdot} \mathbf{H}\left(\bar{\mathbf{Z}}^{n_k}\right) du \xrightarrow{d} 0.$$

As we combine the last two displays together, it follows that the particular limit $\mathbf{z}^*$ a.s. satisfies Eqs. (9)–(11). Also, the mapping $\mathbf{H}$ is Lipschitz continuous and then, by Lemma 1 in [33], Eqs. (9)–(11) have a unique solution. Hence, all particular fluid limits are the same, namely they coincide with the unique solution to Eqs. (9)–(11).

It is left to show that $\mathbf{G}^{n_k}(\cdot) \xrightarrow{d} 0$.

By the LLN,

$$\Pi_1(n\cdot)/n - \cdot \xrightarrow{d} 0 \quad \text{in } D([0,\infty), \mathbb{R}),$$

and also, since $\bar{\mathbf{Z}}^{n_k}(\cdot) \xrightarrow{d} \mathbf{z}^*(\cdot)$ and $\mathbf{z}^*$ is continuous,

$$\int_0^{\cdot} \mu \min\{s, \bar{Z}_Q^{(n_k)}(u)\} du \xrightarrow{d} \int_0^{\cdot} \mu \min\{s, \mathbf{z}^*(u)\} du \quad \text{in } D([0,\infty), \mathbb{R}).$$

Then, by (A.7) and the Random time change theorem in [34],

$$\bar{D}_s^{(n_k)}(t) - \int_0^t \mu \min\{s, \bar{Z}_Q^{(n_k)}(u)\} du \xrightarrow{d} 0 \quad \text{in } D([0,\infty), \mathbb{R}).$$

By the same argument, one can show that the other terms in $G_Q^{(n_k)}(\cdot)$ converge to 0, and that $G_{RC}^{(n_k)}(\cdot)$, $G_{RD}^{(n_k)}(\cdot)$ converge to 0, too. Hence, the proof of Theorem 1 is finished.

References

[1] N. Gans, G. Koole, A. Mandelbaum, Telephone call centers: Tutorial, review, and research prospects, Manuf. Serv. Oper. Manage. 5 (2) (2003) 79–141.
[2] S. Halfin, W. Whitt, Heavy-traffic limits for queues with many exponential servers, Oper. Res. 29 (3) (1981) 567–588.
[3] O. Garnett, A. Mandelbaum, M. Reiman, Designing a call center with impatient customers, Manuf. Serv. Oper. Manage. 4 (3) (2002) 208–227.
[4] S. Aguir, Z. Akşin, F. Karaesmen, Y. Dallery, On the interaction between retrials and sizing of call centers, European J. Oper. Res. 191 (2) (2008) 398–408.
[5] T. Phung-Duc, H. Masuyama, S. Kasahara, Y. Takahashi, A matrix continued fraction approach to multiserver retrial queues, Ann. Oper. Res. 202 (1) (2013) 161–183.
[6] T. Phung-Duc, K. Kawanishi, Performance analysis of call centers with abandonment, retrial and after-call work, Perform. Eval. 80 (2014) 43–62.
[7] T. Phung-Duc, K. Kawanishi, Multiserver retrial queues with after-call work, Numer. Algebra Control Optim. 1 (4) (2011) 639–656.
[8] D. Sze, Or practice—a queueing model for telephone operator staffing, Oper. Res. 32 (2) (1984) 229–249.
[9] G.I. Falin, J.G.C. Templeton, Retrial Queues, Springer, 1997.
[10] Y. Liu, W. Whitt, Stabilizing performance in many-server queues with time-varying arrivals and customer feedback. Technical report, Working paper, 2014.
[11] G. Yom-Tov, A. Mandelbaum, Erlang-R: A time-varying queue with reentrant customers, in support of healthcare staffing, Manuf. Serv. Oper. Manage. 16 (2) (2014) 283–299.
[12] S. Ding, R.D. van der Mei, G. Koole, A method for estimation of redial and reconnect probabilities in call centers, in: Proceedings of the 2013 Winter Simulation Conference, Winter Simulation Conference, 2013, pp. 181–192.
[13] A. Roubos, G. Koole, R. Stolletz, Service-level variability of inbound call centers, Manuf. Serv. Oper. Manage. 14 (3) (2012) 402–413.
[14] W. Whitt, Fluid models for multiserver queues with abandonments, Oper. Res. 54 (1) (2006) 37–54.
[15] A. Mandelbaum, W. Massey, M. Reiman, A. Stolyar, B. Rider, Queue lengths and waiting times for multiserver queues with abandonment and retrials, Telecommun. Syst. 21 (2–4) (2002) 149–171.
[16] A. Mandelbaum, W. Massey, M. Reiman, B. Rider, Time varying multiserver queues with abandonment and retrials, in: Proceedings of the 16th International Teletraffic Conference, Vol. 4, 1999, pages 4–7.
[17] A. Mandelbaum, W. Massey, M. Reiman, Strong approximations for Markovian service networks, Queueing Syst. 30 (1–2) (1998) 149–201.
[18] S. Aguir, F. Karaesmen, Z. Akşin, F. Chauvet, The impact of retrials on call center performance, OR Spectrum 26 (3) (2004) 353–376.
[19] R. Ibrahim, W. Whitt, Real-time delay estimation based on delay history, Manuf. Serv. Oper. Manage. 11 (3) (2009) 397–415.
[20] R. Ibrahim, W. Whitt, Wait-time predictors for customer service systems with time-varying demand and capacity, Oper. Res. 59 (5) (2011) 1106–1118.
[21] W. Massey, J. Pender, Skewness variance approximation for dynamic rate multiserver queues with abandonment, ACM SIGMETRICS Perform. Eval. Rev. 39 (2) (2011) 74.
[22] J. Pender, W. Massey, Approximating and stabilizing dynamic rate Jackson networks with abandonment. Technical Report, 2014, Available via http://www.columbia.edu/~jp3404/Jackson_Stabilize_GVA.pdf (accessed 31.03.15).
[23] G. Pang, W. Whitt, The impact of dependent service times on large-scale service systems, Manuf. Serv. Oper. Manage. 14 (2) (2012) 262–278.
[24] G. Pang, R. Talreja, W. Whitt, Martingale proofs of many-server heavy-traffic limits for Markovian queues, Probab. Surv. 4 (193–267) (2007) 7.
[25] S. Ethier, T. Kurtz, Markov Processes. Characterization and Convergence, John Wiley and Sons, NY, 1986.
[26] H. Shen, J. Huang, Interday forecasting and intraday updating of call center arrivals, Manuf. Serv. Oper. Manage. 10 (3) (2008) 391–410.
[27] R. Ibrahim, P. L'Ecuyer, Forecasting call center arrivals: Fixed-effects, mixed-effects, and bivariate models, Manuf. Serv. Oper. Manage. 15 (1) (2013) 72–85.
[28] Z. Feldman, A. Mandelbaum, W. Massey, W. Whitt, Staffing of time-varying queues to achieve time-stable performance, Manage. Sci. 54 (2) (2008) 324–338.
[29] O. Jennings, A. Mandelbaum, W. Massey, W. Whitt, Server staffing to meet time-varying demand, Manage. Sci. 42 (10) (1996) 1383–1394.
[30] A. Deslauriers, P. L'Ecuyer, J. Pichitlamken, A. Ingolfsson, A. Avramidis, Markov chain models of a telephone call center with call blending, Comput. Oper. Res. 34 (6) (2007) 1616–1645.
[31] A. Roubos, Service-level variability and impatience in call cneters (Ph.D. thesis), Free University Amsterdam, 2012.
[32] S. Borst, A. Mandelbaum, M. Reiman, Dimensioning large call centers, Oper. Res. 52 (1) (2004) 17–34.
[33] J. Reed, A. Ward, A diffusion approximation for a generalized jackson network with reneging, in: Proceedings of the 42nd Annual Allerton Conference on Communication, Control, and Computing, 2004.
[34] P. Billingsley, Convergence of Probability Measures. Vol. 493, Wiley-Interscience, 2009.

Sihan Ding is a Ph.D. student in group Stochastics in Center for Mathematics and Computer Science. He received a Bachelor of Mathematics and Applied Mathematics in Hunan University in 2009, a Master of Mathematics in Leiden University in 2011. He also works in the Business Analytics group at the VU University Amsterdam one day per week. His research interests include service operations management, forecasting and stochastic models.

Maria Remerova is a Post-Doc at the University of Amsterdam. She received her Ph.D. degree at VU University Amsterdam in 2014 for her thesis entitled "Fluid Limit Approximations of Stochastic Networks". In 2013, she worked in Eindhoven University of Technology as a lecturer. Her research interests are stochastic networks with applications in service systems.

Rob van der Mei is the leader of the research theme Logistics and the Industrial Liaison Officer at CWI, and a full professor at the VU University, Amsterdam. Before going to academia, he has been working for over a decade as a consultant and researcher in the ICT industry, working for PTT, KPN, AT&T Labs and TNO ICT. His research interests include queueing theory, performance analysis of ICT systems, health logistics, road traffic management, logistics, grid computing, revenue management, military operations research, sensor networks, call centers and data science. He is the co-author of some 130 papers in journals and refereed proceedings.

Bert Zwart is a researcher at Center for Mathematics and Computer Science, where he leads the Stochastic group. He also holds secondary positions at Eindhoven University of Technology (Professor), Georgia Tech (Adjunct Professor) and the Dutch research center on Stochastics, Eurandom. Before that he was holding a Coca-Cola Chair at the H. Milton Stewart School of Industrial and Systems Engineering at Georgia Institute of Technology. His research is in applied probability and stochastic operations research, inspired by problems in computer, communication, energy and service networks. He is the 2008 recipient of the Erlang prize for outstanding contributions to applied probability by a researcher not older than 35 years old, an IBM faculty award, VENI and VIDI awards from the Dutch Science Foundation NWO, numerous best papers awards, and co-authored more than 100 refereed publications.