Applied Mathematics and Computation 240 (2014) 62–71

Contents lists available at ScienceDirect

Applied Mathematics and Computation

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

Analysis of steady-state and transient delay in discrete-time single-arrival and batch-arrival systems

J. Walraevens *, D. Claeys, H. Bruneel

Department of Telecommunications and Information Processing, Ghent University – UGent, Sint-Pietersnieuwstraat 41, B-9000 Gent, Belgium

ARTICLE INFO

Keywords:
Queueing theory
Generating functions
Transient analysis
Virtual delay

ABSTRACT

We perform an analysis of the transient delay in a discrete-time FIFO buffer with batch arrivals. As transient delay is an ambiguous concept, we first discuss different possible definitions of the term (delay of the kth customer, delay of a customer arriving at time t, etc.). In this paper, we focus on the analysis of the delay of a customer arriving in slot t, also sometimes called virtual delay in single-arrival systems. It turns out that the modeling in batch-arrival systems is more intricate. In analysis, we relate transient delay to transient unfinished work and characterize the latter. Some time-dependent as well as limiting steady-state delay measures are calculated. We also study a variation that is related to active probing measurements. A substantial part of the article finally focuses on some fundamental differences between alternative definitions of transient delay.

1. Introduction

The delay (waiting time, sojourn time) experienced by customers is one of the most important performance measures of a queueing system. There is a vast literature dealing with the analysis of steady-state delay in a broad range of queueing systems. Fewer research efforts have addressed transient delay measures, though it might be of interest to characterize the evolution in time of delay experienced by customers. The latter is the topic of this paper.

The definition of 'transient delay' is not unambiguous. Basically, one can take a *customer-based* approach or a *time-based* approach. With the first approach, transient delay is defined as the delay of the kth arriving (or departing) customer ($k \geqslant 1$). This transient delay is analyzed in for instance [1–8]. With the second approach, transient delay is defined as the delay of a customer arriving at time t ($t \geqslant 0$). In single-arrival systems, this is sometimes called the virtual delay or virtual waiting time at time t, as it is the delay that a 'virtual' customer that would arrive at time t would experience. Examples of analyses of this transient delay can be found in [1,9,10].

We note that the two alternative types of transient delay are substantially different. Only for *single-arrival* systems, the *steady-state* delay distributions are identical, see further. The differences between the *steady-state* delay distributions in case of batch arrivals were discussed by Burke [11]. Burke claims that the customer-based approach (as was done by himself) is the correct one, and the time-based approach (as was used in papers and books before Burke's paper appeared) is erroneous. This is a valid point of view if the steady-state delay is concerned, since one basically wants to characterize the delay of a random customer. However, in case of transient delay, the time-based approach makes sense as well; one does not necessarily know the ordinal number of a particular customer (the customer-based approach), while one might know the time of

* Corresponding author.
E-mail addresses: jw@telin.UGent.be (J. Walraevens), dclaeys@telin.UGent.be (D. Claeys), hb@telin.UGent.be (H. Bruneel).

http://dx.doi.org/10.1016/j.amc.2014.04.071
0096-3003/

arrival of this customer (time-based approach). Therefore, characterization of the delay of a customer arriving at time t is interesting. It is also of interest to analyze the fundamental differences between both approaches. These two issues are the subject of the current contribution.

More precisely, we analyze the delay of a customer arriving during slot t ($t \geqslant 0$) in a discrete-time $Geo^X/G/1$ queue with a FIFO (First-In-First-Out) scheduling discipline. This paper is therefore complementary to our paper [8], where the delay of the kth arriving customer ($k \geqslant 1$) in the same queue is analyzed. We note that almost all articles characterizing transient delay assume *single-arrival* queueing systems. One of the main challenges in the current analysis, however, is how to deal with the batch nature of the arrival process. Indeed, in single-arrival systems, the delay is equal to the sum of the unfinished work at the arrival instant (i.e., the time needed to process all customers present in the system at that time) and the service time of the arriving customer. In batch-arrival systems, other customers can arrive at the same instant as the virtual customer, so we have to deal with this issue. The queueing model is described in more detail in the following section and the analysis of the delay of a customer arriving in slot t is presented in Section 3. We relate the probability generating functions (pgf) of the delay of a customer arriving in slot t and of the unfinished work at the beginning of slot t and use this relation to compute the transform function of the former sequence of pgfs (for all $t \geqslant 0$). In Section 4, we calculate the pgf for $t \to \infty$. The transient mean delay is characterized in Section 5. We then treat a slightly different model in Section 6, where we assume the virtual customer to be an additional 'test' customer. This is, for instance, useful in active probing measurements, as extra customers are injected into the system in this type of measurements and the delay of these customers is measured (see, for instance, [12]). Different definitions of (transient) delay lead to different results. We therefore devote a substantial part of this article (Section 7) to a comparison between the different transient (and steady-state) delays, and investigate fundamental identities and differences. We also provide some numerical examples. We conclude with some final remarks in the last section.

2. Queueing model

We consider a discrete-time single-server queueing system with infinite buffer space. Customers are served on a First-Come-First-Served basis. The numbers of arrivals during the consecutive time units (denoted as slots) are independent and identically distributed (i.i.d.) stochastic variables. Denote the number of arrivals in an arbitrarily chosen slot by a. We use the notations $a(n)$ and $A(z)$ to indicate its probability mass function (pmf) and probability generating function (pgf) respectively, i.e., $a(n) \triangleq \text{Prob}[a = n]$, $n \geqslant 0$ and $A(z) \triangleq \text{E}[z^a] = \sum_{n=0}^{\infty} a(n) z^n$. The service times of customers are defined as the numbers of slots it takes to serve them and are assumed to be generally distributed with pmf $s(n), n \geqslant 1$, and pgf $S(z) = \sum_{n=1}^{\infty} s(n) z^n$. The mean number of arrivals in a slot and the mean service time are given by $\lambda = A'(1)$ and $1/\mu = S'(1)$, respectively. The load equals $\rho = \lambda/\mu$.

3. Analysis

We are interested in the delay of an arbitrary customer arriving in slot t, i.e., from all customers arriving in slot t, we select one customer (the tagged customer) at random and analyze his delay (the number of slots he resides in the system, not including his slot of arrival). Denote this delay by d_t, for all $t \geqslant 0$. Note that slot t is a 'special' slot, as we assume that at least one customer arrives in this slot.

We first introduce the concept 'unit of work'. We may assume that each customer consists of a number of units of work equal to its service time and that the server processes at the rate of one unit of work per slot. We can then relate d_t to the unfinished work w_t in the system at the beginning of slot t (defined as the number of units of work that are in the system at the beginning of slot t), the amount of work units f_t arriving in slot t and to be executed before the tagged customer, and the service time s_t of the tagged customer himself:

$$d_t = (w_t - 1)^+ + f_t + s_t, \tag{1}$$

with $(x)^+ = \max(x, 0)$.

As mentioned, the distribution of the number of arrivals in slot t is different from the distribution of the number of arrivals in a random slot, as it is assumed that at least one customer arrives in slot t. However, since the numbers of per-slot arrivals are i.i.d., the special nature of slot t has no impact on the unfinished work w_t *at the beginning of that slot*, and, therefore, w_t is equally distributed as the unfinished work at the beginning of slot t in a system with a number of per-slot arrivals with pgf $A(z)$ in *all* slots (also in slot t). For that reason, f_t is the only variable of the right-handside of (1) that is affected by the special nature of slot t.

We now translate Eq. (1) to probability generating functions. We have:

$$D_t(z) \triangleq \text{E}[z^{d_t}] = \text{E}\left[z^{(w_t-1)^+ + f_t + s_t}\right] = \frac{S(z)F(z)}{z}(W_t(z) + (z-1)W_t(0)), \tag{2}$$

where $F(z) \triangleq \text{E}\left[z^{f_t}\right]$ and $W_t(z) \triangleq \text{E}[z^{w_t}]$. We express $F(z)$ as a function of $A(z)$ later. Let us first take the z-transform (or x-transform in this case) of the previous equation with respect to t. This will allow us to use results obtained in [13] and to find (semi-) closed-form results. We define