Computers & Operations Research 40 (2013) 1497–1505

Contents lists available at SciVerse ScienceDirect

Computers & Operations Research

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

Tail probabilities of the delay in a batch-service queueing model with batch-size dependent service times and a timer mechanism

Dieter Claeys *, Bart Steyaert, Joris Walraevens [1], Koenraad Laevens, Herwig Bruneel

Stochastic Modelling and Analysis of Communication Systems (SMACS) Research Group, Department of Telecommunications and Information Processing (TELIN), Ghent University, Sint-Pietersnieuwstraat 41, B-9000 Gent, Belgium

ARTICLE INFO

Available online 26 October 2012

Keywords:
Batch service
Batch arrivals
Batch-size dependent
Timer
Customer delay
Tail probabilities

ABSTRACT

We deduce approximations for the tail probabilities of the customer delay in a discrete-time queueing model with batch arrivals and batch service. As in telecommunications systems transmission times are dependent on packet sizes, we consider a general dependency between the service time of a batch and the number of customers within it. The model also incorporates a timer mechanism to avoid excessive delays stemming from the requirement that a service can only be initiated when the number of present customers reaches or exceeds a service threshold. The service discipline is first-come, first-served (FCFS). We demonstrate in detail that our approximations are very useful for the purpose of assessing the order of magnitude of the tail probabilities of the customer delay, except in some special cases that we discuss extensively. We also illustrate that neglecting batch-size dependent service times or a timer mechanism can lead to a devastating assessment of the tail probabilities of the customer delay, which highlights the necessity to include these features in the model. The results from this paper can, for instance, be applied to assess the quality of service (QoS) of Voice over IP (VoIP) conversations, which is typically expressed in terms of the order of magnitude of the probability of packet loss due to excessive delays.

1. Introduction

In many real-life circumstances, customers receive some kind of service in group, which is often referred to as batch service. An elevator can be conceived as a textbook example, since elevators can convey several people simultaneously to another floor. Other examples include transport vehicles, busses, ship locks, ovens in production processes, attractions in amusement parks, etc. Furthermore, in telecommunications, it is often the case that information packets are grouped in larger entities (batches) and these batches are transmitted instead of each packet individually. This is mainly done for efficiency reasons, since only one header per aggregated batch has to be constructed, instead of one header per single information unit, thus leading to an increased throughput. Technologies using packet aggregation include Optical Burst Switched (OBS) networks [1,2] and IEEE 802.11n wireless local area networks (WLANs) [3]. More applications can, for instance, be found in [4].

On account of the wide area of applications, queueing models with batch service have attracted considerable attention. However, the focus was mainly put on the number of waiting customers (see e.g., [5–16]), whereas the waiting time of customers, also called customer delay, has attracted very few attention, especially in the case of batch arrivals.

In [17–19] we have computed the probability generating function (PGF) of the customer delay in distinct discrete-time queueing models with batch arrivals and batch service. Although the established PGFs allow us to calculate various moments of the customer delay, these are not suitable to extract tail probabilities. Nevertheless, this is an important performance measure. For instance, the quality of service (QoS) of Voice Over IP (VoIP) conversations is generally expressed in terms of the (order of magnitude of the) probability that packets arrive too late at the end user (see e.g., [20]). The tail probabilities of the delay in a batch-service queueing model can, among others, be applied to assess the QoS of VoIP conversations in wireless personal area networks (WPANs). The queueing model then represents a node's output and transmission buffer corresponding to a particular destination and QoS: the output buffer is the queue of the batch-service queueing model, the transmission buffer is the (batch) server (one typically places bursts instead of individual packets in the transmission buffer to increase the throughput), and the time that a burst resides in the transmission buffer is the service time.

In view of this, we have established in [21] an approximation for the tail probabilities of the customer delay in a batch-arrival,

* Corresponding author. Tel.: +32 9 264 3411; fax: +32 9 264 4295.
E-mail address: Dieter.Claeys@telin.ugent.be (D. Claeys).
[1] This author is a Postdoctoral Fellow with the Fund for Scientific Research, Flanders (F.W.O.-Vlaanderen), Belgium.

0305-0548/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.cor.2012.10.009

batch-service queueing model with single-slot service times and with a server that only serves full batches (i.e., the server only starts service when at least as many customers are present as the server capacity). In [22], we have considered a more versatile model with a minimum batch size (also called service threshold) l (i.e., service is initiated only if at least l customers are present, with l some value between 1 and the server capacity) and generally distributed service times. In this paper, we extend our previous work [22]. In [22], the service times do not depend on the batch sizes, whereas in actual telecommunications systems transmission times depend on packet sizes. In addition, it has been shown in [22] that in case of light traffic, the delay can be extremely high when a minimum batch size is enforced. Therefore, in the model studied in this paper, we consider a *general dependency between the service time of a batch and the number of customers within it*, and we include a *timer mechanism* that avoids excessive delays in case of light traffic as well. It will turn out that the analysis of these extensions entail various pitfalls and that neglecting those pitfalls leads to inaccurate approximations. In addition, we focus more on an extensive evaluation of the accuracy of our approach. We demonstrate that the established approximations are very useful to assess the order of magnitude of the tail probabilities of the customer delay, except in some peculiar situations which we discuss in detail. Finally, we illustrate that neglecting batch-size dependent service times or a timer mechanism can lead to distorted results, which reflects the importance of including these features in the model.

The remainder of the paper is structured as follows: in Section 2 we describe the model. Then, in Section 3, we deduce approximations for the tail probabilities. The accuracy of our approach is evaluated extensively in Section 4 and the importance of the model is discussed in Section 5. Finally, we draw some conclusions in Section 6.

2. Model description

We consider a discrete-time queueing model. As such, the time axis is divided into fixed-length contiguous time periods, called slots. Customer arrivals during consecutive slots are modelled by a sequence of independent and identically distributed (IID) random variables, with common random variable A whose probability generating function (PGF) is denoted by $A(z)$. The mean value, often referred to as mean arrival rate, is characterized by λ and is by definition equal to $A'(1)$ (we use primes to indicate derivatives). Customers queue up in awaitance of service in a queue of infinite size. The server can serve batches containing up to c customers. We refer to c as the server capacity. Whenever the server is available at the beginning of a slot and finds less than l customers ($l \leq c$), service is initiated with probability β and postponed with probability $1-\beta$. If, on the other hand, at least l customers are present, a service is initiated of a batch containing a maximum of c customers. Service times are synchronized with respect to the slot boundaries, i.e., services always start and end at slot boundaries. Hence, service times last an integral number of slots. The service time of a batch containing n customers is represented by T_n and its corresponding PGF by $T_n(z)$. Under these assumptions, $T_0(z)$ describes the length of a server interruption in an empty system. Finally, the service discipline is first-come, first-served (FCFS).

The results in this paper are valid under the following assumptions:

Assumption 1. The load $\rho \triangleq \lambda T_c'(1)/c < 1$.

This ensures stability of the system.

Assumption 2. The radius of convergence of each PGF is strictly larger than 1.

This implies that all order moments are finite and can be calculated by means of the moment generating property of PGFs. We designate the radius of convergence of some random variable X by $\mathfrak{R}_X$. In addition, we define $\mathfrak{R}_n$ as the radius of convergence of $T_n(A(z))$ and $\mathfrak{R} \triangleq \min\{\mathfrak{R}_n : 0 \leq n \leq c\}$ and $\mathfrak{R}_T \triangleq \min\{\mathfrak{R}_{T_n} : 0 \leq n \leq c\}$.

Assumption 3. $\mathfrak{R}_n \leq \mathfrak{R}_A$, $n = 0, \ldots, c$.

It is worth mentioning that we believe that this assumption is actually a fact, as we have not been able to construct one counterexample.[2] However, as it is tedious to prove that $\mathfrak{R}_n \leq \mathfrak{R}_A$, we mention it as an assumption.

Assumption 4. $z^c - T_c(A(z))$ is aperiodic, i.e., the highest common factor of the set of integers $\{\{c\} \cup \{n \in \mathbb{N} : (\mathrm{d}^n/\mathrm{d}z^n) T_c\ (A(z))\big|_{z=0} \neq 0\}\}$ equals 1.

This assumption ensures that the c unknown boundary probabilities $d(n)$, $n = 0, \ldots, c-1$ (see further) are solutions of a set of c linear independent equations. We thus exclude some special cases (for instance when $c = 2k$, $l = c$, $\beta = 0$ and $A(z) = \sum_{n=0}^{\infty} \Pr[A = 2n] z^{2n}$) in order to present a general solution technique.

Assumption 5. $\lim_{z \uparrow \mathfrak{R}} T_c(A(z))/z^c > 1$.

This assumption will assure that $z^c - T_c(A(z))$ has a zero in the interval $]1, \mathfrak{R}[$. We will show that this entails that the tail probabilities of the customer delay are not dominated by a specific dominant singularity of $T_c(A(z))$ (if any). Although we thus exclude some PGFs $T_c(A(z))$, the commonly adopted PGFs satisfy this assumption. The main advantage is that we can present a general solution whereas otherwise an ad hoc approach would have to be adopted for each PGF $T_c(A(z))$.

3. Deduction of approximation formulas

The delay of a randomly tagged customer is defined as the length of the time period, starting at the end of the slot of arrival, until the customer's batch starts receiving service. It can thus be expressed as an integral number of slots.

In [22] for a system without timer mechanism, we have decomposed the delay W of a randomly tagged customer as the maximum of two parts:

$$W = \max(W_1, W_2).$$

The *queueing delay* W_1 is the time, starting at the beginning of the slot following the slot wherein the tagged customer arrives (i.e., at the same instant that W starts), to serve batches of customers that have arrived before the tagged customer. The *postponing delay* W_2 is the time, starting at the same moment as the queueing delay, until the batch with the tagged customer contains at least l customers. In this particular case, the actual service of a customer can start only if all preceding batches have been processed (FCFS) and if its own batch contains at least l customers; hence the equation $W = \max(W_1, W_2)$.

It seems natural to follow a similar approach, by simply redefining W_2 somewhat to include the timer mechanism. The postponing delay would then represent the time, starting at the same moment as the queueing delay, until the batch containing the tagged customer contains at least l customers or until the

[2] When trying to construct a counterexample, one should verify that the constructed $A(z)$ and $T_n(z)$ are actually PGFs, by checking the normalization condition and verifying that the coefficients in the Taylor series expansions of $A(z)$ and $T_n(z)$ about $z = 0$ are probabilities.