Computers & Operations Research 39 (2012) 2733–2741

Contents lists available at SciVerse ScienceDirect

Computers & Operations Research

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

# Tail distribution of the delay in a general batch-service queueing model

Dieter Claeys *, Bart Steyaert [1], Joris Walraevens [1,2], Koenraad Laevens [1], Herwig Bruneel [1]

*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 17 February 2012

*Keywords:*
Queueing
Batch service
Batch arrivals
Service threshold
Delay
Tail probabilities

ABSTRACT

Batch servers are capable of processing batches of packets instead of individual packets. Although batch-service queueing models have been studied extensively during the past decades, the focus was mainly put on calculating performance measures related to the buffer content, whereas less attention has been paid to the packet delay. In this paper, we focus on the tail probabilities of the delay that a random packet experiences in a general batch-service queueing model. More specifically, we establish approximations for these probabilities, which are highly accurate and easy to calculate. These results, for instance, allow to accurately assess the probability that real-time packets experience an excessive delay in practical telecommunication systems.



## 1. Introduction

Whereas servers in traditional queueing systems serve one packet at a time, batch servers process batches of packets. The maximum number of packets in a served batch is usually finite and is called the server capacity, which we denote by $c$. An inherent feature of batch service is that newly arriving packets cannot join the ongoing service, even if the served batch is not completely filled. In order to reduce the wasted capacity, one often imposes a threshold, $l$ ($1 \le l \le c$), for the minimum amount of packets in a served batch. This implies that the available server solely initiates service when at least $l$ packets have accumulated in the system.

Batch-service queueing models have a wide area of applications, including transportation, production and manufacturing systems (see e.g. [7,17]) and telecommunications (see e.g. [2]). Batch-service queueing models are for instance employed to assess the performance of burst-frame-based MAC protocols for ultra-wideband (UWB) Wireless Personal Area Networks (WPANs) [25]. A node in such a network typically has for each combination of destination and Quality of Service (QoS) an output and a transmission buffer. Upper-layer packets with the same destination and QoS are stored in the same output buffer. When the transmission buffer is empty and at least $l$ packets have accumulated in the output buffer, maximum $c$ of these packets are grouped into a burst and this burst is stored in the transmission buffer (note that the transmission buffer can only store one burst simultaneously). The burst will be removed from the transmission buffer when an ACK frame from the receiver arrives. Although UWB is a high-speed technology, the time spent in the transmission buffer cannot be ignored due to the competition for the channel between the several output queues and the synchronisation (process of synchronising the receiver's clock with the transmitter's clock) time. The batch-service queueing model in this paper can be used to model an 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 server and the time that a burst resides in the transmission buffer is the service time. This application example thus demonstrates that the analysis of the delay in a batch-service queueing system with general service times and a general batch forming policy is important. This theoretical analysis is subject of this paper.

On account of the wide area of applications, batch-service queueing models have been studied extensively. The emphasis was laid on the amount of packets in the system (e.g. [1,5,6,8,17–19,21,24,28–30,32,33]). The packet delay, however, has only attracted attention in [7,13,14,16,22,23,26,27]. In none of these papers, models are studied with the combination of $l > 1$ and batch arrivals.

In [9], we have computed the probability generating function (PGF) of the packet delay in a discrete-time batch-arrival, batch-service queueing model with $l=c$ and single-slot service times. In [10], we have extended this model to geometrically distributed service times and in [12] we have considered generally distributed service times and $1 \le l \le c$. The established PGF's, though, suffer from the drawback that they are not suitable to extract tail probabilities. However, in several cases, this is an important

* Corresponding author. Tel.: +32 9 264 3411; fax: +32 9 264 4295.
*E-mail address:* Dieter.Claeys@telin.ugent.be (D. Claeys).
[1] Tel.: +32 9 264 3411; fax: +32 9 264 4295.
[2] The third 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.
doi:10.1016/j.cor.2012.02.004

performance measure. For instance, consider an output buffer that stores voice packets. Voice packets are delay-sensitive, meaning that when they arrive too late at the end user (for instance after more than 150 ms), they become useless. The quality of the upperlayer conversations is expressed in terms of the (order of magnitude of the) probability of this event (see e.g. [15]).

In view of this, we have established in [11] an approximation for the tail probabilities of the delay that a random packet experiences in the batch-arrival, batch-service queueing model with single-slot service times and $l=c$. In this paper, we extend this previous research by considering the extended model with $l \in [1,c]$ and generally distributed service times. In addition, we also obtain another approximation that allows us to more accurately assess the delay performance in the batch-service queueing model under study. The paper is organised as follows: the model is described in detail in Section 2. The approximations are established in Section 3, while in Section 4, we demonstrate through some examples that these are highly accurate. Hence, the approximation formulas can be adopted to accurately assess the delay performance in practical batch-service queueing systems.

## 2. Model

In this paper, we consider a discrete-time queueing model. Packets arrive one by one and several packets can arrive in a slot. We call this batch arrivals. The number of packet arrivals during consecutive slots is generated by an independent and identically distributed (IID) process. The number of packet arrivals during slot $k$ is denoted by $A_k$; $A$ represents the number of packet arrivals during a random slot and its PGF is denoted by $A(z)$.

The number of packets in a served batch is upper-bounded by the server capacity $c$ and lower-bounded by the threshold $l$ $(1 \le l \le c)$, implying that when the server becomes available and finds less than $l$ packets, he waits to initiate service and leaves the already present packets in the queue until the beginning of the first slot whereby at least $l$ packets have accumulated in the system. When the system contains more than $c$ packets at that time, the server only processes the first $c$ packets and leaves the others in the queue (according to the first-come-first-served policy). Consecutive batch service times do not depend on the number of packets in the served batches, nor on the number of packet arrivals and they constitute an IID process. The service time of any batch is designated by $T$ and its associated PGF by $T(z)$.

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

**Assumption 1.** The load $\rho \triangleq \mathrm{E}[A]\mathrm{E}[T]/c < 1$.

**Assumption 2.** $R > 1$, with $R$ the radius of convergence of $T(A(z))$.

**Assumption 3.** $\lim_{z \uparrow R} T(A(z))/z^c > 1$.

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

Note that Assumption 2 implies that $R_A > 1$ and $R_T > 1$ with $R_A$ and $R_T$ the radii of convergence of $A(z)$ and $T(z)$ respectively. Further, Assumption 3 is always fulfilled if $T(A(z))$ has a finite pole $R$. Vice versa, if Assumption 3 is not fulfilled, then $R$ necessarily is a branch point of $T(A(z))$, and a separate ad hoc analysis of the packet delay tail distribution is required.

## 3. Deducing the approximation formulas

In order to compute the probability that the delay $W$ (being the sojourn time in the queue) of a randomly tagged packet exceeds some large value, we split the delay into two parts. We illustrate this by means of the example depicted in Fig. 1. The tagged packet's arrival slot is denoted by $J$ and $Q_J$ represents the queue content (i.e. the number of packets in the queue, those in service excluded) at the beginning of slot $J$. Further, $B$ ($X$ respectively) represents the number of packet arrivals in slot $J$ arriving before (after respectively) the tagged packet. The first part of the delay, $W_1$, is the time required to serve the batches with previously arrived packets. It is equal to the remaining service time of the batch being served in slot $J$ (if any), plus the sum of $\lfloor (Q_J+B)/c \rfloor$ service times, where $\lfloor . \rfloor$ represents the floor function, i.e. $\lfloor x \rfloor = \max\{n \in \mathbb{N} \mid n \le x\}$. Hence, in the example, $W_1 = 3$, because $T(z) = z$, $c = 10$ and $Q_J + B = 32$. The second part, $W_2$, is the time until enough packets are present to fill the batch of the tagged packet with at least $l$ packets. Mark that exactly $(Q_J+B) \bmod c$ of the previously arrived packets are served in the same batch as the tagged packet. As $l=5$, $((Q_J+B) \bmod c) = 2$, $X=1$, $A_{J+1}=0$ and $A_{J+2}=3$, $W_2$ takes two slots in the example. The total delay of the tagged packet then equals

$$W = \max(W_1, W_2), \tag{1}$$

since the service of the tagged packet's batch can commence only if all preceding batches have been served, and the packet's batch itself contains at least $l$ packets. Calculation of joint probabilities of $W_1$ and $W_2$ is difficult. Therefore, we propose some lower and upper bounds, that only require calculation of marginal tail probabilities of $W_1$ and $W_2$.

On account of (1), we obtain

$$\begin{aligned}\Pr[W > w] &= \Pr[W_1 > w \vee W_2 > w] \\ &= \Pr[W_1 > w] + \Pr[W_2 > w] - \Pr[W_1 > w \wedge W_2 > w].\end{aligned}$$

The following property paves the path towards establishment of a lower bound:

$$\Pr[W_1 > w \wedge W_2 > w] \le \min(\Pr[W_1 > w], \Pr[W_2 > w]). \tag{2}$$

A lower bound is obtained by assuming that the equality in (2) holds, leading to

$$\Pr[W > w] \ge \max(\Pr[W_1 > w], \Pr[W_2 > w]). \tag{3}$$

An upper bound is established by assuming that $\Pr[W_1 > w \wedge W_2 > w] = 0$, leading to

$$\Pr[W > w] \le \Pr[W_1 > w] + \Pr[W_2 > w]. \tag{4}$$

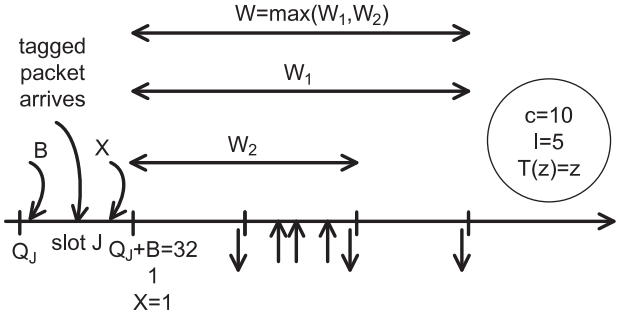


**Fig. 1.** Illustration of $W$, $W_1$ and $W_2$ and introduction of some notations.