Industrial Management & Data Systems

Solving two-dimensional Markov chain model for call centers

Chih-Chin Liang Hsing Luh

Article information:

Users who downloaded this article also downloaded:

For Authors

About Emerald www.emeraldinsight.com

Solving two-dimensional Markov chain model for call centers

Chih-Chin Liang
Graduate Institute of Business and Management, National Formosa University, Hu wei, Taiwan, ROC, and
Hsing Luh
Department of Applied Mathematics, National Chenchi University, Taipei, Taiwan, ROC

Received 4 December 2014
Revised 1 February 2015
18 March 2015
Accepted 23 March 2015

Abstract

Purpose – The purpose of this paper is to develop a novel model of a call center that must treat calls with distinctly different service depending on whether they orginate from VIP or regular customers. VIP calls must be responded to immediately but regular calls can be routed to a retrial queue if the operators are busy.

Design/methodology/approach – This study's proposed model can easily reveal the optimal arrangement of operators while minimizing computational time and without losing any precision of the performance measure when dealing with a call center with more operators.

Findings – Based on the results of the comparison between the exact method and the proposed approximation method, the approach shows that the larger the number of operators or inbound calls, the smaller the error between the two methods.

Originality/value – This investigation presents a computational method and management cost function intended to identify the optimal number of operators for a call center. Because of computational limitations, many operators could not be easily analyzed using the exact method. For the manager of a call center, the sooner the optimal solution is found, the faster business strategies are deployed. This study develops an approximation method and compares it with the exact method.

Keywords Call center, Management cost, Operators management, Two-dimensional Markov chain

Paper type Research paper

1. Introduction

Motivated by a research project involving call centers at a selected company in Taiwan (Kim *et al.*, 2012; Liang *et al.*, 2005, 2009), this study discusses an approximation method suitable for use when an exact solution is not attainable to calculation of the management cost of a call center that involves blocking probability and waiting time (Huang, 2010; Liang and Luh, 2013; Melikov and Babayev, 2006; Xu *et al.*, 2002; Bright and Taylor,1995). A large-scale service sector regards uninterrupted customer service as a key operational target (Kim *et al.*, 2012; Liang *et al.*, 2005, 2009). To meet customer demands, a service company must manage human resources in a call center to answer inbound calls using a computer system. Human resource management is a strategic approach to managing employees who individually and collectively contribute to the achievement of business objectives (Armstrong and Taylor, 2014; Yang *et al.*, 2007). Moreover, a call center charged with satisfying customer inquiries assigns staff members as phone operators (Kim *et al.*, 2012). Additionally, operators should make all customer arrangements during

The authors would like to thank the Ministry of Science and Technology of the Republic of China, Taiwan, for financially supporting this research under Contract No. 103-2410-H-150-004. Additionally, the authors many thanks to Mr Han-Sheng Huang for his work about literature review and data collection.

0263-5577
DOI 10.1108/IMDS-12-2014-0363

customer phone calls (Dudina *et al.*, 2013; Gans *et al.*, 2003). The number of operators represent the capacity of handing inbound calls. Managers can implement actions to reduce management expenses (Armstrong and Taylor, 2014, Benaroch *et al.*, 2010; Li and Chandra, 2007; Lee and Park, 2005). That is, through analyzing the management costs associated with specific solutions, managers can determine the optimal operator assignment strategy (Benaroch *et al.*, 2010; Dubelaar *et al.*, 2005; Lee and Park, 2005). Analysis of the management costs associated with daily call center operations can assist managers in making operator allocation decisions. Operators thus can be assigned efficiently based on analysis of the management costs associated with call center operations.

Obviously, the main daily job of the operator is to answer customer calls. However, not all customers benefit the company equally. On average, the revenues contributed by customers of telecommunication companies in Taiwan range from $7.9 US dollars/month (regular users) to $60 US dollars/month (important users) (National Communication Commission, 2014). This statistic implies that VIP customers can benefit a company far more than regular customers do (Kim *et al.*, 2012; Liang *et al.*, 2005, 2009). Therefore based on financial considerations companies should adopt business strategies by which they respond to VIP customers faster than regular ones. Obviously, the best solution for a company is a system that prioritizes VIP calls (denoted as v-calls) over regular calls (denoted as r-calls). However, given limitations of human resources, how to assign operators to efficiently serve customers of different levels of importance becomes crucial to developing a successful call-center service.

To optimize the assignment of operators in a call center, the guard channel scheme, which suggests the reservation of partial channels for packets with high privilege, could be adopted to help assign operators for v-calls (Choi and Chang, 1999; Do, 2010; Dudina *et al.*, 2013; Gans *et al.*, 2003; Servi, 2002; Winkler, 2013). This investigation considers a novel model for call center operations that is based on a guard channel scheme. Based on the proposed design, v-call are assigned high privilege, while still providing regular customers with satisfactory service. Restated, some operators should be dedicated exclusively to answering v-calls, while regular customers are guaranteed to receive services from the call center. The risk of this approach is that if the regular customer is angry because of the long waiting time, they may spread negative feedback (Kim *et al.*, 2012; Liang *et al.*, 2005, 2009). Therefore, the proposed method should consider waiting time. This study models two types of calls as a two-dimensional Markov chain (Phung-Duc *et al.*, 2010; Tran-Gia and Mandjes, 1997). This investigation applies the V-model to call center behavior (Dudina *et al.*, 2013; Gans *et al.*, 2003). Specifically, two types of customers can enter the call center and be served. Our objective is to derive performance measures of the investigated model, as well as upper and lower bounds of the waiting time of regular customers.

This study presents a model of two call types, v-calls and r-calls, as a two-dimensional Markov chain. Based on analysis of the two-dimensional Markov chain, we can obtain three measures of model performance: the probability that the system fails to serve a regular customer on the first attempt (P_r), the probability that the system fails to serve a VIP on first attempt (P_v), and the average number of regular customers in the retrial group (L_q). Additionally, the upper and lower bounds of the waiting time for regular customers could be found by analysis of the two-dinensional Markov chain.

To identify the optimal assignment among K operators, this investigation constructs a management cost function per unit of time using a queueing model and sets the optimal threshold N. For example, if the number of occupied operators who are too busy to answer customer inquiries is N or larger, the remaining K-N operators are reserved for serving VIPs. On the other hand, if the incoming r-call which finds no operators available on its arrival, the call sent back to the retrial group to wait for the next available operators. Based on the above descriptions, the management cost ($MC(N, K)$) is expressed as follows, where C_w denotes the holding cost of calls in the retrial group, while C_r and C_v represent the opportunity costs of blocking calls of regular customers and VIPs:

$$MC(N, K) = C_w L_q + C_r P_r + C_v P_v \tag{1}$$

From Equation (1), the results could be obtained by computing a two-dimensional Markov chain as explained below. However, applying the proposed model to a real case, it would be difficult to identify the optimal N because of the long computing time. Because the number of operators is generally large, with more than 100 operators in a large-scaled call center, the above complex computation necessitates the expenditure of considerable effort to solve stationary probabilities (Choi and Chang, 1999; Do, 2010; Dudina *et al.*, 2013; Gans *et al.*, 2003; Servi, 2002; Winkler, 2013). Since computers necessarily have limited computational ability, processes might be delayed if performed using a low-end computer. Additionally, such limited computational ability wastes excessive energy together with the excessive waiting time. On the other hand, a high-end computational machine is needed to solve this problem for the large number of operators, if no appropriate solution can be provided. In order to resolve this problem, the investigation provides an approximation method known as the phase merging algorithm, introduced by Melikov and Babayev (2006). Instead of the exact method, the algorithm is rephrased in the form applicable to a two-dimensional Markov chain and applied to the model to approximated stationary probability distributions and measures of their performance.

This study also examines estimated errors between the approximate and precise methods. The numerical results demonstrate that the approximation method enables the manager to save computational time without sacrificing accuracy given numerous operators. Using the approximation method, it is easy to accurately estimate management costs and effectively allocate operators in a large call center.

The remainder of this paper is structured as follows. In Section 2, a quasi-birth-and-death (QBD) process model is designed based on a two-dimensional Markov chain and the estimation of waiting time is further calculated. Section 3 illustrates the computation of stationary probability distribution. Section 4 then describes the approximation method, which is applied to a case to determine the optimum allocation of operators that minimizes management cost. Section 5 discusses the error estimation between the exact and approximation methods. Finally, Section 6 illustrates the conclusions and suggests directions for future research.

2. A queueing model

To estimate the management cost, the queueing model, which describes a call center that deals with calls from two categories of users, is derived to compute the opportunity of incoming calls are the r-calls from regular customers, and the v-calls from VIPs. A total of K operators is considered in a call center, where K is a large finite number. Namely, a maximum of K operators can serve customers. Let λ_γ and λ_v denote the Poisson arrival rates of r-calls and v-calls, respectively. The service rates for both call types by each operator are μ, where the service time is exponentially distributed. If at least one free operator is available on the arrival of a v-call, that call is admitted for immediate service; otherwise, v-call are immediately terminated. Let P_v represent the probability that v-calls are not served immediately.

Figure 1 shows that r-calls can legitimately connect with the call center when the number of busy operators is less than N, N is a nonnegative integer and has a value of K or less. Otherwise, the incoming r-calls are blocked or routed to the retrial group. The r-calls in the retrial group try to re-call following a period that follows an exponential distribution with rate α. Since all waiting calls in the retrial group are stored on a computer system with limited storage capacity, this study assumes that the retrial group has finite capacity, say R. Let P_r represent the probability that r-calls do not receive immediate service. Thus, P_r denotes the probability that r-calls do not receive service when the retrial group is full. Generally, inbound r-calls may leave the system immediately after it learns that no available operators exist for service even if the retrial group is not full. This study thus assumes that r-calls either enter the retrial group with probability H_0 or leave the system forever with probability 1-H_0. The re-call rate of r-calls in the retrial group is α. Regarding the r-calls in the retrial group, following an unsuccessful re-call they either return to the retrial group with probability H_1 or leave the system with probability 1-H_1.

Let $X_1(t)$ represent the random variable representing the number of r-calls in the retrial group and let $X_2(t)$ denote the random variable representing the number of calls in service at time t. Taking a long-term view, let S represent the state space, where $S = \{(X_1, X_2), 0 \leqslant X_1 \leqslant R, 0 \leqslant X_2 \leqslant K\}$. Then (X_1, X_2) denotes a two-dimensional Markov chain whose elements in Q-matrix are represented by:

$$q\left((i,j),\left(i',j'\right)\right) = \begin{cases} \lambda_r+\lambda_v, & \text{if } i' = i,\ j' = j+1,\ 0 \leqslant i \leqslant R,\ 0 \leqslant j \leqslant N-1 \\ \lambda_v, & \text{if } i' = i,\ j' = j+1,\ 0 \leqslant i \leqslant R,\ N \leqslant j \leqslant K \\ i\alpha, & \text{if } i' = i-1,\ j' = j+1,\ 1 \leqslant i \leqslant R,\ 0 \leqslant j \leqslant N-1 \\ i\alpha(1-H_1), & \text{if } i' = i-1,\ j' = j,\ 1 \leqslant i \leqslant R,\ N \leqslant j \leqslant K \\ i\alpha H_1, & \text{if } i' = i,\ j' = j,\ 1 \leqslant i \leqslant R,\ N \leqslant j \leqslant K \\ \lambda_r H_0, & \text{if } i' = i+1,\ j' = j+1,\ 0 \leqslant i \leqslant R,\ N \leqslant j \leqslant K \\ j\mu, & \text{if } i' = i,\ j' = j-1,\ 0 \leqslant i \leqslant R,\ 1 \leqslant j \leqslant K \\ 0, & \text{otherwise} \end{cases}$$

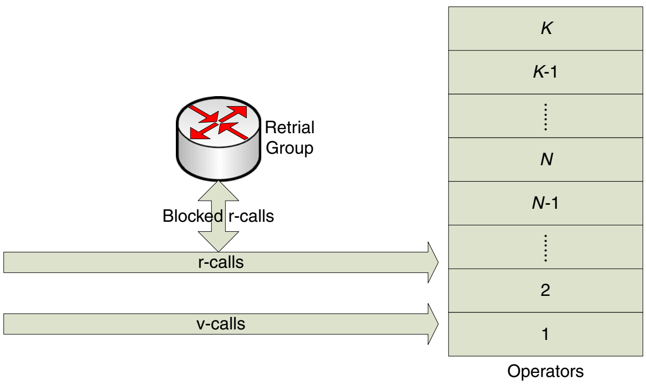


Figure 1.
A queueing system

Figure 2 illustrates the diagram of the state transition rate for call center operations. The parameters of the proposed model are: K, the numbers of operators in the call center; R, the capacity of the retrial group; N, the decision variable of number of operators; λ_r, the arrival rate of r-calls following a Poisson distribution; λ_v: the arrival rate of v-calls following a Poisson distribution; H_0, the probability that a r-call enters the retrial group; α, the rate of associated with a re-call time in the retrial group following an exponential distribution; H_1, the probability of r-calls that re-enter the retrial group after an unsuccessful retry; μ, the service rate of an operator; $\pi(i, j)$, stationary probability at state (i, j), where i is the random variable representing the number of r-calls in the retrial group, j is the random variable representing the number of calls in service from VIP and regular customers; L_q, the average number of calls in retrial group; and λ, the total arrival rate of VIP and regular customers.

The two-dimensional Markov chain resembles the birth and death process model (QBD) (Artalejo and Gómez-Corral, 2008; Artalejo *et al.*, 2005; Gómez-Corral, 2006; Do *et al.*, 2014; Phung-Duc, 2014; Phung-Duc *et al.*, 2013). This investigation establishes the infinitesimal generator (Q-matrix) as follows. Let d_k^i represent the diagonal element on the kth row in the B_i, where i is larger than or equal to zero and smaller than or equal to R, while k is larger than or equal to one and smaller than or equal to K:

$$
\mathrm{Q} = \begin{pmatrix} B_0 & A_0 & 0 & 0 & \cdots & 0 \\ A_2 & B_1 & A_0 & 0 & \cdots & 0 \\ 0 & A_2 & B_2 & A_0 & \ddots & \cdots \\ 0 & 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \vdots & 0 & A_2 & B_{R-1} & A_0 \\ 0 & 0 & 0 & 0 & A_2 & B_R \end{pmatrix}
$$

Let B_i, A_0 and A_2 defined as follows:

$$
\mathrm{B}_i = \begin{array}{c} 0 \\ 1 \\ \vdots \\ N-1 \\ N \\ N+1 \\ \vdots \\ K-1 \\ K \end{array}
\begin{bmatrix}
-d_1^{(i)} & \lambda_v & 0 & 0 & 0 & 0 & \cdots & 0 & 0 \\
\mu & -d_2^{(i)} & \lambda_v & 0 & 0 & 0 & \cdots & 0 & 0 \\
0 & \ddots & \ddots & \ddots & \ddots & \vdots & \vdots & \vdots & \vdots \\
\vdots & \ddots & (N-1)\mu & -d_{N-1}^{(i)} & i\alpha & 0 & \cdots & 0 & 0 \\
0 & 0 & 0 & N\mu & -d_N^{(i)} & \lambda_v & \cdots & 0 & 0 \\
0 & 0 & 0 & 0 & (N+1)\mu & -d_{N+1}^{(i)} & \lambda_v & 0 & 0 \\
\vdots & \vdots & \vdots & \vdots & 0 & \ddots & \ddots & \lambda_v & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & (K-1)\mu & -d_{K-1}^{(i)} & \lambda_v \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & K\mu & -d_K^{(i)}
\end{bmatrix};
$$

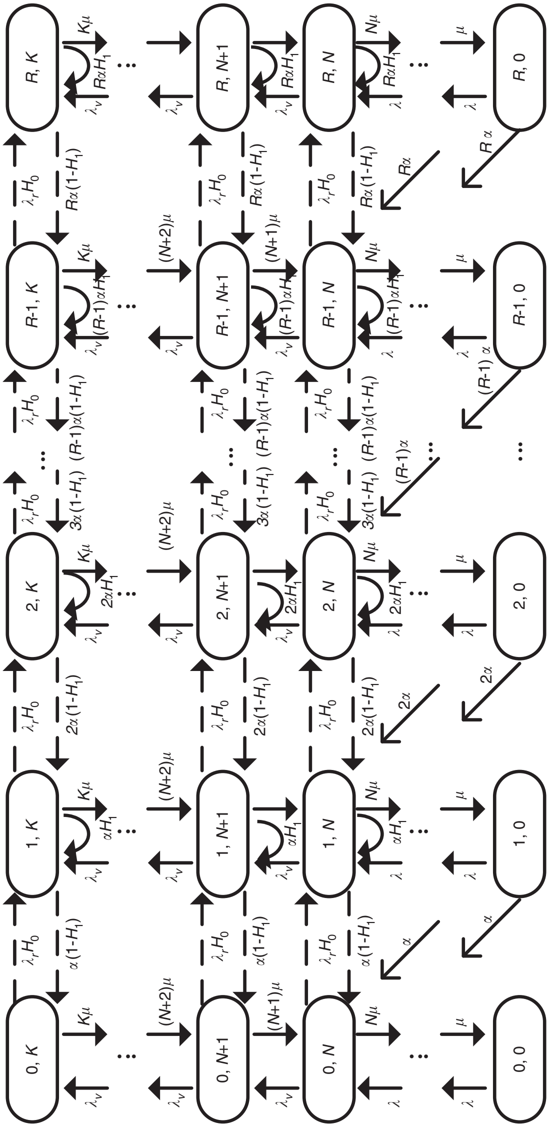


Figure 2.
State transition
diagram

$$A_0 = \begin{array}{c} 0 \\ 1 \\ \vdots \\ N-1 \\ N \\ N+1 \\ \vdots \\ K-1 \\ K \end{array} \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & \ddots & 0 & 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \ddots & 0 & 0 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & 0 & 0 & 0 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & 0 & \lambda_r H_0 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & 0 & \ddots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \ddots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & 0 & 0 & \lambda_r H_0 & \vdots \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \lambda_r H_0 \end{bmatrix};$$

and:

$$A_2 = \begin{array}{c} 0 \\ 1 \\ \vdots \\ N-1 \\ N \\ N+1 \\ \vdots \\ K-1 \\ K \end{array} \begin{bmatrix} 0 & i\alpha & 0 & 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & \ddots & i\alpha & 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \ddots & \ddots & 0 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & 0 & 0 & i\alpha & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & 0 & i\alpha(1-H_0) & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & 0 & i\alpha(1-H_0) & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & 0 & \ddots & \ddots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & 0 & 0 & 0 & i\alpha(1-H_0) & \vdots \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & i\alpha(1-H_0) \end{bmatrix}.$$

Additionally, let $\pi = [\pi_0, \pi_1, \pi_2, \ldots, \pi_{R-1}, \pi_R]$ represent the exact solution, where $\pi_i = (\pi(i,0), \pi(i,1), \ldots, \pi(i,K))$ for $i = 0, 1, 2, \ldots, R$, and $\forall \pi_i \in M_{1\times(K+1)}$, where the above results satisfy equations:

$$\pi Q = 0; \tag{2}$$

and:

$$\sum_{i=0}^{R} \sum_{j=0}^{K} \pi(i,j) = 1. \tag{3}$$

Hence, this investigation adopts the following balance equations:

$$\begin{gathered} \pi_0 B_0 + \pi_1 A_2 = 0 \\ \pi_0 A_0 + \pi_1 B_1 + \pi_2 A_2 = 0 \\ \pi_1 A_0 + \pi_2 B_2 + \pi_3 A_2 = 0 \\ \vdots \\ \pi_{R-2} A_0 + \pi_{R-1} B_{R-1} + \pi_R A_2 = 0 \\ \pi_{R-1} A_0 + \pi_R B_R = 0 \end{gathered}$$

Applying Equations (2) and (3), we could obtain $\pi_i = (\pi(i,\ 0),\ \pi(i,\ 1), \ldots, \pi(i,\ K))$ for $i = 0,\ 1,\ 2, \ldots, R$, by using the Matlab software given relatively small K. The three performance measures derived from the above equations are as follows:

(1) The probability that the system fails to serve a regular customer on the first attempt is $P_r = \sum_{i=0}^{R} \sum_{j=N}^{K} \pi(i,j)$, because calls from regular users will be redirected to the retrial group when N or more operators are busy;

(2) The probability that the system fails to serve a VIP caller-is $P_v = \sum_{i=0}^{R} \pi(i,K)$, because all operators are too busy to answer v-call; and

(3) The average number of regular customers in the retrial group is $L_q = \sum_{i=1}^{R} \sum_{j=0}^{K} i\pi(i,j)$.

3. Computational time

Because of a large $K > 100$, the abovementioned measures require computation through experiments. To determine the computational complexity of the three performance measures, this investigation establishes parameters for an experiment based on a real case. Although the number of operators is 71 ($K = 71$), to clarify the problem of long computational time, this investigation also extended K from 71 to 451. Table II shows the time spent to calculate the blocking probability of calls from regular customers based on the data listed in Table I for K with different capacity through Equations (2) and (3). The experimental results demonstrate that because this queueing model is a two-dimensional Markov chain, the balance equations are so complex that their algorithmic computing time grows exponentially (Figure 3). The experiment is implemented using a mathematical tool (Matlab® 7.0) with a high-end workstation (Windows® 7 Business version, Intel® Core™ i7 CPU at 2.69 GHZ, and three gigabytes of random access memory (RAM)) (Tables I and II).

Obviously, such an exact method becomes highly impractical given large K. Over 20 hours (73,072 seconds) are needed to calculate the results when K is 126. Therefore, Section 3 of this study introduces an approximation known as the phase merging algorithm to calculate Equations (2) and (3) to solve the computational problem.

4. The approximation method

Numerous approximation methods have been proposed to solve the computational problem associated with a large number of channels (exceeding 1,000) with heavy traffic in the telecommunication research literature (Abdrabou and Zhuang, 2011; Allon *et al.*, 2013; Bandi and Bertsimas, 2012; Bicen *et al.*, 2012; Halfin and Whitt, 1981; Nelson, 2012; Sriram and Whitt, 1986; Whitt, 1983). However, call centers have their own limitations and unique features in handling calls from real people, and do

Table I.
Parameter set

Parameter	Value	Parameter	Value
N	K	H_1	0.999
R	15	λ_r	7.78
μ	1/3	λ_v	3.89
α	1.8	λ	11.67
H_0	0.15	K	71

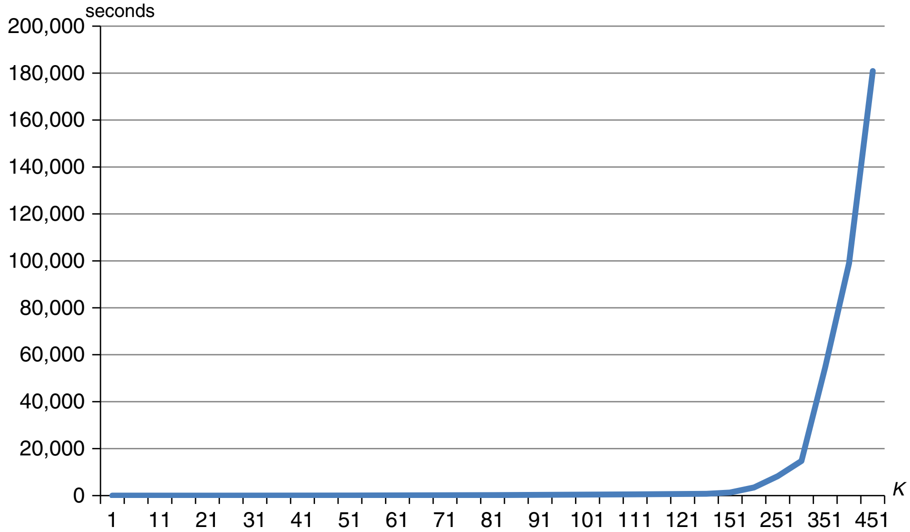


Figure 3. Computational time

Table II. Computational time

K	Seconds	K	Seconds
1	0.02	86	237.56
6	0.17	91	305.57
11	0.56	96	362.80
16	1.31	101	428.32
21	2.76	106	476.88
26	6.30	111	558.36
31	10.25	116	638.95
36	16.44	121	731.54
41	24.79	126	761.41
46	35.30	151	1,325.80
51	49.16	201	3,484.20
56	65.94	251	8,271.80
61	87.36	301	14,773.00
66	111.10	351	54,720.00
71	139.09	401	99,006.70
76	168.04	451	180,890.00
81	208.39		

not merely transmit data. It is necessary to consider simple approximation methods. One such method is known as phase merging algorithm and was introduced by Melikov and Babayev (2006) and adopted by Choi, Melikov and Velibekov (Choi *et al.*, 2008; Liang and Luh, 2013; Ponomarenko *et al.*, 2010) to compute stationary probability. This method can be implemented in call centers. The approximation method is shown as follows.

Application of the phase merging algorithm to this model requires making the following assumptions (based on Figure 2):

$$\lambda_v \gg \lambda_r H_0 \tag{4}$$

$$K\mu \gg R\alpha \tag{5}$$

$$\lambda_v \gg R\alpha(1-H_1) \tag{6}$$

The above assumptions are rational in a call center. When the arrival rate of v-calls significantly exceeds that of r-calls multiplied by the probability of r-calls entering the retrial group being blocked, a high chance of more than N customers in service implies protection for VIPs. In this regard, this study assumes the up vertical flow must significantly exceed the left and right horizontal flow (see Figure 2). Additionally, the down vertical flow significantly exceeds the up slanting flow shown in Figure 2. If the model satisfies the above conditions, then it is possible to adopt the phase merging algorithm.

The steps of the phase merging algorithm are illustrated based on the above assumptions, as follows:

(1) Split state space of the original two-dimensional Markov chain by (Choi *et al.*, 2008):

$$S = \bigcup_{i=0}^{R} S_i, S_i \cap S_{i'} = \emptyset, \quad i \neq i' \tag{7}$$

where $S_i = \{(i, j) \in S : 0 \leqslant j \leqslant K\}$.

(2) The conditional stationary distribution $\pi(j|i) = \frac{\pi(i,j)}{\sum_{j=1}^{K} \pi(i,j)}$ within each class S_i has been defined as an $M/M/K/K$ queueing system given the following arrival rates:

$$\begin{cases} \lambda_r + \lambda_v, & if\ j < N \\ \lambda_v, & if\ j \geqslant N \end{cases}.$$

Therefore:

$$\pi(j|i) = \begin{cases} \frac{\rho^j}{j!}\pi(0|i), & if\ 1 \leqslant j \leqslant N \\ \frac{\rho^N}{j!}\rho_v^{j-N}\pi(0|i), & if\ N \leqslant j \leqslant K \end{cases}.$$

where $\rho = \frac{\lambda_r + \lambda_v}{\mu}$, $\rho_v = \frac{\lambda_v}{\mu}$, $0 \leqslant j \leqslant K$ and:

$$\pi(0|i) = \left(\sum_{j=0}^{N} \frac{\rho^j}{j!} + \rho^N \sum_{j=N+1}^{K} \frac{\rho_N^{j-N}}{j!}\right)^{-1} \tag{8}$$

From Equation (8) (Choi *et al.*, 2008), we find the probability $\pi(j|i)$ which does not depend on i. Therefore, $p(j) \stackrel{\Delta}{=} \pi(j|i)$ is omitted.

(3) All states within subset S_i could be merged into a single merged state $<i>$. The merged model is a birth and death process and the transition rates are manipulated with little algebra, which yields:

$$q\left(<i>, <i'>\right) \begin{cases} \lambda_r H_0 \sum_{J=N}^{K} p(j), & if\ i \geqslant 0,\ i' = i+1, \\ i\alpha\left(\sum_{j=0}^{N-1} p(j) + (1-H_1)\sum_{J=N}^{K} p(j)\right), & if\ i \geqslant 1,\ i' = i-1. \\ 0, & \text{otherwise} \end{cases}$$

The stationary distribution of the merged model $\psi(<i>)$, $<i>, i=0,1,\ldots,R$ is thus obtained. Based on the stationary probability of *M*/*M*/*R*/*K*, this study provides the following equations (Choi *et al.*, 2008):

$$\psi(<i>) = \frac{t^i}{i!}\psi(<0>), \tag{9}$$

where:

$$\psi(<0>) = \left(\sum_{i=0}^{R}\frac{t^i}{i!}\right)^{-1}, \tag{10}$$

$$t = \frac{\lambda_r H_0 \sum_{j=N}^{K} p(j)}{\alpha\left(\sum_{j=0}^{N-1} p(j) + (1-H_1)\sum_{j=N}^{K} p(j)\right)}, \tag{11}$$

Define:

$$p(i,j) = p(j)\psi(<i>). \tag{12}$$

Based on the above descriptions, we obtain the following lemmas.

Based on the assumptions of Equations (4)-(6), the phase merging algorithm can approximate π using Equation (12). It is explained in the following.

Lemma 1. Because $\pi(i,j)$ is the stationary probability of state (i,j) and the expected rates of flow in and out of state (i,j) must be equal, this study considers three cases with given state (i, j) and S_i.

- Case 1. $j > N$.

Flow in state (i,j), it gives:

$$\pi(i+1,j)[(i+1)\alpha(1-H_1)] + \pi(i,j-1)\lambda_v + \lambda_r H_0 + \pi(i,j+1)(j+1)\mu \tag{13}$$

Flow out state (i,j), it gives:

$$\pi(i,j) + (\lambda_v + \lambda_r H_0 + j\mu + i\alpha(1-H_1) \tag{14}$$

Because Equation (13) = Equation (14), by little algebra, we have:

$$\pi(i,j)(\lambda_v + j\mu) = \pi(i,j-1)\lambda_v + \pi(i,j+1)(j+1)\mu + \varepsilon_j \tag{15}$$

where $\varepsilon_j = \pi(i+1,j)[(i+1)\alpha(1-H_1)] + \pi(i-1,j)\lambda_r H_0 - \pi(i,j)\lambda_r H_0 - \pi(i,j)i\alpha(1-H_1)$. Let Equation(15) be divided by λ_v and it produces:

$$\frac{\pi(i,j)(\lambda_v + j\mu)}{\lambda_v} = \frac{\pi(i,j-1)\lambda_v}{\lambda_v} + \frac{\pi(i,j+1)(j+1)\mu}{\lambda_v} + \pi(i+1,j)\frac{(i+1)\alpha(1-H_1)}{\lambda_v} + \pi(i-1,j)\frac{\lambda_r H_0}{\lambda_v} - \pi(i,j)\frac{\lambda_r H_0}{\lambda_v} - \pi(i,j)\frac{i\alpha(1-H_1)}{\lambda_v}$$

The ε_j in Equation (15) might be ignored, because (4) and (6) imply ε_j is approaching to zero. The Equation (16) represents the balance equation within S_i.

- Case 2. $j = N$:

$$\pi(i,j)(\lambda_v + j\mu) = \pi(i,j-1)(\lambda_v + \lambda_r) + \pi(i,j+1)(j+1)\mu + \varepsilon_j \tag{16}$$

where $\varepsilon_j = \pi(i+1,j)[(i+1)\alpha(1-H_1)] + \pi(i-1,j)\lambda_r H_0 - \pi(i,j)\lambda_r H_0 - \pi(i,j)i\alpha(1-H_1)$. Let (16) be divided by $\lambda_v K\mu$ and it produces:

$$\frac{\pi(i,j)(\lambda_v + j\mu)}{\lambda_v K\mu} = \frac{\pi(i,j-1)(\lambda_v + \lambda_r)}{\lambda_v K\mu} + \frac{\pi(i,j+1)(j+1)\mu}{\lambda_v K\mu} + \pi(i+1,j)\frac{(i+1)\alpha(1-H_1)}{\lambda_v K\mu}$$
$$+\pi(i-1,j)\frac{\lambda_r H_0}{\lambda_v K\mu} - \pi(i+1,j-1)\frac{(i+1)\alpha}{\lambda_v K\mu}$$
$$-\pi(i,j)\frac{\lambda_r H_0}{\lambda_v K\mu} - \pi(i,j)\frac{i\alpha(1-H_1)}{\lambda_v K\mu}$$

The ε_j in Equation (16) might be ignored, because (4)-(6) imply ε_j is approaching to zero.

- Case 3. $j < N$:

$$\pi(i,j)(\lambda_v + \lambda_r + j\mu) = \pi(i,j-1)(\lambda_v + \lambda_r) + \pi(i,j+1)(j+1)\mu + \varepsilon_j \tag{17}$$

where $\varepsilon_j = \pi(i+1,j-1)\alpha(i+1) - \pi(i,j)i\alpha$.
Let (17) be divided by $K\mu$ which leads to:

$$\frac{\pi(i,j)(\lambda_v + \lambda_u + j\mu)}{K\mu} = \frac{\pi(i,j-1)(\lambda_v + \lambda_r)}{K\mu} + \frac{\pi(i,j+1)(j+1)\mu}{K\mu}$$
$$+\pi(i+1,j-1)\frac{(i+1)\alpha}{K\mu} - \pi(i,j)\frac{i\alpha}{K\mu}$$

The ε_j in Equation (17) might be ignored, because (5) implies ε_j is approaching to zero. Using Equations (15)-(17) and given S_i, we can ignore the horizontal and slanting flows in the system (see Figure 2) except for vertical flows. Therefore, based on Equations (15)-(17), we conclude the following: If ε_j is small and can be ignored, $p\ (i,\ j)$ is an ε-approximation of $\pi(i, j)$. Based on four approximations, (18)-(21) could be derived for the measures of performance and the average waiting time (Choi *et al.*, 2008):

$$P_r \approx 1 - p(0)\sum_{i=0}^{N-1}\frac{\rho^i}{i!}, \tag{18}$$

$$P_v \approx p(0)\frac{\rho^N \rho_v^{K-N}}{K!}, \tag{19}$$

$$L_q \approx \psi(<0>)\sum_{J=1}^{R}\frac{t^j}{(j-1)!}, \tag{20}$$

Now we compute an effective arrival rate of r-calls in the retrial group denoted by λ_r^*. Because the total calls in the retrial group is completely owing to regular customers, it is calculated by the average number of r-calls admitted in the retrial group minus some of them leave the system with probability (1-H_1) or find no servers available subsequently. Let $P(R)$ denote the probability that the retrial group is full, and P_r^* represents the probability that r-calls may be routed to the retrial group where $P(R) = \sum_{J=0}^{K} \pi(R,j)$ and $P_r^* = \sum_{i=0}^{R-1} \sum_{j=N}^{K} \pi(i,j)$.

Lemma 2. The effective arrival rate of r-calls is given by:

$$\lambda_r^* = \left(\lambda_r P_r^* H_0(1-P(R))\right)\left(\frac{1-P_r^*}{1-H_1P_r^*(1-P(R))}\right) \text{ and } 0 \leqslant \frac{1-P_r^*}{1-H_1P_r^*(1-P(R))} \leqslant 1$$

Proof: Let $\Delta = \lambda_r P_r^* H_0(1-P(R))$. We have:

$$\lambda_r^* = \Delta - \left[\Delta P_r^*(1-H_1) + \Delta P_r^* H_1 P(R)\right] - \left[\Delta P_r^*(1-H_1)H_1P_r^*(1-P(R)) + \Delta P_r^* H_1 P(R) H_1 P_r^*(1-P(R))\right. - \cdots$$

$$= \Delta\left[1 - \frac{P_r^*(1-H_1)}{1-H_1P_r^*(1-P(R))} - \frac{P_r^*H_1P(R)}{1-H_1P_r^*(1-P(R))}\right]$$

$$= \Delta\left(\frac{(1-P_r^*)}{1-H_1P_r^*(1-P(R))}\right)$$

$$= \lambda_r P_r^* H_0(1-P(R))\left(\frac{(1-P_r^*)}{1-H_1P_r^*(1-P(R))}\right)$$

$$\because 0 \leqslant H_1 \leqslant 1 \text{ and } 0 \leqslant P(R) \leqslant 1$$

$$\therefore 0 \leqslant \frac{1-P_r^*}{1-H_1P_r^*(1-P(R))} \leqslant 1$$

Finally, from the abovementioned derivations, we obtain the mean waiting time in the retrial group based on the formula of Little is:

$$\overline{W_r} = \frac{L_q}{\lambda_r^*} \tag{21}$$

The call center must determine the number of servers required to ensure that both types of customers (regular and VIP customers) receive satisfactory service. Satisfactory service (Chen and Henderson, 2001) means that at least q_r percent of r-calls are served within m seconds. This definition of satisfactory service is common in the call center industry. Restated, partial r-calls are answered immediately (set $m = 0$). This investigation considers two waiting time with r-calls: N equals zero and N does not equal zero ($0 < N \leqslant K$).

- Case 1. $N = 0$.

Clearly, any regular customer cannot enter the system to receive service. On the other hand, the *M/M/K/K* queueing system only serves VIP customers. Since VIP customers do not have waiting time in the model, the waiting times for v-calls and r-calls both equal zero. Thus we can conclude that for any time when w is zero or greater, it has:

$$P(\text{waiting time of both calls} > w) = 0. \tag{22}$$

- Case 2. $0 < N \leqslant K$.

Because VIP customers are served immediately, this investigation only needs to consider the waiting time for regular customers in the retrial group. Abate and Whitt have proposed algorithms for numerically calculating tail probabilities. However, the proposed algorithms require complex arithmetic, which is inconvenient for the manager of a call center manager to apply to calculate waiting time. Let W_r denote a random variable of the waiting time of r-calls. Given Markov's inequality for a nonnegative random variable X and constants x, β larger than zero indicates:

$$P(X \geqslant x) \leqslant \frac{E\left(X^{\beta}\right)}{x^{\beta}},$$

where $P(.)$ denotes a probability function (Ross, 2002). Hence, we conclude that:

$$P(W_r > w) \leqslant \frac{\overline{W}_r}{w}. \tag{23}$$

Equation (5) provides an upper bound on the tail probabilities of W_r, and hence a lower bound of $P(W_r > w)$, namely, $P(W_r > w | \lambda_v = 0)$ for any w that exceeds zero.

That is, given $\forall w > 0$ then:

$$P\left(W_r > w \middle| \lambda_v = 0\right) \leqslant P(W_r > w) \leqslant \frac{\overline{W}_r}{w}. \tag{24}$$

Apparently, Equation (24) provides lower and upper bounds for regular customer waiting times. The upper bound is calculated by Equation (21). The manager can use these results to determine the optimum N that satisfies regular customers the service requirements can be met. The procedure used to calculate N for satisfactory service is shown in Algorithm WP. The manager could simply increase N until Equation (23) with $w = m$ falls below (100-q_r) percent. Therefore, we can be reasonably confident of approaching these goals using formulae (21) and (24). WP algorithm is written as follows:

Step 1. Given N, w, q_r.
Step 2. Compute $x = \frac{E(W_r|N)}{w}$
Step 3. If $x < (100 - q_r)$ percent, then STOP.
Step 4. Increase N and go to Step 2.

Based on satisfactory service requirements, we compute the probability of waiting time which is more than w seconds for regular customers when the traffic intensity of regular customers is $\lambda_r / N\mu$. From (24), we can determine the lower bound and upper bound of the probability. The effective arrival rate is λ_r^*. Given w, we have $\overline{W}_r / w = L_q / w\lambda_r^*$. Furthermore, the lower bound can be found when $\lambda_v = 0$. In this case, we consider it as a traditional $M/M/N/K$ queueing model where $K=N+R$. To illustrate satisfactory service requirements, we have set up the parameters in Table III.

The experiment results show that the upper bound of probability of waiting time goes up along with the increasing arrival rate of regular customers. Additionally, under the study it expects that a satisfactory service represents at least 90 percent of r-calls are served within 50 seconds. We adopt WP algorithm to compute N with $w=50$ and $q_r=90$. The experiment results show that the probability of waiting time more than 50 seconds is below 10 percent when $\lambda_r/N\mu = 1.556$ and $N \geqq 48$ (Figure 4). Hence, the call center may control the probability of waiting time of regular customers less than 50 seconds is more than 90 percent by assigning more than 48 servers.

Finally, from Equations (1) and (2), we obtain $\pi_i = (\pi(i, 0), \pi(i, 1), \ldots, \pi(i, K))$ for $i=0, 1, 2, \ldots, R$, by using a mathematical tool (Matlab® 7.0) with a high-end workstation, we can calculate the results using the above method (the exact method). However, it is easy to identify the optimal N when K is small. When K is large, such complex computations might slow the computer system to a standstill. For example, when the number of operators is ⩾70 (the number of the operators of the case

Parameter	Upper bound	Lower bound
K	70	70
R	15	15
N	35	35
$1/\mu$	3	3
α	1.8	1,000
H_0	0.15	1
H_1	0.999	1
λ_r	7.78; 10.37; 12.96; 15.56; 18.15; 20.74; 23.34	7.78; 10.37; 12.96; 15.56; 18.15; 20.74; 23.34
λ_v	3.89; 5.19; 6.48; 7.78; 9.08; 10.37; 11.67	0; 0; 0; 0; 0; 0; 0
w	50 (sec.)	50 (sec.)

Table III. Parameters for upper bound and lower bound tests

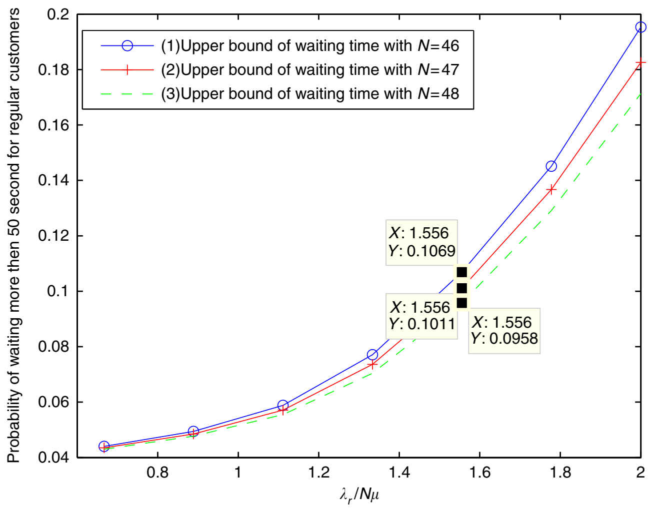


Figure 4. Probability of waiting time more than 50 seconds

company), almost six hours are needed to complete the computation. Therefore, the approximation method must be discussed.

Finally, to clarify the errors between the exact and approximation methods, this investigation obtained the results listed in Table III. In the experiments, we assume $R=15$, $\alpha=1.8$, $\mu=0.33$, $\lambda=11.67$ and $N=K$. This investigation determined that the errors between the exact and approximation methods decrease with increasing K. From Table III, we found almost no error between the exact method and the approximation method once $K>40$. Restated, if K exceeds 51, P_r decreases to zero based on the original assumptions (Table IV). To clarify the comparison of the exact and approximation methods, λ must exceed the original assumptions. This study thus increased λ from 11.67 to 211.67. The results are listed in Table V. Additionally, the

K	Computational time by approximation method (sec.)	Computational time by exact method (sec.)	Speed ratio	P_r by approximation method	P_r by exact method
1	0.0312	0.0624	2.00	0.972	0.991
6	0.0468	0.39	8.33	0.834	0.939
11	0.0468	0.5928	12.67	0.697	0.813
16	0.078	1.2792	16.40	0.563	0.655
21	0.0936	2.652	28.33	0.433	0.502
26	0.1092	5.85	53.57	0.310	0.357
31	0.156	9.4537	60.60	0.198	0.226
36	0.1716	14.3053	83.36	0.107	0.120
41	0.2028	21.7621	107.31	0.044	0.048
46	0.2184	31.8086	145.64	0.012	0.013
51	0.2964	44.9283	151.58	0.002	0.002
56	0.3432	59.14	172.32	0.000	0.000
61	0.3744	76.7993	205.13	0.000	0.000
66	0.4368	98.9358	226.5	0.000	0.000
71	0.4836	123.194	254.74	0.000	0.000

Table IV.
Exact vs approximation ($\lambda=11.67$)

K	Computational time by approximation method (sec.)	Computational time by exact method (sec.)	Speed ratio	P_r by approximation method	P_r by exact method
1	0.016	0.031	2.00	0.998	0.999
6	0.016	0.031	2.00	0.991	0.992
11	0.078	0.593	7.60	0.983	0.985
16	0.094	1.217	13.00	0.975	0.978
21	0.094	2.184	23.33	0.967	0.971
26	0.140	4.976	35.44	0.959	0.964
31	0.125	8.518	68.25	0.951	0.957
36	0.125	13.869	111.13	0.943	0.950
41	0.187	20.904	111.67	0.936	0.943
46	0.250	30.358	121.63	0.928	0.936
51	0.296	41.434	139.79	0.920	0.929
56	0.296	56.940	192.11	0.912	0.922
61	0.343	76.425	222.68	0.904	0.915
66	0.406	98.187	242.08	0.896	0.908
71	0.484	123.184	254.72	0.888	0.901

Table V.
Exact vs approximation ($\lambda=211.67$)

speed ratio (Computational $time_{approximation}$/Computational $time_{exact}$) is used to clarify the outstanding performance of the approximation method.

5. Discussion

However, to merely identify the errors between the exact and approximation methods is insufficient, because the manager needs to know whether their company needs to adopt the approximation method to determine the management costs. The manager's question can be answered by discussing the difference in errors between the approximation and exact methods.

Restated, this study demonstrates the errors of the approximation and exact methods of different performance measures along with the increasing number of total operators. Table VI lists the parameters for finding the errors of P_r, P_v, L_q and *MC* between the approximation and exact methods. Figures 5 and 6 lists the experimental results. Moreover, to identify the errors between the absolute and exact methods, this study assumes a large λ (23.34, which is twice 11.67).

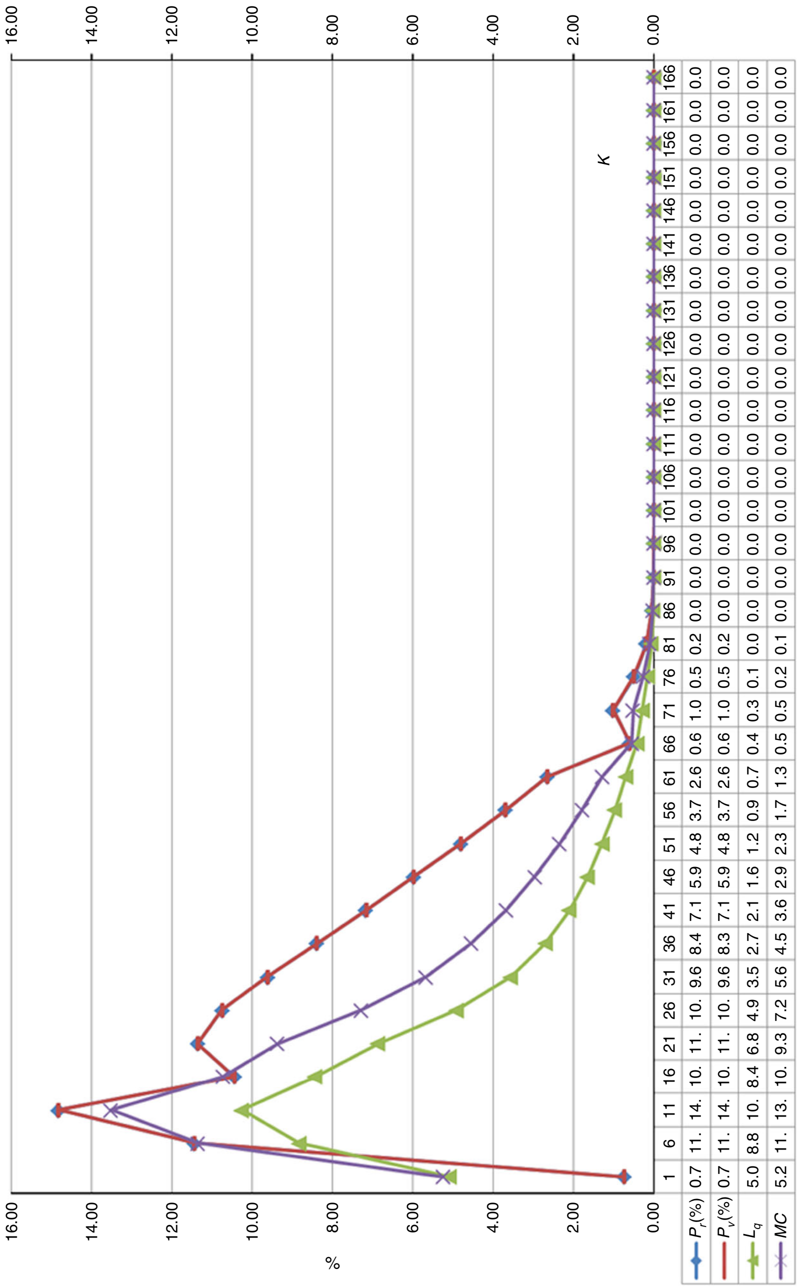
The numerical results are illustrated in Figures 5 and 6 when $\lambda = 11.67$ and 23.34, respectively. The error between approximation and exact methods are increasing at first (the top error values when $\lambda = 11.67$ and 23.34 are $K = 11$ and $K = 6$, respectively) and decreasing gradually. The error between the approximation and exact methods will approach to zero when $\lambda = 11.67$ and 23.34 are $K = 86$ and $K = 51$, respectively. The *MC* could be found quickly with small error (if the acceptable error rate is smaller than 5 percent) when $\lambda = 11.67$ and 23.34 are $K = 36$ and $K = 16$, respectively. The aforementioned numerical results imply that using approximation method is feasible for the manager of a call center to calculate *MC* value in a busy call center. If 5 percent of error rate can be accepted by a manager, the manager can decide the optimal K. Thus, the manager of a call center can hire operators explicitly based on the estimated *MC*. Finally, this study shows the computational results of the *MC* based on different retrial rates (α) that represent the frequency of sending the blocking calls in the waiting queue to operators. The computational results show that *MC* will be 13.5, 61.62 and 70.26 when retrial rates are 1, 30 and 120, respectively.

6. Conclusion

This investigation presents a computational method and management cost function intended to identify the optimal number of operators for a call center. Because of computational limitations, many operators could not be easily analyzed using the exact

Parameter	Value ($\lambda = 11.67$)	Value ($\lambda = 23.34$)
N	25	25
R	15	15
μ	1/3	1/3
α	1.8	1.8
H_0	0.15	0.15
H_1	0.999	0.999
λ_r	7.78	15.56
λ_v	3.89	7.78
C_r	1	1
C_v	2	2
C_w	20	20

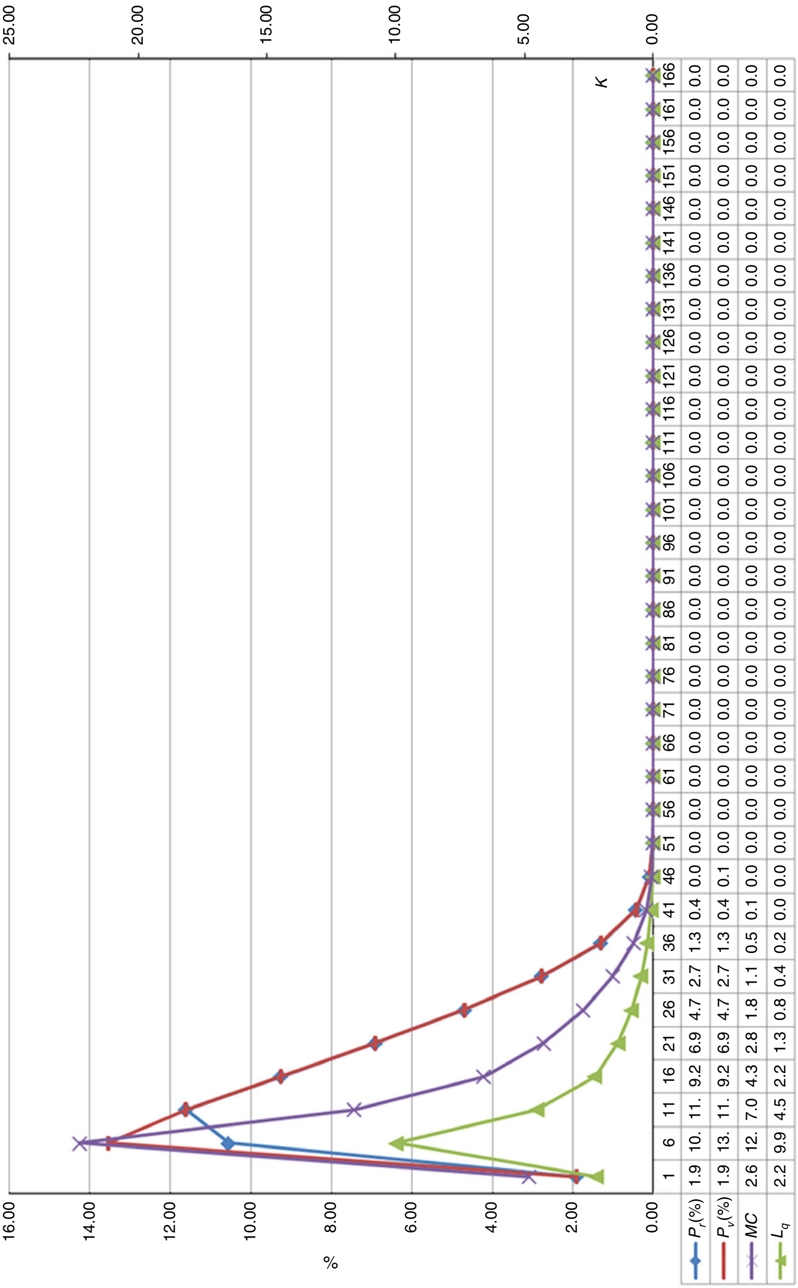
Table VI. Experimental parameters



K	1	6	11	16	21	26	31	36	41	46	51	56	61	66	71	76	81
P_r(%)	0.7	11.	14.	10.	11.	10.	9.6	8.4	7.1	5.9	4.8	3.7	2.6	0.6	1.0	0.5	0.2
P_v(%)	0.7	11.	14.	10.	11.	10.	9.6	8.3	7.1	5.9	4.8	3.7	2.6	0.6	1.0	0.5	0.2
L_q	5.0	8.8	10.	8.4	6.8	4.9	3.5	2.7	2.1	1.6	1.2	0.9	0.7	0.4	0.3	0.1	0.0
MC	5.2	11.	13.	10.	9.3	7.2	5.6	4.5	3.6	2.9	2.3	1.7	1.3	0.5	0.5	0.2	0.1

K	86	91	96	101	106	111	116	121	126	131	136	141	146	151	156	161	166
P_r(%)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
P_v(%)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
L_q	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
MC	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Figure 5.
Absolute errors of approximation minus exact methods with K ($\lambda = 11.67$)



K	1	6	11	16	21	26	31	36	41	46	51	56	61	66	71	76	81
P_r(%)	1.9	10.	11.	9.2	6.9	4.7	2.7	1.3	0.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
P_v(%)	1.9	13.	11.	9.2	6.9	4.7	2.7	1.3	0.4	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
MC	2.6	12.	7.0	4.3	2.8	1.8	1.1	0.5	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
L_q	2.2	9.9	4.5	2.2	1.3	0.8	0.4	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

K	86	91	96	101	106	111	116	121	126	131	136	141	146	151	156	161	166
P_r(%)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
P_v(%)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
MC	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
L_q	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Figure 6.
Absolute errors of approximation minus exact methods with K ($\lambda = 23.34$)

method. For the manager of a call center, the sooner the optimal solution is found, the faster business strategies are deployed. This study develops an approximation method and compares it with the exact method. The approximation method demonstrates outstanding computational performances when the state variables are too big to be handled in matrix.

However, a good call center management scheme must consider not only operational costs, but also traffic conditions and other managerial variables. Future research should include more variables to identify nonlinear optimization with stochastic parameters.

Reference

Abdrabou, A. and Zhuang, W. (2011), "Probabilistic delay control and road side unit placement for vehicular ad hoc networks with disrupted connectivity", *IEEE Journal on Selected Areas in Communications*, Vol. 29 No. 1, pp. 129-139.

Allon, G., Deo, S. and Lin, W. (2013), "The impact of size and occupancy of hospital on the extent of ambulance diversion: theory and evidence", *Operations Research*, Vol. 61 No. 3, pp. 544-562.

Armstrong, M. and Taylor, S. (2014), *Armstrong's Handbook of Human Resource Management Practice*, Kogan Page Publishers, London.

Artalejo, J.R. and Gómez-Corral, A. (2008), *Retrial Queueing Systems: A Computational Approach*, Springer, Berlin.

Artalejo, J.R., Orlovsky, D.S. and Dudin, A.N. (2005), "Multi-server retrial model with variable number of active servers", *Computers & Industrial Engineering*, Vol. 48 No. 2, pp. 273-288.

Bandi, C. and Bertsimas, D. (2012), "Tractable stochastic analysis in high dimensions via robust optimization", *Mathematical programming*, Vol. 134 No. 1, pp. 23-70.

Benaroch, M., Dai, Q. and Kauffman, R.J. (2010), "Should we go our own way? Backsourcing flexibility in IT services contracts", *Journal of Management Information Systems*, Vol. 26 No. 4, pp. 317-358.

Bicen, A.O., Gungor, V.C. and Akan, O.B. (2012), "Delay-sensitive and multimedia communication in cognitive radio sensor networks", *Ad Hoc Networks*, Vol. 10 No. 5, pp. 816-830.

Bright, L. and Taylor, P.G. (1995), "Calculating the equilibrium distribution in level dependent quasi-birth-and-death processes", *Stochastic Models*, Vol. 11 No. 3, pp. 497-525.

Chen B.P.K. and Henderson S.G. (2001), "Two issues in setting call centre staffing levels", *Annals of Operations Research*, Vol. 108 No. 1, pp. 157-192.

Choi, B.D. and Chang, Y. (1999), "Single server retrial queues with priority calls", *Mathematical and Computer modelling*, Vol. 30 No. 3, pp. 7-32.

Choi, B.D., Melikov, A.G.A.S.S.I. and Velibekov, A. (2008), "A simple numerical approximation of joint probabilities of calls in service and calls in the retrial group in a picocell", *Appl. Comput. Math*, Vol. 7 No. 1, pp. 21-30.

Do, T.V. (2010), "A new computational algorithm for retrial queues to cellular mobile systems with guard channels", *Computers & Industrial Engineering*, Vol. 59 No. 4, pp. 865-872.

Do, T.V., Wüchner, P., Bérczes, T., Sztrik, J. and De Meer, H. (2014), "A new finite-source queueing model for mobile cellular networks applying spectrum renting", *Asia-Pacific Journal of Operational Research*, Vol. 31 No. 2, pp. 1-22.

Dubelaar, C., Sohal, A. and Savic, V. (2005), "Benefits, impediments and critical success factors in B2C E-business adoption", *Technovation*, Vol. 25 No. 11, pp. 1251-1262.

Dudina, O., Kim, C. and Dudin, S. (2013), "Retrial queuing system with Markovian arrival flow and phase-type service time distribution", *Computers & Industrial Engineering*, Vol. 66 No. 2, pp. 360-373.

Gans, N., Koole, G. and Mandelbaum, A. (2003), "Telephone call centers: tutorial, review, and research prospects", *Manufacturing & Service Operations Management*, Vol. 5 No. 2, pp. 79-141.

Gómez-Corral, A. (2006), "A bibliographical guide to the analysis of retrial queues through matrix analytic techniques", *Annals of Operations Research*, Vol. 141 No. 1, pp. 163-191.

Halfin, S. and Whitt, W. (1981), "Heavy-traffic limits for queues with many exponential servers", *Operations research*, Vol. 29 No. 3, pp. 567-588.

Huang, H.S. (2010), *A Queueing Model of Call Center by Two-Dimensional Markov Chain Approach in A Case Study*, National ChengChi University, Taipei.

Kim, C., Dudin, A., Dudin, S. and Klimenok, V. (2012), "A queueing system with batch arrival of customers in sessions", *Computers & Industrial Engineering*, Vol. 62 No. 4, pp. 890-897.

Lee, J.H. and Park, S.C. (2005), "Intelligent profitable customers segmentation system based on business intelligence tools", *Expert Systems with Applications*, Vol. 29 No. 1, pp. 145-152.

Li, X. and Chandra, C. (2007), "A knowledge integration framework for complex network management", *Industrial Management & Data Systems*, Vol. 107 No. 8, pp. 1089-1109.

Liang, C.C. and Luh, H. (2013), "Optimal services for content delivery based on business priority", *Journal of the Chinese Institute of Engineers*, Vol. 36 No. 4, pp. 422-440.

Liang, C.C., Hsu, P.Y., Leu, J.D. and Luh, H. (2005), "An effective approach for content delivery in an evolving intranet environment – a case study of the largest telecom company in Taiwan", *Lecture Notes in Computer Science*, Vol. 3806 No. 1, pp. 740-749.

Liang, C.C., Wang, C.H., Luh, H. and Hsu, P.Y. (2009), "Disaster avoidance mechanism for content-delivering service", *Computers & Operations Research*, Vol. 36 No. 1, pp. 27-39.

Melikov, A.Z. and Babayev, A.T. (2006), "Refined approximations for performance analysis and optimization of queueing model with guard channels for handovers in cellular networks", *Computer Communications*, Vol. 29 No. 9, pp. 1386-1392.

National Communication Commission (2014), *The Type I Communications Enterprises*, ROC Government Printing Office, Taipei.

Nelson, B.L. (2012), *Stochastic Modeling: Analysis and Simulation*, Courier Dover Publications, New York.

Phung-Duc, T. (2014), "Impatient customers in power-saving data centers", in *Analytical and Stochastic Modeling Techniques and Applications*, Springer International Publishing, pp. 185-199.

Phung-Duc, T., Masuyama, H., Kasahara, S. and Takahashi, Y. (2010), "A simple algorithm for the rate matrices of level-dependent QBD processes", *Proceedings of the 5th International Conference on Queueing Theory and Network Applications, ACM, July*, pp. 46-52.

Phung-Duc, T., Masuyama, H., Kasahara, S. and Takahashi, Y. (2013), "A matrix continued fraction approach to multiserver retrial queues", *Annals of Operations Research*, Vol. 202 No. 1, pp. 161-183.

Ponomarenko, L., Kim, C.S. and Melikov, A. (2010), *Performance Analysis and Optimization of Multi-traffic on Communication Networks*, Vol. 208, Springer, Heidelberg, Dordrecht, London, and New York, NY.

Ross, S. (2002), *A First Course in Probability*, 8th ed., Pearson, New York.

Servi, L.D. (2002), "Algorithmic solutions to two-dimensional birth-death processes with application to capacity planning", *Telecommunication Systems*, Vol. 21 Nos 2/4, pp. 205-212.

Sriram, K. and Whitt, W. (1986), "Characterizing superposition arrival processes in packet multiplexers for voice and data", *Selected Areas in Communications, IEEE Journal on*, Vol. 4 No. 6, pp. 833-846.

Tran-Gia, P. and Mandjes, M. (1997), "Modeling of customer retrial phenomenon in cellular mobile networks", *IEEE Journal on Selected Areas in Communications*, Vol. 15 No. 8, pp. 1406-1414.

Whitt, W. (1983), "The queueing network analyzer", *Bell System Technical Journal*, Vol. 62 No. 9, pp. 2779-2815.

Winkler, A. (2013), "Dynamic scheduling of a single-server two-class queue with constant retrial policy", *Annals of Operations Research*, Vol. 202 No. 1, pp. 197-210.

Xu, Y., Yen, D.C., Lin, B. and Chou, D.C. (2002), "Adopting customer relationship management technology", *Industrial management & data systems*, Vol. 102 No. 8, pp. 442-452.

Yang, K.H., Lee, S.M. and Lee, S.G. (2007), "Adoption of information and communication technology: impact of technology types, organization resources and management style", *Industrial Management & Data Systems*, Vol. 107 No. 9, pp. 1257-1275.

Corresponding author
Dr Chih-Chin Liang can be contacted at: lgcwow@gmail.com

This article has been cited by:

1. Agassi Melikov, Leonid Ponomarenko, Anar Rustamov, Janos SztrikHierarchical Space Merging Algorithm for Analysis of Two Stage Queueing Network with Feedback 238-249. [CrossRef]