Available online at www.sciencedirect.com
ScienceDirect
European Journal of Operational Research 189 (2008) 1088–1103
EUROPEAN JOURNAL OF OPERATIONAL RESEARCH
www.elsevier.com/locate/ejor

A discrete-time retrial queueing model with one server

Rein Nobel [a,*], Pilar Moreno [b]

[a] *Department of Econometrics, Vrije University, Amsterdam, The Netherlands*
[b] *Departamento de Economía, Métodos Cuantitativos e Historia Económica, Universidad Pablo de Olavide, Sevilla, Spain*

Received 4 April 2006; accepted 15 March 2007
Available online 29 May 2007

Abstract

This paper presents a one-server queueing model with retrials in discrete-time. The number of primary jobs arriving in a time slot follows a general probability distribution and the different numbers of primary arrivals in consecutive time slots are mutually independent. Each job requires from the server a generally distributed number of slots for its service, and the service times of the different jobs are independent. Jobs arriving in a slot can start their service only at the beginning of the next slot. When upon arrival jobs find the server busy all incoming jobs are sent into orbit. When upon arrival in a slot jobs find the server idle, then one of the incoming jobs (randomly chosen) in that slot starts its service at the beginning of the next slot, whereas the other incoming jobs in that slot, if any, are sent into orbit. During each slot jobs in the orbit try to reenter the system individually, independent of each other, with a given retrial probability.

The ergodicity condition and the generating function of the joint equilibrium distribution of the number of jobs in orbit and the residual service time of the job in service are calculated. From the generating function several performance measures are deduced, like the average orbit size. Also the busy period and the number of jobs served during a busy period are discussed. To conclude, extensive numerical results are presented.

Keywords: Late arrival system with delayed access (LAS-DA); Ergodicity condition; Steady-state analysis; Busy period; Number of jobs served during a busy period

1. Introduction

Any branch of science can be subdivided into several categories, each with its own specific problems, and with queueing theory the situation is not different. In queueing theory we can, roughly speaking, identify three types of models which each demonstrate their characteristic ‘congestion properties’, in the language of the queueing theorist, referred to as *delay models*, *loss models*, and *retrial models*. Of course, hybrid forms of these pure types of queueing models can also be considered. In *delay models* jobs which upon arrival do not find an idle server for immediate service, wait in line to be served later according to a prespecified queueing discipline. These models play the dominant role in the literature, they have been widely studied, and form the basis of

* Corresponding author.
E-mail addresses: rnobel@feweb.vu.nl (R. Nobel), mpmornav@upo.es (P. Moreno).

0377-2217/$ - see front matter
doi:10.1016/j.ejor.2007.03.051

most standard courses in queueing theory (for recent monographs we refer to Jain et al. [13], Medhi [17] and Tijms [24]). An important topic in delay models is the study of the waiting-time distribution of an individual job. Many results have been obtained for the well-known $M/G/1$-type queueing models which all show the effect of the variability of the service times on the waiting-time distribution, succinctly expressed in the famous Pollaczek–Khintchine formula. To put it in simple terms, higher irregularity in the service times brings with it longer delays. In *loss models* jobs which upon arrival do not find an idle server available for service are rejected access to the system. These models have also been studied extensively with the $M/G/c/c$ Erlang-loss model as the proverbial example. In these models the main topic has been the calculation of the loss-probability, i.e. the fraction of jobs that will be rejected upon arrival because no idle server is available. A property of many loss models is their insensitivity for the variability of the service times. Many performance measures of the system turn out to depend on the service time only through its mean. These so-called insensitivity results received a lot of attention during the past decades (see e.g. [25]). As mentioned above, also hybrid models in which both delay and loss of jobs occur, e.g. finite-buffer delay systems, have been the subject of much research.

Retrial models received much less attention in the literature. In these models jobs which upon arrival do not find an idle server available for service are not put in a waiting line, neither are they rejected, but they are sent into orbit, i.e. a virtual waiting space from which they try to re-enter the system some (random) time later, usually independent of each other. These models are much more difficult to analyse than the former models, mainly because the arrival stream of the jobs now consists of two types, the primary arrivals which enter for the first time, and the arrivals from the orbit, making the ‘arrival intensity’ dependent of the number of jobs in orbit. As a consequence, overtaking takes place, i.e. jobs are not served according to a specific queueing discipline, which severely complicates the study of the waiting-time distribution of a job, here defined as the total time that the job spends in the orbit. It is probably fair to say that the unpopularity of the research on retrial models is partly due to their intractability, because from a practical point of view retrial models often describe a more realistic picture of many queueing situations than any of the other type of models. Notwithstanding the mathematical difficulties encountered in the study of retrial systems, some models, with the $M/G/1$ retrial queue in a prominent position, have been analysed thoroughly, and we refer to the monograph of Falin and Templeton [12] for an overview of the main results.

Until now we distinguished queueing models with respect to the way congestion is experienced after an unsuccessful attempt to get immediate service upon arrival: either jobs have to ‘suffer’ delay, they are denied access, or jobs are forced to retry access later. We can also distinguish the models with respect to the way ‘time’ is considered, continuous or discrete. Most of the research in queueing theory concerns so-called continuous-time models, with the Poisson process playing a dominant role in the more elementary models like the $M/G/1$ queue, and more complicated Markov modulated arrival processes in advanced models with correlated interarrival times. Also the service times are taken in continuous-time. In one way or the other the exponential distribution is omnipresent in most of the continuous-time queueing models. The motivation for this choice for continuous-time is partly historical (Erlang’s heritage), but probably also mathematical convenience plays a role here. As holds for physics, the mathematical techniques required for studying continuous-time queueing models are well-established and belong to the standard toolkit of every researcher. From a practical point of view it is not so self-evident to direct all the research effort to continuous-time models. Mainly as a consequence of the revolutionary developments in the computer and telecommunication technology, in the eighties and nineties of the past century people started to study also queueing models in discrete-time (the first paper on discrete-time queueing theory dates from 1958, see Meisling [18], but this is an isolated paper). The emphasis of the research has been put on the discrete-time analogues of the well-studied delay models in continuous-time. So, the exponential distribution has been replaced by its discrete-time counterpart, i.e. the geometric distribution, and many papers have been published on (variants of) $Geo/G/1$- and $GI/Geo/1$-type models. As some recent examples we mention the papers of Chaudhry [9] and Chaudhry and Gupta [10]. The $Geo/Geo/c$ queue has been studied by Artalejo and Hernández-Lerma [1]. We refer to Bruneel and Kim [8], Miyazawa and Takagi [19], and Takagi [22] for overviews on discrete-time models.

Until now not much work has been done with respect to discrete-time retrial models. Choi and Kim [11], Li and Yang [14,15], Takahashi et al. [23] and Yang and Li [26] made a start, and not surprisingly they also