Fluid Approximation of a Call Center Model with Redials and Reconnects

Sihan Ding, Maria Remerova, Rob van der Mei, Bert Zwart

November 9, 2018

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 ignoring them 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 model for the purpose of calculating service levels and abandonment rates. The performance of such a procedure is validated in the case of single intervals as well as multiple intervals with changing parameters.

1 Introduction

Nowadays, call centers are important means of communication with the customer. Therefore, the response-time performance of call centers is crucial for the customer satisfaction. For call center managers, making the right staffing decisions (i.e., decide on the right number of agents) is essential to the costs and the performances of call centers. Various models have been developed in order to decide on the right number of agents. One of the most widely used models is the Erlang C model and there is a lot of literature on it (see Gans et al. [2002] and the references therein). A staffing rule called the square-root staffing is proposed by Halfin and Whitt [1981]. Garnett et al. [2002] show that the square-root staffing rule remains valid for the Erlang A model. However, both the Erlang C formula and the square-root staffing formula ignore customer redial (a re-attempt after an abandoned call) behaviors in call centers, while this behavior is quite significant (see Gans et al. [2002] and reference therein). Aguir et al. [2008] discover that ignoring redials can lead to under-staffing or over-staffing, depending on the forecasting assumption being made. Sze [1984] studies a model where abandonments and redials are included, focusing on the traffic loaded systems.

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 Gans et al. [2002], where the reconnect is defined as a revisit. In Ding et al. [2013],

we use real call center data to show that an inbound call can either be a fresh call (a initial attempt), a redial or a reconnect. Also, as argued in Ding et al. [2013], redials and reconnects should be considered and modeled, since ignoring them can lead to significantly inaccurate estimations 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.

Intuitively, when the system is heavily loaded, it would lead to bad service levels (SL). However, for large call centers, especially during the busy hours when the inbound volume is quite large, it is possible that the target SL can be met even in heavy traffic. Further discussions of this effect can be found in Garnett et al. [2002] and Borst et al. [2004].

In this paper, we aim to answer the following question: "In large call centers, what are the SL and the abandonment percentage (AP) if both redialling and reconnection of customers are taken into account?" To this end, 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 Ding et al. [2013]). This dependency effect between the total number of arrivals and the number of agents becomes more complicated in real life, due to the fact that the fresh arrival rate 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 discussion about the SL variability, we refer to the work by Roubos et al. [2012].

Fluid models for call centers have been extensively studied. Whitt [2006] gave an intuitive explanation of the fluid model. He 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 Whitt [2006], the redial behavior is not considered, though. The existence and uniqueness of the fluid limit are given as conjectures. Mandelbaum et al. [2002] 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 Mandelbaum et al. [1999], the authors use the fluid and the diffusion approximation for time varying multiserver queue with abandonment and retrials. They show that the fluid and the diffusion approximation can both be obtained by solving sets of non-linear differential equations, where the diffusion process can provide the confidence bounds for the fluid approximation. The work by Mandelbaum et al. [1998] gives more general theoretical results for the fluid and diffusion approximation for

Markovian service networks. Aguir et al. [2004] extend the model by allowing customer balking behavior, but no formal proof of the fluid limit is given. Besides the application in call center staffing problems, fluid models have also been applied in delay announcement of customers in call centers (see Ibrahim and Whitt [2009, 2011]).

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 that the original stochastic model converges to the fluid model under a proper scaling. We numerically compute the fluid model, and simulate the original model, and compare them in the case of a single interval and multiple intervals, where the parameters are changing per interval and would 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 Figure 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, he 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 $\mathbb{E}\Gamma_{RD} = \delta_{RD} < \infty$. 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 service time B of a customer has an exponential distribution with mean $\mathbb{E}B = 1/\mu < \infty$. After the call has been served, this customer will enter the reconnect orbit with probability q, and he will reconnect after an exponentially distributed time Γ_{RC}, with $\mathbb{E}\Gamma_{RC} = \delta_{RC} < \infty$. 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 time, waiting time and the number 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 Pang and Whitt [2012].

3   Fluid limit approximations

In this section, we first show that the problem of calculating the expected total arrival rate drills 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. Because an arrival can be a fresh arrival, a redial or a reconnect, the following equation must hold

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

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

Figure 1: Call diagram

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 Equation (1). Note that $Z_Q(t)$ does not appear in Equation (1), but we will see later that $Z_{RD}(t)$ and $Z_{RC}(t)$ would depend on $Z_Q(t)$.

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

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

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, by truncating at some large states, this method offers almost exact results. However, for the model we consider, it is very difficult to formulate and solve the global balance equations, and their solution offers no insight about the system. Therefore, for the convenience of practical usage, we will not consider solving this Markov process, but other approximation methods.

3.1 Fluid limit

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

Consider a single interval with the fresh arrival rate remaining constant during this interval (e.g., $\lambda(t) = \lambda$, $t \geq 0$). 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 time $[0,t)$, respectively. $B_j(p)$ is a Bernoulli random variable with success probability p, $j = 1, 2, \ldots, D_a(t)$. $B_j(p) = 1$, if the jth abandoned customer enter the redial orbit; $B_j(p) = 0$, otherwise. Therefore, for given $D_a(t)$, $\sum_{j=1}^{D_a(t)} B_j(p) \sim \mathrm{Bin}(D_a(t), p)$. By the same argument, we have $\sum_{j=1}^{D_s(t)} B_j(q) \sim \mathrm{Bin}(D_s(t), q)$.

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

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

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

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

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

Rigorous proof of these four statements can be given along the lines of Pang et al. [2007], see Lemma 2.1.

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

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

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 $X^{(n)}$ converge 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 Equations (6)-(8)) for a single interval, we divide both sides of Equations (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 A.2. $\qquad\qquad\square$

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

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

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

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

Proof. See Appendix A.3. $\qquad\qquad\square$

We could not obtain analytic expressions of $z_Q(t), z_{RD}(t)$ and $z_{RC}(t)$ from Equations (9)-(11). However, we can solve them numerically by the following iterative procedure. First, Equations (9)-(11) can be rewritten as

$$\mathbf{z}(t) = \Phi(\mathbf{z}(t)).$$

We let $\mathbf{z}^{(0)}(0) = 0$, and then calculate $\mathbf{z}^{(k+1)} = \Phi(\mathbf{z}^{(k)})$, $k = 0, 1, \ldots$, until the difference between $\mathbf{z}^{(k+1)}$ and $\mathbf{z}^{(k)}$ is small enough.

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 [2008] and Ibrahim and L'Ecuyer [2013], 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. Therefore, we now show the fluid limit for multiple intervals, where λ and s vary vary from interval to interval. We assume that other parameters remain constant.

We divide a day 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, $i = 1, 2, \ldots, m$. For the ith interval, i.e., $t_{i-1} \leq t < t_i$, the fluid limit then becomes

$$z_Q(t) = z_Q(t_{i-1}) + \lambda_i (t - t_{i-1}) + \delta_{RD} \int_{t_{i-1}}^t z_{RD}(u)du + \delta_{RC} \int_{t_{i-1}}^t z_{RC}(u)du$$

$$- \mu \int_{t_{i-1}}^t \min\{s_i, z_Q(u)\}du - \theta \int_{t_{i-1}}^t (z_Q(u) - s_i)^+ du, \tag{12}$$

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

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

Numerically solving Equations (12)-(14) is similar to solving Equations (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 Equations (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 fluid limit in stationarity, i.e., we develop conditions under which $\mathbf{z}(t)$ is constant.

By differentiating Equations (9)-(11) and taking into account that $\frac{d}{dt}\mathbf{z}(t) = 0$ for a constant solution, we obtain

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

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

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

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

7

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

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

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

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

The results above would offer some insights. $\rho := \frac{\lambda}{s\mu}$ is the load of the system due to the fresh arrivals. Since $\frac{1}{1-q}$ portion of ρ will reconnect, the total load would be $\hat{\rho} := \frac{\lambda}{(1-q)s\mu}$, if there were no redials. One can notice that in expressions (18)-(20), the value of $\hat{\rho}$ determines whether the fluid model is in heavy traffic or not. Thus, from now on, we use $\hat{\rho}$ to denote the actual load of the system instead of ρ. In the case of $\hat{\rho} < 1$, since the fluid limit is deterministic, $z_Q(\infty) < s$, and $z_{RD}(\infty) = 0$ would hold. This means there is no abandonment at all in the fluid limit when $\hat{\rho} < 1$. In reality, due to the variability of the service duration and patience, abandonments would not be 0 though, but very small. If $\hat{\rho} > 1$, by Equation (18), $z_Q(\infty) > s$. Therefore, in this case, the fluid model indicates that there will be $(z_Q(\infty) - s)$ amount of customers waiting, each with rate θ, and customers will go to the redial orbit with rate $p\theta(z_Q(\infty) - s)$.

4 Validation of the fluid limit

In this section, we will validate the fluid model via simulation both for a single interval and for multiple intervals. We simulate the system for 480 minutes of time, i.e., 8 hours, 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 Equations (9)-(11)) with $\mathbf{Z}(t)$ (simulation results) for different values of $\hat{\rho}$. For each value of $\hat{\rho}$, s changes, while λ and μ remain the same. One example of $\mathbf{z}(t)$ and $\mathbf{Z}(t)$, where $\hat{\rho} = 1.2$, is shown in Figure 2.

One can see from Figure 2 that the system starts with zero customers, and as time passes by, $Z_Q(t), Z_{RD}(t)$ and $Z_{RC}(t)$ gradually build up and reach stationarity. Furthermore, in this parameter setting, the fluid limit offers a close approximation of the original system, especially for $Z_{RQ}(t)$ and $Z_{RC}(t)$. The errors are bigger for $Z_{RD}(t)$.

Figure 2: $z_{RC}(t), z_Q(t), z_{RD}(t)$ (from top to bottom) via fluid limit (red dashed curve) and $Z_{RC}(t), Z_Q(t), Z_{RD}(t)$ via simulation (black solid curve), $\lambda = 40, \mu = 1/4, p = 0.5, q = 0.1, \theta = 1/2, s = 148, \delta_{RD} = 0.05, \delta_{RC} = 0.01$

Obtaining the approximation of $\mathbf{Z}(t)$ 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

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

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

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 Table 1 that for the number of reconnects, the fluid model offers good approximations for all scenarios. However, for the number of redials, the fluid model performs badly when $\hat{\rho} < 1.1$. In the next section, we will show that the consequences of these bad performances are not severe in terms of SL and AP. When $\hat{\rho} \geq 1.1$, the fluid model starts to get more accurate with $e_{RD} \leq 10.3\%$. In cases where $\hat{\rho} \geq 1.2$, e_{RD} is less then 1.1%.

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 minutes of simulation time into 16 intervals with duration 30 minutes. The fresh arrival rate is assumed to be

$\hat{\rho}$	s	e_{RD}	e_{RC}
1.01	176	92.5%	1.7%
1.05	169	35.7%	1.6%
1.1	162	10.3%	0.5%
1.2	148	1.9%	0.5%
1.3	137	1.3%	0.5%
1.4	127	1.4%	0.5%
1.5	119	1.1%	0.7%

Table 1: Approximation errors of different values of $\hat{\rho}$ in single intervals , $\lambda = 40, \mu = 1/4, p = 0.5, q = 0.1, \theta = 1/2, \delta_{RD} = 0.05, \delta_{RC} = 0.01$

piece-wise constants within each interval, but it varies from interval to interval. The fresh arrival pattern is shown in Figure 3. This arrival pattern mimics the situation in reality, where there is a morning peak hour and an afternoon peak hour.

Figure 3: Fresh arrival rate per interval

We omit the figure for $\mathbf{Z}(t)$, since they are similar to the graph in Figure 2. The results for e_{RD} and e_{RC} are shown in Table 2.

Similar to Table 1, one can see from Table 2 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 $\hat{\rho} > 1.1$.

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

$\hat{\rho}$	e_{RD}	e_{RC}
1.01	58.2%	1.7%
1.05	32.2%	1.4%
1.1	12.8%	0.9%
1.2	2.6%	0.5%
1.3	1.3%	0.4%
1.4	1.1%	0.3%
1.5	1.3%	0.7%

Table 2: Approximation errors of different values of $\hat{\rho}$ in multiple intervals, $\mu = 1/4, p = 0.5, q = 0.1, \theta = 1/2, \delta_{RD} = 0.05, \delta_{RC} = 0.01$.

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 distributions of a random customer, which is one of the most used call center performance indicators in call centers. 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, which can be obtained via Equation (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. [2007] and Roubos [2012] for details about the Erlang A formula and calculation details.

Next, we validate the Erlang A approximation of the original model. To save space, we only show the performances in the case of multiple intervals. The arrival pattern is the same as shown in Figure 3. For the given parameters, we compute $\mathbf{z}(t)$ via Equations (12)-(14). After that, $\Lambda(t)$ can be obtained via Equation (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. The SL is defined as percentage of customers that are answered within 30 seconds.

We denote SL^{sim} and SL^{a} as the SL from simulation and from Erlang A formula, respectively. In this section, the SL is specifically set to be the percentage of customers that waited less than 30 seconds. The AP from simulation and from Erlang A formula are denoted as AP^{sim} and AP^{a}, respectively. The comparisons are shown in Table 3.

Based on the results in Table 3, we can see that the Erlang A model offers a close approximation both for the SL and AP in all values of $\hat{\rho}$, with the error less than 3% in SL, and 1.1% in the AP.

One might notice that even though we have large errors in e_{RD} when $\hat{\rho} < 1.1$ in Table 1 and Table 2, the errors in SL and AP are small in Table 3. 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, errors in number of redials would not influence much in $\Lambda(t)$.

11

$\hat{\rho}$	SL^{sim}	SL^a	AP^{sim}	AP^a
1.01	95.5%	96.8%	5.7%	4.8%
1.05	90.2%	92.1%	9.1%	8.2%
1.1	79.7%	80.6%	14.2%	14.0%
1.2	54.9%	53.4%	24.2%	24.6%
1.3	36.0%	34.2%	32.9%	33.2%
1.4	28.0%	26.2%	39.8%	40.2%
1.5	24.9%	23.6%	45.6%	45.8%

Table 3: Approximation errors in SL and AP of different values of $\hat{\rho}$ in multiple intervals, $\mu = 1/4, p = 0.5, q = 0.1, \theta = 1/2, \delta_{RD} = 0.05, \delta_{RC} = 0.01$.

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 of the SL is accurate with error less than 2% in all scenarios, and approximation of AP has errors less than 1% when $\hat{\rho} \le 1.05$ and less than 0.5% when $\hat{\rho} > 1.05$.

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 Borst et al. [2004], Roubos [2012], Sze [1984] 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, 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.

Appendices

Appendix A.1: Notations

Dividing by n on both sides of Equations (6)-(8), we have

$$\bar{Z}_Q^{(n)}(t) = \bar{Z}_Q^{(n)}(0) + G_Q^{(n)}\left(\bar{\mathbf{Z}}^{(n)}\right)(t) + \int_0^t H_Q\left(\bar{\mathbf{Z}}^{(n)}\right)(u)\,du, \tag{21}$$

$$\bar{Z}_{RD}^{(n)}(t) = \bar{Z}_{RD}^{(n)}(0) + G_{RD}^{(n)}\left(\bar{\mathbf{Z}}^{(n)}\right)(t) + \int_0^t H_{RD}\left(\bar{\mathbf{Z}}^{(n)}\right)(u)\,du, \tag{22}$$

$$\bar{Z}_{RC}^{(n)}(t) = \bar{Z}_{RC}^{(n)}(0) + G_{RC}^{(n)}\left(\bar{\mathbf{Z}}^{(n)}\right)(t) + \int_0^t H_{RC}\left(\bar{\mathbf{Z}}^{(n)}\right)(u)\,du, \tag{23}$$

where

$$G_Q^{(n)}\left(\bar{\mathbf{Z}}^{(n)}\right)(t) := \left(\frac{\Pi_{\lambda n}^{(n)}(t)}{n} - \lambda t\right) - \left(\bar{D}_s^{(n)}(t) - \int_0^t \mu\min\{s, \bar{Z}_Q^{(n)}(u)\}du\right)$$

$$- \left(\bar{D}_a^{(n)}(t) - \int_0^t \theta\left(\bar{Z}_Q^{(n)}(u) - s\right)^+ du\right)$$

$$+ \left(\bar{D}_{RD}^{(n)}(t) - \int_0^t \delta_{RD}\bar{Z}_{RD}^{(n)}(u)\,du\right)$$

$$+ \left(\bar{D}_{RC}^{(n)}(t) - \int_0^t \delta_{RC}\bar{Z}_{RC}^{(n)}(u)\,du\right), \tag{24}$$

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

$$- \left(\bar{D}_{RD}^{(n)}(t) - \int_0^t \delta_{RD}\bar{Z}_{RD}^{(n)}(u)\,du\right), \tag{25}$$

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

$$- \left(\bar{D}_{RC}^{(n)}(t) - \int_0^t \delta_{RC}\bar{Z}_{RC}^{(n)}(u)\,du\right), \tag{26}$$

and

$$\bar{D}_s^{(n)}(t) = \Pi_1 \left(n \int_0^t \mu \min\{s, \bar{Z}_Q^{(n)}(u)\} du \right) / n, \tag{27}$$

$$\bar{D}_a^{(n)}(t) = \Pi_2 \left(n \int_0^t \theta \left(\bar{Z}_Q^{(n)}(u) - s \right)^+ du \right) / n, \tag{28}$$

$$\bar{D}_{RD}^{(n)}(t) = \Pi_3 \left(n \int_0^t \delta_{RD} \bar{Z}_{RD}^{(n)}(u) \, du \right) / n, \tag{29}$$

$$\bar{D}_{RC}^{(n)}(t) = \Pi_4 \left(n \int_0^t \delta_{RC} \bar{Z}_{RC}^{(n)}(u) \, du \right) / n, \tag{30}$$

and

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

$$\int_0^t H_{RD} \left(\bar{\mathbf{Z}}^{(n)} \right)(u) \, du := \int_0^t p\theta \left(\bar{Z}_Q^{(n)}(u) - s \right)^+ - \delta_{RD} \bar{Z}_{RD}^{(n)}(u) \, du,$$

$$\int_0^t H_{RC} \left(\bar{\mathbf{Z}}^{(n)} \right)(u) \, du := \int_0^t q\mu \min\{s, \bar{Z}_Q^{(n)}(u)\} - \delta_{RC} \bar{Z}_{RC}^{(n)}(u) \, du.$$

For the convenience of notation, we rewrite Equations (21)-(23) in the vector form

$$\bar{\mathbf{Z}}^{(n)}(t) = \bar{\mathbf{Z}}^{(n)}(0) + \mathbf{G}^{(n)} \left(\bar{\mathbf{Z}}^{(n)} \right)(t) + \int_0^t \mathbf{H} \left(\bar{\mathbf{Z}}^{(n)} \right)(u) \, du, \tag{31}$$

where

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

$$\mathbf{H} \left(\bar{\mathbf{Z}}^{(n)} \right)(u) := \left(H_Q \left(\bar{\mathbf{Z}}^{(n)} \right)(u), H_{RD} \left(\bar{\mathbf{Z}}^{(n)} \right)(u), H_{RC} \left(\bar{\mathbf{Z}}^{(n)} \right)(u) \right)^T.$$

Appendix A.2: Proof of Lemma 1

Proof. 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 Ethier and Kurtz [1986]).

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

$$\limsup_{n \to \infty} \mathrm{P} \left(\omega \left(\bar{\mathbf{Z}}^{(n)}, \delta, T \right) \geq \epsilon \right) \leq \epsilon, \tag{32}$$

where

$$\omega(\mathbf{x}, \delta, T) := \sup_{\substack{\nu, t \in [0,T] \\ |s-\nu|<\delta}} \max_{j \in J} |x_j(t) - x_j(\nu)|,$$

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 Equations (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 (28)-(30).

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 $|t - \nu| \leq \delta$:

$$\int_\nu^t \mu \min\{s, \bar{Z}_Q^{(n)}(u)\} du \leq \mu s \delta =: c_1(\delta),$$

$$\int_\nu^t \theta \left(\bar{Z}_Q^{(n)}(u) - s \right)^+ du \leq \theta V \delta =: c_2(\delta),$$

$$\int_\nu^t \delta_{RD} \bar{Z}_{RD}^{(n)}(u) du \leq \delta_{RD} V \delta =: c_3(\delta),$$

$$\int_\nu^t \delta_{RC} \bar{Z}_{RC}^{(n)}(u) du \leq \delta_{RC} V \delta =: c_4(\delta).$$

Employing formulas (27)-(29), we then get

$$\mathrm{P} \left(\omega \left(\bar{\mathbf{Z}}^{(n)}, \delta, T \right) \geq \epsilon \right) \leq \mathrm{P} \left(\omega \left(\Pi_{\lambda n}^{(n)}(\cdot)/n, \delta, T \right) \geq \epsilon/5 \right)$$

$$+ \sum_{j=1}^4 \mathrm{P} \left(\omega \left(\Pi_j(n \cdot)/n, c_j(\delta), c_j(\delta)T \right) \geq \epsilon/5 \right)$$

$$\to \lambda \delta + \sum_{j=1}^4 c_j(\delta) = \delta(\lambda + \mu s + \delta_{RD} V + \delta_{RC} V),$$

where the convergence holds 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 $\omega(x(\cdot), \delta, T)$, $\omega(x(\cdot), c_j(\delta), c_j(\delta)T)$ with respect to $x(\cdot)$.

Hence, the oscillation control property (32) indeed holds with $\delta = \epsilon/(\lambda + \mu s + \delta_{RD} V + \delta_{RC} V)$.

$\square$

Appendix A.3: Proof of Theorem 1

Proof. 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 Equations (9)-(11), and Equations (9)-(11) have a unique solution, then, due to the arbitrariness of $\mathbf{z}^*(\cdot)$, there must be a unique fluid limit defined by Equations (9)-(11).

We have

$$\bar{\mathbf{Z}}^{n_k}(\cdot) - \bar{\mathbf{Z}}^{n_k}(0) - \int_0^\cdot \mathbf{H} \left(\bar{\mathbf{Z}}^{n_k} \right) du = \mathbf{G}^{n_k}(\cdot). \tag{33}$$

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,

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

On the other hand, below we show that $\mathbf{G}^{n_k}(\cdot) \overset{d}{\to} 0$, and then (33) implies that

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

As we combine the last two displays together, it follows that the particular limit $\mathbf{z}^*$ a.s. satisfies Equations (9)-(11). Also, the mapping $\mathbf{H}$ is Lipschitz continuous and then, by Lemma 1 in Reed and Ward [2004], Equations (9)-(11) have a unique solution. Hence, all particular fluid limits are the same, namely they coincide with the unique solution to (9)-(11).

It is left to show that $\mathbf{G}^{n_k}(\cdot) \overset{d}{\to} 0$.
By the LLN,
$$\Pi_1(n\cdot)/n - \cdot \overset{d}{\to} 0 \text{ in } D([0,\infty), \mathbb{R}),$$

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

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

Then, by (27) and the Random time change Theorem in Billingsley [2009],

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

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. $\quad\square$

17

References

M.S. Aguir, F. Karaesmen, O.Z. Akşin, and F. Chauvet. The impact of retrials on call center performance. *OR Spectrum*, 26(3):353–376, 2004.

M.S. Aguir, O.Z. Akşin, F. Karaesmen, and Y. Dallery. On the interaction between retrials and sizing of call centers. *European Journal of Operational Research*, 191(2):398–408, 2008.

P. Billingsley. *Convergence of probability measures*, volume 493. Wiley-Interscience, 2009.

S. Borst, A. Mandelbaum, and M. Reiman. Dimensioning large call centers. *Operations research*, 52(1):17–34, 2004.

A. Deslauriers, P. L' Ecuyer, J. Pichitlamken, A. Ingolfsson, and A. Avramidis. Markov chain models of a telephone call center with call blending. *Computers & Operations Research*, 34(6):1616–1645, 2007.

S. Ding, R.D. van der Mei, and G. Koole. A method for estimation of redial and reconnect probabilities in call centers. In *Proceedings of the 2013 Winter Simulation Conference, to appear*. Winter Simulation Conference, 2013.

S. Ethier and T. Kurtz. *Markov processes. Characterization and convergence*. NY: John Willey and Sons, 1986.

N Gans, G Koole, and A Mandelbaum. Telephone calls centers: a tutorial and literature review. *European Jour. Oper. Res*, 2002.

O. Garnett, A. Mandelbaum, and M. Reiman. Designing a call center with impatient customers. *Manufacturing & Service Operations Management*, 4(3):208–227, 2002.

S. Halfin and W. Whitt. Heavy-traffic limits for queues with many exponential servers. *Operations research*, 29(3):567–588, 1981.

R. Ibrahim and P. L'Ecuyer. Forecasting call center arrivals: Fixed-effects, mixed-effects, and bivariate models. *Manufacturing & Service Operations Management*, 15(1):72–85, 2013.

Rouba Ibrahim and Ward Whitt. Real-time delay estimation based on delay history. *Manufacturing & Service Operations Management*, 11(3):397–415, 2009.

Rouba Ibrahim and Ward Whitt. Wait-time predictors for customer service systems with time-varying demand and capacity. *Operations research*, 59(5):1106–1118, 2011.

A. Mandelbaum, W. Massey, and M. Reiman. Strong approximations for markovian service networks. *Queueing Systems*, 30(1-2):149–201, 1998.

A. Mandelbaum, W. Massey, M. Reiman, and B. Rider. Time varying multiserver queues with abandonment and retrials. In *Proceedings of the 16th International Teletraffic Conference*, volume 4, pages 4–7, 1999.

A. Mandelbaum, W. Massey, M. Reiman, A. Stolyar, and B. Rider. Queue lengths and waiting times for multiserver queues with abandonment and retrials. *Telecommunication Systems*, 21(2-4):149–171, 2002.

G. Pang, R. Talreja, and W. Whitt. Martingale proofs of many-server heavy-traffic limits for markovian queues. *Probability Surveys*, 4(193-267):7, 2007.

Guodong Pang and Ward Whitt. The impact of dependent service times on large-scale service systems. *Manufacturing & Service Operations Management*, 14(2):262–278, 2012.

J. Reed and A.R. 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.

A. Roubos. *Service-Level Variability and Impatience in Call Cneters*. PhD thesis, Free University Amsterdam, 2012.

Alex Roubos, Ger Koole, and Raik Stolletz. Service-level variability of inbound call centers. *Manufacturing & Service Operations Management*, 14(3):402–413, 2012.

H. Shen and J. Huang. Interday forecasting and intraday updating of call center arrivals. *Manufacturing & Service Operations Management*, 10(3):391–410, 2008.

D. Sze. Or practice - a queueing model for telephone operator staffing. *Operations Research*, 32(2):229–249, 1984.

W. Whitt. Fluid models for multiserver queues with abandonments. *Operations research*, 54(1):37–54, 2006.