Math Meth Oper Res (2008) 67:269–284
DOI 10.1007/s00186-007-0163-z

ORIGINAL ARTICLE

Analytic study of multiserver buffers with two-state Markovian arrivals and constant service times of multiple slots

Peixia Gao · Sabine Wittevrongel · Joris Walraevens · Herwig Bruneel

Received: 11 April 2006 / Revised: 18 April 2007 / Published online: 22 June 2007

Abstract In this paper, we study the behavior of a discrete-time multiserver buffer system with infinite buffer size. Packets arrive at the system according to a two-state Markovian arrival process. The service times of the packets are assumed to be constant, equal to multiple slots. The behavior of the system is analyzed by means of an analytical technique based on probability generating functions (PGF's). Explicit expressions are obtained for the PGF's of the system contents and the packet delay. From these, the mean values, the variances and the tail distributions of the system contents and the packet delay are calculated. Numerical examples are given to show the influence of various model parameters on the system behavior.

Keywords Discrete-time queueing model · Correlated arrivals · Multiple servers · Performance analysis · Generating functions

1 Introduction

Discrete-time queueing models have been used for many years to analyze the behavior and performance of digital communication networks, where buffers are used for the temporary storage of information packets awaiting transmission. In such discrete-time models, time is divided into fixed-length slots and the service or transmission of packets starts and ends at slot boundaries only. In the scientific literature, many results can be found with respect to the analysis of discrete-time single-server queues with various types of (uncorrelated or correlated) packet arrival processes and various

P. Gao (✉)· S. Wittevrongel · J. Walraevens · H. Bruneel
Stochastic Modeling and Analysis of Communication Systems (SMACS) Research Group,
Department of Telecommunications and Information Processing (TELIN),
Ghent University, Sint-Pietersnieuwstraat 41, 9000 Gent, Belgium
e-mail: pg@telin.UGent.be

types of service-time distributions. For systems with multiple servers however fewer results are available. First, most studies of multiserver systems assume constant service times of one slot (see e.g. Bruneel et al. 1992; Bruneel and Kim 1993). Only a limited number of papers consider more general service-time distributions. Multiserver systems with geometrically distributed service times have been studied in Rubin and Zhang (1991), Chaudhry et al. (2001), Artalejo and Hernandez-Lerma (2003), Gao et al. (2003), Chaudhry et al. (2004) and Gao et al. (2004b,c); queues with multiple servers and constant service times of arbitrary length have been studied in Bruneel and Wuyts (1994) and Gao et al. (2004a). Second, in case of multiple servers, mostly an uncorrelated packet arrival process is considered, i.e., the numbers of packet arrivals during the consecutive slots are assumed to be independent (see e.g. Rubin and Zhang 1991; Bruneel et al. 1992; Bruneel and Kim 1993; Bruneel and Wuyts 1994; Artalejo and Hernandez-Lerma 2003; Gao et al. 2004a,b). In Gao et al. (2003) and Gao et al. (2004c), for the case of geometric service times, more general, so-called correlated packet arrival processes are considered, which are more adequate to describe the bursty nature of the traffic in nowadays communication networks.

In the present paper, we investigate the behavior of a discrete-time multiserver buffer system with constant service times of multiple slots and correlated arrivals. Specifically, as a first step, we consider a two-state Markovian arrival process, where the distribution of the number of packet arrivals in a slot is assumed to depend on the value of a Markovian variable that represents the traffic source behavior. There are two possible states, each with geometrically distributed sojourn times. Such an arrival process allows to take into account the correlation between the numbers of packet arrivals during consecutive slots, while still being analytically tractable. For more details we refer to Blondia (1993). Note also that the correlation in the arrival process is strictly Markovian by nature and has an exponential decay over larger time lags. The model qualifies as strictly short range dependent (SRD) and our current analysis does not deal with long range dependent (LRD) traffic; for the latter, other analysis techniques need to be used, we refer to Daniels and Blondia (2000) for the case of single-slot service times. From the above survey, the paper can be seen as a first generalization of Bruneel and Wuyts (1994) and Gao et al. (2004a) to the case of correlated arrivals. It is also an extension of Wittevrongel and Bruneel (1999) to the multiserver case.

The paper is organized as follows. In Sect. 2, we describe the system under study and introduce some notations. In Sect. 3, the PGF's of the partial system contents and system contents are derived, and the mean value, the variance and the tail distribution of the system contents are calculated. In Sect. 4, the characteristics of the packet delay are analyzed. In Sect. 5, some numerical examples and further discussion of the results are given. Finally, the paper is concluded in Sect. 6.

2 System under study

We consider a discrete-time multiserver queueing system with correlated arrivals (see Fig. 1). The specific modeling assumptions are as follows:

(a) The system has an infinite capacity for the storage of packets.
(b) The number of servers (or output channels) is equal to c ($c \geq 1$).

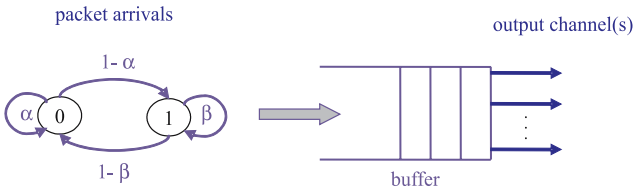

Fig. 1 System model

(c) Time is divided into fixed-length slots. The service (or transmission) of a packet via an output channel can start or end at slot boundaries only, and the service times of the packets are deterministic, equal to s ($s \geq 1$) slots.
(d) Packets are served (transmitted) based on a first-come-first-served (FCFS) queueing discipline.
(e) Packets arrive in the system according to a Markovian arrival process. Specifically, the traffic source has a bursty nature and alternates between two states, state 0 and state 1. Transitions between the states are assumed to occur at slot boundaries. The numbers of consecutive slots during which the source state is 0 or 1 are called 0-times and 1-times, respectively. The 0- and 1-times are independent geometrically distributed random variables with parameters α and β, respectively, i.e.,

$$\text{Prob}[0\text{-time} = n \text{ slots}] = (1-\alpha)\alpha^{n-1}, \quad n \geq 1;$$
$$\text{Prob}[1\text{-time} = n \text{ slots}] = (1-\beta)\beta^{n-1}, \quad n \geq 1.$$

This assumption implies a first-order Markovian correlation in the state of the source, meaning that the probability that the source is in state 0 or state 1 in any given slot is fully determined by the state of the source in the previous slot. In particular, if the source is in state 0 during a slot, it will remain in state 0 with probability α or turn to state 1 with probability $1-\alpha$ during the next slot; if the source is in state 1 during a slot, it will remain in state 1 with probability β or turn to state 0 with probability $1-\beta$ during the next slot (see Fig. 1). The case of uncorrelated source states from slot to slot corresponds to $\gamma = \alpha + \beta - 1 = 0$, where γ is the coefficient of correlation between the source states in two consecutive slots in the steady state. Note also that the autocorrelation with lag k of the source state equals γ^k, so there is an exponential decay over larger time lags. The number of packet arrivals during a slot has an arbitrary distribution which depends only on the source state during the slot. We denote the probability mass functions (PMF's) of the numbers of arrivals during an arbitrary slot where the source state is 0 or 1 by $a_0(n)$ or $a_1(n)$, i.e.,

$$a_m(n) \triangleq \text{Prob}\big[n \text{ arrivals in a slot where the source state is } m\big],$$
$$n \geq 0, \ m = 0, 1,$$

and the corresponding PGF's by $A_0(z)$ and $A_1(z)$, respectively.
(f) The service and arrival processes are assumed to be mutually independent.

With the above assumptions, it is clear that the queueing system can only reach a steady state if

$$\rho = \frac{s\left[\sigma_0\ A_0'(1) + \sigma_1\ A_1'(1)\right]}{c} < 1.$$

Here ρ denotes the load of the system, σ_0 and σ_1 denote the probabilities that the source is in state 0 or state 1, respectively, during an arbitrary slot in the steady state:

$$\sigma_0 = \frac{1-\beta}{2-\alpha-\beta} = \frac{1-\beta}{1-\gamma};$$
$$\sigma_1 = \frac{1-\alpha}{2-\alpha-\beta} = \frac{1-\alpha}{1-\gamma},$$

and $A_0'(1)$ and $A_1'(1)$ are the average arrival rates of packets when the source state is 0 or 1, respectively. In the analysis that follows, we assume this equilibrium condition to be satisfied.

3 System contents and partial system contents

3.1 PGF's of the total and partial system contents

In order to study the system contents, we first introduce the following random variables. We denote by v_k the system contents (i.e., the total number of packets in the buffer system, including the packets under transmission, if any) at the beginning of slot k, a_k denotes the number of packet arrivals during slot k, and t_k is the state of the source during slot k. Furthermore, we let $u_{j,k}$ denote the partial system contents of degree j at slot k, i.e., the number of packets in the system at the beginning of slot k whose service has progressed for at most j slots at the end of slot k. Note that no packets in the system at the beginning of a slot have received more than s slots of service at the end of the slot due to the constant nature of the service times. Then, we have the following system equations:

$$v_k = u_{s,k}, \tag{1}$$
$$u_{j,k+1} = u_{j-1,k} + a_k, \quad 1 \le j \le s, \tag{2}$$
$$u_{0,k} = (u_{s,k} - c)^+, \tag{3}$$

where $(\cdots)^+ = max(0, \ldots)$. Indeed, the right-hand side of (3) is the queue length at the beginning of slot k, i.e., the number of packets present in the system at the beginning slot k whose service has not yet started by the end of the slot. In the steady state, the distributions of the above random variables become independent of the time index k. We denote by $V(z)$ and $U_j(z)$ the PGF's of the random variables v_k and $u_{j,k}$, respectively, when steady state is reached.

The next step is now to transform the system equations (1)–(3) into the z-domain. Since the random variables on the right-hand side of (2) are not independent and a_k is

completely determined by the state t_k of the source in slot k, we define the joint PGF of the random variables $(t_k, u_{j,k})$ as

$$\begin{aligned} Y_{j,k}(x,z) &\triangleq E\left[x^{t_k} z^{u_{j,k}}\right] \\ &= \sum_{m=0}^{1} \sum_{n=0}^{\infty} \text{Prob}\left[t_k = m, u_{j,k} = n\right] x^m z^n. \end{aligned} \tag{4}$$

Using the system equation (2), we then obtain

$$Y_{j,k+1}(x,z) = E\left[x^{t_{k+1}} z^{a_k} z^{u_{j-1,k}}\right], \quad 1 \le j \le s. \tag{5}$$

To further calculate $Y_{j,k+1}(x,z)$ in (5), we need the joint PGF of the random variables (t_{k+1}, a_k). From the arrival process description in Sect. 2, it follows that $\{t_k\}$ is a homogeneous two-state Markov chain and the distribution of a_k depends only on the value of t_k. More specifically, the joint PGF of the random variables (t_{k+1}, a_k) can be written in terms of the PGF of the random variable t_k as follows

$$\begin{aligned} E[x^{t_{k+1}} z^{a_k}] &= E\left[x^{t_{k+1}} z^{a_k} \middle| t_k = 0\right] \cdot \text{Prob}\left[t_k = 0\right] + E\left[x^{t_{k+1}} z^{a_k} \middle| t_k = 1\right] \cdot \text{Prob}\left[t_k = 1\right] \\ &= T_0(x,z) \cdot \text{Prob}\left[t_k = 0\right] + T_1(x,z) \cdot \text{Prob}\left[t_k = 1\right] \\ &= T_0(x,z)\, E\left[\left(\frac{T_1(x,z)}{T_0(x,z)}\right)^{t_k}\right], \end{aligned} \tag{6}$$

where

$$\begin{aligned} T_0(x,z) &= [\alpha + (1-\alpha)x] A_0(z), \\ T_1(x,z) &= [1 - \beta + \beta x] A_1(z). \end{aligned}$$

Combining Eqs. (4)–(6), we can express the steady-state PGF $Y_j(x,z)$ as

$$\begin{aligned} Y_j(x,z) &= \lim_{k\to\infty} T_0(x,z)\, E\left[\left(\frac{T_1(x,z)}{T_0(x,z)}\right)^{t_k} z^{u_{j-1,k}}\right] \\ &= T_0(x,z)\, Y_{j-1}\left(\frac{T_1(x,z)}{T_0(x,z)}, z\right), \quad 1 \le j \le s. \end{aligned} \tag{7}$$

Next, let us introduce the following partial PGF's:

$$Y_{j;m}(z) \triangleq \lim_{k\to\infty} \sum_{n=0}^{\infty} \text{Prob}\left[u_{j,k} = n, t_k = m\right] z^n, \quad 0 \le j \le s.$$

Then the function $Y_j(x,z)$ is expressed as

$$Y_j(x,z) = Y_{j;0}(z) + x Y_{j;1}(z), \quad 0 \le j \le s. \tag{8}$$

Substitution of (8) in the functional equation (7) and identification of the coefficients of equal powers of x on both sides of the resulting equation then yields the following set of two recursive equations for $Y_{j;0}(z)$ and $Y_{j;1}(z)$:

$$\begin{cases} Y_{j;0}(z) = \alpha A_0(z)\, Y_{j-1;0}(z) + (1-\beta) A_1(z)\, Y_{j-1;1}(z), \\ Y_{j;1}(z) = (1-\alpha) A_0(z)\, Y_{j-1;0}(z) + \beta A_1(z)\, Y_{j-1;1}(z), \end{cases}$$

or in a matrix form

$$\begin{bmatrix} Y_{j;0}(z) \\ Y_{j;1}(z) \end{bmatrix} = \begin{bmatrix} \alpha A_0(z) & (1-\beta) A_1(z) \\ (1-\alpha) A_0(z) & \beta A_1(z) \end{bmatrix} \cdot \begin{bmatrix} Y_{j-1;0}(z) \\ Y_{j-1;1}(z) \end{bmatrix}, \quad 1 \le j \le s. \tag{9}$$

By repeated use of Eq. (9), we find

$$\begin{bmatrix} Y_{j;0}(z) \\ Y_{j;1}(z) \end{bmatrix} = M^j \begin{bmatrix} Y_{0;0}(z) \\ Y_{0;1}(z) \end{bmatrix}, \quad 1 \le j \le s, \tag{10}$$

where

$$\begin{aligned} M^j &= \begin{bmatrix} M_j^{00}(z) & M_j^{01}(z) \\ M_j^{10}(z) & M_j^{11}(z) \end{bmatrix} \\ &\triangleq \begin{bmatrix} \alpha A_0(z) & (1-\beta) A_1(z) \\ (1-\alpha) A_0(z) & \beta A_1(z) \end{bmatrix}^j. \end{aligned}$$

In a similar way, now using the system equations (1) and (3), we get

$$\begin{aligned} Y_{0,k}(x,z) &= E\left[x^{t_k} z^{u_{0,k}}\right] \\ &= E\left[x^{t_k} z^{(v_k - c)^+}\right] \\ &= z^{-c} E\left[x^{t_k} z^{v_k} \middle| v_k \ge c\right] \text{Prob}[v_k \ge c] + E\left[x^{t_k} \middle| v_k < c\right] \text{Prob}[v_k < c]. \end{aligned} \tag{11}$$

The steady-state PGF $Y_0(x,z)$ is then obtained as

$$Y_0(x,z) = z^{-c} Y_s(x,z) + \sum_{m=0}^{1} \sum_{n=0}^{c-1} v(n,m) x^m (1 - z^{n-c}), \tag{12}$$

where

$$\begin{aligned} v(n,m) &\triangleq \lim_{k \to \infty} \text{Prob}[v_k = n, t_k = m] \\ &= \text{Prob}[v = n, t = m], \quad m = 0, 1; \ 0 \le n \le c-1. \end{aligned}$$

Using (8), we find

$$\begin{bmatrix} Y_{0;0}(z) \\ Y_{0;1}(z) \end{bmatrix} = z^{-c} \left\{ \begin{bmatrix} Y_{s;0}(z) \\ Y_{s;1}(z) \end{bmatrix} + \begin{bmatrix} \sum_{n=0}^{c-1} v_{n0}(z) \\ \sum_{n=0}^{c-1} v_{n1}(z) \end{bmatrix} \right\}, \tag{13}$$

where

$$v_{nm}(z) \triangleq v(n,m)(z^c - z^n), \quad m = 0, 1; \ 0 \le n \le c-1.$$

Combination of Eqs. (10) and (13) finally gives

$$z^c \begin{bmatrix} Y_{j;0}(z) \\ Y_{j;1}(z) \end{bmatrix} = M^j \left\{ \begin{bmatrix} Y_{s;0}(z) \\ Y_{s;1}(z) \end{bmatrix} + \begin{bmatrix} \sum_{n=0}^{c-1} v_{n0}(z) \\ \sum_{n=0}^{c-1} v_{n1}(z) \end{bmatrix} \right\}, \quad 0 \le j \le s. \tag{14}$$

Note that the correctness of Eq. (14) can also be seen as follows. First, $z^{-c}\left[Y_{s;i}(z) + \sum_{n=0}^{c-1} v_{ni}(z)\right]$ is the partial PGF of the number of waiting packets while the arrival process is in state i [see also (13)]. Also, the transposed of M^j holds the PGF's of the number of arrivals in a time interval of length j slots. Therefore, since none of the packets waiting at the beginning of a slot have received j slots of service j slots further in time, Eq. (14) readily follows. The entries of the matrix M^j can be expressed in terms of the 2 eigenvalues λ_1 and λ_2 of the matrix M, by using the property that ${\lambda_1}^j$ and ${\lambda_2}^j$ are the 2 eigenvalues of the matrix M^j, as follows:

$$\begin{aligned} M_j^{00}(z) &= \frac{\lambda_1^{j+1} - \lambda_2^{j+1} + \beta A_1(z)(\lambda_2^j - \lambda_1^j)}{\lambda_1 - \lambda_2}; \\ M_j^{01}(z) &= \frac{(1-\beta) A_1(z)(\lambda_1^j - \lambda_2^j)}{\lambda_1 - \lambda_2}; \\ M_j^{10}(z) &= \frac{(1-\alpha) A_0(z)(\lambda_1^j - \lambda_2^j)}{\lambda_1 - \lambda_2}; \\ M_j^{11}(z) &= \frac{\lambda_1^{j+1} - \lambda_2^{j+1} + \alpha A_0(z)(\lambda_2^j - \lambda_1^j)}{\lambda_1 - \lambda_2}, \end{aligned}$$

where

$$\begin{aligned} \lambda_\tau \ &= \frac{\alpha A_0(z) + \beta A_1(z)}{2} \\ &\quad \pm \frac{\sqrt{[\alpha A_0(z) + \beta A_1(z)]^2 - 4\gamma A_0(z) A_1(z)}}{2}, \quad \tau = 1, 2, \end{aligned}$$

with $\pm$ being $+$ for $\tau = 1$ and $-$ for $\tau = 2$. Note that λ_1 and λ_2 are functions of z. However, we write λ_τ instead of $\lambda_\tau(z)$ to ease the notation. For the detail of the calculations of the matrix M^j, we refer to Gao (2006).

When $j = s$, Eq. (14) leads to a set of linear equations for $Y_{s;0}(z)$ and $Y_{s;1}(z)$, from which the partial PGF's $Y_{s;0}(z)$ and $Y_{s;1}(z)$, as well as the PGF of the system contents $V(z) = Y_{s;0}(z) + Y_{s;1}(z)$ can be calculated. Substitution of the results for $Y_{s;0}(z)$ and $Y_{s;1}(z)$ in (14) moreover enables the calculation of the PGF $U_j(z) = Y_{j;0}(z) + Y_{j;1}(z)$ of the partial system contents of degree j, $0 \le j \le s$. By means of

$$\lambda_1 + \lambda_2 = \alpha A_0(z) + \beta A_1(z);$$
$$\lambda_1 \lambda_2 = (\alpha + \beta - 1) A_0(z) A_1(z),$$

and some straightforward, but rather tedious mathematical manipulations, the following expressions are obtained:

$$V(z) = \frac{1}{\lambda_1 - \lambda_2} \sum_{n=0}^{c-1} \left\{ \left[\frac{\lambda_1 \lambda_2^s}{z^c - \lambda_2^s} - \frac{\lambda_2 \lambda_1^s}{z^c - \lambda_1^s} \right] v_n(z) + \frac{z^c (\lambda_1^s - \lambda_2^s)[A_0(z) v_{n0}(z) + A_1(z) v_{n1}(z)]}{(z^c - \lambda_1^s)(z^c - \lambda_2^s)} \right\}; \tag{15}$$

$$U_0(z) = z^{-c} \left[V(z) + \sum_{n=0}^{c-1} v_n(z) \right]; \tag{16}$$

$$U_j(z) = \frac{\lambda_1^j - \lambda_2^j}{\lambda_1^s - \lambda_2^s} V(z) - \frac{\lambda_1^j \lambda_2^s - \lambda_2^j \lambda_1^s}{\lambda_1^s - \lambda_2^s} U_0(z), \quad 0 \le j \le s, \tag{17}$$

where

$$v_n(z) \triangleq v_{n0}(z) + v_{n1}(z) = v(n)(z^c - z^n);$$
$$v(n) \triangleq v(n, 0) + v(n, 1), \quad 0 \le n \le c - 1.$$

In order to determine $V(z)$ completely, we need to find the $2c$ unknown constants $v(n, 0)$ and $v(n, 1)$ $(0 \le n \le c - 1)$ in (15). 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 (Kleinrock 1975), it can be shown that the factor $(z^c - \lambda_1^s)(z^c - \lambda_2^s)$ in the denominator of $V(z)$ has exactly $2c - 1$ roots inside the unit disk. We denote these roots by z_i, $1 \le i \le 2c - 1$. Since $V(z)$ is analytic for $|z| < 1$, the numerator of $V(z)$ must also be zero at these points. Thus, we have

$$(\lambda_1^s - \lambda_2^s) z^c \sum_{n=0}^{c-1} \Big\{ \left[\lambda_1 \delta(z^c - \lambda_2^s) + \lambda_2 \delta(z^c - \lambda_1^s) \right] v_n(z) - A_0(z) v_{n0}(z) - A_1(z) v_{n1}(z) \Big\} \Big|_{z=z_i} = 0, \quad 1 \le i \le 2c - 1, \tag{18}$$

where $\delta(\cdot)$ is the Kronecker delta function, which is 1 when its argument is zero and 0 otherwise. From the normalization condition $V(1) = 1$ and Eq. (15), we moreover

find that

$$\sum_{n=0}^{c-1}(c-n)v(n) = c - s\lambda_1'(1) = c(1-\rho), \tag{19}$$

where $\lambda_1'(1) = \sigma_0 A_0'(1) + \sigma_1 A_1'(1)$ is the first-order derivative of λ_1 at $z = 1$, which also denotes the mean number of packet arrivals during an arbitrary slot. With Eqs. (18) and (19), the constants $v(n, 0)$ and $v(n, 1)$ $(0 \le n \le c-1)$ can be found. Note that this requires the calculation of the roots z_i of the denominator of $V(z)$ inside the unit disk. For medium-sized problems this can be done by standard algebra software packages such as e.g. MAPLE. For problems of higher dimensions, one has to turn to more specific algorithms, such as described e.g. in Kravanja and Van Barel (2000).

The expressions in the above analysis of the system contents have been derived under the implicit assumption that $\lambda_1(z) \neq \lambda_2(z)$. In case for a given z, $\lambda_1(z) = \lambda_2(z)$, the analysis can be adapted based on techniques from matrix theory described e.g. in (Gantmacher 1998, p.101). It turns out that this comes down to applying de l'Hospital's rule to the obtained expressions, which can be done e.g. by taking the limit for $\lambda_1 \to \lambda_2$, considering λ_2 to be a constant and λ_1 the variable.

3.2 Moments and tail distribution of the system contents

Once $V(z)$ is determined, some important performance measures for the system, such as the mean value, the variance and the tail distribution of the system contents, can be calculated. The mean system contents $E[v]$ can be obtained by taking the first-order derivative of Eq. (15) with respect to z in $z = 1$. Using de l'Hospital's rule twice, we get

$$\begin{aligned} E[v] &= V'(1) \\ &= \frac{\sum_{n=0}^{c-1}[A_0'(1)\, v(n,0) + A_1'(1)\, v(n,1)](c-n)}{c\,(1-\gamma)(1-\rho)} \\ &\quad - \frac{\rho c}{s(1-\gamma)} + \frac{s\lambda_1''(1) + \sum_{n=0}^{c-1}(c^2-n^2)v(n)}{2c(1-\rho)} \\ &\quad - \frac{c(1-\rho)}{2} + \frac{\rho(s-c\rho)}{2s(1-\rho)}, \end{aligned}$$

where $\lambda_1''(1)$ is the second-order derivative of λ_1 with respect to z at $z = 1$. Higher-order moments of the system contents can be derived in a similar way. For instance, the variance of the system contents follows from the relation

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

Another important performance characteristic for a buffer is the tail distribution of the system contents, i.e., the probability that the system contents equals a given value n, for sufficiently large n. In principle, the tail distribution of a discrete random variable can be determined by applying the inversion formula for z-transforms and

Cauchy's residue theorem from complex analysis (see e.g. Kleinrock 1975) 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 argued in Bruneel and Kim (1993), the system-contents distribution exhibits a geometric tail behavior. That is, for sufficiently large values of n, the tail distribution of the system contents can be approximated as

$$\text{Prob}[v = n] \approx -C_v \, z_v^{-n-1}, \tag{20}$$

where z_v is the pole of $V(z)$ with the smallest modulus (outside the unit disk), and the constant C_v is the residue of $V(z)$ at $z = z_v$. The dominant pole z_v must necessarily be real and positive in order to ensure that the tail distribution is nonnegative anywhere (Bruneel and Kim 1993). From (15), it follows that z_v is a real positive zero of the denominator of $V(z)$. The residue C_v can be calculated from (15) as

$$C_v = \begin{cases} \dfrac{\sum_{n=0}^{c-1} \{A_0(z)v_{n0}(z) + A_1(z)v_{n1}(z) - \lambda_2 v_n(z)\}}{(\lambda_1 - \lambda_2)\,[c/z - s\lambda_1'(z)/\lambda_1]}\Big|_{z=z_v}, & \text{when } {z_v}^c = \lambda_1(z_v)^s; \\ \dfrac{\sum_{n=0}^{c-1} \{A_0(z)v_{n0}(z) + A_1(z)v_{n1}(z) - \lambda_1 v_n(z)\}}{(\lambda_2 - \lambda_1)\,[c/z - s\lambda_2'(z)/\lambda_2]}\Big|_{z=z_v}, & \text{when } {z_v}^c = \lambda_2(z_v)^s. \end{cases} \tag{21}$$

From (20), the probability that the system contents exceeds a given threshold N, for large N, follows as

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

This probability (for an infinite buffer model) is often used to estimate the packet loss probability or buffer overflow probability that would be observed in case of a buffer with a finite storage capacity N (see e.g. Bisdikian et al. 1993).

4 Packet delay

4.1 PGF of the packet delay

The delay of a packet is defined as the total number of slots between the end of the slot during which the packet arrives in the system and the end of the slot where the packet finishes its transmission and leaves the system. Let $D(z)$ be the PGF of the delay d that an arbitrary packet experiences in the system. In this section, we analyze the characteristics of the packet delay by means of the general relationship between partial system contents and packet delay established in Gao et al. (2005) and Gao (2006). Specifically, it has been shown in Gao et al. (2005) and Gao (2006) that for any discrete-time multiserver system with constant service times of multiple slots and a FCFS queueing discipline, the PGF $D(z)$ can be expressed in terms of the PGF's of the partial system

contents as

$$D(z^c) = \frac{1-z^c}{c\sigma} \sum_{j=0}^{c-1} \frac{\theta^j z^s}{(1-\theta^j z^s)^2} \cdot \sum_{i=0}^{s-1} z^{ci} \Big[U_{s-i-1}(\theta^j z^s) - U_{s-i}(\theta^j z^s) \Big], \quad (22)$$

where $\theta = \exp(2\pi I/c)$ with $I^2 = -1$. The relationship (22) holds regardless of the exact nature of the arrival process, and therefore it can also be applied to derive the delay characteristics for the considered system with a two-state (first-order Markovian) correlated traffic source. Here σ, the mean number of packet arrivals per slot, equals $\lambda_1'(1)$. Combination of (22) and (15)–(17) finally gives

$$\begin{aligned} D(z^c) = {} & \frac{1-z^c}{c\lambda_1'(1)} \sum_{j=0}^{c-1} \frac{\theta^j z^s}{(1-\theta^j z^s)^2 (z^c - \lambda_1)(z^c - \lambda_2)} \\ & \times \sum_{n=0}^{c-1} \Big\{ (z^c + \lambda_1\lambda_2 - \lambda_1 - \lambda_2) v_n(\theta^j z^s) + (1-z^c) \\ & \times \Big[A_0(\theta^j z^s) v_{n0}(\theta^j z^s) + A_1(\theta^j z^s) v_{n1}(\theta^j z^s) \Big] \Big\}. \end{aligned} \quad (23)$$

Note that in Eq. (23) λ_1 and λ_2 are functions of $\theta^j z^s$, i.e., functions $\lambda_1(\theta^j z^s)$ and $\lambda_2(\theta^j z^s)$.

4.2 Moments and tail distribution of the packet delay

The mean value of the packet delay can be found from (23) by evaluation of the first-order derivative of the PGF $D(z^c)$ with respect to z at $z = 1$. Specifically, we get

$$E[d] = D'(1) = \frac{1}{c} \frac{dD(z^c)}{dz}\Big|_{z=1} = \frac{E[v]}{\lambda_1'(1)},$$

in agreement with Little's theorem. In a similar way, we can also obtain higher-order moments of the packet delay, by calculating the appropriate higher-order derivatives of $D(z^c)$ at $z = 1$. For instance, the variance of the packet delay (delay jitter) can be obtained as

$$\begin{aligned} \mathrm{Var}[d] &= D''(1) + D'(1) - D'(1)^2 \\ &= \frac{1}{c^2} \frac{d^2 D(z^c)}{dz^2}\Big|_{z=1} + \frac{1}{c} D'(1) - D'(1)^2. \end{aligned}$$

In order to derive the tail distribution of the delay of a packet, we use a similar procedure as for the system contents. However, from expression (23) for $D(z^c)$, we note that this function does not satisfy the condition that it has only one pole with minimal modulus. Indeed, if z_v is the dominant pole of $V(z)$, i.e., the zero of $[z^c - \lambda_1(z)^s] \cdot [z^c - \lambda_2(z)^s]$

with smallest modulus outside the unit disk, then $z_d(0) \triangleq z_v^{1/s}$ is the zero with minimal modulus outside the unit disk of the factor $[z^c - \lambda_1(z^s)][z^c - \lambda_2(z^s)]$ in the denominator of $D(z^c)$. Due to $\theta^{mc} = 1$ for any integer value of m, z^c remains unchanged when z is multiplied by θ^{-m}, and therefore $z_d(m) = \theta^{-m} z_v^{1/s}$ $(0 \leq m \leq c-1)$ 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 $[z^c - \lambda_1(\theta^j z^s)][z^c - \lambda_2(\theta^j z^s)]$ 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 $z_d(m)$, $0 \leq m \leq c-1$, and keeping in mind that $\text{Prob}[d = n]$ is the coefficient of z^{cn} in the series expansion of $D(z^c)$, we finally get

$$\begin{aligned} \text{Prob}[d = n] &\approx -\sum_{m=0}^{c-1} \frac{b_m}{z_d(m)} [z_d(m)]^{-cn} \\ &= -\sum_{m=0}^{c-1} \frac{b_m}{z_d(m)} z_v^{-cn/s} \\ &= -C_d \, {z_v}^{-cn/s}, \end{aligned} \tag{24}$$

for sufficiently large n. In (24), b_m is the residue of $D(z^c)$ at the point $z = z_d(m)$ and is given by

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

where $N_m(z)$ and $R_m(z)$ are the numerator and the denominator, respectively, of the term in (23) corresponding to the index value $j = (ms) \bmod c$. Using the expressions (23) and (21), we find

$$C_d = \sum_{m=0}^{c-1} \frac{b_m}{z_d(m)} = \frac{z_v^{-c/s}}{\lambda_1'(1)} \left(\frac{1 - z_v^{c/s}}{1 - z_v} \right)^2 C_v.$$

The probability that the packet delay exceeds a given threshold T follows from (24) as

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

5 Discussion of results

In order to illustrate the influence of various parameters of the model, such as the degree of correlation in the arrival process, the number of servers and the length of the service times, on the system behavior, we present a number of numerical examples in this section.

Table 1 The three sets of arrival distributions

Set	1	2	3
$A_0(z)$	$\dfrac{1}{1+\Lambda_1-\Lambda_1 z}$	$1-\Lambda_2+\Lambda_2 z$	$1-\Lambda_3+\Lambda_3 z$
$A_1(z)$	1	$\dfrac{1}{1+2\Lambda_2-2\Lambda_2 z}$	$\dfrac{1}{1+\Lambda_3-\Lambda_3 z}$

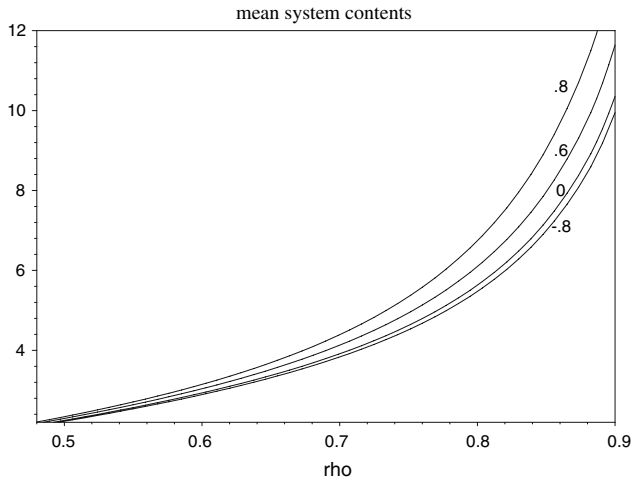

Fig. 2 Mean system contents versus load ρ

Throughout this section, we assume that the packet arrivals during states 0 and 1 are governed by the sets of distributions shown in Table 1.

In the first set, packet arrivals are governed by a geometric distribution with arrival rate Λ_1 during state 0 and there are no packet arrivals when the source is in state 1. In the second set, packet arrivals are governed by a Bernoulli distribution with rate Λ_2 during state 0 and a geometric distribution with rate $2\Lambda_2$ during state 1. In the third set, the arrival distributions are of the same type as for the second set, but with the same arrival rate Λ_3 during both states.

In Fig. 2, we have plotted the mean system contents versus the load ρ, for $c = 4$, $s = 4$, $\alpha = \beta$, arrival distributions of set 2, and various values of the source state correlation coefficient γ, namely $\gamma = -0.8,\ 0,\ 0.6,\ 0.8$. The figure clearly shows that for a given ρ, the mean system contents increases as γ increases. Especially, for higher loads ρ, the system contents may be heavily underestimated when the (positive) correlation between the source states in two consecutive slots is not taken into account.

In Figs. 3–5, we assume $\alpha = 0.7$ and $\beta = 0.8$. The source state correlation coefficient γ then equals 0.5. In Fig. 3, the overflow probability Prob$[v > N]$ is shown as a function of N (see the solid lines), for $\rho = 0.8$, $c = 4$, $s = 4$ and the three sets of arrival distributions. The mean service rate is equal to $c/s = 1$. For comparison, we have also plotted (by means of the analysis of Gao et al. 2004c) the corresponding curves in case the service times of packets are geometrically distributed with parameter $1 - \mu$, where we set $\mu = 0.25$, so that the mean service rate $c\mu$ is also equal to 1

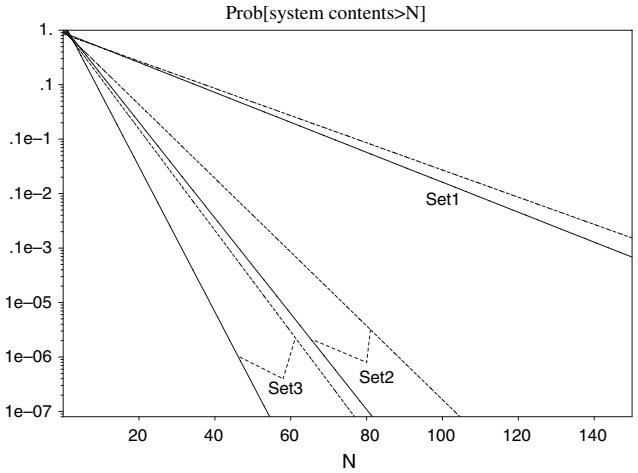

Fig. 3 Overflow probability, Prob[$v > N$], versus N

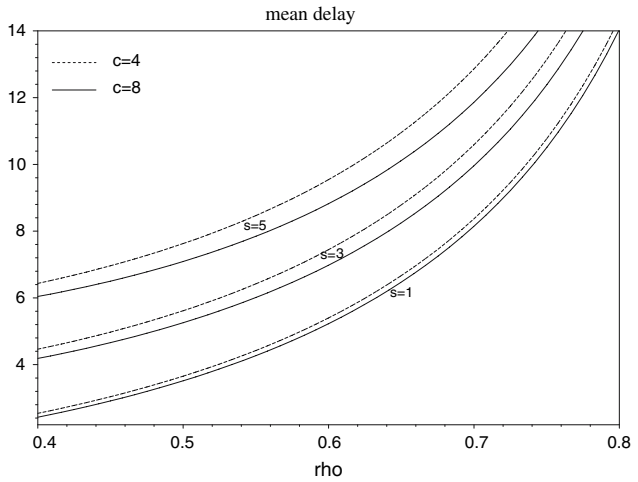

Fig. 4 Mean packet delay versus load ρ

(see the dashed lines). We note that for a given type of service-time distribution, the first set of arrival distributions gives the highest overflow probability, while the third set gives the smallest value. This observation can be understood intuitively from the fact that the variance of the number of arrivals per slot decreases in the order of set 1, set 2 and set 3. Indeed, the higher the variance of the number of arrivals and, hence, the more fluctuation of the arrival process, the higher we expect the buffer contents to be. For a given set of arrival distributions, the overflow probability is higher for geometric service times than for constant service times. This is also intuitively clear since the variance of the service times is higher for the geometric distribution. The required buffer size N to satisfy a given loss bound can also be estimated from Fig. 3.

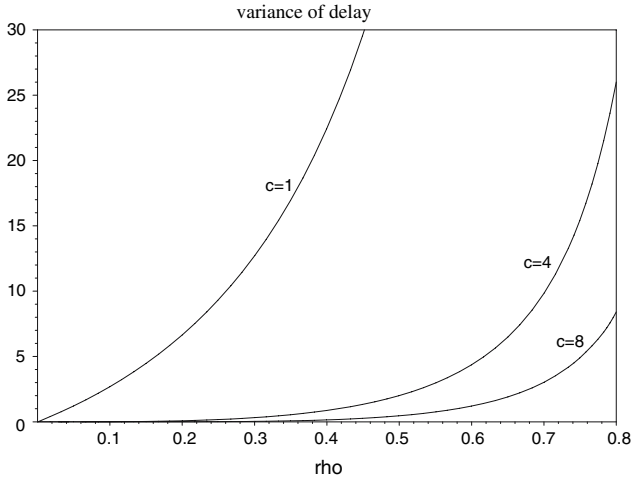

Fig. 5 Variance of the packet delay versus load ρ

In Fig. 4, the mean packet delay is plotted versus the load ρ, for the arrival distributions of set 1, for $s = 1,\ 3,\ 5$ and $c = 4, 8$. For given values of c and s, we see that the mean packet delay increases as ρ increases. For a given ρ, the mean delay increases as the service times become longer and/or the number of servers decreases. We also observe that the longer the service times, the higher the impact of the number of servers on the packet delay, especially when the load gets higher.

In Fig. 5, the variance of the packet delay is shown versus ρ, for the arrival distributions of set 3, for $s = 8$ and $c = 1,\ 4,\ 8$. Clearly, for a given value of ρ, the delay jitter decreases as the number of servers increases.

6 Concluding remarks

In this paper, we have studied the behavior of a discrete-time infinite-capacity buffer system with multiple servers and constant service times of multiple slots. Packet arrivals to the system are described by a general but state-dependent arrival process. Specifically, a two-state traffic source with a first-order Markovian correlation in the state of the source is considered. An analytical technique based on generating functions has been presented for the analysis of the system. Closed-form expressions for the mean values, the variances and the tail distributions of the system contents and the packet delay have been derived.

Some numerical results have been presented to illustrate the analysis. The numerical examples show that the characteristics of the system contents and the packet delay are sensitive to both the arrival process and the service mechanism. The analysis method presented in this paper relies on the fact that the arrival process is a two-state Markovian process; the obtained results are expressed in terms of the eigenvalues of a 2×2 matrix; these eigenvalues can be calculated analytically in case of a 2×2 matrix. As future work, we intend to investigate how the analysis can be generalized to the case

of an m-state ($m > 2$) Markovian arrival process. It is expected that basically similar methods could be used in this case. However, as the dimensions of the occurring matrices will grow, more numerical work will be involved.

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

References

Artalejo J, Hernandez-Lerma O (2003) Performance analysis and optimal control of the Geo/Geo/c queue. Perform Eval 52(1):15–39

Bisdikian C, Lew J, Tantawi A (1993) On the tail approximation of the blocking probability of single server queues with finite buffer capacity. In: Proceedings of the 2nd international conference on queueing networks with finite capacity, Research Triangle Park, pp 267–80

Blondia C (1993) A discrete-time batch Markovian arrival process as B-ISDN traffic model. Belg J Oper Res Stat Comput Sci (JORBEL) 32(3–4):3–23

Bruneel H, Kim BG (1993) Discrete-time models for communication systems including ATM. Kluwer, Boston

Bruneel H, Steyaert B, Desmet E, Petit G (1992) An analytical technique for the derivation of the delay performance of ATM switches with multiserver output queues. Int J Digit Analog Commun Syst 5:193–201

Bruneel H, Wuyts I (1994) Analysis of discrete-time multiserver queueing models with constant service times. Oper Res Lett 15:231–236

Chaudhry ML, Gupta UC, Goswami V (2001) Modeling and analysis of discrete-time multiserver queues with batch arrivals: GI(X)/Geom/m. INFORMS J Comput 13(3):172–180

Chaudhry ML, Gupta UC, Goswami V (2004) On discrete-time multiserver queues with finite buffer: GI/Geom/m/N. Comput Oper Res 31(13):2137–2150

Daniels T, Blondia C (2000) Tail transitions in queues with long range dependent input. Lect Notes Comput Sci 1815:264–274

Gantmacher FR (1998) The theory of matrices, vol 1. AMS Chelsea Publishing, Providence

Gao P (2006) Discrete-time multiserver queues with generalized service times. PhD thesis, Ghent University

Gao P, Wittevrongel S, Bruneel H (2003) Delay against system contents in discrete-time G/Geom/c queue. Electron Lett 39(17):1290–1292

Gao P, Wittevrongel S, Bruneel H (2004a) Delay analysis for a discrete-time GI-D-c queue with arbitrary-length service times. In Proceedings of EPEW 2004, Lecture Notes in Computer Science, vol 3236, Toledo, pp 184–195

Gao P, Wittevrongel S, Bruneel H (2004b) Discrete-time multiserver queues with geometric service times. Comput Oper Res 31(1):81–99

Gao P, Wittevrongel S, Bruneel H (2004c) On the behavior of multiserver buffers with geometric service times and bursty input traffic. IEICE Trans Commun E87-B(12):3576–3583

Gao P, Wittevrongel S, Bruneel H (2005) Relationship between delay and partial system contents in multiserver queues with constant service times. In: Booklet of abstracts of ORBEL 19, Louvain-la-Neuve, pp 72–74

Kleinrock L (1975) Queueing systems, volume I: theory. Wiley, New York

Kravanja P, Van Barel M (2000) Computing the zeros of analytic functions. Lect Notes Math 1727:1–59

Rubin I, Zhang Z (1991) Message delay and queue-size analysis for circuit-switched TDMA systems. IEEE Trans Commun 39:905–914

Wittevrongel S, Bruneel H (1999) Discrete-time queues with correlated arrivals and constant service times. Comput Oper Res 26(2):93–108