ON THE INTERACTION BETWEEN RETRIALS AND SIZING OF CALL CENTERS

M. Salah Aguir∗, O. Zeynep Aşin†
Fikri Karaesmen‡‡, Yves Dallery∗

∗ Laboratoire Génie Industriel
Ecole Centrale Paris
Grande Voie des Vignes
92295 Chatenay-Malabry Cedex FRANCE

† Graduate School of Business
Koç University
34450, Sariyer - Istanbul, Turkey

‡‡ Department of Industrial Engineering
Koç University
34450, Sariyer - Istanbul, Turkey

salah.aguir@lgi.ecp.fr, zaksin@ku.edu.tr, fkaraesmen@ku.edu.tr, dallery@lgi.ecp.fr

Corresponding author: O. Zeynep Aşin

June 2004
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 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 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 Aksin 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.

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 $\lambda$, and service times are exponentially distributed with rate $\mu$. 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 $\theta$. 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 them to call back later. We assume that such a customer will call back or retry with a probability $p$ after an exponential delay of rate $\delta$. 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). Figure 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 Figure 1 can be modeled as a Continuous Time Markov Chain (CTMC) in two dimensions, as shown in Figure 2. 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 $K_1$. 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 $(m, n)$, $m = 0, 1, \ldots, K_1$, $n = 0, 1, \ldots$ 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 $\delta$.

In order to analyze this Markov Chain, the infinite dimensional orbit will be truncated at a sufficiently large $K_2$ value, such that states corresponding to larger orbit values have negligible probabilities. The non-zero transition probabilities $P_{(m,n)(i,j)}$ of this truncated CTMC can be stated as follows:

- If $n = 0$:
  \[
  \begin{align*}
  P_{(m,n)(m+1,n)} &= \lambda \text{ if } m < K_1 \\
  P_{(m,n)(m-1,n)} &= m \mu \text{ if } 0 < m \leq C \\
  P_{(m,n)(m-1,n)} &= C \mu + (m - C)\theta \text{ if } C < m \leq K_1 \\
  P_{(m,n)(m,n+1)} &= p \lambda \text{ if } m = K_1
  \end{align*}
  \]
If 0 < n ≤ K₂:

\[
\begin{align*}
P_{(m,n)(m+1,n)} &= \lambda \text{ if } m < K_1 \\
P_{(m,n)(m+1,n-1)} &= n \delta \text{ if } m < K_1 \\
P_{(m,n)(m-1,n)} &= m \mu \text{ if } 0 < m \leq C \\
P_{(m,n)(m-1,n)} &= C \mu + (m-C)\theta \text{ if } C < m \leq K_1 \\
P_{(m,n)(n,n+1)} &= p \lambda \text{ if } m = K_1 \\
P_{(m,n)(m,n-1)} &= n (1-p) \delta \text{ if } m = K_1
\end{align*}
\]

If n = K₂:

\[
\begin{align*}
P_{(m,n)(m+1,n)} &= \lambda \text{ if } m < K_1 \\
P_{(m,n)(m+1,n-1)} &= n \delta \text{ if } m < K_1 \\
P_{(m,n)(m-1,n)} &= m \mu \text{ if } 0 < m \leq C \\
P_{(m,n)(m-1,n)} &= C \mu + (m-C)\theta \text{ if } C < m \leq K_1 \\
P_{(m,n)(m,n-1)} &= n (1-p) \delta \text{ if } m = K_1
\end{align*}
\]

Steady state probabilities Π_{(m,n)} for this system can be calculated via known general numerical methods for Markov Chains. In order to render the numerical computation more efficient, we propose to adapt a special method that exploits the structure of the CTMC proposed by Tran-Gia and Mandjes (1997) for the study of the retrial phenomenon in cellular mobile networks. Their model differs from ours in that there is a finite customer population in their model and no queue for waiting customers. Our proposed algorithm develops a recursive algorithm that computes steady-state probabilities for a given number of clients n, in the orbit. The value of n varies between K₂ and 0 and, for each n value the probability of a state (m, n), 1 ≤ m ≤ K₁ is calculated as a function of all states (i, n), 0 ≤ i ≤ m, and also all states (i, n + 1) for n < K₂. Finally the steady-state will be determined as a function of the state (0, K₂). The complete pseudo-code of the algorithm is provided in Appendix A.

4 Numerical Analysis of the Retrial Phenomenon

In this Section, we first validate the truncation of the CTMC that was performed in the previous subsection, via a comparison to results obtained from a discrete-event simulation. Using the same example that is used for the validation, we then demonstrate the significance of the retrial phenomenon.
Figures 3 and 4 show the evolution of the steady state retrial rate as a function of fresh call volume, where the steady state retrial rate, $T_r$, is given by:

$$T_r = \sum_{n=1}^{K_2} n \delta \sum_{m=0}^{K_1} \Pi_{m,n}$$  \hspace{1cm} (1)

These curves are obtained both through the numerical computation described above, and via a simulation of the original system. The call centers shown in these figures are identical except for their size in terms of number of servers (100 versus 25) and the queue size. The capacity $K$ 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 ranging from 70 % to 130 %. All other parameters are derived from real data.

The first observation we make is that the curves obtained via our truncated CTMC calculation coincide with the simulation results for the original system. This demonstrates that the truncation does not distort the results in terms of the retrial rate estimation. We further observe 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 31 for a fresh call rate of 39 resulting in an observed call rate of 70. For the second example, the retrial rate is 7.8 for a fresh call rate of 9.8 leading to an observed call rate of 17.6. For both cases, such doubling of the observed demand with respect to first-attempts will have a significant impact on call center staffing. The qualitative nature of the curves are similar for the two examples. However we observe that for the larger call center, retrials are very close to zero until the system load reaches 93 %. This number is 77 % for the small center, illustrating the well known fact that stochastic effects diminish as we increase call center size.

We next explore the sensitivity of the retrial rate to various system parameters. Especially for parameters that capture client behavior like $p$ and $\delta$, it is important to understand the sensitivity of the results to estimation errors in these parameters. Figure 5 depicts the effect of the retrial probability $p$ on the retrial rate for different values of $\lambda$. This call center has $C = 100$ and $K - C = 5$. 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. These curves are steeper for higher values
of the fresh call rate, indicating a higher sensitivity in busy call centers.

The graph in Figure 6 explores the sensitivity of the retrial rate to the individual retrial rate parameter $\delta$. 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 $\delta$ the shorter will be the length of the orbit. However at the same time high $\delta$ 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 first-attempt arrival rate, 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 $\delta$.

We next look at some capacity related parameters and their impact on the retrial rate. Since our model captures retrials by blocked customers, the queue size is an important parameter to analyze. Figure 7 depicts the effect of increasing the waiting space $K_1 - C$ on the retrial rate, for a call center with $C = 100$ and $p = 0.8$. The graph demonstrates the significant level of retrials that can be experienced by a busy call center with a small blocking threshold.

Our final analysis is to understand the impact of the number of servers on the 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 $\rho$ is kept constant (equal to 0.9) for different values of $C$ by adjusting the arrival rate $\lambda$. The ratio $(K - C)/C$ will also be kept constant (equal to 0.05) in these examples. Figure 8 depicts the evolution of the ratio of the retrial rate to the fresh call rate, as a function of $C$. As expected, the ratio diminishes as a function of size approaching zero for 200 servers. Figure 9 illustrates the same analysis for a system where $\rho$ has been increased to 1.3. This time even for 200 servers, the ratio rests around 78%. It is apparent that even for a large system, if the system load exceeds one, the retrial phenomenon will have a significant impact.
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 first-time-acceptance rate. The latter determines the proportion of clients who are not blocked and who do not abandon. We compare the optimum number of servers obtained for the model with retrials and compare it to the optimization of an M/M/C/K+M queue without retrials. 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 first-time-acceptance rate can be written as

\[
\text{first-time-acceptance (without retrials)} = 1 - \left( \frac{\lambda \Pi_K + \theta Q_W}{\lambda} \right),
\]

where \(\Pi_K\) denotes the stationary probability of finding \(K\) clients in the system, and \(Q_W\) is the average queue length.

For this system without retrials, calculating the stationary probabilities is straightforward. The system is a birth-death process, and has been analyzed by Akşin and Harker (2000) and Garnett et al. (2002). Using the stationary probabilities and a given number of servers, the first-time-acceptance rate can be obtained using Equation (2), where \(\Pi_K\) and \(Q_W\) take the corresponding values in an M/M/C/K+M queue. An inverse recursive procedure can then be used to determine the number of servers necessary to obtain a desired service level, expressed in terms of the first-time acceptance rate.

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 a recursive numerical procedure detailed in Appendix B. Table 1 demonstrates an example where the observed call volume 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 1: $\lambda$ as a function of $C$ for a system where $K = C + 5$, $\lambda_o = 15$, $\mu = 0.3$, $\theta = 0.3$, $\delta = 1$, $p = 0.8$

<table>
<thead>
<tr>
<th>$C$</th>
<th>25</th>
<th>30</th>
<th>35</th>
<th>40</th>
<th>45</th>
<th>50</th>
<th>55</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\lambda$</td>
<td>9.79</td>
<td>10.81</td>
<td>11.78</td>
<td>12.69</td>
<td>13.48</td>
<td>14.15</td>
<td>14.62</td>
</tr>
</tbody>
</table>

Once the fresh call rate $\lambda$ is thus obtained, the first-time-acceptance rate can be determined using the expression:

First-time-acceptance (with retrials) = $1 - \left( \frac{\lambda_{obs} \sum_{n=0}^{K_2} \Pi_{K_1,n} + \theta Q_W}{\lambda_{obs}} \right)$.

In this equation, the probabilities $\Pi_{K_1,n}$, $n = 0, 1, \ldots, K_2$ are those corresponding to the Markov Chain in Figure 2, and require knowledge of the fresh call rate $\lambda$. This time $Q_W$ represents the average queue length for the model with retrials. It can be calculated using the relationship:

$Q_w = \sum_{i=C+1}^{K_1} \left( (i - C) \sum_{n=0}^{K_2} \Pi_{i,n} \right)$.

Figure 10 illustrates the impact of ignoring retrials in dimensioning a call center. In this example, the optimal number of servers that ensure a desired first-time-acceptance rate are shown for three different systems. For all three systems $\lambda_o$ 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 Equation (2) 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 1. This time using the formula in (3) in conjunction with an
inversion algorithm the dimensioning of the call center is performed. 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.

We observe in Figure 10 that by ignoring the presence of retrials, one can under-size or over-size the system. If we consider for example the case of a first-time-acceptance rate of 90%, the optimal system size can go from 48 servers if retrials are ignored to 34 or 51 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 λ₀. 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 λ₀ will be very close to each other (since λ₀/Cµ < 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 first-time-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.

To complete our study of the interaction between retrials and staffing in call centers, we next focus on a staffing method, known as the square-root staffing rule, that has been proposed as a simple and robust tool to dimension call centers. For a detailed overview of the origins of this approach, as well as its technical development, the reader is referred to Gans et al. (2002) and references therein. Our objective is to compare the optimal size that we obtain in the system with retrials to the one that would be suggested if instead a square root staffing rule were being used. This rule determines the appropriate size for a call center given an offered load λ/µ and a service level objective β. It is particularly appropriate to study the impact of a change in λ on the required number of servers, when the service level is kept constant. For certain queueing systems, this approach can be used to determine the optimal number of servers required to satisfy a certain service
level like a desired waiting probability for example, by using explicit characterizations of \( \beta \) (see for example Halfin and Whitt, 1981; and Garnett et al., 2002).

In this paper, we are going to restrict our attention to the M/M/C+M system studied in Garnett et al. (2002). Note that this system takes abandonments into account, but does not consider retrials. For this system, the optimal \( \beta \) has not been characterized. We will make use of the following relationship between \( \beta \) and the number of servers \( C \):

\[
\beta = \left( C^* - \frac{\lambda}{\mu} \right) \sqrt{\frac{\mu}{\lambda}}. \tag{4}
\]

Then for a given \( \beta \) the number of servers as a function of the arrival rate \( \lambda \) is given by

\[
C = \frac{\lambda}{\mu} + \beta \sqrt{\frac{\lambda}{\mu}}. \tag{5}
\]

We next look at an example where we compare the model with retrials using Equation (3) to the Square Root Staffing method. All parameters not mentioned take the same values as earlier in this Section. Suppose that the fresh call arrival rate is \( \lambda = 30 \) calls per minute with a first-time-acceptance rate of 80%. For these values we obtain the optimal number of servers as \( C^* = 95 \). Now using Equation (4), the service level can be quantified as \( \beta = -0.5 \). Using this in Equation (5) yields the number of servers for the square root staffing method. In Table 2 we compare the number of servers obtained using the two approaches for different values of \( \lambda \). We note that the results are relatively close to each other, with the difference increasing as the call volume increases. The square-root staffing method, that disregards retrials, always over-dimensions the system.

<table>
<thead>
<tr>
<th>( \lambda )</th>
<th>40</th>
<th>50</th>
<th>60</th>
<th>70</th>
<th>80</th>
<th>90</th>
<th>100</th>
</tr>
</thead>
<tbody>
<tr>
<td>( C ) (square root)</td>
<td>128</td>
<td>160</td>
<td>193</td>
<td>226</td>
<td>259</td>
<td>291</td>
<td>324</td>
</tr>
<tr>
<td>( C^* ) (with retrials)</td>
<td>127</td>
<td>159</td>
<td>191</td>
<td>222</td>
<td>254</td>
<td>286</td>
<td>317</td>
</tr>
</tbody>
</table>

The previous example, due to the high level of first-time-acceptance rate chosen, reflects a system that would not experience too many retrials. In order to perform the comparison in a setting where retrials would be more important, we look at the same
example with first-time-acceptance rate equal to 60% for \( \lambda = 30 \). This time \( \beta = -1.5 \) for \( C^* \) obtained as 85. A similar comparison to the one above yields the numbers in Table 3 below, illustrating a wider gap between the two methods. Both of these examples

<table>
<thead>
<tr>
<th>( \lambda )</th>
<th>40</th>
<th>50</th>
<th>60</th>
<th>70</th>
<th>80</th>
<th>90</th>
<th>100</th>
</tr>
</thead>
<tbody>
<tr>
<td>( C ) (square root)</td>
<td>116</td>
<td>147</td>
<td>179</td>
<td>210</td>
<td>242</td>
<td>274</td>
<td>306</td>
</tr>
<tr>
<td>( C^* ) (with retrials)</td>
<td>115</td>
<td>144</td>
<td>173</td>
<td>203</td>
<td>232</td>
<td>261</td>
<td>291</td>
</tr>
</tbody>
</table>

assumed that the fresh call arrival rate \( \lambda \) is known. However if we consider the same call center as before, where the database contains data on the observed call rate only, a manager using the square root staffing method that does not consider retrials may not be aware of the fresh call rate \( \lambda \). Our final example compares the staffing with retrials to the square root method where \( \lambda_o \) is used instead of \( \lambda \). Consider the example with first-time-acceptance rate equal to 80%. Recall that for \( \lambda = 30 \) the number of servers necessary is \( C^* = 95 \). This allows one to compute \( \lambda_o = 35.93 \). Now using this arrival rate in Equation (4) we obtain the service level \( \beta = -2.26 \). Replacing \( \lambda \) by \( \lambda_o \) in Equation (5) the corresponding number of servers for the square root staffing method can be obtained. Table 4 tabulates the comparison between the two methods. The gap between the two methods is much larger in this case, and increasing as a function of \( \lambda \).

<table>
<thead>
<tr>
<th>( \lambda_o )</th>
<th>40</th>
<th>50</th>
<th>60</th>
<th>70</th>
<th>80</th>
<th>90</th>
<th>100</th>
</tr>
</thead>
<tbody>
<tr>
<td>( C ) (square root)</td>
<td>107</td>
<td>137</td>
<td>168</td>
<td>199</td>
<td>230</td>
<td>261</td>
<td>292</td>
</tr>
<tr>
<td>( C^* ) (with retrials)</td>
<td>99</td>
<td>107</td>
<td>114</td>
<td>120</td>
<td>127</td>
<td>133</td>
<td>140</td>
</tr>
</tbody>
</table>
6 Concluding Remarks

In this paper, a model that explicitly considers retrials generated by blocked calls in a call center with abandonments is formulated. An algorithm is proposed to resolve the resulting two-dimensional Markov Chain. Using this algorithm in numerical examples that reflect the operations of a real call center, the magnitude that retrials can attain in similar call centers is demonstrated. It is shown that retrials can have an important effect on call center performance. A numerical sensitivity analysis establishes call center characteristics that can lead to similar performance effects. By comparing the model to others where the retrial phenomenon is ignored, the paper establishes that call center dimensioning can be highly distorted in the latter case. The inability to monitor fresh call arrival rates further exacerbates this problem. The qualitative nature of the distortion in terms of an under or over-sizing is affected by the staffing in earlier periods from which historical data is obtained. This last result points to the need of considering the interaction between retrials and sizing in a multi-period setting.

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.

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.

Acknowledgments: 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.

References


**Appendix A: The algorithm for calculation of the stationary distribution**

Determination of the non-normalized stationary probabilities for the top row of the orbit ($K_2$ customers in the orbit)

\[ \Pi_{0,K_2} := 1 \]

\[ \Pi_{1,K_2} := \frac{1}{\mu} (\lambda + K_2 \delta) \quad \Pi_{0,K_2} = \frac{1}{\mu} (\lambda + K_2 \delta) \]

For $m := 2$ to $C$

\[ \Pi_{m,K_2} := \frac{1}{m \mu} ((\lambda + K_2 \delta + (m - 1) \mu) \quad \Pi_{m-1,K_2} - \lambda \Pi_{m-2,K_2}) \]

End for
For $m := C + 1$ to $K_1$

$$
\Pi_{m,K_2} := \frac{1}{C\mu + (m-C)\theta} \left( (\lambda + K_2\delta + C\mu + (m - C - 1)\theta) \Pi_{m-1,K_2} - \lambda \Pi_{m-2,K_2} \right)
$$

End for

**Determination of the non-normalized stationary probabilities corresponding to customers in the orbit:**

For $n := K_2 - 1$ down to 0

The stationary probability of state $K_1,n$ is given by the following expression:

$$
a := \begin{cases} 
(\frac{C\mu+(K_1-C)\theta+K_2(1-p)\delta}{p\lambda}) \Pi_{K_1,K_2-\lambda} \Pi_{K_1-1,K_2} & \text{if } n = K_2 - 1 \\
(\frac{C\mu+(K_1-C)\theta+(n+1)(1-p)\delta+p\lambda}{p\lambda}) \Pi_{K_1,n+1-\lambda} \Pi_{K_1-1,n+1-(n+2)(1-p)\delta} \Pi_{K_1-1,n+2} & \text{otherwise}
\end{cases}
$$

Note that the coefficient $a$ takes the following form:

$$
a =: c \Pi_{0,n} + b, \text{ where } b = f \text{ (coefficients from row } n+1 \text{ of the orbit)}
$$

$$
\Pi_{0,n} := 0 \text{ (in order to express the probabilities in row } n \text{ as a function of the probabilities of row } n+1)
$$

$$
\Pi_{1,n} := \frac{1}{\mu} (\lambda + n\delta) \; \Pi_{0,n} = 0
$$

for $m := 2$ to $C$

$$
\Pi_{m,n} := \frac{1}{m \mu} ((\lambda + n\delta + (m - 1)\mu) \; \Pi_{m-1,n} - \lambda \Pi_{m-2,n} - (n + 1)\delta \; \Pi_{m-2,n+1})
$$

end for

for $m := C + 1$ to $K_1$

$$
\Pi_{m,n} := \frac{(\lambda + n\delta + C\mu + (m - C - 1)\theta)}{C\mu + (m-C)\theta} \; \Pi_{m-1,n} - \lambda \Pi_{m-2,n} - (n + 1)\delta \; \Pi_{m-2,n+1}
$$

end for

$$
b := \Pi_{K_1,n}
$$

$$
\Pi_{0,n} := 1 \text{ (in order to momentarily eliminate the effect of row } n+1)
$$

$$
\Pi_{1,n} := \frac{1}{\mu} (\lambda + n\delta) \; \Pi_{0,n} = \frac{1}{\mu} (\lambda + n\delta)
$$

16
for $m := 2$ to $C$

$$\Pi_{m,n} := \frac{1}{m \mu} (\lambda + n\delta + (m-1)\mu) \Pi_{m-1,n} - \lambda \Pi_{m-2,n}$$

end for

for $m := C+1$ to $K_1$

$$\Pi_{m,n} := \frac{(\lambda + n\delta + C\mu + (m-C-1)\theta) \Pi_{m-1,n} - \lambda \Pi_{m-2,n}}{C\mu + (m-C)\theta}$$

end for

$$c := \Pi_{K_1,n}$$

$$\Pi_{0,n} := \frac{a-b}{c}$$

$$\Pi_{1,n} := \frac{1}{\mu} (\lambda + n\delta) \Pi_{0,n}$$

for $m := 2$ to $C$

$$\Pi_{m,n} := \frac{1}{m \mu} (\lambda + n\delta + (m-1)\mu) \Pi_{m-1,n} - \lambda \Pi_{m-2,n} - (n+1)\delta \Pi_{m-2,n+1}$$

end for

for $m := C+1$ to $K_1$

$$\Pi_{m,n} := \frac{(\lambda + n\delta + C\mu + (m-C-1)\theta) \Pi_{m-1,n} - \lambda \Pi_{m-2,n} - (n+1)\delta \Pi_{m-2,n+1}}{C\mu + (m-C)\theta}$$

end for

End for

*calculation of the normalization constant $\Sigma$*

$$\Sigma := \sum_{0 \leq i \leq K_1, 0 \leq j \leq K_2} \Pi_{i,j}$$

*normalizing probabilities by the normalization constant $\Sigma$*

For $m := 0$ to $K_1$
for $n := 0$ to $K_2$

$\Pi_{m,n} := \frac{\Pi_{m,n}}{\Sigma}$

end for

End for

**Appendix B: The algorithm for calculation of the fresh call rate from the observed call rate**

Let $f(\lambda, \mu, c, K_1, K_2, \theta, \delta, p)$ denote the function that calculates the observed call rate $\lambda_{obs}$ when the other parameters are given. $\lambda_0$ denotes the desired observed call rate and $\epsilon$ the desired precision:

*initialization*

$x_{low} := 0$

$x_{high} := \lambda_0$

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
Figure 1: A call center with blocking, abandonments, and retrials
Figure 2: The system as a CTMC

Figure 3: Evolution of the retrial rate as a function of the fresh call rate for a system with 100 servers
Figure 4: Evolution of the retrial rate as a function of the fresh call rate for a system with 25 servers

Figure 5: The effect of the probability of retrying on the retrial rate
Figure 6: The effect of $\delta$ on the retrial rate
Figure 7: The Effect of a finite waiting space on the retrial rate

Figure 8: The Effect of the system size $C$ for $\rho = 0.9$ and $K - C = 5\%$ of $C$
Figure 9: The effect of the system size $C$ for $\rho = 1.3$ and $K - C = 5\%$ of $C$

Figure 10: The effect of ignoring retrials in dimensioning a call center