Computers & Industrial Engineering 96 (2016) 61–71

Contents lists available at ScienceDirect

Computers & Industrial Engineering

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

Priority tandem queueing system with retrials and reservation of channels as a model of call center

Chesoong Kim [a,*], V.I. Klimenok [b,*], A.N. Dudin [b]

[a] Department of Industrial Engineering, Sangji University, Wonju, Kangwon 220-702, Republic of Korea
[b] Belarusian State University, 4, Nezavisimosti Ave., Minsk 220030, Belarus

ARTICLE INFO

Article history:
Received 4 March 2015
Received in revised form 19 February 2016
Accepted 12 March 2016
Available online 22 March 2016

Keywords:
Tandem retrial queue
Impatience
Priority
Servers reservation
Optimization

ABSTRACT

We study a tandem queueing system consisting of two multi-server stations and finite intermediate buffer. Customers arrive at the first station of the tandem according to a Markovian Arrival Process. The first station does not have a buffer. Customers, who do not succeed to enter the service immediately upon arrival, retry for the service after a random amount of time. After service completion at the first station, a customer leaves the system permanently or moves for the service at the second station. A part of customers entering the second station have a priority over other customers. The priority is provided by means of reservation of a part of servers for the service of priority customers only. Usually, non-priority customers are not allowed to occupy the reserved servers. However, if the queue of non-priority customers in the intermediate buffer becomes larger than some preassigned threshold while there are free servers, a non-priority customer is picked-up for the service. Customers staying in the buffer are impatient. Priority and non-priority customers have different patience time and may leave the system or return to the first station if the patience time expires. The system is analyzed in steady state. A condition for existence of the stationary regime in the system is derived, the steady state distribution and various performance measures of the system are calculated, some illustrative numerical examples are discussed. A tandem queue under consideration is suitable, e.g., for modeling call centers with Interactive Voice Response Machines. Analysis presented in this paper was implemented in borders of the applied project funded by one of the banks.

1. Introduction

At the present days, call centers are complex socio-technical systems that play an increasingly important role in the modern world society. They are used to provide services in information and emergency centers, help-desks, tele-marketing, etc. To offer high quality services, call center managers and designers should consider the complex of the factors associated with arrivals of customers at random instants and a variety of customer requirements for quality of the service. Queueing models can be effectively used for call centers design and support of their management. The surveys of research works devoted to mathematical modeling of call centers can be seen in the papers by Aksin, Armony, and Mehrotra (2007), Jouini, Dallery, and Aksin (2009), Jouini, Pot, Koole, and Dallery (2010), Khudyakov, Feigin, and Mandelbaum (2010) and references therein. In this paper, we consider a tandem queueing system that takes into account such important features of call centers as: possibility that a customer may need service from several sequentially located servers; the presence of multi-class customers with different Quality of Service (QoS) requirements; the impatience of customers and the retrial phenomena.

Tandem queueing systems can be used for modeling real-life queueing networks as well as for validation of general decomposition algorithms in networks. These systems have found great attention in the literature, since the practical importance and mathematical complexity of such systems make them attractive for researchers in the field of queuing theory and its applications, see, e.g., the papers by Balsamo, Persone, and Inverardi (2003) and Perros (1989).

Retrial queueing systems differ from the systems with waiting room or losses in the fact that a customer that does not succeed to get an access to the service facility immediately upon arrival neither enters a buffer, nor leaves the system permanently. He/she enters the so-called orbit (the virtual room for such customers) from which he/she makes repeated attempts to reach the service

* Corresponding authors.
E-mail addresses: dowoo@sangji.ac.kr (C. Kim), vklimenok@yandex.ru (V.I. Klimenok), dudin@bsu.by (A.N. Dudin).

http://dx.doi.org/10.1016/j.cie.2016.03.012
0360-8352/

facility in a random amount of time. There are many publications in the field of queueing theory and telecommunications devoted to investigation of retrial queues. The state of the art in research in the field of retrial queueing systems is partially presented by Gomez-Corral (2006), Artalejo and Gomez-Corral (2008), Artalejo (2010) and Kim and Kim (2016).

In the literature, there is only a relatively small number of publications dealing with tandem queues with retrials, although the effect of retrials is an integral part of many real telecommunication systems and ignorance of this effect may lead to significant errors in the design and evaluation of their performance. To the best of our knowledge, tandem queuing systems with retrials were considered only in the papers by Falin (2013), Gomez-Corral and Martos (2002), Kim, Klimenok, and Taramin (2010a), Kim, Park, Dudin, Klimenok, and Tsarenkov (2010b), Klimenok and Savko (2013), Moutzoukis and Langaris (2013) and Phung-Duc (2012). The papers by Falin (2013) and Phung-Duc (2012) deal with exponential retrial tandem queue consisting of two single-server stations without intermediate buffer. For this quite simple queue, the explicit formulas for steady-state probabilities are obtained. Moutzoukis and Langaris (2013) have considered the system with stationary Poisson input, single-server stations and constant intensity of repeated attempts from the orbit. In the work by Klimenok and Savko (2013), the tandem queue with stationary Poisson flow, two types of customers and reservation of channels for the priority customers at the second station is studied. Gomez-Corral and Martos (2002) deal with queueing system with Markovian Arrival Process (*MAP*), single-server stations and phase type service time distributions. In the paper by Kim et al. (2010b), more general model is under study. The arrival flow is a *BMAP*, the service time distribution at Station 1 is arbitrary. The paper by Kim et al. (2010a) is devoted to the tandem retrial queue with *BMAP*, single-server first station, an additional flow at multi-server second station and reservation of channels for the customers from this flow.

The phenomena of impatience of customers is an important feature of many telecommunication systems including call centers and contact centers. For references relating to application of queueing models with the impatience to analysis of call centers we refer the reader to the works by Dudin, Kim, and Dudina (2013), Garnett, Mandelbaum, and Reiman (2002), Jouini et al. (2010), Kim, Dudin, Dudin, and Dudina (2013) and Roubus and Jouini (2013) and references therein. But we know only the paper by Klimenok and Savko (2013) where a tandem queue with retrials and the impatience of customers is dealt.

In the present paper, we consider a tandem consisting of two stations, Station 1 and Station 2. Station 1 is represented by a multi-server queue with retrials. A customer who finds all lines busy upon arrival joins the orbit and retries for the service after a random amount of time independently of other orbital customers. Such a behavior is typical for customers calling to a call center and receiving a signal that all trunk lines are busy.

All incoming calls at Station 1 (primary and repeated), can finish the service in the tandem system at this (first) stage or require an additional service at Station 2 (second stage). The first stage can be considered, e.g., as dial phase or IVR-Interactive Voice Response (for definition see the papers by Khudyakov et al. (2010) and by Dudin et al. (2013)). Unsatisfied clients are directed to the agents at the second stage to resolve their problems. Kim and Park (2010) propose also the following interpretation of two-stage service in a call center. All customers are initially handled by the agents at the first stage of call center (servers of Station 1) and a part of the customers can be satisfied with the service at this stage. However, the service of some customers cannot be completed by these agents due their limited authority and knowledge. These customers are transferred to experts at the second stage who can handle them.

Station 2 is represented by a multi-server queue with a finite buffer. We assume that all agents (servers of Station 2) are flexible enough to answer all requirements of service. But the customers to be served at Station 2 are divided into two different classes according to their ability to wait for the connection to the agent. We assume that the company that owns the call center provides preferences to high-valued clients who are more impatient while staying in the buffer. To prevent the loss of most of these clients, management of call center may decide that some group of agents will be reserved for service of high-valued (priority) clients only.

In this paper we study the operation of the system in steady state and discuss the question "how many agents should be reserved to provide the best QoS for various types of customers in the call center?"

The rest of the paper is organized as follows. In Section 2, the queueing system under consideration is defined. In Section 3, the multi-dimensional continuous time Markov chain, which describes the behavior of this system, is constructed. The generator of this chain is presented and the fact that this Markov chain belongs to the class of the asymptotically quasi-Toeplitz Markov chains is proved. In Section 4, the ergodicity condition of this Markov chain is derived. The algorithm for computing the stationary probabilities is outlined in Section 5. In Section 6, formulas for some key performance measures of the system are given. Numerical examples are presented and an optimization problem is formulated and solved in Section 7. Section 8 concludes the paper.

2. Model description

We consider a tandem queueing system consisting of two stations in series, Station 1 and Station 2. The structure of the system is presented in Fig. 1.

Station 1 is represented by the N-server retrial queue without a buffer. Station 2 is represented by $(K+R)$-server queue with a finite buffer of capacity M.

Customers arrive at Station 1 according the Markovian Arrival Process (MAP). The *MAP* is governed by some underlying process $\nu_t, t \geqslant 0$, which is an irreducible continuous time Markov chain with state space $\{0, 1, .., W\}$. The transition rates of this chain transitions, which are accompanied by an arrival (no arrival) of a customer, are combined into the matrices $D_1(D_0)$ of size $(W+1)\times(W+1)$. The matrix $D = D_0 + D_1$ is the infinitesimal generator of the Markov chain $\nu_t, t \geqslant 0$. The vector $\boldsymbol{\theta}$ of the stationary state distribution of the chain is the unique solution of the system $\boldsymbol{\theta} D = \mathbf{0}, \boldsymbol{\theta}\mathbf{e} = 1$, where $\mathbf{0}$ is a zero row vector and $\mathbf{e}$ is a column vector consisting of $1's$. The average intensity λ(fundamental rate) of the *MAP* is defined as $\lambda = \boldsymbol{\theta} D_1 \mathbf{e}$. The coefficient of variation c_{var} of intervals between arrivals is defined by $c_{var}^2 = 2\lambda\boldsymbol{\theta}(-D_0)^{-1}\mathbf{e} - 1$. The coefficient of correlation c_{cor} of the successive intervals between arrivals is given by

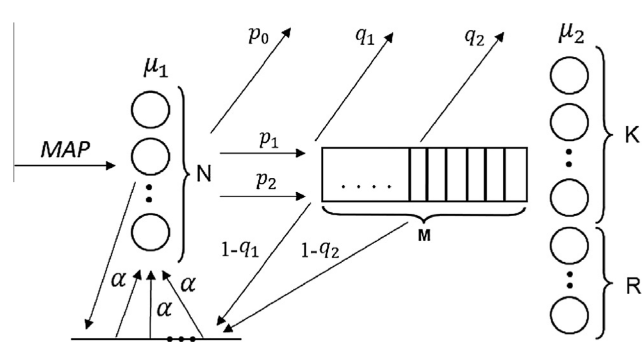

Fig. 1. The structure of the system.