Available at
www.ElsevierComputerScience.com
POWERED BY SCIENCE DIRECT®
Computer Networks 44 (2004) 335–352
COMPUTER NETWORKS
www.elsevier.com/locate/comnet

Modeling IP traffic: joint characterization of packet arrivals and packet sizes using BMAPs

Paulo Salvador [a], António Pacheco [b,*], Rui Valadas [a]

[a] Institute of Telecommunications Aveiro, University of Aveiro, Campus de Santiago, 3810-193 Aveiro, Portugal
[b] Department of Mathematics and Centre for Mathematics and its Applications, Instituto Superior Técnico, Technical University of Lisbon, Av. Rovisco Pais 1, 1049-001 Lisboa, Portugal

Received 25 October 2002; received in revised form 20 June 2003; accepted 9 October 2003

Responsible Editor: H.L. Truong

Abstract

This paper proposes a traffic model and a parameter fitting procedure that are capable of achieving accurate prediction of the queuing behavior for IP traffic exhibiting long-range dependence. The modeling process is a discrete-time batch Markovian arrival process (dBMAP) that jointly characterizes the packet arrival process and the packet size distribution. In the proposed dBMAP, packet arrivals occur according to a discrete-time Markov modulated Poisson process (dMMPP) and each arrival is characterized by a packet size with a general distribution that may depend on the phase of the dMMPP. The fitting procedure is designed to provide a close match of both the autocovariance and the marginal distribution of the packet arrival process, using a dMMPP; a packet size distribution is fitted individually to each state of the dMMPP. A major feature of the procedure is that the number of states of the fitted dBMAP is not fixed a priori; it is determined as part of the procedure itself. In this way, the procedure allows establishing a compromise between the accuracy of the fitting and the number of parameters, while maintaining a low computational complexity.

We apply the inference procedure to several traffic traces exhibiting long-range dependence. Very good results were obtained since the fitted dBMAPs match closely the autocovariance, the marginal distribution and the queuing behavior of the measured traces. Our results also show that ignoring the packet size distribution and its correlation with the packet arrival process can lead to large errors in terms of queuing behavior.

Keywords: IP traffic modeling; Packet size distribution; BMAP

1. Introduction

Traffic characterization and modeling is a crucial activity towards an efficient dimensioning and resource management of IP networks. Accurate modeling of IP traffic requires matching closely not only the packet arrival process but also the packet

* Corresponding author.
E-mail addresses: salvador@av.it.pt (P. Salvador), apacheco@math.ist.utl.pt (A. Pacheco), rv@det.ua.pt (R. Valadas).

1389-1286/$ - see front matter © 2003 Elsevier B.V. All rights reserved.
doi:10.1016/j.comnet.2003.10.004

size distribution. Surprisingly, while the arrival process has received considerable attention [1–10], very few works have addressed the packet size distribution and, more so, its joint characterization with the arrival process [11,12]. However, as it will be made clear in this paper, a joint characterization is required for accurate prediction of the queuing behavior (i.e., the packet loss ratio or average packet delay observed in a queuing system).

The queuing behavior is one of the most important criteria to assess the suitability of traffic models (and associated parameter fitting procedures), since it addresses the effect of traffic on network performance. The analysis consists in comparing the curves of packet loss ratio (or average packet delay) versus buffer size, obtained with the measured traces (through trace-driven simulation) and with the inferred traffic model (using again trace-driven simulation or numerical computation of these performance measures whenever possible). When dealing with models that characterize only the arrival process, it is common practice to assume that the packet size is fixed and equal to the average packet size of the measured trace. This may lead to large errors when the packets have variable size, such as in IP traffic.

Accurate prediction of the queuing behavior also requires detailed modeling of the first and second order statistics of the packet arrival process. This motivated the inference procedures for circulant Markov modulated Poisson processes (MMPPs) presented in [2] and for other special cases of MMPPs presented in [10]. Also in this direction, the work in [6] discusses the limitations of using only the mean and the autocorrelation function as statistical descriptors of the input traffic, for the purpose of analyzing queuing performance. The authors show that the mean queue length can vary substantially when the parameters of the input process are varied, subject to the same mean arrival rate and autocorrelation function. Thus, in general, accurate prediction of the queuing behavior requires detailed modeling of the marginal distribution, and not just of the mean arrival rate, in addition to the autocovariance modeling.

Since the work by Leland et al. [13] several studies have shown that network traffic may exhibit properties of self-similarity [13–19]. More recently, a multifractal behavior, which is typically associated with networking mechanisms operating on small time scales, was discovered in several traces of Internet WAN traffic [20–24]. In general, these characteristics include the long-range dependence (LRD) property and can have a significant impact on the network performance. However, as pointed out in [25–27] matching the LRD is only required within the time scales of interest to the system under study. For example, in order to analyze queuing behavior, the selected traffic model needs only to capture the correlation structure of the source up to the so-called critical time scale or correlation horizon, which is directly related to the maximum buffer size. One of the consequences of this result is that more traditional traffic models, such as MMPPs, can still be used to model traffic exhibiting LRD.

We introduce a batch Markovian arrival process (BMAP) (see, e.g., [28–30]), that jointly characterizes the packet arrival process and the packet size distribution, and we develop a parameter fitting procedure that is capable of achieving accurate prediction of queuing behavior for IP traffic exhibiting LRD behavior. We consider the discrete-time version of the BMAP process, denoted hereby by dBMAP. In this dBMAP, packet arrivals occur according to a discrete-time Markov modulated Poisson process (dMMPP) and each arrival is further characterized by a packet size with a general distribution that may depend on the phase of the dMMPP. The fitting procedure starts by matching both the autocovariance and the marginal distribution of the packet arrival process using a dMMPP; it then matches the packet size distribution in each state of the associated dMMPP in order to fully characterize the dBMAP. This allows having a packet size distribution closely related to the packet arrival process, and is in contrast with the approach followed by [11] where the packet size distribution is fitted prior to the matching of the packet arrival rates. In addition, the number of states of the associated dMMPP is not fixed a priori; it is determined as part of the procedure. In this way, the procedure allows establishing a compromise between the accuracy of the fitting and the number of parameters, while maintaining a low computational complexity.

The associated dMMPP fits the first and second order statistics of the packet arrival process. Specifically, we work with the counts of the packet arrival process, i.e., the number of packet arrivals in a pre-defined sampling interval (also called time slot). Matching simultaneously the autocovariance and marginal distribution of the packet arrival process is a difficult task since every dMMPP parameter has an influence on both characteristics. With the purpose of achieving some degree of decoupling when matching these two statistics, we construct the dMMPP as a superposition of L 2-dMMPPs and one M-dMMPP (where k-dMMPP represents a dMMPP with k states); L and M are determined as part of the fitting procedure. We will denote the resulting process as $M2^L$-dMMPP. We use the L 2-dMMPPs to match the autocovariance, where each 2-dMMPP models a characteristic time constant of this function, and the M-dMMPP to match the marginal distribution (taking into account the contribution of the L 2-dMMPPs). While, in general, the addition of a dMMPP to a process changes both the marginal distribution and the autocovariance function, we devised a procedure where this is done in a controlled way. Specifically, the procedure assures that the M-dMMPP has null autocovariance, thus forcing the autocovariance of the combined process to equal that of the superposition of the L 2-dMMPPs. Finally, the dBMAP is constructed by associating a packet size distribution to each state of the associated dMMPP.

We apply the fitting procedure to several traffic traces exhibiting LRD. The LRD characteristics are analyzed using the wavelet based estimator of [31]. Results show that the dBMAPs obtained through the fitting procedure are capable of modeling the LRD behavior present in data, and closely match the first and second order statistics of the packet arrival process, the packet size distribution and the queuing behavior.

This paper is organized as follows. In Section 2 we give some background on discrete-time BMAPs and define the specific dBMAP that we propose. In Section 3 we describe the fitting procedure. A study concerning the importance of the packet size fitting is carried out in Section 4. In Section 5 we present the results of applying the fitting procedure to measured traffic traces. Finally, in Section 6 we conclude the paper.

The main contributions of this paper are the following. First, we introduce a traffic model and a fitting procedure that provide a detailed joint characterization of the packet size distribution and the first and second order statistics of the packet arrival process; the process is both accurate and numerically efficient. Second, we show that ignoring the packet size distribution and its correlation with the packet arrival process can lead to large errors in terms of queuing behavior. Finally, we show that the proposed traffic model and fitting procedure are capable of matching closely traffic traces with LRD characteristics.

2. Background

The (continuous-time) BMAP was introduced by Lucantoni [28] as a generalization of the (simple) Markovian arrival process introduced in [32]. The BMAP is a very general arrival process that, as remarked by Pacheco and Prabhu [30], achieves the full generality of the univariate Markov additive processes of arrivals when the associated Markov component has finite state space. In addition, it enjoys the many good properties of Markov additive processes as, e.g., being closed for the superposition of independent processes. Moreover, Asmussen and Koole [33] proved that any marked point process whose marks are real-valued can be obtained as the weak limit, i.e., limit in distribution, of a sequence of Markovian arrival processes. The fitting of the BMAP to IP traffic is used in [11] with relative success. For a history of the BMAP and its applications, and a very extensive list of references, see [29].

The discrete-time version of the BMAP (which we denote by dBMAP, usually denoted by D-BMAP) was proposed by Blondia and Casals [34] and has received a great deal of attention (see, e.g., [35–39] and references therein), although not as many as its continuous-time counterpart. The dBMAP also belongs to the class of univariate Markov additive processes and, thus, similarly enjoys the many good properties of this class of processes, some of which are addressed in Section 2.1. Among other

applications, the dBMAP has been used to model bursty sources and in speech recognition. Moreover, it has been shown that LRD may result from the superposition of dBMAPs [38].

A very important particular case of the dBMAP is the dMMPP [10]. In the dMMPP the number of arrivals at each instant has a Poisson distribution with parameter depending on the phase (of the modulator discrete-time Markov chain).

For the modeling of IP traffic we propose to use a dBMAP such that: arrivals (of packets) occur according to a dMMPP; and, each arrival is characterized by a batch whose size has a general distribution that may depend on the phase of the dMMPP describing the packet arrival process. This is in contrast with [11], where the packet size is fitted in advance, in part, to avoid a prohibitive number of parameters to be estimated. However we do not have such a problem since the packet size distribution for each phase of the dMMPP that models the arrival of packets is fitted only after the parameters of the dMMPP have been adjusted. Our procedure is especially well suited to accurately fit the packet size distribution.

2.1. Characterization of the discrete-time batch Markovian arrival process used

The dBMAP may be regarded as an Markov random walk whose additive component takes values on the non-negative integers, $\mathbb{N}_0$. Thus, we say that a Markov chain $(Y,J) = \{(Y_k, J_k), k \in \mathbb{N}_0\}$ on the state space $\mathbb{N}_0 \times S$ is a dBMAP if

$$P(Y_{k+1} = m, J_{k+1} = j \mid Y_k = n, J_k = i) = \begin{cases} 0, & m < n, \\ p_{ij} q_{ij}(m-n), & m \geqslant n, \end{cases} \quad (1)$$

where $\mathbf{P} = (p_{ij})_{i,j\in S}$ is a stochastic matrix and, for each pair $(i,j) \in S^2$, $q_{ij} = \{q_{ij}(n), n \in \mathbb{N}_0\}$ is a probability function over $\mathbb{N}_0$, and we let $\mathbf{Q}(n) = (q_{ij}(n))_{i,j\in S}$. This implies, in particular that J is a Markov chain, called the *Markov component* or *phase* of (Y,J) and S is the set of modulating states or the phase set. When the dBMAP (Y,J) is used to model an arrival process, Y_k may be interpreted as the total number of arrivals until instant k.

Following [30], we will say that the dBMAP (Y,J) described by (1) has parametrization $(\mathbf{P}, \{\mathbf{Q}(n), n \in \mathbb{N}\})$. [1]

In the paper, we will refer to two dBMAPs, (X,J) and (Y,J), where X_n (Y_n) will represent the total number of packets (bytes) that arrive until instant n. Similarly, J will represent the state of a non-observable environment that affects both the number of packets that arrive at each instant and the corresponding packet sizes (in bytes).

An important particular case of the dBMAP is the dMMPP. We say that the process (X,J) on the state space $\mathbb{N}_0 \times S$ is a dMMPP with parameters $(\mathbf{P}, \mathbf{\Lambda})$, where $\mathbf{P} = (p_{ij})_{i,j\in S}$ is a stochastic matrix and $\mathbf{\Lambda} = (\lambda_{ij})_{i,j\in S} = (\lambda_i 1_{\{i=j\}})_{i,j\in S}$ is a diagonal matrix with non-negative entries (i.e., $\lambda_i \geqslant 0$, $i \in S$), if it is a dBMAP with parametrization $(\mathbf{P}, \{\mathbf{Q}(n), n \in \mathbb{N}\})$, where

$$q_{ij}(n) = \mathrm{e}^{-\lambda_j} \frac{\lambda_j^n}{n!} \quad (2)$$

for $i,j \in S$ and $n \in \mathbb{N}$; i.e., $q_{ij} = \{q_{ij}(n), n \in \mathbb{N}_0\}$ is the probability function of a Poisson random variable with mean λ_j. Thus a dMMPP is a dBMAP for which the number of arrivals in a given instant of time is only a function of the current phase of the dBMAP and when the process is in phase j the number of arrivals at an instant has a Poisson distribution with mean λ_j; the parameter λ_j may be null, in which case no arrivals occur in phase j.

In our analysis a dMMPP will be used to model the packet arrival process. We will consider additionally that the packets have independent sizes, with the size of packets arriving in phase i having probability function $q_i = \{q_i(n), n \in \mathbb{N}\}$. Accordingly, if we let (X,J) denote the dMMPP, on the state space $\mathbb{N}_0 \times S$ and having parametrization $(\mathbf{P}, \mathbf{\Lambda})$, that models the packet arrival process, then the byte arrival process (Y,J) is a dBMAP, on the state space $\mathbb{N}_0 \times S$, satisfying (1) with

[1] We note that this parametrization does not correspond to the most common parametrization used for dBMAPs, which is of the form $\{\mathbf{D}_n, n \in \mathbb{N}_0\}$, where the matrices $\mathbf{D}_n$ are finite non-negative square matrices such that $\sum_{n=0}^{\infty} \mathbf{D}_n$ is a stochastic matrix [34]. However, the parametrization we will use is more adequate for the dBMAPs we consider in the paper.

$$q_{ij}(n) = \sum_{l=0}^{+\infty} \mathrm{e}^{-\lambda_j} \frac{\lambda_j^l}{l!} q_j^{(l)}(n) \tag{3}$$

for $i, j \in S$ and $n \in \mathbb{N}_0$, where $q_j^{(l)}$ denotes de convolution of order l of q_j. Thus, (Y,J) is a dBMAP on the state space $\mathbb{N}_0 \times S$, such that, for $n, m \in \mathbb{N}_0$,

$$P(Y_{k+1} = m+n, J_{k+1} = j \mid Y_k = m, J_k = i) = p_{ij} \sum_{l=0}^{+\infty} \mathrm{e}^{-\lambda_j} \frac{\lambda_j^l}{l!} q_j^{(l)}(n), \tag{4}$$

which we express by saying that (Y,J) has *type-II parametrization* $(\mathbf{P}, \mathbf{\Lambda}, \{q_i, i \in S\})$.

The alternative type-II parametrization is the most convenient for the subclass of dBMAPs that we will consider in the paper, since $(\mathbf{P}, \mathbf{\Lambda}, \{q_i, i \in S\})$ contains: (i) the parameters $(\mathbf{P}, \mathbf{\Lambda})$ of the associated dMMPP modeling the packet arrival process and (ii) the distribution of the size of packets arriving in phase i, q_i, for $i \in S$. However, as our construction highlights, the type-II parametrization is only valid for dBMAPs satisfying (4). Given a dBMAP (Y,J) with phase set S and type-II parametrization $(\mathbf{P}, \mathbf{\Lambda}, \{q_i, i \in S\})$ we write

$$(Y,J) \sim \mathrm{dBMAP}_S(\mathbf{P}, \mathbf{\Lambda}, \{q_i\}). \tag{5}$$

Similarly if (X,J) is a dMMPP with phase set S and having parameters $(\mathbf{P}, \mathbf{\Lambda})$, we write

$$(X,J) \sim \mathrm{dMMPP}_S(\mathbf{P}, \mathbf{\Lambda}). \tag{6}$$

If S has cardinality r, we say that (Y,J) $((X,J))$ is a dBMAP (dMMPP) of order r, r-dBMAP (r-dMMPP). When, in particular, $S = \{1, 2, \ldots, r\}$ for some $r \in \mathbb{N}$, then

$$\mathbf{P} = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1r} \\ p_{21} & p_{22} & \cdots & p_{2r} \\ \cdots & \cdots & \cdots & \cdots \\ p_{r1} & p_{r2} & \cdots & p_{rr} \end{bmatrix}, \quad \mathbf{\Lambda} = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & \lambda_r \end{bmatrix}, \tag{7}$$

and we write simply $(Y,J) \sim \mathrm{dBMAP}_r(\mathbf{P}, \mathbf{\Lambda}, \{q_i\})$ and $(X,J) \sim \mathrm{dMMPP}_r(\mathbf{P}, \mathbf{\Lambda})$.

2.2. Auxiliary results for the fitting procedure

The fitting of a $\mathrm{dBMAP}_{M2^L}(\mathbf{P}, \mathbf{\Lambda}, \{q_i\})$ to the input traffic is made in two steps: (i) fitting of the $\mathrm{dMMPP}_{M2^L}(\mathbf{P}, \mathbf{\Lambda})$ modeling the packet arrival process and (ii) fitting of the packet size distribution in phase i, q_i, for $i = 1, 2, \ldots, M2^L$. Thus, we next proceed with a description of some transformations on dMMPPs which we will use to obtain the process $\mathrm{dMMPP}_{M2^L}(\mathbf{P}, \mathbf{\Lambda})$.

We consider the superposition of L independent 2-dMMPPs

$$(X^{(l)}, J^{(l)}) \sim \mathrm{dMMPP}_2(\mathbf{P}^{(l)}, \mathbf{\Lambda}^{(l)}), \quad l = 1, 2, \ldots, L \tag{8}$$

and one M-dMMPP

$$(X^{(L+1)}, J^{(L+1)}) \sim \mathrm{dMMPP}_M(\mathbf{P}^{(L+1)}, \mathbf{\Lambda}^{(L+1)}), \tag{9}$$

which is illustrated in Fig. 1.

Note that, in particular, for $l = 1, 2, \ldots, L$,

$$\mathbf{P}^{(l)} = \begin{bmatrix} p_{11}^{(l)} & p_{12}^{(l)} \\ p_{21}^{(l)} & p_{22}^{(l)} \end{bmatrix}, \quad \mathbf{\Lambda}^{(l)} = \begin{bmatrix} \lambda_1^{(l)} & 0 \\ 0 & \lambda_2^{(l)} \end{bmatrix} \tag{10}$$

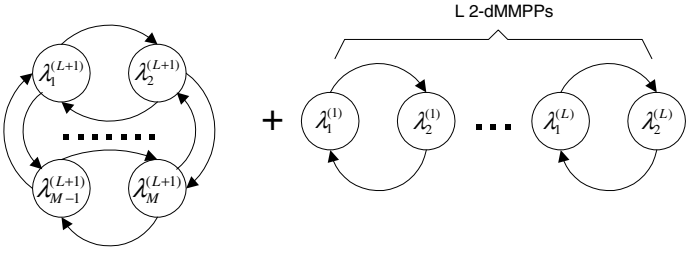

Fig. 1. Superposition of one M-dMMPP and L 2-dMMPP processes.

and we assume that $p_{12}^{(l)} + p_{21}^{(l)} < 1$. In addition, we consider $J^{(1)}, J^{(2)}, \ldots, J^{(L+1)}$ to be ergodic chains in steady-state. For $l = 1, 2, \ldots, L$ we denote by $\pi^{(l)} = [\pi_1^{(l)} \quad \pi_2^{(l)}]$ the stationary distribution of $J^{(l)}$. Similarly, we denote by $\pi^{(L+1)} = [\pi_1^{(L+1)} \quad \pi_2^{(L+1)} \quad \cdots \quad \pi_M^{(L+1)}]$ the stationary distribution of $J^{(L+1)}$.

The result of the superposition is the process

$$(X, J) = \left(\sum_{l=1}^{L+1} X^{(l)}, (J^{(1)}, J^{(2)}, \ldots, J^{(L+1)}) \right) \sim \mathrm{dMMPP}_{M2^L}(\mathbf{P}, \mathbf{\Lambda}), \tag{11}$$

where

$$\mathbf{P} = \mathbf{P}^{(1)} \otimes \mathbf{P}^{(2)} \otimes \cdots \otimes \mathbf{P}^{(L+1)}, \tag{12}$$

$$\mathbf{\Lambda} = \mathbf{\Lambda}^{(1)} \oplus \mathbf{\Lambda}^{(2)} \oplus \cdots \oplus \mathbf{\Lambda}^{(L+1)} \tag{13}$$

with $\oplus$ and $\otimes$ denoting the Kronecker sum and the Kronecker product, respectively. Note that the Markov chain J is also in steady-state. We refer to (X, J) as being the $M2^L$-dMMPP.

In our approach L and M are not fixed a priori but instead are computed as part of the fitting procedure. However, in the rest of this section they may be thought as being fixed. We want that the L 2-dMMPPs capture the autocovariance function of the increments of the packet arrival process (X) and, discounting for the effect of the other L 2-dMMPPs, the M-dMMPP approximates the distribution of the increments of the packet arrival process. To explain how this may be accomplished, it is convenient to define the increment processes associated to $X^{(1)}, X^{(2)}, \ldots, X^{(L+1)}$, and X, which we denote by $Z^{(1)}, Z^{(2)}, \ldots, Z^{(L+1)}$, and Z, respectively. Thus,

$$Z_k^{(l)} = X_{k+1}^{(l)} - X_k^{(l)}, \quad l = 1, 2, \ldots, L+1 \tag{14}$$

and

$$Z_k = X_{k+1} - X_k \tag{15}$$

for $k = 0, 1, \ldots$ Note that Z_k is the (total) number of packet arrivals at sampling interval k and $Z_k^{(l)}$ is the number of packet arrivals that are due to the lth packet arrival process, so that, in particular,

$$Z_k = \sum_{l=1}^{L+1} Z_k^{(l)}, \quad k = 0, 1, 2, \ldots \tag{16}$$

Moreover $Z^{(1)}, Z^{(2)}, \ldots, Z^{(L+1)}$, and Z, are stationary sequences.

In order to characterize the marginal distributions of the processes $Z^{(1)}, Z^{(2)}, \ldots, Z^{(L+1)}$, and Z, we denote, respectively, by $\{f_l(k), k = 0, 1, 2, \ldots\}$, $l = 1, 2, \ldots, L+1$, and $\{f(k), \ k = 0, \ 1, 2, \ldots\}$, their (marginal) probability functions. As the univariate distributions of $Z^{(1)}, Z^{(2)}, \ldots, Z^{(L+1)}$ are mixtures of Poisson distributions, we denote the probability function of a Poisson random variable with mean μ by $\{g_\mu(k), k = 0, 1, 2, \ldots\}$, for $\mu \in [0, +\infty)$, so that

$$g_\mu(k) = \mathrm{e}^{-\mu} \frac{\mu^k}{k!}, \quad k = 0, 1, 2, \ldots \tag{17}$$

For $l = 1, 2, \ldots, L$, the (stationary) marginal distribution of $Z^{(l)}$ (that is, the distribution of $Z_k^{(l)}$, for $k = 0, 1, \ldots$) is a mixture of two Poisson distributions with means $\lambda_1^{(l)}$ and $\lambda_2^{(l)}$ and weights $\pi_1^{(l)}$ and $\pi_2^{(l)}$, respectively, since $\pi_1^{(l)}$ and $\pi_2^{(l)}$ are the stationary probabilities of states 1 and 2 and $\lambda_1^{(l)}$ and $\lambda_2^{(l)}$ are the corresponding Poisson packet arrival rates. Thus the probability functions of $Z^{(l)}$, $l = 1, 2, \ldots, L$, are given by

$$f_l(k) = \pi_1^{(l)} g_{\lambda_1^{(l)}}(k) + \pi_2^{(l)} g_{\lambda_2^{(l)}}(k), \quad k = 0, 1, 2, \ldots \tag{18}$$

and their autocovariance functions are

$$\gamma_k^{(l)} = \mathrm{Cov}\left(Z_0^{(l)}, Z_k^{(l)}\right) = \pi_1^{(l)} \pi_2^{(l)} \left(\lambda_2^{(l)} - \lambda_1^{(l)}\right)^2 \mathrm{e}^{k c_l}, \quad k = 1, 2, \ldots, \tag{19}$$

where $c_l = \ln(1 - p_{12}^{(l)} - p_{21}^{(l)})$. Note that, in particular, the autocovariance functions of $Z^{(1)}, Z^{(2)}, \ldots, Z^{(L)}$ exhibit an exponential decay to zero.

As we want the M-dMMPP to approximate the distribution of the increments of the packet arrival process but to have no contribution to the autocovariance function of the increments of the $M2^L$-dMMPP, we choose to make $J^{(L+1)}$ a Markov chain with no memory whatsoever. This is accomplished by choosing

$$\mathbf{P}^{(L+1)} = \begin{bmatrix} \pi_1^{(L+1)} & \pi_2^{(L+1)} & \cdots & \pi_M^{(L+1)} \\ \pi_1^{(L+1)} & \pi_2^{(L+1)} & \cdots & \pi_M^{(L+1)} \\ \cdots & \cdots & \cdots & \cdots \\ \pi_1^{(L+1)} & \pi_2^{(L+1)} & \cdots & \pi_M^{(L+1)} \end{bmatrix}. \tag{20}$$

Note that this implies that $Z^{(L+1)}$ is an independent and identically distributed sequence of random variables whose distribution is a mixture of M Poisson random variables with means $\lambda_i^{(L+1)}$ and weights π_i^{L+1}, for $i = 1, 2, \ldots, M$. As a consequence, the probability function of $Z^{(L+1)}$ is given by

$$f_{L+1}(k) = \sum_{j=1}^{M} \pi_j^{(L+1)} g_{\lambda_j^{(L+1)}}(k), \quad k = 0, 1, 2, \ldots \tag{21}$$

and the autocovariance function of $Z^{(L+1)}$ is null for all positive lags; i.e.,

$$\gamma_k^{(L+1)} = \mathrm{Cov}(Z_0^{(L+1)}, Z_k^{(L+1)}) = 0, \quad k \geqslant 1. \tag{22}$$

Taking into account (16), it follows that the probability function of Z is given by

$$f(k) = (f_1 \oplus f_2 \oplus \cdots \oplus f_{L+1})(k), \tag{23}$$

where $\oplus$ denotes the convolution of probability functions. Z is a sequence of random variables whose distribution is a mixture of Poisson random variables (note that the sum of independent mixtures of Poisson random variables is also a mixture of Poisson random variables), and the probability function of Z may be written in the following way:

$$f(k) = \sum_{j_1=1}^{2} \sum_{j_2=1}^{2} \cdots \sum_{j_L=1}^{2} \sum_{j_{L+1}=1}^{M} \left(\prod_{l=1}^{L+1} \pi_{j_l}^{(l)} \right) \times g_{\sum_{l=1}^{L+1} \lambda_{j_l}^{(l)}}(k). \tag{24}$$

Moreover, from (16) and taking into account (19) and (22), we conclude that the autocovariance function of Z is given by

$$\gamma_k = \mathrm{Cov}(Z_0, Z_k) = \sum_{l=1}^{L+1} \mathrm{Cov}(Z_0^{(l)}, Z_k^{(l)}) = \sum_{l=1}^{L} \pi_1^{(l)} \pi_2^{(l)} (\lambda_2^{(l)} - \lambda_1^{(l)})^2 \, \mathrm{e}^{k c_l} \tag{25}$$

for $k = 1, 2, \ldots$

3. Inference procedure

In the rest of the paper we will refer to $Z^{(1)}, Z^{(2)}, \ldots, Z^{(L)}$ as the 2-dMMPPs that are used to match the autocovariance of packet arrival counts, to $Z^{(L+1)}$ as the M-dMMPP that is incorporated in order to match the marginal distribution of packet arrivals counts, to Z as the counts of X, the aggregate $M2^L$-dMMPP that models the packet arrival process, and to Y as the $M2^L$-dBMAP that models the arrival process of bytes.

The inference procedure can be divided in four steps: (i) approximation of the empirical autocovariance of packet arrival counts by a weighted sum of exponentials and identification of time scales, (ii) inference of the M-dMMPP probability function and of the 2-dMMPPs parameters, (iii) inference of the M-dMMPP packet arrival rates and transition probabilities, and subsequent calculation of the $M2^L$-dMMPP parameters, and (iv) fitting of the packet size (in bytes) distribution for each of the $M2^L$ phases and calculation of the final $M2^L$-dBMAP parameters. The flow diagram of this procedure is represented in Fig. 2. In the following four subsections we describe these four steps in detail.

3.1. Autocovariance approximation and time scales identification

Our approach is to approximate the autocovariance (of packet arrival counts) by a large number of exponentials and then aggregate exponentials with a similar decay into the same time scale. This is close to the approaches considered in [5,8,40]. As a first step, we approximate the empirical autocovariance by a sum of K exponentials with real positive weights and negative real time constants. We chose K as $\sqrt{k_{\max}}$, where $k_{\max}$ represents the number of points of the empirical autocovariance. This is accomplished through a modified Prony algorithm [41]. The Prony algorithm returns two vectors,

$$\vec{a} = [\, a_1 \quad \cdots \quad a_K \,], \quad \vec{b} = [\, b_1 \quad \cdots \quad b_K \,],$$

which correspond to the approximating function

$$C_k(\vec{a}, \vec{b}) = \sum_{i=1}^{K} a_i \, \mathrm{e}^{-b_i k}, \quad k = 0, 1, 2, \ldots \tag{26}$$

At this point we identify the components of the autocovariance that characterize the different time

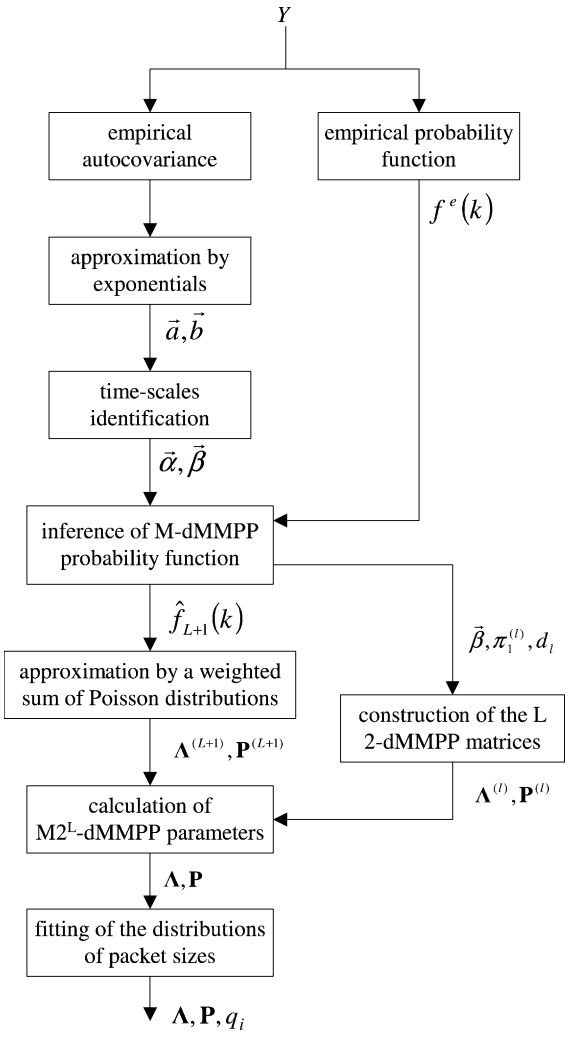

Fig. 2. Flow diagram of the inference procedure.

scales. We define L different time scales, in which the autocovariance decays, b_i, $i = 1,\ldots,K$, fall in the same logarithmic scale. To explain how this is accomplished it is useful to order the b_j coefficients in non-decreasing order, i.e., $b_j \leqslant b_{j+1}$, $1 \leqslant j \leqslant K-1$, and let $\lceil\cdots\rceil$ denote the integer round towards plus infinity. The value L is computed through the following iterative process. Starting with $l = 1$ and $i_l = 1$ compute i_{l+1} through

$$i_{l+1} = \min\{K+1, \inf\{j : i_l < j \leqslant K \wedge \lceil\log_{10}(b_j)\rceil > \lceil\log_{10}(b_{j-1})\rceil\}\}.$$

If $i_{l+1} > K$ then make $L = l$ and stop; otherwise increment l by one and repeat the procedure. Note that, in particular,

$$\lceil\log_{10}(b_{i_l})\rceil = \lceil\log_{10}(b_{i_l+1})\rceil = \cdots = \lceil\log_{10}(b_{i_{l+1}-1})\rceil,$$

but, if $i_{l+1} \leqslant K$,

$$\lceil\log_{10}(b_{i_l})\rceil < \lceil\log_{10}(b_{i_{l+1}})\rceil.$$

For $l = 1,2,\ldots,L$, we consider that the decays b_{i_l} to $b_{i_{l+1}-1}$ characterize the same traffic time scale and we aggregate the $n_l = i_{l+1} - i_l$ components in one component with the following parameters:

$$\alpha_l = \sum_{k=i_l}^{i_{l+1}-1} a_k \quad \text{and} \quad \beta_l = \frac{\sum_{k=i_l}^{i_{l+1}-1} a_k b_k}{\alpha_l}. \tag{27}$$

Taking into account (26) and (27), the autocovariance function of the 2-dMMPP $Z^{(l)}$, $\gamma_k^{(l)}$, is fitted by $\alpha_l e^{k\beta_l}$. Thus, in view of (19), it results that

$$\alpha_l = d_l^2 \pi_1^{(l)} \pi_2^{(l)} \quad \text{and} \quad \beta_l = c_l, \tag{28}$$

where $d_l = \lambda_2^{(l)} - \lambda_1^{(l)}$, and the fitted autocovariance function of $Z_1 + Z_2 + \cdots + Z_L$ is

$$\sum_{l=1}^{L} \alpha_l e^{k\beta_l}, \quad k = 1,2,\ldots \tag{29}$$

The parameters α_l and β_l are obtained via the autocovariance approximation just described; in the next steps, the inference procedure will restrict d_l^2, $\pi_1^{(l)}$ and $\pi_2^{(l)}$ to satisfy (28).

3.2. Inference of the M-dMMPP probability function and of the L 2-dMMPP parameters

The next step is the inference of the M-dMMPP probability function from the empirical probability function of the original data trace. The relation between the probability functions of the 2-dMMPPs, the M-dMMPP and the $M2^L$-dMMPP is defined by (23).

In order to simplify the deconvolution of $f_{L+1}(k)$ and $f_l(k)$, $l = 1,\ldots,L$, we reduce the number of parameters to be fitted by considering that the Poisson packet arrival rate is zero in one state of each 2-dMMPP source; that is, $\lambda_1^{(l)} = 0$ and $\lambda_2^{(l)} = d_l$, for $l = 1,\ldots,L$. From (28),

$$d_l = \sqrt{\frac{\alpha_l}{\pi_1^{(l)}\pi_2^{(l)}}}, \quad l = 1, 2, \ldots, L. \tag{30}$$

The probability function of the M-dMMPP, f_{L+1}, is inferred from the empirical probability function of the packet data, denoted by f^{e}, and the L 2-dMMPP probability functions, denoted by $\hat{f}_l$, $l = 1, 2, \ldots, L$, based on the fitted parameters, after fixing the probabilities $\pi_1^{(l)}$, $l = 1, 2, \ldots, L$, through (23). More precisely, f_{L+1} is fitted jointly with the parameters $\pi_1^{(l)}$, $l = 1, \ldots, L$, through the following constrained minimization process:

$$\underset{\{\pi_1^{(l)}, l=1,\ldots,L\},\{f_{L+1}(k),k=0,1,\ldots\}}{\text{minimize}} \sum_k |o^{\mathrm{e}}(k)|, \tag{31}$$

where

$$o^{\mathrm{e}}(k) = f^{\mathrm{e}}(k) - (\hat{f}_1 \oplus \cdots \oplus \hat{f}_L \oplus f_{L+1})(k) \tag{32}$$

subject to (28) and

$$\begin{aligned} &0 < \pi_1^{(l)} < 1, \quad l = 1, 2, \ldots, L, \\ &f_{L+1}(k) \geqslant 0, \quad k = 0, 1, \ldots \\ &\text{and} \quad \sum_{k=0}^{+\infty} f_{L+1}(k) = 1. \end{aligned} \tag{33}$$

We denote by $\hat{f}_{L+1}$ the fitted probability function of the M-dMMPP. Note that $\pi_1^{(l)}$ is not allowed to be 0 or 1 because, in both cases, the lth 2-dMMPP would degenerate into a Poisson process. The empirical probability function, $f^{\mathrm{e}}(k)$, is inferred for the range of values defined by

$$k = 0, 1, \ldots, \max\left\{1.1\sum_{l=1}^{L} d_l \Delta t, \max_k Z_k\right\},$$

where Δt is the sampling interval and Z_k is the number of packet arrivals at sampling interval k. The probability functions $\hat{f}_l(k)$, $l = 1, \ldots, L$ are also calculated for the same range of values. However, in order to the reduce the number of points used in the convolutions, only the subrange of values defined by

$$\begin{aligned} k = \min_j \{\hat{f}_l(j) > \xi_l\}, \min_j \{\hat{f}_l(j) > \xi_l\} + 1, \\ \ldots, \max_j \{\hat{f}_l(j) > \xi_l\}, \end{aligned}$$

where $\xi_l = 10^{-4}\max(\hat{f}_l)$, is considered. The constrained minimization process given by (31)–(33) is a non-linear programming problem and in general, it is computationally demanding to obtain the global optimal solution. Accordingly, to solve this problem we consider two approximations: (i) we make $\pi_1^{(l)} = \pi_1^{(l+1)}$, $l = 1, \ldots, L-1$ and (ii) we restrict the range of possible $\pi_1^{(l)}$ solutions to be discrete and such that $\pi_1^{(l)} = 0.001k$, $k = 1, \ldots, 999$. Then a search process is used to find the minimum value of the objective function. The complexity of the minimization process is essentially proportional to L, due to the convolution of the L 2-dMMPPs. The considered approximations have had negligible impact on the results obtained so far with the fitting procedure, in particular on those presented in Section 5.

At this point all parameters of the 2-dMMPPs, $X^{(1)}, X^{(2)}, \ldots, X^{(L)}$, have been determined and their corresponding 2-dMMPP matrices

$$\{(\mathbf{P}^{(l)}, \mathbf{\Lambda}^{(l)}), l = 1, 2, \ldots, L\}$$

can be constructed in the following way:

$$\mathbf{P}^{(l)} = \begin{bmatrix} 1 - \pi_2^{(l)}(1 - \mathrm{e}^{\beta_l}) & \pi_2^{(l)}(1 - \mathrm{e}^{\beta_l}) \\ \pi_1^{(l)}(1 - \mathrm{e}^{\beta_l}) & 1 - \pi_1^{(l)}(1 - \mathrm{e}^{\beta_l}) \end{bmatrix},$$

$$\mathbf{\Lambda}^{(l)} = \begin{bmatrix} 0 & 0 \\ 0 & d_l \end{bmatrix}.$$

3.3. Inference of the M-dMMPP packet arrival rates and transition probabilities

The next step is the inference of the number of states and Poisson packet arrival rates of the M-dMMPP from $\hat{f}_{L+1}$. To do this we infer $\hat{f}_{L+1}$ as a weighted sum of Poisson probability functions.

The matching is carried out through an algorithm that progressively subtracts a Poisson probability function from $\hat{f}_{L+1}$. This algorithm is described in the flowchart of Fig. 3. We represent the ith Poisson probability function with mean φ_i by $g_{\varphi_i}(k)$. We define $h^{(i)}(k)$ as the difference between $\hat{f}_{L+1}(k)$ and the weighted sum of Poisson probability functions at the ith iteration. Initially, we set $h^{(1)}(k) = \hat{f}_{L+1}(k)$. In each step, we first detect the maximum of $h^{(i)}(k)$. The corresponding k-value, $\varphi_i = \arg\max_k h^{(i)}(k)$, will be considered the ith Poisson rate of the M-dMMPP. We then calculate the weights of each Poisson probability

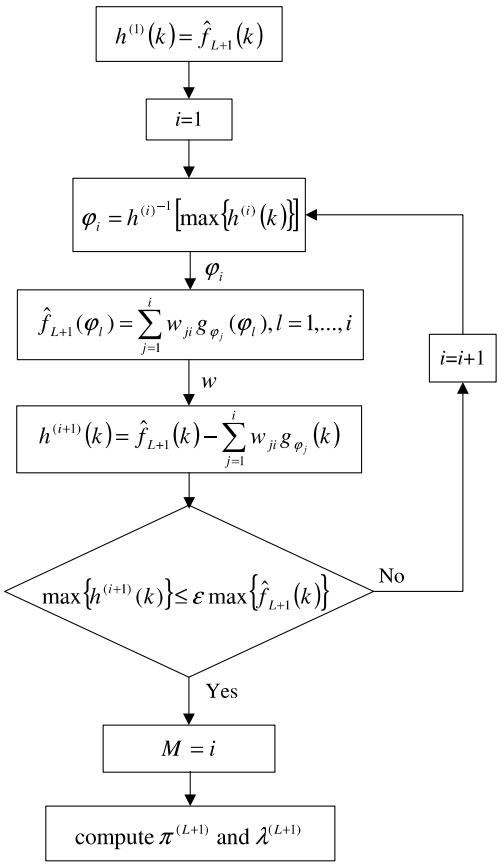

Fig. 3. Algorithm for calculation of the number of states and Poisson packet arrival rates of the M-dMMPP.

function, $\vec{w}_i = [w_{1i}, w_{2i}, \ldots, w_{ii}]$, through the following set of linear equations:

$$\hat{f}_{L+1}(\varphi_l) = \sum_{j=1}^{i} w_{ji} g_{\varphi_j}(\varphi_l), \quad l = 1, \ldots, i.$$

This assures that the fitting between $\hat{f}_{L+1}(k)$ and the weighted sum of Poisson probability functions is exact at φ_i points, for $l = 1, 2, \ldots, i$. The final step in each iteration is the calculation of the new difference function

$$h^{(i)}(k) = \hat{f}_{L+1}(k) - \sum_{j=1}^{i} w_{ji} g_{\varphi_j}(k).$$

The algorithm stops when the maximum of $h^{(i)}(k)$ is lower than a pre-defined percentage ε of the maximum of $\hat{f}_{L+1}(k)$ and M is made equal to i.

Other methods for parameter estimation of finite Poisson mixtures with an unknown number of components, have been proposed, e.g., based on moment estimation or maximum likelihood (see, e.g., [42] and references therein). These methods lead to the solution of a system of non-linear equations with a number of variables equal to twice the number of components of the Poisson mixture model. By contrast, in the algorithm described above the rates of the Poisson mixture model are fitted directly and the weights are fitted by solving a system of linear equations with the weights as variables, thus obtaining a more computationally efficient procedure. Moreover, the proposed procedure leads to the fitting of a small number of components for the fitted IP traces. Note also that the proposed procedure is in line with the global heuristic approach followed in the paper, which seeks a fast method to fit a particular class of MMPPs instead of a purely statistical fitting of an MMPP.

After M has been determined, the parameters of the M-dMMPP, $\{(\pi_j^{(L+1)}, \lambda_j^{(L+1)}), j = 1, 2, \ldots, M\}$, are then set equal to

$$\pi_j^{(L+1)} = w_{jM} \quad \text{and} \quad \lambda_j^{(L+1)} = \varphi_j.$$

Finally, the $M2^L$-dMMPP process can be constructed using Eqs. (12) and (13), where $\mathbf{\Lambda}^{(L+1)}$, $\mathbf{P}^{(L+1)}$, $\mathbf{\Lambda}^{(i)}$ and $\mathbf{P}^{(i)}$, $i = 1, \ldots, L$, were calculated in the last two subsections.

3.4. Packet size distribution fitting and calculation of the final parameters

The packet size characterization is made independently for each state of the inferred $M2^L$-dMMPP. There are two steps: (i) association of each time slot to one of the $M2^L$-dMMPP states and (ii) inference of a packet size distribution for each state of the $M2^L$-dMMPP.

In the first step, we scan all time slots of the empirical data. A time slot in which k packet arrivals were observed is randomly assigned to a state, according to the probability vector $\vec{\theta}(k) = \{\theta_1(k), \ldots, \theta_{M2^L}(k)\}$, where $\theta_i(k)$ represents the probability that the observed k packet arrivals were originated in state i. This is given by

$$\theta_i(k) = \frac{\pi_i g_{\lambda_i}(k)}{\sum_{j=1}^{M2^L} \pi_j g_{\lambda_j}(k)}, \quad (34)$$

where λ_j represents the Poisson packet arrival rate of the jth state of the $M2^L$-dMMPP and π_j the corresponding steady-state probability (as stated before, $g_\lambda(y)$ represents a Poisson probability distribution function with mean λ).

The inference of the packet size distribution in each state resorts to histograms. The inference of each histogram uses only the packets that arrived during the time slots previously associated with the state for which we are inferring the packet size distribution. Note that some low-probability states may have no packets associated with them, making impossible the packet characterization specifically for these states. We associate a packet size distribution to these states that considers all data packets, i.e., the packet size distribution unconditioned on the $M2^L$-dMMPP states. The histograms result in the packet size distributions $q_i = \{q_i(n), n \in \mathbb{N}\}$ for $i = 1, 2, \ldots, M2^L$.

This step completes the inference procedure. According to (5), the $M2^L$-dBMAP that models the input traffic is fully characterized by the q_i, $i = 1, 2, \ldots, M2^L$, inferred in this section and by matrices $\mathbf{\Lambda}$ and $\mathbf{P}$ inferred in previous sections.

4. The importance of fitting the packet size

The purpose of this section is to illustrate, through a worked example, the impact in terms of queuing behavior of modeling the packet size and its correlation with the packet arrivals. We compare the packet loss ratio versus buffer size curves obtained with three dBMAPs where the packet size distribution differs but the packet arrival process is kept the same. The packet arrival process is modeled by a 2-dMMPP with parameters $\pi_1 = 0.2$, $\pi_2 = 0.8$, $\lambda_1 = 1000$ pkts/s and $\lambda_2 = 300$ pkts/s. We consider two possible packet sizes: 50 and 1000 bytes. In the first dBMAP, which we take as our reference in this study, the packet size distribution is 80% of 50 bytes and 20% of 1000 bytes in state 1, i.e., $q_1(50) = 0.8$ and $q_1(1000) = 0.2$, and 50% of both sizes in state 2, i.e., $q_2(50) = 0.5$ and $q_2(1000) = 0.5$. In the second dBMAP, we consider that the packet size distribution is the same in both states, and is given by the unconditional packet size distribution of the first dBMAP, i.e., the packet size distribution that considers all packets irrespective of the 2-dMMPP states. Thus the packet size distribution is, in both states, 63.63% of 50 bytes and 36.37% of 1000 bytes. In the third dBMAP, we assume that the packet size is fixed and equal to the average packet size of the first and second dBMAPs (395 bytes). Note that the second and third dBMAPs correspond to simplifications of the first one, which are usually taken when characterizing the input traffic and assessing queuing behavior. In particular, both simplifications consider independence between packet arrivals and packet sizes.

We estimated the packet loss ratio via trace-driven simulation using traces of 1 million packets generated from the three dBMAPs described above. The sampling period was 0.1 s, the service rate was 210 Kbytes/s and the buffer size was varied from 1 to 100 Kbytes. The results are shown in Fig. 4. As can be observed, there is a large difference between the packet loss ratio obtained with the first dBMAP and with the second and third ones. For example, for a buffer size of 20 Kbytes the packet loss ratios were 1.39×10^{-3}, 3.79×10^{-2} and 6.36×10^{-2} for the first, second and third dBMAP traces, respectively. This clearly indicates that the simplifications subjacent to the second and third dBMAPs can lead to large errors in

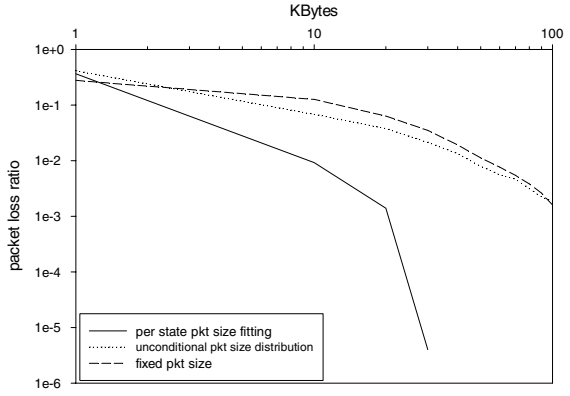

Fig. 4. Packet loss ratio versus buffer size, dBMAP.

terms of queuing behavior. Thus detailed modeling of the packet size and of the correlations with the packet arrivals is clearly required. In the next section, we complete this study through the consideration of measured data traces.

5. Results

We have applied our fitting procedure to five traces of IP traffic: (i) the well known pAug.TL Bellcore trace [13] and (ii) four traces measured at the University of Aveiro (UA), all of them exhibiting LRD. The UA traces are representative of Internet access traffic produced within a University campus environment. The University of Aveiro is connected to the Internet through a 10 Mbit/s ATM link and the measurements were carried out in a 100 Mbit/s Ethernet link connecting the border router to the firewall, which only transports Internet access traffic.

We assess the fitting procedure using several performance evaluation criteria. In the case of the packet arrival process, we compare both the probability and autocovariance functions of the packet arrival counts obtained with the fitted dBMAPs (theoretical) and with the original data traces. The sampling interval of the counting process was 0.1 s. In the case of the packet size distribution, we compare the probability function of the original trace and of a trace generated from the fitted dBMAP (hereafter called fitted trace).

We analyze the presence of LRD behavior, in both the original and fitted traces, using the method described in [31]. This method resorts to the so-called logscale diagram which consists in the graph of y_j against j, together with confidence intervals about the y_j, where y_j is a function of the wavelet discrete transform coefficients at scale j. Traffic is LRD if, within the limits of the confidence intervals, the y_j fall on a straight line with slope $\alpha \in (0,1)$, in a range of scales from some initial value j up to the largest one present in data. The Hurst parameter is related to the slope α by $H = (\alpha + 1)/2$. Thus LRD behavior implies an Hurst parameter $H \in (0.5, 1)$. To relate scales with seconds we note that octave j corresponds to 0.1×2^j s.

Table 1 indicates the number of states of the fitted dBMAPs and summarize the fitting results relative to the probability function (PF) and autocovariance function (AF) of the packet arrival counts. The fitting error is defined as the quotient between the integral of the absolute difference between the curves of the original and fitted traces, and the integral of the original trace curve. The results indicate that a close match was obtained in case of the PF; the match is not so good for the AF due to the oscillatory behavior in the case of the original data, but this has negligible impact on the queuing behavior. This type of behavior has been observed in other data traces. Table 2 lists the estimated Hurst parameters of both original and fitted traces, showing that all traces exhibit LRD behavior and that the fitted dBMAPs were able to capture this behavior. The table also includes the range of largest scales where alignment (in the logscale diagram) was obtained.

We also analyze the queuing behavior by comparing the packet loss ratio, obtained through trace-driven simulation, using four types of input traffic: (i) original traces, (ii) traces generated according to the fitted dBMAP, (iii) traces where the arrival instants were generated according to

Table 1
Fitting results for the packet arrival process

Trace name	Inferred model	π_1	L	Mean rate (pkts/s)	PF IC (%)	AF IC (%)
pAug.TL	56-dBMAP	0.1	3	318	7.25	29.05
UA1	24-dBMAP	0.1	3	776	10.02	7.31
UA6	12-dBMAP	0.2	2	527	13.62	5.70
UA10	16-dBMAP	0.1	2	1111	9.70	17.88
UA19	12-dBMAP	0.2	2	1125	10.75	5.23

Table 2
Estimated Hurst parameters of original and fitted traces

Trace name	Hurst parameter (original)	Hurst parameter (fitted)
pAug.TL	0.866 (3–11)	0.880 (4–11)
UA1	0.985 (7–15)	0.956 (8–15)
UA6	0.969 (6–15)	0.943 (7–15)
UA10	0.986 (6–15)	0.981 (6–15)
UA19	0.982 (6–15)	0.977 (5–15)

the fitted dMMPP arrival process and the packet size according to the unconditional packet size distribution of the fitted dBMAP and (iv) traces where the arrival instants were also generated according to the fitted dMMPP arrival process but the packet size is fixed and equal to the average packet size of the original trace. The results of trace-driven simulation for the fitted traces were based on 10 replicas.

5.1. Belcore trace

A 56-dBMAP was fitted to the pAug.TL trace. In Figs. 5 and 6 we show the results of the packet arrival fitting. There is an excellent agreement in terms of probability functions. In terms of autocovariance, it can be seen that the fitted model is able to reproduce the average behavior of the empirical autocovariance (but not its oscillatory behavior).

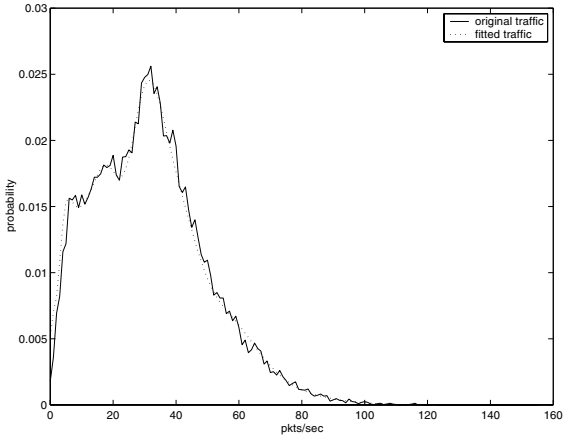

Fig. 5. Probability function of the packet arrival rate, trace pAug.TL.

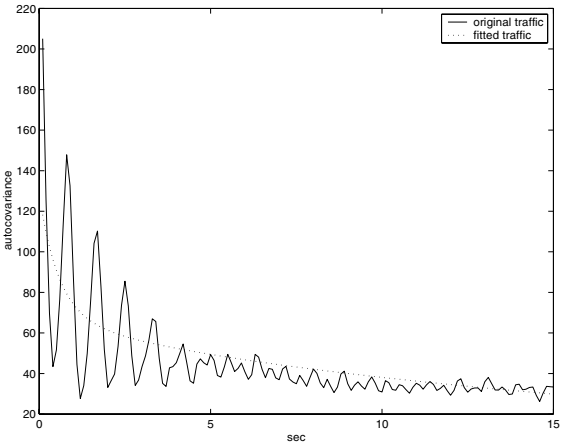

Fig. 6. Autocovariance of the packet arrival rate, trace pAug.TL.

In order to analyze the queuing behavior we considered a queue with a service rate of 175 Kbytes/s, corresponding to a link utilization of $\rho = 0.79$, and varied the buffer size from 100 Kbytes to 8 Mbytes. As it can be observed in Fig. 7, there is a close agreement between the curves corresponding to the original trace and to the trace generated according to the fitted 56-dBMAP, for all buffer size values. In contrast, for the other two curves corresponding to traces where the packet size is fitted independently of the packet arrival process, significant deviations are obtained. This confirms the conclusions of the previous section.

The packet size distributions of the original and fitted traces are shown in Figs. 8 and 9, respectively.

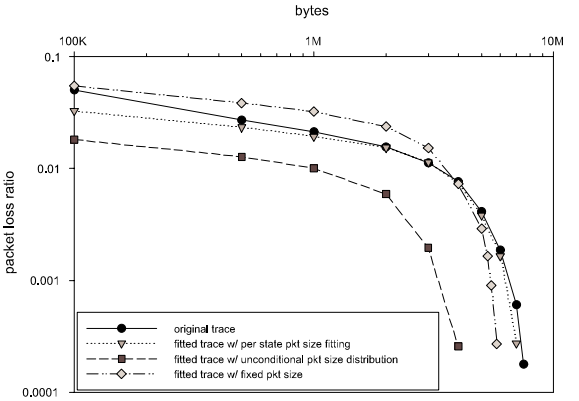

Fig. 7. Packet loss ratio versus buffer size, trace pAug.TL.

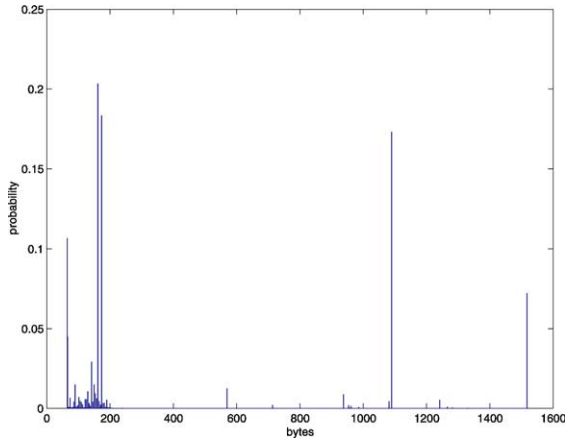

Fig. 8. Packet size histogram of original data, trace pAug.TL.

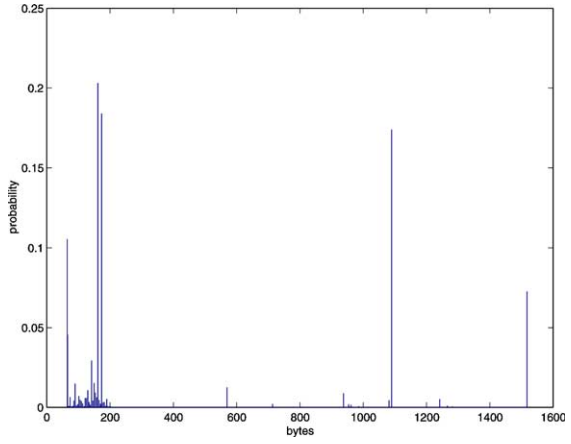

Fig. 9. Packet size histogram of fitted trace, trace pAug.TL.

An excellent agreement between the original and fitted distributions was obtained.

5.2. UA traces

The main characteristics of the UA traces are described in Tables 3 and 4. Each UA trace corresponds to a different day period. All measurements captured 20 million packets. However, to assure stationarity, traces UA6 and UA10 were truncated to 10 million packets. The traffic analyzer was a 1.2 GHz AMD Athlon PC, with 1.5 Gbytes of RAM, running WinDump. The measurements recorded the arrival instant and the IP header of each packet.

Table 4
Main statistics of measured traces

Trace name	Mean rate (pkts/s)	Mean rate (Kbytes/s)	Mean pkt size (bytes)
UA1	766	461.0	598
UA6	533	441.8	692
UA10	1074	645.4	600
UA19	1138	629.0	557

In the following we will make a detailed analysis of the results relative to trace UA19. Fig. 10 shows that the trace has LRD, since there is alignment from scale 6 up the maximum one present in data. The slope of the straight line is $\alpha = 0.964$, leading to an Hurst parameter of $H = 0.982$. This analysis considered the traffic in bytes/s, i.e., including arrivals and packet sizes. Similar results were obtained when considering only the arrival process (traffic in packets/s).

A 12-dBMAP was fitted to the trace. The parameter estimation took less than 30 s, using a MATLAB implementation running in the PC described above. This shows that the fitting procedure is computationally very efficient.

In Figs. 11 and 12 we show the results of the packet arrival fitting. There is an excellent agreement in terms of probability functions. In terms of autocovariance, the empirical function shows some oscillatory behavior. However, the dBMAP succeeded in capturing the main trend and, as will be seen later, this is enough for accurate queuing

Table 3
Main characteristics of measured traces

Trace name	Capture date	Capture interval	Trace size (pkts)
UA1	Tue, July 3rd 2001	8.00 p.m. to 3.15 a.m.	20 millions
UA6	Thu, July 5th 2001	3.58 a.m. to 9.11 a.m.	10 millions
UA10	Fri, July 6th 2001	12.41 p.m. to 3.16 p.m.	10 millions
UA19	Tue, July 10th 2001	10.15 a.m. to 3.08 p.m.	20 millions

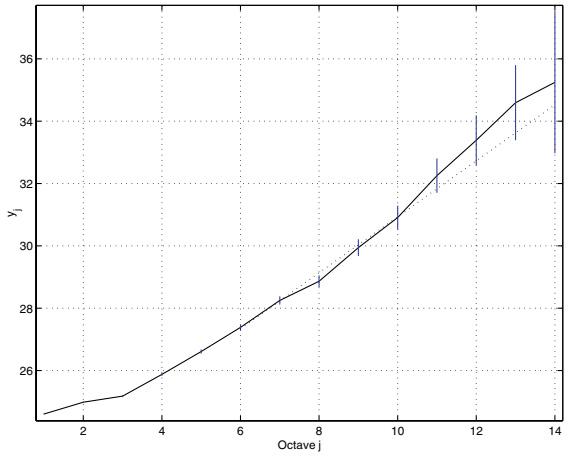

Fig. 10. Logscale diagram for the bytes/s process, trace UA19.

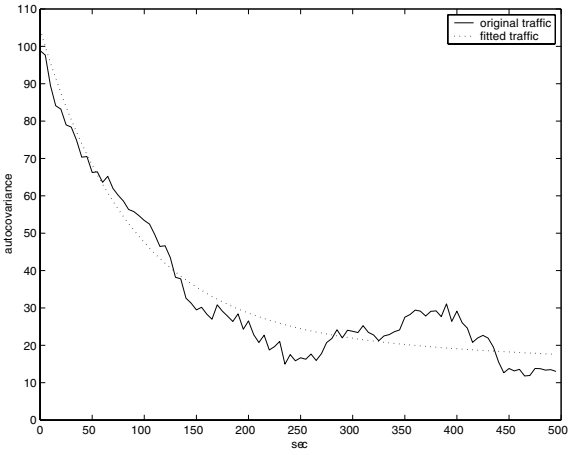

Fig. 12. Autocovariance of the packet arrival rate, trace UA19.

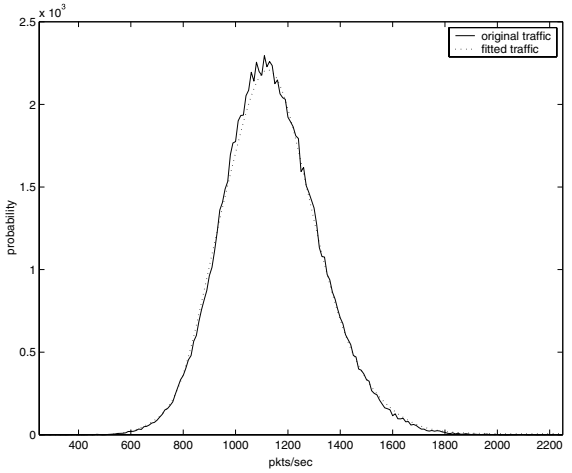

Fig. 11. Probability function of the packet arrival rate, trace UA19.

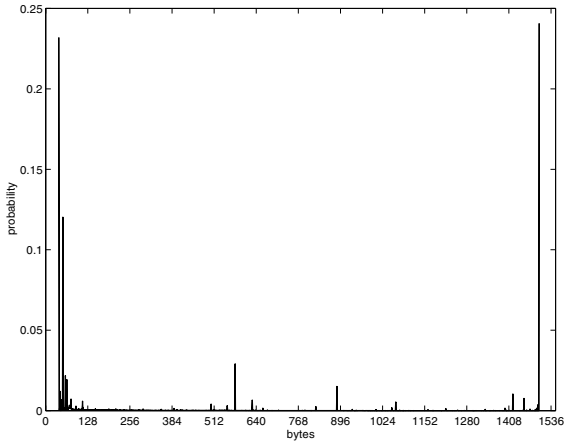

Fig. 13. Packet size histogram of original data, trace UA19.

behavior assessment. Note that the autocovariance was fitted up to time lags of 500 s. In general, the accuracy of packet arrival fitting is similar to the ones reported in [2] for the case of the autocovariance and marginal distribution and in [5] for the case of the autocovariance and mean arrival rate.

The packet size distributions of the original and fitted traces are shown in Figs. 13 and 14, respectively. The distribution is essentially bimodal with two pronounced peaks around 40 and 1500 bytes; it also presents non-negligible values at 576 and 885 bytes. We note that the minimum IP packet size is 40 bytes and that, in many implementations, the maximum is 1500 bytes. Again, an excellent agreement between the original and fitted distributions was obtained. Note, in particular, the fitting accuracy on the lowest probability packet sizes.

In order to analyze the queuing behavior we considered a queue with a service rate of 700 Kbytes/s, corresponding to a link utilization of $\rho = 0.90$, and varied the buffer size from 10 Kbytes to 60 Mbytes. As it can be observed in Fig. 15, there is a close agreement between the curves

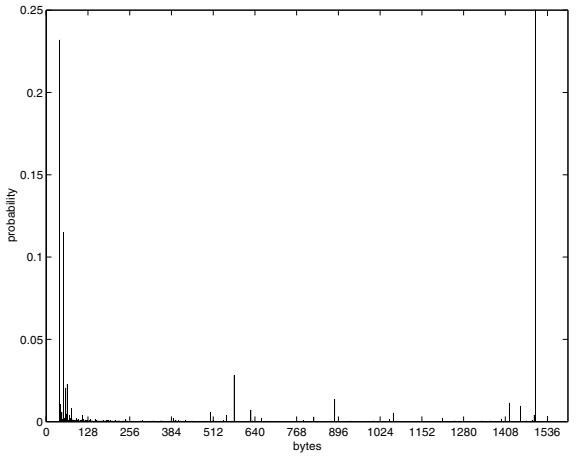

Fig. 14. Packet size histogram of fitted trace, trace UA19.

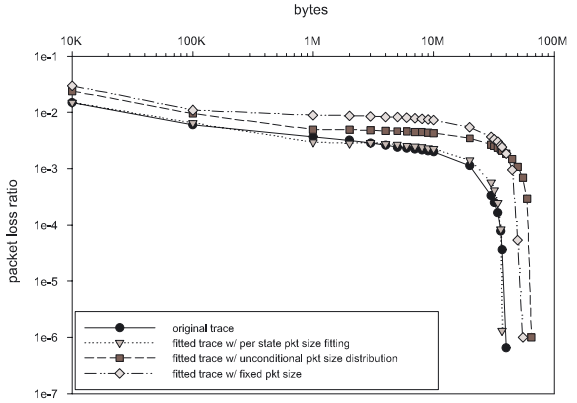

Fig. 15. Packet loss ratio versus buffer size, trace UA19.

corresponding to the original trace and to the trace generated according to the fitted 24-dBMAP, for all buffer size values. In contrast, for the other two curves corresponding to traces where the packet size is fitted independently of the packet arrival process, significant deviations are obtained. This confirms the conclusions of the previous section.

The results obtained with the other three traces are similar. The queuing results for the three traces are shown in Figs. 16–18. The service rate was 625 Kbytes/s ($\rho = 0.72$), 500 Kbytes/s ($\rho = 0.88$) and 650 Kbytes/s ($\rho = 0.99$) for traces UA1, UA6 and UA10, respectively. In all cases the buffer size was varied from 10 Kbytes to 60 Mbytes. As in the case of the UA19 trace, there exist a close agreement

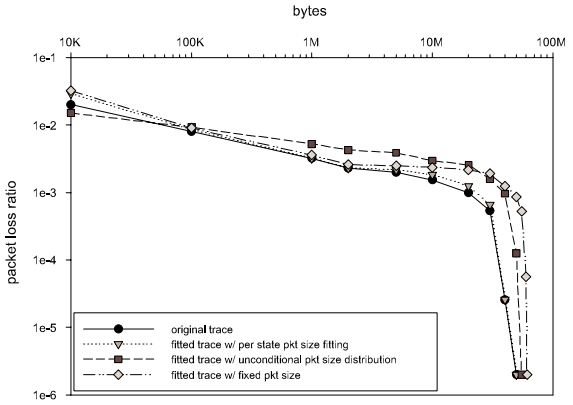

Fig. 16. Packet loss ratio versus buffer size, trace UA1.

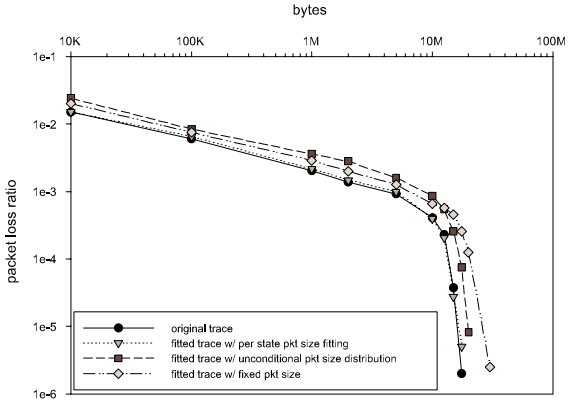

Fig. 17. Packet loss ratio versus buffer size, trace UA6.

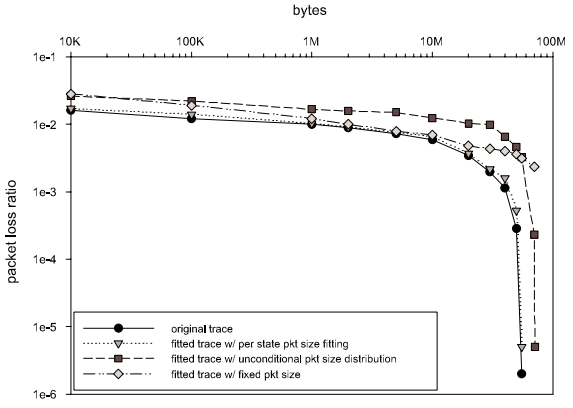

Fig. 18. Packet loss ratio versus buffer size, trace UA10.

for the fitted dBMAPs, but large differences when the packet size is fitted independently of the packet

arrival process. When compared with the queuing results reported in [11], our results show a closer fitting, despite the fact that we are considering higher link utilizations.

The number of states of the resulting dBMAPs can, in some cases, restrict the applicability of the model to very simple analytical performance studies. However, we note that the number of states is determined as part of the fitting procedure and that it can be reduced by alleviating the accuracy requirements on the fitting of the M-dMMPP probability function.

6. Conclusions

We have developed a fitting procedure for discrete-time batch Markovian arrival processes (dBMAPs), which allows for the simultaneous matching of both: (i) the autocovariance and marginal distribution of the packet arrival process and (ii) the distribution of the packet size.

The procedure was applied to several measured traces exhibiting LRD. Our numerical results, have shown that the procedure matches closely the autocovariance and the probability function of the packet arrival process and also the packet size distribution. The queuing behavior, as assessed by the packet loss ratio suffered by data and fitted traces, also shows a very good agreement, when the packet size distribution is fitted individually for each state of the dMMPP modeling packet arrivals. Our results also point out that matching only the packet arrival process and considering that packets have a fixed size or distribution can lead to significant errors in the estimation of the packet loss ratio.

A direction for future work is the consideration of continuous-time BMAPs. In fact, continuous-time BMAPs are popular for a large range of performance models and there are performance modeling tools available using continuous-time BMAPs. Thus, it is of interest to develop traffic modeling procedures similar to the one presented in the paper using continuous-time BMAPs. As our approach is based at first on the fitting of the autocovariance function of packet arrival counts, a discrete-time characteristic, a first step in this direction would be to obtain a proper continuous-time function substitute of the autocovariance function of packets counts.

Acknowledgements

The authors would like to thank the anonymous referees for their valuables comments and suggestions. This research was supported in part by Fundação para a Ciência e a Tecnologia, the project POSI/42069/CPS/2001, and the grant BD/19781/99.

References

[1] K. Meier-Hellstern, A fitting algorithm for Markov-modulated Poisson process having two arrival rates, European Journal of Operational Research 29 (1987) 370–377.
[2] S. Li, C. Hwang, On the convergence of traffic measurement and queuing analysis: a statistical-match and queuing (SMAQ) tool, IEEE/ACM Transactions on Networking 5 (1) (1997) 95–110.
[3] S. Robert, J.L. Boudec, A modulated Markov process for self-similar traffic, in: Proceedings of Saarbrucken Schloss Dagstuhl, Germany, 1995, pp. 1–14.
[4] S. Robert, J.L. Boudec, New models for self-similar traffic, Performance Evaluation 30 (1–2) (1997) 57–68.
[5] A. Andersen, B. Nielsen, A Markovian approach for modeling packet traffic with long-range dependence, IEEE Journal on Selected Areas in Communications 16 (5) (1998) 719–732.
[6] B. Hajek, L. He, On variations of queue response for inputs with the same mean and autocorrelation function, IEEE/ACM Transactions on Networking 6 (5) (1998) 588–598.
[7] R. Riedi, M. Crouse, V. Ribeiro, R. Baraniuk, A multifractal wavelet model with application to network traffic, IEEE Transactions on Information Theory 45 (4) (1999) 992–1018.
[8] T. Yoshihara, S. Kasahara, Y. Takahashi, Practical time-scale fitting of self-similar traffic with Markov-modulated Poisson process, Telecommunication Systems 17 (1–2) (2001) 185–211.
[9] P. Salvador, A. Nogueira, R. Valadas, Modeling multifractal traffic with stochastic L-Systems, in: Proceedings of GLOBECOM 2002, 2002.
[10] P. Salvador, R. Valadas, A. Pacheco, Multiscale fitting procedure using Markov modulated Poisson processes, Telecommunications Systems 23 (1–2) (2003) 123–148.
[11] A. Klemm, C. Lindemann, M. Lohmann, Traffic modeling of IP networks using the batch Markovian arrival process, Performance Evaluation 54 (2) (2003) 149–173.

[12] J. Gao, I. Rubin, Multifractal analysis and modeling of long-range-dependent traffic, in: Proceedings ICC'99, 1999, pp. 382–386.
[13] W. Leland, M. Taqqu, W. Willinger, D. Wilson, On the self-similar nature of Ethernet traffic (extended version), IEEE/ACM Transactions on Networking 2 (1) (1994) 1–15.
[14] J. Beran, R. Sherman, M. Taqqu, W. Willinger, Long-range dependence in variable-bit rate video traffic, IEEE Transactions on Communications 43 (2–4) (1995) 1566–1579.
[15] M. Crovella, A. Bestavros, Self-similarity in World Wide Web traffic: evidence and possible causes, IEEE/ACM Transactions on Networking 5 (6) (1997) 835–846.
[16] V. Paxson, S. Floyd, Wide-area traffic: the failure of Poisson modeling, IEEE/ACM Transactions on Networking 3 (3) (1995) 226–244.
[17] K. Park, G. Kim, M. Crovella, On the relationship between file sizes, transport protocols, and self-similar network traffic, in: Proceedings of the 4th International Conference on Networks Protocols (ICNP'96), 1996, pp. 171–180.
[18] W. Willinger, M. Taqqu, R. Sherman, D. Wilson, Self-similarity through high-variability: statistical analysis of Ethernet LAN traffic at the source level, IEEE/ACM Transactions on Networking 5 (1) (1997) 71–86.
[19] W. Willinger, V. Paxson, M. Taqqu, Self-similarity and Heavy Tails: Structural Modeling of Network Traffic, A Practical Guide to Heavy Tails: Statistical Techniques and Applications, Birkhauser, Boston, 1998.
[20] R. Riedi, J. Véhel, Multifractal properties of TCP traffic: a numerical study, Technical Report No. 3129, INRIA Rocquencourt, France. Available from <www.stat.rice.edu/~riedi>.
[21] A. Feldmann, A. Gilbert, W. Willinger, Data networks as cascades: investigating the mullifractal nature of Internet WAN traffic, in: Proceedings of SIGCOMM, 1998, pp. 42–55.
[22] A. Feldmann, A. Gilbert, P. Huang, W. Willinger, Dynamics of IP traffic: a study of the role of variability and the impact of control, in: SIGCOMM, 1999, pp. 301–313.
[23] P. Abry, P. Flandrin, M. Taqqu, D. Veitch, Wavelets for the analysis, stimation and synthesis of scaling data, in: K. Park, W. Willinger (Eds.), Self-Similar Network Traffic Analysis and Performance Evaluation, Wiley, New York, 2000.
[24] A. Erramilli, O. Narayan, A. Neidhardt, I. Saniee, Performance impacts of multi-scaling in wide area TCP/IP traffic, in: Proceedings of INFOCOM'2000, 2000.
[25] B. Ryu, A. Elwalid, The importance of long-range dependence of VBR video traffic in ATM traffic engineering: myths and realities, ACM Computer Communication Review 26 (1996) 3–14.
[26] M. Grossglauser, J.C. Bolot, On the relevance of long-range dependence in network traffic, IEEE/ACM Transactions on Networking 7 (5) (1999) 629–640.
[27] A. Nogueira, R. Valadas, A simulation study on the relevant time scales of the input traffic for a tandem network, in: Proceedings of IEEE ICC 2002, 2002.
[28] D. Lucantoni, New results on the single server queue with a batch Markovian arrival process, Stochastic Models 7 (1) (1991) 1–46.
[29] D. Lucantoni, The BMAP/G/1 queue: a tutorial, in: L. Donatiello, R. Nelson (Eds.), Models and Techniques for Performance Evaluation of Computer and Communication Systems, Springer, Berlin, 1993, pp. 330–358.
[30] A. Pacheco, N.U. Prabhu, Markov-additive processes of arrivals, in: J. Dshalalow (Ed.), Advances in Queueing: Theory and Methods, CRC, Boca Raton, FL, 1995, pp. 167–194 (Chapter 6).
[31] D. Veitch, P. Abry, A wavelet based joint estimator for the parameters of LRD, IEEE Transactions on Information Theory 45 (3) (1999) 878–897.
[32] D. Lucantoni, K. Meier-Hellstern, M. Neuts, A single-server queue with server vacations and a class of non-renewal arrival processes, Advances in Applied Probability 22 (1990) 676–705.
[33] S. Asmussen, G. Koole, Marked point processes as limits of Markovian arrival streams, Journal of Applied Probability 30 (2) (1993) 365–372.
[34] C. Blondia, O. Casals, Statistical multiplexing of VBR sources: a matrix-analytic approach, Performance Evaluation 16 (1–3) (1992) 5–20.
[35] C. Blondia, A discrete-time batch Markovian arrival process as B-ISDN traffic model, Belgian Journal of Operations Research, Statistics and Computer Science 32 (1993) 3–23.
[36] A. Alfa, S. Chakravarthy, A discrete queue with the Markovian arrival process and phase type primary and secondary services, Stochastic Models 10 (2) (1994) 437–451.
[37] N. Rananand, Markov approximations to D-BMAPs: information-theoretic bounds on queueing performance, Stochastic Models 11 (4) (1995) 713–734.
[38] F. Geerts, C. Blondia, Superposition of Markov sources and long range dependence, in: F. Aagesen, H. Botnevik, D. Khakhar (Eds.), Information Network and Data Communications: Proceedings of the IFIP/ICCC International Conference on Information Network and Data Communication, Chapman & Hall, London, 1996.
[39] A. Ridder, Fast simulation of discrete time queues with Markov modulated batch arrivals and batch departures, AEU International Journal of Electronics and Communications 52 (1998) 127–132.
[40] A. Feldmann, W. Whitt, Fitting mixtures of exponentials to long-tail distributions to analyze network performance models, Performance Evaluation 31 (3–4) (1997) 245–279.
[41] M. Osborne, G. Smyth, A modified prony algorithm for fitting sums of exponential functions, SIAM Journal on Scientific and Statistical Computing 16 (1995) 119–138.
[42] D. Karlis, E. Xekalaki, Robust inference for finite Poisson mixtures, Journal of Statistical Planning and Inference 93 (2001) 93–115.