View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Ghent University Academic Bibliography

Calculation of delay characteristics for multiserver queues with constant service times

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

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

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.

Editor: Artalejo

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

[*] Corresponding author, Phone: +32-9-264 89 01, Fax: +32-9-264 42 95, E-mail: sw@telin.UGent.be

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 Sect. 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], [8], [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 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 Sect. 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 Sect. 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 Sect. 4. In Sect. 5, the performance measures for the customer delay are presented. In Sect. 6, some numerical examples are given to illustrate the analysis and the usefulness of the results. Finally, the paper is concluded in Sect. 7.

2 Mathematical Model

In this paper, we consider a discrete-time multiserver queueing system with c ($c \geq 1$) 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 ($s \geq 1$) 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 $u_{j,k}$ ($0 \leq j \leq s-1$) 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:

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

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

and

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

where $(.)^+ = \max(0, .)$. 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. Equations (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 $(1 \leq i \leq c-1)$ are the $c-1$ zeros inside the unit disk $\{z : |z| < 1\}$ of $z^c - A(z)^s$, and

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

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:

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

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 $(1 \leq i \leq s)$ 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 $u_{s-i-1,J} + f \geq c$, where f is the number of arrivals in slot J before P, the waiting time of P will be at least i slots. For each extra group of c customers in $u_{s-i-1,J} + f$, P will have to wait another s slots extra. We may therefore conclude that

$$w \geq \ell s + i \quad \Leftrightarrow \quad u_{s-i-1,J} + f \geq \ell c + c ,$$

or equivalently

$$w \geq \ell s + i \quad \Leftrightarrow \quad \widetilde{q}_i \geq \ell c\,, \quad \text{for } \ell \geq 0\,, \quad 1 \leq i \leq s\,, \tag{9}$$

with

$$\widetilde{q}_i \triangleq f + u_{s-i-1,J} - c\,. \tag{10}$$

Secondly, we transform the relationship (9) between the random variables w and $\widetilde{q}_i$ ($1 \leq i \leq s$) into a relationship between their PGFs. To this end, we use the identity

$$\sum_{n=1}^{\infty} \text{Prob}[w \geq n] z^n = \frac{z[W(z) - 1]}{z - 1}\,. \tag{11}$$

From this identity and equation (9), it follows that

$$\begin{aligned} \frac{z^c[W(z^c) - 1]}{z^c - 1} &= \sum_{i=1}^{s} \sum_{\ell=0}^{\infty} \text{Prob}[w \geq \ell s + i] z^{c(\ell s + i)} \\ &= \sum_{i=1}^{s} \sum_{\ell=0}^{\infty} \text{Prob}[\widetilde{q}_i \geq \ell c] z^{s\ell c + ci} \\ &= \sum_{i=1}^{s} z^{ci} \sum_{m=0}^{\infty} \text{Prob}[\widetilde{q}_i \geq m] z^{sm} \sum_{\ell=0}^{\infty} \delta(m - \ell c)\,, \end{aligned} \tag{12}$$

where $\delta(.)$ is the Kronecker delta function, which is zero unless its argument is zero, in which case it is equal to 1. Since $m \geq 0$ in the above expression, the lower limit of the sum over ℓ in (12) can be replaced by $-\infty$ without any influence on the result. We can then eliminate the Kronecker delta functions from (12) by using the following identity:

$$\frac{1}{c} \sum_{j=0}^{c-1} \beta^{Kj} = \sum_{\ell=-\infty}^{\infty} \delta(K - \ell c)\,, \quad \text{with } \beta \triangleq \exp(2\pi I / c)\,. \tag{13}$$

This identity expresses that the left-hand side of (13) equals zero unless the integer K is a multiple of c, in which case it is equal to 1. By using (13) in (12), we obtain

$$\begin{aligned}\frac{z^c[W(z^c)-1]}{z^c-1} &= \frac{1}{c}\sum_{i=1}^{s} z^{ci}\sum_{j=0}^{c-1}\sum_{m=0}^{\infty}\text{Prob}[\widetilde{q}_i \geq m](z^s\beta^j)^m \\ &= \frac{1}{c}\sum_{j=0}^{c-1}\sum_{i=1}^{s}\frac{1-\beta^j z^s\widetilde{Q}_i(\beta^j z^s)}{1-\beta^j z^s}z^{ci}\,, \end{aligned} \tag{14}$$

where $\widetilde{Q}_i(z)$ is the PGF of $\widetilde{q}_i$, and where we have also used the identity

$$\sum_{m=0}^{\infty}\text{Prob}[\widetilde{q}_i \geq m]z^m = \frac{1-z\widetilde{Q}_i(z)}{1-z}\,. \tag{15}$$

Thirdly, we relate the PGFs $\widetilde{Q}_i(z)$, $1 \leq i \leq s$, to the PGF $V(z)$ of the system content v at the start of an arbitrary slot. This can be done based on the definition (10) of $\widetilde{q}_i$. In view of the uncorrelated nature of the customer arrival process, the random variables f and $u_{s-i-1,J}$ on the right-hand side of (10) are statistically independent and the PGF of $u_{s-i-1,J}$ at the beginning of *slot* J is given by the PGF $U_{s-i-1}(z)$ at the beginning of an *arbitrary slot* in the steady state. Hence, we have

$$\widetilde{Q}_i(z) = \frac{U_{s-i-1}(z)F(z)}{z^c}\,, \tag{16}$$

where $F(z)$ is the PGF of f, which can be shown to be (see e.g. [1])

$$F(z) = \frac{A(z)-1}{A'(1)(z-1)}\,. \tag{17}$$

Combination of (16) and (6) yields

$$\widetilde{Q}_i(z) = \frac{F(z)V(z)}{z^cA(z)^i}\,, \quad \text{for } 1 \leq i \leq s\,. \tag{18}$$

Finally, substituting (18) into (14), working out the sum over i, using the property that $\beta^{jc} = 1$ regardless of the value of j and the identity

$$\frac{1}{c}\sum_{j=0}^{c-1}\frac{1}{1-\beta^j z^s} = \frac{1}{1-z^{cs}}\,, \tag{19}$$

which is easily shown based on equation (13), we obtain

$$W(z^c) = \frac{1-z^c}{cz^{cs}} \sum_{j=0}^{c-1} \frac{\beta^j z^s}{1-\beta^j z^s} \frac{[A(\beta^j z^s)^s - z^{cs}]F(\beta^j z^s)V(\beta^j z^s)}{A(\beta^j z^s)^s[A(\beta^j z^s) - z^c]} \,. \tag{20}$$

Combination of (20) with (8) then leads to the desired relationship between the steady-state PGFs of the system content v at the beginning of an arbitrary slot and the delay d of an arbitrary customer. □

5 Characteristics of the Customer Delay

From relationship (7) between the PGFs of system content and delay and from the known expression (5) for the PGF $V(z)$ of the system content, we find the following explicit formula for the PGF of the delay experienced by an arbitrary customer in the steady state:

$$D(z^c) = \frac{1-\rho}{A'(1)} \sum_{j=0}^{c-1} \frac{1-z^c}{1-(\beta^j z^s)^{-1}} \frac{A(\beta^j z^s)-1}{A(\beta^j z^s)-z^c} \prod_{i=1}^{c-1} \frac{\beta^j z^s - z_i}{1-z_i} \,. \tag{21}$$

In the rest of this section, we will use the expression for $D(z^c)$ to derive some important characteristics of the customer delay.

5.1 Moments of the Delay

The mean value of the customer delay can be found by evaluation of the first-order derivative of the PGF $D(z^c)$ with respect to z at $z = 1$. Specifically,

$$E[d] = D'(1) = \frac{1}{c} \left. \frac{\mathrm{d}D(z^c)}{\mathrm{d}z} \right|_{z=1} \,, \tag{22}$$

where $D(z^c)$ is given in (21). After some mathematical manipulations, we find

$$E[d] = \frac{E[v]}{A'(1)} \,, \tag{23}$$

which proves that our result fully agrees with Little's theorem ([6]). In a similar way, we can also obtain higher-order moments of the customer delay from (21), by calculating the appropriate higher-order derivatives of $D(z^c)$ at $z = 1$. For instance, the variance of the customer delay (delay jitter) can be expressed as

$$\mathrm{var}[d] = D''(1) + D'(1) - D'(1)^2 \,, \tag{24}$$

where $D''(1)$, the second derivative of $D(z)$ in $z = 1$, can be obtained from (21) as

$$D''(1) = \frac{1}{c^2} \left. \frac{\mathrm{d}^2 D(z^c)}{\mathrm{d}z^2} \right|_{z=1} - \frac{c-1}{c} D'(1) \,. \tag{25}$$

5.2 Tail Probabilities of the Delay

The aim of this subsection is to determine the tail distribution of the customer delay, i.e., the probability that the delay equals a given value n, for a sufficiently large value of n. In principle, we can determine the tail distribution of a discrete random variable by applying the inversion formula for z-transforms and Cauchy's residue theorem from complex analysis (see e.g. [13]) on its generating function and keeping only the contribution of the pole (or poles) of the PGF with smallest modulus outside the unit disk, as explained in [1]. From the expression (21) for $D(z^c)$, we find that $D(z^c)$ has c poles with the same smallest modulus. These poles are given by

$$z_d(m) = \beta^{-m} z_v^{1/s} \,, \quad \text{for } m = 0, \ldots, c-1 \,, \tag{26}$$

where z_v is the dominant pole of the PGF $V(z)$ of the system content, i.e., the zero of $z^c - A(z)^s$ outside the unit disk with the smallest modulus. Indeed, it is easy to show that $z_d(0) = z_v^{1/s}$ is the zero with minimal modulus outside the unit disk of the factor $[A(z^s) - z^c]$ in the denominator of $D(z^c)$. Moreover, since z^c remains unchanged when z is multiplied by β^{-m}, it is clear that $z_d(m) = \beta^{-m} z_v^{1/s}$ is also a pole of $D(z^c)$ with the same modulus ${z_v}^{1/s}$. In particular, it can be shown that the pole $z_d(m)$ is a zero of the factor $[A(\beta^j z^s) - z^c]$ in the denominator of $D(z^c)$ for which $j = (ms) \bmod c$, i.e., for which j equals the remainder of the division of ms by c. Taking into account all the poles $z_d(m)$ of $D(z^c)$ with minimal modulus and keeping in mind that $\mathrm{Prob}[d = n]$ is the coefficient of z^{cn} in the series expansion of $D(z^c)$, we finally obtain the following expression for $\mathrm{Prob}[d = n]$

for sufficiently large n:

$$\begin{aligned}\text{Prob}[d=n] &\approx -\sum_{m=0}^{c-1} \frac{b_m}{z_d(m)} [z_d(m)]^{-cn} \\ &= -\sum_{m=0}^{c-1} \frac{b_m}{z_d(m)}\; z_v^{-cn/s} \\ &= -C_d\; {z_v}^{-cn/s}\,, \end{aligned} \tag{27}$$

where b_m is the residue of $D(z^c)$ in the point $z = z_d(m)$ and where we have used the property that $\beta^{mc} = 1$. The residue b_m is given by

$$b_m = \frac{N_m(z_d(m))}{{R_m}'(z_d(m))}\,,$$

where $N_m(z)$ and $R_m(z)$ are the numerator and the denominator, respectively, of the term of (21) corresponding to the index value $j = (ms) \bmod c$. Using the expression (21), we find

$$C_d = \sum_{m=0}^{c-1} \frac{b_m}{z_d(m)} = \frac{c(1-\rho)}{A'(1)}\, \frac{1-z_v^{c/s}}{1-{z_v}^{-1}}\, \frac{A(z_v)-1}{s z_v A'(z_v) - c{z_v}^{c/s}} \prod_{i=1}^{c-1} \frac{z_v - z_i}{1-z_i}\,. \tag{28}$$

The probability that the customer delay exceeds a given threshold T can be easily derived from (27) as

$$\text{Prob}[d>T] \approx -C_d \frac{z_v^{-cT/s}}{{z_v}^{c/s}-1}\,. \tag{29}$$

6 Numerical Examples and Discussion

In order to illustrate the results obtained above, we discuss a number of numerical examples in this section. In a first set of examples, we assume the number of customers that arrive during a slot has a geometric distribution, i.e.,

$$A(z) = \frac{1}{1+\lambda-\lambda z}\,,$$

where λ denotes the mean number of customer arrivals per slot. The basic influence of the number of servers c and the length of the service times s on the delay characteristics is illustrated in the Figs. 1-3.

In Fig. 1, the mean customer delay is plotted versus the load ρ for various values of c and s. For given values of c and s, we observe that the mean customer delay increases with increasing values of ρ. We also note that all the curves have a vertical asymptote at $\rho = 1$. For a given ρ, the mean delay increases as the service times become longer, although a higher number of servers can compensate this effect to some extent.

In Fig. 2, the variance of the customer delay is shown versus ρ, for $c = 4$, 8 and for $s = 1$, 4, 8. We see that for given values of c and ρ, the variance of the customer delay also increases as the length of the service times increases. For given values of s and ρ, the variance of the delay is higher when less servers are used. Also note that the variance of the delay for $c = 4$, $s = 4$ is almost the same as for $c = 8$, $s = 8$, especially for high load. As can be observed from Fig. 1, this is not the case for the mean customer delay, which always consists of at least s slots. Remark that for $c = s$, the delay is still slightly less variable for higher c and s, as expected intuitively.

In Fig. 3, the probability that the delay exceeds some given threshold T is plotted versus T for $\rho = 0.8$, $s = 5$ and for three different numbers of servers, namely $c = 1$, 4 and 8. Clearly, the probability of having long delays decreases as the number of servers increases, in accordance with our intuitive feeling.

As a second, more practical example we consider an $N \times N(c)$ packet switch as shown in Fig. 4. The switch has N inlets and N outlets. The outlets are organized in $\frac{N}{c}$ different destination groups, each group containing exactly c outlets, and one separate output buffer is provided for each destination group. We assume here that packets enter the switch via the inlets according to independent Bernoulli arrival streams and incoming packets are routed independently and uniformly to one of the $\frac{N}{c}$ destination groups. The PGF of the number of packet arrivals per slot in an output buffer of the switch is then given by

$$A(z) = \left(1 - \frac{pc}{N} + \frac{pc}{N}z\right)^N ,$$

where p denotes the probability of a packet arrival on a switch inlet. Our analysis can then be used to study the packet delay in an output buffer.

In Fig. 5, we have plotted the probability that the packet delay exceeds some given threshold T versus T, for $N = 8$, $c = 2$, $s = 5$ and several values of the load $\rho = ps$. These curves can be used to determine the 10^{-n} quantile of the packet delay, for some integer value of n, i.e., the value T such that

$$\text{Prob}[d > T] = 10^{-n} \, .$$

For instance, for a load $\rho = 0.7$, the 10^{-6} quantile of the delay is given by about 52 slots.

Alternatively, the results of our analysis can be used to estimate the admissible traffic load in order to satisfy a specified delay constraint. In Fig. 6, the maximum load ρ such that $\text{Prob}[d > T] \leq 10^{-6}$ is plotted versus the value of the threshold T, for $N = 8$, $s = 5$ and various values of c. As expected, the maximum admissible load is higher for higher T and higher c. We see that the maximum load can be much lower than 0.5 when the delay constraint is stringent (low T) and when the number of servers is low. The increase of the maximum load with increasing T is largest for small T and small c, while the curves for high c are almost horizontal lines for high T. This type of results are easily obtainable from our analysis and are highly interesting for a network operator.

7 Concluding Remarks

In this paper, we have studied the delay characteristics of a discrete-time infinite-capacity queueing system with multiple servers, constant service times of arbitrary length and a general independent arrival process. The study is an extension of previous work ([3]), which was concerned with the analysis of the system content for this type of multiserver queueing system. In the present paper, we have established a relationship between the PGFs of the customer delay and the system content, using an analytical technique based on generating functions. Then from the result for the PGF of the system content, known from [3], we have obtained an explicit expression for the PGF of the customer delay, as well as closed-form expressions for the mean value, the variance and the tail distribution of the customer delay. The obtained analytical results are easy to evaluate numerically. Some numerical results have been presented to illustrate the analysis and the usefulness of the results.

The analyzed queueing model has practical applications in the domain of digital communication networks, such as circuit-switched TDMA systems, switching elements and traffic concentrators. We note that besides the characteristics of the system content studied in [3] for the considered model, the derived results for the customer delay (i.e., performance measures like the mean delay, the delay jitter and the probability that the customer delay exceeds a given threshold) are also very important for a network designer to guarantee the quality of service of the network.

Acknowledgement

The third author is a Postdoctoral Fellow with the Fund for Scientific Research, Flanders (F.W.O.-Vlaanderen), Belgium.

Appendix: Alternative Derivation of $V(z)$

In this Appendix, we give an alternative, more intuitive derivation (as compared to [3]) of the steady-state PGF $V(z)$ of the system content at the start of an arbitrary slot.

We observe the evolution of the system content over a frame of s consecutive slots. If the system content at the beginning of the frame is at least c, exactly c customers will leave the system during the frame. On the other hand, if there are less than c, say n, customers in the system at the beginning of the frame, there will be n customer departures during the frame. We thus have the following system equation:

$$v_{k+s} = (v_k - c)^+ + \sum_{j=k}^{k+s-1} a_j \, .$$

By transforming this equation into the z-domain, we then immediately get an equation for the steady-state PGF $V(z)$:

$$V(z) = A(z)^s \left\{ \sum_{n=0}^{c-1} \text{Prob}[v = n](1 - z^{n-c}) + z^{-c} V(z) \right\} \, .$$

Solving this equation for $V(z)$ and determining the boundary probabilities as explained in [3], we obtain expression (5). So, the PGF $V(z)$ can be derived without the introduction of the random variables $u_{j,k}$ (see Sect. 3). Note however that these variables are needed for the delay analysis.

References

[1] H. Bruneel and B.G. Kim, Discrete-time models for communication systems including ATM, Kluwer Academic Publishers, Boston, 1993.

[2] H. Bruneel, B. Steyaert, E. Desmet and G.H. Petit, An analytical technique for the derivation of the delay performance of ATM switches with multiserver output queues, International Journal of Digital and Analog Communication Systems, vol. 5 (1992), pp. 193-201.

[3] H. Bruneel and I. Wuyts, Analysis of discrete-time multiserver queueing models with constant service times, Operations Research Letters, vol. 15, no. 5 (1994), pp. 231-236.

[4] M.L. Chaudhry and N.K. Kim, A complete and simple solution for a discrete-time multi-server queue with bulk arrivals and deterministic service times, Operations Research Letters, vol. 31, no. 2 (2003), pp. 101-107.

[5] H. Daduna, Queueing networks with discrete time scale: explicit expressions for the steady state behavior of discrete time stochastic networks, Springer, New York, 2001.

[6] D. Fiems and H. Bruneel, A note on the discretization of Little's result, Operations Research Letters, vol. 30, no. 1 (2002), pp. 17-18.

[7] P. Gao, S. Wittevrongel and H. Bruneel, Delay against system contents in discrete-time G/Geom/c queue, Electronics Letters, vol. 39, no. 17 (2003), pp. 1290-1292.

[8] P. Gao, S. Wittevrongel and H. Bruneel, Discrete-time multiserver queues with geometric service times, Computers & Operations Research, vol. 31, no. 1 (2004), pp. 81-99.

[9] P. Gao, S. Wittevrongel and H. Bruneel, On the behavior of multiserver buffers with geometric service times and bursty input traffic, IEICE Transactions on Communications, vol. E87-B, no. 12 (2004), pp. 3576-3583.

[10] P. Gao, S. Wittevrongel and H. Bruneel, Delay analysis for a discrete-time GI-D-c queue with arbitrary-length service times, Proceedings of the First European Performance Engineering Workshop, EPEW 2004 (Toledo), Lecture Notes in Computer Science, vol. 3236 (2004), pp. 184-195.

[11] P. Harrison, Performance engineering and stochastic modelling, Proceedings of the Second European Performance Engineering Workshop, EPEW 2005 (Versailles), Lecture Notes in Computer Science, vol. 3670 (2005), pp. 1-14.

[12] J.J. Hunter, Mathematical techniques of applied probability, Volume 2, Discrete time models: techniques and applications, Academic Press, New York, 1983.

[13] L. Kleinrock, Queueing systems, Volume I: Theory, Wiley, New York, 1975.

[14] T.G. Robertazzi, Computer networks and systems: queueing theory and performance evaluation, Springer, New York, 2000.

[15] I. Rubin and Z. Zhang, Message delay and queue-size analysis for circuit-switched TDMA systems, IEEE Transactions on Communications, vol. 39, no. 6 (1991), pp. 905-914.

[16] H. Takagi, Queueing analysis, A foundation of performance evaluation, Volume 3: discrete-time systems, North-Holland, Amsterdam, 1993.

[17] B. Vinck and H. Bruneel, Delay analysis of multiserver ATM buffers, Electronics Letters, vol. 32, no. 15 (1996), pp. 1352-1353.

[18] M.E. Woodward, Communication and computer networks: modelling with discrete-time queues, Pentech Press, London, 1993.

Figure captions

Figure 1 Mean customer delay versus load ρ.

Figure 2 Variance of the customer delay versus load ρ. For each value of s, the upper curve corresponds to $c = 4$ and the lower curve to $c = 8$.

Figure 3 Tail distribution of customer delay, $\text{Prob}[d > T]$, versus T, for $\rho = 0.8$ and $s = 5$.

Figure 4 $N \times N(c)$ switch with output buffers.

Figure 5 Tail distribution of packet delay, $\text{Prob}[d > T]$, versus T, for $N = 8$, $c = 2$, $s = 5$ and $\rho = 0.1, 0.2, \ldots, 0.9$.

Figure 6 Maximum load ρ such that $\text{Prob}[d > T] \leq 10^{-6}$ versus T, for $N = 8$ and $s = 5$.

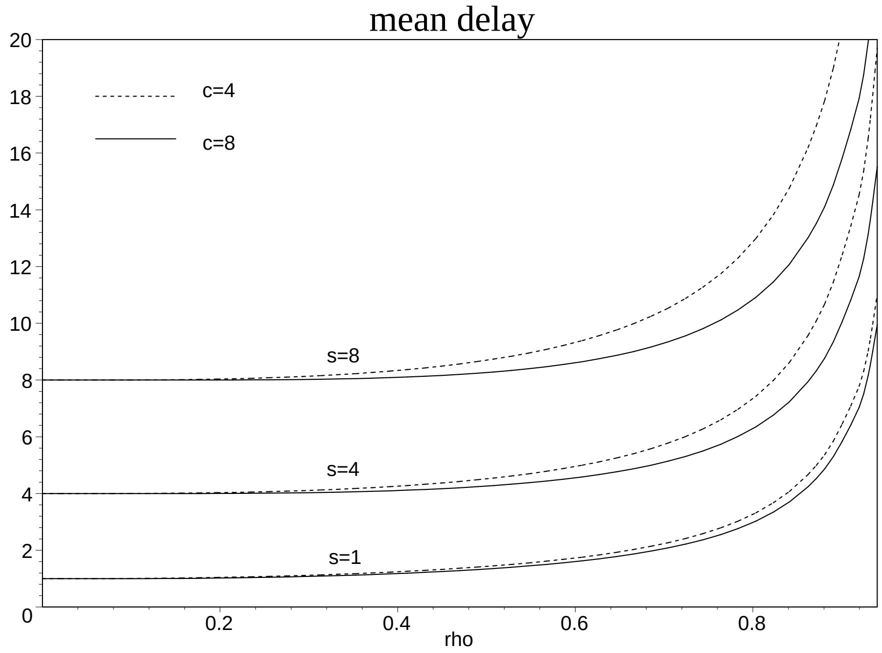

Figure 1: Mean customer delay versus load ρ.

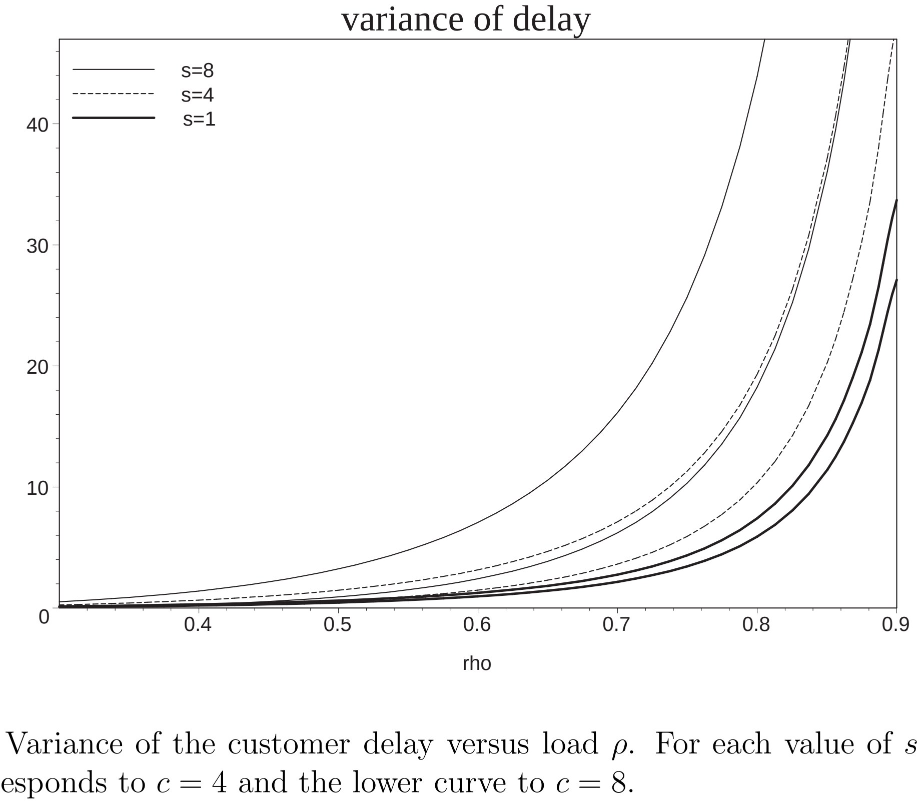

Figure 2: Variance of the customer delay versus load ρ. For each value of s, the upper curve corresponds to $c = 4$ and the lower curve to $c = 8$.

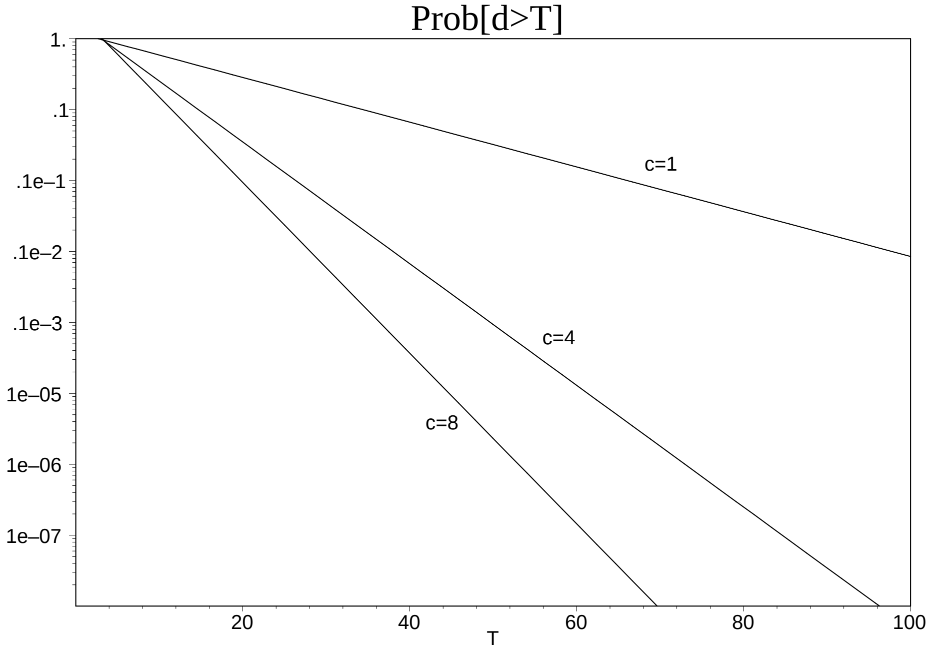

Figure 3: Tail distribution of customer delay, Prob$[d > T]$, versus T, for $\rho = 0.8$ and $s = 5$.

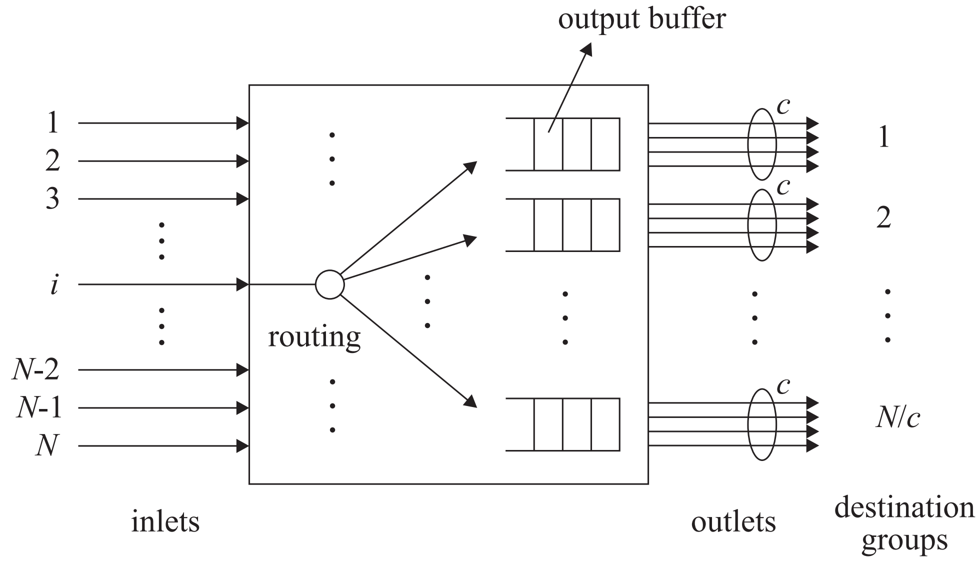

Figure 4: $N \times N(c)$ switch with output buffers.

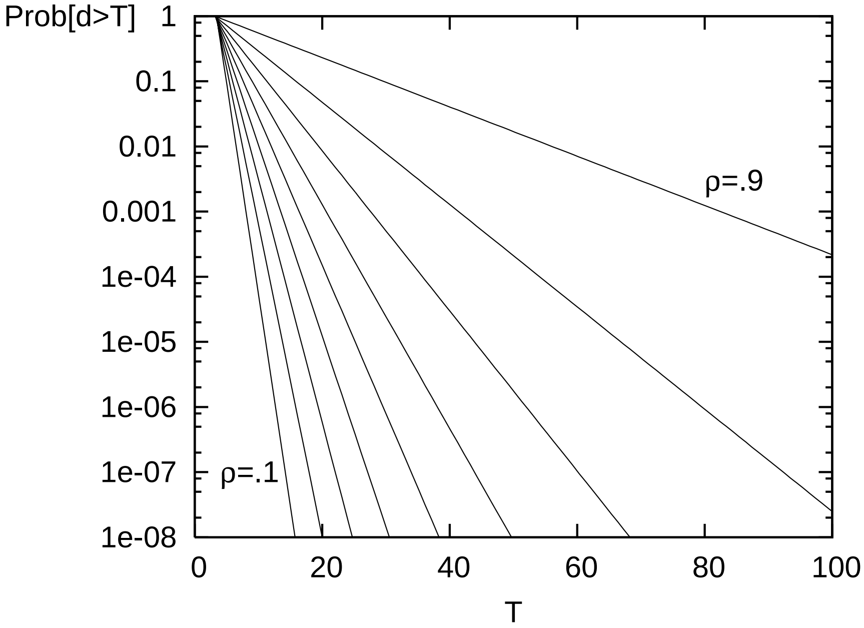

Figure 5: Tail distribution of packet delay, Prob$[d > T]$, versus T, for $N = 8$, $c = 2$, $s = 5$ and $\rho = 0.1, 0.2, \ldots, 0.9$.

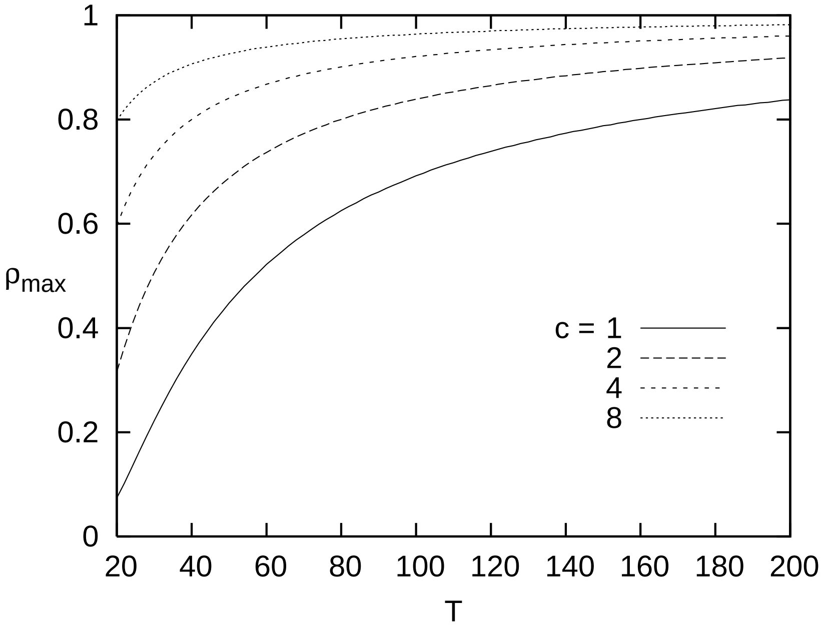

Figure 6: Maximum load ρ such that Prob$[d > T] \leq 10^{-6}$ versus T, for $N = 8$ and $s = 5$.