European Journal of Operational Research 199 (2009) 170–175

Contents lists available at ScienceDirect

European Journal of Operational Research

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

Stochastics and Statistics

Calculation of delay characteristics for multiserver queues with constant service times

Peixia Gao, Sabine Wittevrongel *, Joris Walraevens, Marc Moeneclaey, Herwig Bruneel

Department of Telecommunications and Information Processing (TELIN), Ghent University, Sint-Pietersnieuwstraat 41, B-9000 Gent, Belgium

ARTICLE INFO

Article history:
Received 28 June 2007
Accepted 21 October 2008
Available online 29 October 2008

Keywords:
Queueing
Discrete time
Multiple servers
Constant service times
Delay analysis

ABSTRACT

We consider a discrete-time infinite-capacity queueing system with a general uncorrelated arrival process, constant-length service times of multiple slots, multiple servers and a first-come-first-served queueing discipline. Under the assumption that the queueing system can reach a steady state, we first establish a relationship between the steady-state probability distributions of the system content and the customer delay. Next, by means of this relationship, an explicit expression for the probability generating function of the customer delay is obtained from the known generating function of the system content, derived in previous work. In addition, several characteristics of the customer delay, namely the mean value, the variance and the tail distribution of the delay, are derived through some mathematical manipulations. The analysis is illustrated by means of some numerical examples.

1. Introduction

Discrete-time queueing models have received considerable attention during the past years, see e.g. the books [1,5,12,14,16,18] and the references therein. A main reason is the applicability of these models in the performance evaluation of packet-based high-speed telecommunication networks, where buffers are used for the temporary storage of information packets which cannot be transmitted to their destination immediately. The information packets then constitute the customers of the queueing system and the transmission of packets corresponds to the service of customers. In discrete-time queueing models, the time axis is divided into fixed-length intervals, referred to as slots, and the service of customers can start or end at slot boundaries only. The latter implies that the service times of the customers consist of an integer number of slots.

Usually, the performance of a queueing system is expressed in terms of such quantities as the system content (i.e., the total number of customers present in the queueing system) and the delay of a customer (i.e., the time (in slots) spent by a customer in the system). Especially when multimedia applications in packet-based networks are concerned, it is important to be able to accurately predict the characteristics of the packet delay, such as the mean delay and the delay jitter, in order to guarantee acceptable delay boundaries for the admitted network traffic. The analysis of delay characteristics in the current internet thus is an important research topic. There are a number of performance analysis techniques for discrete-time systems, ranging from computer simulation to the numerical solution of the associated set of balance equations and various types of analytical methods. Computer simulation often suffers from long run times and requires a new run for each parameter setting. Hence, for performance engineering purposes, analytical methods are preferred [11], since these lead to closed-form expressions for the performance measures of interest and therefore allow a fast performance prediction.

If we focus our attention on analytical performance studies, many results have been obtained for both the system content and the customer delay in a single-server environment. In case systems with multiple servers are considered, fewer analytical results are however available, although such systems occur in many practical applications, for instance, in output-buffering switches in the nodes of packet-based networks (see Section 6 for more details). Most studies of multiserver systems assume constant service times equal to one slot, see e.g. [2 and 17]. Multiserver systems with geometrically distributed service times have been considered in [7–9 and 15]. In [3 and 4], discrete-time queueing models with multiple servers and constant service times of multiple slots have been studied, but only results in connection with the system content have been derived. This deterministic service-time distribution has several applications, for instance, in the performance analysis of packet switches with a different internal and external transfer mode, as explained in [3].

In this paper, we will extend the analysis of [3] in order to investigate the characteristics of the delay, which is one of the most important performance metrics from a user perspective [11]. First, a relationship between the steady-state probability distributions of the customer delay and the system content is

* Corresponding author. Tel.: +32 9 264 89 01; fax: +32 9 264 42 95.
E-mail address: sw@telin.UGent.be (S. Wittevrongel).

0377-2217/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.ejor.2008.10.017

established. Then, from the results for the system content derived in [3], an explicit expression for the probability generating function (PGF) of the customer delay is obtained. Finally, from this PGF several delay-related characteristics, namely the mean delay, the variance of the delay and the probability that the delay exceeds a given threshold, are calculated. A preliminary version of this work can be found in [10].

The remainder of the paper is organized as follows: in Section 2, we describe the class of discrete-time queueing systems under study and introduce some notations. Some results of [3], which will be used in the paper, are summarized in Section 3. For the considered class of queueing systems, we establish a relationship between the steady-state PGFs of the system content and the customer delay in Section 4. In Section 5, the performance measures for the customer delay are presented. In Section 6, some numerical examples are given to illustrate the analysis and the usefulness of the results. Finally, the paper is concluded in Section 7.

2. Mathematical model

In this paper, we consider a discrete-time multiserver queueing system with $c(c \geqslant 1)$ servers. The time axis is divided into fixed-length intervals, referred to as slots. Customers arrive at the input of the system according to a general independent arrival process, i.e., the numbers of customer arrivals during the consecutive slots are assumed to be independent and identically distributed (i.i.d.) random variables; we denote their common PGF by $A(z)$. Customers are then queued until they can be served by one of the c servers based on a first-come-first-served (FCFS) discipline. The queue has an infinite storage capacity for customers. The service of a customer can start or end at slot boundaries only. In this paper, the service times of the customers are assumed to be constant equal to $s(s \geqslant 1)$ slots. Moreover, the service and arrival processes are assumed to be mutually independent. Finally, in the analysis that follows it is assumed that the queueing system can reach a steady state. Such a steady state exists if the mean number of customer arrivals during an arbitrary slot $(A'(1))$ is strictly less than the mean number of customers that can be served per slot (c/s), i.e., if the load

$$\rho \triangleq \frac{sA'(1)}{c} < 1. \tag{1}$$

3. Preliminary results

Let us denote by v_k the system content (i.e., the total number of customers in the queueing system, including the customers in service, if any) at the beginning of slot k and by a_k the number of arriving customers during slot k. Furthermore, let $u_{j,k}(0 \leqslant j \leqslant s-1)$ indicate the total number of customers in the system at the beginning of slot k whose service has progressed for at most j slots. Note that no customers in the system have received more than $s-1$ slots of service due to the constant nature of the service times (customers who have received s slots of service are no longer in the system). In [3], it was shown that the following set of system equations can then be established:

$$v_k = u_{s-1,k}, \tag{2}$$

$$u_{j,k+1} = u_{j-1,k} + a_k, \quad \text{for } 1 \leqslant j \leqslant s-1, \tag{3}$$

and

$$u_{0,k+1} = (u_{s-1,k} - c)^+ + a_k, \tag{4}$$

where $(\cdot)^+ = \max(0,\cdot)$. We moreover introduce the notation $u_{-1,k} = (u_{s-1,k} - c)^+$ to indicate the number of customers in the system at the beginning of slot k and not being served during slot k. In the steady state the distributions of the random variables v_k and $u_{j,k}$ become independent of the time index k. We denote by $V(z)$ and $U_j(z)$ the equilibrium PGFs of v_k and $u_{j,k}$, respectively. Eqs. (2)–(4) were used in [3] to derive the following expressions for the PGFs $V(z)$ and $U_j(z)$:

$$V(z) = c(1-\rho)\frac{(z-1)A(z)^s}{z^c - A(z)^s}\prod_{i=1}^{c-1}\frac{z-z_i}{1-z_i}, \tag{5}$$

where $z_i(1 \leqslant i \leqslant c-1)$ are the $c-1$ zeros inside the unit disk $\{z : |z| < 1\}$ of $z^c - A(z)^s$, and

$$U_j(z) = \frac{V(z)}{A(z)^{s-j-1}}, \quad \text{for } -1 \leqslant j \leqslant s-1. \tag{6}$$

In the Appendix, we give an alternative, more intuitive derivation of $V(z)$. In the main part of this paper, we will study the delay characteristics for the considered queueing model.

4. Relationship between system content and customer delay

We define the delay of a customer as the total number of slots between the end of the slot during which the customer arrived in the system and the end of the slot where the service of the customer finishes and the customer leaves the system. In this section, we prove the following relationship between the steady-state PGF $V(z)$ of the system content at the start of an arbitrary slot and the steady-state PGF $D(z)$ of the delay of an arbitrary customer:

$$D(z^c) = \frac{z^{cs}(1-z^c)}{cz^{cs}A'(1)}\sum_{j=0}^{c-1}\frac{\beta^j z^s}{(1-\beta^j z^s)^2} \times \frac{[z^{cs} - A(\beta^j z^s)^s][A(\beta^j z^s) - 1]}{A(\beta^j z^s)^s[A(\beta^j z^s) - z^c]}V(\beta^j z^s), \tag{7}$$

with $\beta \triangleq \exp(2\pi I/c)$, and where I is the imaginary unit $(I^2 = -1)$.

Proof. Let us consider an arbitrary customer P (referred to as the tagged customer), that arrives in the queueing system during some slot J in the steady state. Let d with PGF $D(z)$ denote the delay of P. Also define the waiting time of a customer as the number of slots between the end of the customer's arrival slot and the beginning of the slot where the service of the customer starts. The delay of a customer is equal to the sum of the waiting time and the service time of the customer and thus we can express the PGF $D(z)$ as

$$D(z) = z^s W(z), \tag{8}$$

where $W(z)$ denotes the PGF of the waiting time w of P.

We now concentrate on the derivation of the PGF $W(z)$. First, we make the following observations.

- The waiting time of the tagged customer P depends on the customers in the system right after slot J with service priority over P.
- As long as there are at least c customers in the system with service priority over P, P is still waiting for service and the c servers are all busy serving customers.
- Since each customer requires exactly s slots of service, there will be exactly c departures during each frame of s consecutive slots as long as P is still waiting for service.
- In view of the FCFS discipline, the number of customers in front of P right after slot J that still need to receive at least $i(1 \leqslant i \leqslant s)$ slots of service at the beginning of slot $J+1$ consists of the $u_{s-i-1,J}$ customers that arrived before slot J on the one hand, and the customers that arrived in slot J but before P on the other hand.

Based on these observations, it is then easily seen that if and only if $u_{s-i-1,J} + f \geqslant c$, where f is the number of arrivals in slot J