Computers & Operations Research 46 (2014) 69–77

Contents lists available at ScienceDirect

Computers & Operations Research

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

An improved truncation technique to analyze a *Geo/PH/1* retrial queue with impatient customers

Sherif I. Rabia

Department of Engineering Mathematics and Physics, Faculty of Engineering, Alexandria University, Alexandria 21544, Egypt

ARTICLE INFO

Available online 21 January 2014

Keywords:
Retrial queues
Discrete-time setting
Impatient customers
Truncation approaches
Matrix decomposition

ABSTRACT

This paper considers a discrete-time retrial queue with impatient customers. We establish the global balance equations of the Markov chain describing the system evolution and prove that this queueing system is stable as long as the customers are strict impatient and the mean retrial time is finite. Direct truncation with matrix decomposition is used to approximate the steady-state distribution of the system state and hence derive a set of performance measures. The proposed matrix decomposition scheme is presented in a general form which is applicable to any finite Markov chain of the *GI/M/1*-type. It represents a generalization of the Gaver–Jacobs–Latouche's algorithm that deals with QBD process. Different sets of numerical results are presented to test the efficiency of this technique compared to the generalized truncation one. Moreover, an emphasis is put on the effect of impatience on the main performance measures.

1. Introduction

In the classical queueing theory [1,2], a customer arriving at a full system departs immediately from the system and has no further influence on it. In many applications (such as establishing a telephone call between two consumers or starting a communication between two nodes in a network), those blocked customers will return back after some random time to retry getting the required service. Following the standard terminology, such customers are assumed to join a hypothetical place called "orbit" from which a stream of retrials arrives at the system besides the original stream of customers. Tackling such phenomenon and analyzing its effect on the performance of queueing systems is the issue of the theory of retrial queues. The mathematical aspects of this theory are explained in [3] while computational procedures are described in [4]. A number of survey papers appeared monitoring the work done in retrial queues (e.g., [5–7]).

Research in the area of retrial queues focused mainly on the continuous-time setting. However, communication systems such as (Asynchronous Transfer Mode) ATM-based systems work in slotted-time setting where time axis is divided into slots and events may occur only at the boundaries of these slots [8,9].

Yang and Li [10] gave the first attention to discrete-time retrial queues. They analyzed a *Geo/G/1* retrial queue with geometric retrial times and were able to derive the generating function of the distribution of number of customers in the system, develop recursive formulas for extracting such distribution and prove that the continuous-time *M/G/1* retrial queue can be approximated by a discrete-time *Geo/G/1* retrial queue.

Different papers appeared applying a similar methodology to analyze different queueing features such as batch arrivals [11], general retrial times [12–14], server breakdowns [13,15,16], multiplicative retrial policy [17], negative customers [18], DBMAP arrivals [19] and customer collision with preemptive resume [20]. The numerical inversion and the maximum entropy techniques were employed to improve the numerical computations in [21]. Simulation was used to analyze the multiple servers case with a finite customer population [22]. An algorithmic treatment of a *Geo/Geo/c* retrial queue was presented in [23]. A study directed to investigate the distributions of successful and blocked retrials/external arrivals during a busy period appeared in [24]. The reader is referred to the reference list of these papers for related work.

All these models assume that customers are persistent. They continue retrying until the required service is obtained. In many applications (such as call centers and communications systems), some customers may choose to leave the orbit (i.e., stop retrying) after a number of unsuccessful retrials and return back to the infinite population; perhaps to initiate a new arrival later on. In this case customers are impatient. The special case of balking customers was analyzed in [25–28]. A simulation-based analysis that considers the general impatience case appeared in [29].

The first main issue of the present work is to extend the theory of discrete-time retrial queues to include the impatient customers case and to solve the added analysis complexity due to impatience.

E-mail addresses: sherif.rabia@alexu.edu.eg, shrfrabia@hotmail.com

0305-0548/$ - see front matter © 2014 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.cor.2014.01.004

As noted in [23], the retrial phenomenon causes a spatial heterogeneity in the underlying process which complicates the analysis of retrial queues. Analyzing discrete-time retrial queues has an added complication. In continuous-time setting, the occurrence of multiple events at the same time epoch has a zero probability. In discrete-time setting, this event has a positive probability. Hence, analyzing such systems is a challenging problem. These complications increase when customers are impatient. In the persistent customers case (with a single arrival stream), system transitions are of the birth-death type in the sense that the number of customers in the orbit increases or decreases by at most one during any time slot. When customers are impatient, multiple departures from the orbit may occur during a single time slot and hence the system evolution is no longer birth-death (although it may still be Markovian).

The second main issue of the present work is to generalize the Gaver–Jacobs–Latouche's algorithm [30]. Such algorithm is a variation of block-Gaussian elimination which is applicable to finite QBD (Quasi-Birth-and-Death) processes. Here, we extend the applicability of this algorithm to the finite Markov chains of the *GI/M*/1-type.

We consider a *Geo/PH*/1 retrial queue with impatient customers and geometric retrial times. We establish the global balance equations of the Markov chain describing the system evolution and prove its stability condition. It is shown that the system with strict (really) impatient customers is stable as long as the mean retrial time is finite. Direct truncation approach [3,4] is used to build up an algorithmic approximation of the steady-state distribution of this Markov chain and hence derive a set of performance measures. The special structure of the transition probability matrix is utilized to construct an efficient matrix decomposition scheme which is in fact applicable to any finite Markov chain of the *GI/M*/1-type. This scheme is discussed in both pure algebraic and probabilistic perspectives. A numerical study is presented to compare the efficiency of this technique with the generalized truncation technique and to investigate the effect of impatience on the system performance.

As far as we know, the present work is the first one to consider level-dependent Markov chains of the GI/M/1-type. Recently, a couple of papers appear treating the level-dependent QBD processes. For example, in [31] the authors consider an infinite level-dependent QBD process where not only the transition probabilities changes between levels but also the number of states increases as the level number increases. Moreover, this model is heavy-tailed in the sense that a great part of the probability mass function is distributed far away from the lower levels. In [32], the authors consider an infinite level-dependent QBD process where sums of Kronecker products are used to represent the non-zero blocks of the transition rate matrix. An analysis and an extensive numerical study are undertaken to arrive at an optimal memory usage. We hope that the present work opens the door for analyzing more-involved level-dependent Markov chains of the GI/M/1-type as done for the level-dependent QBD processes.

The rest of this paper is organized as follows. In Section 2, we give a detailed description of the queueing system under investigation and introduce some notations that will be used throughout this work. In Section 3, we present the global balance equations of the two-dimensional Markov chain describing the evolution of our queueing system and investigate its stability condition. Section 4 describes the details of the direct truncation technique where Section 5 is devoted to the generalized truncation technique. Numerical results are presented in Section 6. We conclude this work in Section 7.

2. System description

We consider a discrete-time retrial queue with impatient customers. In discrete-time setting, it is assumed that the time axis is divided into slots. Systems events occur only at the boundaries of these slots. Since multiple events may occur at the same time epoch, some order of occurrence must be imposed for the system state to be completely identified at any time epoch. Throughout this work, it is assumed that the system evolution is controlled by the early arrival scheme which is also known as the departure first rule [9]. According to this scheme, departures are assumed to take precedence over arrivals. More specifically, at the time epoch m, it is assumed that potential arrivals (from outside and from the orbit) occur in the interval (m, m^+) whereas potential departure takes place in the interval (m^-, m). As will be explained shortly, this scheme is extended here to allow for abandonments (see Fig. 1).

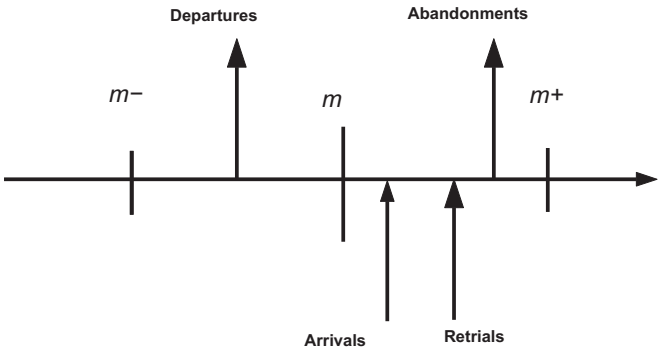

Fig. 1. Early arrival scheme with abandonments.

Customers arrive according to a geometrical arrival process. At the beginning of each time slot, a customer arrives at the system with probability p. Hence, time between arrivals (inter-arrival time) has a geometric distribution with parameter p. If the server is idle upon his arrival, the customer proceeds immediately to get the required service. Service time has a phase-type (PH) distribution [33] with H phases and representation ν and T where ν is an initial probability sub-vector of length H and T is a substochastic matrix for the transient states $\{1, 2, \ldots, H\}$. We assume that $I-T$ is non-singular which means that absorbing into state 0 is certain. Following standard notations [33], the column vector t of length H is defined as $t = e - Te$ which represents the absorbing probability into state 0 from any state i, $1 \le i \le H$, where e is used to denote a unit column vector of an appropriate length. The mean service time $1/\mu$ is given by $\nu(I-T)^{-1}e$.

A customer arriving at a busy server joins the orbit with probability α or departs immediately without being served with the complementary probability $1-\alpha$. Customers in the orbit (orbiting customers) make retrials independently of each other. During any time slot, an orbiting customer makes a retrial with probability $1-r$. Hence, time between retrials (retrial time) is geometrically distributed with parameter $1-r$. Upon making a retrial, if the customer finds the server busy, he chooses between joining the orbit again (with probability α) and departing from the system without being served (with probability $1-\alpha$). We assume that an arriving customer has access to the server before returning customers. Moreover, if more than one customer make retrials during the same time slot and the server is idle, one of them is selected at random and is allowed for service while the other customers see the server busy. The order of different events is shown in Fig. 1.

Inter-arrival times, service times and retrial times are mutually independent. We use the notation $\overline{x}$ to denote $1-x$, e. g., $\overline{p} = 1-p$, $\overline{r} = 1-r$, $\overline{\alpha} = 1-\alpha$. It is assumed that $0 < p \le 1$, $0 \le r < 1$ and $0 < \alpha \le 1$. When $\alpha < 1$, we say that customers are strict (really) impatient.

3. Mathematical model

The system state at the time epoch $m^+, m \ge 0$, is given by $X_m = (V_m, N_m)$ where V_m is the server state (0 if the server is idle or