Queueing Analysis of Internet-Based Call Centers with Interactive Voice Response and Redial

Kosuke Hashizumea, Tuan Phung-Dueb, Shoji Kasaharaa and Yutaka Takahashia

aGraduate School of Informatics, Kyoto University
Yoshida Honmachi, Sakyo-ku Kyoto 606–8501, Japan
bDepartment of Mathematical and Computing Sciences, Tokyo Institute of Technology
Ookayama, Tokyo 152-8552, Japan
hashizume@sys.i.kyoto-u.ac.jp, tuan@is.titech.ac.jp, {shoji, takahashi}@i.kyoto-u.ac.jp

Abstract—Recently, call centers are important from a customer service point of view because they have a significant impact on customer satisfaction. In general, most of the management cost for call centers is labor cost for Customer Service Representatives (CSRs), and it is important for companies to manage CSRs in a cost-effective manner, keeping a high quality of customer service. Therefore, most of companies install Interactive Voice Response systems (IVRs) in order to not only reduce CSR management cost, but also to provide high-quality customer service. In this paper, focusing on call centers with IVRs, we investigate the impact of the service time at IVRs and number of IVRs on the utilization of a CSR. To this end, we model the call center with IVRs by a queueing system with retrials, analyzing the steady state probability by a continuous-time Markov chain. We derive performance measures such as the mean number of redialing customers, the blocking probability and the mean sojourn time. Numerical results show that under a low arrival rate of new calls, the utilization of a CSR is insensitive to the IVR service time as well as to the number of IVRs, and that the utilization grows proportionally with the increase in the probability that a customer in an IVR leaves for the CSR service.

Index Terms—call center, two-stage service, interactive voice response, redial, retrial queues

I. INTRODUCTION

A call center is a customer service department which uses telephones to interactively communicate with customers in order to resolve their problems. Because call centers provide a direct communication channel between customers and a company, they play an important role in keeping a good relationship between customers and a company and have a significant impact on customer satisfaction.

It is reported that almost 60–80% of the operational costs for a call center involves personnel expenses [1]. Therefore, an efficient deployment of Customer Service Representatives (CSR) keeping a high quality of service is indispensable for reducing operational expenses in a call center. In order to improve this trade-off, most companies install Interactive Voice Response (IVR) [2], enabling self-service of customers through an automatic voice guidance by an IVR system. An IVR system is expected to reduce the operational costs while improving service levels and profits.

In this paper, we focus on a two-stage call center which handles a call first by an IVR and then by a CSR if the call needs a further assistance after finishing service by an IVR.

Examples of call centers with IVR include telephone banking, airline reservation systems and technical support. The call processing flow in a call center with IVR is given in Figure 1.

Some performance modelling works have been done for call centers with IVRs and two-stage service. In [3], call centers handling simple and complex services are considered. This system is modeled as a queueing system where some servers can serve only a simple service. By analyzing the queueing system, the authors find an optimal strategy for outsourcing a part of the simple service. From a two-stage service queueing point of view, the model in [3] is close to that of this paper. However, the corresponding call center presented in [3] provides both services by a server, while that presented here serves customers by IVRs and CSRs, separately. Moreover, in [3] redial calls are not considered.

In [2] and [4], the performance of call centers with IVRs is analyzed. Call centers presented in [4] integrate an IVR with a telephone line. It means that if a call needs a further service by an operator, the telephone line is not released and thus the corresponding IVR cannot accommodate a new call. These traditional call centers are based only on telephone lines where CSRs work in the same place. In contrast, inspired by new Internet technology, nowadays CSRs are not necessarily concentrated in the same place. They can work in any place where Internet connection is available allowing a more flexible work style for CSRs. This type of call centers is referred to as virtual call centers [5] (Internet-based call centers) which is going to be popular in service industry. In these call centers,
calls are first connected to IVRs where an automatic guidance is made. After receiving a service from an IVR, a call is forwarded to a CSR and the IVR that had been used is released for accommodating a newly arrived call.

In practice, if a call which is not connected upon arrival or cannot complete service due to the unavailability of the CSR, the call may reattempt after some random time [6]. This fact motivates us to take retrial phenomenon of calls into account.

In [7] and [8], a tandem queue with retrial close to our model is presented for performance analysis of short TCP connections. In [7], under the assumption that the retrial of calls does not depend on the number of redialing calls, the authors obtained explicit expressions for the case of single server in each stage. In [8], an extension to multi-stage tandem queue with multiple servers in each stage is presented. The authors in [8] analyze the model using mean value analysis and fixed point approximation. The analysis in [7] is valid in a light load regime while a call center is expected to operate and fixed point approximation. The analysis in [8] is valid in a heavy load regime in practice in order to achieve a high productivity. Furthermore, since the model in [8] is motivated from modelling of TCP flows, all packets (calls) are assumed to pass all the stages. However, in a call center context, some calls may complete service at the IVR level.

In this study, we present a two-stage tandem queue with retrials for Internet-based call centers with IVRs. Our model extends that of [8] by taking into account the possibilities that a call may complete service at an IVR and that a blocked call may not retry. We evaluate the impact of the service time at an IVR, the number of IVRs and the probability that a call needs an assistance by a CSR on the performance of a call center. We formulate the queueing system using a level-dependent quasi-birth-and-death (QBD) process whose stationary distribution is calculated using the algorithm in [9]. This methodology yields accurate results for any load regime. We derive the performance measures such as the utilizations of a CSR and an IVR, the mean number of redialing calls, the blocking probabilities and the mean sojourn time of a call.

The rest of this paper is organized as follows. We present a queueing model for a call center with IVR and redial calls in Section II. In Section III, we formulate the queueing model using a level-dependent QBD process and derive performance measures. In Section IV, numerical results are extensively presented to show the effects of parameters on the performance of call centers. Finally, concluding remarks are presented in Section V.

II. THE MODEL

In this section, we present a two-stage tandem queueing system with retrials for call centers with IVRs, as depicted in Figure 2. New calls arrive at a call center according to a Poisson process with rate $\lambda$. In the call center, there are $c_1$ IVRs and $c_2$ CSRs. All the calls are first processed by an IVR and some of them are forwarded to a CSR for receiving further service. A new call either occupies an idle IVR if any or is blocked if all the $c_1$ IVRs are occupied upon arrival. The processing time in each IVR is assumed to follow an exponential distribution with mean $1/\mu_1$. After being processed in an IVR, a call is either forwarded to a CSR for an assistance with probability $\eta$ or completes the service and departs from the call center with probability $1 - \eta$. A call finding an idle CSR is served immediately, otherwise it is blocked. The conversation with a CSR is assumed to follow another exponential distribution with mean $1/\mu_2$.

After being blocked at IVRs or CSRs, a call either gives up forever or enters an orbit to retry for service again after some exponentially distributed time with mean $1/\nu$. We assume that new calls and retrial calls blocked at IVRs retry with probabilities $\theta_1$ and $\theta_2$, respectively, while those blocked at CSRs retry with probability $\theta_3$.

III. PERFORMANCE ANALYSIS

A. Derivation of the stationary distribution

In this section, we analyze the model described in Section II. Let $I(t)$, $J(t)$ and $K(t)$ denote, respectively, the number of busy IVRs, the number of busy CSRs and the number of calls in the orbit at time $t \geq 0$. Under the assumptions made in Section II, it is easy to confirm that $\{X(t) = (K(t), I(t), J(t)): t \geq 0\}$ form an irreducible continuous-time Markov chain on the state space $S = \{0, 1, \ldots, c_1\} \times \{0, 1, \ldots, c_2\} \times \mathbb{Z}_+$, where $\mathbb{Z}_+$ represents the set of nonnegative integers. We sort the state space as $S = \{(l(0), l(1), \ldots)\}$, where $l(k) = \{(i, j, k); (i, j) \in \{0, 1, \ldots, c_1\} \times \{0, 1, \ldots, c_2\}\}$ $k \in \mathbb{Z}_+$ and $(i, j)$ is further arranged according to the lexicographic order. Under this arrangement of the state space, the infinitesimal generator $Q$ of $\{X(t); t \geq 0\}$ is given by

$$Q = \begin{bmatrix}
A_0 & B & O & \cdots & \cdots \\
C_1 & A_1 & B & O & \cdots \\
& C_2 & A_2 & B & \cdots \\
& & C_3 & A_3 & \cdots \\
& & & \cdots & \cdots
\end{bmatrix},$$

where the details of the block matrices are given in the appendix. This type of Markov chain is referred to as a level-dependent QBD process where the level and the phase are $K(t)$ and $(I(t), J(t))$, respectively.

It is well known that the ergodic condition for a general level-dependent QBD process is challenging and has not been
derived yet. In this paper, although the necessary and sufficient condition for the ergodicity of \( \{X(t)\} \) has not been derived, we assume that \( \lambda < c_1 \mu_1 (1-\eta) + c_2 \mu_2 \), which means that the arrival rate is smaller than the maximum departure rate in the stationary state. Under the ergodic condition, let \( \pi_{i,j,k} = \lim_{n \to \infty} \Pr(I(t) = i, J(t) = j, K(t) = k) \) \( ((i,j,k) \in S) \) denote the stationary distribution. Using the algorithm proposed in [9], we can efficiently compute the stationary distribution from the infinitesimal generator \( Q \). It is well known that the stationary distribution of a level-dependent QBD process can be expressed in terms of a series of rate matrices. In [9], a simple algorithm based on matrix continued fractions for these rate matrices is proposed.

### B. Performance measures

In this section, we derive some performance measures from the stationary distribution obtained in Section III-A.

- Let \( B_{\text{IVR}} \) and \( B_{\text{CSR}} \) denote the blocking probability at IVRs and that at CSRs, respectively. We have

\[
B_{\text{IVR}} = \Pr(I = c_1) = \sum_{0 \leq j \leq c_2, k \geq 0} \pi_{c_1,j,k},
\]

\[
B_{\text{CSR}} = \Pr(J = c_2) = \sum_{0 \leq i \leq c_1, k \geq 0} \pi_{i,c_2,k}.
\]

- Let \( L_{\text{IVR}} \) and \( L_{\text{CSR}} \) denote the mean number of busy IVRs and that of busy CSRs, respectively. We have

\[
L_{\text{IVR}} = E[I] = \sum_{(i,j,k) \in S} i \pi_{i,j,k},
\]

\[
L_{\text{CSR}} = E[J] = \sum_{(i,j,k) \in S} j \pi_{i,j,k}.
\]

We then obtain the utilizations of an IVR \( (U_{\text{IVR}}) \) and that of a CSR \( (U_{\text{CSR}}) \) by \( L_{\text{IVR}}/c_1 \) and \( L_{\text{CSR}}/c_2 \), respectively.

- Let \( L_{\text{orbit}} \) denote the mean number of customers in the orbit. \( L_{\text{orbit}} \) is given by

\[
L_{\text{orbit}} = E[K] = \sum_{(i,j,k) \in S} k \pi_{i,j,k}.
\]

- Let \( W \) denote the sojourn time that a customer spends from arrival to departure. We then have

\[
W = \frac{L_{\text{IVR}} + L_{\text{CSR}} + L_{\text{orbit}}}{\lambda},
\]

according to the Little’s law.

### IV. NUMERICAL EXPERIMENT

In this section, we investigate the impact of various parameters on the performance measures. In particular, we show the influence of \( \mu_1, c_1 \) and \( \eta \) on all the performance measures presented in Section III-B.

#### A. Parameter setting

We use the Algorithm 3 in [9] with \( k_n = 2^n \) in order to obtain an approximation to the stationary distribution. In the algorithm in [9], there are two parameters: truncation point \( N \) and \( \epsilon \). The truncation point \( N \) is the maximum level that we consider and \( \epsilon \) is used when computing the rate matrix at level \( N \). We fix \( \epsilon = 10^{-10} \) and select a sufficiently large \( N \). We also conduct a simulation in order to validate the selection of \( N \) and \( \epsilon \). In the simulation, the number of samples is 30 and the confidence interval is 95%. Fixed parameters are presented in Table I, where retrial probability is \( \theta_1 = \theta_2 = \theta_3 = \theta \). Variable parameters are shown in Table II. In the following graphs, analysis and simulation results are plotted together for each parameter set.

#### B. Utilization and blocking probability

Figure 3 shows the utilization of a CSR \( (U_{\text{CSR}}) \) and the blocking probability at CSRs \( (B_{\text{CSR}}) \) against \( \eta \) for several \( c_1 \) where \( \mu_1^{-1} = 1 \). We find that analytical results match the corresponding simulation ones well, implying the validation of the truncation point \( N \) and \( \epsilon \). This validation is further confirmed in the graphs below. We observe from Figure 3(a) that the utilization of a CSR is almost constant against \( \epsilon \) and is sensitive to \( c_1 \); this is because the arrival rate from IVRs to CSRs is nearly equal to \( \eta \lambda \) for any \( c_1 \). We observe from Figure 3(b) that the blocking probability at CSRs increases with \( \eta \) and is sensitive to \( \eta \). These observations suggest that keeping \( \eta \) small is a way to reduce the utilization of a CSR and the blocking probability at CSRs.

Figure 4(a) and 4(b) show the utilization of an IVR \( (U_{\text{IVR}}) \) and the blocking probability at IVRs \( (B_{\text{IVR}}) \), respectively, against \( c_1 \), in case of \( \eta = 0.2 \). We find from Figure 4(a) that when \( c_1 \) is small or \( \mu_1^{-1} \) is large, the utilization of an IVR is high. Furthermore, comparing Figure 4(a) and Figure 4(b), we observe that the higher utilization of an IVR \( (U_{\text{IVR}}) \), the larger the blocking probability at IVRs \( (B_{\text{IVR}}) \). This is because a small \( c_1 \) or a large \( \mu_1^{-1} \) imply a large load to IVRs, leading to the increase in the blocking probability at IVRs.

#### C. Mean number of calls in the orbit and mean sojourn time

Figure 5 shows the mean number of customers in the orbit against \( c_1 \) for \( \eta = 0.2 \). It is observed from Figure 5 that \( L_{\text{orbit}} \) increases with the increase in \( \mu_1^{-1} \) and the decrease in \( c_1 \).
This is because a small \( c_1 \) or a large \( \mu_1^{-1} \) imply a large load to IVRs. As a result, the blocking probability at IVRs is large, leading to the increase in the number of calls in the orbit.

Figure 6 shows the mean sojourn time for a call from arrival to departure (\( W \)). Figure 6(a) demonstrates the mean sojourn time (\( W \)) against \( \eta \) for \( c_1 = 50 \), while Figure 6(b) represents \( W \) against \( c_1 \) for \( \eta = 0.2 \). We read from Figure 6(a) that \( W \) increases proportionally in \( \eta \). This is because the number of calls served by a CSR is proportional to \( \eta \). We also find that \( W \) increases proportionally in \( \mu_1^{-1} \) since every customer needs service in an IVR with mean \( \mu_1^{-1} \).

Figure 6(b) shows the sojourn time against the number of IVRs (\( c_1 \)). We observe from Figure 6(b) that the mean sojourn time is almost constant with respect to \( c_1 \) except for only the case of \( \mu_1^{-1} = 2, c_1 = 30, \eta = 0.2 \). This is because when \( c_1 \) is large and \( \mu_1^{-1} \) is small, blocking at IVRs rarely occurs. As a result, the sojourn time is in fact the total of service times in IVRs and CSRs. However, when \( c_1 \) is small and \( \mu_1^{-1} \) is large, the fraction of calls that are blocked at IVRs increases, leading to some additional time in the orbit.

V. CONCLUSION

In this study, we have analyzed the performance of a call center with IVRs and redial customers. To this end, we have modeled a call center with IVRs and redials by a two-stage queueing system with retrial, formulating the model by a level-dependent QBD process. We have derived the performance measures such as utilizations, mean number of redialing calls, blocking probabilities and mean sojourn time. Numerical results have shown that under a low arrival rate of new calls, the mean busy rate of CSRs is insensitive to
the number IVRs and their processing times, while growing proportionally with the forwarding probability from IVRs to CSRs. This suggests that reducing the forwarding probability is effective to keep the utilization and the blocking probability by the followings.

APPENDIX

Block matrices of the infinitesimal generator $Q$ are given by the followings.

$$B = \begin{bmatrix} B^0 & \mathbf{O} \\ \mathbf{O} & B^0 \end{bmatrix}.$$ 

$$(B^0)_{i,i'} = \begin{cases} \lambda \theta_1, & (i,i') = (c_1,c_1), \\ 0, & \text{otherwise.} \end{cases}$$

$$(B^0)_{i,c_2} = \begin{cases} \lambda \theta_1, & (i,i') = (c_1,c_1), \\ 0, & \text{otherwise.} \end{cases}$$

$$(B^0)_{c_2,i'} = \begin{cases} \lambda \theta_1, & (i,i') = (c_1,c_1), \\ \mu_{i+1}\theta_2, & (i,i') = (i,i+1), 1 \leq i \leq c_1, \\ 0, & \text{otherwise.} \end{cases}$$

$$C_k = \begin{bmatrix} C^0_k & \mathbf{O} \\ \mathbf{O} & C^0_k \end{bmatrix}.$$ 

$$(C^0_k)_{i,i'} = \begin{cases} k \nu(1 - \theta_3), & (i,i') = (c_1,c_1), \\ k \nu, & (i,i') = (i,i+1), 0 \leq i \leq c_1 - 1, \\ 0, & \text{otherwise.} \end{cases}$$

$$(A^0)_{i,i'} = \begin{cases} \lambda, & (i,i') = (i,i+1), 0 \leq i \leq c_1 - 1, \\ \eta \mu_1(1 - 2 \eta), & (i,i') = (i,i-1), 1 \leq i \leq c_1, \\ 0, & \text{otherwise.} \end{cases}$$

$$(A^0_{c_2})_{i,i'} = \begin{cases} \lambda, & (i,i') = (i,i+1), 0 \leq i \leq c_1 - 1, \\ \eta \mu_1(1 - 2 \eta), & (i,i') = (i,i-1), 1 \leq i \leq c_1, \\ 0, & \text{otherwise.} \end{cases}$$

$$(A^+)_{i,i'} = \begin{cases} \lambda \mu_1, & (i,i') = (i,i-1), 1 \leq i \leq c_1, \\ 0, & \text{otherwise.} \end{cases}$$

$$(A^c)_{i,i'} = \begin{cases} \lambda \mu_2, & (i,i') = (i,i), 0 \leq i \leq c_1, \\ 0, & \text{otherwise.} \end{cases}$$

For a square matrix $X$, let $D_X$ denote the diagonal matrix whose diagonal elements are equal to the sums of the rows of $X$. We have $A_0 = A - D_A - D_B$ and $A_k = A - D_A - D_B - D_{C_k}$.

REFERENCES


Fig. 6. Mean sojourn time