Available online at www.sciencedirect.com
ScienceDirect
European Journal of Operational Research xxx (2007) xxx–xxx
EUROPEAN JOURNAL OF OPERATIONAL RESEARCH
www.elsevier.com/locate/ejor

Stochastics and Statistics

On the interaction between retrials and sizing of call centers

M. Salah Aguir [a], O. Zeynep Akşin [b,*], Fikri Karaesmen [c], Yves Dallery [a]

[a] *Laboratoire Génie Industriel, Ecole Centrale Paris, Grande Voie des Vignes, 92295 Chatenay-Malabry Cedex, France*
[b] *College of Administrative Sciences and Economics, Koç University, 34450 Sariyer – Istanbul, Turkey*
[c] *Department of Industrial Engineering, Koç University, 34450 Sariyer – Istanbul, Turkey*

Received 26 June 2004; accepted 22 June 2007

Abstract

This paper models a call center as a Markovian queue with multiple servers, where customer impatience, and retrials are modeled explicitly. The model is analyzed as a continuous time Markov chain. The retrial phenomenon is explored numerically using a real example, to demonstrate the magnitude it can take and to understand its sensitivity to various system parameters. The model is then used to assess the impact of disregarding existing retrials in the staffing of a call center. It is shown that ignoring retrials can lead to under-staffing or over-staffing with respect to the optimal, depending on the forecasting assumptions being made.

Keywords: Queueing; Retrials and abandonments; Call centers

1. Introduction

This paper is motivated by a large European call center's problem. As in any call center, a certain number of servers answer customer calls that are placed to this center. 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. 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. 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 the disconnected 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.

Current management policy is to keep the blocking threshold at very low values. Given this choice, the center experiences a lot of calls that are redialing

* Corresponding author. Tel.: +90 212 338 15 45; fax: +90 212 338 16 53.
E-mail addresses: salah.aguir@lgi.ecp.fr (M.S. Aguir), zaksin@ku.edu.tr (O.Z. Akşin), fkaraesmen@ku.edu.tr (F. Karaesmen), dallery@lgi.ecp.fr (Y. Dallery).

0377-2217/$ - see front matter © 2007 Published by Elsevier B.V.
doi:10.1016/j.ejor.2007.06.051

customers who were unable to enter the system on a first attempt. The information system in place in this center does not allow one to distinguish between first-time attempts and redialing customers. As a result call volume forecasts are distorted by the retrials of blocked customers. The staff planning process, which takes call volume forecasts as an input, further exacerbates this distortion. The objective of this study was to document the magnitude of the retrial phenomenon resulting from blocked customers, and to assess the impact of this unknown retrial rate on the subsequent staff planning. Given the importance of staffing costs in the overall budget of a call center, this type of an analysis would serve as a first step in assessing the cost implications of a low-blocking threshold policy.

The results obtained in this paper were instrumental in reorganizing the staffing function of the call center that motivated this research. In particular, being alerted to the magnitude of the retrial phenomenon under a low-blocking threshold policy, communication between the staffing and forecasting functions was enhanced, and regression-based estimators were developed to account for retrials in forecasting.

The following section provides a review of related literature. The model is formulated and analyzed in Section 3. A numerical analysis in Section 4 based on parameters that are representative of the call center in question, demonstrates the significance of the retrial phenomenon, and explores sensitivity to model parameters. Section 5 formalizes the relationship between distorted call volumes and staffing. It is shown that ignoring the existence of redialing customers can lead to erroneous staff planning, and that the error can lead to higher or lower staff levels compared to the optimal.

2. Literature review

We model a call center as a finite queue with blocking, abandonments, and retrials by blocked customers. This model was first formulated in Aguir et al. (2003). Motivated by call centers, Baccelli and Hebuterne (1981) and Brandt and Brandt (1999) 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 (1994). Focusing on exponential abandonment times, Akşin and Harker (2001) and Garnett et al. (2002) treat impatience within specific call center applications. Whitt (1999) analyzes a call center with balking and abandonments. None of these models consider the retrials by blocked customers. In this paper, we show that for certain call centers, ignoring the retrials can lead to significant under or over-staffing.

There is an extensive literature on so-called *retrial queues* (Yang and Templeton, 1987; Falin, 1995; Falin and Templeton, 1997; Artalejo, 1999). Most of the models in this literature do not consider abandonment behavior. Hoffman and Harris (1986) incorporate abandonments and retrials in a model which is also motivated by the problem of a call center. Our analysis is similar, however, we focus explicitly on characterizing the interaction between retrials and staffing. Mandelbaum et al. (1999) consider multi-server systems with abandonments and retrials and propose a fluid approximation for their analysis. Given our objective of understanding the extent of the retrial phenomenon due to blocked calls, we have focused on a steady-state Markov chain analysis herein. The results of this paper, that show the significance of the retrial phenomenon and its impact on staffing, motivated a closer look at the problem of estimating the retrial rate in this call center. In particular, Aguir et al. (2004) investigate the same problem under non-stationary call arrivals. Finally, a recent paper by Artalejo and Pla (2007) investigates approximations based on different truncation schemes for a similar system.

3. Model formulation and analysis

3.1. Problem description

We consider a call center with C service representatives. Customer arrivals are assumed to be a Poisson process with rate λ, and service times are exponentially distributed with rate μ. Customers who are unable to find an idle server upon arrival will be put on hold. In order to limit the number of waiting customers, the size of the call center is limited to K. This implies that the number of customers waiting on hold cannot be more than $K - C$. Waiting customers will abandon the system if their patience threshold is exceeded. We assume that customers abandon with an exponential rate θ. The resulting model is an $M/M/C/K + M$ queue where the last M denotes exponential abandonments.

A customer who arrives when there are K customers already in the system will hear a message, asking him to call back later. We assume that such

Please cite this article in press as: Aguir, M.S. et al., On the interaction between retrials and sizing of call centers, European Journal of Operational Research (2007), doi:10.1016/j.ejor.2007.06.051

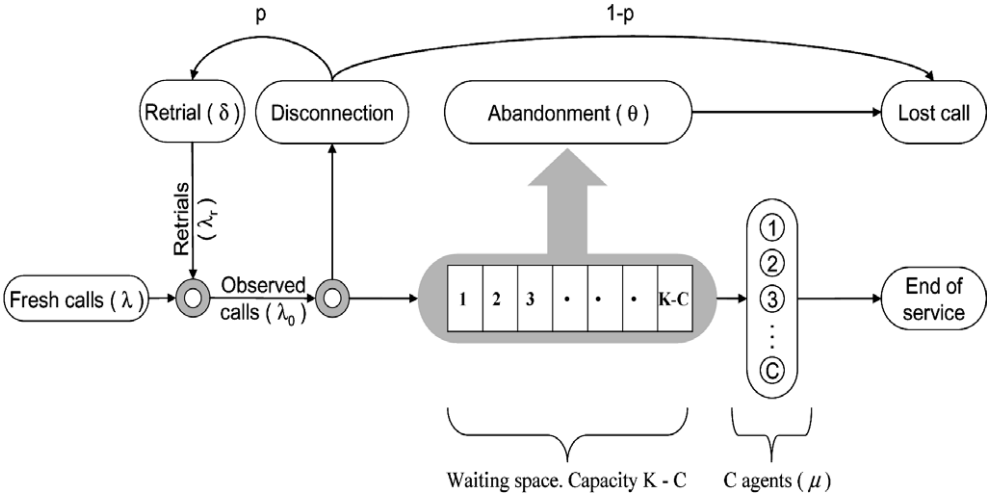

Fig. 1. A call center with blocking, abandonments, and retrials.

a customer will call back or retry with a probability p after an exponential delay of rate δ. It is assumed that customers who abandon will not retry. This assumption is made primarily because the objective of the study is to assess the impact of retrials from blocked calls, resulting from the policy of keeping $K - C$ at low values. For more precise estimation purposes, the case of retrials from calls that have abandoned is modeled in Aguir et al. (2004). Fig. 1 illustrates the functioning of the call center. Note that call arrivals to this center, or observed calls, will consist of first-time attempts (fresh calls) and retrials.

3.2. Modeling as a continuous time Markov chain

The system in Fig. 1 can be modeled as a continuous time Markov chain (CTMC) in two dimensions. The first dimension corresponds to the real queue, consisting of the C servers and the waiting space. The total number of customers in this queue cannot exceed K. The second dimension corresponds to the queue of customers who have been blocked and who are waiting to reattempt their call. In the retrial literature, this queue is known as the orbit. The dimension of this queue is, in general, assumed to be infinite. Thus the state (m, n), $m = 0, 1, \ldots, K$, $n = 0, 1, \ldots$ corresponds to a system with m customers in the real queue and n customers in orbit. These latter customers will retry after an exponentially distributed time with rate δ. Let $\pi_{m,n}$ denote the steady-state probability of being in state (m, n). The balance equations of this CTMC are given by:

For $n = 0$:

$$\lambda\pi_{0,0} = \mu\pi_{1,0},$$
$$(\lambda + m\mu)\pi_{m,0} = \lambda\pi_{m-1,0} + \delta\pi_{m-1,1} + (m+1)\mu\pi_{m+1,0}, 1 \leqslant m \leqslant C-1,$$
$$(\lambda + C\mu + (m-C)\theta)\pi_{m,0} = \lambda\pi_{m-1,0} + \delta\pi_{m-1,1} + (C\mu + (m+1-C)\theta)\pi_{m+1,0}, C \leqslant m \leqslant K-1,$$
$$(p\lambda + C\mu + (K-C)\theta)\pi_{K,0} = \lambda\pi_{K-1,0} + \delta\pi_{K-1,1} + (1-p)\delta\pi_{K,1}$$

and for $n \geqslant 1$:

$$(\lambda + n\delta)\pi_{0,n} = \mu\pi_{1,n},$$
$$(\lambda + m\mu + n\delta)\pi_{m,n} = \lambda\pi_{m-1,n} + (n+1)\delta\pi_{m-1,n+1} + (m+1)\mu\pi_{m+1,n}, 1 \leqslant m \leqslant C-1$$
$$(\lambda + C\mu + (m-C)\theta + n\delta)\pi_{m,n} = \lambda\pi_{m-1,n} + (n+1)\delta\pi_{m-1,n+1} + (C\mu + (m+1-C)\theta)\pi_{m+1,n}, C \leqslant m \leqslant K-1$$
$$(p\lambda + C\mu + (K-C)\theta + n(1-p)\delta)\pi_{K,n} = \lambda\pi_{K-1,n} + (n+1)\delta\pi_{K-1,n+1} + (n+1)(1-p)\delta\pi_{K,n+1} + p\lambda\pi_{K,n-1}$$

The steady-state probabilities $\pi_{m,n}$ for this system can be calculated by truncating the infinite dimensional orbit at some value K_2, and then using known general numerical methods for Markov chains (see Artalejo and Pla, 2007). In order to perform the calculation efficiently, we adapt a method by Tran-Gia and Mandjes (1997) that exploits the special struc-

Please cite this article in press as: Aguir, M.S. et al., On the interaction between retrials and sizing of call centers, European Journal of Operational Research (2007), doi:10.1016/j.ejor.2007.06.051

ture of this Markov chain. Details of this method are omitted from this paper, but can be obtained from the authors.

4. Numerical analysis of the retrial phenomenon

In this section, we first demonstrate the significance of the retrial phenomenon. Table 1 illustrates, for several examples, the evolution of the steady-state retrial rate as a function of fresh call volume, where the steady-state retrial rate, λ_r, is given by:

$$\lambda_r = \sum_{n=1}^{K_2} n\delta \sum_{m=0}^{K} \pi_{m,n}. \quad (1)$$

This table is obtained through the numerical computation described above. The truncation of the CTMC performed in the previous subsection was validated via a comparison to results obtained from a discrete-event simulation. The call centers shown in the columns of the table are identical except for their size in terms of number of servers and the queue size. The capacity $K - C$ of the queue was chosen for each example in order to provide a realistic number with respect to call center size, and to emulate the low-blocking threshold policy in use. The fresh call rate was varied to cover a system load ρ ranging from 70% to 130%. Here, $\rho = \lambda/C\mu$ is different from the effective system load $\lambda_0/C\mu$, where $\lambda_0 = \lambda + \lambda_r$ denotes the observed call rate. All other parameters are derived from real data. For the rest of this section, the values of the main parameters are given by Table 2: when not mentioned, parameters will take values from this table.

We observe from Table 1 that for loads higher than 100%, the retrial rate is of similar magnitude to the fresh call rate. Thus, for the first example, the retrial rate is 7.82 for a fresh call rate of 9.75 resulting in an observed call rate of 17.57. For the last example, the retrial rate is 31.06 for a fresh call rate of 39.00 leading to an observed call rate of 70.06. For all cases, such doubling of the observed demand with respect to first attempts will have a significant impact on call center staffing. The qualitative behavior of the system is similar for the four examples. However, we observe that for the larger call centers, retrials remain close to zero for higher loads compared to the smaller centers, illustrating the well known fact that stochastic effects diminish as we increase call center size.

The results of Table 1 are shown in Fig. 2 which depicts the ratio λ_r/λ for the four cases as a function of the system load ρ. We can observe that the ratio is a non-decreasing function of ρ. For more important loads, the proportion of retrials become higher than fresh calls.

We next explore the sensitivity of the retrial rate to various system parameters. Especially for parameters that capture client behavior like p and δ, sensitivity to estimation errors in these parameters is important. Fig. 3 depicts the effect of the retrial probability p on the retrial rate for different values of ρ. All parameters, except p, take the values given in Table 2. We note that for high values of p, the results are very sensitive to the value of this parameter and an estimation error can lead to an important difference in the retrial rate. We also note that these curves are steeper for higher values of the system load, indicating a higher sensitivity in busy call centers.

The graph in Fig. 4 explores the sensitivity of the retrial rate to the individual retrial rate parameter δ.

Table 2
Parameter values

C	K	μ	δ	θ	p
100	105	0.3	1	0.3	0.8

Table 1
Evolution of the retrial rate λ_r as a function of the fresh call rate and the system size

ρ (%)	$C = 25$ and $K = C + 2$		$C = 50$ and $K = C + 3$		$C = 75$ and $K = C + 4$		$C = 100$ and $K = C + 5$	
	Fresh call rate	Retrial rate	Fresh call rate	Retrial rate	Fresh call rate	Retrial rate	Fresh call rate	Retrial rate
70.00	5.25	0.06	10.50	0.01	15.75	0.00	21.00	0.00
80.00	6.00	0.27	12.00	0.14	18.00	0.07	24.00	0.03
90.00	6.75	0.79	13.50	0.76	20.25	0.65	27.00	0.53
100.00	7.50	1.80	15.00	2.50	22.50	2.94	30.00	3.25
110.00	8.25	3.35	16.50	5.74	24.75	7.89	33.00	9.93
120.00	9.00	5.40	18.00	10.26	27.00	15.04	36.00	19.82
130.00	9.75	7.82	19.50	15.54	29.25	23.29	39.00	31.06

Please cite this article in press as: Aguir, M.S. et al., On the interaction between retrials and sizing of call centers, European Journal of Operational Research (2007), doi:10.1016/j.ejor.2007.06.051

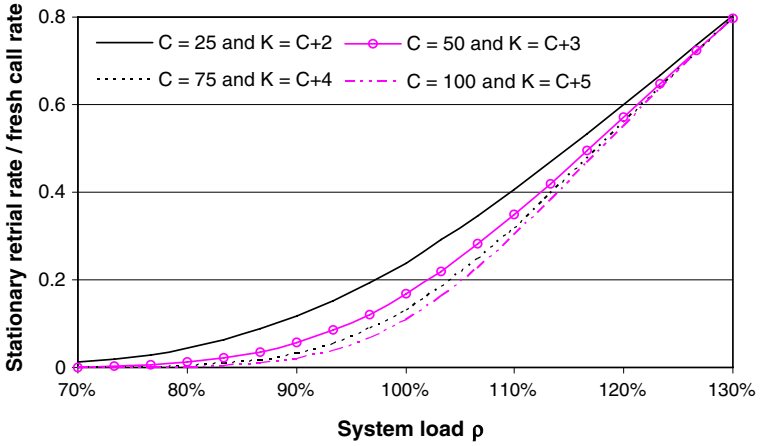

Fig. 2. Evolution of the retrial rate as a function of the fresh call rate.

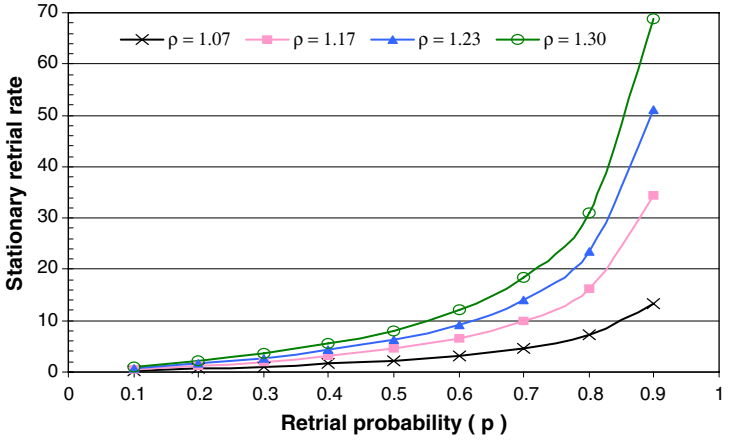

Fig. 3. Evolution of the retrial rate as a function of the retrial probability for a system with 100 servers.

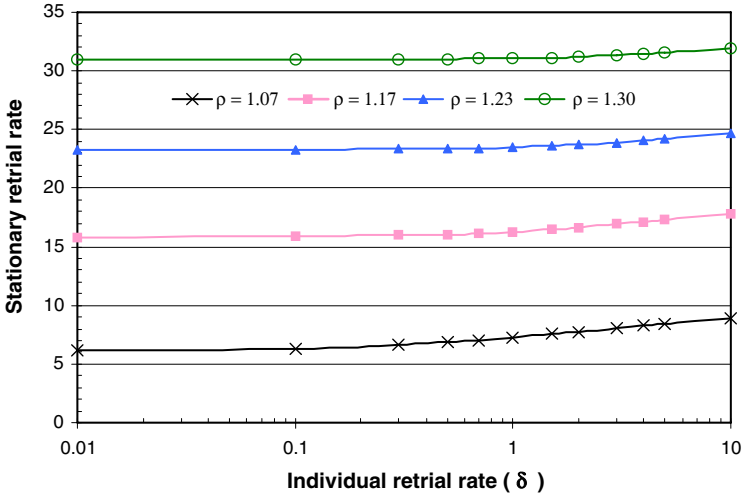

Fig. 4. The effect of δ on the retrial rate.

Please cite this article in press as: Aguir, M.S. et al., On the interaction between retrials and sizing of call centers, European Journal of Operational Research (2007), doi:10.1016/j.ejor.2007.06.051

It is interesting to note that these curves are relatively flat. This is a consequence of the double role that this parameter plays: the larger δ is, the shorter will be the length of the orbit. However, at the same time high δ implies that customers retry rapidly, which balances the effect via an increase in the real queue size. The curves become flatter as a function of the system load, implying that for high-call volume systems this insensitivity property will be more pronounced. These results are encouraging, since estimating the parameter p is easier than estimating δ. The particular shape of the curves suggests an investigation of the system behavior at 0 and ∞.

When δ has a very high value, we can assume that a disconnected customer will retry immediately with probability p and will leave the system with probability $1-p$. Repeating this argument, we can see that the number of retrials per disconnected customer is geometrically distributed and therefore, the average number of attempts is $p/(1-p)$. Thus we can estimate the stationary retrial rate for large values of δ by:

$$\lambda_r = \frac{p\lambda}{(1-p)}\pi_K, \tag{2}$$

where π_K refers to the stationary probability of finding K customers in the $M/M/C/K+M$ (without retrials) system. The examples in Table 3, show that this method gives an exact stationary retrial rate for sufficiently large values of δ, and a good upper bound for realistic values of δ, especially when the system load is important.

For very low values of δ, since customers will redial after a long period of time, we can assume that the retrial flow is Poisson and is independent of the fresh calls. In this case our analysis becomes similar to that in So and Tang (1996). We first express the stationary retrial rate as a function of the observed rate:

$$\lambda_r = p\lambda_0\pi_K, \tag{3}$$

where π_K denotes here the stationary probability of finding K customers in the $M/M/C/K+M$ (without retrials) system. This time, the arrival rate will be equal to λ_0. Thus, π_K is a function of λ_0. Since $\lambda_0 = \lambda + \lambda_r$, we can say that π_K is a function of λ_r. As in So and Tang (1996), a fixed point equation is then obtained from Eq. (3). The fixed point procedure to determine the stationary retrial rate is the following:

1. Begin by setting $\lambda_0^1 = \lambda$.
2. Compute λ_r^1 with Eq. (3).
3. Set $\lambda_0^{i+1} = \lambda + \lambda_r^i$.
4. Repeat the last two steps until $|\lambda_r^{i+1} - \lambda_r^i| \leqslant \epsilon$.

Table 3
Upper and lower bounds for the stationary retrial rate

ρ	δ			
	$\delta \to 0$	0.01	10	$\delta \to \infty$
1.07	6.16	6.18	8.95	10.65
1.17	15.82	15.82	17.83	19.82
1.23	23.26	23.26	24.66	26.67
1.3	30.96	30.96	31.92	33.86

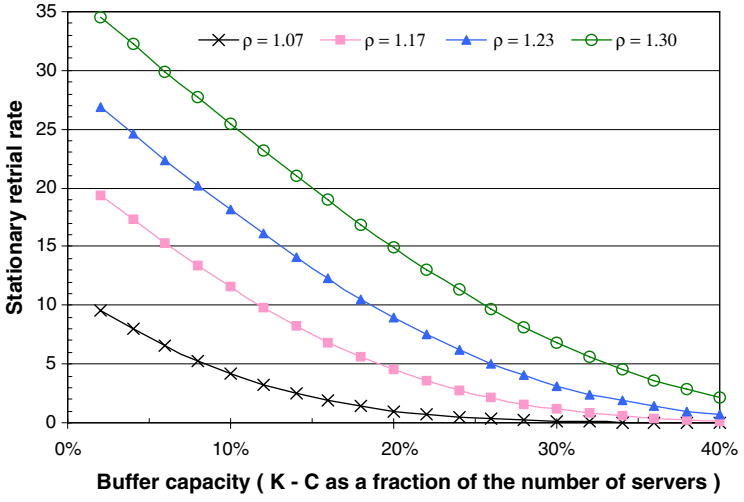

Fig. 5. The effect of a finite waiting space on the retrial rate.

Please cite this article in press as: Aguir, M.S. et al., On the interaction between retrials and sizing of call centers, European Journal of Operational Research (2007), doi:10.1016/j.ejor.2007.06.051

As shown in Table 3 this approach assures a good lower bound for realistic values of δ.

We next look at some capacity related parameters and their impact on the retrial rate. Fig. 5 depicts the effect of increasing the waiting space $K - C$ on the retrial rate, for a call center with the parameters of Table 2. The graph demonstrates the significant level of retrials that can be experienced by a busy call center with a small blocking threshold.

We now look at the impact of the number of servers on the stationary retrial rate. In order to avoid mixing the effect of a change in $K - C$ or the system utilization $\rho = \lambda/C\mu$ as we change C, we perform this analysis such that ρ is kept constant for different values of C by adjusting the arrival rate λ. The ratio $K - C/C$ will also be kept constant (equal to 0.05) in these examples. Fig. 6 depicts the evolution of the ratio of the retrial rate to the fresh call rate for several values of ρ, as a function of C. As expected, the ratio diminishes as a function of size since stochastic effects diminish according to system size. Here we can observe that this ratio approaches zero for $\rho = 0.9$ and $C = 200$. We note, however, that even for a large system, if the system load exceeds one, the retrial phenomenon will have a significant effect. It is possible to estimate the retrial rate making use of a fluid approximation proposed in Aguir et al. (2004). This approximation works well for large number of servers and large system loads. For a system with $\rho = 1.17$ and $C = 40$ it underestimates the retrial rate by about 80%, however, for $\rho = 1.3$ and $C = 200$, the error becomes less than 2%.

5. Ignoring retrials in call center staffing

In this section, we are going to demonstrate the impact an explicit modeling of retrials will have on staff sizing in a call center. We will assume that the call center is determining its staff size in order to minimize the number of servers, while attaining a given completion rate, α_{NR}. The completion rate, alternately called the acceptance rate below, gives the proportion of calls who are connected to a service representative (i.e. who are not blocked and who do not abandon). We will determine the optimum number of servers for the model with retrials and compare it to optimizations of the $M/M/C/K + M$ queue without retrials done by resolving the steady-state distribution of the corresponding Markov chain, thus assessing the lack of accuracy from ignoring retrials while staffing call centers.

In practice, the data necessary for such a comparison would be taken from a data base containing historical data, which is typically used to obtain call volume forecasts. We suppose that such a data base has call volumes experienced in a time period along with the corresponding number of servers present during that time period.

Disregarding retrials, the acceptance rate, α_{NR} can be written for an $M/M/C/K + M$ queue as

$$\alpha_{NR} = 1 - \frac{\lambda\pi_K + \theta L_Q}{\lambda}, \tag{4}$$

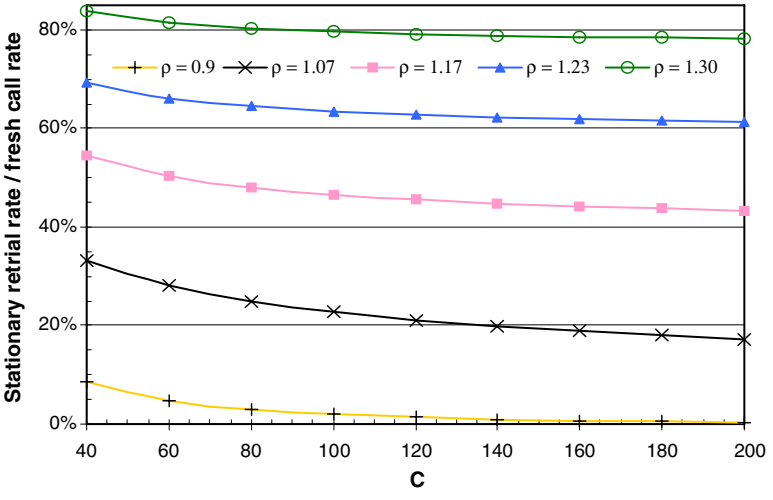

Fig. 6. The Effect of the system size C for $K - C = 5\%$ of C.

Please cite this article in press as: Aguir, M.S. et al., On the interaction between retrials and sizing of call centers, European Journal of Operational Research (2007), doi:10.1016/j.ejor.2007.06.051

where π_K denotes the stationary probability of finding K clients in the system (blocking probability), and L_Q is the average queue length.

For this system without retrials, calculating the stationary probabilities is straightforward (Akşin and Harker, 2000; Garnett et al., 2002). An inverse recursive procedure is then used to determine the number of servers necessary to obtain a desired service level, expressed in terms of the acceptance rate given by Eq. (4). Note here that there is no distinction between fresh calls and observed calls.

In the case when we acknowledge the presence of retrials, we first need to extract the fresh call rate from the observed call volume data in the database. This can be done as long as the corresponding number of servers for a given time period is also known. The extraction is performed using the following recursive procedure. Let $f(\lambda, \mu, c, K_1, K_2, \theta, \delta, p)$ denote the function that calculates the observed call rate λ_{obs} when the other parameters are given. λ_0 denotes the desired observed call rate and ϵ the desired precision:

initialization
$x_{low} := 0$ and $x_{high} := \lambda_0$
While error$> \epsilon$,
  $\lambda := (x_{low} + x_{high})/2$
  $\lambda_{obs} := f(\lambda, \mu, c, K_1, K_2, \theta, \delta, p)$
  error $:= |\lambda_{obs} - \lambda_0|$
  If(error $> \epsilon$)
    If($\lambda_0 > \lambda_{obs}$)
      then $x_{low} := (x_{low} + x_{high})/2$
    else $x_{high} := (x_{low} + x_{high})/2$
    end if
  end if
End while

Table 4 demonstrates an example where the observed call volume in the historical data base is 15. Each column of the Table provides the corresponding fresh call rate, assuming a different number of servers present during that time period: if for example 25 servers answered calls on the day when this data was collected, then the fresh call rate would be 9.79 for an observed call volume of 15.

Table 4
λ as a function of C for a system where $K = C + 5, \lambda_o = 15, \mu = 0.3, \theta = 0.3, \delta = 1, p = 0.8$

C	25	30	35	40	45	50	55
λ	9.79	10.81	11.78	12.69	13.48	14.15	14.62

Let us denote by α_R the proportion of all calls that connect to a service representative. This, of course, is a service level that can be calculated easily from available data (number of calls completed/number of calls observed). Once the fresh call rate λ is obtained, α_R can be determined using the expression:

$$\alpha_R = \frac{\lambda_0 - \lambda \sum_{n=0}^{K_2} \pi_{K,n} - \delta \sum_{n=1}^{K_2} n\pi_{K,n} - \theta L_Q}{\lambda_0}. \quad (5)$$

In this equation, the probabilities $\pi_{K,n}$, $n = 0, 1, \ldots, K_2$ are those corresponding to the Markov chain in Section 3.2 , and require knowledge of the fresh call rate λ. This time L_Q represents the average queue length for the model with retrials. It can be calculated using the relationship:

$$L_Q = \sum_{i=C+1}^{K} \left((i - C) \sum_{n=0}^{K_2} \pi_{i,n} \right).$$

The numerator of the right hand side in Eq. (5) expresses the number of calls that are connected to a service representative by subtracting the total number of blocked calls (first attempts and retrials) and abandoned calls. Note that the numerator can equivalently be written as the rate of calls served by the call center, μ_D. We can then rewrite α_R as:

$$\alpha_R = \frac{\mu_D}{\lambda_0} = \frac{\mu \sum_{i=1}^{K} \min(i, C) \sum_{n=0}^{K_2} \pi_{i,n}}{\lambda_0}. \quad (6)$$

Fig. 7 and Table 5 illustrate the impact of ignoring retrials in dimensioning a call center. In this example, the optimal number of servers that ensure a desired acceptance rate are shown for three different systems. For all three systems λ_0 is taken to be 15. The first system corresponds to a system that does not consider retrials, either because their presence is ignored or unknown. For this system, the fresh call arrival rate is going to be equal to the observed call volume. As a result, the number of servers will be a function of the forecasted observed call volume. Using Eq. (4) in conjunction with a numerical inversion procedure, the optimal number of servers to achieve the desired service level can be obtained. The second system represents the case where retrials are explicitly modeled. It is assumed that the historical data comes from a time period where 25 servers were present answering calls. Thus, for this system the fresh call arrival rate will be 9.79 by Table 4. This time the dimensioning of the call center is performed using the formula in (5) in

Please cite this article in press as: Aguir, M.S. et al., On the interaction between retrials and sizing of call centers, European Journal of Operational Research (2007), doi:10.1016/j.ejor.2007.06.051

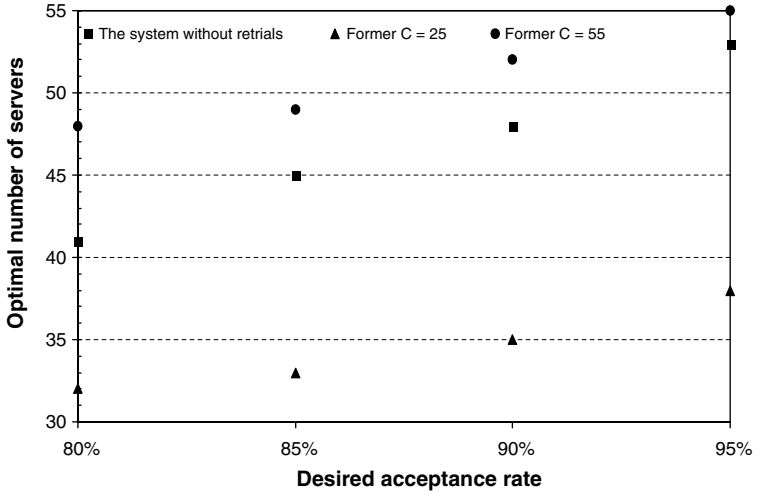

Fig. 7. The effect of ignoring retrials in dimensioning a call center.

Table 5
The effect of ignoring retrials on the optimum number of servers and effective acceptance rate

Desired α_R	Without retrials	Former $C = 25$	Former $C = 30$	Former $C = 35$	Former $C = 40$	Former $C = 45$	Former $C = 50$	Former $C = 55$
80%	41	32 (98.7%)	35	39	41	44	46	48 (58.8%)
85%	45	33 (99.7%)	37	40	43	46	48	49 (73.1%)
90%	48	35 (99.9%)	39	42	45	48	50	52 (82.6%)
95%	53	38 (99.9%)	41	45	48	51	53	55 (93.5%)

conjunction with a numerical method. The third system is identical to the second one except for the assumption that there were 55 servers on the day the call volumes were observed. The optimal number of servers is summarized in Table 5.

We observe in Fig. 7 that by ignoring the presence of retrials, one can under-size or over-size the system. If we consider for example the case of an acceptance rate of 90%, the optimal system size can go from 48 servers if retrials are ignored to 35 or 52 depending on the historical number of servers assumed. The over-sizing occurs if on the day that the call volume data was recorded, there were a small number of servers present. This leads to an important difference between λ and λ_0. By ignoring this difference, the system without retrials will thus over-size with respect to what is optimal. On the other hand, when the data is assumed to be recorded on a day when many servers are present, then the two arrival rates λ and λ_0 will be very close to each other (since $\lambda_0/C\mu < 1$ on the day the data was recorded). When this is the case, the two systems will consider very similar call arrival rates, however, by explicitly accounting for the additional calls that will be generated by retrying customers, the model with retrials will experience a higher overall arrival rate. As a result, the model ignoring retrials will under-size the system. This difference is higher for lower acceptance levels, since one will have more customers that retry in such a system. This example illustrates the close interaction between forecasting and staffing in a call center with retrials where the retrial rate cannot be directly measured. In a dynamic setting, the call center might oscillate between over and under-staffing. However, even then, service levels will suffer and staffing costs can not be optimized.

Since ignoring retrials in Fig. 7 leads to inappropriate staffing, we illustrate in Table 5, columns 3 and 9 the effective acceptance rate when retrials are ignored. As shown by this table, the difference can be significant. In fact, if the desired level is, for example, 80%, then the realized level can be 98.7% or 58.8%. Thus, we can say that ignoring retrials can lead to either an overstaffed system (with the underlying costs) with an excess of Quality of Service, or to an understaffed system with a lack in the QoS (with unhappy customers).

Please cite this article in press as: Aguir, M.S. et al., On the interaction between retrials and sizing of call centers, European Journal of Operational Research (2007), doi:10.1016/j.ejor.2007.06.051

The service level we have used in our analysis, $\alpha_R = \mu_D/\lambda_0$, is a measure of all calls that are served. This measure is used in a setting where the call center observes all calls and cannot distinguish between fresh calls and retrials. However, from a customer perspective, the more appropriate service level measure would have been one that represents the service level experienced by customers, i.e. $\alpha'_R = \mu_D/\lambda$. This type of a service measure is based on the rate of first-time attempts. For call centers that do not have the technical capability of distinguishing retrials from first-time attempts, it is a service level that will not be directly measurable, making it less attractive from a managerial standpoint.

If we consider the example studied in Fig. 7, we note that for the system that ignores retrials, both service measures will lead to the same staffing, since λ_0 is taken to be the same as λ in this case. Let us consider the settings where retrials are taken into account in the planning. We want to explore the staffing level C for a given service level α under both settings. So we set $\alpha_R = \alpha'_R = \alpha$ or equivalently $\mu_D/\lambda_0 = \mu'_D/\lambda$. Since $\lambda \leqslant \lambda_0$ it follows that $\mu'_D \leqslant \mu_D$ implying that our procedure would yield $C(\alpha'_R) \leqslant C(\alpha_R)$. In other words, adopting a different service level measure that is based on customer experience, we would observe the points for the cases that take retrials into account in Fig. 7 to shift downwards. For the example in question, this downward shift is illustrated in Fig. 8. We observe that ignoring the retrials leads to over-staffing in both cases with this service level measure. Indeed, this service level measure will lead to over-staffing all the time, irrespective of the former staff level. To illustrate the reason, consider a setting where we assume an infinite staff level on the day call volumes were observed. This would ensure that $\lambda_0 = \lambda$. Now, for a given staff level C, the system with retrials will serve more calls than the system without retrials, since the blocked calls are lost in the latter. This implies that for identical C we would have $\mu_D(\text{with retrials}) > \mu_D(\text{no retrials})$, leading to $\alpha'_R > \alpha'_{NR}$. It is clear that for the same service level $\alpha'_R = \alpha'_{NR}$, we get $C(\text{with retrials}) < C(\text{no retrials})$.

It is also possible to compare staffing obtained from the model with retrials to a well known simple and robust rule used to dimension call centers: the *square root staffing rule* (Gans et al., 2003). This rule determines the appropriate size for a call center given an offered load λ/μ and a service grade objective. It is shown for various systems ($M/M/C$, $M/M/C/K$, $M/M/C+M$) that the optimal staffing level is given by: $(\lambda/\mu) + \beta\sqrt{(\lambda/\mu)}$, where the coefficient β reflects the desired service level or the profit-cost tradeoffs (Halfin and Whitt, 1981; Massey and Wallace, Forthcoming; Garnett et al., 2002). Using the square root staffing rule for the $M/M/C+M$ model (Garnett et al., 2002) for the example in Fig. 7, we note that the same staffing levels as those obtained by the exact analysis that ignores retrials are obtained, despite the fact that blocked calls are ignored by the square root staffing analysis. In other examples with larger offered loads and lower acceptance rates as the service grade objective (i.e. in settings where we expect to see more retrials) square root staffing tends to over-staff compared to the system with retrials. We do not know whether this over-staffing is due to the retrials that are being ignored by the square root staffing rule, or the blocked calls that are ignored by the

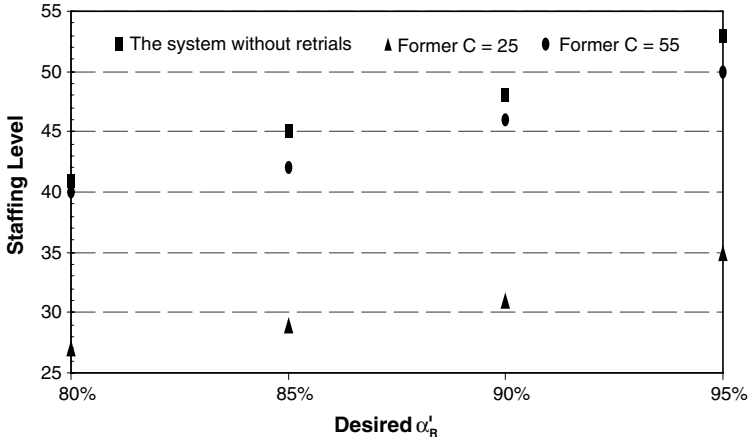

Fig. 8. The effect of ignoring retrials in dimensioning with α'_R as service level measure.

Please cite this article in press as: Aguir, M.S. et al., On the interaction between retrials and sizing of call centers, European Journal of Operational Research (2007), doi:10.1016/j.ejor.2007.06.051

$M/M/C+M$ model. As a result, we do not report these examples numerically herein. In this instance, the more appropriate comparison would have been the square root staffing rule developed for an $M/M/C/K+M$ model or better still such a model that also considers retrials. Approximate staffing rules for both of these systems remain to be explored and establish an important direction for future research.

6. Concluding remarks

In parallel work, Aguir et al. (2004) develops the analysis in this paper in a multi-period setting. A fluid approximation is proposed to analyze this system to overcome some of the computational issues involved with a Markov chain analysis. The fluid approximation allows for a non-stationary analysis of the system. Developing an optimal staffing rule for a non-stationary system with retrials remains as an interesting future research direction.

Acknowledgements

The authors would like to acknowledge helpful discussions with Fabrice Chauvet, Rabie Nait-Abdallah and Thierry Prat (Bouygues Telecom) during the course of this project, and thank anonymous referees for their comments that helped improve the paper.

References

Aguir, M.S., Karaesmen, F., Akşin, Z., Chauvet, F., 2003. Analyse du problème des rappels et dimensionnement dans un centre d'appels. In: Proceedings of MOSIM'03, Toulouse, France.

Aguir, M.S., Karaesmen, F., Akşin, Z., Chauvet, F., 2004. The impact of retrials on call center performance. OR Spectrum 26, 353–376.

Akşin, O.Z., Harker, P.T., 2000. Computing performance measures in a multi-class, multi-resource, processor shared loss system. European Journal of Operational Research 123 (1), 61–72.

Akşin, O.Z., Harker, P.T., 2001. Modeling a phone center: Analysis of a multi channel multi-resource processor shared loss system. Management Science 47, 324336.

Artalejo, J.R., 1999. Accessible bibliography on retrial queues. Mathematical and Computer Modeling 30, 1–6.

Artalejo, J.R., Pla, V., 2007. On the impact of customer balking, impatience and re-trials in telecommunication systems. Working Paper, Department of Statistics and Operations Research, Complutense University of Madrid.

Baccelli, F., Hebuterne, G., 1981. On queues with impatient customers. In: Kylstra, F.J. (Ed.), Performance '81. North-Holland, Amsterdam, pp. 159–179.

Boxma, O.J., de Waal, P.R., 1994. Multi-server queues with impatient customers. In: Labetouille, J., Roberts, J.W. (Eds.), The Fundamental Role of Teletraffic in the Evolution of Telecommunications Networks (Proceedings of ITC-14). North-Holland, Amsterdam, pp. 743–756.

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, 191210.

Falin, G., 1995. Estimation of retrial rate in a retrial queue. Queueing Systems 19, 231–246.

Falin, G.I., Templeton, J.G.C., 1997. Retrial Queues. Chapman & Hall.

Gans, N., Koole, G.M., Mandelbaum, A., 2003. Telephone call centers: Tutorial, review, and research prospects. Manufacturing & Service Operations Management 5, 79–141.

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

Halfin, S., Whitt, W., 1981. Heavy-traffic limits for queues with many exponential servers. Operations Research 29, 567–587.

Hoffman, K.L., Harris, C.M., 1986. Estimation of a caller retrial rate for a telephone information system. European Journal of Operational Research 27, 207–214.

Mandelbaum, A., Massey, W.A., Reiman, M.I., Rider, R., 1999. Time varying multi-server queues with abandonments and retrials. In: Key, P., Smith, D. (Eds.), Proceedings of the 16th International Teletraffic Conference.

Massey, W.A., Wallace, R.B., forthcoming. An optimal design of the M/M/C/K queue for call centers. Queueing Systems.

So, K.C., Tang, C.S., 1996. On managing operating capacity to reduce congestion in service systems. European Journal of Operational Research 92, 83–98.

Tran-Gia, P., Mandjes, M., 1997. Modeling of customer retrial phenomenon in cellular mobile networks. IEEE Journal on Selected Areas in Communications 15, 1406–1414.

Whitt, W., 1999. Improving service by informing customers about anticipated delays. Management Science 45, 192–207.

Yang, T., Templeton, J.G.C., 1987. A survey on retrial queues. Queueing Systems 2, 201–233.

Please cite this article in press as: Aguir, M.S. et al., On the interaction between retrials and sizing of call centers, European Journal of Operational Research (2007), doi:10.1016/j.ejor.2007.06.051