ELSEVIER

Computer Communications 24 (2001) 1400–1410

computer communications

www.elsevier.com/locate/comcom

Research

Multifractal modeling of counting processes of long-range dependent network traffic

Jianbo Gao*, Izhak Rubin

Department of Electrical Engineering, University of California, Los Angeles, CA 90095, USA

Received 5 May 2000; revised 19 December 2000; accepted 20 December 2000

Abstract

Source traffic streams as well as aggregated traffic flows often exhibit long-range-dependent (LRD) properties. In this paper, we study traffic streams through their counting process representation. We first study the condition for the measured LRD traffic, as described by the interarrival time and packet size sequences, to be sufficiently well approximated by a synthesized stream formed by recording the counting state of the traffic at the start of each time slot. We then demonstrate that the burstiness of the counting processes is not well characterized by the Hurst parameter. We model a counting process by constructing a multiplicative multifractal process, which contains only one or two parameters. We study the LRD property of such processes, and show that the model has well-defined burstiness descriptors, and are easy to construct. We consider a single server queueing system, which is loaded, on one hand, by the measured processes, and, on the other hand, by properly parameterized multifractal processes. In comparing the system-size tail distributions, we demonstrate our model to effectively track the behavior exhibited by the system driven by the actual traffic processes. Our study may help resolve a hot debate on the modeling of an often used trace of VBR video traffic.

Keywords: Network traffic modeling; Long-range-dependence; LAN, WAN, WWW, and VBR video traffic; Multifractal

1. Introduction

Recent analysis of high-quality traffic measurements have revealed the prevalence of long-range-dependent (LRD) (or self-similar) features in traffic processes loading packet switching communications networks. Included are local area networks (LANs) [10], wide area networks (WANs) [13], variable-bit-rate (VBR) video traffic [1,8], and world wide web (WWW) traffic [2].

With LRD traffic measured in many data networks, two related questions arise. One is how to parsimoniously model LRD traffic? The other is: what is the impact of LRD traffic on network performance? A distinguished issue along the first line is the hot debate on the modeling of LRD VBR video traffic. The finding of the LRD features in VBR video traffic [1,8] is thought to imply that Markovian processes may not be effective in modeling measured video traffic. However, Heyman and Lakshman [9] showed that Markovian processes with suitable degree of complexity can fit the system-size tail distribution quite well. We show in this paper that a key root of the underlying modeling difficulty relates to the inefficiency involved in using the Hurst parameter to characterize the burstiness of LRD traffic.

To gain an understanding of the first question, we shall base our study on carrying out queueing performance analysis, using both the measured and the modeled traffic processes to drive a queueing system. For the modeling of LRD traffic, an issue of much interest is whether and how multifractal can be employed to model the LRD feature of measured traffic [4,14]. Using Telcordia's LAN and WAN traffic trace data, we have shown [5] that the associated interarrival time series and packet length sequences are LRD, and are multifractals over certain finite time scale ranges. We have developed a method [5] to model the interarrival time series and the packet length sequences using two multiplicative multifractals.

Network traffic is often measured by collecting interarrival-time and packet-length statistics. For reference purposes, we refer to such a description as the customary model for network traffic. Aggregated traffic flows measured at a network node are presented as a stochastic counting process. The counting process is a more compact representation of a network traffic process. We examine in this paper whether a counting process can sufficiently well represent a network traffic flow. If so, we study here whether the counting process can be well approximated by a

* Corresponding author.

E-mail addresses: jbgao@ee.ucla.edu (J. Gao), rubin@ee.ucla.edu (I. Rubin).

0140-3664/01/$ - see front matter
PII: S0140-3664(01)00297-3

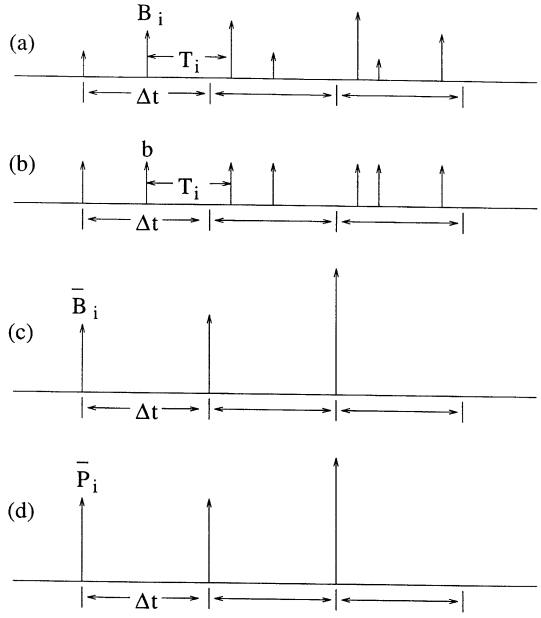

Fig. 1. Schematics for the pattern of (a) Traffic B, (b) Traffic P, (c) Traffic $\bar{B}$ and (d) Traffic $\bar{P}$.

multiplicative multifractal. We demonstrate here that a counting process can indeed sufficiently well represent LRD traffic flow, provided that the length of the time slot used to obtain the counting process is not longer than the mean packet delay time. The key contribution of this paper is to demonstrate that LRD counting processes, as in the case for interarrival time series and the packet length sequences of a network traffic [5], can be well approximated by multiplicative multifractals. Furthermore, we demonstrate here that the burstiness features of LRD traffic may not be properly characterized by the Hurst parameter. This finding implies that the LRD nature of many observed VBR video traffic streams does not by itself imply that such a process is too bursty to be properly modeled by a Markovian process. Indeed, we have found that among the four types of measured traffic streams studied here, namely, LAN, WAN, WWW, and VBR video traffic processes, video traffic is the least bursty. Hence, Markovian models can be quite effective in describing the statistics of such video traffic streams [9]. We show here that multiplicative multifractal processes provide another effective alternative for modeling such LRD processes.

We have chosen four different types of measured LRD network traffic: LAN, WAN, WWW, and VBR video traffic processes, for use in this study. The LAN and VBR video traffic data has been obtained from Telcordia. They are denoted as pAug.TL and MPEG.data, respectively[1]. The LAN traffic pAug.TL contains 1 million points representing measured values for packet arrival time stamps and packet sizes. It covers a time span of 3142.8 s. The video data consists of 174,136 integers, representing the number of bits per video frame (at 24 frames/s for approximately 2 h). The WAN and WWW trace data were collected on the FDDI ring of the UCLA campus backbone. Two such trace data flows, denoted as Sample-B (WAN) and S3p80 (WWW), will be used in this study. Sample-B represents measured values for 3.7 million arrival time stamps and packet sizes. It covers a time span of 989.6 s. S3p80 represents measured values for 2 million arrival time stamps and packet sizes. It covers a time span of 782.0 s. We demonstrate here that all of the traffic processes considered here can be readily and effectively modeled as multiplicative multifractal counting processes.

The remaining of the paper is organized as follows. In Section 2, we consider, first, how to represent a LRD traffic by its counting process, and second, the condition for this representation to be valid. We point out a limitation involved in using the Hurst parameter to characterize the burstiness of LRD traffic, and discuss its implications to the modeling of VBR video traffic processes. In Section 3, we outline the procedure used for constructing a multiplicative multifractal. We then discuss the LRD property of such processes. In Section 4, we show that the burstiness exhibited by a multifractal traffic process can be well expressed in terms of certain model descriptors. This feature enables us to develop a systematic approach for modeling measured LRD counting processes. In Section 5, we consider a single server queueing system that is loaded, on one hand, by a measured LRD process, and, on the other hand, by a properly parameterized multiplicative multifractal process. In comparing the system-size tail distributions of both systems, we demonstrate our model to effectively track the behavior exhibited by the system driven by measured traffic process. Conclusions are drawn in Section 6.

2. Representing a LRD traffic by counting processes

Two random processes, the interarrival time series, $\{T_i\}$, where T_i denotes the ith interarrival time between two successive packet arrivals, and the packet length sequence, $\{B_i\}$, where B_i represents the length of the ith packet, are used to define a traffic process. This is schematically shown in Fig. 1(a). We denote this traffic characterization by Traffic B.

For many network systems, such as ATM networks, messages are segmented into network layer PDUs (i.e., packets) that have relatively short maximum packet lengths. For such networks, packets are virtually of fixed size. Hence, we define a traffic presentation model under which we set the ith packet length (B_i) equal to the mean of the packet length, $b = \sum_{i=1}^{n} B_i/n$, while $\{T_i\}$ is used as before. This is schematically shown in Fig. 1(b). We denote this traffic pattern by Traffic P.

Two counting random processes can also be defined. These processes represent the number of bits and the

[1] This is available at ftp.telcordia.com under the directory /pub/world/wel/lan_traffic and /pub/vbr.video.trace.

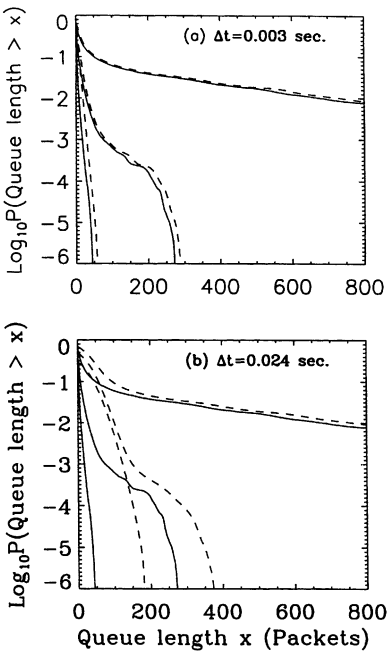

Fig. 2. (a) System-size tail probabilities obtained when Traffic B (solid lines) and Traffic $\bar{B}$ (dashed lines) of the WWW traffic S3p80 are used to drive a queueing system. Three curves, from top to bottom, correspond to $\rho = 0.7$, 0.5, and 0.3, respectively. The length of the time slot Δt used to generate Traffic $\bar{B}$ is 0.003 s; (b) same as (a) except now the length of the time slot is 0.024 s.

number of packets arriving every Δt seconds. We denote them by $\bar{B} = \{\bar{B}_i,\ i = 1, 2, 3, \ldots\}$ and $\bar{P} = \{\bar{P}_i,\ i = 1, 2, 3, \ldots\}$, respectively. The variable $\bar{B}_i$ represents the number of packet bits arriving during the ith interval. The variable $\bar{P}_i$ represents the number of packets arriving across the ith time slot, each packet of size b bits. As is evident, these processes can be thought of as being derived from Traffic B and Traffic P, respectively.

For the LAN, WAN, and WWW traffic trace data described in Section 6, we obtain 2^{18} intervals for the underlying counting process models. Hence, $\Delta t = 3.77 \times 10^{-3}$, 1.20×10^{-2}, and 2.98×10^{-3} s for WAN data Sample-B, LAN data pAug.TL, and WWW data S3p80, respectively.

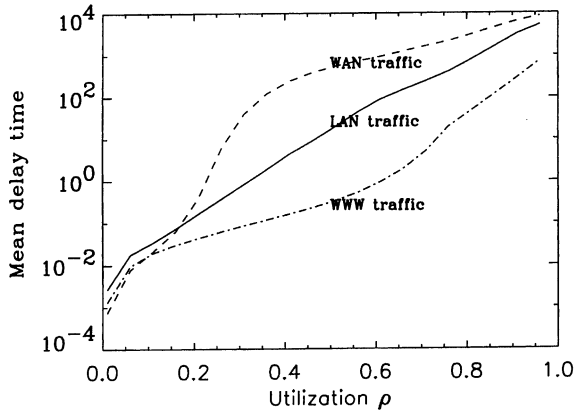

Fig. 3. Ratio (in logarithmic scale) of the mean delay time and the length of the time slots chosen for constructing the counting processes vs. the utilization ρ.

To associate a time of arrival with the data included in each interval, we select the following simple model. We record the counts $\bar{B}_i$ and $\bar{P}_i$ to occur at the start of the ith time slot (we shall observe later that one can use any other distribution of the counts across the interval), as shown schematically in Fig. 1(c) and (d). We denote the resulting processes as Traffic $\bar{B}$ and Traffic $\bar{P}$. Note the VBR video traffic process presented in Section 1 is already in the form of Traffic $\bar{B}$.

As demonstrated below, when the system is loaded by traffic processes exhibiting LRD behavior, for many network applications and regular queueing performance analysis purposes, Traffic $\bar{B}$ and Traffic $\bar{P}$ are equivalent to Traffic B and Traffic P, respectively.

Consider a single server queueing system using a FIFO service discipline and an infinite buffer. To compare the performance of the system when it is loaded by Traffic B and Traffic $\bar{B}$ or by Traffic P and Traffic $\bar{P}$, we proceed as follows. We use them to drive the queueing system, and then compare the system-size tail probabilities under different utilization levels. We have observed that for LAN traffic data pAug.TL and WAN traffic data Sample-B, when the utilization level is $\rho \geq 0.3$, the system-size tail probabilities are the same for both queueing systems. For WWW traffic S3p80, however, we do observe small differences in the system-size tail probabilities. This is shown in Fig. 2(a), where the solid and dashed curves are obtained when Traffic B and Traffic $\bar{B}$ of S3p80 are used to drive the queueing system, respectively. Three curves, from top to bottom, correspond to utilization levels $\rho = 0.7$, 0.5, and 0.3. Since even in this worst case, the difference in the system-size tail probabilities between the two queueing systems can still be ignored, we can safely conclude that Traffic $\bar{B}$ and Traffic $\bar{P}$ can sufficiently well represent Traffic B and Traffic P, for the traffic data considered here.

At first sight, this result may sound counterintuitive. After all, Traffic $\bar{B}$ is not identical to Traffic B. Neither is Traffic $\bar{P}$ is identical to Traffic P. How can we understand this result?

Note that the difference between the mean packet delay times for the counting process representation and the original traffic can be at most Δt, the length of the time slot chosen for obtaining Traffic $\bar{B}$ or Traffic $\bar{P}$. Hence, when Δt is much smaller than the mean packet delay time, which is true for LRD traffic when the utilization is not too low, then counting processes can sufficiently well represent their corresponding network traffic, at least in terms of system performance. This is the underlying mechanism for the above observation (Fig. 2(a)) to be true. This understanding immediately suggests that, if we approximate Traffic B of WWW traffic S3p80 by Traffic $\bar{B}$ using a longer time slot, then the approximation will be worse. This is indeed the

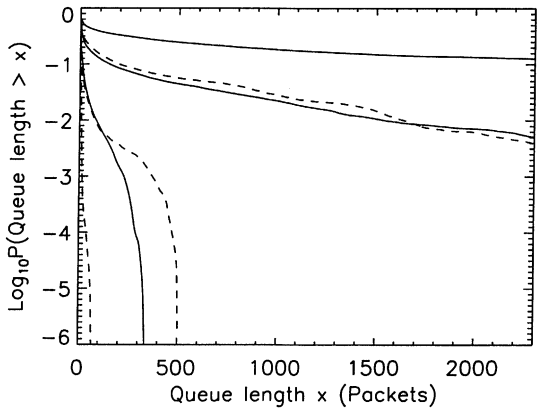

Fig. 4. System-size tail probabilities obtained when Traffic $\bar{B}$ (solid lines) and Traffic $\bar{P}$ (dashed lines) of LAN traffic pAug.TL are used to drive a queueing system. Three curves, from top to bottom, correspond to $\rho = 0.7$, 0.5, and 0.3, respectively.

case, as shown in Fig. 2(b), where $\Delta t'$ is eight times the value of Δt ($\Delta t' = 2.40 \times 10^{-2}$ s). As before, the solid and dashed curves are obtained when Traffic B and Traffic $\bar{B}$ of S3p80 are used to drive the queueing system, respectively. The three curves, from top to bottom, correspond to utilization levels $\rho = 0.7$, 0.5, and 0.3. To be quantitative, we compute the dimensionless mean delay time (in units of the length of the time slot) vs. the utilization level ρ for the LAN, WAN, and WWW traffic data considered here. This is shown in Fig. 3. We observe that for high utilization levels, the mean delay time is orders of magnitude longer than the length of the time slots. This result points out that to construct a traffic process from $\{\bar{B}_i\}$ or $\{\bar{P}_i\}$ time series, we can actually record them anywhere inside the corresponding time slot. Note this result is qualitatively equivalent to a result obtained by Erramilli et al. [3], where they showed that locally shuffling the order of arriving packets leads to only a slight change in the mean delay time.

Fig. 3 also explains why the WWW traffic is the worst

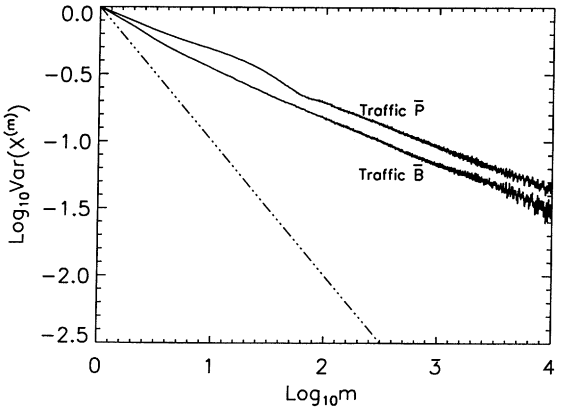

Fig. 5. Variance–time plots (in logarithmic scale) for Traffic $\bar{B}$ and Traffic $\bar{P}$ of LAN traffic pAug.TL. The dash dot dot dot line is the 'Reference' line with slope −1.

case. For light and medium utilization levels, the mean delay time is actually smaller than the length of the chosen time slot Δt. This results in the difference in the system-size tail probabilities shown in Fig. 2(a). If we choose $\Delta t' = 8\Delta t$, then for low and medium utilization levels, the mean delay time is much smaller than $\Delta t'$, and Traffic $\bar{B}$ does not represent Traffic B (Fig. 2(b)) well. However, for high utilization level, $\rho = 0.7$, a model using either time slot Δt or $\Delta t'$ yields good result (Fig. 2(a) and (b)).

Before ending this section, we point out a limitation of using the Hurst parameter for characterizing the burstiness of the LRD traffic. It is argued [10] that a larger H value corresponds to more bursty traffic. For example, the heavy-tailed ON/OFF model exhibits such behavior when the burstiness is defined by the mean queue size [12]. Recently, this idea (i.e., using H as an index of burstiness) has been slightly corrected by Neidhardt and Wang [11] based on the study of the fractional Brownian motion model. Our observation is that a burstier traffic is not necessarily associated with a larger value for the Hurst parameter. This point is readily demonstrated by considering, for example, Traffic $\bar{B}$ and Traffic $\bar{P}$ of LAN traffic pAug.TL. Fig. 4 shows the system-size tail probabilities (used henceforth as a measure of burstiness) when Traffic $\bar{B}$ (solid lines) and Traffic $\bar{P}$ (dashed lines) are used to drive the queueing system. Three curves, from top to bottom, correspond to utilization levels $\rho = 0.7$, 0.5, and 0.3. While we observe that both Traffic $\bar{B}$ and Traffic $\bar{P}$ are very bursty, Traffic $\bar{B}$ is more bursty. Using the variance-time relation [10]: $\mathrm{Var}(X^{(m)}) \sim m^{2(H-1)}$ to estimate the Hurst parameter H for Traffic $\bar{B}$ and Traffic $\bar{P}$, we find, however, that the value for H estimated from Traffic $\bar{P}$ is quite close to that estimated from Traffic $\bar{B}$, as shown in Fig. 5. This result suggests that video traffic processes that have LRD properties may not be characterized as too bursty flows (in terms of their induced queueing tail behavior). This is indeed so, as will be shown in Section 5. In Section 3, we demonstrate that the key parameter(s) for multifractal traffic streams (i.e., those characterizing the multiplier functions) can be used to better indicate the burstiness feature of a measured traffic process.

3. Multifractal modeling of counting processes

In this section, We recapitulate the procedure for constructing a multiplicative multifractal from a multiplier function and discuss the LRD properties of the multiplicative multifractal counting processes.

3.1. Construction of multiplicative multifractals

Consider a unit interval. Associate it with a unit mass. Divide the unit interval into two (say, left and right) segments of equal length. Also partition the mass into two fractions, r and $1 - r$, and assign them to the left and right segments, respectively. The parameter r is in general a random variable, governed by a probability density function

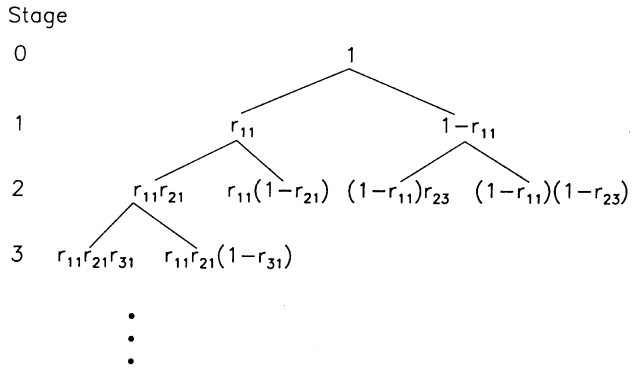

Fig. 6. A schematic illustrating the construction rule of a multiplicative multifractal.

$P(r)$, $0 \le r \le 1$. The fraction r is called the multiplier, and $P(r)$ is called the multiplier function. Each new subinterval and its associated weight (or mass) are further divided into two parts following the same rule. This procedure is schematically shown in Fig. 6, where the multiplier r is written as r_{ij}, with i indicating the stage number. Note the scale (i.e., the interval length) associated with stage i is 2^{-i}. We assume that $P(r)$ is symmetric about $r = 1/2$, and has successive moments $\mu_1, \mu_2, \ldots$. Hence r_{ij} and $1 - r_{ij}$ both have marginal distribution $P(r)$. The weights at the stage N, $\{w_n, n = 1, \ldots, 2^N\}$, can be expressed as $w_n = u_1 u_2 \ldots u_N$, where u_l, $l = 1, \ldots, N$, are either r_{ij} or $1 - r_{ij}$. Thus, $\{u_i, i \ge 1\}$ are independent identically distributed random variables having multiplier function $P(r)$. When $w_n(N)$ is interpreted as the loading to a network (representing the total count of message units) in a time slot of length $2^{-N}T$, where T is the total time period one is interested in, then this process becomes a counting traffic process model. The multifractality of the multiplicative process refers to the fact that $M_q(\epsilon) = E(\sum_{n=1}^{2^N} (w_n(N))^q) \sim \epsilon^{\tau(q)}$, with $\epsilon = 2^{-N}$, $\tau(q) = -\ln(2\mu_q)/\ln 2$ [6].

3.2. Properties of multiplicative multifractals

For the weights at stage N, we prove the following properties to hold (for a shorter account of these results, see also [6]):

(i)

$$E(w) = E(w_n) = E(u_1 u_2 \ldots u_N) = 2^{-N}, \qquad n = 1, \ldots, 2^N. \tag{1}$$

(ii)

$$E(w^q) = E((u_1 u_2 \ldots u_N)^q) = \mu_q^N. \tag{2}$$

In particular

$$E(w^2) = \mu_2^N \tag{3}$$

and

$$\mathrm{Var}(w) = \mathrm{Var}(w_n) = \mu_2^N - 2^{-2N}, \qquad n = 1, \ldots, 2^N. \tag{4}$$

(iii)

$$\mathrm{Var}(W^{(m)}) = \mu_2^N (4\mu_2)^{-k} - 2^{-2N}, \tag{5}$$

where $W^{(m)} = (w_{im-m+1} + \cdots + w_{im})/m$, $m = 2^k$, $k = 1, 2, \ldots$, and $i \ge 1$. This is proven by expressing $W^{(m)} = 2^{-k}x$, where x is a weight at stage $N - k$.

Eq. (5) expresses a variance-time relation. For LRD traffic [10], $\mathrm{Var}(W^{(m)}) \sim m^{2H-2}$, where $1/2 < H < 1$ is the Hurst parameter. For multiplicative multifractal processes, when N is large and $\mu_2 > 0$, the term $\mu_2^N (4\mu_2)^{-k}$ dominates. When the term 2^{-2N} in Eq. (5) is dropped, the functional variation of $\log \mathrm{Var}(W^{(m)})$ vs. $\log m$ is linear. The resulting slope, $-\log(4\mu_2)/\log 2$, provides an estimate of $2H - 2$. A moment of thinking will convince us that this slope is an upper bound for $2H - 2$. Hence

$$H \le -\tfrac{1}{2}\log_2 \mu_2. \tag{6}$$

Since the multiplier distribution $P(r)$ is defined for $0 \le r \le 1$, and is symmetric about 1/2, hence its mean is 1/2, and its variance is upper bounded by 1/4. We thus have $(1/2)^2 \le \mu_2 \le (1/2)^2 + (1/4)$. Therefore $1/2 \le H \le 1$, with $H = 1$ corresponding to deterministic time series (i.e., $P(r) = \delta(r - 1/2)$). We thus observe that a multiplicative multifractal traffic stream also possesses LRD property.

Let us check how good the linearity defined by Eq. (5) is. For this purpose, we consider three different functional forms for the multiplier function, namely, double exponential with parameter α_e

$$P(r) \sim \mathrm{e}^{-\alpha_e |r - 1/2|}. \tag{7}$$

Gaussian with parameter α_g

$$P(r) \sim \mathrm{e}^{-\alpha_g (r - 1/2)^2} \tag{8}$$

and a function being of the form

$$P(r) = \begin{cases} q + p\delta(r - 1/2), & 1/2 - d \le r \le 1/2 + d, \\ 0, & \text{otherwise}, \end{cases} \tag{9}$$

where $0 \le d \le 1/2$. The last function is schematically shown in Fig. 7. Note that the three parameters d, p, and q

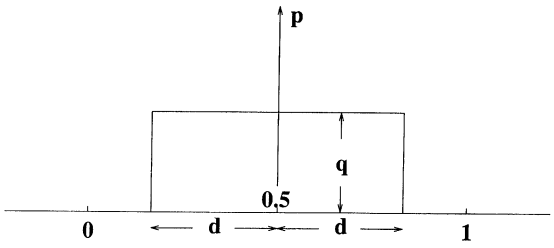

Fig. 7. A schematic showing the form for the multiplier function as described by Eq. (9).

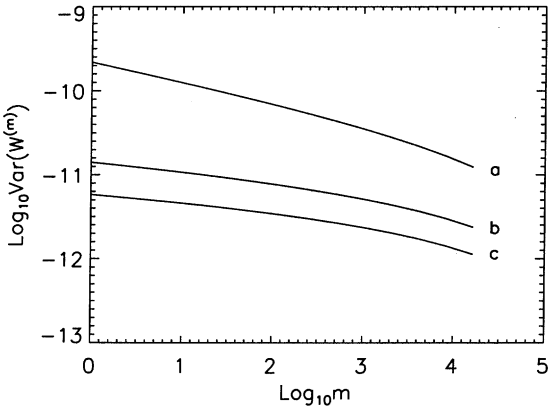

Fig. 8. log$Var(W^{(m)})$ vs. log m curves for (a) $\alpha = 10$, (b) $\alpha = 50$, and (c) $\alpha = 100$.

are related by the equation, $p + 2qd = 1$. Hence, the function contains two independent parameters. We shall choose p and d as the two basic parameters. Note one may introduce a parameter equivalent to d for the functions characterized by Eqs. (7) and (8). Due to exponential decay of Eqs. (7) and (8), however, such a parameter is not too interesting. Also note that we may rewrite Eqs. (7) and (8) as $P(r) \sim e^{-\alpha|r-1/2|^{\beta}}$, with $\beta = 1$ for the double exponential, and 2 for the Gaussian. Hence, Eqs. (7) and (8) really contain two parameters with a prefixed parameter β.

We generate a number of realizations of multiplicative processes with $P(r)$ given by one of the above three forms. We then compute the variance-time relation from the generated time series. Some examples of the variance-time curves are shown in Fig. 8, with $P(r)$ given by Eq. (8) and $N = 18$. We observe that indeed the variance-time curves are approximately linear, with the degree of linearity being better for larger μ_2.

We can furthermore check how tight the bound determined by Inequality (6) is by estimating H from the

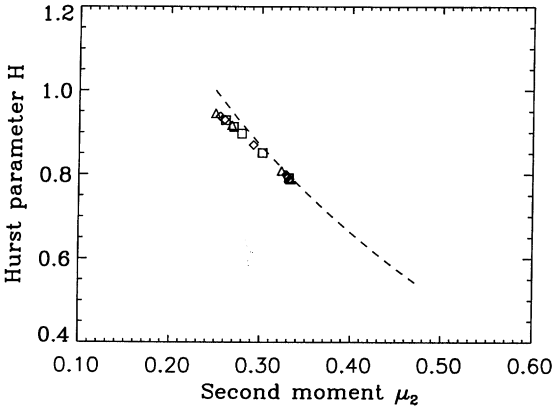

Fig. 9. Hurst parameter vs. the second moment μ_2. The dashed line is computed according to the right-hand side of Inequality (6), while the points designated by squares, diamonds and triangles are directly estimated from multiplicative processes with their multiplier distributions given by Eqs. (7)–(9).

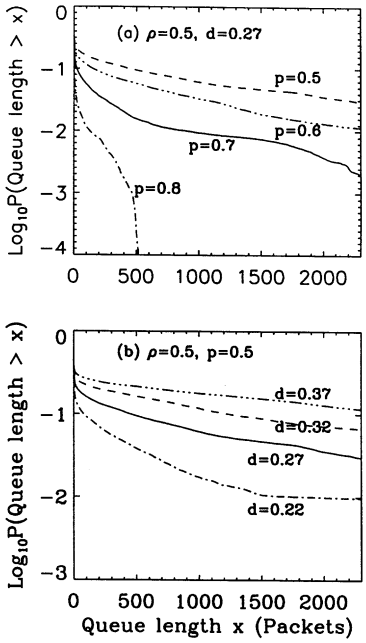

Fig. 10. For the fixed utilization level $\rho = 0.5$, the system-size tail probabilities obtained when different multiplicative multifractal traffic are used to drive a queueing system. The parameters for the different multifractal traffic are (a) $d = 0.27$, and $p = 0.5$, 0.6, 0.7, 0.8; and (b) $p = 0.5$, and $d = 0.37$, 0.32, 0.27, 0.22.

variance-time curves and comparing it with the right-hand side of Inequality (6). Fig. 9 shows such a comparison, where the dashed curve is generated from equation $H = -(1/2)\log_2\mu_2$. The points denoted by diamonds, triangles, and squares are estimated from multiplicative processes with $P(r)$ given by Eqs. (7)–(9), respectively. We observe that the bound given by Inequality (6) is very tight, especially for not too small values of μ_2. We shall further show in Section 4 that a burstier multifractal traffic is associated with a larger value for μ_2, hence a smaller value for the Hurst parameter.

4. Burstiness of the multiplicative multifractal counting processes

In this section, we discuss how the parameters in the multiplier functions described by Eqs. (7)–(9) control the burstiness of the multiplicative multifractal traffic.

We first consider the physical meaning of the α parameters in Eqs. (7) and (8). Consider $P(r) \sim e^{-\alpha|r-1/2|^{\beta}}$, with $\alpha \to \infty$. Then $P(r) \to \delta(r - 1/2)$. If we choose $P(r) = \delta(r - 1/2)$, then all the weights are identical. They constitute a non-bursty (or deterministic) traffic. This indicates that the burstiness of the multiplicative multifractal traffic described by Eqs. (7) and (8) decreases with the α parameter. For more details on this feature, we refer to our earlier paper [5].

The physical meaning of the parameters p and d in Eq. (9)

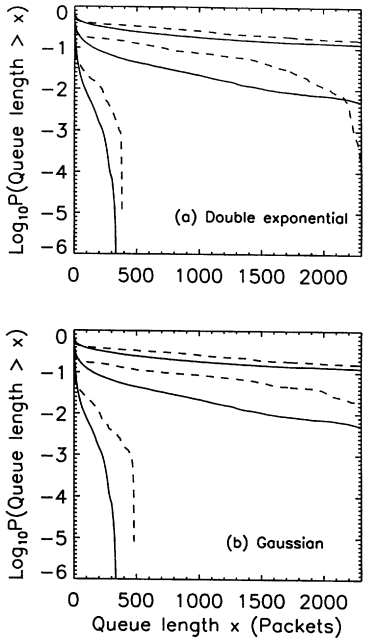

Fig. 11. System-size tail probabilities obtained when Traffic $\bar{B}$ of pAug.TL (solid lines both in (a) and (b)), and multifractal Traffic Processes constructed from a multiplier function being (a) double exponential with parameter $\alpha_e = 11$, and (b) Gaussian with parameter $\alpha_g = 36$, are used to drive a queueing system. Three curves, from top to bottom, correspond to $\rho = 0.7$, 0.5, and 0.3, respectively.

is as follows. Parameter p indicates the mean of the counting process; parameter d describes the variation of the traffic around the mean function. A direct consequence of this physical interpretation is that the burstiness of the modeled traffic increases with d when p is fixed; and decreases with p when d is fixed. This property can be readily demonstrated by feeding the multifractal counting processes into a queueing system, and examining the system-size tail probabilities. Fig. 10(a) shows the system-size tail probabilities when d is fixed to be 0.27 and the utilization is set equal to $\rho = 0.5$. The four illustrated curves, from top to bottom, correspond to $p = 0.5$, 0.6, 0.7 and 0.8. The results for a fixed value of $p = 0.5$ and $\rho = 0.5$ are shown in Fig. 10(b), where four curves, from top to bottom, correspond to $d = 0.37$, 0.32, 0.27, and 0.22, respectively. Clearly, we observe that the burstiness of the multifractal counting processes increases with d when p is fixed, and decreases with p when d is fixed.

The above burstiness behavior implies that a burstier multifractal traffic is associated with a larger value for μ_2. Because $H \approx -(1/2)\log_2\mu_2$ (Inequality(6) and Fig. 9), hence a burstier multifractal traffic is associated with a smaller value for the Hurst parameter. This contrasts sharply with the heavy-tailed ON/OFF model or the fractional Brownian motion model, where a burstier traffic is associated with a larger value for the Hurst parameter.

For our subsequent discussion on the modeling of measured traffic, we need to first select a multiplier function. For example, we can use one of the multiplier functions described by Eqs. (7)–(9). For this purpose, we consider the system-size tail probabilities of a single server queueing system, which is loaded, on one hand, by Traffic $\bar{B}$ of the LAN traffic pAug.TL, and on the other hand, by the multiplicative multifractal traffic characterized by the multiplier function described by Eqs. (7)–(9) with suitable corresponding parameters (the discussion on the selection of the parameters will be postponed to Section 5). The results are shown in Fig. 11(a) and (b) for the multiplier functions described by Eqs. (7) and (8), and Fig. 12(a) for the multiplier function described by Eq. (9). The solid (and dashed) curves display the results obtained when Traffic $\bar{B}$ of pAug.TL (and its multiplicative multifractal model) are used to drive the queueing system. The three depicted curves, from top to bottom, correspond to utilization levels $\rho = 0.7$, 0.5, and 0.3, respectively. The parameters used to generate the multiplicative multifractal processes are, $\alpha_e = 11$, for Fig. 11(a), $\alpha_g = 36$, for Fig. 11(b), and $(p, d) =$

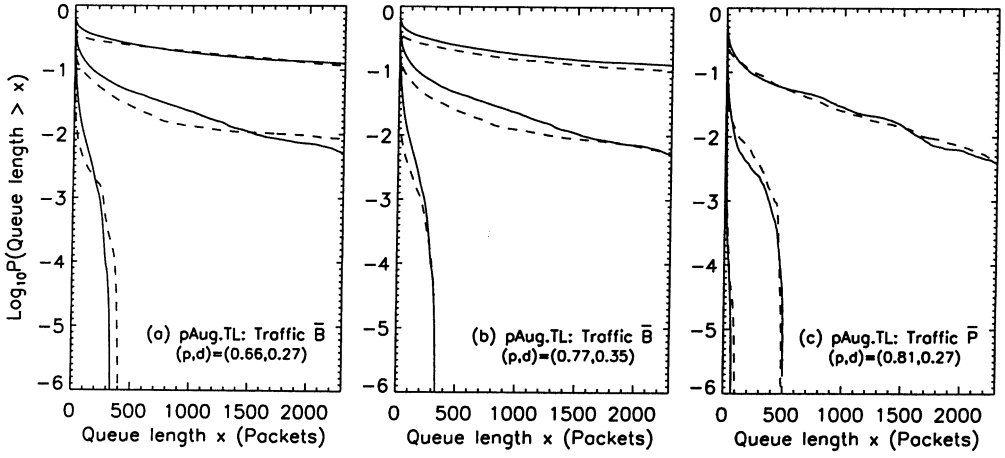

Fig. 12. System-size tail probabilities obtained when (a, b) Traffic $\bar{B}$ and (c) Traffic $\bar{P}$ (solid curves) of pAug.TL, and multifractal traffic processes constructed from a multiplier function characterized by Eq. (9) with parameter pair (a) $(p, d) = (0.66, 0.27)$, (b) $(p, d) = (0.77, 0.35)$, and (c) $(p, d) = (0.81, 0.27)$ (all dashed curves), are used to drive the queueing system. Three curves, from top to bottom, correspond to $\rho = 0.7$, 0.5, and 0.3, respectively.

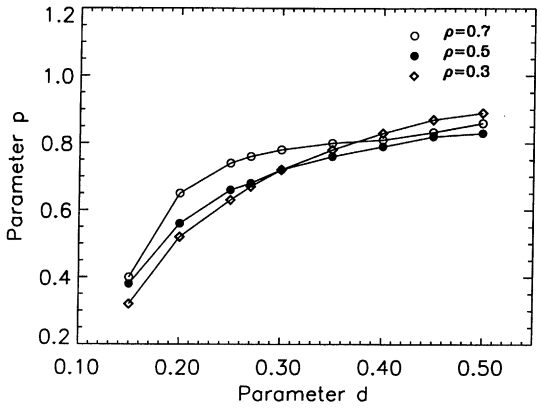

Fig. 13. Equi-burstiness parameter curves corresponding to pAug.TL for three utilization levels, $\rho = 0.3$, 0.5, and 0.7.

(0.66, 0.27), for Fig. 12(a). We first note that all three fittings to the system-size tail probabilities, each using a different multiplier function, are good. This indicates that, given a careful selection of the multiplier function parameter(s), the exact functional form for the multiplier function may not be important in fitting the system-size tail probabilities.

For the traffic process we modeled in our study, the fitting that uses the multiplier function described by Eq. (9) yields the best result. This is understandable, since Eqs. (7) and (8) only contain one adjustable parameter, while Eq. (9) contains two adjustable parameters, hence it is more flexible. Hence, we select the multiplier function to be characterized by Eq. (9) for modeling measured traffic studied in the following section.

5. Multifractal modeling of measured counting processes

The system-size tail distribution serves as a key performance measure in the engineering, analysis, and design of communication network systems. This statistic is readily measurable even without collecting a traffic trace data. Hence, our purpose for modeling a measured counting process is to find a single parameter pair (p,d) for the multiplier function so that when the multiplicative multifractal counting process is used to drive a queueing system, the system-size tail probabilities under light, medium, and high loading conditions are simultaneously very close to those of a queueing system driven by the measured traffic trace data. Note that the system-size tail distribution provides an accurate measure on the degree of the burstiness of a network traffic. If we can indeed find a single parameter pair (p,d) for the multiplier function, then we have found a simple and accurate method of characterizing the burstiness of a traffic.

Since there is no a priori guarantee that a single parameter pair (p,d) would exist such that the queueing system driven by the multifractal traffic under different loading conditions would exhibit similar system-size tail behavior to a queueing system driven by a measured traffic process, we need to develop a systematic approach to check whether our goal is achievable. And if it is achievable, will the systematic approach also tell us how to select the proper parameter pair (p,d)?

Such systematic approaches do exist. We describe a simple one here. Recall that the burstiness of the multifractal traffic increases with d when p is fixed, and decreases with p when d is fixed. This property implies that if a specific parameter pair (p_0, d_0) fits some tail distribution of a system under certain loading condition $\rho = \rho_0$, then we should also be able to find different parameter pairs (p,d), with $p > p_0$, $d > d_0$, or $p < p_0$, $d < d_0$, to provide similarly good fit to the system-size tail distribution under that particular loading condition. These different parameter pairs would trace out a curve in the parameter plane. Since different parameter pairs on this curve correspond to the same degree of burstiness of a measured traffic (under the specified loading condition), we call such a curve equi-burstiness parameter curve under the loading condition. If different equi-burstiness parameter curves under different loading conditions are very close together at certain parameter pair values, then clearly our goal is achievable, with the proper parameter pair values given by where the equi-burstiness parameter curves are very close together.

Hence, the problem is reduced to whether an equi-burstiness parameter curve exists for the network under a particular loading condition, and how can we find this curve if it does exist. The first part of the question is easy to answer if we note the following. The range of the burstiness of the multifractal traffic is lower-bounded by the non-bursty traffic characterized by the multiplier function $P(r) = \delta(r - 1/2)$ (corresponding to the parameter pair $(p, d) = (1, 0)$), and upper-bounded by the most bursty traffic characterized by the multiplier function being uniform on the unit interval [0,1] (corresponding to the parameter pair $(p, d) = (0, 1/2)$). If the burstiness of a measured traffic belongs to this range, then we conclude there is at least one parameter pair (p,d) to generate a multifractal traffic having the same degree of burstiness as the measured traffic. This then implies, by an earlier argument, that an equi-burstiness parameter curve exists.

If an equi-burstiness parameter curve exists, then one has two ways to find that curve. One can proceed by trial and error, guided by the property that the burstiness of the multifractal traffic increases with d when p is fixed, and decreases with p when d is fixed. Alternatively, one can first simply assume the mean of the multiplicative multifractal counting process to be one unit. Then one can construct multifractal counting processes parameterized by a series of different pairs of (p,d), and use them to drive a queueing system to obtain system-size tail probabilities under different utilization levels. One then saves those system-size tail probabilities corresponding to different parameter pairs (p,d) as a data base. To find a desired parameter pair (p,d) for a measured traffic trace data, one need simply to use the

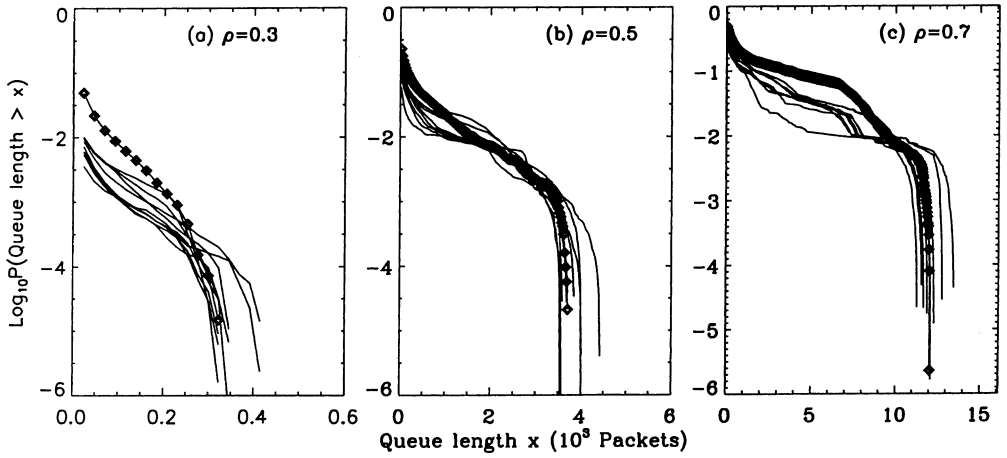

Fig. 14. System-size tail distributions for three loading conditions. The thick curves (symboled as diamonds) are obtained when the measured traffic trace data pAug.TL is used to drive a queueing system. Other solid curves are obtained when the multiplicative multifractal traffic processes with parameter pair values indicated as open circles, dark circles, and diamonds in Fig. 13 are used to drive the queueing system.

measured trace data to drive a queueing system to obtain the system-size tail probabilities, and then compare them with the data base. It should be obvious that when one has many measured traffic trace data to model, the second method is far superior.

We illustrate the above procedure by modeling the Traffic $\bar{B}$ of the LAN traffic pAug.TL in some detail. We choose $\rho = 0.3$, 0.5, and 0.7 as representatives of low, medium, and high loading conditions. Fig. 13 shows the equi-burstiness parameter curves for the loading conditions considered. Under these loading conditions, the system-size tail distributions are shown in Fig. 14(a)–(c), where the thick curves (symboled as diamonds) are obtained when the measured traffic pAug.TL is used to drive the queueing system. Other solid curves are obtained when the multiplicative multifractal traffic processes, with parameter pair values indicated as open circles, solid circles, and diamonds in Fig. 13, are used to drive the queueing system. We observe that the multifractal traffic indeed has the same degree of burstiness as the measured traffic under a specified loading condition. We also observe that the equi-burstiness parameter curves under different loading conditions are very close together. This suggests that multiple good parameter pairs (p,d) exist for the measured traffic pAug.TL. Recall that the overall system-size tail distributions obtained when the multifractal traffic is parameterized by $(p, d) = (0.66, 0.27)$ have been shown in Fig. 12(a). Another similarly good example is displayed in Fig. 12(b), where $(p, d) = (0.77, 0.35)$ is used to generate the multiplicative multifractal.

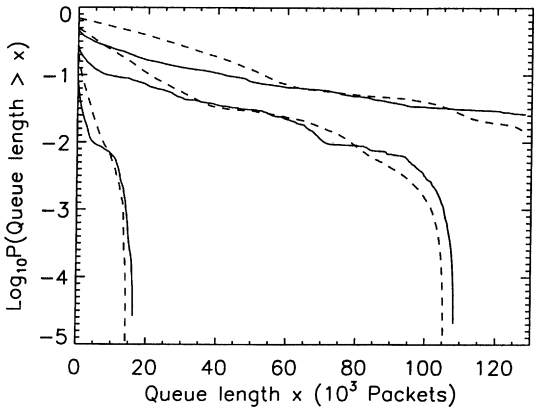

Fig. 15. System-size tail probabilities obtained when WAN traffic Sample-B (solid curves) and its corresponding multifractal traffic process (dashed curves) are used to drive the queueing system. Three curves, from top to bottom, correspond to $\rho = 0.7$, 0.5, and 0.3, respectively.

By the same procedure, we have found that multiple parameter pairs (p,d) also exist for other measured traffic processes. For example, $(p, d) = (0.81, 0.27)$, for the Traffic $\bar{P}$ of the LAN traffic pAug.TL; $(p, d) = (0.69, 0.45)$, for the Traffic $\bar{B}$ of WAN traffic Sample B; $(p, d) = (0.84, 0.22)$, for the *Traffic* $\bar{B}$ of WWW traffic S3p80; and $(p, d) = (0.73, 0.22)$, for the VBR video traffic. A different set of good parameter pairs will be given later when we discuss further on the characterization of the burstiness of measured traffic.

Fig. 12(c) shows a comparison in the system-size tail probabilities between queueing systems loaded by the Traffic $\bar{P}$ of pAug.TL, and the multiplicative multifractal traffic parameterized by $(p, d) = (0.81, 0.27)$. The solid and dashed curves display the results obtained when Traffic $\bar{P}$ of pAug.TL and its multiplicative multifractal model are used to drive the queueing system. Three curves, from top to bottom, correspond to utilization levels, $\rho = 0.7$, 0.5, and 0.3. Clearly, the queueing systems exhibit very similar system-size tail distributions when loaded by the measured traffic and the multiplicative multifractal traffic.

The burstiness of the Traffic $\bar{B}$ and Traffic $\bar{P}$ of pAug.TL is indicated by the parameter pairs $(p, d) = (0.66, 0.27)$, and

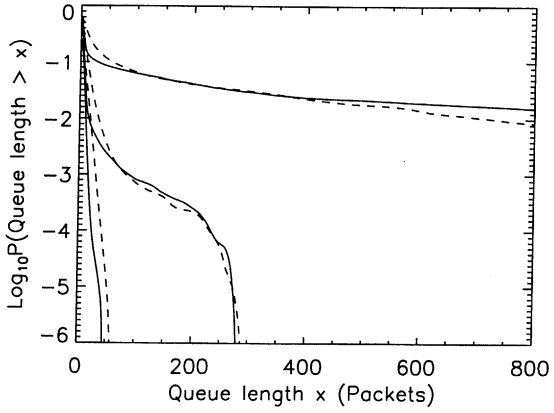

Fig. 16. System-size tail probabilities obtained when WWW traffic S3p80 (solid curves) and its corresponding multifractal traffic process (dashed curves) are used to drive the queueing system. Three curves, from top to bottom, correspond to $\rho = 0.7$, 0.5, and 0.3, respectively.

(0.81,0.27), respectively. According to the results shown in Fig. 10, we see that Traffic $\bar{B}$ is more bursty than Traffic $\bar{P}$. This is consistent with the result shown in Fig. 4. Hence, our multiplicative multifractal counting process model has indeed overcome the inconsistency problem associated with using the Hurst parameter to characterize the burstiness of traffic, as studied in Section 2.

Comparisons in the system-size tail probabilities between queueing systems loaded by the measured traffic Sample-B, S3p80, and MPEG.data, and their corresponding multiplicative multifractal models are shown in Fig. 15–17, respectively. The solid and dashed curves display the results obtained when Traffic $\bar{B}$ of the measured traffic and their corresponding multiplicative multifractal models are used to drive the queueing system. Three curves, from top to bottom, correspond to utilization levels, $\rho = 0.7$, 0.5, and 0.3. We observe that in all cases the queueing systems exhibit very similar system-size tail distributions when loaded

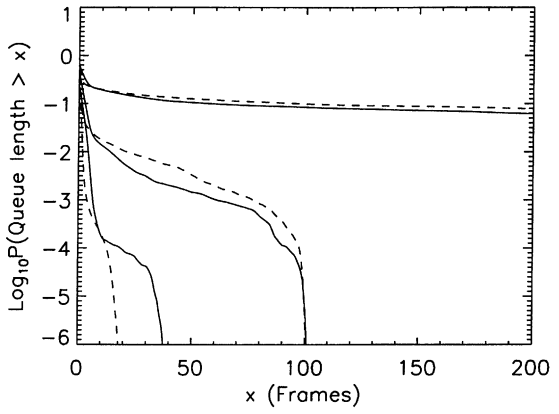

Fig. 17. System-size tail probabilities obtained when VBR video traffic MPEG.data (solid curves) and its corresponding multifractal traffic process (dashed curves) are used to drive the queueing system. Three curves, from top to bottom, correspond to $\rho = 0.7$, 0.5, and 0.3, respectively.

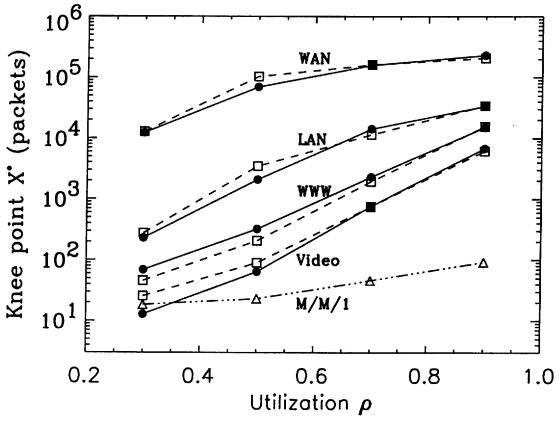

Fig. 18. Variation of the knee point, X^*, with the utilization level, ρ. Open squares connected by dashed lines are obtained when the measured traffic, WAN, LAN, WWW, and VBR video, is used to drive a queueing system. Open triangles connected by dash–dot–dot–dot curve are computed from a M/M/1 queueing system. Dark circles connected by solid lines are computed from multiplicative multifractal traffic processes.

by the measured traffic and the multiplicative multifractal traffic.

Note an interesting feature exhibited by Figs. 12, 15–17. For not too low utilization levels, the system size tail probabilities drop almost vertically after the buffer size exceeds a certain size, X^*. We call this special point on the system size tail probability curve the knee point. Knee points are of crucial importance for engineering purposes, since they characterize accurately the notion of burstiness of network traffic, and are easily measurable.

Fig. 18 shows the variation of the knee point with the utilization level. Points designated by open squares and connected by dashed lines are obtained when the measured traffic, WAN, LAN, WWW, and VBR video, are used to drive a queueing system. Note that the burstiness of these traffic processes decrease in the order WAN, LAN, WWW, and VBR video. For comparison, the knee points for a M/M/1 queueing system are also shown as open triangles connected by dash–dot–dot–dot curve. Note that M/M/1 queueing system is the least bursty when compared to the measured traffic processes.

As can be expected from the results shown in Figs. 12, 15–17, multiplicative multifractal traffic processes will give very similar knee point curves as their corresponding measured traffic processes. This is indeed so, as shown by solid circles connected by solid curves in Fig. 18.

6. Conclusions

We analyze traffic flow traces taken from LANs, WANs, and WWW. These traffic processes are described by their interarrival-time and packet-size sequences. They have been shown to exhibit LRD features. We show in this paper that these traffic streams are well represented by counting

process models. For this purpose, one must set the duration used to collect the traffic count to be not larger than the mean packet delay time. We also show that the burstiness of these LRD counting processes cannot be effectively characterized by the Hurst parameter. We then introduce a new model, the multiplicative multifractal process, to characterize the counting traffic process. We develop a number of properties exhibited by such processes. In particular, we prove them to have LRD properties. We show that a burstier multiplicative multifractal traffic process is associated with a smaller value for the Hurst parameter. This is in sharp contrast with the commonly held belief that a burstier LRD traffic is often associated with a larger value for the Hurst parameter. To simplify the modeling process, we select a multiplicative multifractal model that employs one or two basic parameters. We show that this model has well defined burstiness descriptors, and is easy to construct. We consider a single server queueing system that is loaded, on one hand, by the measured LAN, WAN, WWW, and VBR video traffic processes, and, on the other hand, by the corresponding properly parameterized multifractal processes. In comparing the system-size tail distributions, we demonstrate our model to effectively track the behavior exhibited by the system driven by the actual traffic processes. By using the parameters mentioned above, this model can be calibrated to fit the different types of network traffic processes studied here. We also describe a systematic approach for the selection of these parameters.

Our finding concerning the inefficiency of the Hurst parameter (in characterizing the burstiness of LRD traffic processes) sheds light on a modeling approach for VBR video traffic. The LRD features included in the measured video traffic stream do not imply that the measured traffic flow cannot be described by Markovian models. Indeed, we have found that among the four types of measured LRD traffic processes studied here (namely, LAN, WAN, WWW, and video traffic), the video traffic is the least bursty among all. LRD traffic models such as the multiplicative multifractal processes presented here provide another simple and effective alternative. This is especially so when one wishes the model to contain as few parameters as possible. (An even simpler multifractal model, which contains a single parameter, can be found in [7].)

Acknowledgements

Our thanks to the Telcordia researchers (Drs. Leland and Garret) for making available their Ethernet traffic trace data and the VBR video data. Thanks also to R. Ritke of computer science department of UCLA, who provided the WAN and WWW traffic data studied here. This work was supported by SBC Pacific Bell and University of California MICRO research grants No. 96-157, 97-152, 98-131, and by Army Research Office contract No. DAAG55-9801-0338.

References

[1] J. Beran, R. Sherman, M.S. Taqqu, W. Willinger, Long-range-dependence in variable-bit-rate video traffic, IEEE Trans. Commun. 43 (1995) 1566–1579.

[2] M.E. Crovella, A. Bestavros, Self-similarity in world wide web traffic: evidence and possible causes, IEEE/ACM Trans. Networking 5 (1997) 835–846.

[3] A. Erramilli, O. Narayan, W. Willinger, Experimental queueing analysis with long-range dependent packet traffic, IEEE/ACM Trans. Networking 4 (1996) 209–223.

[4] A. Feldmann, A.C. Gilbert, W. Willinger, Data networks as cascades: Investigating the multifractal nature of Internet WAN traffic, Proceedings of the ACM/SIGCOMM'98, Vancouver, BC, 1998.

[5] J.B. Gao, I. Rubin, Multiplicative multifractal modeling of long-range-dependent traffic, Proceedings of ICC'99, Vancouver, Canada, June 1999.

[6] J.B. Gao, I. Rubin, Statistical properties of multiplicative multifractal processes in modeling telecommunications traffic streams, Electron. Lett. 36 (2000) 101–102.

[7] J.B. Gao, I. Rubin, Multifractal analysis and modeling of VBR video traffic, Electron. Lett. 36 (2000) 278–279.

[8] M.W. Garret, W. Willlinger, Analysis, modeling and generation of self-similar VBR video traffic, in Proceedings of the ACM SIGCOMM, London, England, 1994.

[9] D.P. Heyman, T.V. Lakshman, What are the implications of long-range dependence for VBR-video traffic engineering, IEEE/ACM Trans. Networking 4 (1996) 301–317.

[10] W.E. Leland, M.S. Taqqu, W. Willinger, D.V. Wilson, On the self-similar nature of Ethernet traffic (extended version), IEEE/ACM Trans. Networking 2 (1994) 1–15.

[11] A.L. Neidhardt, J.L. Wang, The concept of relevant time scales and its application to queuing analysis of self-similar traffic (or is Hurst naughty or nice?), SIGMETRICS'98/PERFORMANCE'98, Joint International Conference on Measurement and Modeling of Computer Systems, Madison, WI, 1998.

[12] K.H. Park, G. Kim, M. Crovella, On the effect of traffic self-similarity on network performance, Proceedings of the SPIE International Conference on Performance and Control of Network Systems, November 1997, pp. 296–310.

[13] V. Paxson, S. Floyd, Wide area traffic — The failure of Poisson modeling, IEEE/ACM Trans. Networking 3 (1995) 226–244.

[14] M.S. Taqqu, V. Teverovsky, W. Willinger, Is network traffic self-similar or multifractal?, Fractals 5 (1997) 63–73.