European Journal of Operational Research 235 (2014) 149–158

Contents lists available at ScienceDirect

European Journal of Operational Research

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

Stochastics and Statistics

Class clustering destroys delay differentiation in priority queues

Herwig Bruneel, Tom Maertens *, Joris Walraevens

Ghent University (UGent), Department of Telecommunications and Information Processing (TELIN), SMACS Research Group, Sint-Pietersnieuwstraat 41, B-9000 Ghent, Belgium

ARTICLE INFO

Article history:
Received 11 December 2012
Accepted 3 December 2013
Available online 12 December 2013

Keywords:
Priority queueing
Multiclass
Discrete-time
Interclass correlation
Delay differentiation

ABSTRACT

This paper considers a discrete-time priority queueing model with one server and two types (classes) of customers. *Class-1* customers have absolute (service) priority over *class-2* customers. New customer batches enter the system at the rate of one batch per slot, according to a general independent arrival process, i.e., the batch sizes (total numbers of arrivals) during consecutive time slots are i.i.d. random variables with arbitrary distribution. All customers entering the system during the same time slot (i.e., belonging to the same arrival batch) are of the same type, but customer types may change from slot to slot, i.e., from batch to batch. Specifically, the types of consecutive customer batches are correlated in a Markovian way, i.e., the probability that any batch of customers has type 1 or 2, respectively, depends on the type of the previous customer batch that has entered the system. Such an arrival model allows to vary not only the relative loads of both customer types in the arrival stream, but also the amount of correlation between the types of consecutive arrival batches. The results reveal that the amount of delay differentiation between the two customer classes that can be achieved by the priority mechanism strongly depends on the amount of such *interclass correlation* (or, *class clustering*) in the arrival stream. We believe that this phenomenon has been largely overlooked in the priority-scheduling literature.

1. Introduction

In multiclass queueing systems, where multiple types (or classes) of customers compete for the use of the same resources, scheduling disciplines are used to determine the order of service for customers of different types. One particular type of scheduling that has received substantial attention in the queueing literature is priority scheduling (Jaiswal, 1968; Takagi, 1991; Williams, 1980; De Clercq, De Turck, Steyaert, & Bruneel, 2011; Min & Yih, 2010; Jin & Min, 2007; Zeltyn, Feldman, & Wasserkrug, 2009; Feng & Umemura, 2009; Walraevens, Fiems, Wittevrongel, & Bruneel, 2009; Gamarnik & Katz, 2009; Walraevens, Fiems, & Bruneel, 2006, 2008; Walraevens, Steyaert, & Bruneel, 2005; Maertens, Walraevens, & Bruneel, 2007; Abate & Whitt, 1997; Maertens, Walraevens, & Bruneel, 2008; Chang & Harn, 1992; Choi, Choi, Lee, & Sung, 1998; Drekic & Grassmann, 2002; Fiems, Walraevens, & Bruneel, 2007; Hashida & Takahashi, 1991; Laevens & Bruneel, 1998; Lee & Lee, 2003; Mehmet Ali & Song, 2004; Subramanian & Srikant, 2000; Sugahara, Takine, Takahashi, & Hasegawa, 1995; Takine, 1999; Tham, Yao, & Jiang, 2002; Van Houdt & Blondia, 2006; Walraevens, Steyaert, & Bruneel, 2004; Walraevens, Steyaert, & Bruneel, 2002; Zhao, Li, Cao, & Ahmad, 2006; Chen & Zhang, 2000; Adan, Sleptchenko, & Van Houtum, 2009; Maertens, Bruneel, & Walraevens, 2012), where customers are partitioned in a number of distinct classes and the order of service is based on the classes customers belong to. E.g., class-1 customers have (service) priority over all other classes, class-2 customers have priority over all classes except class 1, etc. Typical applications where priority scheduling is used are, amongst others, packet switches in modern telecommunication networks (Choi, Lee, & Un, 1997; Walraevens et al., 2005; Fayza, 2010), where delay-sensitive packets (e.g. telephony, teleconferencing, video) are given preferential treatment above delay-tolerant packets (e.g. file transfer, e-mail), or emergency services in hospitals, where patients requiring urgent intervention are given priority over regular patients (Min & Yih, 2010; Jacobson, Argon, & Ziya, 2012).

In classical priority queueing models, it is generally assumed that the different classes of customers occur randomly and independently in the arrival stream of customers into the system. In this paper, however, we explicitly wish to examine the effect of *interclass correlation* (or *class clustering*) in the arrival process of a two-class priority queue on the performance of this queue. In particular, we are interested to know whether the degree to which customers of the same type have the tendency to arrive closely together (i.e., "clustered"), or, conversely, the degree to which such customers have the tendency to be spread in time and mixed with customers of the other type, have a substantial impact on the performance of a two-class priority queueing system. In order to do so, we superimpose a two-state Markovian interclass correlation model (with arbitrary transition probabilities) on top of a regular

* Corresponding author. Tel.: +32 92648901.
E-mail addresses: hb@telin.UGent.be (H. Bruneel), tmaerten@telin.UGent.be (T. Maertens), jw@telin.UGent.be (J. Walraevens).

0377-2217/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.ejor.2013.12.005

general independent arrival-process model for the aggregated customer stream.

For this model, we first derive the probability generating function (pgf) of the *total number* of customers in the system, as well as the pgf of the delay (system time) of an arbitrary customer. These results can be easily retrieved from the well-known analysis of a single-class discrete-time queueing model with general i.i.d. arrivals and deterministic single-slot service times (Bruneel & Kim, 1993; Vinck & Bruneel, 1995). Various performance measures of practical use, such as the total mean system content and the (global) mean delay of an arbitrary customer, can be easily derived from this pgf. Next, we analyze the queueing performance of the *high-priority* customer class (i.e., class 1). This requires the solution of a single-class queueing model with Markovian arrival interruptions. We are able to obtain explicit expressions for the pgf's of the class-1 system content and the class-1 customer delay, and from this, performance measures such as the mean number of class-1 customers in the system and the mean delay of an arbitrary class-1 customer. Finally, combining all the latter results, we are also able to derive explicit expressions for the mean number of *low-priority* (i.e., class-2) customers in the system and the mean delay of an arbitrary class-2 customer.

The resulting formulas and a number of numerical examples reveal that the priority mechanism does what it is designed for, i.e., favor high-priority customers in terms of a lower mean delay than low-priority customers, as long as the interclass correlation is sufficiently low. For high to very high interclass correlation, however, the delay-differentiation capabilities of the priority mechanism are reduced significantly, and even disappear completely when the interclass correlation approaches +1. We believe that this phenomenon is not well recognized in most of the priority-queueing literature, because the arrival process of the various types of customers is usually chosen quite arbitrarily and not given much attention. The current paper proves – by means of closed-form results – that class clustering is not to be neglected in the context of priority queues.

2. Mathematical model

We consider a discrete-time queueing system with infinite waiting room, one server, and two types (classes) of customers, named 1 and 2. As in all discrete-time models, the time axis is divided into fixed-length intervals referred to as *slots*. New customers may enter the system at any given (continuous) point on the time axis, but services are synchronized to (i.e., can only start and end at) slot boundaries.

In order not to complicate matters, we model the service process of the system as simply as possible in this paper. Specifically, we assume that the service times of all customers (belonging to either class 1 or class 2) are deterministically given by one slot each. More general models for the service process will be the subject of future investigations.

The arrival process of new customers in the system, however, which is the main concern of this paper, is characterized in two steps.

First, we model the total (aggregated) arrival stream of new customers by means of a sequence of i.i.d. non-negative discrete random variables with common probability mass function (pmf) $a(n)$ and common probability generating function (pgf) $A(z)$. More specifically,

$$a(n) \triangleq \text{Prob}[n \text{ arrivals in one slot}], \quad n \geqslant 0,$$

$$A(z) \triangleq \sum_{n=0}^{\infty} a(n) z^n.$$

We call the total number of arrivals in one slot an *arrival batch* in the sequel.

The mean batch size, i.e., the (total) mean number of arrivals per slot, in the sequel referred to as the (total) mean arrival rate, is given by

$$\lambda = A'(1). \tag{1}$$

Next, we describe the occurrence of the two types (1 and 2) in the sequence of the consecutively arriving customer batches. First of all, we assume in this study that each arrival batch contains only one type of customers, i.e., either all type-1 customers or all type-2 customers. The case where customers of both types may be present in the same batch has been considered in many existing studies and is therefore not included here; we further comment on this in Section 6. We further assume that both customer classes are "mixed" in the arrival stream, but that there may be some degree of "class clustering" in the arrival process, i.e., customer batches of any given type may (or may not) have a tendency to "arrive back-to-back". Mathematically, this means that the types of two consecutive batches may be non-independent. Specifically, we assume a first-order Markovian type of correlation between the types of two consecutive batches, which basically means that the probability that the next batch belongs to a given class depends on the type of the previous batch.

It should be noted that our arrival model is a non-classical "mixture" of independent arrivals (on the aggregated level) and correlated arrivals (for each customer class individually). It is not to be confused with classical single-class Markovian arrival models such as, for instance, in Ali Khan and Gani (1968), Pakes and Phatarfod (1978), Bruneel (1985, 1988), Bruneel and Steyaert (1996), Mehmet Ali and Song (2004), Gao, Wittevrongel, Walraevens, and Bruneel (2008), and Claeys, Steyaert, Walraevens, Laevens, and Bruneel (2013). There are two main reasons for this choice. First, the lack of correlation in the aggregated arrival process makes the analysis easier; secondly, it is our explicit intention to study the impact of interclass correlation, i.e., correlation between the arrivals of the two customer types, as "purely" as possible, i.e., without possible interference of other sources of correlation. For the same reason, we have kept the service process as simple as possible (in particular, uncorrelated from customer to customer).

Let t_k denote the type (i.e., 1 or 2) of the batch arriving during slot k. The transition probabilities of the Markov chain that determines the types of the consecutive batches are then defined as (see Fig. 1)

$$\begin{aligned} &\text{Prob}[t_{k+1} = 1 | t_k = 1] = \alpha; \quad \text{Prob}[t_{k+1} = 2 | t_k = 1] = 1 - \alpha, \\ &\text{Prob}[t_{k+1} = 1 | t_k = 2] = 1 - \beta; \quad \text{Prob}[t_{k+1} = 2 | t_k = 2] = \beta. \end{aligned} \tag{2}$$

It is well-known (Bruneel & Kim, 1993; Bruneel, 1988) that for a two-state Markov chain of this type, the steady-state probabilities π_1 and π_2 of finding the chain in state 1 or 2 respectively, are given by

$$\begin{aligned} \pi_1 &\triangleq \lim_{k\to\infty} \text{Prob}[t_k = 1] = \frac{1-\beta}{2-\alpha-\beta}, \\ \pi_2 &\triangleq \lim_{k\to\infty} \text{Prob}[t_k = 2] = \frac{1-\alpha}{2-\alpha-\beta}. \end{aligned} \tag{3}$$

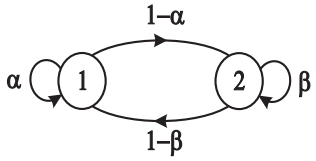

Fig. 1. Two-state Markov chain of the customer types.