OR Spectrum (2004) 26: 353–376
DOI: 10.1007/s00291-004-0165-7

OR Spectrum

The impact of retrials on call center performance*

Salah Aguir[1], Fikri Karaesmen[2], O. Zeynep Akşin[3], and Fabrice Chauvet[4]

[1] Laboratoire Génie Industriel, Ecole Centrale Paris, Grande Voie des Vignes, 92295 Chatenay-Malabry Cedex, France (e-mail: salah.aguir@lgi.ecp.fr)
[2] Department of Industrial Engineering, Koç University, 34450 Sariyer, Istanbul, Turkey (e-mail: fkaraesmen@ku.edu.tr)
[3] Graduate School of Business, Koç University, 34450 Sariyer, Istanbul, Turkey (e-mail: zaksin@ku.edu.tr)
[4] Bouygues Telecom R&D, 10 rue Paul Dautier, 78944 Velizy Cedex, France (e-mail: fchauvet@bouyguestelecom.fr)

Abstract. This paper models a call center as a Markovian queue with multiple servers, where customer balking, impatience, and retrials are modeled explicitly. The resulting queue is analyzed both in a stationary and non-stationary setting. For the stationary setting a fluid approximation is proposed, which overcomes the computational burden of the continuous time markov chain analysis, and which is shown to provide an accurate representation of the system for large call centers with high system load. An insensitivity property of the retrial rate to key system parameters is established. The fluid approximation is shown to work equally well for the non-stationary setting with time varying arrival rates. Using the fluid approximation, the paper explores the retrial phenomenon for a real call center. The model is used to estimate the real arrival rates based on demand data where retrials cannot be distinguished from first time calls. This is a common problem encountered in call centers. Through numerical examples, it is shown that disregarding the retrial phenomenon in call centers can lead to huge distortions in subsequent forecasting and staffing analysis.

Keywords: Call centers – Multi-server queues – Retrials

1 Introduction

This paper is motivated by the problem of a major European telecommunications service provider's call center. The call center is managed as a cost center, and minimizing the staffing costs associated with this call center is an important business

* The authors would like to acknowledge helpful discussions with Yves Dallery (Ecole Centrale Paris), Rabie Nait-Abdallah and Thierry Prat (Bouygues Telecom) during the course of this project.
Correspondence to: F. Karaesmen

concern. At the same time, the mobile telephony market within which this call center operates is a highly competitive one, and providing good service to customers constitutes a competitive necessity. Good service is measured in many different ways, but fundamentally it implies meeting the needs of the customers in the best possible way. For a call center, this involves at a very basic level matching the capacity of the center to uncertain demand.

The operations of the call center in question can be described as follows. There are a certain number of servers that answer customer calls. When a customer call arrives, it will be served immediately if a server is available. If all servers are busy with other calls, the customer will be put on hold, and will be asked to wait until a server becomes available. The call center may choose to announce an expected waiting time to the customer at this point. Some customers are patient enough to wait for a server to become available, while others will hang-up or abandon after waiting for some time or immediately upon hearing the waiting time announcement. Management would like to limit the time customers wait for service, and as a result whenever the number of customers waiting to be served exceeds a threshold value, the call will automatically be disconnected and the customer will be asked to call back later. A portion of these customers will redial and try to access the call center. Customers do not like waiting, being disconnected, or attempting a call several times, so from a customer service standpoint management tries to determine the number of servers and the disconnection or blocking threshold such that costs are minimized while certain service levels are satisfied. The use of queueing models as the basis for this type of analysis is common in call centers. However before any optimization can be performed using such queueing models, basic data pertaining to call arrivals and their characteristics are necessary to model the performance of the system.

Call centers track detailed information pertaining to calls. The total number of calls attempted, blocked, or abandoned can all be viewed in a typical database for thirty minute intervals over several years. This data, along with data pertaining to call lengths, abandonment time lengths, waiting time lengths can be used to estimate arrival processes, abandonment behavior, and service processes. Only one type of data is not available, which also constitutes the main motivation of this paper. Looking at the historical call data, the company is unable to tell what proportion of calls during a given thirty minute interval are first attempts or primary calls and what proportion consists of customers who are redialing or retrying to receive service, having abandoned or been blocked in an earlier attempt. This implies that when historical data is used to estimate arrival rates, and staffing levels are determined by embedding the resulting queuing model in an optimization problem, the resulting numbers may be distorted due to historical errors in staffing levels. If in a given period the staffing was done such that realized demand heavily exceeded capacity, this would generate a lot of retrials in subsequent thirty minute intervals. However, not being able to distinguish between first time attempts and retrials, the forecasting method would treat this as an increase in arriving calls, rather than an artifact of the bad staffing decision in an earlier period, potentially leading to future bad dimensioning decisions. Given the importance of staffing costs, the call center would like to determine to what extent this type of interaction between staffing and

forecasting is taking place, and whether one can extract the first time attempts from the total number of calls using the historical data and a queueing model. These two basic questions have motivated the ensuing analysis.

We model such a system as an M/M/C+M queue with retrials and balking, where the $+M$ denotes exponential abandonments. The following section provides a review of the literature and positions our model with respect to models that have appeared earlier. Section 3 formulates the model. In Section 4, we consider systems where parameters like the arrival rates are not time-dependent and the system can be approximated by a stationary analysis in a single time period. We analyze the system using both a continuous time Markov chain analysis and a fluid approximation. The latter approach is shown to overcome the computational difficulties associated with the Markov Chain analysis. It is illustrated how one can estimate the proportion of calls that are first attempts, using this model and the rate of total observed call arrivals. Subsequently, in Section 5 the assumption about a stationary single period is relaxed. In the multiple period setting, the system is analyzed using the fluid approximation. We illustrate that this approximation performs well in terms of representing the performance of such a system by comparing it to simulation. The interaction between staffing and retrials are explored through the help of numerical examples that are partially based on real data. The paper ends with concluding remarks in Section 6.

2 Literature review

There is a rapidly growing literature on queueing models that address call center design and management problems. An excellent survey of this literature can be found in Gans, Koole, and Mandelbaum [12]. Herein, we only describe some work that is of immediate relevance to the model being considered.

Queueing models for call centers differ in the types of customer behavior that they model. As described in the previous section, a customer that is not served can balk, i.e. leave the system immediately, can abandon or renege, i.e. leave the system after waiting for some time, and in both cases can decide to call back in order to access service. Motivated by call centers Baccelli and Hebuterne [6] and Brandt and Brandt [9] treat the case with general impatience times, and characterize performance of such systems. General impatience times are analyzed in the context of telecommunication systems in Boxma and de Waal [7]. Focusing on exponential abandonment times Akşin and Harker [2] and Garnett et al. [13] treat impatience within specific call center applications.

There is an extensive literature on so called *retrial queues* (Yang and Templeton [22]; Falin [10]; Falin and Templeton [11]). Most of the models in this literature do not consider abandonment behavior. Hoffman and Harris [16] incorporate abandonments and retrials in a model which is also motivated by the problem of estimating real arrivals as in our case. We attempt to make this estimation more precise herein. Similarly, Mandelbaum et al. [19] consider multi-server systems with abandonments and retrials and propose a fluid approximation for their analysis. Their model differs from the one herein in that balking behavior is not modeled and the queues are of infinite capacity. Through an insensitivity property that we

are able to show, we illustrate that the approximation we have developed for infinite capacity systems also works well when there is no balking and queue capacity is finite. De Véricourt and Zhou [20] consider retrials that are generated by quality problems associated with call content. We restrict ourself to retrials that are due to a mismatch in demand and capacity.

Artalejo [5] considers a multi-server system with balking and retrials. Customers that find all servers busy are assumed to balk or quit the system with a probability that depends on the number of customers waiting in the queue. This model does not consider abandonment behavior. The systems modeled by Whitt [21] combine balking and abandonment behavior. The objective of the paper is to compare the performance of two systems: one in which state information is communicated to the customers, and another where no information is provided. In the system with no information, a proportion of the customers balk while the remaining may abandon later. In the system where state information is communicated, customers balk with a higher probability such that all reneging is replaced by balking. Our model resembles the case that provides information to customers upon arrival, however in our system customers can balk, renege, and try to call back later. We model retrials explicitly. Hui and Tse [17] have performed an experimental analysis, in order to analyze the growing practice of announcing some form of waiting time or state related information to customers in call centers or other service systems. They show that the appropriate information depends on waiting times, where one would like to communicate the position in the queue for long waiting times, the expected duration of the wait for medium waiting times, and no information for very short waiting times. In our analysis, we assume that expected waiting times are announced, which is a practice being considered by the motivating call center. Armony and Maglaras [4] consider the impact of expected waiting time announcement on the option of postponed or call-back service, rather than a retrial by the customer.

With the exception of Mandelbaum et al. [19], all of the models described so far assume systems with stationary parameters, or systems that can be analyzed as a single period stationary system. Given the inherent transient nature of the interaction between staffing and retrials that we would like to study, we also consider systems with nonstationary arrivals as studied by Green et al. [15] and Green and Kolesar [14]. Rather than using a pointwise stationary approximation, we propose a fluid approximation and illustrate under what conditions this constitutes a precise estimate of real performance. The use of fluid approximations for the analysis of call centers is not new. In addition to Mandelbaum et al. [19], Altman et al. [3] and Jimenez and Koole [18] have analyzed fluid approximations and their application to call center problems.

In this paper, we extend the analysis in Aguir et al. [1] to a multi-period setting and propose a fluid approximation to evaluate system performance both in the single period and multi-period settings. To summarize, we combine balking, abandonment, and retrial behavior of customers in call centers, thereby bringing together different features of earlier studied models. Our aim is to explore the issue of estimating primary call demand from historical data, as in Hoffman and Harris

[16], and to explore the impact of the retrial phenomenon on call center performance as being done for the case of impatience in Garnett et al. [13].

3 The model

Let us consider the following model of a call center that has C Customer Service Representatives (CSR)'s. Service times (including the talk time and the after talk wrap-up time) are assumed to be exponentially distributed with rate μ. First-attempt calls arrive to the system according to a Poisson process with rate λ. These calls will be alternately labeled as *primary calls*. Customers who can access a free CSR at the time of arrival are immediately served and depart the system. Customers who cannot access a free CSR at the time of arrival may balk immediately with probability β. These balking customers reattempt their call with probability p after an exponentially distributed amount of time with rate δ. In practice p may not be a constant probability for subsequent reattempts, however for tractability purposes this will be assumed herein. Customers who join the queue (with probability $1-\beta$) abandon if they cannot access a CSR within a delay that is exponentially distributed with rate θ_1. Abandoning customers are assumed to have identical retrial behaviour as balking customers; they reattempt their call with probability p and after an exponentially distributed amount of time with rate δ. It should be noted that data from the motivating call center supports the assumption that the retrial probability for both balking and abandoning customers is approximately equal. In accordance with the "retrial queueing" literature, we refer to the pool of customers that are waiting to repeat their call as the *orbit*. Note that, under the above assumptions the time that a customer spends in the orbit is exponentially distributed with rate δ. A summary of the notation can be found below:

λ: Arrival rate of first-attempt (primary) calls
λ_o: Total call arrival rate (observed call rate)
C: Number of CSRs
μ: Service (talk and wrap-up) rate
β: Instantaneous balking probability for customers who cannot access a free CSR at the time of arrival
θ_1: Abandonment rate of customers who join the queue
p: Retrial probability for balking or abandoning customers
δ: Retrial rate for balking or abandoning customers

In reality all of these parameters may be time dependent. We address the fluctuations in the arrival rate and the number of CSRs in the second part of this paper. The fluctuations in the other parameters are milder and are not modeled here.

Finally, it is important to note that the total call arrivals to the center are constituted of two separate flows: primary (first-attempt) calls (with rate λ) and repeated calls (retrials). We denote by λ_o (where $\lambda_o \geq \lambda$), the total call arrival rate (also referred to as *observed* call rate).

Let us now adapt this model to the more general case where customers are informed about their anticipated waiting times at the time of arrival. Whitt [21] presents a thorough explanation of the model and the underlying assumptions.

In this case, customers are announced their anticipated waiting time before accessing a CSR. The precise information announced to the customer may be an estimate of the expected waiting time or a related measure. In any case, the announced information is based on the actual number of customers waiting in the queue in front of the arriving customer. As explained by Whitt [21], the customer balking behaviour is then a function of the number of customers waiting in the queue. Let $r(k)$ $(k \geq C)$ be the probability that an arriving customer balks when there are $k - C$ waiting customers in the queue. This represents the probability that the announced waiting time exceeds customer expectations. This probability can be expressed as:

$$r(k) = \beta + (1 - \beta)P(T < S_k) \tag{1}$$

In this equation T is a random variable that represents the patience threshold of the customer and S_k is a random variable that represents the time between the arrival of a customer and the time this customer accesses a CRS. Since service times are exponential, S_k has an Erlang distribution with $k - C + 1$ stages each with rate $C\mu$. As in Whitt [21], it is helpful to approximate $P(T < S_k)$ in equation (1) by $P(T < E[S_k])$ where:

$$E[S_k] = \frac{k - C + 1}{C\mu}$$

If we now assume that the patience threshold T is exponentially distributed with rate θ_1, we reach:

$$r(k) = 1 - (1 - \beta)e^{-\theta_1 \frac{k-C+1}{C\mu}} \tag{2}$$

The above analysis assumes that, a customer deciding to join the queue does not abandon thereafter. We may assume that customers joining the queue may still abandon with rate θ where the new abandonment rate is less than the initial (uninformed) abandonment rate θ_1. The above analysis is then approximately valid, if θ is small enough not to modify S_k significantly. Finally, note that a system where waiting times are not announced is just a special case of the more general model, with $r(k) = \beta$.

4 Stationary analysis

4.1 The stochastic model

Under the previously stated assumptions, a Markov chain can be employed to model the call center with repeated calls. Let the state of the system be $(m, n), (m,\ n \geq 0)$, where m represents the number of customers in the real system (those in service plus those who are in the queue) and n represents the number of customers in the orbit who repeat their call with (exponential) rate δ. Unless the real queue has finite buffers, both m and n are unbounded. For numerical solution purposes, we approximate the unbounded system by a truncated system where m is truncated at K_1 and n is truncated at K_2.

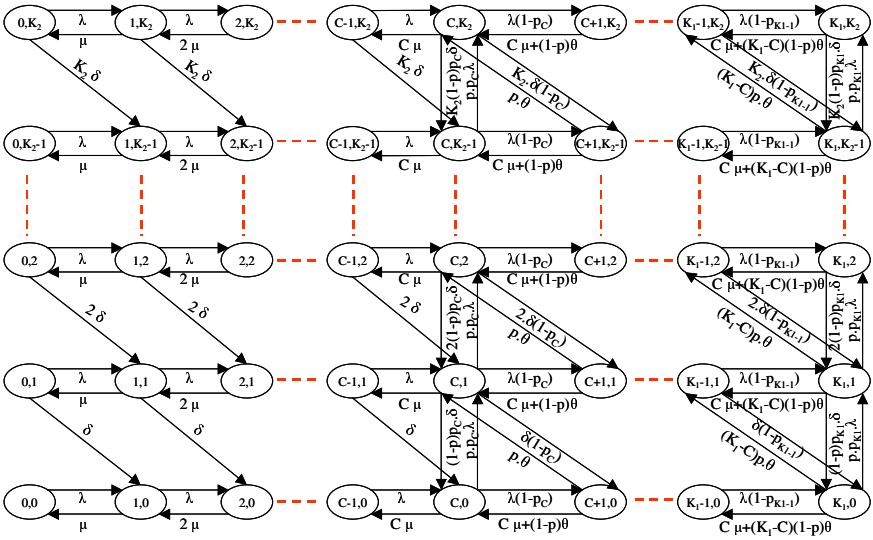

Fig. 1. The transition diagram of the Markov chain

Let us now describe the transition rates of the Markov chain. Let $Q_{(m,n)(m',n')}$ be the transition rate from state (m, n) to state (m', n'), m, m' = 0, 1, ..., K_1, n, n' = 0, 1, ..., K_2. The non-zero transition rates are as follows:

$$Q_{(m,n)(m+1,n)} = \begin{cases} \lambda \text{ for } 0 \le m < C, \ 0 \le n \le K_2 \\ \lambda(1 - r(m)) \text{ for } C \le m < K_1, \ 0 \le n \le K_2 \end{cases}$$

$$Q_{(m,n)(m-1,n)} = \begin{cases} m\mu \text{ for } 0 < m \le C, \ 0 \le n \le K_2 \\ C\mu + (m - C)(1 - p)\theta \text{ for } C < m \le K_1, \ 0 \le n \le K_2 \end{cases}$$

$$Q_{(m,n)(m+1,n-1)} = \begin{cases} n\delta \text{ for } 0 \le m < C, \ 0 < n \le K_2 \\ n\delta(1 - r(m)) \text{ for } C \le m < K_1, \ 0 < n \le K_2 \end{cases}$$

$$Q_{(m,n)(m-1,n+1)} = (m - C)p\theta \text{ for } C < m \le K_1, \ 0 \le n < K_2$$

$$Q_{(m,n)(m,n+1)} = p\, r(m)\lambda \text{ for } C \le m \le K_1, \ 0 \le n < K_2$$

$$Q_{(m,n)(m,n-1)} = n(1 - p)\, r(m)\delta \text{ for } C \le m \le K_1, \ 0 < n \le K_2$$

Figure 1 depicts the state transition diagram of the above Markov chain.

There does not seem to be a special structure and hence no easy analytical solution for the above Markov chain. Standard numerical methods are employed to obtain a numerical solution for the stationary distribution by truncating the state space. Because our objective is to find an accurate approximation for the infinite state-space problem, in the implementation we experiment with increasing truncation limits K_1 and K_2 until further increases do not affect the stationary distribution. Once this is done, the performance measures of interest follow. Let $\pi_{m,n}$, m = 0, , ..., K_1, n = 0, 1, ..., K_2, be the stationary probability of being in state

(m, n) and let T_r be the *stationary retrial rate* (number of repeated calls per unit time). The retrial rate is given by:

$$T_r = \sum_{n=1}^{K_2} n\delta \sum_{m=0}^{K_1} \pi_{m,n} \tag{3}$$

The retrial rate can hence be obtained using the numerical solution of the Markov chain and expression (3) but this computation is numerically intensive and rather time consuming. The next subsection proposes an approximate procedure for this computation.

Below, we propose an alternative expression for T_r based on the stationary flow balance of the system (similar to Hoffman and Harris [16]). The result is summarized in the following proposition:

Proposition 1 *Let $E[B]$ be the average number of busy servers. The retrial rate, T_r of the system can be expressed as:*

$$T_r = \frac{p}{1-p}\left(\lambda - E[B]\ \mu\right) \tag{4}$$

Proof. Let us consider the stationary flows in the system. The total incoming flow, λ_o is composed of two streams, the primary arrivals and the retrials. We then have $\lambda_o = \lambda + T_r$. The outgoing flow is composed of three streams, let us denote the outgoing flow due to abandonments by A, and the one due to balking by R. The outgoing flow due to service completions is given by the average effective service rate $E[B]\ \mu$.

In the stationary equilibrium, the outgoing flow must be equal to the incoming flow which implies:

$$\lambda_o = A + R + E[B]\ \mu \tag{5}$$

In addition the average flow into and out of the orbit must also be equal. A call that joins the orbit is either due to an abandonment or due to balking. The inflow to the orbit is then: $p(A + R)$. Since the departure rate from the orbit is T_r, we can write:

$$T_r = p(A + R). \tag{6}$$

Using the flow balance into and out of the orbit (equations (5) and (6):

$$\lambda_o = \frac{T_r}{p} + E[B]\mu \tag{7}$$

Since $\lambda_o = \lambda + T_r$, equation (7) enables us to write:

$$\lambda + T_r = \frac{T_r}{p} + E[B]\ \mu$$

which yields the result. □

Figure 2 represents the outgoing and incoming flows used in the above proof. Although proposition 1 does not facilitate the exact computation of the retrial rate ($E[B]$ still has to be computed), it suggests a way of estimating it from real data. If p, λ, μ and $E[B]$ are known, Proposition 1 provides a simple estimator for T_r. It may also be employed as the basis of approximate techniques such as the ones in Hoffman and Harris [16]. We postpone, however, the discussion of an approximation based on this proposition to the next subsection where a different approach is presented.

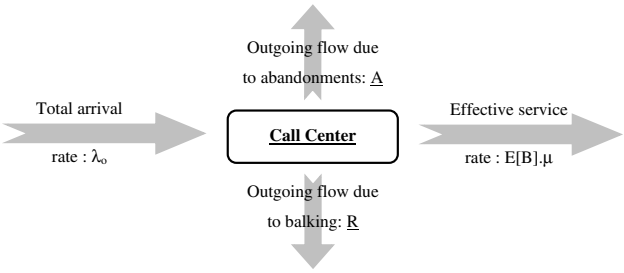

Fig. 2. The incoming and outgoing flows

4.2 The fluid approximation

Although the stochastic model of the preceding subsection can be used to numerically calculate the performance measures related to retrials, it is computationally burdensome. Below, we introduce a simple approximation that replaces the transition rates of the Markov chain by deterministic flow rates. Because the resulting model is a deterministic continuous flow model, the approximation is referred to as the *fluid approximation*. Mandelbaum et al. [19] obtain such an approximation as a formal limit of a related stochastic model and show through numerical examples that the approximation is accurate.

In order to describe the approximation, let us begin by replacing the discrete state space of the original stochastic model by a continuous state space. In the approximate model, the state is represented by $(x_1(t),x_2(t))$, where $x_1(t)$ is the (continuous) level of the real buffer and $x_2(t)$ the (continuous) level of the orbit at time t. Under this continuous deterministic approximation, the total arrival rate to the system is given by:

$$\lambda_o(t) = \delta x_2(t) + \lambda(t) \tag{8}$$

The incoming flow to $x_1(t)$ depends on $\lambda_o(t)$ and the balking rate at time t. Using the deterministic approximation, the rate of increase in $x_1(t)$ is then : $(1 - r\,(x_1(t)))\,\lambda_o(t)$, where $r(x)$ represents the balking probability when the real buffer's level is x. To define this probability, we extend the previous definition (equation 2) to a continuous state space as follows:

$$r(x) = \begin{cases} 1 - (1-\beta)e^{-\theta_1 \frac{x-C+1}{C\mu}} & \text{if } x \geq C \\ 0 \quad \text{otherwise} \end{cases} \tag{9}$$

The buffer level $x_1(t)$ decreases through service completions and abandonments. With the approximation, the rate of decrease in $x_1(t)$ due to service completions and abandonments is equal to :

$$\mu\ Min\,(x_1(t), C) + \theta\ Max\,(x_1(t) - C, 0)\,.$$

We can now express the total rate of change of $x_1(t)$ as follows :

$$\frac{dx_1}{dt} = (1 - r\,(x_1(t)))\,\lambda_o(t) - \mu Min\,(x_1(t), C) - \theta Max\,(x_1(t) - C, 0) \tag{10}$$

An identical reasoning leads to following differential equation for the rate of change in x_2, the level of the orbit:

$$\frac{dx_2}{dt} = p\left(r\left(x_1(t)\right)\lambda_o(t) + \theta\ Max\left(x_1(t) - C, 0\right)\right) - \delta x_2(t) \tag{11}$$

Replacing $\lambda_o(t)$ by the right-hand side of equation (8), we can obtain $(x_1(t),x_2(t))$ as the solution of the following differential system:

$$\begin{cases} \dfrac{dx_1}{dt} = \left(1 - r\left(x_1(t)\right)\right)\left(\delta x_2(t) + \lambda(t)\right) - \mu\ Min\left(x_1(t), C\right) \\ \qquad\qquad -\theta\ Max\left(x_1(t) - C, 0\right) \\ \dfrac{dx_2}{dt} = pr\left(x_1(t)\right)\left(\delta x_2(t) + \lambda(t)\right) - \delta x_2(t) + \theta p\ Max\left(x_1(t) - C, 0\right) \end{cases} \tag{12}$$

In particular, in the stationary regime we have:

$$\lim_{t\to\infty} \frac{dx_1(t)}{dt} = \lim_{t\to\infty} \frac{dx_2(t)}{dt} = 0 \tag{13}$$

which enables us to obtain the stationary buffer levels x_1 and x_2 as follows:

$$\begin{cases} \lim\limits_{t\to\infty} x_1(t) = x_1 \\ \lim\limits_{t\to\infty} x_2(t) = x_2 \end{cases}$$

It is easy to see that, if the average load $\rho\ (= \lambda/C\mu)$ is less than or equal to 1, the stationary level of the orbit, x_2, is zero. In this case, the approximation can provide no information about the retrial rate of the system. Let us focus on the more interesting case where $\rho > 1$. In this case, it can be verified that $x_1 \geq C$. This enables us to write: Min $(x_1, C) = C$ and $Max\left(x_1 - C, 0\right) = x_1 - C$. In the stationary regime, the system (12) then becomes:

$$\begin{cases} \left(1 - r(x_1)\right)\left(\delta x_2 + \lambda\right) = C\mu + \theta\left(x_1 - C\right) \\ \delta x_2 = pr(x_1)\left(\delta x_2 + \lambda\right) + \theta p\left(x_1 - C\right) \end{cases}$$

$$\Leftrightarrow \begin{cases} p\left(\delta x_2 + \lambda\right) = pC\mu + \theta p\left(x_1 - C\right) + pr(x_1)\left(\delta x_2 + \lambda\right) \\ \delta x_2 = pr(x_1)\left(\delta x_2 + \lambda\right) + \theta p\left(x_1 - C\right) \end{cases}$$

which finally leads to:

$$\begin{cases} p\left(\delta x_2 + \lambda\right) - \delta x_2 = pC\mu \\ \delta x_2 = pr(x_1)\left(\delta x_2 + \lambda\right) + \theta p\left(x_1 - C\right) \end{cases} \tag{14}$$

The above expression enables us to obtain the stationary buffer levels. The stationary level of the real buffer is the solution of:

$$\left(\lambda - p\ C\mu\right) r(x_1) + \theta(1 - p)(x_1 - C) = \lambda - C\mu \tag{15}$$

Equation (15) can be solved using standard numerical methods to determine the value of x_1. On the other hand, it turns out that there is a simpler solution for x_2. Interestingly, x_2 does not depend on θ, the customer patience threshold rate. This is a useful property since the other parameters (C, μ, λ, δ and p) are easier to estimate

in practice. Another useful consequence of the above property (insensitivity to θ) is that the same insensitivity is also true for the stationary retrial rate. Moreover, this finding is also true for the stochastic model of the previous subsection. This result is presented in the following proposition:

Proposition 2 *i. The retrial rate, T_r(fluid), obtained through the fluid model does not depend on θ*
ii. The fluid approximation for the retrial rate is asymptotically correct for the stochastic model (i.e. there exists λ such that T_r(stochastic)$-T_r$(fluid) $< \epsilon$, ($\epsilon > 0$)) as λ increases.

Proof. From equations (14), the stationary level of the orbit, x_2 is given by:

$$x_2 = \frac{p}{1-p}\frac{\lambda - C\mu}{\delta}. \tag{16}$$

Now, noting that $T_r(\text{fluid}) = \delta x_2$, we obtain:

$$T_r(\text{fluid}) = \frac{p}{1-p}(\lambda - C\mu) \tag{17}$$

which does not depend on θ.

In order to prove part ii., consider the expression for T_r(stochastic) given in Proposition 1 (equation 4) for the stochastic model. We can express:

$$T_r(\text{stochastic}) - T_r(\text{fluid}) = \frac{\mu p}{1-p}(C - E[B])$$

A direct sample path comparison indicates that $E[B]$ is increasing in λ. This implies that for any desired difference ϵ, there exists a λ which makes $(C - E[B]) < ((\mu p)/(1-p))\epsilon$. □

Proposition 2 suggests the following: the approximation for the retrial rate given by equation (17) must be accurate for overloaded call centers. In fact, the error of the approximation is due to replacing C by $E[B]$ in the corresponding formula for the stochastic model. For overloaded systems, $E[B]$ is reasonably close to C. As mentioned before, the approximation does not provide any useful information when $\lambda \leq C\mu$. Nevertheless for large call centers, the retrial phenomenon is essentially due to overload. Indeed in a large call center that is not overloaded it is well known (see for example Borst et al. [8]) that both blocking and waiting will occur at a minimum level, resulting in a negligible potential for retrials.

In the next subsection, we assess the accuracy of this approximation under various conditions. The convergence rate of the fluid approximation is an important issue, which is explored through the numerical examples in the following subsection.

4.3 Numerical assessment for the fluid approximation

In order to assess the performance of the fluid approximation, we show via numerical examples that the error between the retrial rate $\delta \cdot x_2$ (given by equation (17)) and

the exact retrial rate(given by equation (3)) diminishes rapidly as a function of the number of servers and system load. In other words, the retrial rate obtained through the stochastic model (exact) will be compared to that obtained by the fluid model (approximation). The exact retrial rates were obtained through a numerical solution of the corresponding Markov Chain.

4.3.1 A call center with balking

Consider first the case where the system load ρ is set to 133% and the number of servers C is varied. Since $\rho = \lambda/C\mu$, λ is also varied proportionally in order to ensure a constant system load. The results of the comparison between the

Table 1. The comparison of the approximation with respect to the actual retrial rate for $\rho = 1.33$

C	T_r	δx_2 (fluid)	% Error
5	0.69	0.50	27.86
10	1.20	1.00	16.96
15	1.70	1.50	11.71
20	2.19	2.00	8.64
25	2.68	2.50	6.64
30	3.17	3.00	5.26
35	3.66	3.50	4.26
40	4.15	4.00	3.52
45	4.64	4.50	2.94
50	5.13	5.00	2.49

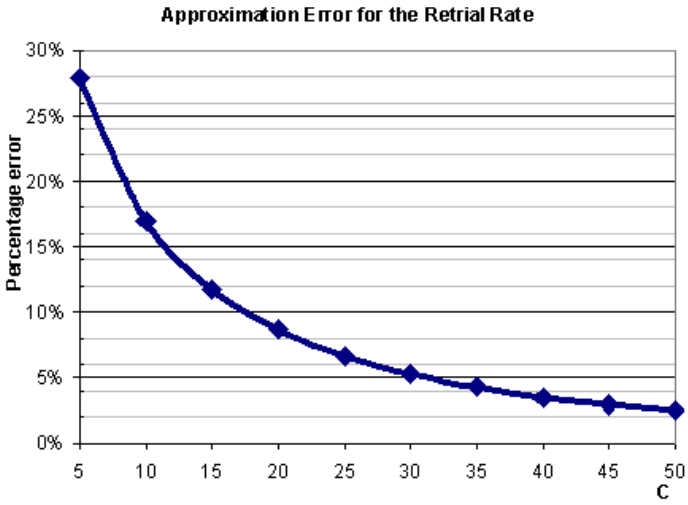

Fig. 3. Percentage error of the fluid approximation for the retrial rate as a function of the number of servers C for ρ = 133 %

two models is given in Table 1. In this table and in the remaining parts of this paper a unit of time is taken to be one minute. As such, $\mu = 0.3$ implies a service rate of 0.3 calls per minute. Using the results in Table 1 we can calculate the error made by the fluid approximation as a relative percentage of the exact (stochastic) retrial rate Tr (i.e. percentage error=(| T_r(fluid)-T_r(stochastic)|/T_r(stochastic))$\times$ 100). This error is shown in Figure 3 as a function of the number of servers.

We observe that for a system load fixed at 133 % , the precision of the fluid approximation increases as a function of the number of servers. For fifty servers the error is already less than 2.5 % . For small number of servers, the error rate

Table 2. The comparison of the approximation with respect to the actual retrial rate for $C = 40$

ρ	T_r	δx_2 (fluid)	% Error
100 %	0.98	0.00	100.00
110 %	1.76	1.20	31.96
120 %	2.72	2.40	11.69
130 %	3.78	3.60	4.69
140 %	4.90	4.80	2.01
150 %	6.05	6.00	0.91
160 %	7.23	7.20	0.42
170 %	8.42	8.40	0.20
180 %	9.61	9.60	0.10
190 %	10.80	10.80	0.05
200 %	12.00	12.00	0.02

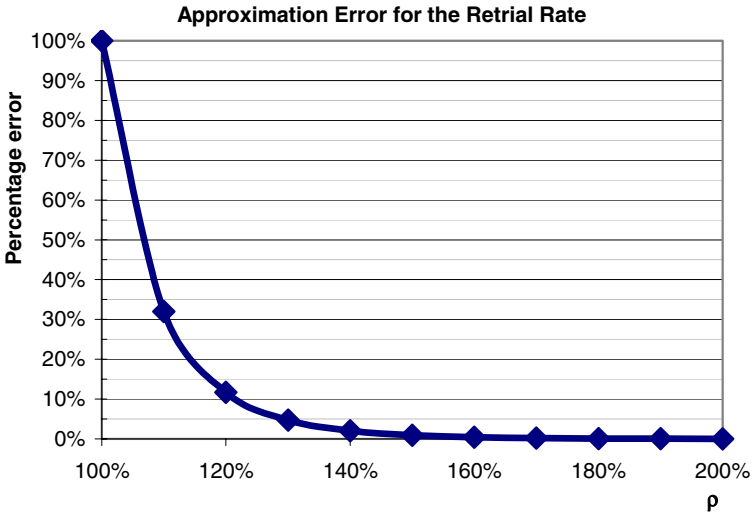

Fig. 4. Percentage error of the fluid approximation for the retrial rate as a function of the number of the system load ρ for $C = 40$

can be higher than 25 % . Given these results, one would clearly prefer the fluid approximation for large centers as the one being analyzed in the example, given the simplicity of the analytical evaluation that needs to be performed.

Next, we look at an example where the number of servers C is fixed at 40, and we vary the system load by increasing the primary arrival rate λ. Table 2 compares the retrial rates obtained by the stochastic and fluid models. As before, Figure 4 shows the evolution of the percentage error as a function of the system load. The figure confirms that the precision of the fluid approximation increases rapidly as a function of the system load. This is not surprising by Proposition 2 which states that the fluid approximation asymptotically leads to an exact result. Note that the number of servers in this example represents a small call center in reality. A further observation to be made is in the case of $\rho = 100\%$. For this case, the error of the fluid approximation is seen to be 100%. This is not surprising, given the fact that the fluid approximation completely ignores the stochasticity in the system and predicts zero retrials. The consolation is that in this particular case the retrial rate of the stochastic system is very small (0.98 per minute) as observed in Table 2.

The two examples illustrate that the performance of the fluid approximation improves with higher system load and higher number of servers. In a setting where both effects are combined, the convergence to high levels of precision occurs even faster. As a result, we conclude that for medium to large sized call centers with a heavy system load, the fluid approximation constitutes an effective means to evaluate the retrial rate. Similar to some recent results in Jimenez and Koole [18], it may be possible to show this property of scale economies formally using fluid limits. This is identified as a direction for future research.

4.3.2 A call center without balking

Proposition 2 states another interesting feature of the fluid approximation. Accordingly, the approximate retrial rate obtained by the approximation does not depend on the balking function $r(k)$. The latter can take another form than that in (2). Using this property, we will next analyze the performance of the approximation for another system, given by a balking function $r_1(k)$ of the form:

$$r_1(k) = \begin{cases} 0 \text{ if } k < K_1 \\ 1 \text{ otherwise} \end{cases},$$

with K_1 integer and strictly larger than C. In practice many call centers restrict the length of the queue of waiting customers to a finite length, in order to avoid long delays in the queue, thereby preferring to block a customer rather than seeing them abandon. The artificial balking probability $r_1(k)$ allows us to indirectly model an $M/M/C/K_1 + M$ call center, that models the impatience of customers in a finite queue of length K_1, where abandoning calls will retry with a probability p after a time that is exponentially distributed with rate δ. This new system does not allow for balking. Furthermore, it constitutes a special case of the generic model, and as a result we expect the fluid approximation to work well for this case as well.

In order to verify this claim, we compare the fluid approximation to the stochastic model for this new system. We consider two examples, one in which the waiting

Table 3. The comparison of the approximation with respect to the actual retrial rate for $\rho = 1.33$ and for varying buffer sizes

	$K_1 = C + 5$			$K_1 = C + 10$		
C	T_r	δx_2 (fluid)	% Error	T_r	δx_2 (fluid)	% Error
5	0.59	0.50	15.43	0.58	0.50	14.36
10	1.09	1.00	8.22	1.07	1.00	6.34
15	1.59	1.50	5.42	1.55	1.50	3.49
20	2.08	2.00	3.96	2.04	2.00	2.19
25	2.58	2.50	3.07	2.54	2.50	1.50
30	3.08	3.00	2.48	3.03	3.00	1.09
35	3.57	3.50	2.06	3.53	3.50	0.83
40	4.07	4.00	1.75	4.03	4.00	0.65
45	4.57	4.50	1.51	4.52	4.50	0.52
50	5.07	5.00	1.32	5.02	5.00	0.43

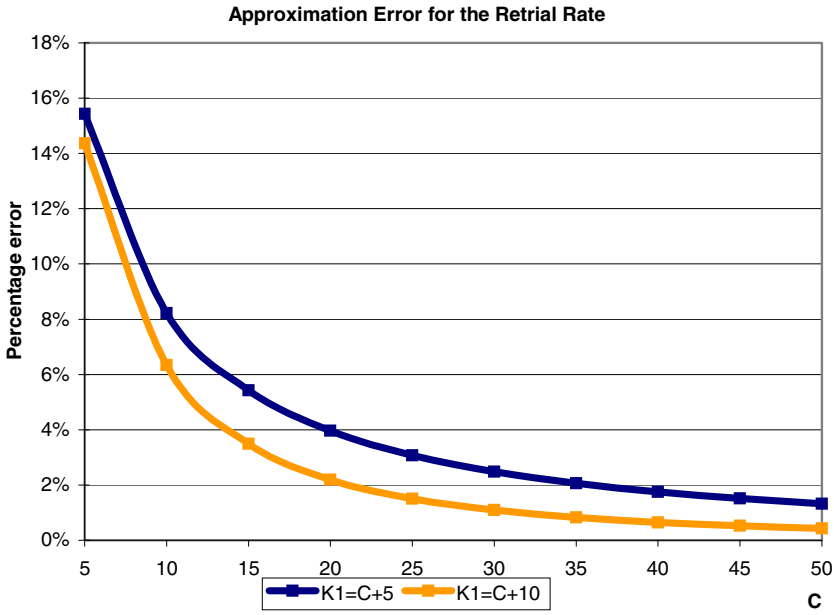

Fig. 5. Percentage error of the fluid approximation for the retrial rate as a function of the number of servers C for ρ = 133 % in a system with finite queue capacity

space is restricted to 5 customers ($K_1 = C + 5$) and another to 10 customers ($K_1 = C + 10$). The parameters and results obtained by the two models are tabulated in Table 3 for a system load of 133%. Figure 5 graphs the error for both systems. We note that even for a system with finite queue length, the precision of the fluid approximation improves rapidly as a function of the system size. In fact the precision is higher compared to the case with infinite waiting room and balking

Table 4. The comparison of the approximation with respect to the actual retrial rate for $C = 40$ and for varying buffer sizes

	$K_1 = C + 5$			$K_1 = C + 10$		
ρ	T_r	δx_2 (fluid)	% Error	T_r	δx_2 (fluid)	% Error
100 %	0.81	0.00	100.00	0.72	0.00	100.00
110 %	1.60	1.20	25.16	1.50	1.20	19.92
120 %	2.59	2.40	7.40	2.51	2.40	4.36
130 %	3.69	3.60	2.46	3.64	3.60	1.03
140 %	4.84	4.80	0.91	4.81	4.80	0.26
150 %	6.02	6.00	0.37	6.00	6.00	0.07
160 %	7.21	7.20	0.16	7.20	7.20	0.02
170 %	8.41	8.40	0.08	8.40	8.40	0.01
180 %	9.60	9.60	0.04	9.60	9.60	0.00
190 %	10.80	10.80	0.02	10.80	10.80	0.00
200 %	12.00	12.00	0.01	12.00	12.00	0.00

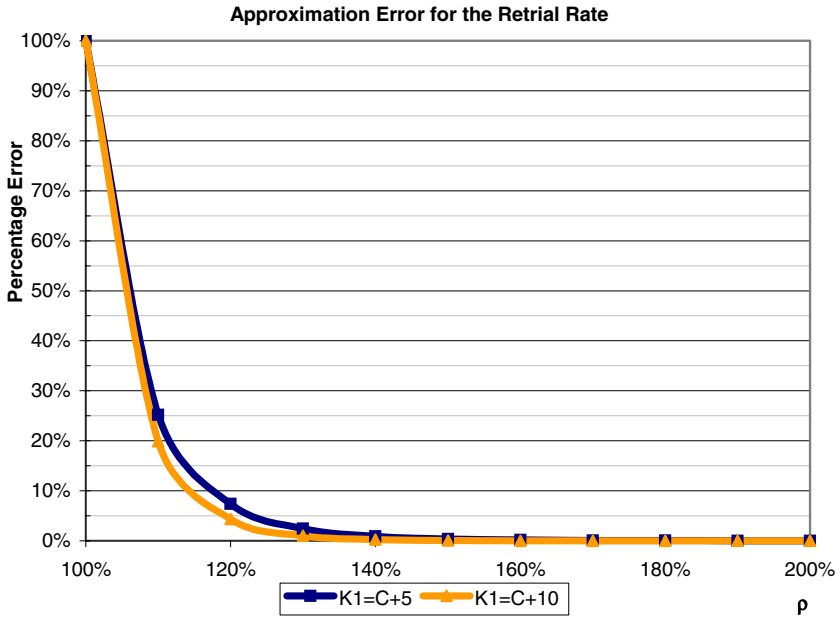

Fig. 6. Percentage error of the fluid approximation for the retrial rate as a function of the number of the system load ρ for $C = 40$ in a system with finite queue capacity

as modeled previously. Similar results are obtained for the case where we fix the number of servers $C = 40$ and vary the system load, as demonstrated by Table 4 and Figure 6.

5 Analysis of the non-stationary system

In most cases, the parameters that we use to model a call center, vary over time. In the case of arrivals, this represents a changing desire or need by customers to call a center during the day, and is modeled as non-stationary arrivals. It is quite common to have beginning or end of the day peaks in call centers that do not operate twenty four hours a day. Similarly, for twenty four hour a day operational call centers arrival rates clearly vary between the day and night. The number of servers will also vary throughout a day, partially in response to changing arrival rates, and partially as a result of workforce scheduling requirements and constraints. Call center staffing problems typically treat a day as consisting of multiple time periods, during which the number of servers remains constant. In this section, we are going to consider such a period to be of length thirty minutes, also representing the planning period for most call centers.

Consider now any such period during the day with duration T. During this period, the parameters of the system $\lambda(t)$, C, μ, θ, p, δ and $r(x)$ remain constant. We assume that the period begins with a real queue of size x_1^0 and an orbit of size x_2^0. The system of differential equations

$$\begin{cases} \dfrac{dx_1}{dt} = (1 - r\,(x_1(t)))\,(\delta x_2(t) + \lambda(t)) - \mu.Min\,(x_1(t), C) \\ \quad -\theta\, Max\,(x_1(t) - C, 0) \\ \dfrac{dx_2}{dt} = p\, r\,(x_1(t))\,(\delta x_2(t) + \lambda(t)) \\ \quad -\delta x_2(t) + \theta\, p\, Max\,(x_1(t) - C, 0) \end{cases} \tag{18}$$

allows us to evaluate the evolution of the queue within the period being considered. In particular, we can compute the size of the real queue x_1^T as well as the size of the orbit x_2^T at the end of the period with duration T.

Since an analytical calculation is not possible, we resort to a numerical analysis. The following algorithm describes how the equations in (18) can be used to evaluate the system for multiple planning periods like the one described above. The basic idea is to link the periods together by setting the appropriate initial levels x_1^0 and x_2^0 for each period.

Algorithm for multi-period analysis

- Step 1: Initialization: $x_1^0 = x_2^0 = 0$. Using the differential equations in (18) determine $x_1(t)$ and $x_2(t)$ for $0 \leq t \leq T$, with the corresponding period's parameter values(C, λ, ...). Let $x_1^T = x_1(T)$ and $x_2^T = x_2(T)$.
- Step 2: Initialization: $x_1^0 = x_1^T$ and $x_2^0 = x_1^T$. Using the differential equations in (18) determine $x_1(t)$ and $x_2(t)$ for $T \leq t \leq 2T$. Let $x_1^T = x_1(2T)$ and $x_2^T = x_2(2T)$.

- Step i, $i \geq 3$: Initialization: $x_1^0 = x_1^T$ and $x_2^0 = x_1^T$. Using the differential equations in (18) determine $x_1(t)$ and $x_2(t)$ for $(i-1)T \leq t \leq (i)T$. Let $x_1^T = x_1(iT)$ and $x_2^T = x_2(iT)$.
- Continue until $i = n$.

5.1 Numerical examples

A numerical example that uses data from the telecommunication service provider's call center is presented next. The retrial rate throughout one day of operations at this call center is determined using the proposed algorithm and the expression $\delta x_2(t)$ for the retrial rate. These calculations were performed using a standard multi-step finite approximation method. This is compared to a discrete-event simulation of the system in question in order to validate the fluid approximation. The simulations were performed using the discrete event simulation software Arena 5.0. For the simulation experiments, we collected data in intervals of 2.5 minutes and performed 10000 replications. Below, we report the average retrial rate for each interval over these 10000 replications for all of the systems considered. We also calculated the standard deviation of the retrial rate over the replications for assessing the accuracy of the estimates. The standard deviations for each interval vary during the day and range from 2-5% (of the average value) for peak periods to 5-12% for periods that generate few retrials. The standard error of the estimator of the mean (standard deviation of the sample/$\sqrt{\text{sample size}}$) is hence extremely small (i.e. under 0.12%). The comparisons are made for three systems, which we can view as the current system (System 1), the current system under higher arrival rates (System 2), and a system where the total number of servers have been distributed equally between the periods in a day (System 3). Common parameters for the example are tabulated in Table 5. Table 6 tabulates the arrival rates and number of servers for each system for a particular day. The total cumulative number of servers for all three systems is 3135.

Figure 7 compares the case of System 1 and 2. For both systems, the simulation results practically coincide with the numerical analysis. The only interval where they are slightly separated is between 15:30 and 16:30. Notice that the lowest utilization ($\rho = \lambda/C\mu$) (between 0.95 and 1.04) for both systems happens during the interval 14:30 and 16:30. Between 14:30 and 15:30 there are still calls in the orbit, so the actual utilization that the system will experience will be higher. As a result, during this interval the approximation still works well. In the subsequent time period, the system will experience the low utilization rates, and in this case

Table 5. The parameters that are constant during the day for the analyzed systems

μ	θ	δ	p	β	θ_1
0.3	0.5	0.1	0.6	0.2	1

Table 6. The parameters that vary during the day for the analyzed systems

Period		System 1		System 2		System 3	
start	end	C	λ	C	λ	C	λ
09:00	09:30	86	68	86	110	175	68
09:30	10:00	114	75	114	110	175	75
10:00	10:30	177	101	177	115	175	101
10:30	11:00	180	87	180	100	174	87
11:00	11:30	197	82	197	94	174	82
11:30	12:00	192	80	192	89	174	80
12:00	12:30	169	73	169	75	174	73
12:30	13:00	155	74	155	78	174	74
13:00	13:30	169	67	169	72	174	67
13:30	14:00	124	74	124	80	174	74
14:00	14:30	140	70	140	80	174	70
14:30	15:00	238	68	238	71	174	68
15:00	15:30	231	72	231	70	174	72
15:30	16:00	235	69	235	67	174	69
16:00	16:30	215	67	215	64	174	67
16:30	17:00	214	69	214	73	174	69
17:00	17:30	163	69	163	74	174	69
17:30	18:00	136	73	136	88	174	73

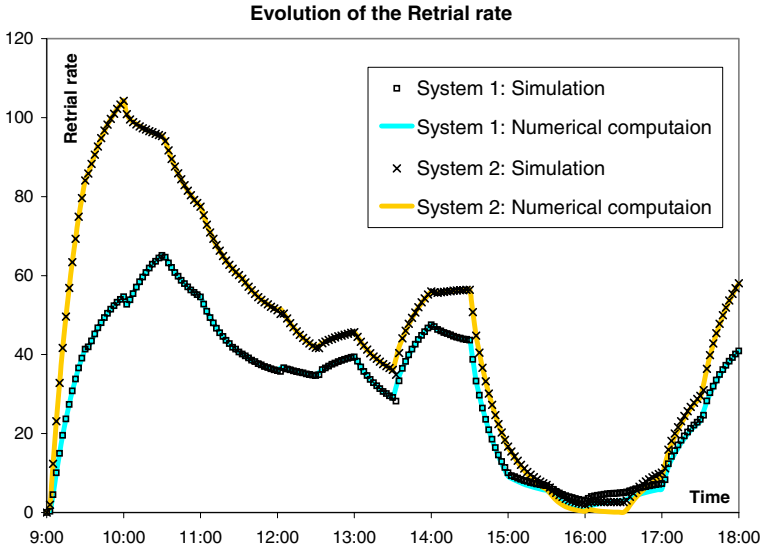

Fig. 7. Comparison of simulation results and numerical analysis results for Systems 1 and 2

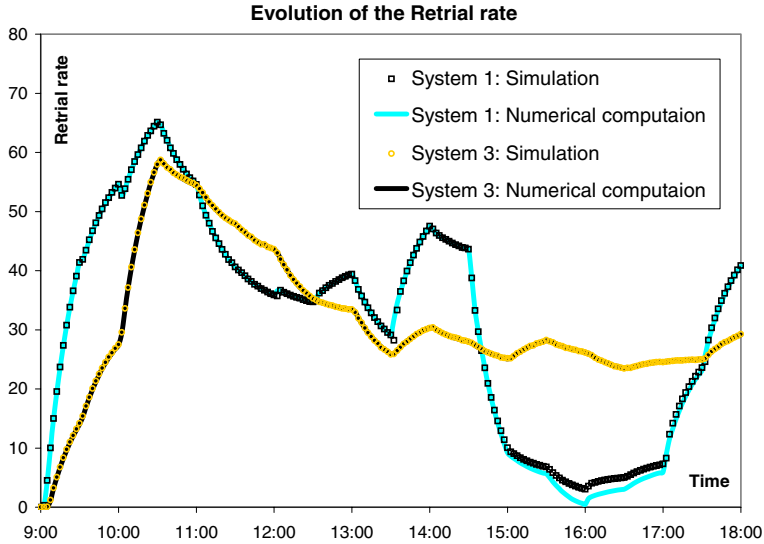

Fig. 8. Comparison of simulation results and numerical analysis results for Systems 1 and 3

the numerical analysis is no longer as precise. We observe that System 2 has higher retrial rates compared to System 1, though the qualitative pattern of the retrial curves for both systems resemble each other.

Figure 8 compares System 1 and 3. Recall that both of these systems have identical arrival rates, and the difference comes from the distribution of the total number of servers during the day. Once again, there is a good fit between the numerical analysis and simulation curves. In fact, since the redistribution of servers results in utilization rates that are higher than one in each period, we observe that the approximation results coincide with the simulation for all periods. This time the shape of the retrial rate curve is visibly different. The example illustrates the importance of staffing decisions on the resulting retrial phenomenon. Indeed, improved staffing can lead to a lower and less variable retrial rate curve.

5.2 Estimating primary calls from observed calls

Once $x_1(t)$ and $x_2(t)$, $0 \leq t \leq T$, have been determined, one can evaluate the observed call rate for the entire horizon using the relation

$$\lambda_o(t) = \lambda + \delta x_2(t). \tag{19}$$

The average rate of observed calls is then given by

$$\bar{\lambda}_o = \frac{1}{T} \int_0^T \lambda_o(t)\, dt$$

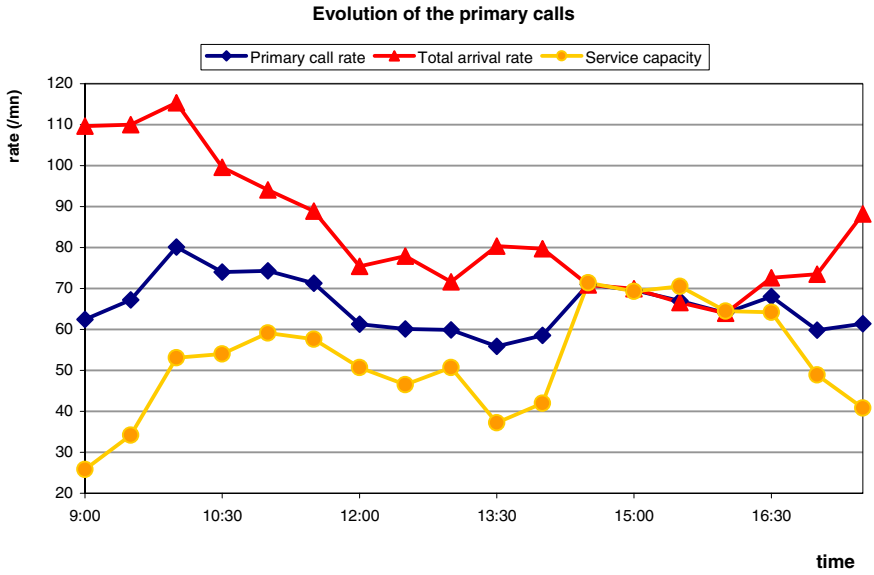

Fig. 9. Evolution of the primary call rate λ as a function of the observed call rate λ_o

and is calculated using the expression

$$\bar{\lambda_o}(\lambda) = \lambda + \frac{\delta}{T} \int_0^T x_2(t)\, dt \tag{20}$$

In many call centers, managers have information about the observed calls λ_o, and not about the primary calls λ. Thus, one would like to obtain an estimate for λ using λ_o. This can be done by a numerical inversion of Equation (20), using the differential equations in (18). In order to find the rates for the entire day, it is sufficient to link the periods together, using the end values of a period for the initial values in the next.

Figure 9 illustrates an example where λ has been determined with a knowledge of λ_o for a case that is representative of a real call center. In particular, the same parameters given in Table 5 have been used, along with the number of servers as tabulated for Systems 1 and 2 in Table 6. The arrival rates that were treated as primary calls in the previous numerical example, were taken as the observed arrival rates here. Using these parameters, the primary arrival rate that is obtained is shown in the figure. We observe that one can have huge differences between the two arrival rates as shown for the beginning of the day in the example. These periods represent cases where the arrivals exceed service capacity, which results in an accumulation of retrial calls. As service capacity is increased, the two curves come closer together coinciding once the capacity is sufficient to satisfy both incoming primary calls and retrials from before. For this particular example, we note that the primary calls curve represents a relatively flat curve. If one were to use the observed calls to determine staffing needs, it is clear that one would end up with a system where capacity varies much more than necessary throughout the day, and mismatches between demand and supply occur.

6 Concluding remarks

We have analyzed the phenomenon of retrials in a call center with abandonments and balking. The system is analyzed both for a single-period and a multi-period setting. For the single period setting a fluid approximation is proposed to estimate the retrial rate in the system. This approximation results in an easy to compute analytical expression for the retrial rate. Using this analytical expression, we illustrate an insensitivity of this rate to key parameters such as the abandonment rate, the individual retrial probability, and the balking probability distribution. This insensitivity property allows us to use the same approximation for models that have an finite buffer. It is shown through numerical examples that the fluid approximation works very well for large call centers that have a utilization that is greater than one. For ρ less than one, the approximation is not appropriate. One may be able to use it coupled with a good approximation of the expected number of busy servers. This will be explored in future research. However note that we are less interested in systems with ρ less than one, since these will lead to much lower number of retrials, that will have an insignificant impact on system performance.

The fluid approximation is also shown to provide an effective method of analysis in the multi-period setting. The retrial probability can no longer be estimated using an analytical formula, and numerical analysis is required. A comparison with simulation illustrates the effectiveness of the method. Using data that resembles the call center in question, numerical examples illustrate that the retrial phenomenon can be substantial and will have an important effect on performance analysis and subsequent system optimization if not taken into account. By comparing systems with different arrival rates and allocation of staff across time periods, it is shown that there is a significant impact on the retrial rate both of arrival rates and staffing distribution. Thus having the wrong arrival rate estimate or improper allocation of staff across time periods will have an effect on the retrials that are generated. Furthermore, it is shown that the shape of the observed arrivals call pattern can be qualitatively different from the shape of the primary calls curve, emphasizing the importance of explicitly modeling retrial behavior in non-stationary systems. This is an important conclusion for call center managers, since typical call center models and software used by call centers ignore the retrial phenomenon. For call centers that operate under moderate or low utilization this will not pose a problem, however for call centers that operate under heavy utilization this will lead to distortions that increase operational costs. We further show how one can estimate the rate of first attempts using the historical data on total call volumes. While future software may make it possible for call centers to track retrials directly, this result is still useful in cleaning up historical data.

In this paper we have demonstrated that there is a strong and significant interaction between staffing and retrials, and as a result that retrials can have an important impact on call center performance. In Aguir et al. [1] it is illustrated how one may think about staffing in the presence of retrials in a single period setting. In future work, we would like to consider staffing schemes for the multi-period setting, that explicitly account for retrials in a call center.

References

1. Aguir MS, Karaesmen F, Aksin OZ, Chauvet F (2003) Analyse du problème des rappels et dimensionnement dans un centre d'appels. Proceedings of MOSIM'03, Toulouse (France)
2. Akşin OZ, Harker PT (2001) Modeling a phone center: Analysis of a multi-channel multi-resource processor-shared loss system. Management Science 47: 324–336
3. Altman E, Jimenez T, Koole G (2001) On the comparison of queueing systems with their fluid limits. Probability in the Engineering and Informational Sciences 15: 165–178
4. Armony M, Maglaras C (2004) On customer contact centers with a call-back option: customer decisions, routing rules, and system design. Operations Research 52: 271–292
5. Artalejo JR (1995) A queueing system with returning customers and waiting line. Operations Research Letters 17: 191–199
6. Baccelli F, Hebuterne G (1981) On queues with impatient customers. In: Kylstra FJ (ed) Performance '81, pp 159–179. North-Holland, Amsterdam
7. Boxma OJ, de Waal PR (1994) Multiserver queues with impatient customers. In: Labetouille J, Roberts JW (eds) The fundamental role of teletraffic in the evolution of telecommunications networks (Proc. ITC-14), pp 743–756. North-Holland, Amsterdam
8. Borst S, Mandelbaum A, Reiman MI (2004) Dimensioning large call centers. Operations Research 52: 17–34
9. Brandt A, Brandt M (1999) On a two-queue priority system with impatience and its application to a call center. Methodology and Computing in Applied Probability 1: 191–210
10. Falin G (1995) Estimation of retrial rate in a retrial queue. Queueing Systems 19: 231–246
11. Falin G, Templeton JGC (1997) Retrial queues. Chapman and Hall, London
12. Gans N, Koole GM, Mandelbaum A (2002) Telephone call centers: Tutorial, review, and research prospects. Manufacturing & Service Operations Management 5: 97–141
13. Garnett O, Mandelbaum A, Reiman M (2002) Designing a call center with impatient customers. Manufacturing & Service Operations Management 4: 208–227
14. Green LV, Kolesar PJ, Svoronos A (1991) Some effects of nonstationarity on multiserver markovian queueing systems. Operations Research 39: 502–511
15. Green LV, Kolesar PJ (1997) The lagged PSA for estimating peak congestion in multiserver markovian queues with periodic arrival rates. Management Science 43: 80–87
16. Hoffman KL, Harris CM (1986) Estimation of a caller retrial rate for a telephone information system. European Journal of Operational Research 27: 207–214
17. Hui MK, Tse DK (1996) What to tell consumers in waits of different lengths: an integrative model of service evaluation. Journal of Marketing 60: 81–90
18. Jimenez T, Koole G (2004) Scaling and comparison of fluid limits of queues applied to call centers with time-varying parameters. OR Spectrum 26: 415–423
19. Mandelbaum A, Massey WA, Reiman MI, Rider R (1999) Time varying multiserver queues with abandonments and retrials. In: Key P, Smith D (eds) Proceedings of the 16th International Teletraffic Conference

20. de Véricourt F, Zhou Y (2003) Managing response time and service quality in a call routing problem. Fuqua School of Business, Duke University, Working paper
21. Whitt W (1999) Improving service by informing customers about anticipated delays. Management Science 45: 192–207
22. Yang T, Templeton JGC (1987) A survey on retrial queues. Queueing Systems 2: 201–233