Performance Evaluation 80 (2014) 43–62

Contents lists available at ScienceDirect

Performance Evaluation

journal homepage: www.elsevier.com/locate/peva

Performance analysis of call centers with abandonment, retrial and after-call work

Tuan Phung-Duc [a], Ken'ichi Kawanishi [b,*]

[a] Department of Mathematical and Computing Sciences, Tokyo Institute of Technology, Ookayama, Tokyo 152-8552, Japan

[b] Division of Electronics and Informatics, Gunma University, 1-5-1 Tenjin-cho, Kiryu, Gunma 376-8515, Japan

ARTICLE INFO

Article history:
Available online 25 March 2014

Keywords:
Multiserver retrial queue
Abandonment
After-call work
Call center
Human behaviors
Level-dependent QBD process

ABSTRACT

This paper considers a multiserver queueing model with abandonment, retrial and after-call work for call centers. Upon a phone call, customers that find a free call line occupy the line immediately while those who see all the call lines busy are blocked and join an orbit. Customers holding a call line are served according to the first-come first-served discipline. After completing a call, the customer leaves the system while the server must start an after-call work and the call line is released for a newly arrived customer. Waiting customers may abandon after some waiting time and then either join the orbit or leave forever. Customers in the orbit retry to hold a free call line after some time. We formulate the queueing system using a continuous-time level-dependent quasi-birth-and-death process for which a sufficient condition for the ergodicity is derived. We obtain a numerical solution for the stationary distribution based on which performance measures such as the waiting time distribution and the blocking probability are derived. Using Little's law, we obtain explicit formulae which verify the accuracy of the numerical solution. We compare our model with some simpler models which do not fully take into account some human behaviors. The comparison shows significant differences implying the importance of our model. Numerical results show various insights into the performance of call centers.

1. Introduction

Nowadays, call centers are one of the core parts for the customer support of companies and organizations because they serve as a channel for direct communication with customers. Examples for call centers include telephone shopping, ticket reservation and telephone banking. Recently, call center business is also important because it provides a large amount of employment in many countries. Therefore, optimal design and management of call centers are interesting from both theoretical and practical points of view. Call centers have been attracting much attention of researchers from both academia and industry where a large number of papers have been published and a huge number of patents have been granted. For an extensive survey of the state-of-the-art in call centers, we refer to [1,2] and the references therein. From a modeling point of view, call centers can be naturally seen as queueing systems where call agents and calls correspond to servers and customers, respectively. Because the running cost is dominated by the labor cost of call agents, suitable scheduling and staffing policies balancing the cost with the quality of service (QoS) are indispensable for the management of call centers. To this

* Corresponding author. Tel.: +81 277301838.
E-mail addresses: tuan@is.titech.ac.jp (T. Phung-Duc), kawanisi@cs.gunma-u.ac.jp (K. Kawanishi).

http://dx.doi.org/10.1016/j.peva.2014.03.001
0166-5316/

end, we need to develop a queueing model which captures the behaviors of customers as well as call agents for a careful design of call centers. In this paper, we consider such a queueing model with retrial, abandonment and after-call-work for call centers.

Retrial queues are characterized by the fact that arriving customers that find the service facility fully occupied enter an orbit and retry for their luck after some random time. Recently, retrial queues have been paid much attention to because they have applications in various telecommunication systems, service systems and call centers [2–4]. The authors in [2–4] state that retrial phenomena cannot be disregarded in a careful design of these systems. Furthermore, numerical results in [5] show that there is a large difference between the blocking probability obtained by a multiserver retrial queue and that computed by an equivalent loss model when the retrial rate is large.

In call centers, *after-call work* is an additional operation that should be done by a call agent immediately after finishing a call. An after-call work (also known as post-call activity and wrap-up) includes entering or updating data into the customer database to complete the transaction. It should be noted that a call agent cannot answer a new call while handling an after-call work; however a call line is released. As a result, an arriving call can occupy the released call line in order to wait for a free call agent. If the system capacity (i.e., the number of call lines) is infinite, the after-call work can be regarded as a part of the service time. However, since the system capacity is limited in real world call centers, the blocking probability and other performance measures are influenced by the after-call work. In fact, the effects of the after-call work on the performance of queueing models are discussed by several authors [6–8]. The authors of these papers conclude that the blocking probability computed by a queueing model with after-call work is smaller than that obtained by the corresponding queueing model where the duration of after-call work is included in the service time.

Abandonment of customers is another typical phenomenon in call centers, which occurs when a customer has to wait to connect to a call agent [4]. In particular, we consider a situation where a customer makes a call while all the call agents are busy with either a call or an after-call work. In such a case, if a free call line is available, the customer can hold the line in order to wait for a free call agent instead of being blocked. The customer may abandon receiving service if the waiting time is too long. Managers of call centers wish to keep the abandonment rate as small as possible under a minimal staffing level (number of call agents).

There are a number of interesting papers dealing with multiserver queues with abandonments [3,9–13]. Garnett et al. [9] and Mandelbaum et al. [11] extend the so-called Halfin–Whitt regime for M/M/*s* queues with abandonment. These authors derive some simple approximate formulae for the staffing level of call centers with a large number of call agents. Whitt [10] carries out a sensitivity analysis of M/M/*s* queues with abandonment. Artalejo and Pla [3] and Shin [12] analyze the stationary distribution for multiserver retrial queues with infinite capacity waiting room and orbit using a direct truncation method and a generalized truncation method, respectively. Wuchner et al. [13] investigate multiserver retrial queues with finite population, impatient customers, balking and orbital search using the modeling language MOSEL. We refer to [14] for an extensive survey of queueing models with impatient customers.

In all the work mentioned above, abandonment, retrial and after-call work are separately considered. However, in practice, retrial, after-call work and abandonment coexist in a call center and they influence each other. Therefore, these phenomena should be taken into account concurrently in order to obtain an accurate performance evaluation of a call center. In our previous work [15], we have proposed and analyzed a retrial queueing model with after-call work. In this paper, we extensively extend our work to take abandonment into account in order to quantify the mutual effects of these phenomena on the performance of call centers. There is no doubt that the consideration of approximate analysis for large scale call centers is interesting and important [9–11,16]. This paper aims at an accurate modeling and analysis of call centers with arbitrary number of servers and call lines taking into account the most important human behaviors of customers and call agents. To the best of our knowledge, such a detailed model has never been investigated in queueing and call center literature.

The main contribution our paper is to provide an exhaustive stationary analysis of a detailed but numerically tractable model taking into account the most important human behaviors in call centers. We formulate the queueing system by a three-dimensional Markov chain which is a level-dependent quasi-birth-and-death (QBD) process, where the level is referred to as the number of customers in the orbit. Although our model is complicated, we are able to establish an ergodic condition under which the stationary distribution exists. We derive a sufficient condition for the ergodicity of the Markov chain by exploiting the special structure of the model and by using the approach of Diamond and Alfa [17]. Under the ergodic condition, we analyze the stationary distribution of the Markov chain. As is well known, analytical solutions for the stationary distributions of retrial queues are difficult and are obtained in a few special cases [18].

In this paper, we focus on a numerical solution to the stationary distribution from which we compute some performance measures. In particular, we obtain a numerical solution to the stationary distribution of our level-dependent QBD process by the direct-truncation algorithm presented in [19] whose truncation point is determined based on that of [15]. The numerical solution is highly accurate as it satisfies various explicit formulae obtained by Little's law. We obtain various performance measures beyond the stationary distribution which is the main objective of our previous work [15]. In particular, we derive the waiting time distribution which is the most important performance measure of call centers. The waiting time distribution is derived using a two-dimensional absorbing Markov chain following the methodology in [20,21].

In addition, we derive various explicit formulae expressing the relations between performance measures by Little's law. These explicit formulae are used to validate the numerical solution and to obtain some performance measures which cannot directly measured in real-world call centers such as the average number of retrials per customer. We compare our