Applied Mathematical Modelling 50 (2017) 1–16

Contents lists available at ScienceDirect

Applied Mathematical Modelling

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

Retrial queue with discipline of adaptive permanent pooling

A. Brugno[a], A.N. Dudin[b,c,*], R. Manzo[a]

[a] Department of Information Engineering, Electrical Engineering and Applied Mathematics, University of Salerno, Via Giovanni Paolo II, 132, 84084 Fisciano, SA, Italy
[b] Belarusian State University, 4, Nezavisimosti Ave., Minsk 220030, Belarus
[c] Peoples Friendship University of Russia, 6, Miklukho-Maklaya st., 117198 Moscow, Russia

ARTICLE INFO

Article history:
Received 11 October 2015
Revised 12 April 2017
Accepted 11 May 2017
Available online 19 May 2017

Keywords:
Retrial queueing system
Group service
Permanent pooling
Stationary distribution
Optimization

ABSTRACT

A novel customer service discipline for a single-server retrial queue is proposed and analysed. Arriving customers are accumulated in a pool of finite capacity. Customers arriving when the pool is full go into orbit and attempt to access the service later. It is assumed that customers access the service as a group. The size of the group is defined by the number of customers in the pool at the instant the service commences. All customers within a group finish receiving the service simultaneously. If the pool is full at the point the service finishes, a new service begins immediately and all customers from the pool begin to be served. Otherwise, the customer admission period starts. The duration of this period is random and depends on the number of customers in the pool when the admission period begins. However, if the pool becomes full before the admission period expires, this period is terminated and a new service begins. The system behaviour is described by a multi-dimensional Markov chain. The generator and the condition of ergodicity of this Markov chain are derived, and an algorithm for computing the stationary probability distribution of the states of the Markov chain is given. Formulas for computing various performance measures of the system are presented, and the results of numerical experiments show that these measures essentially depend on the capacity of the pool and the distribution of the duration of the admission period. The advantages of the proposed customer service discipline over the classical discipline and the discipline in which customers cannot enter the pool during the service period are illustrated numerically.

1. Introduction

The phenomenon of customer retrials in the case when a customer cannot be served upon arrival is typical in many real-world systems, particularly in telecommunications. Hence, retrial queues represent an important and popular subject of analysis in the theory of queues; see, e.g. the surveys in [1–5] and some recent papers [6,7] .

In the analysis of single-server retrial queues, it is usually assumed that when an arriving customer meets an idle server, service begins immediately. Otherwise, the arriving customer moves to some virtual place called an orbit and attempts to obtain service at a later time. Customers in the orbit make repeated attempts to be served after random periods of time, independently of other customers in orbit. Therefore, a very typical situation in retrial queueing models is the following: after each service completion, the server is idle during some random time while customers wait in orbit.

* Corresponding author at: Belarusian State University, 4, Nezavisimosti Ave., Minsk 220030, Belarus.
E-mail addresses: abrugno@unisa.it (A. Brugno), dudin@bsu.by, dudin_alex@yahoo.com (A.N. Dudin), rmanzo@unisa.it (R. Manzo).

http://dx.doi.org/10.1016/j.apm.2017.05.019

In a recent paper [8], a new and effective discipline of customer service in retrial queues was proposed. This discipline assumes that service to customers is provided in groups, but not individually. After completing the service of a group, the system starts the so-called customer admission period. The duration of this period is random and service is not provided until this period expires. During the admission period, all customers arriving from the outside (primary customers) and from orbit (retrial customers) are accumulated in the so-called *pool*, which is some (probably virtual) place having a finite capacity, until the pool becomes full. Once the pool is full, all arriving customers go into orbit. When the admission period expires, all customers stockpiled in the pool start being served simultaneously. The duration of the service time depends on the number of customers in the serviced group. Customers arriving at the system during service go into orbit.

Such a service discipline is realistic in some real-world systems, in particular, telecommunication networks with multicast communication, transportation systems with group customer delivery, and electronic commerce systems for delivering goods to customers. In such systems, a group of requests generated by customers can often be processed contemporaneously in parallel and the processing of the entire group is considered to have finished if the service of all peculiar requests allocable to the group is completed. Hence, the service time of a group is the maximum of the independent service times of customers allocable to this group. It is known that the mean value of the maximum of several independent random variables can be much less than the sum of the mean values of these variables. Therefore, the expectation of the time dedicated to servicing an individual customer (defined as the ratio of the average service time of a group to the average group size) under the proposed service discipline may be essentially smaller than under the classical discipline. Hence, the throughput of the system is higher, and other performance measures of the system may be better, than in the classical service discipline.

In this paper, we propose and quantitatively justify a more effective discipline of customer admission and service than that considered in [8]. We introduce three significant improvements:

- It is assumed in [8] that service is resumed only after the admission period expires (even if the pool becomes full earlier). This assumption holds true for many real-world systems in which the admission period cannot be terminated ahead of schedule, e.g. in polling systems, where the server indeed provides service to many queues and to other queues during the admission period. This is also true in systems where a certain sequence of technological operations should be implemented during the admission period and no single operation can be cancelled. However, in some other real-world systems, it is possible to terminate the admission period when the pool becomes full. Here, we deal with the case where the termination of the admission period is possible. In the model considered in [9], it is also assumed that the termination of the admission period is possible.
- In [9] and in [8], it is assumed that the admission of customers can only start when service finishes. During the service period, customers are not admitted to the pool and go into orbit. Here, we assume that customers can be admitted to the pool during the service period. Therefore, if the pool becomes full during the service, a new service starts immediately after the current one has been completed.
- In [9] and in [8], the pool is always empty at the moment of service completion. Thus, the duration of admission periods is an independent and identically distributed random variable. In the model under study, it is possible that the pool will not be empty at the moment of service completion. We assume that the duration of the admission period depends on the number of customers presenting in the pool at the beginning of this period.

The proposed improvements in our model make the analytical and numerical analysis more technically difficult than for [9] and [8] because of the higher dimension of the Markov chain that describes the behaviour of the system and the presence of more non-zero blocks in the generator of the chain.

Because the considered model describes the operation of many real-world systems, but has not yet been analysed in the queueing literature, the results presented in this paper have great practical importance. Even small improvements in the customer service discipline can lead to essential increases in system performance. Our results help to optimally adjust the parameters of the service discipline (capacity of the pool and distributions of the admission period and service time) as well as the parameters of the arrival process.

The systems with joint service of an entire group of customers, but not of individual service, were analysed in the literature as a bulk service discipline more than 60 years ago [10]. Currently, there are a great number of studies devoted to bulk queues. We restrict ourselves to mentioning the work of Neuts [11], Chaudhry and Templeton [12], and Chakravarthy [13–17]. For references, see also the recent paper [18]. In these papers, it is usually assumed that service can be provided to groups of size at least L and at most K where L and $K, K \geq L$, are certain thresholds. The essential difference between our model and the papers cited above is the following. We allow service to start after a certain random time (admission period), even if the number of customers ready for service does not reach the threshold value. This makes the admission discipline more flexible and provides a better quality of service to customers. Last, but not least, practically all papers in the field of bulk queues consider systems with finite or infinite buffers. We deal with customer retrials. The Markov chain describing the behaviour of the considered system is not space homogeneous, and so the well-known and powerful tool for the study of multi-dimensional Markov chains, namely, the theory of Quasi-Birth-and-Death Processes, is not applicable. Note that queues with bulk service and the queueing system introduced in this paper are in some sense very close to the so-called queues with assembly-like service [19]. In queues with assembly-like service, service only starts when all existing buffers in the system are occupied. Thus, the set of places at the head of each buffer can be interpreted as a pool.