European Journal of Operational Research 000 (2015) 1–12

Contents lists available at ScienceDirect

European Journal of Operational Research

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

Stochastics and Statistics

Queueing models with optional cooperative services

Srinivas R. Chakravarthy*

Department of Industrial and Manufacturing Engineering, Kettering University, Flint, MI-48504, USA

A R T I C L E I N F O

Article history:
Received 31 January 2015
Accepted 10 August 2015
Available online xxx

Keywords:
Queueing
Cooperative services
Matrix-analytic method
Simulation

A B S T R A C T

In this paper we consider a single-server queueing model in which the customers arrive according to a versatile point process that includes correlated arrivals. An arriving customer can either request for an individual service or for a cooperative service (to be offered along with other customers with similar requests) with some pre-specified probabilities. There is a limit placed on the number of customers requiring cooperative services at any given time. Assuming the service times to be exponentially distributed with possibly different parameters depending on individual or cooperative services, we analyze this model using matrix-analytic method. Second, we simulate this model to obtain a couple of key performance measures which are difficult to compute analytically as well as numerically to show the benefit of cooperative services in queueing. Interesting numerical examples from both analytical and simulated models are discussed. We believe this type of queueing model, which is very much applicable in service areas, has not been studied in the literature.

1. Introduction

The motivation for the study of this queueing model comes out of situations seen in capacitated inventory situations wherein the customers coordinate their orders so as to minimize the ordering costs (see e.g., Fiestras-Janeiro, García-Jurado, Meca, & Mosquera, 2014; 2015). While cooperative services are widely used in inventory modeling, they are very rarely considered in queueing literature. The rare situations where cooperative services are considered in queueing literature arise from the point of view of the service providers as opposed to the customers' points of view (see e.g., Anily & Haviv, 2010; Garcia-Sanz, Fernandez, Fiestras-Janeiro, García-Jurado, & Puerto, 2008). Basically, in these queueing situations the resources (i.e., service providers) of many queueing systems are pooled in one form or the other to efficiently provide services to the pooled customers.

It should be pointed out that the model studied here does not belong to any of the various types of queueing models with batch (or group) services studied extensively in the literature. The study of batch service queueing model was initiated by Bailey (1954) and has grown significantly over the years with the works of Neuts (1967), Powell and Humblet (1986), Chakravarthy (1992), Gold and Tran-Gia (1993), Baba (1996), Adan and Resing (2000), Banik, Chaudhry, and Gupta (2008), and Banik, Gupta, and Chaudhry (2009). The book by Chaudhry and Templeton (1983) provides a comprehensive study on batch service queues. These sample references are meant to point out different types of batch service queueing models studied so far and how our model is totally different from these.

Thus, in our study here we focus on (certain) customers requesting cooperative services so as (a) to minimize their waiting time in the system; (b) to maximize the chances of staying in the system shorter than the average time it takes in the corresponding queueing model with first-come-first-served basis; (c) to share the costs associated with services with fellow customers; or (d) a combination of two or more of (a–c). To the best of our knowledge there is no literature on this type of queueing models. Hence, we view our paper as a first step toward a different class of queueing models which will generalize some well-known queueing models as special cases through proper choice of the key parameters.

We will briefly present an application of this model and similar applications can be found in other service areas. These days the online purchasing has become so prevalent that even small and upcoming companies are becoming integral part of this e-commerce business. According to the U.S. Census Bureau (see www.census.gov) the annual retail e-commerce (which includes online shopping) has grown from 4.984 billion dollars in 1998 to 260.669 billion dollars in 2013. Thus, more and more retailers want a piece of action in this business. Consider a small to medium retailer handling customers' orders over the phone. Suppose that two types of customers, say, Type 1 and Type 2, place orders over the phone. The orders are processed by a receiving attendant (server) and the processing times (not including the shipping and receiving times) are the actual service times. Type 1 customers prefer to get their orders shipped directly to their addresses whereas Type 2 customers, in order to save money in shipping and processing, prefer to get their orders shipped

* Tel.: +1 810 7627 906; fax: +1 810 6108 391.
E-mail address: schakrav@kettering.edu

http://dx.doi.org/10.1016/j.ejor.2015.08.015
0377-2217/

to a particular location to be picked up by them. The orders from Type 2 customers are processed as soon as the number of orders hits L on a non-preemptive priority (over Type 1) basis. This is to make sure that exactly L orders are put in the same package for shipping to a specific location for all Type 2 customers to individually pick their order. Thus, the orders are processed on a FIFO basis within Type 1 and on a non-preemptive priority basis for Type 2. Similar applications can be found in other areas notably in service systems requiring parts for offering services and these parts are ordered as and when the inventory level reaches a certain point.

The rest of the paper is organized as follows. In Section 2 the model under study is described. The steady-state analysis of the model is performed in Section 3 and in this section we also display some key system performance measures. Special cases of the model under study are presented in Section 4. Illustrative numerical examples are discussed in Section 5. In Section 6 we simulate the model under study. The main purposes of this simulation approach is to see the effectiveness of cooperative service model as compared to the classical queueing model. The effectiveness comparison of the models is done through two system performance measures, of which one is very difficult to obtain analytically and hence simulation plays an important role. Further, our simulated model will help in future analytical study of the model for more general cases. Some concluding remarks are given in Section 7.

2. Model description

We consider a single server queueing system in which customers arrive according to a Markovian arrival process (MAP) with representation (D_0, D_1, D_2) of order m. The generator D, defined by $D = D_0 + D_1 + D_2$, governs the underlying Markov chain of the MAP such that D_0 accounts for the transitions corresponding to no arrival; D_1 governs those corresponding to an arrival of a customer who requires individual services, and D_2 governs those corresponding to an arrival of a customer who requires cooperative services. By assuming D_0 to be a nonsingular matrix, the interarrival times will be finite with probability one and the arrival process does not terminate. Hence, we see that D_0 is a stable matrix. Henceforth, we will refer to customers requiring individual services as Type 1 customers and those requiring cooperative services will be referred to as Type 2 customers.

We assume that Type 1 customers have an infinite waiting space while Type 2 customers have a finite waiting space of size L, $1 \le L < \infty$, where L is a pre-determined threshold. An arriving Type 1 customer finding the server idle will get into service immediately; otherwise will enter into Type 1 buffer and will wait for the server to be free to offer Type 1 services. An arriving Type 2 customer will either (a) get into service immediately along with other similar customers provided the server is idle and there are $L - 1$ such customers already waiting in the system to receive cooperative services; (b) get into Type 2 buffer of size L provided there is enough space available irrespective of whether the server is busy or idle; (c) with probability γ, $0 \le \gamma \le 1$, will become a Type 1 customer and enter into Type 1 buffer since Type 2 buffer is full; or (d) be lost with probability $1 - \gamma$ since Type 2 buffer is full. On becoming a Type 1, the customer remains as Type 1 until leaving the system with a service.

We assume that the service times are exponential with parameter depending on the type of customer(s) in service. Let μ_1 and μ_2, respectively, denote the rate of services for Type 1 and Type 2 customers. The cooperative services are offered in batches of size L. Note that with our assumption the system can have no more than $2L$ Type 2 customers (L waiting for service and L in service) at any given time. We also assume that Type 2 customers have non preemptive services over Type 1 customers. Thus, upon completion of a service the server will offer a service to a group of Type 2 customers if there are L Type 2 customers waiting; otherwise the server will offer a service to a Type 1 customer, if any, or become idle.

The MAP, first introduced by Neuts (1979) as a versatile Markovian point process, is a rich class of point processes that includes many well-known processes such as Poisson, PH-renewal processes, and Markov-modulated Poisson process. For further details on MAP and their usefulness in stochastic modelling, we refer to Lucantoni, Meier-Hellstern, and Neuts (1990) Lucantoni (1991) Neuts (1989) 1992) and for a review and recent work on MAP we refer the reader to Artalejo, Gomez-Correl, and He (2010) Chakravarthy (2010) Chakravarthy and Krishnamoorthy (2001).

Let $\boldsymbol{\delta}$ be the stationary probability vector of the Markov process with irreducible generator D. That is, $\boldsymbol{\delta}$ is the unique (positive) probability vector satisfying.

$$\boldsymbol{\delta} D = \mathbf{0}, \boldsymbol{\delta}\boldsymbol{e} = 1. \tag{1}$$

The constant $\lambda = \boldsymbol{\delta}(D_1 + D_2)\boldsymbol{e}$, referred to as the **fundamental rate**, gives the expected number of arrivals per unit of time in the stationary version of the MAP. Often, in model comparisons, it is convenient to select the time scale of the MAP so that λ has a certain value. That is accomplished, in the continuous MAP case, by multiplying the coefficient matrices D_0, D_1, and D_2, by the appropriate common constant. We define $\lambda_1 = \boldsymbol{\delta} D_1 \boldsymbol{e}$ and $\lambda_2 = \boldsymbol{\delta} D_2 \boldsymbol{e}$ so that these give, respectively, the rates of Type 1 and Type 2 customers arriving to the system. For use in sequel, we take $p = \frac{\lambda_1}{\lambda}$, $q = 1 - p$.

In the sequel we need the following notations. By $\boldsymbol{e}$ we will denote a column vector (of appropriate dimension) of 1's; $\boldsymbol{e}_i$ we will denote a unit column vector (of appropriate dimension) with 1 in the ith position and 0 elsewhere; and I an identity matrix (of appropriate dimension). We will display the dimension should there be a need to emphasize it. For example, if there is a need to display the dimension of an identity matrix of order m, we will do so by writing I_m rather than I; a unit vector of dimension m will be denoted as $\boldsymbol{e}(m)$ rather than $\boldsymbol{e}$. The symbol $\otimes$ denotes the Kronecker product of matrices. Thus, if A is a matrix of order $m \times n$ and if B is a matrix of order $p \times q$, then $A \otimes B$ will denote a matrix of order $mp \times nq$ whose (i, j)th block matrix is given by $a_{ij}B$. For details and properties on Kronecker products we refer the reader to Graham (1981) Marcus and Minc (1964) Steeb and Hardy (2011).

3. The steady-state analysis

In this section we will analyze the queueing model described in Section 2 in steady-state. Let $N_1(t)$, $N_2(t)$, $N_3(t)$, and $N_4(t)$ denote, respectively, the number of Type 1 customers in the queue, the number of Type 2 customers in the queue, the status of the server, and the phase of the arrival process. By taking $N_3(t) = 0, 1, 2$, respectively, we will denote the status of the server to be idle, busy with a Type 1 customer, and busy with Type 2 customers. The process $\{(N_1(t), N_2(t), N_3(t), N_4(t)) \colon t \ge 0\}$ is a continuous-time Markov chain with state space given by

$$\Omega = \{(0, i_2, 0, k) : 0 \le i_2 < L, 1 \le k \le m\} \bigcup \{(i_1, i_2, i_3, k) : i_1 \ge 0, 0 \le i_2 \le L, i_3 = 1, 2, 1 \le k \le m\}.$$

We now define the set of states as follows:

$$\underline{\mathbf{0}}^* = \{(0, i_2, 0, k), 0 \le i_2 < L, 1 \le k \le m\}$$

$$\underline{\boldsymbol{i}} = \{(i, i_2, 1, k) : 0 \le i_2 \le L, 1 \le k \le m\} \bigcup \{(i, i_2, 2, k) : 0 \le i_2 \le L, 1 \le k \le m\}, \ i \ge 0.$$

It can easily be verified that the set of states, $\underline{\mathbf{0}}^*$, corresponds to the case when the server is idle with i_2, $0 \le i_2 < L$, Type 2 customers in the queue and keeping track of the phase of the arrival process. The set of states, $\underline{\boldsymbol{i}}$, corresponds to the case when the server is busy (with a Type 1 customer when $N_3(t) = 1$ and busy with a group of L Type 2 customers when $N_3(t) = 2$) with i, $i \ge 0$, Type 1 customers and i_2, $0 \le i_2 \le L$, Type 2 customers waiting in the queue and keeping track of the phase of the arrival process.