Discrete-Time Multiserver Queues with Geometric Service Times

Peixia Gao[*]**, Sabine Wittevrongel, Herwig Bruneel**

SMACS[†] **Research Group,**

Department of Telecommunications and Information Processing,

Ghent University, Sint-Pietersnieuwstraat 41, B-9000 Gent, Belgium.

Scope and purpose

Discrete-time queueing models are widely used to study the behavior of various types of telecommunication and computer systems. Most of the existing studies of discrete-time multiserver queueing models assume that the service times of the customers are constant. Recently, however, there has been an increased interest in discrete-time models with non-deterministic service times, due to the more and more complicated and irregular service mechanisms in nowadays telecommunication networks. In view of this, this paper focuses on a discrete-time queueing system with multiple servers and geometrically distributed service times. We show that the queueing model can be analyzed by means of a generating-functions approach. The results of the analysis are important to evaluate performance measures such as packet loss and delay in communication and computer networks.

[*] Corresponding author. Tel.: +32-9-264-89-01; fax: +32-9-264-42-95.

E-mail address: pg@telin.rug.ac.be (P. Gao)

[†] *SMACS: Stochastic Modeling and Analysis of Communication Systems*

Abstract

In this paper, a discrete-time multiserver queueing system with infinite buffer size and general independent arrivals is considered. The service times of a packet served by one of the servers are assumed to be independent and identically distributed according to a geometric distribution. Each packet gets service from only one server. In the paper, the behavior of the queueing system is studied analytically by means of a generating-functions approach. This results in closed-form expressions for the mean values, the variances and the tail distributions of the system contents and the packet delay. Some numerical examples are given to illustrate the analysis.

Keywords: Discrete-time queueing, multiserver system, geometric service times, system contents, delay

1 Introduction

Discrete-time queueing theory has received considerable attention during the last decades because of its direct applicability in the study of many computer and communication systems in which time is divided into fixed-length time intervals ("slots"). The reader is referred to the books [1]-[6] and the references therein for an extensive treatment of a wide variety of discrete-time queueing models. There are a number of analysis techniques for discrete-time models, ranging from computer simulation to the numerical solution of the associated set of linear balance equations and various types of analytical or semi-analytical methods. In case of computer simulation, a program is built to simulate the behavior of the system in software; a typical simulation technique is discrete event simulation ([7]). With respect to the mathematical solution techniques, a distinction can be made between algorithms that allow to efficiently calculate the performance measures numerically (e.g. matrix-analytic methods ([8], [9])) and analytical methods that lead to (exact or approximate) closed-form expressions (e.g. probability generating functions based techniques ([4], [10]) or methods based on the maximum entropy principle ([11])).

In this paper we focus our attention on the specific class of discrete-time queues with multiple servers (or output channels). To the best of our knowledge, in most of the existing literature on discrete-time multiserver queueing models, the service (or transmission) times of customers (or packets) are assumed to be constant, equal to one slot (see [12]-[15]) or multiple slots ([16]). These assumptions are appropriate when modeling for instance buffers in ATM-based integrated-services digital networks, in view of the fixed packet (cell) length used in these networks ([17]). Only a few authors have considered more general discrete-time multiserver models with non-deterministic (variable) service times, although there are both fundamentally theoretic as well as more practical motives to do this, given the more and more complicated and irregular service mechanisms in the nowadays internet. We mention [18], where the service times are assumed to be geometrically distributed; somewhat related models with randomly interrupted servers are studied in [19] and [20]. Geometric service times are also considered in [21] and [22], while [23]-[25] deal with general service times, but only for the case of one single server.

In line with the above, we are concerned in this paper with generalizing the service-time distribution. More specifically, as a first step, we consider a discrete-time queueing system with geometric service times and a general uncorrelated arrival process, and present an analytical technique based on generating functions for the performance analysis of the system. The model under study has also been treated in [18]. However, as opposed to [18], we obtain explicit expressions for the probability generating functions (pgf's) of not only the system contents, but also the total delay experienced by customers. Furthermore, these pgf's are used to derive various additional performance measures. Specifically, expressions are derived for the mean values, the variances and the tail probabilities of system contents and delay.

The outline of the paper is as follows. In section 2, the queueing model under study is described. In section 3, we focus on the system contents and derive expressions for the pgf, the mean value, the variance and the tail probabilities of the system contents. The packet delay is analyzed in section 4, and again the pgf, mean value, variance and tail probabilities are obtained. In section 5, some special cases are considered in order to check the obtained results. In section 6, some numerical examples are given and discussed. Finally, the paper is concluded in section 7.

2 Queueing Model

In this paper, we consider a discrete-time queueing system with multiple output channels and an infinite storage capacity for packets. The number of output channels (servers) will be denoted by c ($c > 0$). Packets arrive at the input of the system in a stochastic manner, wait in a buffer for some time and are then transmitted via one of the c output channels, after which they finally leave the system. The time axis is assumed to be divided into fixed-length time intervals, referred to as slots and chronologically labeled. New packets enter the buffer according to a general uncorrelated arrival process, i.e., the numbers of packets arriving in the buffer during the consecutive slots are modeled as independent and identically distributed (i.i.d.) random variables, with common pgf $A(z)$.

The service process is modeled as follows. The service (or transmission) of a packet can start or end at slot boundaries only. Hence, the service time of a packet always consists of an integer number of full slots. This also implies that the service of a packet which arrives in an empty system cannot start before the end of the packet's arrival slot. One packet can only get service from one of the servers. The service times of the packets are assumed to be i.i.d. random variables and geometrically distributed with parameter $1-\mu$ ($0<\mu\leq 1$), i.e., with common probability mass function (pmf)

$$\begin{aligned} g(n) &= \text{Prob[service of a packet takes } n \text{ slots]} \\ &= \mu(1-\mu)^{n-1}, \quad n\geq 1, \end{aligned} \tag{1}$$

and corresponding pgf

$$G(z)=\sum_{n=1}^{\infty} g(n)z^n = \frac{\mu\ z}{1-(1-\mu)\ z}. \tag{2}$$

Moreover, the service and arrival processes are assumed to be mutually independent. Packets are served according to a first-come-first-served (FCFS) queueing discipline.

It is well known (see e.g. [2]) that a queueing system can reach a steady state if and only if the average amount of “work” (expressed in slots of service time) entering the system per slot is strictly less than what the system can handle per slot. In view of the above model description, this equilibrium condition can be expressed as

$$A^{'}(1) < c\mu, \tag{3}$$

where we have used primes to indicate derivatives. In the analysis that follows, we assume that the condition (3) is satisfied.

3 System Contents

This section deals with the analysis of the system contents. First, we calculate the pgf of the system contents and next, we use this pgf to derive the mean, the variance and the tail probabilities of the system contents.

3.1 The pgf of the system contents

Let us denote by v_k the system contents (i.e., the total number of packets in the queueing system, including the ones being transmitted, if any) at the beginning of slot k, and by a_k the number of packet arrivals during slot k. Then, in view of the assumptions of the model in section 2, we have

$$v_{k+1} = v_k - t_k + a_k, \tag{4}$$

where

$$t_k = \sum_{j=1}^{(v_k,c)^-} t_{k,j}, \tag{5}$$

with $(.,.)^- \overset{\triangle}{=} min(.,.)$. The random variable t_k denotes the total number of departures at the end of slot k, i.e., the total number of packets whose service is completed at the end of slot k.

Note that $t_{k,j}$ corresponds to the number of departures at the end of slot k via the j-th output channel when there is a packet receiving service from this output channel. In view of the geometric distribution of the packet service times, $t_{k,j}$ is a Bernoulli distributed random variable with parameter μ, i.e.,

$$t_{k,j} = \begin{cases} 1 & \text{with probability } \mu, \\ 0 & \text{with probability } 1-\mu. \end{cases}$$

We define the conditional pgf $T_i(z)$ as

$$T_i(z) \overset{\triangle}{=} E[z^{t_k} | (v_k, c)^- = i], \; i = 0, \; 1, \; \dots , \; c,$$

where $E[\cdot]$ denotes the expected value of the expression between square brackets. Then, from (5) we have

$$T_i(z) = (1 - \mu + \mu z)^i. \tag{6}$$

Now let $V_k(z)$ denote the pgf of v_k. Using standard z-transform techniques, we can translate the system equation (4) into the z-domain, as follows:

$$\begin{aligned} V_{k+1}(z) =&A(z)\ E\left[z^{v_k - t_k}\right] \\ =&A(z)\left\{V_k(z)\ T_c(\frac{1}{z}) + \sum_{i=0}^{c-1} \text{Prob}[v_k = i] z^i\ \left[T_i(\frac{1}{z}) - T_c(\frac{1}{z})\right]\right\}. \end{aligned} \tag{7}$$

When the steady state is reached, the pgf's $V_k(z)$ and $V_{k+1}(z)$ converge to their equilibrium version $V(z)$, which is the pgf of the system contents v at the beginning of an arbitrary slot in the steady state. Taking limits for $k \to \infty$ in (7) and solving the resulting equation for $V(z)$, we then obtain the following expression:

$$\begin{aligned} V(z) =& \frac{A(z)\ \sum_{i=0}^{c-1}\ [\ T_i(\frac{1}{z}) - T_c(\frac{1}{z})\ ] v(i) z^i}{1 - T_c(\frac{1}{z}) A(z)} \\ =& \frac{A(z) z^c H(z)^c\ \sum_{i=0}^{c-1} [\ H(z)^{i-c} - z^{i-c}\ ]\ v(i)}{z^c - H(z)^c A(z)}, \end{aligned} \tag{8}$$

where $H(z) = \mu + (1-\mu) z$, and the constants $v(i)$ are probabilities, defined as $v(i) \triangleq \text{Prob}[v = i],\ i = 0, 1, ..., c-1$.

In order to determine $V(z)$ completely, we need to find the c unknown constants $v(i)$, $i = 0, 1, ..., c-1$. These can be obtained by invoking the analyticity of the pgf $V(z)$ inside the unit disk $(z : |z| < 1)$ of the complex z-plane and the normalization condition $V(1) = 1$. Specifically, by means of Rouché's theorem ([26]), it can be shown that the denominator of the right-hand side of (8) has exactly $c-1$ roots inside the unit disk. We denote these roots by $z_j,\ j = 1, 2, ..., c-1$. Since $V(z)$ is analytic for $|z| < 1$, the numerator of the right-hand side of (8) must also be zero at these points. Hence, we have

$$\sum_{i=0}^{c-1} [\ H(z_j)^{i-c} - z_j^{i-c}\ ]\ v(i) = 0, \quad j = 1, 2, ..., c-1. \tag{9}$$

From the normalization condition $V(1) = 1$ and equation (8), we find that

$$\mu \sum_{i=0}^{c-1} (c - i)\ v(i) = c\mu - A^{'}(1). \tag{10}$$

Note that whenever the condition (3) for the existence of a steady state is fulfilled, the right-hand side of (10) is strictly positive. From equations (9) and (10), the constants $v(i)$

$(i = 0, 1, ..., c-1)$ can be calculated. Once $V(z)$ is determined, some important performance measures of the queueing system can be obtained, for instance the mean value, the variance and the tail distribution of the system contents.

3.2 Mean value and variance of the system contents

The mean system contents can be obtained by taking the first-order derivative of equation (8) with respect to z in $z = 1$. By applying the rule of de l'Hospital twice to (8), we get

$$\begin{aligned} E[v] &= V^{'}(1) \\ &= \frac{\mu(2-\mu)}{2\ [\ c\mu - A^{'}(1)\ ]} \sum_{i=0}^{c-1} (c^2 - i^2)\ v(i) \\ &\quad + \frac{A^* + 2cA^{'}(1) - (2-\mu)\ [\ c^2\mu - A^{'}(1)\ ]}{2\ [\ c\mu - A^{'}(1)\ ]}, \end{aligned} \tag{11}$$

where $A^* = A^{''}(1) - 2A^{'}(1)^2$ and $A^{''}(1)$ is the second-order derivative of $A(z)$ with respect to z at $z = 1$. With equation (11), the calculation of the mean system contents is straightforward, once the unknown probabilities $v(i)$ $(i = 0, 1, ..., c-1)$ have been obtained. Higher-order moments of the system contents can be derived in a similar way. For instance, taking the second-order derivative of equation (8) with respect to z at $z = 1$ and using (10) and (11), we find

$$\begin{aligned} V^{''}(1) = &-\frac{1}{3} \frac{\mu\ \mu_{[3]}}{c\mu - A^{'}(1)} \sum_{i=0}^{c-1} (c^3 - i^3)\ v(i) \\ &- \frac{(\mu_{[3]} + c)\ [\ 2A^{'}(1)V^{'}(1) + A^*\ ] - c(c+1)\ \mu_{[2]}\ [\ \mu V^{'}(1) - A^{'}(1)\ ]}{(2-\mu)[\ c\mu - A^{'}(1)\ ]} \\ &+ \frac{\mu_{[3]}\ [\ c^3\mu - A^{'}(1)\ ] - 3\ A^*\ [\ V^{'}(1) - 2A^{'}(1)\ ] + 6A^{'}(1)^3 - A^{'''}(1)}{3\ [\ c\mu - A^{'}(1)\ ]}, \end{aligned} \tag{12}$$

where $\mu_{[\zeta]} = \mu^2 - \zeta\mu + \zeta$. Expression (12) can be used to calculate the variance of the system contents through the relation

$$Var[v] = V^{''}(1) + V^{'}(1) - V^{'}(1)^2. \tag{13}$$

3.3 Tail probabilities of the system contents

Another important performance characteristic for a queueing system is the tail distribution of the system contents. This can be used to estimate the packet loss probability that would be observed in case of a buffer with finite storage capacity, namely by approximating the packet loss probability for a system of size N by the probability that the system contents in the corresponding infinite system exceeds N (see e.g. [27]). We will use here the technique presented in [15] and [28] to derive the tail distribution of the system contents. Specifically, we have that for sufficiently large values of N, the tail distribution of the system contents can be approximated as

$$\text{Prob}[v > N] \approx -C_v \, \frac{z_v^{-N-1}}{z_v - 1}. \tag{14}$$

In the above expression, z_v is the real positive pole of $V(z)$ with the smallest modulus outside the unit disk, i.e., the dominant pole of $V(z)$, and C_v is the residue of $V(z)$ at $z = z_v$. From equation (8), it follows that z_v is a real positive zero of the denominator of $V(z)$. The residue C_v can be calculated from (8) as

$$C_v = -\frac{z_v H(z_v) A(z_v)^2}{Y} \sum_{i=0}^{c-1} \left[H(z_v)^i - \frac{z_v^i}{A(z_v)} \right] v(i), \tag{15}$$

where $Y = z_v H(z_v)\ A^{'}(z_v) - c\mu\ A(z_v)$.

4 Packet Delay

In this section, we derive the pgf of the delay experienced by packets in the queueing system. The delay of a packet is defined as the total number of slots between the end of the slot during which the packet arrived in the system and the end of the slot where the packet's transmission finishes and the packet leaves the system. From the pgf of the delay, we will be able to calculate various performance measures, such as the mean value, the variance and the tail distribution of the packet delay.

4.1 The pgf of the delay

Let us consider an arbitrary packet, denoted by P (called the tagged packet). Upon arrival, it will find a number of other packets in the queueing system. Just after the end of the arrival slot of P, all packets that arrived in the previous slots, but have not been taken into service yet, and all packets that arrived in the same slot as the tagged packet, but before this packet, are waiting in the queue in front of the tagged packet (due to the FCFS queueing discipline). Whenever there are output channels available for transmission at the beginning of a slot, the packet at the head of the queue is selected for service, until eventually the tagged packet itself is served and in the end leaves the system. We now assume that the packet P arrives in the system during slot k. Let us denote by τ_k its arrival instant, by v_k the number of packets present in the system at the beginning of slot k, and by q_k the number of packets staying in the system at τ_k, except the tagged packet itself and the ones that leave the system at the end of slot k, if such packets exist. Then q_k can be expressed as follows:

$$q_k = v_k - t_k + f_k, \tag{16}$$

where f_k denotes the number of packets arriving during slot k before P. It can be shown (see [2]) that the pgf $F(z)$ of f_k is given by

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

Owing to the independent nature of the arrival process, we have (i) that v_k has the same probability distribution as v, which denotes the system contents at the beginning of a random slot in the steady state, and (ii) that v_k and f_k are statistically independent of each other. From equations (16) and (4), we can derive the pgf $Q(z)$ of q_k when the steady state is reached as

$$Q(z) = \frac{F(z)V(z)}{A(z)}. \tag{18}$$

We now first consider the waiting time experienced by the tagged packet P. The waiting time of a packet is defined as the number of slots between the end of the packet's arrival slot and the beginning of the slot when the packet's transmission starts. We define w_k as the waiting time that P experiences if it arrives during slot k. In order to derive the pgf of w_k, let us briefly discuss a close relationship between the waiting time w_k and the random variable q_k (i.e., the number of packets present in the system right after the arrival slot of the tagged packet P, that have been or will be selected for service before P).

Let us observe the slots following the k-th slot. During the $(k+1)$-th slot, there will be $\widetilde{t}_{k+1}$ $(= \sum_{j=1}^{c} t_{k+1,j})$ packets leaving the system at the end of slot $k+1$, if $q_k \geq c$ (i.e., just after the end of that slot or at the beginning of slot $k+2$, there will be $q_k - \widetilde{t}_{k+1}$ packets in the system with priority over the tagged packet P to be taken into service), or the tagged packet will get into service, if $q_k < c$ (i.e., the tagged packet stops waiting). If the tagged packet didn't get into service, then a similar remark can be made for the $(k+2)$-th slot: either there will be $q_k - \widetilde{t}_{k+1} - \widetilde{t}_{k+2}$ packets with service priority over the tagged packet at the end of slot $k+2$, if $q_k - \widetilde{t}_{k+1} \geq c$, or the packet P will get into service, if $q_k - \widetilde{t}_{k+1} < c$. We observe that the tagged packet will still be waiting for service during the $(k+i+1)$-th slot only if $q_k - \widetilde{t}_{k+1} - \widetilde{t}_{k+2} - ... - \widetilde{t}_{k+i} \geq c$. This leads to the following relationship between w_k and q_k:

$$w_k > i \;\Leftrightarrow\; q_k - \widetilde{t}_{k+1} - \widetilde{t}_{k+2} - ... - \widetilde{t}_{k+i} \geq c.$$

If we introduce the following random variables :

$$s_{k,0} \triangleq c;\;\; s_{k,i} \triangleq \widetilde{t}_{k+1} + \widetilde{t}_{k+2} + ... + \widetilde{t}_{k+i} + c,\;\; i \geq 1,$$

then the above equation for the waiting time can be rewritten as

$$w_k > i \;\Leftrightarrow\; q_k \geq s_{k,i}. \tag{19}$$

Note that since the variables $\widetilde{t}_k$ are i.i.d. random variables, the pgf of $s_{k,i}$ can be expressed as

$$S_{k,i}(z) \triangleq E[z^{s_{k,i}}] = z^c T_c(z)^i. \tag{20}$$

We now begin with the derivation of the pgf $W_k(z)$ of the waiting time w_k. From equation (19) and the easily proven identity

$$\frac{W_k(z)-1}{z-1} = \sum_{i=0}^{\infty} z^i \mathrm{Prob}[w_k > i],$$

we can write

$$\begin{aligned}\frac{W_k(z)-1}{z-1} &= \sum_{i=0}^{\infty} z^i \mathrm{Prob}[q_k \geq s_{k,i}] \\ &= \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} z^i \mathrm{Prob}[q_k = j]\ \mathrm{Prob}[q_k \geq s_{k,i} | q_k = j].\end{aligned}$$

Since q_k and the $s_{k,i}$'s are statistically independent, this can be rewritten as

$$\frac{W_k(z)-1}{z-1} = \sum_{j=0}^{\infty} \mathrm{Prob}[q_k = j]\ \sum_{n=c}^{j}\sum_{i=0}^{\infty} \mathrm{Prob}[s_{k,i} = n] z^i. \tag{21}$$

Next, by using (20) and the probability generating property of pgf's, i.e.,

$$\mathrm{Prob}[s_{k,i} = n] = \frac{1}{n!}\frac{d^n}{dx^n} S_{k,i}(x)\Big|_{x=0} = \frac{1}{n!}\ \frac{d^n}{dx^n}\ \left[\ x^c T_c(x)^i\ \right]\Big|_{x=0},$$

and working out the sum over i, we find

$$\begin{aligned}\frac{W_k(z)-1}{z-1} &= \sum_{j=0}^{\infty} \mathrm{Prob}[q_k = j]\ \sum_{n=c}^{j}\sum_{i=0}^{\infty} \frac{1}{n!}\frac{\partial^n}{\partial x^n}\left(x^c\left[\ T_c(x) z\right]^i\right)\Big|_{x=0} \\ &= \sum_{j=0}^{\infty} \mathrm{Prob}[q_k = j]\ \sum_{n=c}^{j} \frac{1}{n!}\ \frac{\partial^n}{\partial x^n}\left(\frac{x^c}{1 - zT_c(x)}\right)\Big|_{x=0}.\end{aligned} \tag{22}$$

Note that working out the sum over i requires that $|zT_c(x)| < 1$ in the neighborhood of $x = 0$. This condition will always be fulfilled for $|z| \leq 1$, since $|T_c(x)| < 1$ for $|x| < 1$. In order to derive the partial derivatives in (22), we replace the argument by its partial fraction expansion. Considering z to be a constant and x the variable of interest, we find

$$\frac{x^c}{1-zT_c(x)} = -\frac{1}{\mu^c z} + \sum_{p=0}^{c-1} \frac{-x_p^c}{zT_c^{'}(x_p)\ (x-x_p)}, \tag{23}$$

where the x_p's are the solutions of the equation

$$1 - zT_c(x) = 0. \tag{24}$$

Clearly, the x_p's are functions of z. However, to ease the notation, we write x_p instead of $x_p(z)$. Since $T_c(x)$ is a polynomial of degree c, equation (24) has c solutions, which we assume to be distinct. Notice that (24) has a solution at $x = 1$, when $z = 1$. We denote this solution as x_0. Since $T_c^{'}(1) > 0$, this solution is unique. Substituting (23) in equation (22), we find after some manipulations that

$$\frac{W_k(z)-1}{z-1} = \sum_{p=0}^{c-1} \frac{-x_p^c}{zT_c^{'}(x_p)\ (1-x_p)} \sum_{j=c}^{\infty} \text{Prob}[q_k = j](x_p^{-c} - x_p^{-1-j}).$$

When the steady state is reached, $W_k(z)$ becomes independent of k and converges to a limit function $W(z)$. By using the definition of the pgf $Q(z)$, we then get

$$\begin{aligned} \frac{W(z)-1}{z-1} =& \sum_{p=0}^{c-1} \frac{-x_p^c}{zT_c^{'}(x_p)\ (1-x_p)} \left[x_p^{-c} - \frac{1}{x_p}\ Q(\frac{1}{x_p}) \right] \\ &+ \sum_{j=0}^{c-1} \text{Prob}[q=j] \sum_{p=0}^{c-1} \frac{1 - x_p^{c-j-1}}{zT_c^{'}(x_p)\ (1-x_p)}. \end{aligned}$$

Using the partial fraction expansions

$$\begin{aligned} \frac{1}{1-zT_c(x)} &= \sum_{p=0}^{c-1} \frac{-1}{zT_c^{'}(x_p)\ (x-x_p)}; \\ \frac{1-x^{c-j-1}}{1-zT_c(x)} &= \sum_{p=0}^{c-1} \frac{-(1-x_p^{c-j-1})}{zT_c^{'}(x_p)\ (x-x_p)} \end{aligned}$$

at $x = 1$, we finally find the following expression for the pgf of the packet waiting time:

$$W(z) = (z-1)\sum_{p=0}^{c-1} \frac{x_p^{c-1}}{zT_c^{'}(x_p)\ (1-x_p)} Q(\frac{1}{x_p}). \tag{25}$$

Next, we derive the expression for the pgf of the packet delay. Let the random variable d indicate the delay experienced by the tagged packet P (when the system has reached equilibrium), and let $D(z)$ denote the corresponding pgf. Note that the packet delay is equal to the sum of the packet waiting time and the packet transmission time. From (2) and (25), we have

$$\begin{aligned} D(z) &= W(z)\cdot G(z) \\ &= \frac{\mu(z-1)}{1-(1-\mu)z} \ \sum_{p=0}^{c-1} \frac{x_p^{c-1}}{T_c^{'}(x_p)(1-x_p)} Q(\frac{1}{x_p}). \end{aligned} \tag{26}$$

We now proceed with calculating some performance measures related to the delay.

4.2 Mean value and variance of the delay

In a similar way as for the mean system contents, the mean delay experienced by packets can be found by calculating the first derivative of $D(z)$ at $z=1$. From equation (26), we find

$$E[d] = D^{'}(1) = W^{'}(1) + G^{'}(1). \tag{27}$$

Clearly, the derivation of $W^{'}(1)$ is a major contribution to get the mean delay. In order to calculate $W^{'}(1)$, we take the first-order derivative of $W(z)$ in (25) with respect to z at $z=1$, using the rule of de l'Hospital twice because for the term where $p=0$, $x_0=1$ if $z=1$. With the following derivative, found by differentiating (24) with respect to z:

$$\frac{dx_p(z)}{dz} = -\frac{1}{z^2\ T_c^{'}(x_p(z))},$$

we find

$$W^{'}(1) = \frac{1}{c\mu}\left[\frac{(c-1)(\mu-2)}{2} + Q^{'}(1)\right] + \sum_{p=1}^{c-1} \frac{x_p^{c-1}}{T_c^{'}(x_p)\ (1-x_p)} Q(\frac{1}{x_p})\ \Big|_{z=1}, \tag{28}$$

where $Q^{'}(1)$ is the first-order derivative of $Q(z)$ evaluated at $z = 1$. Combining (8), (17) and (18), the second part of equation (28) can be expressed as:

$$\sum_{p=1}^{c-1} \frac{x_p^{c-1}}{T_c^{'}(x_p)\ (1-x_p)} Q(\frac{1}{x_p}) \Big|_{z=1} = \frac{1}{A^{'}(1)} \sum_{i=0}^{c-1} v(i) \ \sum_{p=1}^{c-1} \frac{x_p^{c-i}\,[\ 1-(1-\mu+\mu x_p)^i\ ]}{T_c^{'}(x_p)\ (1-x_p)^2}\Big|_{z=1}. \tag{29}$$

Using partial fraction expansion again, we have:

$$\frac{x^{c-i}\ [\ 1-(1-\mu+\mu x)^i\ ]}{1-T_c(x)} = \mu^{i-c} + \sum_{p=0}^{c-1} \frac{-x_p^{c-i}\,[\ 1-(1-\mu+\mu x_p)^i\ ]}{T_c^{'}(x_p)\ (x-x_p)}\Big|_{z=1}. \tag{30}$$

Taking the derivative on both sides of (30) with respect to x, moving the term with $p = 0$ to the left-hand side and then taking the limit for $x \rightarrow 1$, we find (by using the rule of de l'Hospital several times) that

$$\sum_{p=1}^{c-1} \frac{x_p^{c-i}\,[\ 1-(1-\mu+\mu x_p)^i\ ]}{T_c^{'}(x_p)\ (1-x_p)^2}\Big|_{z=1} = \frac{2-\mu}{2c}\ i(c-i). \tag{31}$$

Combining (31), (29), (28), (11) and (10), we finally have

$$W^{'}(1) = \frac{E[v]}{A^{'}(1)} - \frac{1}{\mu}.$$

It is clear from the above equation and (27) that the mean packet delay and the mean system contents are related through Little's theorem ([26]), which is a check for the correctness of our analysis.

Calculation of the second-order derivative of $D(z)$ at $z = 1$ yields the variance of the delay through the relation

$$Var[d] = D^{''}(1) + D^{'}(1) - D^{'}(1)^2, \tag{32}$$

which is also called the delay jitter. From (26), we have

$$D^{''}(1) = W^{''}(1) + 2\ W^{'}(1)\ G^{'}(1) + G^{''}(1). \tag{33}$$

Taking the second-order derivative of (25) and applying the rule of de l'Hospital to the term for $p = 0$, we have

$$\begin{aligned} W^{''}(1) = {} & \frac{2}{A^{'}(1)} \sum_{p=1}^{c-1} -T_1(x_p(1)) x_p(1)^{c-3} \left\{ \frac{T_1(x_p(1))\ S_1}{c^3 \mu^3 X_{p_1}^2} + \left[1 - \frac{1}{\mu} + \frac{x_p(1)}{A(\frac{1}{x_p(1)}) - 1} \right. \right. \\ & \left. \left. +(1-c) x_p(1)^2 + \frac{T_1(x_p(1))}{c\mu X_{p_1}} + \frac{x_p(1)^3}{X_{p_1} T_1(x_p\ (1))} \right] \ \frac{T_1(x_p(1))\ S_2}{c^2 \mu^2 X_{p_1}^2} \right\} \\ & + \frac{1}{c^2 \mu^2} \left\{ V^{''}(1) + \frac{2\xi A^{'}(1) + A^*}{A^{'}(1)} V^{'}(1) + \frac{\xi - A^{'}(1)}{A^{'}(1)} A^* + \frac{A^{'''}(1)}{3A^{'}(1)} \right. \\ & \left. -A^{''}(1) - (c-1) \Big[\ \frac{(c+1)\mu^2}{6} + \xi - \mu \ \Big] \right\}, \end{aligned} \tag{34}$$

where

$$\begin{aligned} S_1 &= \sum_{i=0}^{c-1} \left\{ [(c+i)(\mu - 1) - c\mu x_p(1)] \ \left[\frac{T_1(x_p(1))}{x_p(1)} \right]^i \right. \\ & \qquad \left. + [(c+i)(\mu-1) - i\mu\ x_p(1)] \ \left[\frac{1}{x_p(1)} \right]^i \right\} v(i), \\ S_2 &= \sum_{i=0}^{c-1} \left\{ \left[\frac{T_1(x_p(1))}{x_p(1)} \right]^i - \left[\frac{1}{x_p(1)} \right]^i \right\} v(i), \\ X_{p_1} &= x_p(1) - 1, \\ \xi &= 2 - \mu - c. \end{aligned}$$

From the analysis above, an explicit expression for the packet delay jitter or the variance of the packet delay can be obtained.

4.3 Tail probabilities of the delay

The tail distribution of the packet delay, for a sufficiently large value of T, can be expressed as

$$\text{Prob}[d > T] \approx -C_d \ \frac{z_d^{-T-1}}{z_d - 1}, \tag{35}$$

when $D(z)$ has a singular dominant pole z_d, and where C_d is the residue of $D(z)$ at $z = z_d$, or

$$\text{Prob}[d > T] \approx - \left[C_{1d} - \left(2 + T + \frac{1}{z_d - 1} \right) \frac{C_{2d}}{z_d} \right] \frac{z_d^{-T-1}}{z_d - 1}, \tag{36}$$

when $D(z)$ has a dominant pole z_d with multiplicity 2, and where C_{1d} and C_{2d} are determined by

$$C_{1d} = \lim_{z \to z_d} \frac{d}{dz} \left[(z - z_d)^2 D(z) \right] , \; C_{2d} = \lim_{z \to z_d} (z - z_d)^2 D(z). \tag{37}$$

In order to find the dominant pole z_d, note that from (2), (25) and (26), it follows that poles of $D(z)$ are to be found among the following z's:

- z so as to make $1 - (1 - \mu)z = 0$. This requires $z = 1/(1 - \mu)$.
- z so as to make $1/x_p(z)$ (for some p) a pole of $Q(z)$. This requires $z = 1/T_c(\frac{1}{z_*})$, where z_* is a pole of $Q(z)$.
- z so as to make $x_p(z) - 1 = 0$ (for some p). This case, however, does not yield a pole since it requires $z = 1$ (see (24)).

As in [20], it can be shown that if z_* is a pole of $Q(z)$ with $z_* \neq z_v$, then $|1/T_c(\frac{1}{z_*})| > |1/T_c(\frac{1}{z_v})|$, and hence a pole of $D(z)$ given by $z = 1/T_c(\frac{1}{z_*})$ cannot be the dominant pole. Thus, in order to find z_d, we need to compare the modulus of $1/(1 - \mu)$ and $1/T_c(\frac{1}{z_v})$. As a result, we have to distinguish the following three cases:

Case 1. When $1/T_c(\frac{1}{z_v}) < 1/(1-\mu)$, there is a dominant pole z_d with multiplicity 1, given by

$$z_d = \frac{1}{T_c(1/z_v)} = A(z_v),$$

and the tail distribution of the delay is given by the geometric form (35), where

$$C_d = C_v \, \frac{G(A(z_v)) \; [A(z_v) - 1]^2}{A'(1) \; z_v^c \; (z_v - 1)^2}.$$

We notice that this is always valid for $c = 1$ because of the monotonically increasing character of $T_c(z)$ along the positive real axis, i.e. $1/T_1(\frac{1}{z_v}) < 1/T_1(0) = 1/(1 - \mu)$.

Case 2. When $1/T_c(\frac{1}{z_v}) > 1/(1 - \mu)$, the dominant pole z_d also has a multiplicity 1 and is equal to

$$z_d = \frac{1}{1-\mu}. \tag{38}$$

Hence, in this case, we also have a tail distribution of the form (35), where

$$C_d = -\frac{\mu}{(1-\mu)^2}\, W(\frac{1}{1-\mu}).$$

Case 3. When $1/T_c(\frac{1}{z_v}) = 1/(1-\mu)$, we get a dominant pole z_d of multiplicity 2, given by

$$z_d = \frac{1}{1-\mu} = A(z_v) = 1/T_c(\frac{1}{z_v}), \tag{39}$$

and the tail distribution of the delay has the form (36), where C_{1d} and C_{2d} are obtained from (26) and (37) as

$$\begin{aligned}
C_{1d} =& \frac{\mu\,[1-A(z_v)]}{1-\mu} \Bigg\{ \frac{C_v\,[A(z_v)-1]}{c^2\mu^2 A'(1) z_v^c\,(z_v-1)^2} \Bigg[-\frac{z_v H(z_v)}{c\mu C_v} \sum_{i=0}^{c-1} \Big[\,[\,c\mu - iH(z_v)\,]\,z_v^i \\
&+ (1-\mu) z_v A(z_v)\ i\ H(z_v)^i\,\Big] v(i) + \frac{z_v H(z_v)}{2Y}\,\Big[-[\,\mu + c\mu - 2H(z_v)\,]A'(z_v) \\
&+ z_v H(z_v) A''(z_v)\Big] - \frac{Y}{c\mu A(z_v)}\,\Big[\,\frac{(3c-1)z_v - (1+2c+cz_v)H(z_v)}{z_v - 1} \\
&+ \frac{c\mu + z_v H(z_v) A'(z_v)}{A(z_v)-1}\,\Big]\Bigg] + \sum_p \frac{x_p^{c-1}\,Q(1/x_p)}{T_c'(x_p)(1-x_p)} \Bigg\},
\end{aligned}$$

$$C_{2d} = -C_v \frac{\mu\,A(z_v)}{(1-\mu)A'(1)z_v^c}\,\frac{[A(z_v)-1]^2}{(z_v-1)^2}.$$

In the last term of the expression for C_{1d}, x_p is decided by $T_c(x_p) = 1/A(z_v)$, except the term for $1/x_p = z_v$.

5 Special Cases

In this section, we consider some special cases ($c = 1$ and $\mu = 1$) in order to check the results obtained above.

When $c = 1$, i.e., for a discrete-time queueing system with one output channel and geometric service times, the pgf of the system contents can be derived from (8) as

$$V(z) = \frac{[\ \mu - A^{'}(1)\ ]\ (z-1)A(z)}{z - H(z)A(z)}. \tag{40}$$

From (26) and considering the fact that x_0 is the root of the equation $1 - zT_1(x) = 0$, the pgf of the packet delay can be expressed as

$$\begin{aligned} D(z) =& G(z)\ Q\left(G(z)\right) \\ =& \frac{[\mu - A^{'}(1)]\ z\ [\ A(G(z)) - 1]}{A^{'}(1)\ [\ z - A(G(z))\ ]}. \end{aligned} \tag{41}$$

Equations (40) and (41) correspond exactly to the results found in [2].

When $\mu = 1$, i.e., for a queueing system with c output channels and deterministic service times of one slot, the pgf of the system contents can be derived from (8) as

$$V(z) = \frac{A(z)\ \sum_{i=0}^{c-1}\ (z^c - z^i)\ v(i)}{z^c - A(z)}. \tag{42}$$

Considering the fact that the x_p's are the roots of the equation $1 - zx^c = 0$, the pgf of the packet delay can be simplified from (25) and (26) as

$$D(z^c) = \frac{1}{c} \sum_{j=0}^{c-1} \frac{z^c - 1}{1 - (\alpha^j z)^{-1}} Q(\alpha^j z), \tag{43}$$

where $\alpha = \exp(2\Pi\iota/c)$ (ι is the imaginary unit). Equations (42) and (43) correspond exactly to the results found in [14].

6 Numerical Examples

In order to illustrate the results obtained above, let us consider a number of numerical examples. Throughout this section, we assume that the number of output channels c equals 1, 2, 4 or 16. The number of packets that arrive during a slot has a geometric distribution, i.e.,

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

In Figure 1, the mean system contents (divided by c) is plotted versus the average carried load ρ ($= \lambda/c\mu$) for the four different numbers of output channels. All curves have the same $\mu = 0.75$. Note that the scaling of the mean system contents by a factor c is done to allow a fair comparison between systems with dedicated or shared buffers. We observe that for a given ρ, when one shared buffer with multiple output channels is used to transport a given mean number of packets per slot, the equivalent mean system contents (per output channel) is less than in case several dedicated buffers with a single output channel are used. This also confirms the conclusion that the shared resource policy has an advantage over the non-shared one. For a given number of output channels, the mean system contents increases with increasing values of ρ. A similar figure can be plotted for the mean packet delay in terms of ρ, in view of Little's law.

In Figure 2, the variance of the packet delay is shown (for $\mu = 0.75$ and $c = 1,\ 2,\ 4,\ 16$) versus ρ. Clearly, for a given value of ρ, the delay jitter for the system with more output channels is less than the jitter for systems with less output channels.

In Figures 3 and 4, we have plotted the tail distributions of the system contents and the packet delay, for $\mu = 0.75$ and $\rho = 0.25$. To fairly compare systems with dedicated or shared buffers, the tail distribution of the system contents Prob$[v > N]$ is shown in Figure 3 as a function of N/c. This can be used to approximate the equivalent packet loss probability in a system with c output channels and a finite capacity N/c per output channel. Once again, we observe that for a system with more output channels, its equivalent packet loss probability is much less than for systems with less output channels. In Figure 4, the probability that the delay exceeds some given threshold T is plotted versus T for the four different numbers of output channels. Clearly, for a given T, the tail distribution of the delay decreases as the number of the output channels increases.

In Figure 5, we have plotted the quantiles of the system contents, i.e., the amount of buffer space required to yield a loss probability of 10^{-6} and 10^{-12}, versus the value of ρ, for $\mu = 0.75$ and $c = 4$. For high values of ρ (near the limiting value $\rho = 1.0$), a large storage capacity will be needed to keep the loss probability at an acceptable level.

7 Concluding Remarks

In this paper, we have developed an analytical technique for the analysis of a discrete-time multiserver queueing system with geometric service times and a general uncorrelated arrival process. Our approach is based on the use of probability generating functions and leads to closed-form expressions for the mean values, the variances and the tail distributions of the system contents and the packet delay. The obtained expressions are easy to evaluate numerically.

Several further generalizations of the model could be investigated. For instance, more general non-geometric service-time distributions could be allowed or some degree of correlation between the numbers of packet arrivals in different slots could be introduced in the model. A first result in this direction is reported in [29], where our analytical technique has been adapted to deal with the case of a correlated two-state Markovian arrival process.

Acknowledgements

The authors wish to thank the anonymous referees for their valuable comments and suggestions.

References

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

[2] Bruneel H, Kim BG. Discrete-time models for communication systems including ATM. Boston: Kluwer Academic Publishers, 1993.

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

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

[5] Robertazzi TG. Computer networks and systems: queueing theory and performance evaluation. New York: Springer-Verlag, 2000.

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

[7] Kurose JF, Mouftah HT. Computer-aided modeling, analysis, and design of communication networks. IEEE Journal on Selected Areas in Communications 1988;6(1):130-45.

[8] Neuts MF. Matrix-geometric solutions in stochastic models: an algorithmic approach. Baltimore: Johns Hopkins University Press, 1981.

[9] Blondia C, Casals O. Statistical multiplexing of VBR sources: a matrix analytic approach. Performance Evaluation 1992;16:5-20.

[10] Wittevrongel S, Bruneel H. Discrete-time ATM queues with independent and correlated arrival streams. In: Kouvatsos D (Ed.). Performance evaluation and applications of ATM networks. Boston: Kluwer Academic Publishers, 2000. p.387-412.

[11] Kouvatsos DD, Tabet-Aouel NM, Denazis SG. ME-based approximations for general discrete-time queueing models. Performance Evaluation 1994;21(1-2):81-109.

[12] Chu WW. Buffer behavior for Poisson arrivals and multiple synchronous constant outputs. IEEE Transactions on Computers 1970;C-19:530-4.

[13] Li S-Q. A general solution technique for discrete queueing analysis of multimedia traffic on ATM. IEEE Transactions on Communications 1991;39:1115-32.

[14] Bruneel H, Steyaert B, Desmet E, Petit G. 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 1992;5:193-201.

[15] Bruneel H, Steyaert B, Desmet E, Petit G. Analytic derivation of tail probabilities for queue lengths and waiting times in ATM multiserver queues. European Journal of Operational Research 1994;76:563-72.

[16] Bruneel H, Wuyts I. Analysis of discrete-time multiserver queueing models with constant service times. Operations Research Letters 1994;15:231-6.

[17] De Prycker M. Asynchronous transfer mode: solution for broadband ISDN. New York: Ellis Horwood, 1991.

[18] Rubin I, Zhang Z. Message delay and queue-size analysis for circuit-switched TDMA systems. IEEE Transactions on Communications 1991;39:905-14.

[19] Georganas ND. Buffer behavior with Poisson arrivals and bulk geometric service. IEEE Transactions on Communications 1976;24:938-40.

[20] Laevens K, Bruneel H. Delay analysis for discrete-time queueing systems with multiple randomly interrupted servers. European Journal of Operational Research 1995;85:161-77.

[21] Hsu J. Buffer behavior with Poisson arrival and geometric output processes. IEEE Transactions on Communications 1974;22:1940-1.

[22] Gao P, Wittevrongel S, Bruneel H, Zhang S. A discrete-time queueing system with correlated arrivals and geometric service times. Belgian Journal of Operations Research, Statistics and Computer Science (JORBEL) 2002;41(3).

[23] Briem U, Theimer TH, Kröner H. A general discrete-time queueing model: analysis and applications. Proceedings of the 13th International Teletraffic Congress, ITC 13, Copenhagen, 1991. p. 13-9.

[24] Bruneel H. Performance of discrete-time queueing systems. Computers & Operations Research 1993;20(3):303-20.

[25] Herrmann C. The complete analysis of the discrete time finite DBMAP/G/1/N queue. Performance Evaluation 2001;43(2-3):95-121.

[26] Kleinrock L. Queueing systems, Volume I: theory. New York: Wiley, 1975.

[27] Bisdikian C, Lew JS, Tantawi AN. On the tail approximation of the blocking probability of single server queues with finite buffer capacity. Proceedings of the Second International Conference on Queueing Networks with Finite Capacity, Research Triangle Park, 1993. p. 267-80.

[28] Woodside CM, Ho EDS. Engineering calculation of overflow probabilities in buffers with Markov-interrupted service. IEEE Transactions on Communications 1987;35:1272-7.

[29] Gao P, Wittevrongel S, Bruneel H. Discrete-time multiserver buffer systems with correlated arrivals and geometric service times. Submitted for publication.

Peixia Gao has been working as a Ph.D. student at the SMACS Research Group, Department TELIN (Telecommunication and Information Processing), Faculty of Applied Sciences, Ghent University, Belgium since 1999. Her main research interests include the stochastic modeling of communication networks and the analysis of discrete-time queueing models.

Sabine Wittevrongel was born in Gent, Belgium, in 1969. She received the M.S. degree in Electrical Engineering and the Ph.D. degree in Applied Sciences from Ghent University, Belgium, in 1992 and 1998, respectively. Since September 1992, she has been with the SMACS Research Group, Department TELIN, Faculty of Applied Sciences, Ghent University, first in the framework of various projects, from October 1994 to September 2001, as a researcher of the Fund for Scientific Research - Flanders (Belgium) (F.W.O.), and since October 2001 as a full time Professor. Her main research interests include discrete-time queueing theory, the performance evaluation of communication networks, the study of traffic control mechanisms, and the analysis of ARQ protocols.

Herwig Bruneel was born in Zottegem, Belgium, in 1954. He received the M.S. degree in Electrical Engineering, the degree of Licentiate in Computer Science, and the Ph.D. degree in Computer Science in 1978, 1979 and 1984 respectively, all from Ghent University, Belgium. He is full time Professor in the Faculty of Applied Sciences and head of the Department TELIN at the same university. He also leads the SMACS Research Group within this department. His main personal research interests include stochastic modeling and analysis of communication systems, discrete-time queueing theory, and the study of ARQ protocols. He has published more than 170 papers on these subjects and is coauthor of the book *H. Bruneel and B.G. Kim, "Discrete-Time Models for Communication Systems Including ATM"* (Kluwer Academic Publishers, Boston, 1993). Since October 2001, he serves as the Academic Director for Research Affairs at Ghent University.

Figure captions

Figure 1. Mean system contents versus average carried load ρ.

Figure 2. Variance of the packet delay versus average carried load ρ.

Figure 3. Tail distribution of the system contents, $\text{Prob}[v > N]$, versus N/c.

Figure 4. Tail distribution of the packet delay, $\text{Prob}[d > T]$, versus T.

Figure 5. Buffer space N required to yield an overflow probability of 10^{-6} and 10^{-12}, versus the average carried load ρ.

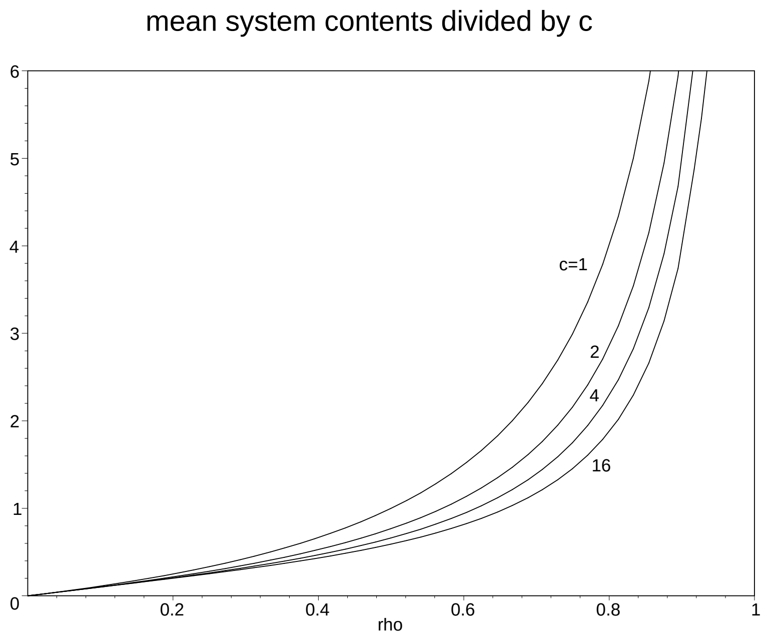
mean system contents divided by c
6
5
4
3
2
1
0
0.2
0.4
0.6
0.8
1
rho
c=1
2
4
16

Figure 1

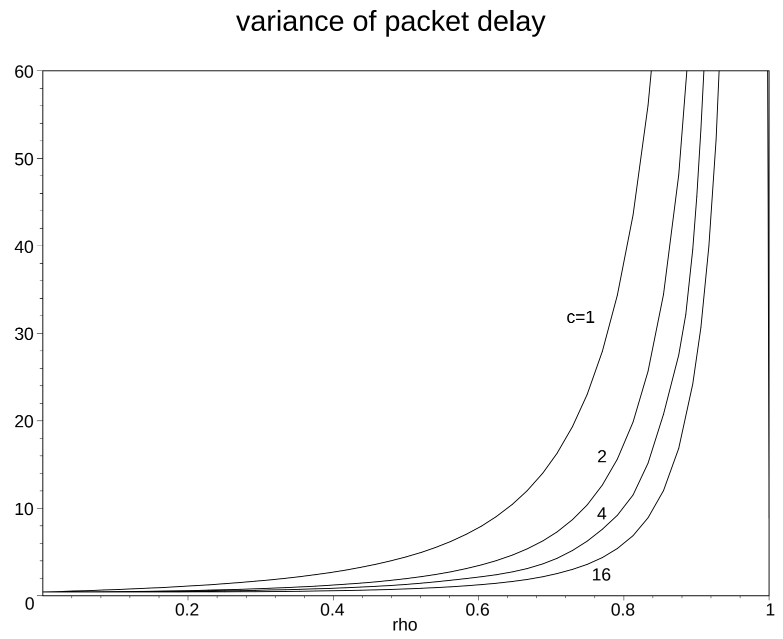
variance of packet delay
60
50
40
30
20
10
0
0.2
0.4
0.6
0.8
1
rho
c=1
2
4
16

Figure 2

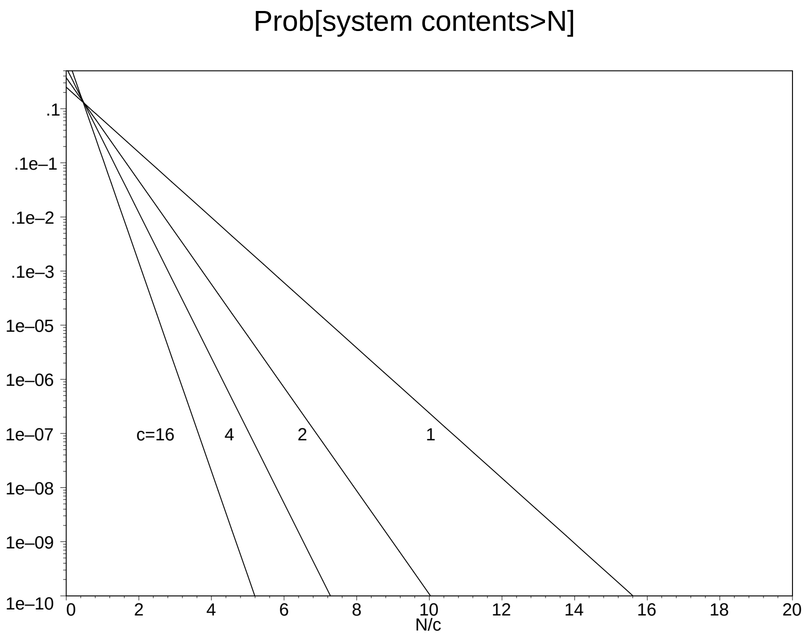
Prob[system contents>N]
.1
.1e–1
.1e–2
.1e–3
1e–05
1e–06
1e–07
1e–08
1e–09
1e–10
0
2
4
6
8
10
12
14
16
18
20
N/c
c=16
4
2
1

Figure 3

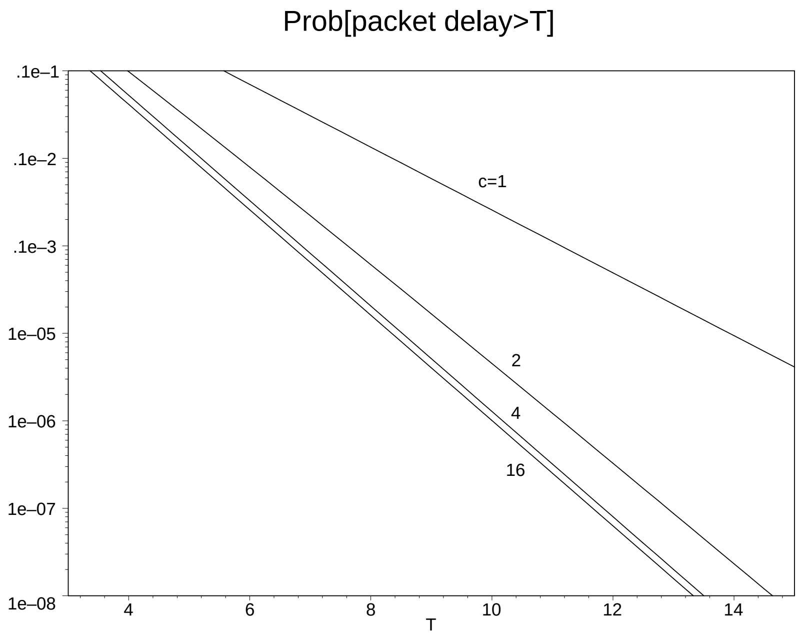
Prob[packet delay>T]
.1e–1
.1e–2
.1e–3
1e–05
1e–06
1e–07
1e–08
4
6
8
10
12
14
T
c=1
2
4
16

Figure 4

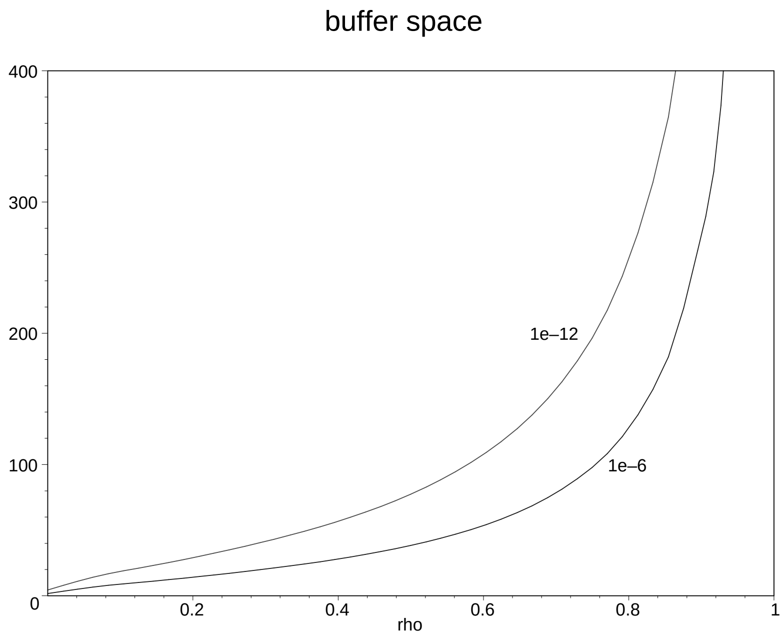

Figure 5