Analysis of Discrete-Time Buffers with General Session-Based Arrivals

Sabine Wittevrongel, Stijn De Vuyst, and Herwig Bruneel

SMACS Research Group*, Department of Telecommunications
and Information Processing (TELIN), Ghent University,
Sint-Pietersnieuwstraat 41, B-9000 Gent, Belgium
Tel.: +32-9-264 89 01, Fax: +32-9-264 42 95
{sw,sdv,hb}@telin.ugent.be

Abstract. Session-based arrival streams are a new approach for modelling the traffic generated by users in a telecommunication network. In this paper, we analyze the behavior of a discrete-time buffer with one output line, an infinite storage capacity and session-based arrivals. Users from an infinite user population can start and end sessions during which they are active and send packets to the buffer. Each active user generates a random but strictly positive number of packets per time slot. Unlike in previous work, there are T different session types and for each type, the session-length distribution is general. The resulting discrete-time queueing model is analyzed by means of an analytical technique, which is basically a generating-functions approach that uses an infinite-dimensional state description. Expressions are obtained for the steady-state probability generating functions of both the buffer content and the packet delay. From these, the mean values and the tail distributions of the buffer content and the packet delay are derived as well. Some numerical examples are shown to illustrate the influence of the session-based packet arrival process on the buffer behavior.

Keywords: Discrete-time queueing model, Session-based arrivals, General session lengths, Analytic study, Buffer content and delay.

1 Introduction

In many subsystems of packet-based telecommunication networks, buffers are used for the temporary storage of information packets. In order to assess the behavior of these buffers, appropriate traffic models need to be considered. In particular, there is a continuing need for models that can accurately capture the correlated nature of the traffic streams in modern telecommunication networks. Session-based arrival streams are a new traffic modelling approach. We consider an infinite user population where each user can start and end sessions. During a session a user is active and sends information packets through the communication system. Since we focus on discrete-time models, we assume time is divided

* SMACS: Stochastic Modeling and Analysis of Communication Systems.

K. Al-Begain, D. Fiems, and G. Horváth (Eds.): ASMTA 2009, LNCS 5513, pp. 189–203, 2009.

into fixed-length slots. Each active user generates a random but strictly positive number of packets per slot. Note that such session-based packet generation introduces time correlation in the packet arrival process. Session-based arrivals are illustrated in Fig. 1. Here the term session length denotes the number of consecutive slots during which a user remains active, whereas the number of packets generated per slot during a session is referred to as the session bandwidth.

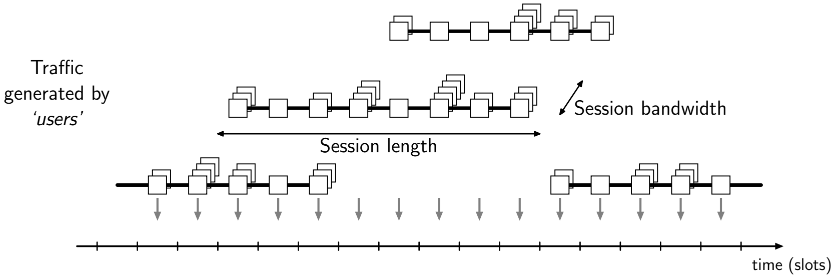

Fig. 1. Session-based packet arrivals: session length and bandwidth

A possible application of session-based arrival processes is depicted in Fig. 2. A web server accepts requests from users for a certain web page or file and responds by sending the requested file to the user. The web server is connected to the internet through a gateway and this gateway contains a buffer for outgoing data from the server to the internet. If we define the download of a file by a user as one session, the traffic towards the output buffer of the web server can be adequately described by a session-based arrival process.

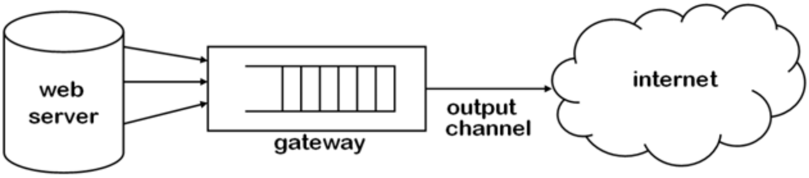

Fig. 2. A web server connected to the internet through a gateway

In previous work [10,11], we have analyzed a discrete-time queue with session-based arrivals and geometrically distributed session lengths. The train arrival process, where messages (the equivalent of what we consider sessions) arrive to the queue at the rate of exactly one packet per slot, is considered in [3, 5, 7, 14, 15, 16]. Also somewhat related are the on/off-type arrival models studied in [8, 12, 18], where a finite number of users generate one packet per slot during on-periods and no packets during off-periods. In [6], messages consisting of a fixed number of packets are considered in case of an uncorrelated packet arrival process. A related continuous-time model is analyzed in [1].

The aim of the present paper is to further extend the previous analyses to a discrete-time queueing system with general session-based arrivals. Specifically, unlike in previous work, we consider *heterogeneous* sessions of T different types with *general* type-dependent session-length distributions. This extension allows e.g. to take into account the fact that files on a web server are typically either small or very large [2]. A model with generally distributed session lengths moreover makes it possible to investigate the impact of the nature of the session-length distributions on the buffer behavior. We develop a mathematical analysis technique, that makes extensive use of probability generating functions (PGFs). As opposed to our previous work for geometric session lengths (see [10, 11]), an infinite-dimensional state description is required in case of generally distributed session lengths, which seriously complicates the analysis.

The outline of the paper is as follows. In Sect. 2, we describe the queueing model under study. In Sect. 3, a set of state variables is defined and the system equations are established. A functional equation for the joint PGF of the system state vector is obtained in Sect. 4. Some further characteristics of the session-based packet arrival process are studied in Sect. 5. Section 6 concentrates on the derivation of the mean value, the PGF and the tail distribution of the buffer content from the functional equation. The packet delay is analyzed in Sect. 7, for a first-come-first-served (FCFS) queueing rule for packets. Some numerical examples are discussed in Sect. 8. Finally, the paper is concluded in Sect. 9.

2 Queueing Model Description

We study a discrete-time queueing system with one single output line and an infinite storage capacity for packets. As usual for discrete-time models (see e.g. [4,13]), the time axis is divided into fixed-length slots and transmissions from the buffer can only start at slot boundaries. Therefore, when the queueing system is empty at the beginning of a slot, no packet can leave the buffer at the end of that slot, even if some packets have arrived to the buffer during the slot.

A session-based arrival process is considered. Users from an infinite user population can start and end sessions during which they are active and send packets to the queueing system. When a user starts a session, he generates a random but strictly positive number of packets per slot. The session ends when the user has no more data left to send.

There are T different session types. For sessions of type t $(1 \leq t \leq T)$, the session lengths (expressed in slots) are assumed to be independent and identically distributed (i.i.d.) random variables with the following probability mass function (PMF) and PGF:

$$\ell_t(i) = \text{Prob}[\text{session length of type } t \text{ is } i \text{ slots}] \ , \quad i \geq 1 \ ; \tag{1}$$

$$L_t(z) = \sum_{i=1}^{\infty} \ell_t(i)\, z^i \ . \tag{2}$$

The numbers of new sessions of type t started by the user population during the consecutive slots are assumed to be i.i.d. random variables with common PGF $S_t(z)$. Since in normal conditions, internet users act independently from each other, this seems like a realistic assumption. The numbers of packets generated per slot during a session of type t are assumed to be i.i.d. with PGF $P_t(z)$, where $P_t(0)$ equals zero, since at least one packet is generated per slot per session. Sessions of different types are assumed to be independent.

The queueing system has an unreliable output line subject to random failures that are assumed to occur independently from slot to slot. The output line availability is modelled by a parameter σ. Specifically, σ is the probability that the output line is available during a slot. The transmission times of the packets from the buffer are assumed to be constant, equal to one slot per packet. So, whenever the queueing system is nonempty at the beginning of a slot, a packet will leave the buffer at the end of this slot with probability σ and no packet will leave with probability $1-\sigma$, independently from slot to slot. Note that these assumptions result in a geometric distribution (with parameter $1-\sigma$) for the effective transmission times required for the successful transmission of a packet from the queueing system and the mean effective transmission time of a packet equals $1/\sigma$.

3 System Equations

The goal of this section is to introduce a Markovian state description for the queueing system described above. In order to do so, we first take a closer look at the packet arrival process. Let us define $s_k(t)$ as the number of new sessions of type t started during slot k. In view of the model description of Sect. 2, for a given t, the random variables $s_k(t)$ are i.i.d. with common PGF $S_t(z)$. Let $a_{n,k}(t)$ be the random variable representing the number of active sessions of type t that are already active for exactly n slots during slot k. Then the following relationships hold:

$$a_{1,k}(t) = s_k(t) \ ; \tag{3}$$

$$a_{n,k}(t) = \sum_{i=1}^{a_{n-1,k-1}(t)} c_{n-1,k}^{i}(t)\,, \quad n > 1 \ . \tag{4}$$

The random variable $c_{n-1,k}^{i}(t)$ in (4) takes on the values 0 or 1, and equals 1 if and only if the ith active session of type t that was in its $(n-1)$th slot during slot $k-1$, remains active in slot k. We define $\pi_t(n-1)$ as the probability that a session of type t that was $n-1$ slots long continues by at least one more slot, i.e.,

$$\pi_t(n-1) \triangleq \frac{1-\sum\limits_{i=1}^{n-1} \ell_t(i)}{1-\sum\limits_{i=1}^{n-2} \ell_t(i)} \ . \tag{5}$$

Hence, we have that for given n and t, the $c^i_{n-1,k}(t)$'s are i.i.d. random variables with common PGF

$$C_{n-1,t}(z) \triangleq E\left[z^{c^i_{n-1,k}(t)}\right] = 1 - \pi_t(n-1) + \pi_t(n-1)\,z\,, \quad n > 1 \ , \tag{6}$$

where $E[.]$ is the expected value of the argument between square brackets.

Next, let m_k denote the total number of packets generated during slot k. Then m_k can be expressed as

$$m_k = \sum_{t=1}^{T} \sum_{n=1}^{\infty} \sum_{i=1}^{a_{n,k}(t)} p^i_{n,k}(t) \ , \tag{7}$$

where $p^i_{n,k}(t)$ represents the number of packets generated during slot k by the ith session of type t that is already active for exactly n slots. From Sect. 2, it follows that for a given t, the random variables $p^i_{n,k}(t)$ are i.i.d. with PGF $P_t(z)$.

Finally, let u_k denote the buffer content (i.e., the total number of packets in the queueing system, including the packet in transmission, if any) after slot k. The evolution of the buffer content is governed by the following system equation:

$$u_k = (u_{k-1} - r_k)^+ + m_k \ , \tag{8}$$

where $(.)^+ = \max(.,0)$ and the r_k's are i.i.d. Bernoulli random variables equal to zero with probability $1-\sigma$ and equal to one with probability σ, in view of the random interruptions of the output line.

From the above system equations (3)–(8) it is easily seen that the set of vectors $\left\{\left(\underline{\mathbf{a}}_{1,k}, \ldots, \underline{\mathbf{a}}_{T,k}, u_k\right)\right\}$, where $\underline{\mathbf{a}}_{t,k} = (a_{1,k}(t), a_{2,k}(t), \ldots)$, constitutes a Markov chain. The state of the queueing system after slot k can hence be fully described by the infinite-dimensional vector $\left(\underline{\mathbf{a}}_{1,k}, \ldots, \underline{\mathbf{a}}_{T,k}, u_k\right)$.

4 Functional Equation

We start the analysis of the buffer behavior by defining the joint PGF of the state vector $\left(\underline{\mathbf{a}}_{1,k}, \ldots, \underline{\mathbf{a}}_{T,k}, u_k\right)$:

$$Q_k(\underline{\mathbf{x}}_1, \ldots, \underline{\mathbf{x}}_T, z) \triangleq E\left[\left(\prod_{t=1}^{T} \prod_{n=1}^{\infty} {x_{n,t}}^{a_{n,k}(t)}\right) z^{u_k}\right] \ , \tag{9}$$

where $\underline{\mathbf{x}}_t = (x_{1,t}, x_{2,t}, \ldots)$. With this definition, (7) and (8), $Q_k(\underline{\mathbf{x}}_1, \ldots, \underline{\mathbf{x}}_T, z)$ can then be obtained as

$$Q_k(\underline{\mathbf{x}}_1, \ldots, \underline{\mathbf{x}}_T, z) = E\left[\left(\prod_{t=1}^{T} \prod_{n=1}^{\infty} (x_{n,t}\, P_t(z))^{a_{n,k}(t)}\right) z^{(u_{k-1}-r_k)^+}\right] \ .$$

Next, by using (3) and (4), and by averaging over the distributions of the $c^i_{n-1,k}(t)$'s, defined in (6), we can transform the expression for $Q_k(\underline{\mathbf{x}}_1, \ldots, \underline{\mathbf{x}}_T, z)$ further into

$$Q_k(\underline{\mathbf{x}}_1,\ldots,\underline{\mathbf{x}}_T,z) = \left(\prod_{t=1}^{T} S_t(x_{1,t}\,P_t(z))\right)$$
$$\cdot\, E\left[\left(\prod_{t=1}^{T}\prod_{n=1}^{\infty} G_{n,t}(\underline{\mathbf{x}}_t,z)^{a_{n,k-1}(t)}\right) z^{(u_{k-1}-r_k)^+}\right] \ , \qquad (10)$$

where

$$G_{n,t}(\underline{\mathbf{x}}_t,z) \triangleq C_{n,t}(x_{n+1,t}\,P_t(z)) \ , \quad n \geq 1 \ , \quad 1 \leq t \leq T \ . \qquad (11)$$

In order to remove the operator $(.)^+$, we need to distinguish between the case where $r_k = 0$, the case where $r_k = 1$, $u_{k-1} > 0$ and the case where $r_k = 1$, $u_{k-1} = 0$. Moreover, we note that $u_{k-1} = 0$ implies that no packets have arrived during slot $k-1$, and hence $a_{n,k-1}(t) = 0$ ($n \geq 1$, $1 \leq t \leq T$), owing to the fact that a packet can never leave the buffer at the end of its arrival slot. With this property, the right-hand side of (10) can be further expressed in terms of the Q_{k-1}-function.

We now assume that the equilibrium condition is satisfied so that the queueing system can reach a steady state. In the steady state, $Q_k(\underline{\mathbf{x}}_1,\ldots,\underline{\mathbf{x}}_T,z)$ becomes independent of k. As a result, we then obtain the following functional equation for the steady-state PGF $Q(\underline{\mathbf{x}}_1,\ldots,\underline{\mathbf{x}}_T,z)$:

$$z\,Q(\underline{\mathbf{x}}_1,\ldots,\underline{\mathbf{x}}_T,z) = \left(\prod_{t=1}^{T} S_t(x_{1,t}\,P_t(z))\right)$$
$$\cdot\,\{\Phi(z)\,Q(\underline{\mathbf{G}}_1(\underline{\mathbf{x}}_1,z)\,,\ldots,\underline{\mathbf{G}}_T(\underline{\mathbf{x}}_T,z)\,,z) + \sigma\,(z-1)\,p_0\} \ , \quad (12)$$

where

$$\underline{\mathbf{G}}_t(\underline{\mathbf{x}}_t,z) \triangleq (G_{1,t}(\underline{\mathbf{x}}_t,z)\,,G_{2,t}(\underline{\mathbf{x}}_t,z)\,,\ldots) \ , \quad 1 \leq t \leq T \ ; \qquad (13)$$
$$\Phi(z) \triangleq \sigma + (1-\sigma)\,z \ , \qquad (14)$$

and p_0 is the steady-state probability of an empty buffer. Note that $\underline{\mathbf{G}}_t(\underline{\mathbf{x}}_t,z)$ only depends on $\underline{\mathbf{x}}_t$ and z, which is due to the fact that sessions of different types are assumed to be independent. In principle, (12) fully describes the steady-state buffer behavior. In the next sections, we will use (12) to derive several explicit results.

5 Packet Arrival Process

First, we study some further characteristics of the session-based packet arrival process. These will prove to be useful for the buffer-content analysis further in the paper. Let $a_n(t)$ denote the steady-state version of $a_{n,k}(t)$. The joint PGF $A(\underline{\mathbf{x}}_1,\ldots,\underline{\mathbf{x}}_T)$ of the $a_n(t)$'s is then given by $Q(\underline{\mathbf{x}}_1,\ldots,\underline{\mathbf{x}}_T,1)$. Putting $z=1$ in (12), we obtain

$$A(\underline{\mathbf{x}}_1,\ldots,\underline{\mathbf{x}}_T) = \left(\prod_{t=1}^{T} S_t(x_{1,t})\right) A(\underline{\mathbf{G}}_1(\underline{\mathbf{x}}_1,1)\,,\ldots,\underline{\mathbf{G}}_T(\underline{\mathbf{x}}_T,1)) \ . \qquad (15)$$

Successive applications of (15) then lead to

$$A(\underline{\mathbf{x}}_1, \ldots, \underline{\mathbf{x}}_T) = \prod_{t=1}^{T} \prod_{j=0}^{\infty} S_t\left(\sum_{i=1}^{j} \ell_t(i)\,(1 - x_{j+1,t}) + x_{j+1,t} \right) \quad . \tag{16}$$

Here we have used the definitions (11) and (13) and the following relationships:

$$C_{1,t}(C_{2,t}(\ldots C_{j,t}(x_{j+1,t})\ldots)) = \sum_{i=1}^{j} \ell_t(i)\,(1 - x_{j+1,t}) + x_{j+1,t} \ ; \tag{17}$$

$$\lim_{j \to \infty} C_{n,t}(C_{n+1,t}(\ldots C_{j,t}(x_{j+1,t})\ldots)) = 1\,, \quad n \geq 1 \ , \tag{18}$$

which can be derived from (5) and (6). The marginal PGF $A_{n,t}(z)$ of $a_n(t)$ can be obtained from (16) as

$$A_{n,t}(z) = S_t\left(\sum_{i=1}^{n-1} \ell_t(i)\,(1 - z) + z \right) \quad . \tag{19}$$

The average number of sessions of type t that are in their nth slot during an arbitrary slot is then given by

$$E[a_n(t)] = A'_{n,t}(1) = S'_t(1) \left(1 - \sum_{i=1}^{n-1} \ell_t(i) \right) \quad , \tag{20}$$

i.e., the mean number of new sessions of type t started during a slot times the probability of having a session length of at least n slots.

Let m denote the total number of packet arrivals during an arbitrary slot in the steady state. Then the PGF $M(z)$ of m can be derived from (16) as

$$M(z) = A(\underline{\mathbf{x}}_1, \ldots, \underline{\mathbf{x}}_T)\big|_{x_{n,t} = P_t(z),\, n \geq 1,\, 1 \leq t \leq T} \quad . \tag{21}$$

The mean number of packet arrivals per slot is then obtained as

$$E[m] = M'(1) = \sum_{t=1}^{T} S'_t(1)\, L'_t(1)\, P'_t(1) \ . \tag{22}$$

The equilibrium condition states that the load ρ of the queueing system has to be strictly smaller than one:

$$\rho = \frac{M'(1)}{\sigma} < 1 \ . \tag{23}$$

6 Buffer Content

In this section, we focus on the buffer content u after a slot in the steady state. Starting from the functional equation (12), we derive expressions for the mean value, the PGF and the tail distribution of the buffer content.

6.1 Mean Buffer Content

We can find the mean buffer content if we consider those values of $x_{n,t}$ ($n \geq 1$, $1 \leq t \leq T$) and z for which the arguments of the Q-functions on both sides of (12) are equal to each another, i.e., for which

$$x_{n,t} = G_{n,t}(\underline{\mathbf{x}}_t, z) \quad , \tag{24}$$

or more explicitly,

$$x_{n,t} = 1 - \pi_t(n) + \pi_t(n)\, x_{n+1,t}\, P_t(z) \quad . \tag{25}$$

These relationships can be solved for the $x_{n,t}$'s in terms of z. Denoting the solution for $x_{n,t}$ by $X_{n,t}(z)$, we obtain

$$X_{n,t}(z) = \frac{\sum\limits_{j=n}^{\infty} \ell_t(j)\, P_t(z)^{j-n}}{1 - \sum\limits_{j=1}^{n-1} \ell_t(j)} \ , \quad n \geq 1 \, , \quad 1 \leq t \leq T \ . \tag{26}$$

Note in particular that

$$X_{1,t}(z) = \frac{L_t(P_t(z))}{P_t(z)} \tag{27}$$

and

$$X_{n,t}(1) = 1 \, , \quad n \geq 1 \ . \tag{28}$$

Choosing $x_{n,t} = X_{n,t}(z)$ ($n \geq 1$, $1 \leq t \leq T$) in (12), we then get a linear equation for the function $Q(\underline{\mathbf{X}}_1(z), \ldots, \underline{\mathbf{X}}_T(z), z)$, which has the following solution:

$$Q(\underline{\mathbf{X}}_1(z), \ldots, \underline{\mathbf{X}}_T(z), z) = \frac{\sigma\,(z-1)\, p_0\, S(z)}{z - S(z)\, \Phi(z)} \quad , \tag{29}$$

where $\underline{\mathbf{X}}_t(z) = (X_{1,t}(z), X_{2,t}(z), \ldots)$ and the function $S(z)$ is defined as

$$S(z) \triangleq \prod_{t=1}^{T} S_t(L_t(P_t(z))) \quad . \tag{30}$$

The probability p_0 in (29) can be calculated from the normalization condition $Q(\underline{\mathbf{X}}_1(z), \ldots, \underline{\mathbf{X}}_T(z), z)|_{z=1} = 1$. By using de l'Hôpital's rule, we obtain

$$p_0 = 1 - \rho \quad , \tag{31}$$

where ρ is the load of the system.

In order to obtain the mean buffer content, we calculate the first derivative of (29) with respect to z in the point $z = 1$. This leads to

$$\sum_{t=1}^{T} \sum_{n=1}^{\infty} E[a_n(t)]\ X'_{n,t}(1) + E[u] = \frac{\mathrm{d}}{\mathrm{d}z} \left\{ \frac{\sigma\,(z-1)\, p_0\, S(z)}{z - S(z)\, \Phi(z)} \right\} \Bigg|_{z=1} \quad , \tag{32}$$

where $E[a_n(t)]$ is given by (20). With (26)–(28), after some further calculations, we finally find the following explicit expression for the mean buffer content:

$$\begin{aligned} E[u] = &-\sum_{t=1}^{T} \frac{1}{2}\, S_t'(1)\, P_t'(1)\, \left[\sigma_{L,t}^2 - L_t'(1) + L_t'(1)^2\right] \\ &+ \frac{1}{2\sigma\,(1-\rho)} \left\{ \rho\,\sigma\,(2-\rho\,\sigma) + \sum_{t=1}^{T} \left(\sigma_{S,t}^2\, L_t'(1)^2\, P_t'(1)^2 + \sigma_{L,t}^2\, S_t'(1)\, P_t'(1)^2\right) \right. \\ &\left. + \sum_{t=1}^{T} \left(\sigma_{P,t}^2 - P_t'(1)\right)\, S_t'(1)\, L_t'(1) \right\} \ , \end{aligned} \tag{33}$$

where $\sigma_{L,t}^2$, $\sigma_{S,t}^2$ and $\sigma_{P,t}^2$ are the variances of the session length, the number of new sessions and the session bandwidth respectively, for sessions of type t.

6.2 PGF of the Buffer Content

The PGF $U(z)$ of u is given by $Q(1,\ldots,1,z)$. Successive applications of (12) then allow to express $U(z)$ in terms of the function $Q(\underline{\mathbf{X}}_1(z),\ldots,\underline{\mathbf{X}}_T(z),z)$, given in (29). As a result, we obtain

$$\begin{aligned} U(z) = {}& Q(\underline{\mathbf{X}}_1(z),\ldots,\underline{\mathbf{X}}_T(z),z) \left(\prod_{j=1}^{\infty} \frac{\Phi(z)}{z}\, g_j(z)\right) \\ &+ \sigma\,(z-1)\,p_0 \sum_{k=1}^{\infty} \frac{1}{\Phi(z)} \left(\prod_{j=1}^{k} \frac{\Phi(z)}{z}\, g_j(z)\right) \ , \end{aligned} \tag{34}$$

where we have used the property that

$$\lim_{j\to\infty} C_{n,t}(P_t(z)\, C_{n+1,t}(\ldots P_t(z)\, C_{j,t}(P_t(z))\ldots)) = X_{n,t}(z) \ , \tag{35}$$

$n \geq 1$, $1 \leq t \leq T$, as can be shown from (5), (6) and (26). The function $g_j(z)$ in (34) is defined as

$$g_j(z) \triangleq \prod_{t=1}^{T} S_t(P_t(z)\, C_{1,t}(P_t(z)\, C_{2,t}(\ldots P_t(z)\, C_{j-1,t}(P_t(z))\ldots))) \ , \tag{36}$$

and can be further calculated with (5) and (6) as

$$g_j(z) = \prod_{t=1}^{T} S_t\left(P_t(z)^j + \sum_{i=1}^{j-1} \ell_t(i)\, \left(P_t(z)^i - P_t(z)^j\right)\right) \ . \tag{37}$$

Combination of (29) and (34) then leads to the following explicit expression for the PGF $U(z)$:

$$U(z) = \frac{\sigma\,(z-1)\,p_0\, H(z)}{z - S(z)\,\Phi(z)} \ , \tag{38}$$

where $H(z)$ is given by

$$H(z) = S(z)\left(\prod_{j=1}^{\infty} \frac{\Phi(z)}{z}\, g_j(z)\right) + [z - S(z)\,\Phi(z)] \sum_{k=1}^{\infty} \frac{1}{\Phi(z)} \left(\prod_{j=1}^{k} \frac{\Phi(z)}{z}\, g_j(z)\right) . \tag{39}$$

6.3 Tail Distribution of the Buffer Content

In order to derive an expression for the tail distribution of the buffer content, we will use an approximation technique described in [4]. Specifically, from the inversion formula for z-transforms, it follows that the PMF Prob$[u = i]$ of u can be expressed as a weighted sum of negative ith powers of the poles of $U(z)$. As the modulus of all these poles is larger than 1, since $U(z)$ is a PGF, it is clear that for large values of i, Prob$[u = i]$ is dominated by the contribution of the pole of $U(z)$ having the smallest modulus. Let z_0 denote this dominant pole of $U(z)$. The pole z_0 must necessarily be real and positive in order to ensure that the tail distribution is nonnegative anywhere. From (38), it follows that z_0 is a real root of $z - S(z)\,\Phi(z) = 0$. The PMF Prob$[u = i]$ can then be approximated by the following geometric form:

$$\text{Prob}[u = i] \approx -\frac{\theta_0}{z_0}\left(\frac{1}{z_0}\right)^i , \tag{40}$$

for i sufficiently large, where the constant θ_0 is the residue of $U(z)$ in the point $z = z_0$. This residue can be calculated from (30), (38) and (39) as

$$\begin{aligned} \theta_0 &= \lim_{z \to z_0} (z - z_0)\, U(z) = \frac{\sigma\,(z_0 - 1)\, p_0\, \Phi(z_0)\, H(z_0)}{\sigma - S'(z_0)\, \Phi(z_0)^2} \\ &= \frac{\sigma\, z_0\,(z_0 - 1)\, p_0 \left(\prod\limits_{j=1}^{\infty} \frac{\Phi(z_0)}{z_0}\, g_j(z_0)\right)}{\sigma - \sum\limits_{t=1}^{T} \frac{S_t'(L_t(P_t(z_0)))\, L_t'(P_t(z_0))\, P_t'(z_0)\, z_0\, \Phi(z_0)}{S_t(L_t(P_t(z_0)))}} . \end{aligned} \tag{41}$$

Notice the infinite product in the above expression for θ_0. We know however from (37) that $\lim\limits_{j \to \infty} g_j(z) = S(z)$. Due to the definition of z_0, we moreover have that $S(z_0)\,\Phi(z_0) = z_0$. Therefore, we see that

$$\lim_{j \to \infty} \frac{\Phi(z_0)}{z_0}\, g_j(z_0) = 1 , \tag{42}$$

i.e., the factors of the infinite product in (41) go to 1, as j goes to infinity. Hence, we can calculate the residue θ_0 numerically up to any desired precision by taking the product over a sufficiently large number of factors.

A quantity of considerable interest is the probability that the buffer content exceeds a certain threshold U. Indeed, this quantity is often used to approximate the packet loss ratio, i.e., the fraction of the arriving packets that is lost upon arrival because of buffer overflow, in a buffer model with a finite storage capacity (for U waiting packets), see e.g. [17]. From (40), we get

$$\mathrm{Prob}[u > U] \approx -\frac{\theta_0}{z_0 - 1}\left(\frac{1}{z_0}\right)^{U+1} , \quad \text{for large } U \ . \tag{43}$$

7 Packet Delay

In this section, we assume a FCFS queueing discipline for packets. We define the delay of a packet as the time interval (expressed in slots) between the end of the packet's arrival slot and the end of the slot during which the packet is transmitted.

In [9], it has been shown that for any discrete-time single-server infinite-capacity queueing system with an FCFS queueing discipline and geometrically distributed packet transmission times (with parameter $1 - \sigma$), regardless of the nature of the arrival process, the following relationship exists between the PGF $D(z)$ of the delay d of an arbitrary packet that arrives in the buffer during a slot in the steady state and the PGF $U(z)$ of the buffer content u after an arbitrary slot in the steady state:

$$D(z) = \frac{U(B(z)) - p_0}{\rho} \ , \tag{44}$$

where $B(z) = \frac{\sigma\, z}{1-(1-\sigma)\, z}$ is the PGF of the geometric transmission times.

Since the effective packet transmission times in our model have a geometric distribution, the above relationship is also valid here. It allows us to express the mean value and the tail distribution of the packet delay in terms of the previously derived mean value and tail distribution of the buffer content. In particular, the mean packet delay follows from (44) as

$$E[d] = D'(1) = \frac{U'(1)}{\rho\,\sigma} = \frac{E[u]}{E[m]} \ , \tag{45}$$

in accordance with Little's theorem. For i sufficiently large, the PMF of the packet delay can be approximated as

$$\mathrm{Prob}[d = i] \approx -\frac{\theta_D}{z_D}\left(\frac{1}{z_D}\right)^{i} \ . \tag{46}$$

From (44), it follows that the dominant pole z_D of $D(z)$ is related to the dominant pole z_0 of $U(z)$ as $z_0 = B(z_D)$, or equivalently,

$$z_D = \frac{z_0}{\sigma + (1-\sigma)\, z_0} \ . \tag{47}$$

The residue θ_D of $D(z)$ in the point $z = z_D$ can be calculated from (44) as

$$\theta_D = \lim_{z \to z_D} (z - z_D)\, D(z) = \frac{\theta_0}{\rho\, B'(z_D)} = \frac{\theta_0\, z_D\, (z_D - 1)}{\rho\, z_0\, (z_0 - 1)} \ . \tag{48}$$

8 Numerical Results and Discussion

In Fig. 3, we assume a single session type ($T = 1$) and show the mean buffer content $E[u]$ as a function of the load ρ. The mean length of the sessions is equal to 100 slots for all of the shown curves, but the distribution of the session lengths is different for each curve. That is, the session lengths respectively are constant, uniform between 1 and 201, negative binomial with two stages, geometric and mixed geometric. For the latter, the mixed geometric distribution of the session lengths has two weighted parallel phases, i.e., the PGF and the mean value are given by

$$L_1(z) = w \, \frac{\gamma_{1,1} \, z}{1 - (1 - \gamma_{1,1}) \, z} + (1 - w) \, \frac{\gamma_{1,2} \, z}{1 - (1 - \gamma_{1,2}) \, z} \;\; ; \tag{49}$$

$$L_1'(1) = w \, \frac{1}{\gamma_{1,1}} + (1 - w) \, \frac{1}{\gamma_{1,2}} \;\; . \tag{50}$$

The mean $1/\gamma_{1,1}$ of the first phase is taken to be 50 slots, while the second phase has a mean of 200 slots. The weight w is chosen in order to ensure that $L_1'(1) = 100$. For all curves, the bandwidth of the sessions is fixed at 2 packets per slot, i.e., $P_1(z) = z^2$ and in each slot a new session starts with probability 1/4000. As in all further examples, the load is increased on the horizontal axis by increasing the mean effective transmission time $1/\sigma$ of the packets. The plot clearly illustrates that the mean buffer content $E[u]$ depends not only on the first moment $L_1'(1)$ of the session-length distribution but on the second-order moment as well. Specifically, (33) predicts a linear impact on the mean buffer content of the variance $\sigma^2_{L,1}$ of the session lengths, which for the considered session-length distributions is 0, 3300, 4999.5, 9900 and 14900 respectively.

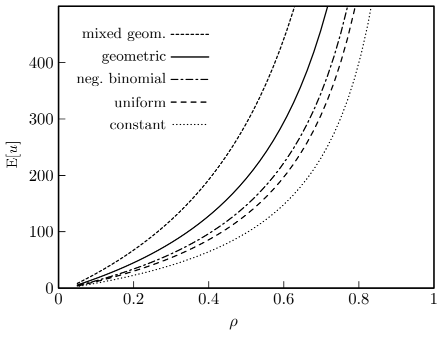

Fig. 3. Mean buffer content as a function of the load ρ in the homogeneous case ($T = 1$) for different session-length distributions $L_1(z)$ with a mean of 100 slots. Bandwidth and session starts have PGF $P_1(z) = z^2$ and $S_1(z) = \frac{3999+z}{4000}$.

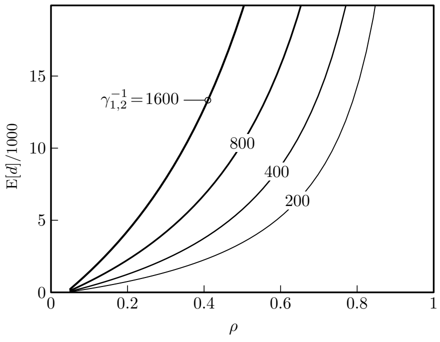

Fig. 4. Mean packet delay as a function of the load ρ in the homogeneous case ($T = 1$) with $P_1(z) = \frac{5}{6}z\,(1 - \frac{z}{6})^{-1}$ and $S_1(z) = \frac{2399+z}{2400}$. The session lengths are mixed geometric with mean 100 and first phase mean $1/\gamma_{1,1} = 50$.

This effect is illustrated further in Fig. 4, where the mean packet delay $E[d]$ is shown as a function of the load ρ. Again, the distributions of the session starts and the session bandwidth are the same for all curves, as well as the mean session length $L_1'(1)$ which is 100 slots. The distribution $L_1(z)$ is chosen to be mixed geometric of the form (49) with the first phase mean equal to 50 slots. The plot shows the impact on $E[d]$ if the second phase mean of the session-length distribution is increased, i.e., $1/\gamma_{1,2} = 200$, 400, 800, 1600. For the same configuration and $\rho = 0.8$, the tail distribution of the buffer content $\text{Prob}[u = i]$ is shown in Fig. 5, together with the corresponding mixed geometric distributions of the session lengths.

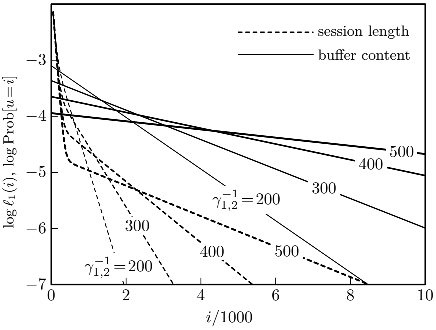

Fig. 5. Logarithmic plot of the mixed geometric session-length distribution and the corresponding tail distribution of the buffer content for load $\rho = 0.8$

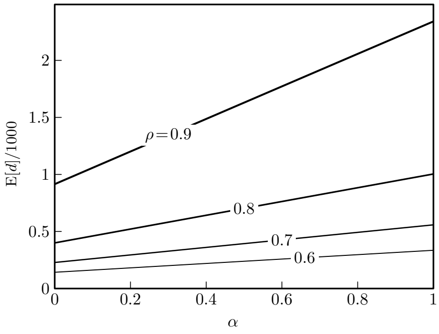

Fig. 6. Mean packet delay as a function of the session mix α. On the left, all sessions are of type 1, while all sessions are of type 2 on the right.

In Fig. 6, we assume heterogeneous traffic with two types of sessions, i.e., $T = 2$. The sessions of type 1 have a constant length of 25 slots, shifted geometric bandwidth $P_1(z) = \frac{z}{2-z}$ with a mean of 2 packets per slot and a Bernoulli session-start distribution with mean $\frac{1-\alpha}{200}$. The sessions of type 2 have a length that is uniformly distributed between 1 and 201, a bandwidth of exactly one packet per slot and a Poisson start distribution with mean $\frac{\alpha}{400}$. The session mix α $(0 \leq \alpha \leq 1)$ indicates the fraction of the load due to sessions of type 1. If $\alpha = 0$, all arrivals are of type 1, while if $\alpha = 1$, there are only sessions of type 2. In Fig. 6, the mean packet delay is shown for load $\rho = 0.6$, 0.7, 0.8, 0.9. We observe a linear dependence of $E[d]$ (and thus also $E[u]$) on the session mix, which is predicted by (33).

In Fig. 7, we consider sessions of T different types and show $E[d]$ as a function of T in case the load is $\rho = 0.6$, 0.7, 0.8, 0.9. For a certain number of types T, the sessions of type t $(1 \leq t \leq T)$ receive an equal share ρ/T of the total load. The sessions of all types have mean length $L_t'(1) = 100$, shifted geometric bandwidth with a mean of $P_t'(1) = 2$ packets per slot and a Bernoulli session-start distribution with mean $\frac{\rho\,\sigma}{100 \cdot 2 \cdot T}$. Also, the session-length distribution for all types is mixed geometric of the form (49), with first phase mean $1/\gamma_{t,1} = 50$ slots. The tail of the session length however is chosen to be larger for higher types: the second phase mean of type t is $1/\gamma_{t,2} = 100 + 2^t$. Again, we observe a clear impact of the variance of the session lengths on the performance of the system. For $T = 5$, 10 and load $\rho = 0.8$, the tail distributions of the buffer content Prob$[u = i]$ and the packet delay Prob$[d = i]$ are shown in Fig. 8.

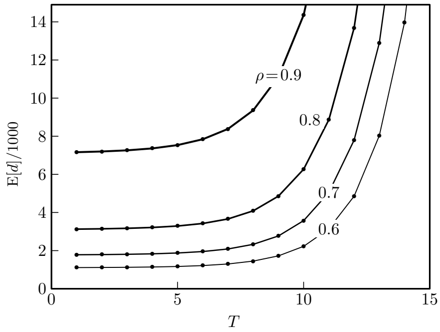

Fig. 7. Mean packet delay as a function of the number of session types T. For type t, the session-length distribution is mixed geometric with mean 100. The phase means are $1/\gamma_{t,1} = 50$ and $1/\gamma_{t,2} = 100 + 2^t$.

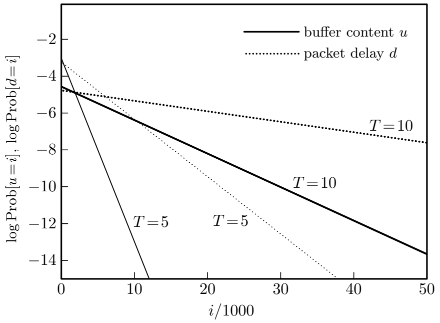

Fig. 8. Logarithmic plot of the tail distributions of the buffer content and the packet delay for $T = 5, 10$ heterogeneous session types and $\rho = 0.8$

9 Conclusions

We have presented an analytical technique for the performance evaluation of a buffer with session-based arrival streams. Differently from previous work, there are T session types and for each type, the session lengths may have a general distribution. Expressions have been obtained for the PGFs, the mean values and the tail distributions of the buffer content and the packet delay. By means of some numerical examples, the impact of the session-based packet arrival process on the performance has been investigated. The results indicate that the buffer behavior strongly depends on the session-length characteristics.

As future work, we plan to study the delay of a session. Also, we intend to apply the model with general session-based arrivals to evaluate the performance of a web server by fitting the model parameters to traces of web traffic.

References

1. Altman, E., Jeanmarie, A.: The distribution of delays of dispersed messages in an M/M/1 queue. In: 14th IEEE Conference on Computer Communications, INFOCOM 1995, Boston (1995)

2. Arlitt, M., Williamson, C.: Internet Web Servers: Workload Characterization and Performance Implications. IEEE ACM Transactions on Networking 5, 631–645 (1997)
3. Bruneel, H.: Packet Delay and Queue Length for Statistical Multiplexers with Low-Speed Access Lines. Computer Networks and ISDN Systems 25, 1267–1277 (1993)
4. Bruneel, H., Kim, B.G.: Discrete-Time Models for Communication Systems Including ATM. Kluwer Academic Publishers, Boston (1993)
5. Choi, B.D., Choi, D.I., Lee, Y., Sung, D.K.: Priority Queueing System with Fixed-Length Packet-Train Arrivals. IEE Proceedings-Communications 145, 331–336 (1998)
6. Cidon, I., Khamisy, A., Sidi, M.: Delay, Jitter and Threshold Crossing in ATM Systems with Dispersed Messages. Performance Evaluation 29, 85–104 (1997)
7. Daigle, J.: Message Delays at Packet-Switching Nodes Serving Multiple Classes. IEEE Transactions on Communications 38, 447–455 (1990)
8. Elsayed, K., Perros, H.: The Superposition of Discrete-Time Markov Renewal Processes with an Application to Statistical Multiplexing of Bursty Traffic Sources. Applied Mathematics and Computation 115, 43–62 (2000)
9. Gao, P., Wittevrongel, S., Bruneel, H.: Delay against System Contents in Discrete-Time G/Geom/c Queue. Electronics Letters 39, 1290–1292 (2003)
10. Hoflack, L., De Vuyst, S., Wittevrongel, S., Bruneel, H.: System Content and Packet Delay in Discrete-Time Queues with Session-Based Arrivals. In: 5th International Conference on Information Technology, ITNG 2008, Las Vegas (2008)
11. Hoflack, L., De Vuyst, S., Wittevrongel, S., Bruneel, H.: Modeling Web Server Traffic with Session-Based Arrival Streams. In: Al-Begain, K., Heindl, A., Telek, M. (eds.) ASMTA 2008. LNCS, vol. 5055, pp. 47–60. Springer, Heidelberg (2008)
12. Kamoun, F.: Performance Analysis of a Discrete-Time Queuing System with a Correlated Train Arrival Process. Performance Evaluation 63, 315–340 (2006)
13. Takagi, H.: Queueing Analysis – A Foundation of Performance Evaluation. Discrete-Time Systems, vol. 3. North-Holland, Amsterdam (1993)
14. Walraevens, J., Wittevrongel, S., Bruneel, H.: A Discrete-Time Priority Queue with Train Arrivals. Stochastic Models 23, 489–512 (2007)
15. Wittevrongel, S.: Discrete-Time Buffers with Variable-Length Train Arrivals. Electronics Letters 34, 1719–1721 (1998)
16. Wittevrongel, S., Bruneel, H.: Correlation Effects in ATM Queues due to Data Format Conversions. Performance Evaluation 32, 35–56 (1998)
17. Woodside, C., Ho, E.: Engineering Calculation of Overflow Probabilities in Buffers with Markov-Interrupted Service. IEEE Transactions on Communications 35, 1272–1277 (1987)
18. Xiong, Y., Bruneel, H.: Buffer Behaviour of Statistical Multiplexers with Correlated Train Arrivals. International Journal of Electronics and Communications 51, 178–186 (1997)