Cybernetics and Systems Analysis, Vol. 46, No. 3, 2010

INVESTIGATION OF CALL CENTERS AS RETRIAL QUEUING SYSTEMS[1]

S. V. Pustova

UDC 519.872

Abstract. *The paper analyzes the mathematical models of call centers that take into account repeated calls and the specific features and key aspects of modeling call centers as retrial queues. Queuing systems $M/M/c/0/L/H_j$, $M/M/c/0///E_2$, and $M/M/c/0/L//E_2$ are considered as models of call centers. Numerical results for the performance indices of these systems are given.*

Keywords: *call center, queuing system, repeated calls*.

Introduction. Development of innovation technologies for telecommunication systems gave rise to and promoted active introduction of such systems as call centers into almost all social and business spheres. In control theory and practice, call centers are often regarded as a synonym of the telephone service of customers, as a factor of success for a firm to be recognized by consumers. Call centers are a standard in servicing and are widely used in telecommunication and trading companies, credit-card banking, hotels, social and emergency services, etc. Call centers have become an irreplaceable communication facility and efficient client management tool [1].

The majority of organizations where communication with people prevails (these may be both private companies and official institutions) reorganize their infrastructure by introducing from one to several call centers. Such systems allow using the company's resources (employees, communication lines, equipment, software) in the best way to service client calls.

Recently, call centers have been renamed contact centers since telephone means are supplemented with the Internet, e-mail, fax, chat, database technology, etc. These factors considerably influence the operation of call centers, the information transmitted, and clients' solutions since the clients can address the call center more often and obtain more diversified information. Call centers are an important communication tool in the modern world; therefore, developing adequate mathematical models is one of the key tasks at the design stage.

The studies of functioning and administration of call centers are based on queuing theory [2]. Note that retrial queuing systems (QSs) are the most adequate method to model call centers [3]. Such QSs can take into account the primary, secondary, and other call flows arriving at the input of a call center and influencing its performance indices [4]. Most publications on modeling call centers disregard repeated calls (retrials). The studies that take retrials into account disregard their dependence on the flow of primary calls.

Modeling call centers as retrial systems is addressed in [5, 6], where the consequences of ignoring repeated calls in calculating the optimal number of operators of the call center are shown. The paper [5] models a call center as a multiserver QS where the following operations are explicitly modeled: queuing a client failure, client's impatience, and repeated calls. The resultant QS is analyzed for stationary and nonstationary modes. For the stationary mode, use is made of the fluid approximation, which facilitates the analysis of a Markov chain continuous in time and is used for the exact mapping of systems of large call centers with heavy traffic. The fluid approximation method is used to study the phenomenon of retrials for a real call center. The model is applied to estimate the call arrival rate based on the statistical data where retrials cannot be distinguished from primary calls. This is a common problem for all call centers. Numerical methods showed that ignoring the phenomenon of repeated calls at call centers may cause significant errors in the subsequent prediction and determination of the optimal number of agents.

[1]The study was supported by the Ministry of Education and Science of Ukraine (NDR of April 22, 2009, Contract No. M/202-2009).

National Aviation University, Kyiv, Ukraine, *p82004@ukr.net*. Translated from Kibernetika i Sistemnyi Analiz, No. 3, pp. 162–168, May–June 2010. Original article submitted November 24, 2009.

 1060-0396/10/4603-0494

The paper [6] models a call center as a Markov chain considering clients' impatience and repeated calls. It shows that ignoring retrials may result in the insufficient or redundant (depending on the prediction) staffing compared with the optimal one.

Call Center as an $M/M/c/0/L/H_j$ QS. In practice, the number of subscribers of a call center is always finite and not greater than the number of subscribers of the telephone network of the country. Therefore, the orbit capacity is finite too. Let us consider the model of a call center with the orbit capacity bounded by a given constant L. If the orbit capacity is L, the calls arriving at the system are lost and do not influence the operation of the system.

Let an arrived call leave the system after several failed attempts. Let also H_j be the probability that the jth failed attempt will be followed by the $(j+1)$th attempt. Assume that the probability of repeated calls after a failed attempt does not depend on the number of previous attempts (i.e., $H_2 = H_3 = \ldots$).

Let a Poisson flow of primary calls arrive at c service channels (servers) with the rate λ. If any of the c channels is free at the instant of arrival of the primary call, the call is served and leaves the system. Otherwise, the call leaves the system without service with probability $1-H_1$ and with probability $H_1 > 0$ arrives at the orbit if at least one of L places in the orbit is free or leaves the system (the call is lost) if all the places in the orbit are occupied. The service times are exponentially distributed with the parameter μ.

Calls in the Orbit are a Poisson Process with the Rate ν. If any service channel is free at the time of arrival of a repeated call, the latter leaves the system and disappears from the orbit after service. Otherwise, the call will leave the system with probability $1-H_2$ or will repeat an attempt to be served with probability H_2.

Setting Up an Analytical Model. The operation of an $M/M/c/0/L/H_j$ system as a model of a call center can be described by a two-dimensional process $(C(t), N(t))$, where $C(t)$ is the number of busy channels, $N(t)$ is the number of retrials in the orbit at the time t. The process $(C(t), N(t))$ is Markov and defined on the set of states $S^{(L)} = \{0, 1, \ldots, c\} \times \{0, 1, \ldots, L\}$. Its infinitesimal rates $q_{(ij)(nm)}$ of the transition from the state (i, j) to the state (n, m) are specified as follows:

for $0 \le i \le c-1,\ 0 \le j \le L$

$$q_{(ij)(nm)} = \begin{cases} \lambda & \text{if } (n,m) = (i+1, j), \\ i\mu & \text{if } (n,m) = (i-1, j), \\ j\nu & \text{if } (n,m) = (i+1, j-1), \\ -(\lambda + i\mu + j\nu) & \text{if } (n,m) = (i,j), \\ 0 & \text{otherwise;} \end{cases}$$

for $i = c,\ 0 \le j \le L-1$

$$q_{(cj)(nm)} = \begin{cases} \lambda H_1 & \text{if } (n,m) = (c, j+1), \\ i\nu(1-H_2) & \text{if } (n,m) = (c, j-1), \\ c\mu & \text{if } (n,m) = (c-1, j), \\ -(\lambda H_1 + j\nu(1-H_2) + c\mu) & \text{if } (n,m) = (c,j), \\ 0 & \text{otherwise;} \end{cases}$$

for $i = c,\ j = L$

$$q_{(cL)(nm)} = \begin{cases} c\mu & \text{if } (n,m) = (c-1, L), \\ L\nu(1-H_2) & \text{if } (n,m) = (c, L-1), \\ -(c\mu + L\nu(1-H_2)) & \text{if } (n,m) = (c,j), \\ 0 & \text{otherwise.} \end{cases}$$

Since the set of states of the process $(C(t), N(t))$ is finite, it is always ergodic [3]. Its stationary distribution $p_{ij} = P(C(t) = i, N(t) = j)$ (p_{ij} is the probability that the system is in the state (i, j)) can be found as the solution of the following system of equations:

$$(\lambda + i\mu + j\nu)p_{ij} = \lambda\, p_{i-1,j} + (j+1)\nu\, p_{i-1,j+1} + (i+1)\mu p_{i+1,j},\ 0 \le i < c,\ 0 \le j < L; \quad (1)$$

$$(\lambda + i\mu + L\nu)p_{iL} = \lambda p_{i-1,L} + (i+1)\mu p_{i+1,L},\ 0 \le i < c; \quad (2)$$

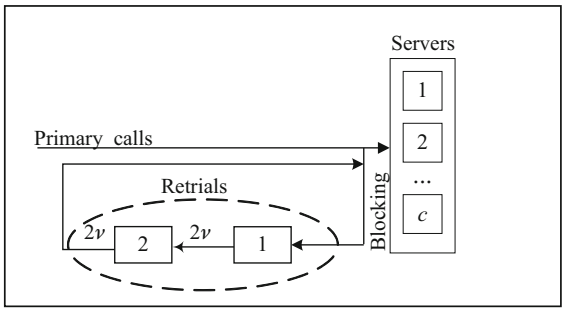

Fig. 1. Call center as an $M/M/c/0///E_2$ QS.

$$(\lambda H_1 + j\nu(1-H_2)+c\mu)p_{cj} = \lambda\, p_{c-1,\,j} + (j+1)\nu\, p_{c-1,\,j+1} + \lambda H_1\, p_{c,\,j-1} + (j+1)\nu(1-H_2)\, p_{c,\,j+1},\ 0 \le j < L,\ 0 \le j < L; \tag{3}$$

$$(c\cdot\mu + L\nu(1-H_2))p_{c,L} = \lambda\, p_{c-1,L} + \lambda H_1\, p_{c,L-1}. \tag{4}$$

These equations satisfy the normalization condition

$$\sum_{i=0}^{c}\sum_{j=0}^{L} p_{ij} = 1. \tag{5}$$

The most important performance indices of a call center are: (i) stationary probability of server occupation $B = \lim\limits_{t\to\infty} P\{C(t)=c\}$; (ii) average number of repeated calls $N = \lim\limits_{t\to\infty} \mathrm{E}N(t)$; (iii) average number of busy servers in the stationary mode $Y = \lim\limits_{t\to\infty} \mathrm{E}C(t)$; and (iv) the average waiting time in the orbit $W = \dfrac{N}{\lambda}$ (by Little's formula).

Using generating functions, we obtain the following formulas for the average number of calls in the orbit:

for $H_2 < 1$

$$N = \frac{\dfrac{\lambda}{\mu}H_2 + \dfrac{\lambda}{\mu}(H_1 - H_2)B - H_2 Y - \dfrac{\lambda}{\mu}H_1 p_{cL}}{\dfrac{\nu}{\mu}(1-H_2)};$$

for $H_2 = 1$

$$N = \frac{1+\dfrac{\nu}{\mu}}{\dfrac{\nu}{\mu}\left(c - \dfrac{\lambda}{\mu}H_1\right)}\left[\frac{\lambda}{\mu} + \left(\frac{\lambda}{\mu}\right)^2 - \mathrm{E}(C(t))^2 - \frac{\lambda}{\mu}(1-H_1)\left(\frac{\lambda}{\mu} + c + 1 - \frac{\dfrac{\lambda}{\mu}H_1}{1+\dfrac{\nu}{\mu}}\right)\cdot B\right.$$

$$\left. + \left(\frac{\lambda}{\mu} + c + 1 + \frac{L\dfrac{\nu}{\mu} + \dfrac{\lambda}{\mu}(1-H_1)}{1+\dfrac{\nu}{\mu}}\right)\frac{\lambda}{\mu}H_1 p_{cL}^{(L)}\right].$$

Model of a Call Center as an $M/M/c/0///E_2$ QS. Let us consider a call center as a multichannel system with a Poisson input call flow, exponentially distributed service time, without waiting places, with unlimited orbit, loss-free, and with two-phase Erlang distribution of the call flow in the orbit (Fig. 1).

Let a Poisson flow of primary calls arrive at c service channels with the rate λ (the density function $a(x) = \lambda e^{-\lambda x}$). If at least one of the servers is idle when a call arrives, the call immediately occupies it, is serviced, and then leaves the system. Otherwise the call becomes the source of retrials.

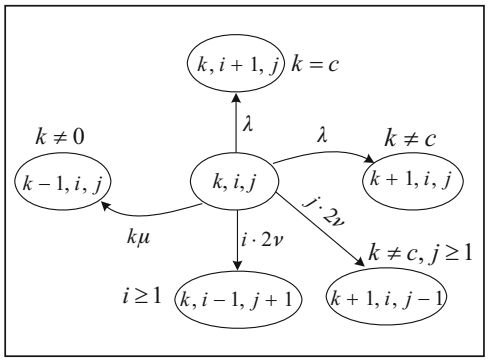

Fig. 2. Transition diagram for an $M/M/c/0//NL/E_2$ QS for transitions from the state (k, i, j).

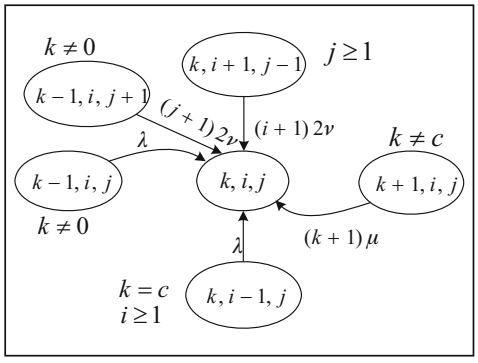

Fig. 3. Transition diagram for an $M/M/c/0//NL/E_2$ QS for transitions to the state (k, i, j).

Each such source creates a two-phase Erlang process of retrials with the parameter ν (the density function $d(x) = (2\nu)^2 x e^{-(2\nu)x}$). If any server is idle at the time of arrival of a retrial, the call is served and then leaves the system, and the source of retrials disappears. Otherwise the call will make a new attempt to be served.

The service times are assumed to be exponentially distributed with the parameter μ (the density function $b(x) = \mu e^{-\mu x}$).

Setting Up an Analytical Model. The operation of the system can be described by a three-dimensional process $(X(t), Y(t), Z(t))$, where $X(t)$ is the number of busy servers (in a single-channel system, the server is busy/idle), $Y(t)$ is the number of calls in the orbit at the first phase, $Z(t)$ is the number of calls in the orbit at the second phase at the instant of time t; the sum $Y(t) + Z(t)$ is the number of calls in the orbit at the instant of time t. The process $(X(t), Y(t), Z(t))$ is defined on the set of states $S = \{0, 1, \ldots, c\} \times \{0, 1, \ldots\} \times \{0, 1, \ldots\}$.

Let us set up the diagrams of transition states of an $M/M/c/0///E_2$ QS as a model of a call center (Figs. 2, 3). We will write the transition rates of the process $(X(t), Y(t), Z(t))$ in the time interval $(t, t+dt)$, $t \geq 0$. The system can pass from a state (k, i, j), $k = \overline{0, c}$, $i \geq 0$, $j \geq 0$, to another state in time dt with certain probability:

- λdt passes to a state $(k+1, i, j)$, $k \neq c$ (a new primary call has arrived and has obtained service at once);
- $j \cdot 2\nu dt$ passes to a state $(k+1, i, j-1)$, $k \neq c$, $j \geq 1$ (one of j repeated calls from the second phase has made a successful attempt to be served);
- $i \cdot 2\nu dt$ passes to a state $(k, i-1, j+1)$, $i \geq 1$ (one of i repeated calls from the first phase has passed to the second phase);
- $k\mu dt$ passes to a state $(k-1, i, j)$, $k \neq 0$ (the call service has been completed, one of the channels became idle);
- λdt passes to a state $(k, i+1, j)$, $k = c$ (a new primary call arrived, all the channels appeared busy, and it passed in the orbit to the first phase).

Then the transition rates $q_{(k,i,j)(g,n,m)}$, $k, g = \overline{1, c}$, $i, j, n, m = \{0, 1, \ldots\}$ (infinitesimal transition rates) of the process $(X(t), Y(t), Z(t))$ from the state (k, i, j) to the state (g, n, m) are given as follows

for $0 \leq k \leq c-1$

$$q_{(kij)(g,n,m)} = \begin{cases} \lambda & \text{if } (g,n,m) = (k+1, i, j), \\ j \cdot 2\nu & \text{if } (g,n,m) = (k+1, i, j-1), j \geq 1, \\ i \cdot 2\nu & \text{if } (g,n,m) = (k, i-1, j+1), i \geq 1, \\ k\mu & \text{if } (g,n,m) = (k-1, i, j), k \neq 0, \\ -(\lambda + j \cdot 2\nu + i \cdot 2\nu + k\mu) & \text{if } (g,n,m) = (k, i, j), \\ 0 \quad \text{otherwise;} \end{cases} \tag{6}$$

for $k = c$

$$q_{(cij)(g,n,m)} = \begin{cases} \lambda & \text{if } (g,n,m) = (c, i+1, j), \\ i \cdot 2\nu & \text{if } (g,n,m) = (c, i-1, j+1), i \geq 1, \\ c\mu & \text{if } (g,n,m) = (c-1, i, j), \\ -(\lambda + i \cdot 2\nu + c\mu) & \text{if } (g,n,m) = (c, i, j), \\ 0 \quad \text{otherwise.} \end{cases} \tag{7}$$

In the stationary mode, an $M/M/c/0///E_2$ QS satisfies the Kolmogorov system of equations (see Figs. 2, 3) ($p_{kij} = 0$ if $\forall k, \forall i, \forall j < 0$):

$$(\lambda + j \cdot 2\nu + i \cdot 2\nu + k\mu) p_{kij} = \lambda p_{k-1,i,j} + (k+1)\mu p_{k+1,i,j} + (i+1)2\nu p_{k,i+1,j-1}$$
$$+ (j+1)2\nu p_{k-1,i,j+1} + \lambda p_{k-1,i,j}, \quad 0 \le k \le c-1, \quad i \ge 0, \quad j \ge 0; \tag{8}$$

$$(\lambda + i \cdot 2\nu + c\mu) p_{cij} = \lambda p_{c-1,i,j} + (i+1)2\nu p_{c,i+1,j-1}$$
$$+ (j+1)2\nu p_{c-1,i,j+1} + \lambda p_{c,i-1,j}, \quad k = c, \quad i \ge 0, \quad j \ge 0, \tag{9}$$

and the normalization condition

$$\sum_{k=0}^{c} \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} p_{kij} = 1. \tag{10}$$

Model of a Call Center as an $M/M/c/0/L//E_2$ QS. Since it is rather difficult to obtain the analytical solution for a system $M/M/c/0///E_2$, we restrict the orbit capacity to a sufficiently large constant L (Wilkinson method). This model will adequately describe the operation of a call center since the orbit capacity is always a finite number.

Setting Up an Analytical Model. In view of the above constraints, formulas (6) and (7) become as follows:

for $0 \le k \le c-1, \ i + j \le L$

$$q_{(kij)(g,n,m)} = \begin{cases} \lambda & \text{if } (g,n,m) = (k+1, i, j), \\ j \cdot 2\nu & \text{if } (g,n,m) = (k+1, i, j-1), \ j \ge 1, \\ i \cdot 2\nu & \text{if } (g,n,m) = (k, i-1, j+1), \ i \ge 1, \ j < L, \\ k\mu & \text{if } (g,n,m) = (k-1, i, j), \ k \ne 0, \\ -(\lambda + j \cdot 2\nu + i \cdot 2\nu + k\mu) & \text{if } (g,n,m) = (k, i, j), \\ 0 & \text{otherwise;} \end{cases} \tag{11}$$

for $k = c, \ i + j \le L$

$$q_{(cij)(g,n,m)} = \begin{cases} \lambda & \text{if } (g,n,m) = (c, i+1, j), \ i \le L-1, i+j < L, \\ i \cdot 2\nu & \text{if } (g,n,m) = (c, i-1, j+1), \ i \ge 1, \ j < L, \\ c\mu & \text{if } (g,n,m) = (c-1, i, j), \\ -(L \cdot 2\nu + c\mu) & \text{if } (g,n,m) = (c, i, j), \ i = L, j = 0, \\ -(\lambda + c\mu) & \text{if } (g,n,m) = (c, i, j), \ i = 0, j = L, \\ -(\lambda + i \cdot 2\nu + c\mu) & \text{if } (g,n,m) = (c, i, j), \ i \le L-1, j \le L-1, \\ 0 & \text{otherwise.} \end{cases} \tag{12}$$

According to (11) and (12), we have the following system for formulas (8)–(10) ($p_{kij} = 0$ if $\forall k, \forall i, \forall j < 0$ or $\forall k, \forall i, \forall j > L$, or $i + j \le L$):

$$(\lambda + j \cdot 2\nu + i \cdot 2\nu + k\mu) p_{kij} = \lambda p_{k-1,i,j} + (k+1)\mu p_{k+1,i,j} + (i+1)2\nu p_{k,i+1,j-1}$$
$$+ (j+1)2\nu p_{k-1,i,j+1} + \lambda p_{k-1,i,j}, \quad 0 \le k \le c-1, \quad i \ge 0, \quad j \ge 0, \quad i + j \le L; \tag{13}$$

$$(\lambda + i \cdot 2\nu + c\mu) p_{cij} = \lambda p_{c-1,i,j} + (i+1)2\nu p_{c,i+1,j-1} + (j+1)2\nu p_{c-1,i,j+1} + \lambda p_{c,i-1,j}, \tag{14}$$
$$k = c, \ 0 \le i \le L-1, \ 0 \le j \le L-1, \ i + j \le L;$$

$$(L \cdot 2\nu + c\mu) p_{cL0} = \lambda p_{c-1,L,0} + 2\nu p_{c-1,L,1} + \lambda p_{c,L-1,0}, \quad k = c, \quad i = L, \quad j = 0; \tag{15}$$

$$(\lambda + c\mu) p_{c0L} = \lambda p_{c-1,0,L} + 2\nu p_{c,1,L-1}, \quad k = c, \quad i = 0, \quad j = L, \tag{16}$$

and the normalization condition

$$\sum_{k=0}^{c} \sum_{i=0}^{L} \sum_{j=0}^{L} p_{kij} = 1, \ i + j \le L. \tag{17}$$

TABLE 1

λ	L	Results	
		B Markovian	B Erlang2
0.5	1	0.4000	0.4044
	2	0.4400	0.4458
	3	0.4643	0.4697
	4	0.4790	0.4833
	5	0.4878	0.4908
	50	0.4999	0.5
0.9	1	0.5473	0.5534
	2	0.6018	0.6137
	3	0.6438	0.6604
	4	0.6771	0.697
	5	0.7040	0.7262
	50	0.8949	0.8978
2.0	1	0.7096	0.707
	2	0.7434	0.7396
	3	0.7705	0.7662
	4	0.7927	0.7882
	5	0.7112	0.8064
	50	0.9637	0.9557
	55	0.9667	0.9588
	70	0.9733	0.9659

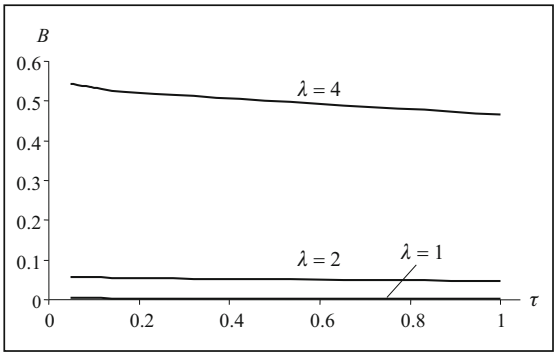

Fig. 4. Dependence of B on τ and λ for $c = 5$, $L = 50$, $\mu = 1$.

Since the orbit is limited by the constant L, the system of equations (13)–(17) is finite and the QS is ergodic for any conditions.

Numerical Solution of the System. System (13)–(17) can be solved on a computer by standard procedures. However, the memory size required for the matrix of the system is significant though the majority of matrix elements are zero. Therefore, system (13)–(17) was solved with the use of sparse matrix technology in the Matlab environment.

Some Numerical Results. Table 1 summarizes the dependence of the probability $B = \lim\limits_{t\to\infty} P\{X(t) = c\}$ of channel occupation on the orbit capacity L for $M/M/c/0/L//M$ (B Markov) and $M/M/c/0/L//E_2$ (B Erlang2) systems for $\nu = 0{,}5$, $\mu = 1$, and $c = 1$. Note that as L increases, the difference between the probabilities of channel occupation decreases and is a small quantity.

Figure 4 plots the dependence of the probability B of channel occupation on $\tau = 1/\mu$ and λ. As is seen, retrials influence the probability of channel occupation: the more the interval between retrials, the less the probability.

Conclusions. The paper considered various models of call centers that take into account repeated calls such as $M/M/c/0/L/H_j$, $M/M/c/0///E_2$, and $M/M/c/0/L//E_2$. It is these models that can adequately describe the operation of call centers since they take into account secondary, tertiary, etc. call flows arriving at the system. The numerical results showed that retrials influence the performance indices of call centers.

REFERENCES

1. N. Gans, G. Koole, and A. Mandelbaum, "Telephone call centers: tutorial, review, and research prospects," Manufacturing and Service Operations Management (M&SOM), **5**, No. 2, 79–141 (2003).
2. A. Mandelbaum, Call Centers (Centres): Research Bibliography with Abstracts, Version 7 (2006), http://iew3.technion.ac.il/serveng/References/US7_CC_avi.pdf.
3. G. I. Falin and J. G. C. Templeton, Retrial Queues, Chapmen & Hall, London (1997).
4. S. V. Pustova, "Dependence of performance indices of a call center on the distribution calls' sojourn time in the orbit," Cybern. Syst. Analysis, **45**, No. 2, 314–325 (2009).
5. M. S. Aguir, F. Karaesmen, Z. Aksin, and F. Chauvet, "The impact of retrials on call center performance," Oper. Res., **26**, 353–376 (2004).
6. M. S. Aguir, O. Z. Aksin, F. Karaesmen, and Y. Dallery, "On the interaction between retrials and sizing of call centers," Europ. J. Oper. Res., **191**, No. 2, 398–408 (2008).