Performance Evaluation 85–86 (2015) 52–70

Contents lists available at ScienceDirect

Performance Evaluation

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

A discrete-time queue with customers with geometric deadlines

Herwig Bruneel, Tom Maertens *

SMACS Research Group, Department of Telecommunications and Information Processing, Ghent University - UGent, Belgium

ARTICLE INFO

Article history:
Received 21 December 2012
Received in revised form 8 December 2014
Accepted 19 January 2015
Available online 7 February 2015

Keywords:
Queueing
Discrete-time
Deadlines
Closed-form results
Polynomial approximation

ABSTRACT

This paper studies a discrete-time queueing system where each customer has a maximum allowed sojourn time in the system, referred to as the "deadline" of the customer. More specifically, we model the deadlines of the consecutive customers as independent and geometrically distributed random variables. Customers enter the system according to a *general* independent arrival process, i.e., the numbers of arrivals during consecutive time slots are i.i.d. random variables with arbitrary distribution. Service times of the customers are deterministically equal to one slot each. For this queueing model, we are able to obtain exact formulas for such quantities as the generating function and the expected value of the system content, the mean customer delay and the deadline-expiration ratio. These formulas, however, contain infinite sums and infinite products, which implies that truncations are required to actually compute numerical values. Therefore, we also derive some easy-to-evaluate approximate results for the main performance measures, based on a *polynomial approximation* technique. We believe this technique, in its own right, is also one of the major (methodological) contributions of the paper.

Possible applications of this type of queueing model are numerous: the (variable) deadlines could model, for instance, the fact that customers may become impatient and leave the queue unserved if they have to wait too long in line, but they could also reflect the fact that the service of a customer is not useful anymore if it cannot be delivered soon enough, etc.

1. Introduction

In a typical queueing model, customers present themselves near some service facility to receive some kind of service and – if they cannot be served immediately upon arrival – wait patiently in a queue until the server is available for them. In some cases, however, customers may leave (or *abandon*) the queue unserved if their time in the queue becomes too big. Although sometimes referred to as "queues with abandonments" or "queues with reneging" in the literature, this type of queues is usually known as "queues with customer impatience".

The motivations for studying queues with customer impatience are legion. Early papers on the topic were written in the context of telephone traffic and call centers. One of the pioneering works is that of Palm [1]. He considered an unlimited M/M/n queue and assumed that each individual customer stays in the queue as long as his waiting time does not exceed an exponentially distributed impatience time (i.e., M/M/n+M, where the "+M" specifies the impatience law). This is the so-called *Erlang-A* model; it is the simplest model including abandonments. Amongst other results, he represented the steady-state distribution of the number of customers in the Erlang-A system, and some of its important performance measures, in

* Corresponding author.
E-mail addresses: hb@telin.UGent.be (H. Bruneel), tmaerten@telin.UGent.be (T. Maertens).

http://dx.doi.org/10.1016/j.peva.2015.01.009
0166-5316/

terms of incomplete Gamma functions and the blocking probability in the Erlang-B (i.e., M/M/n/n) system. Several authors extended his results, in various directions and sometimes independently of each other. Readers are referred to the invited review paper of Gans et al. [2] and to the Ph.D. thesis of Zeltyn [3] for extensive literature reviews (up to 2003–2004) and relevant results on call center research in general, including models of customers' impatience, and specifically on M/M/n queues with exponentially and generally distributed patient times, respectively. More recent references on customer impatience research in the context of call centers can be found in [4]. In [4], moreover, the authors propose an extension of the Erlang-A model in which the possibility of balking (refusing to join the queue) is included. This simple extension makes the performance prediction by the queueing model much more accurate. Furthermore, they study a number of different service level definitions, including all those used in practice, and show how to explicitly compute their performance measures by using existing results on the *virtual waiting time*, i.e., the waiting time that a customer with infinite patience would experience.

Customer impatience (or, more general, abandonments), however, is also not to be ignored in, for example, real-time telecommunication applications (see, e.g., [5–8]), inventory management (see, e.g., [9–11]), emergency situations, staffing decisions (see, e.g., [12–14]), parking policy (see, e.g., [15]), etc. Usually, it is the customer that takes the decision to abandon prematurely, e.g., because the customer (usually, a human being in this case) does not like to or cannot wait any longer. A call center is the most obvious example of this type of abandonments. On the other hand, also the system itself may decide to remove customers from the queue, e.g., if servicing those customers is deemed not to be useful any more after some time in the queue or if the customers are "expired" (e.g., perishable goods). The first situation may appear in audio or video streaming applications (see, e.g., [16–18]). In particular, when packets belonging to such applications would not arrive soon enough at their next destination if they have to wait any longer, they are removed from the buffer (see, e.g., [6,19,20]). The second situation can be of great importance in inventory management (see, e.g., [10,11]). There are many examples of perishable products such as food items, chemicals, pharmaceutical products, blood, etc. Understanding such systems and investigating the impact of the finiteness of product lifetimes on production and inventory control decisions is thus clearly necessary in a society in which waste is less and less accepted and in which extra costs (e.g., for cleaning up waste) are more and more avoided. For other examples and situations in which abandonments play an important role, we refer to [21,22].

There is clearly no shortage of *continuous-time* models to study queues with customer impatience (see, e.g., Zeltyn and Mandelbaum [23] and references therein for a good overview). In the present paper, however, we make a rare attempt to investigate a *discrete-time* queueing model with customer impatience, by means of a simple analytical model. Specifically, we study a GI/1/1 queue where the patience times (or "deadlines") of the customers are independent and geometrically distributed. We are able to obtain exact formulas for the probability generating function and the mean of the system content, the mean customer delay and the deadline-expiration ratio. These formulas, however, contain infinite sums and infinite products, and are thus not quite useful to see the impact of the various system parameters. Therefore, we also derive some easy-to-evaluate approximate results for the main performance measures. Jean-Marie and Hyon [24] consider the same model, but are interested in optimization rather than in in-depth structural analysis. They show that the optimal control of service in the GI/1/1+Geo queue is a threshold policy and they give the value of this threshold. In Kim et al. [25], furthermore, the authors study a discrete-time multi-server queue in which the customers arrive according to a simple Bernoulli process, in which the service times are geometrically distributed, and in which the customers wait for service for a limited time with a general distribution. They present exact expressions for the loss probability and the queue-length distribution. Van Velthoven et al. [26] derive an expression for the probability of abandonment in a Geo/Geo/1+GI queue and show that systems with a smaller patience distribution in the convex-ordering sense give rise to fewer abandonments (due to impatience), irrespective of whether customers become patient when entering the service facility. Finally, Wu et al. [27] combine the concept of customer impatience with the concepts of retrials and priorities in a discrete-time Geo/G/1 queue. They analyze the Markov chain underlying the considered queueing system and obtain the system-state distribution as well as the orbit-size and the system-size distributions in terms of their generating functions. Besides, they investigate a stochastic decomposition property and the corresponding continuous-time queueing system.

The contributions of the present paper concern the specific model that is considered and the methods that are used to obtain exact and approximate results for the main performance measures. As for the model, it is, as far as we know, the first attempt to perform a structural analysis of a *discrete-time* queue with customer impatience and a *general* independent arrival process. We are totally aware of the simplicity of the service process and of the abandonment process. However, since we want the focus of this paper to lie on the proposed approximation method, we have kept these processes as simple as possible. Of course, we will focus on generalizations of these processes in the future. It is commonly known from the continuous-time literature on queues with customer impatience that exact results of even simple models are often far from user-friendly and that more general models may soon become analytically intractable, so *approximations* have to be proposed to study these systems. We mention a few of them; we refer to Xiong et al. [22] and Sakuma et al. [28] for more. Boxma and de Waal [29] were amongst the first who developed several approximations for the probability to abandon in the (continuous-time) M/G/n+G queue. These approximations, based on intuition and observations, use the exact results obtained for the M/M/n+M and M/M/n+D cases. Extensive tests of these approximations reveal a near-insensitivity of the overflow probability with respect to the service-time distribution, and – apart from a small traffic region – a rather weak sensitivity with respect to the patience distribution. In [22], the authors propose a methodology for approximating the mean waiting time by mapping a multi-server queue to a single server queue with an augmented service rate. The main objective of [28], finally, is to provide an approximation for the waiting-time distribution in an analytically tractable form.