Study of the Effect of the Wireless Gateway on Incoming Self-Similar Traffic

Jie Yu, *Member, IEEE*, and Athina P. Petropulu, *Senior Member, IEEE*

***Abstract*—It has been well established by now that high-speed wireline traffic exhibits self-similar behavior. Due to the important consequences of traffic self-similarity in network design, several studies have assumed that wireless traffic is also self-similar and looked at its effects on network performance. However, due to factors such as power limitations and the wireless channel, it is not straightforward that wireline traffic will remain self-similar as it enters the wireless network. This paper provides an analytical study of the propagation of traffic characteristics as wireline traffic is passed to the wireless network through a gateway. The analysis takes into account buffering and repacking operations performed at the gateway, and models for wireline traffic and the wireless channel. We consider two server models, an instant transfer model, and an energy-conserving one. We show that in most cases, in response to self-similar wireline traffic the gateway will produce self-similar wireless traffic. However, when the gateway operates under an energy-conserving mode and if it has a large buffer, wireline traffic such as non-real-time variable-bit-rate traffic will result in non-self-similar wireless traffic. We also study the delays of packets passing through a gateway that is fed by self-similar traffic and show that their survival function has an asymptotically power-law tail with index smaller than 2.**

***Index Terms*—Impulsive traffic, multimedia traffic, packet delays, self-similar network, wireless network traffic.**

I. Introduction

STATISTICAL modeling of traffic is very important in network engineering, and a substantial body of literature has been devoted to it. Traffic here is defined as bits per unit interval. Over the past decade, a number of empirical studies have established that traffic generated by multimedia applications exhibits a bursty outlook over a wide range of timescales [37]. Indeed, looking at local area network (LAN) traffic defined as bits per 10 s, or LAN traffic defined as bits per 1 s, one sees the same trends. The former is produced by averaging the latter every 10 s, a process referred to as aggregation. This behavior of LAN data traffic is in sharp contrast to circuit-switched voice traffic, which is smoothed out when aggregated. A process that is not smoothed out by aggregation (or time-scaling) is referred to as *self-similar* (SS).

Self-similarity can degrade network performance by causing large delays, packet dropping and by requiring large buffers [21], [27]. With traditional teletraffic models being nonapplicable to self-similar traffic, several new models have been proposed. Models driven by application level dynamics are based on the concept of the ON/OFF process [12], [15], [20], [37], which is also called alternating fractal renewal process (AFRP) [18]. According to an ON/OFF process, traffic alternates between periods of constant rate (ON states), and periods of silence (OFF states). The ON states durations are heavy-tail distributed, which is consistent with the high variability of files sizes associated with multimedia applications. It is those heavy-tailed durations that give rise to the traffic self-similarity. To capture the traffic marginal statistics as well as self-similarity the extended AFRP (EAFRP) model was proposed in [38] for LAN traffic. This is an ON/OFF model that has heavy-tail distributed rates during the ON states. An improvement to the EAFRP model was the rate-limited EAFRP [40], [41], which applies to all traffic load situations and takes into account finite rate limits imposed by real networks.

In addition to application level dynamics, protocol dynamics and the network are also factors that play a role in shaping the statistics of traffic [30], [33], [34]. In [30], [31] the effect of TCP on traffic was investigated, and it was concluded that the TCP's retransmission and congestion control mechanism, and in particular, its timeout and exponential backoff mechanisms can lead to self-similarity in aggregated TCP flows.

Recently, several studies hypothesized that wireless high-speed traffic will also be self-similar and looked into the consequences of self-similarity in the network in terms of performance and resource allocation [10], [32], [35], [42]. It is true that recent advances in wireless networks can enable high-speed communication and that high-speed wireless users tend to have the same bandwidth requirements as wired users when accessing the Internet [26]. However, that might not be sufficient justification for assuming that wireless traffic will exhibit self-similarity. There are big differences between wireless and wired transmission, such as the unreliable wireless channel [16], and severe power limitations imposed to the wireless users [4], [9], [11].

One plausible justification of SS in wireless traffic could be TCP effects. The loss rate here is higher than that of wireline traffic, thus, TCP could generate self-similarity in the wireless traffic even more so as it did in the wireline case. Another justification could be the propagation of application level characteristics at the moment when wireline traffic is fed to the wireless network via a gateway and has to face the wireless channel. We here investigate the latter possibility. In general, packet sizes are different over a heterogeneous collection of networks, and the access point (through which the wireless clients pass) acting as

Manuscript received January 8, 2005; revised October 25, 2005. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Zidong Wang. This work has been supported by ONR under grant N00014-03-1-0123.

The authors are with the Department of Electrical and Computer Engineering, Drexel University, Philadelphia, PA 19104 USA (e-mail: yujie@cbis.ece.drexel.edu; athina@ece.drexel.edu).

Digital Object Identifier 10.1109/TSP.2006.879329

gateway should provide a means by which all packets can be fragmented and reassembled [23]. In this paper, we propose a model for the repacking operation and use it along with a model for the radio channel [16] to study the dependence structure of traffic that leaves the gateway in response to the incoming self-similar traffic.

We also study queuing delays which are experienced by packets at a gateway that is fed by SS traffic. Among the various forms of packet delays, e.g., processing delays, queueing delays, transmission delays, propagation delays, queueing delay is the dominant factor in shaping queuing performance. Related works on this topic include [39], where for a G/M/1 queueing model with self-similar inputs, the average delay at a router was shown to be longer than that of an M/M/1 queueing system. In [39], it was also shown that the delay exhibits a rise as the degree of self-similarity increases. In [38] and [41], a GI/G/1 queueing model with SS input was studied and the stationary distribution of queue length was derived. In this paper, we derive an expression for the probability $P(\text{delay} > x)$ and show that for large x the probability decays in a power-law fashion with tail exponent less than 2. This is in contrast to the delay corresponding to Poisson traffic, for which the survival function decays in an exponential fashion.

The paper is organized as follows: We provide necessary statistical background in Section II. We describe the models of the wireless gateway and the incoming/outgoing traffic in Section III. In Section IV, we study the impact of gateway on the propagation of self-similar traffic. In Section V, we derive statistics of the delay experienced at the gateway by SS input traffic. Finally, in Section VI, we present simulation results to validate our findings.

Notation: $\langle a,b\rangle = \max(a,b)$; $a \wedge b = \min(a,b)$; $\lceil . \rceil$ is ceiling operation.

II. Mathematical Preliminaries

A. Heavy-Tail Distributions

A random variable X is heavy-tail distributed [7] with index α if there exists a slowly varying function at infinity, $L(x)$, such that, as $x \to \infty$

$$P(|X| \geq x) \sim \frac{L(x)}{x^\alpha} \tag{1}$$

where $0 < \alpha < 2$; $P(.)$ denotes probability; $L(x)$ is such that $\lim_{x\to\infty} L(bx)/L(x) = 1$ for any positive b. Typically, $L(x)$ is a constant, or ratio of two polynomials with identical degrees. Note that, in this paper, the notation $\sim$ means logarithmic convergence.

Heavy-tail distributed random variables have infinite variance, which implies that the random variable fluctuates far away from its mean value (defined only when $\alpha > 1$), with nonnegligible probability.

One example of heavy-tail distributions is the Pareto distribution, which is defined in terms of its complementary distribution function (survival function) as [2]:

$$\bar{F}(x;\alpha,K) = P(X \geq x) = \begin{cases} \left(\frac{K}{x}\right)^\alpha, & x \geq K \\ 1, & x < K \end{cases} \tag{2}$$

where K is a positive constant and $0 < \alpha < 2$. The corresponding probability density function is

$$f(x;\alpha,K) = \alpha K^\alpha x^{-(\alpha+1)}, \ \ x \geq K. \tag{3}$$

Note that if $0 < \alpha < 1$, the Pareto distribution has an unbounded mean.

Another well-known member of that class is the α-stable distribution [19], [28].

B. Long-Range Dependence and Self-Similarity

A wide-sense stationary process $\{X_k\}_{k\in Z}$ is said to possess long memory, or long-range dependence with Hurst parameter H, if its autocorrelation function $R_x(\tau)$, for all $\tau \in Z$, satisfies [5]

$$\lim_{\tau\to\infty} \frac{R_x(\tau)}{\tau^{2H-2}} = c, \quad 1/2 < H < 1 \tag{4}$$

for some positive constant c. Thus, a long-memory process is characterized by an autocorrelation that decays hyperbolically as the lag τ increases, in contrast to a short memory processes, e.g., ARMA, whose autocorrelation decays in an exponential fashion.

For discrete-time processes, self-similarity, in a strict sense is described by means of distributional invariance upon aggregation and scaling. The aggregate process of X_k of degree m is a running average of nonoverlapping blocks of X_k with length m. While the aggregate of a short-range dependent process has variance that decays as $c_1 m^{-1}$ with $m \to \infty$, the aggregate of a long-range dependent process has variance that decays much slower, as $c_2 m^{(2H-2)}$ with $m \to \infty$ (here c_1 and c_2 are constants). Aggregation is equivalent to time scaling. Thus, under time scaling, a long-range dependent process is smoothed out much slower in comparison to a short-range dependent process. In other words, the time-scaled long-range dependent process maintains **similarity** to the original process, thus indicating a relationship between long-range dependence and self-similarity.

Network traffic can be viewed as aggregation of traffic over smaller time intervals. It is in the sense of (4) that wireline network traffic is characterized as self-similar.

C. ON/OFF *Process*

The ON/OFF process is used to model single user traffic. The ON/OFF process alternates between two states: the ON, during which the source generates traffic at a rate A_j, and the OFF, during which the source remains silent. Let X_j and Y_j denote the duration of the jth ON and OFF state, respectively. Mathematically, the ON/OFF process can be expressed as

$$S(t) = \sum_{j=0}^{\infty} A_j 1_{[S_j, S_j+X_j)}(t), \quad t \geq 0, \tag{5}$$

where we have the following.

- S_j is a so-called regenerative point [3], denoting the onset of the jth ON period. It holds $S_j = S_0 + \sum_{i=1}^{j-1}(X_i + Y_i)$, $j \geq 1$, where S_0 represents the starting time of the first ON period, which will here be taken as $S_0 = 0$.
- $1_{[s_1,s_2)}(t)$ is the indicator function, which is nonzero and equals to one only for $t \in [s_1, s_2)$.

In most ON/OFF-type models, each of the X_j, Y_j are assumed to be independent identically distributed (i.i.d.). The X_j's are heavy-tail distributed, while the Y_j's, depending on the application can be heavy-tail distributed, or can have finite variance [13], [37].

In the original ON/OFF model [37], A_j is constant for all j. In the EAFRP [38] the A_j's are i.i.d. Pareto distributed, while in the rate-limited EAFRP [41], the A_j's are random i.i.d., cutoff Pareto distributed.

The Hurst parameter [17] of the ON/OFF process equals [36]:

$$H = \frac{3 - \min(\alpha_0, \alpha_1)}{2} \tag{6}$$

where α_1 and α_0 are the tail indexes of the ON and OFF durations, respectively. If the ON or OFF durations have finite variance then the corresponding tail index is taken as 2 when applying (6) [37]. The ON/OFF process is self-similar if the corresponding Hurst parameter satisfies $1/2 < H < 1$, or equivalently, when at least one of the ON or OFF durations is heavy-tail distributed.

III. System Model

A. Incoming Traffic

Let us denote by $S(t)$ the traffic that arrives at the gateway. $S(t)$ is a superposition of multiple traffic streams. In [41], it was shown that such traffic can be modeled by a single ON/OFF process referred to as mixture cutoff Pareto process, defined as follows. The ON/OFF durations, X_j^S and Y_j^S, are independent of each other, and each one is i.i.d. according to $\bar{F}(x; \alpha_1, K_1)$ and $\bar{F}(x; \alpha_0, K_0)$, respectively. We will here assume that $1 < \alpha_1 < 2$ and $1 < \alpha_0 \leq 2$. The ON duration rates, A_j are i.i.d. according to

$$\begin{aligned} &\bar{F}_A(x; \alpha_I, K_I, \alpha_{II}, L, R) \\ &\quad = P(X \geq x) \\ &\quad = \begin{cases} 1 & 0 \leq x < K_I \\ \left(\frac{K_I}{x}\right)^{\alpha_I} & K_I \leq x < L \\ \left(\frac{K_{II}}{x}\right)^{\alpha_{II}} & L \leq x < R \\ 0 & x > R \end{cases} \end{aligned} \tag{7}$$

where $u(.)$ is the unit step function; L and R represent two data limits imposed on single-user traffic and overall traffic, respectively [41]; $K_{II} = \exp\{(1/\alpha_{II}) \ln[K_I^{\alpha_I} L^{(\alpha_{II} - \alpha_I)}]\}$.

The density function corresponding to (7) is:

$$\begin{aligned} &f_A(x; \alpha_I, K_I, \alpha_{II}, L, R) \\ &\quad = f(x; \alpha_I, K_I)\,[1 - u(x - L)] \\ &\qquad + f(x; \alpha_{II}, K_{II})\,[u(x - L) - u(x - R)] \\ &\qquad + \left(\frac{K_{II}}{R}\right)^{\alpha_{II}} \delta(x - R) \end{aligned} \tag{8}$$

where $f(.)$ denotes the Pareto density function (see (3)); $\delta(.)$ is the Dirac function; $u(.)$ is the unit step function.

Due to the limit R, the mixture cutoff Pareto distribution has finite variance.

This model was shown in [41] to capture both correlation structure and marginal statistics of traffic, and applies to a wide range of traffic loads. For small to medium number of users, such as in LAN networks, the model matches the non-Gaussian impulsive characteristics of traffic, while as the number of users increases it results in Gaussian traffic, both of which are consistent with real network measurements [41].

B. Outgoing Traffic

Let us denote by $T(t)$ the traffic that leaves the gateway. It is also modeled as an ON/OFF process in the sense that it alternates between ON and OFF states. The ON/OFF durations will be denoted by X_j^T, Y_j^T, and their distributions will be derived in the next section. The ON state rate is constant and is determined by the channel capacity.

Due to the resampling function performed at the gateway, there is no one-to-one mapping between packets of incoming traffic and outgoing traffic. We will refer to outgoing packets as "cells."

For mathematical tractability, we will consider the stationary distributions of X_j^T, Y_j^T; thus, we will refer to

$$X^T \stackrel{d}{=} \lim_{j \to \infty} X_j^T \quad Y^T \stackrel{d}{=} \lim_{j \to \infty} Y_j^T$$

where $\stackrel{d}{=}$ represents equality in distribution.

C. Wireless Channel

We model the wireless channel based on the two-state Markovian model of [16], according to which, the channel alternates between *good* and *bad* states, corresponding to bit-error rates (BERs) $P_{e,\text{good}}$ and $P_{e,\text{bad}}$ where $P_{e,\text{good}} \ll P_{e,\text{bad}}$. The periods of good and bad states are i.i.d. exponentially distributed with means $1/\beta$ and $1/\gamma$, respectively. The model incorporates automatic repeat request/forward-error correction (ARQ/FEC) into the outgoing cells and assumes that all transmission errors can be detected. Retransmissions are requested if errors cannot be recovered by the FEC, until the cell is finally correctly received.

Let e be the maximum number of erroneous bits that can be corrected in an FEC protected cell. The probability that a received cell contains a noncorrectable error, in which case the transmission fails, is

$$P_{\text{fail,state}} = \sum_{i=e+1}^{w} \binom{w}{i} P_{e,\text{state}}^i (1 - P_{e,\text{state}})^{w-i} \qquad \text{state} = \{\text{good,bad}\} \tag{9}$$

where w is the size of a FEC protected cell.

At most, one cell will be sent out through the channel per time slot, while no cell will be sent out if the buffer is empty. The probability of successful transmission is $1 - P_{\text{fail,state}}$. If the transmission failed, that cell will be kept in the buffer and retransmitted until successfully received. The service time of a cell includes all transmission/retransmission attempts. As far as throughput is concerned, there are two deterministic service rates: c_g during *good* states and c_b during *bad* states are approximated to simplify the question, specifically as

$$c_g = P\frac{k}{w}(1 - P_{\text{fail,good}}) \quad c_b = P\frac{k}{w}(1 - P_{\text{fail,bad}}) \tag{10}$$

where P is the overall bit rate, including information bits and redundancy, and (w, k) FEC code is assumed, where w and k are the size in bits of code word and payload, respectively. Since we assumed $P_{e,\text{good}} \ll P_{e,\text{bad}}$, it holds $c_g \gg c_b$.

We should note that the two-state model is a simplistic model. When applied to a realistic environment, each state would model the average behavior during good and bad channel states. Extensions to multistate channel model will be discussed in Section IV-D.

D. Gateway Modeling

The gateway serving the overall traffic will be viewed as a buffering system. We use the traditional Kendall's notation to denote this system, i.e., **G/G/1/B**, where the first **G** represents the general input traffic, the second **G** represents the general statistics for service times, **1** represents one server, and **B** represents buffer size.

Traffic streams $S(t)$ from multiple connections (corresponding to clients sharing the same access point) are fed into a finite buffer of size B until the buffer overflows, in which case the excess data are discarded. Let us consider a time-slotted model with the time slot denoted by τ. In the sequel we will use the notation $S(n)$, $T(n)$ instead of $S(t)$, $T(t)$, where n is the slot index. Each ON or OFF period consists of several slots. In the beginning of the nth time slot, assuming that $S(n) > 0$, one packet arrives with $S(n)$ information bits in it. Thereafter, the gateway resamples the packet into bits, stores them in the buffer, and repacks $c\tau$ bits into one cell to be transmitted. At the end of the slot n, at most one cell leaves the buffer.

As mentioned in the previous subsection, the instant channel capacity alternates between c_g and c_b. In the sequel, cell size refers to the number of information bits in a packet or cell, which excludes overhead and redundancy.

Two server models are considered in this paper:

Server Model 1 (SM1): If the data in the buffer are less than $c\tau$, the server takes no action and waits until enough data have come in to form a cell.

Server Model 2 (SM2): The server sends out cells whenever there are bits in the buffer; trivial bits are added if needed to form a cell.

SM1 is an energy-conserving model. It avoids trivial bits in the outgoing cells, and also reduces the time the wireless receiver needs to be active receiving data. However, it increases transmission delays. SM2 on the other hand, results in smaller delays due to instant transmission of information at the expense of bandwidth and energy efficiency.

SM1 and SM2 can be viewed as two extreme cases with respect to delay and power consumption. Any real server can be viewed as somewhere in between, achieving tradeoff between delay and power consumption. Therefore, these two server models will be studied here in order to determine upper and lower bounds of the performance of the gateway.

IV. Impact of the Buffering System on the Outgoing Traffic

Let us consider a slow-varying wireless channel so that the service rate of the gateway can be assumed constant within several ON/OFF periods of $S(n)$.

We can view the gateway approximately as the statistical multiplexing of two buffer systems, both having the same buffer size but serving at two different rates: c_g, corresponding to *good* channel states, and c_b, corresponding to *bad* channel states. As a first step, we will derive the statistics of the outgoing traffic for a buffering system that serves at a constant rate.

Let $Q(n)$ denote the buffer content at the slot n, with initial value $Q(0) = 0$.

A. Server Model 1

$T(n)$ and $Q(n)$ are updated on a slot-by-slot basis as follows:

$$T(n) = \begin{cases} c, & \text{if } S(n)\tau + Q(n-1) \geq c\tau \\ 0, & \text{if } S(n)\tau + Q(n-1) < c\tau \end{cases} \tag{11}$$

$$Q(n) = \begin{cases} \langle S(n)\tau + Q(n-1) - c\tau, 0\rangle \wedge B, \\ \quad \text{if } S(n)\tau + Q(n-1) \geq c\tau \\ \langle S(n)\tau + Q(n-1), 0\rangle \wedge B, \\ \quad \text{if } S(n)\tau + Q(n-1) < c\tau \end{cases} \tag{12}$$

where B denotes buffer size. Thus, the buffer content satisfies: $0 \leq Q(n) \leq B$.

For mathematical simplicity in the sequel we only study two extreme cases: the *small* buffer system $(B = c\tau)$ and *large* buffer system $(B \gg c\tau)$. If B is moderately larger than $c\tau$, the analysis is rather intractable. However, our simulations indicate that if $B > 5c\tau$ the corresponding buffering system acts more like a *large* buffer system, in which case the analysis shown next still applies.

1) Small Buffer System: In this case, the buffer holds at most one cell at a time. Unless the buffer is full, i.e., $Q(n) = B$, there is not enough data to form a cell, so no traffic leaves the buffer. In addition, during all OFF periods of $S(n)$, $T(n) = 0$. During the ON periods of $S(n)$, as new bits come into the buffer with rates A_j, the buffer content is updated according to (12), and thus $T(n)$ changes according to (11). The buffer system will not keep one-to-one mapping between X_j^S and X_j^T, or Y_j^S and Y_j^T. Instead, while waiting for enough data to accumulate in the buffer, $2M + 1$ consecutive ON/OFF periods of $S(n)$ could be combined to form a bigger OFF period in $T(n)$, i.e., $Y_k^T = Y_{j-1}^S + \sum_{i=j}^{j+M-1}(X_j^S + Y_j^S)$. We will refer to such action as "*combining*" action.

It holds

$$P(X^T > x) = P(X^T > x | A \geq c)P(A \geq c) + P(X^T > x | A < c)P(A < c) \tag{13}$$

and

$$P(Y^T > y) = P(Y^T > y | A \geq c)P(A \geq c) + P(Y^T > y | A < c)P(A < c). \tag{14}$$

Let us make the following assumptions.

A1) The minimum rate during an ON period of the incoming traffic, i.e., K_I, satisfies $K_I \ll c$. To see the implication of this assumption, let $K_I < (c/z)$, where z is a large number. Then, $(P(A < c)/P(A \geq c)) = (1/P(A \geq c)) - 1 > z^{\alpha_I} - 1 \gg 1$. Thus, under A1), it holds $P(A < c) \gg P(A \geq c)$. Assumption A1) will be valid in a low-to-moderate traffic load scenario. To see this, consider the traffic intensity for the rate-limited EAFRP process [41]

$$\begin{aligned}\rho &= \frac{E\{S(n)\}}{c} \\ &= \frac{\mu_1}{(\mu_1+\mu_0)}\left(\frac{K_I}{c}\right) \\ &\quad \times \left\{\frac{\alpha_I}{1-\alpha_I}\left[K_I^{(\alpha_I-1)}L^{(1-\alpha_I)}-1\right]+\frac{\alpha_{II}}{(1-\alpha_{II})K_I}\right. \\ &\qquad \left. \times K_{II}^{\alpha_{II}}\left[R^{(1-\alpha_{II})}-L^{(1-\alpha_{II})}\right]\right\} \end{aligned} \tag{15}$$

where μ_1 and μ_0 are means of ON/OFF durations, respectively. The above equation implies that for fixed K_I, as c increases the traffic load decreases. In the following, we assume that A1) holds during both good and bad channel states. At first look, it might seem rather restrictive to assume A1) during bad states. However, let us consider the real LAN traffic data that will be used in the simulations section. For those data, it holds $K_I = 48$ bps. For wireless traffic, the overall data rate 1.27 Mb/s and bit error rate (BER) 10^{-2} (10) yields $c_b = 290$ Kb/s. Although the wireless traffic rate used was rather conservative (IEEE 802.11a can achieve 54 Mb/s, or around 20 Mb/s without overhead), even in bad states $K_I \ll c_b$.

A2) We ignore the previous buffer content, $Q(n-1)$ when calculating $T(n)$ by (11). For this small buffer system, the previous buffer content can cause the ON/OFF durations of $T(n)$ to increase/decrease by at most one time slot. On the other hand, X^S and Y^S are Pareto distributed and thus can take very large value with nontrivial probability. Therefore, the previous buffer content has relatively small impact on the distribution of X^T and Y^T and thus can be ignored. Then, (11) can be simplified as

$$T(n) = \begin{cases} c & S(n) \geq c \\ 0 & S(n) < c \end{cases}. \tag{16}$$

The above approximation is made mainly for mathematical convenience. In the simulations section we will provide simulation results to confirm its validity.

A3) For a real queue, it always holds that $L > c$ for all practical values of traffic intensity ρ.

Via assumption A2), if $A < c$, then $T(n) = 0$ for the entire duration of the ON states of $S(n)$; thus, $P(X^T > x | A < c) = 0$. In addition, based on A1), we can assume that $P(A < c) \approx 1$. Combining A1) and A2), we approximate (13) and (14) as $P(X^T > x) \approx P(X^T > x | A \geq c)P(A \geq c)$, and $P(Y^T > y) \approx P(Y^T > y | A < c)$, respectively.

Proposition 1: For the small buffer system $(B = c\tau)$ under SM1 and assumptions A1), A2), and A3), it holds

$$\begin{aligned} P(X^T > x) &\approx P(X^S > x) \\ P(Y^T > y) &= \sum_{m=1}^{\infty} P\left(\sum_{i=1}^{m} X_i^S + \sum_{i=1}^{m+1} Y_i^S > y\right) \\ &\qquad \times \left[1-\left(\frac{K_I}{c}\right)^{\alpha_I}\right]^m \left(\frac{K_I}{c}\right)^{2\alpha_I} \end{aligned} \tag{17}$$

$$\overset{y\to\infty}{\sim} L(y)y^{-\min\{\alpha_1,\alpha_0\}}. \tag{18}$$

Proof: See Appendix A.

Based on Proposition 1 and due to our assumption that $1 < \alpha_1 < 2$ and $1 < \alpha_0 \leq 2$, the minimum between the tail indexes of ON and OFF durations of $T(n)$ is still less than 2. Thus, $T(n)$ is self-similar.

2) Large Buffer System: Assumption A2) is no longer valid in the large buffer case. When $B \gg c\tau$, the previous buffer content $Q(n-1)$ can take very large values and thus cannot be ignored. In this case, $T(n)$ can be in ON state even when $A < c$, and thus $P(X^T > x | A < c)$ can take nonzero values.

Based on assumption A1), we can approximate (13) and (14) as $P(X^T > x) \approx P(X^T > x | A < c)$ and $P(Y^T > y) \approx P(Y^T > y | A < c)$, respectively.

In addition to assumption A1), we will also assume the following

A1) The queue is stable, which implies that $c > E[S(n)] = (\mu_A \mu_1/\mu_1 + \mu_0)$. Under A4), we can study the stationary distribution of the buffer content at the regenerative points S_j, defined as

$$Q_e \overset{d}{=} \lim_{j\to\infty} Q(S_j)$$

where $\overset{d}{=}$ represents equality in distribution and Q_e represents stationary distribution of $Q(S_j)$.

Proposition 2: For the large buffer system (i.e., $B \gg c\tau$) and under SM1, and assumptions A1), A3), and A4), it holds

$$\begin{aligned} P(X^T > x) &\overset{x\to\infty}{\sim} C_1 x^{-(\alpha_1+1)} \\ P(Y^T > y) &\approx P(Y^S > y) \end{aligned} \tag{19}$$

where C_1 is a constant and α_1 is the tail index of the ON durations of the incoming traffic.

Proof: See Appendix B.

Proposition 2 suggests that both ON and OFF durations will be power-law distributed. Also, the tail index of the ON durations of the outgoing traffic will increase by one, while the tail index of the OFF durations will be the same as in the incoming traffic. The outgoing traffic will be self-similar if the smaller among the tail indexes of X^T, Y^T is less than 2. Thus, if $\alpha_0 < 2$ the outgoing traffic will be self-similar, while if $\alpha_0 = 2$ it will not be self-similar.

B. Server Model 2

Based on the definition of SM2, $T(n)$ and $Q(n)$ are updated on a slot-by-slot basis as (20) and (21), shown at the bottom of the next page.

As long as the input traffic is in ON state, i.e., $S(n) > 0, T(n)$ is always in ON state, i.e., $T(n) = c$. Therefore, at the end of the jth ON period of $S(n)$, the buffer content is

$$Q\left(S_j + X_j^S\right) = \left\langle (A_j - c)X_j^S + Q(S_j), 0\right\rangle \wedge B.$$

If $Q(S_j + X_j^S) > 0$, then $T(n)$ remains in the ON state even when $S(n)$ enters its jth OFF period. The length of the corresponding extension is $\delta_j = \lceil (Q(S_j + X_j^S)/c) \rceil \wedge Y_j^S$.

Let us assume A1), i.e., the event "$A < c$" is dominant over "$A \geq c$." Taking into account that $A < c$, we can see that the buffer content decreases during the ON period of $S(n)$, i.e., $Q(S_j + X_j^S) \leq Q(S_j)$, with high probability. During the subsequent OFF period, no new bits arrive while the server continues sending packets until the buffer is empty. After several alternations of ON/OFF states, $Q(S_j) = 0$. Since $Q(S_j + X_j^S) \leq Q(S_j)$ and the buffer content is nonnegative, $Q(S_j + X_j^S) = 0$ with high probability. In other words, $\delta_j = 0$ holds with high probability, and thus the jth ON period of $T(n)$ (i.e., X_j^T) ends at the time of $S_j + X_j^S$, which is the beginning of the subsequent jth OFF period of $S(n)$.

In summary, for SM2 and under assumption A1), the server always sends cells during the ON periods of $S(n)$ and remains silent during the OFF periods (since no inputs arrived in the meantime), i.e., $P(X^T > x) \approx P(X^S > x)$ and $P(Y^T > x) \approx P(Y^S > x)$. Thus, the tail indexes of X^T, Y^T are the same with those of X^S, Y^S, respectively.

C. Buffering System Serving the Wireless Channel

As it was already mentioned, we view the gateway action as statistical multiplexing of two buffering systems that share the same buffer but have different service rates determined by the capacities of wireless channel, i.e., c_g and c_b.

It holds that

$$\begin{aligned} P(X^T > x) = P(X^T > x|c = c_g)P(c = c_g) \\ +P(X^T > x|c = c_b)P(c = c_b) \end{aligned} \quad (22)$$

and

$$\begin{aligned} P(Y^T > y) = P(Y^T > y|c = c_g)P(c = c_g) \\ +P(Y^T > y|c = c_b)P(c = c_b). \end{aligned} \quad (23)$$

With the alternation of channel capacities, the gateway can be categorized into the following two types.

1) $B = c_g\tau \gg c_b\tau$. In this case, the gateway alternates under small and large buffer model during good and bad channel states, respectively. Substituting (18) and (19) into (22) and (23) and we get

$$\begin{aligned} P(X^T > x) \overset{x\to\infty}{\sim} & K_1^{\alpha_1} x^{-\alpha_1} P(c = c_g) \\ & + C_2 x^{-(\alpha_1+1)} P(c = c_b) \\ \sim & \left[K_1^{\alpha_1} P(c = c_g)\right] x^{-\alpha_1} \end{aligned} \quad (24)$$

and

$$\begin{aligned} P(Y^T > x) \overset{x\to\infty}{\sim} & x^{-\min\{\alpha_1,\alpha_0\}} P(c = c_g) \\ & + K_0^{\alpha_0} x^{-\alpha_0} P(c = c_b) \\ \sim & x^{-\min\{\alpha_1,\alpha_0\}} \end{aligned} \quad (25)$$

where $P(c = c_g) = \gamma/(\beta + \gamma)$, $P(c = c_b) = \beta/(\beta + \gamma)$ and $1/\beta$, $1/\gamma$ are mean durations of good and bad states. Based on Propositions 1 and 2, the tail index of X^T will be α_1, and the tail index of Y^T will be $\min(\alpha_1, \alpha_0) < 2$. Thus, since we assumed that $1 < \alpha_1 < 2$ and $1 < \alpha_0 \leq 2$, the outgoing traffic will be self-similar.

2) $B \gg c_g\tau, c_b\tau$. In this case, the gateway always falls under a large buffer model. According to Proposition 2, the gateway increases the tail exponent of ON durations by one, while it leaves the tail of the OFF durations unchanged. If $\alpha_0 < 2$, then the outgoing traffic will be self-similar. However, if $\alpha_0 = 2$ the outgoing traffic will not be self-similar.

D. Summary of Findings and Discussion

We are concerned with a gateway fed by self-similar traffic of low to moderate rate. The results depend on the gateway protocol considered.

For the instant transferring protocol, SM2, when the sum of incoming data and previous buffer content is greater than the cell size, the excess data is stored in the buffer and waits for transmission in the next available slot. Otherwise, even if the total data is not enough to form a cell, a cell always leaves the buffer. Based on the results presented above, a gateway operating under SM2, when excited with self-similar traffic, will produce self-similar traffic. This is also in agreement with common wisdom that views buffering as a low-pass filter, which should let long-range dependence go through unaltered.

However, a gateway operating under an energy conserving protocol, SM1, can, under certain circumstances, generate traffic that is not self-similar. In particular, if the gateway is equipped with a buffer that can store a large number of packets (large buffer model), the statistics of the ON/OFF durations of the output traffic can change in comparison to the corresponding states of the incoming traffic. In other words, we find that the energy conserving mechanism has a stronger impact on the change of statistics as compared to just the buffering that

$$T(n) = \begin{cases} c, & \text{if } S(n)\tau + Q(n-1) > 0 \\ 0, & \text{if } S(n)\tau + Q(n-1) = 0 \end{cases} \quad (20)$$

$$Q(n) = \begin{cases} [S(n)\tau + Q(n-1) - c\tau] \wedge B, & \text{if } S(n)\tau + Q(n-1) > c\tau \\ 0, & \text{if } S(n)\tau + Q(n-1) \leq c\tau \end{cases} \quad (21)$$

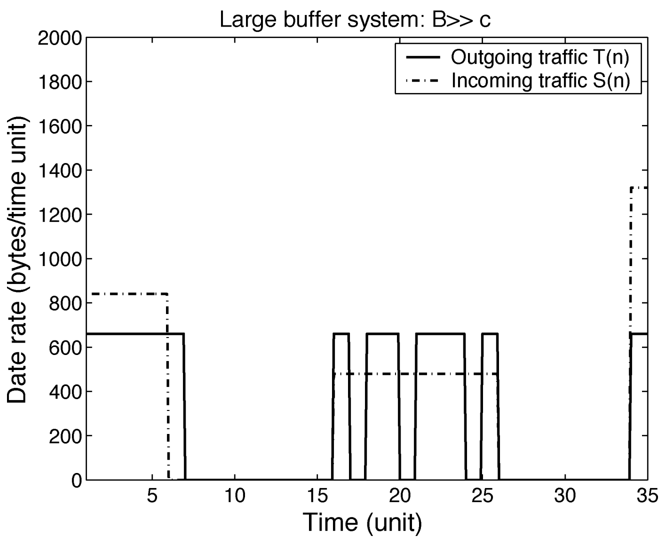

Fig. 1. For a gateway operation under an energy conserving mode (SM1), long ON states in the input traffic can give rise to a number of shorter ones at the output.

occurs in SM2. An intuitive explanation for this change can be obtained by considering the effect of the large buffer model on the incoming data. Since the incoming rate is assumed to be smaller than the outgoing rate (see assumption A1)), and due to the energy conserving operation of SM1, long ON states in the input traffic can give rise to a number of shorter ones at the output (see Fig. 1 for an illustration of the effect of the gateway). As a result, the probability that the output ON states duration will exceed a large number reduces, thus the tail index of the output ON durations increases. The loss of self-similarity because of the truncation of long bursts has also been reported by a simulation study in [29] conducted on wireline traffic. There, the truncation occurred due to packet dropping.

For LAN traffic, where the tail indexes of both ON and OFF durations of the incoming traffic are less than 2, the change in the tail indexes of output traffic ON and OFF states can change the degree of self-similarity (Hurst parameter). For example, if $\alpha_1 < \alpha_0 < 2$, the Hurst parameter of the incoming traffic is $H = (3 - \alpha_1)/2 < 1$, and that of the outgoing traffic is $H = (3 - \alpha_0)/2 < 1$. However, for such traffic, the output traffic will still be self-similar. On the other hand, for non-real-time variable-bit-rate (VBR) multimedia traffic [13], where the OFF states have finite variance (i.e., $\alpha_0 = 2$), the change in the statistics will be such that it will result in loss of self-similarity.

Although our claims on self-similarity loss in the outgoing traffic are based on the tails of ON and OFF periods of the outgoing traffic, in practice these periods might not be independent. Since a more realistic analysis might be intractable, in the simulations section (see Fig. 7), we show that by considering time-variance plots for the overall outgoing traffic we can arrive to the same conclusion.

The above analysis employed a simple two-state model for the wireless channel. The model can be extended to include more states, i.e., $c_1, \ldots c_n$. For each c_i, depending on the relative value of B and $c_i\tau$, the gateway can be categorized as either a small buffer case (i.e., $B = c_i\tau$), large buffer case (i.e., $B \gg c_i\tau$), or something in between. The above analysis can straightforwardly apply to the multiple-state channel model, if the c_i's are well separated and the B is appropriately chosen so that the gateway alternates only between the small and large buffer cases. Unfortunately, our analysis does not apply for the case of B moderately larger than $c_i\tau$.

V. Analysis of Delays

In addition to throughput performance, statistics of packet delays are also of great interest, especially in delay-sensitive applications, such as online gaming or real-time audio/video, just to name a few.

In this section, we study the delay of packets of incoming SS traffic, which is experienced due to the gateway protocol and due to the wireless channel. For buffer size B and channel capacities c_g and c_b, we provide analytic expressions for the upper and lower bounds of the probability that the delay will reach a certain value. These bounds can be useful for admission control and scheduling.

We will make the following assumptions.

B1) The buffering system is first-in-first-out (FIFO), and the buffer capacity is large enough to hold all incoming data, i.e., no cell will be discarded due to buffer overflow.

B2) All retransmission attempts are assumed to be independent of each other.

B3) All cells will be kept in the buffer until they are received correctly.

B4) We assume that the durations of channel states are relatively large (i.e., slow-varying channel) so that all retransmission attempts of a cell can be completed within one channel state duration.

Assumption B3) is made to simplify the analysis. SM1 can introduce very long delays. In practical cases, e.g., real-time or streaming applications, there is a timeout mechanism that avoids long delays by discarding packets whose waiting times exceed a certain threshold. The queueing analysis of the practical system that discards packets, or equivalently, of a finite buffer system, is rather intractable. Compared with a system that does not discard any packets, the practical system will encounter smaller packet delays. In the simulations section, we will compare the analysis based on assumption B3) to simulations results that impose an upper limit on delays.

Due to the resampling and repacking operations of the gateway, bits of the same packet of $S(n)$ might be separated into several consecutive cells of $T(n)$, or vice versa.

Consider the j_1j_2th packet, i.e., the data arriving during the j_2th slot of the j_1th ON period of $S(n)$. Let us define the following:

- $n^1_{j_1j_2}$: the index of the time slot when the j_1j_2th packet comes into the buffer (we assume that all bits of this packet come in at the same time);
- $n^2_{j_1j_2}$: the index of the time slot when the cell that includes the last bit of the j_1j_2th packet is received successfully by the receiver.

The one-way delay via the gateway [1] of the j_1j_2th packet of $S(n)$ is

$$\tau^S_{j_1j_2} = n^2_{j_1j_2} - n^1_{j_1j_2}. \quad (26)$$

Let us first consider the service time for one cell that holds the first place in the buffer and is ready for transmission.

Let τ_j^T denote the service time of that first cell. Here, τ_j^T includes all transmission/retransmission attempts. Due to the Markovian characteristic of wireless channel, and the identical sizes of all cells, we can assume that the τ_j^T's are i.i.d. and denote them as τ^T. Their distribution function conditioned on the channel state is

$$P(\tau^T = i|\text{state}) = (1 - P_{\text{fail,state}})P_{\text{fail,state}}^{i-1} \quad i = 1, 2, \ldots \quad \text{state} = \{\text{good,bad}\}. \tag{27}$$

According to B4), it holds

$$\begin{aligned} P(\tau^T = i) &= P(\tau^T = i|\text{state = good})P(\text{state = good}) \\ &\quad + P(\tau^T = i|\text{state = bad})P(\text{state = bad}) \\ &= (1 - P_{\text{fail,good}})P_{\text{fail,good}}^{i-1}\frac{1/\beta}{1/\gamma + 1/\beta} \\ &\quad + (1 - P_{\text{fail,bad}})P_{\text{fail,bad}}^{i-1}\frac{1/\gamma}{1/\gamma + 1/\beta} \\ &= (1 - P_{\text{fail,good}})P_{\text{fail,good}}^{i-1}\frac{\gamma}{\gamma + \beta} \\ &\quad + (1 - P_{\text{fail,bad}})P_{\text{fail,bad}}^{i-1}\frac{\beta}{\gamma + \beta}. \end{aligned} \tag{28}$$

The energy conserving SM1 will incur the longest delay, while SM2 will achieve the smallest delay. Since the delay analysis is difficult, in particular now that the incoming traffic is self-similar, we will next consider two bounds: the upper bound achieved by SM1 and the lower bound achieved by SM2.

A. Upper Bound: Achieved by Server Model 1

Consider the j_1j_2th packet with size $S(n_{j_1j_2}^1)$, which comes into the buffer at time $n_{j_1j_2}^1$. Since we assume the data rate over the same ON period to be constant, we have $S(n_{j_1j_2}^1) = A_{j_1}$.

In response to its arrival, the total buffer is updated as

$$\begin{aligned} Q\left(n_{j_1j_2}^1\right) &= Q\left(n_{j_1j_2}^1 - 1\right) + S\left(n_{j_1j_2}^1\right)\tau \\ &= Q\left(n_{j_1j_2}^1 - 1\right) + A_{j_1}\tau && (29) \\ &= c\tau N_{j_1j_2} + \delta_{j_1j_2} && (30) \end{aligned}$$

where $N_{j_1j_2}=\lfloor(Q(n_{j_1j_2}^1)/c\tau)\rfloor$ and $\delta_{j_1j_2}=\text{Mod}(Q(n_{j_1j_2}^1), c\tau)$.

By the time the $N_{j_1j_2}$ cells have been transmitted successfully, the excess $\delta_{j_1j_2}$ bits may need extra waiting time if, in the mean time, not enough bits to form a cell have arrived, or equivalently, if $\delta_{j_1j_2} + \sum_{i=1}^{N_{j_1j_2}\tau^T} S(n_{j_1j_2}^1 + i)\tau < c\tau$.

Let us express the entire delay of the j_1j_2th packet of $S(n)$ as

$$\tau_{j_1j_2}^S = N_{j_1j_2}\tau^T + W_{j_1j_2} + \tau^T \tag{31}$$

where τ^T is the service time of the last cell; $W_{j_1j_2}$ is the extra waiting time for the more incoming bits to form a cell, or equivalently, the smallest n ($n > n_{j_1j_2}^1$) for which

$$Q(n) = \delta_{j_1j_2} + \sum_{i=n_{j_1j_2}^1+1}^{n} S(i)\tau > c\tau. \tag{32}$$

The excess waiting time $W_{j_1j_2}$ falls under the following cases.

C1) $W_{j_1j_2} = -\tau^T$, if $Q(n_{j_1j_2}^1 - 1) + A_{j_1}\tau = c\tau N_{j_1j_2}$, which implies that the buffer content at the time $n = n_{j_1j_2}^1$, i.e., $Q(n_{j_1j_2}^1)$, is exactly a multiple of the cell size.

C2) $W_{j_1j_2} = 0$, if $\delta_{j_1j_2} + \sum_{i=1}^{N_{j_1j_2}\tau^T} S(n_{j_1j_2}^1 + i)\tau = \delta_{j_1j_2} + N_{j_1j_2}\tau^T A_{j_1}\tau > c\tau$ and $N_{j_1j_2}\tau^T \leq \text{Res\_}X_{j_1}^S$, where $\text{Res\_}X_{j_1}^S$ represents the residue life the j_1th ON period of $S(n)$ after the time $n = n_{j_1j_2}^1$. The first condition implies that (32) is satisfied by the time the $N_{j_1j_2}$ cells have been serviced, and the second one implies that all transmission of $N_{j_1j_2}$ cells is finished before the end of the ongoing ON period (i.e., j_1th ON period of $S(n)$).

C3) $W_{j_1j_2} \leq \max\{\text{Res\_}X_{j_1}^S - N_{j_1j_2}\tau^T, 0\}$, if $\delta_{j_1j_2} + \sum_{i=1}^{\text{Res\_}X_{j_1}^S} S(n_{j_1j_2}^1 + i)\tau = \delta_{j_1j_2} + \text{Res\_}X_{j_1}^S A_{j_1}\tau > c\tau$. The condition implies that (32) is satisfied before the end of the j_1th ON period.

C4) $W_{j_1j_2} \leq \max\{\text{Res\_}X_{j_1}^S + \sum_{i=0}^{m-1}[Y^S(j_1+i) + X^S(j_1 + i + 1)] - N_{j_1j_2}\tau^T - \text{Res\_}X_{j_1+m}^S, 0\}$, where m is the smallest integer that satisfies the inequality

$$\sum_{i=j_1+1}^{j_1+m} A_i X_i \tau > c\tau - \delta_{j_1j_2} - \text{Res\_}X_{j_1}^S A_{j_1}\tau.$$

The condition implies that (32) is satisfied before the end of the $(j_1 + m)$th ON period. The max function in C3) and C4) guarantees that $W_{j_1j_2}$ takes nonnegative values.

For mathematical convenience, we again turn to the stationary distributions

$$\begin{aligned} &Q \overset{d}{=} \lim_{n\to\infty} Q(n); \quad S \overset{d}{=} \lim_{n\to\infty} S(n); \\ &\tau^S \overset{d}{=} \lim_{j\to\infty} \tau_j^S; \quad W \overset{d}{=} \lim_{j_1\to\infty, j_2\to\infty} W_{j_1j_2}; \\ &N \overset{d}{=} \lim_{j_1\to\infty, j_2\to\infty} N_{j_1j_2}; \quad \text{Res\_}X^S \overset{d}{=} \lim_{j\to\infty} \text{Res\_}X_j^S \end{aligned}$$

where $\overset{d}{=}$ represents equality in distribution. The distribution of Q was found in [41] to be heavy-tailed with tail index $\alpha_1 - 1$ (α_1 is the tail index of ON durations), i.e., $P[Q > q] \overset{q\to\infty}{\sim} C' q^{1-\alpha_1}$, where C' is a constant. So the pdf of Q is

$$f_Q[x] \sim C'(\alpha_1 - 1)x^{-\alpha_1} \triangleq C_Q x^{-\alpha_1}. \tag{33}$$

We also assume i.i.d. ON rates and ON and OFF duration for $S(n)$. We set $\delta_{j_1j_2} = 0$ in the following to simplify the above expressions, and by doing this, we tend to overestimate the value of excess waiting time W. Taking the stationary versions of C1)–C4) and considering the upper limits of the excess waiting time, we get

$$\begin{aligned} &P\{W = w\} \\ &= P\{w = -\tau^T|Q + A\tau = c\tau L\}P\{Q + A\tau = c\tau L\} \\ &\quad + P\{w = 0|N\tau^T A > c\}P\{N\tau^T A > c\} \\ &\quad + P\left\{w = \max\{\text{Res\_}X^S - N\tau^T, 0\}|\text{Res\_}X^S A > c\right\} \\ &\quad \times P\{\text{Res\_}X^S A > c\} \\ &\quad + P\left\{w = \max\left\{\text{Res\_}X_{j_1}^S + mY^S + (m-1)X^S - N\tau^T - \text{Res\_}X_{j_1+m}^S, 0\right\}\right. \\ &\qquad \left. |mAX^S > c - \text{Res\_}X_{j_1}^S A > (m-1)AX^S\right\} \\ &\quad \times P\left\{mAX^S > c - \text{Res\_}X_{j_1}^S A > (m-1)AX^S\right\} \end{aligned} \tag{34}$$

where $N = \lfloor (Q + A\tau)/(c\tau) \rfloor$; $\text{Res\_}X^S_{j_1}$ and $\text{Res\_}X^S_{j_1+m}$ are i.i.d. and the subscripts are added to indicate that they belong to different ON periods.

To find the stationary distribution of the excess waiting time W, we propose a semiexperimental approach (such an approach was also used in [25]) for computing $P\{W = w\}$ as follows. We perform Monte Carlo runs, in each run generating Q_j, N_j, τ^T_j, A_j, X_j, Y_j, $\text{Res\_}X^S_j$ $(j = 1, \ldots, M)$ according to their distributions given, respectively, by (33), (54), (28), $\bar{F}_A(x; \alpha_I, K_I, \alpha_{II}, L, R)$, $\bar{F}(x; \alpha_1, K_1)$, $\bar{F}(x; \alpha_0, K_0)$, (56). The channel capacity takes the value c_g with probability $\gamma/(\beta + \gamma)$, and c_b with probability $\beta/(\beta + \gamma)$. With the generated samples, we compute the excess waiting time W_j as (35), shown at the bottom of the page. Then, we generate τ^S_j as the sum $(N_j + 1)\tau^T_j + W_j$ $(j = 1, \ldots, M)$ via (31). Finally, we compute the empirical cumulative density function of the generated τ^S's.

We should note here that one could model the outgoing traffic as an ON/OFF process, and then compute packet delays and find their statistics. We will refer to this approach as the "simulation approach." The semiexperimental method offers an advantage as compared to the simulation approach, as it can lead to the statistics in significantly less computation time. For the simulation approach, we need to update the values of $T(n)$, $Q(n)$ (via (11) and (12) for SM1 and (20) and (21) for SM2) per time slot and then compute the delay of the j_1j_2th packet as $\tau^S(j_1j_2) = n^2_{j_1j_2} - n^1_{j_1j_2}$ (see Section V). It takes $\tau^S_{j_1j_2}$ steps to get one sample of τ^S. As it will be shown in Fig. 9, τ^S is asymptotically heavy-tailed; thus, τ^S could be very large. On the other hand, with the semiexperimental method, we can get one sample per step (via (35)). As a comparison of the two approaches, for the results that are shown in Section VI to get $M = 10^6$ samples of τ^S, the simulation method takes 7655 seconds while the semiexperimental one takes 78 seconds.

B. Lower Bound: Achieved by Server Model 2

When the SM2 is applied, the last few bits of the j_1j_2th packet will be repacked into the next $N_{j_1j_2}$th cell of $T(n)$, where

$$N_{j_1j_2} = \left\lceil \frac{Q\left(n^1_{j_1j_2}\right)}{c\tau} \right\rceil. \quad (36)$$

The stationary version of the above equation is: $N = \lceil (Q + A\tau)/(c\tau) \rceil$. We assume the independence between N and τ^T for mathematical convenience, and we can validate this assumption by the simulation. We show results in Fig. 9 (left). In that figure, we did the semiexperiment (details can be found in Section V-A) and compared that with the result of real traffic. The two curves ("Analytic result of SM2" and "Simulation of SM2") are in close agreement, which suggests the analytical result is accurate.

The delay of the packet can be expressed as $\tau^S = N\tau^T$, so the distribution function of τ^S, along the lines of Appendix C, equals

$$\begin{aligned} P(\tau^S = k) &= P(N\tau^T = k) \\ &= \sum_{i_1 i_2 = k,\ i_1, i_2 = 1, 2, \ldots} \Bigg\{ c_g^{1-\alpha_1} i_1^{-\alpha_1} (1 - P_{\text{fail,good}}) \\ &\qquad \times P^{i_2-1}_{\text{fail,good}} \frac{\gamma}{\gamma+\beta} \\ &\qquad + c_b^{1-\alpha_1} i_1^{-\alpha_1} (1 - P_{\text{fail,bad}}) \\ &\qquad \times P^{i_2-1}_{\text{fail,bad}} \frac{\beta}{\gamma+\beta} \Bigg\}. \end{aligned} \quad (37)$$

Although it is difficult to get the closed form of $P(\tau^S > x)$, the simulations that will be presented in the next section, and also the form of (37), suggest that the delay has asymptotically a power-law tail (see Fig. 9). In summary, the packets of self-similar traffic input may experience longer queueing delay more frequently than those of Poisson traffic input. This is because for Poisson traffic $P(\tau^S > x)$ decays exponentially with x [8], [22].

As we have already mentioned, SM1 achieves the upper bound due to its energy-conserving approach and SM2 achieves the lower bound. However, when considering longer delays, i.e., $\tau \to \infty$, the two models converge in terms of $P(\tau_S > \tau)$. The occurrences of large delays are due to the occurrence of $A > c$ and the accumulation of excess data (i.e., $(A - c)$ bits per time slot) in the buffer over long ON periods. When $A > c$, no excess waiting time is necessary by SM1, and thus the two-server models behave nearly the same. Therefore, they have similar delays, which implies that large delays occur with nearly the same probability for both server models.

$$W_j = \begin{cases} -\tau^T_j, \\ \quad \text{if } Q_j + A_j\tau = c_j\tau N_j, \\ 0, \\ \quad \text{if } N_j\tau^T_j A_j > c_j, \\ \max\left\{\text{Res\_}X^S_j - N_j\tau^T_j, 0\right\}, \\ \quad \text{if } \text{Res\_}X^S_j A_j > c_j, \\ \max\left\{\text{Res\_}X^S_j + m_jY^S_j + (m_j - 1)X^S_j - N_j\tau^T_j - \text{Res\_}X^S_{j+m_j}, 0\right\}, \\ \quad \text{where } m_j = \left\lceil \frac{c_j - \text{Res\_}X^S_j A_j}{A_j X^S_j} \right\rceil, \\ \quad \text{if } m_jA_jX^S_j > c_j - \text{Res\_}X^S_j A_j > (m_j - 1)A_jX^S_j. \end{cases} \quad (35)$$

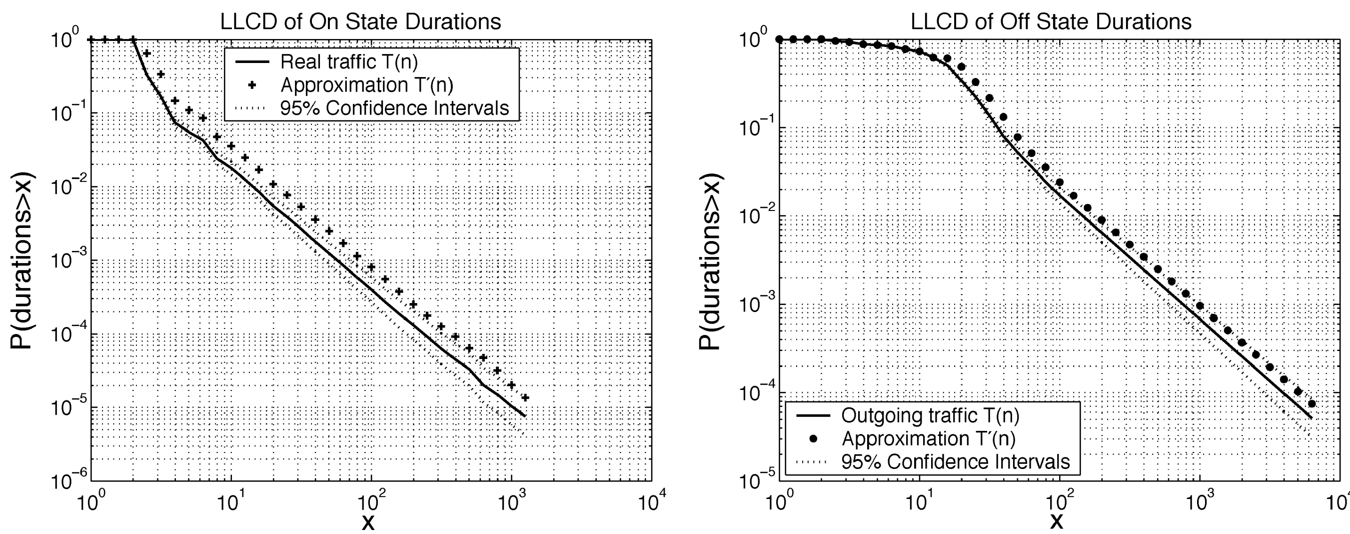

Fig. 2. Small buffer system: LLCD of ON durations (left) and OFF durations (right) of traffic obtained based on the approximation of (16) (dotted line), and outgoing process $T(n)$ obtained via (11) and (12) (solid line).

VI. Simulation Results

In this section, we provide simulation results to support the results on the effect of the gateway on the incoming traffic.

A. Data Synthesis

We generated the incoming traffic $S(n)$ as explained in Section III-A. The model parameters were estimated based on a data trace collected at the 100 Mb/s network of the Department of Electrical and Computer Engineering at Drexel University, Philadelphia, PA, as discussed in [41]. The ON/OFF durations were Pareto distributed as $\bar{F}\ (x; \alpha_1 = 1.6, K_1 = 1)$ and $\bar{F}(x; \alpha_0 = 1.4, K_0 = 1)$, respectively. The ON state rate was taken from a mixture cutoff Pareto distribution with survival function $\bar{F}_A\ (x; \alpha_I = 1.19, K_I = 48, \alpha_{II} = 6.5, L = 10^{4.64}, R = 1.25 \times 10^5)$. The time unit τ was taken as $\tau = 0.001$ s.

The channel state durations were taken to be independent exponentially distributed, with mean $1/\beta = 0.1$ s for good states, and $1/\gamma = 0.0333$ s for the bad states [16]. The bit error rate of good states was taken to be $P_{e,\text{good}} = 10^{-6}$ and that of bad states $P_{e,\text{bad}} = 0.01$. We adopted $\text{BCH}(n, k)$ (Bose–Chaudhuri–Hocquenghem) code [24] for forward-error correction (FEC) purposes. For good states, we chose a BCH(127,120) code with codeword length $n = 127$, payload $k = 120$, and maximum number of correctable bits per code $e = 1$. For the bad states, more redundancy is required to recover errors; thus, for those states, we chose a BCH(127,29) code, which can recover error bits up to $e = 21$. We set $P = 1.27$ (Mb/s), which corresponds to the amount of data during one time slot (1 ms) given an overall rate of 1.27 Mb/s. The cell consisted of ten blocks, with each block containing 127 bits encoded by the above described BCH code schemes. Based on the above parameters and via (10), we can get c_g (information bits/time slot) and c_b (information bits/time slot).

In the sequel, we will use the term *bits* (or bit rate) instead of *information bits* (or information bit rate).

B. Impact of the Buffering System on the Degree of Self-Similarity of Outgoing Traffic

1) Service Model 1—Small Buffer System: Here, we set $B = 1200$ bits and choose BCH(127,120) code for FEC. Via (10), we get $c = 1.2$ Mb/s and thus $c\tau = 1200$ bits. When deriving the analytical expression for this case we assume that the contribution of the previous buffer content can be ignored when deriving the distributions of $T(n)$. To verify this assumption, we performed the following simulation. Based on the synthesized incoming traffic, $S(n)$, we generated the outgoing traffic $T(n)$ based on (11) and (12). We also generated an approximation of the outgoing traffic, denoted by $T'(n)$, using (16) by ignoring the previous buffer content $Q(n-1)$. The log–log complementary distributions (LLCD) (both with 95% confidence intervals) corresponding to the ON and OFF durations of $T(n)$ and $T'(n)$ are shown in Fig. 2. One can see that the approximation falls in the 95% confidence intervals indicating that the approximation holds reasonably well.

For a small buffer system, the ON durations of $T(n)$ have the same distribution as those of $S(n)$, and the OFF durations satisfy (18). These results are supported by our simulations, as seen in Fig. 3 (with 95% confidence intervals), where the LLCD of ON/OFF duration of $T(n)$ $S(n)$ are plotted. Fig. 3 shows that $P(Y^T > y)$ is asymptotically power law with tail exponent $\min\{\alpha_1, \alpha_0\} = 1.4$. We also performed simulations for incoming traffic with $\alpha_1 < \alpha_0$, i.e., $\alpha_1 = 1.4$, $\alpha_0 = 1.6$. Fig. 4 shows the LLCD of ON/OFF duration of $T(n)$, where one can see that in accordance with Proposition 1, $\alpha_{X_T} = \alpha_1$ and $\alpha_{Y_T} = \min\{\alpha_1, \alpha_0\} = \alpha_1$.

2) Service Model 1—Large Buffer System: We here set $B = 1200$ bit, $c\tau = 290$ bit.

$T(n)$ was generated as in the previous case. According to Proposition 2, the tail index of X^T should be $\alpha_1 + 1 = 2.6$, and the tail index of Y^T should be equal to $\alpha_0 = 1.4$. Both results are confirmed by our simulations, which are presented in Fig. 5.

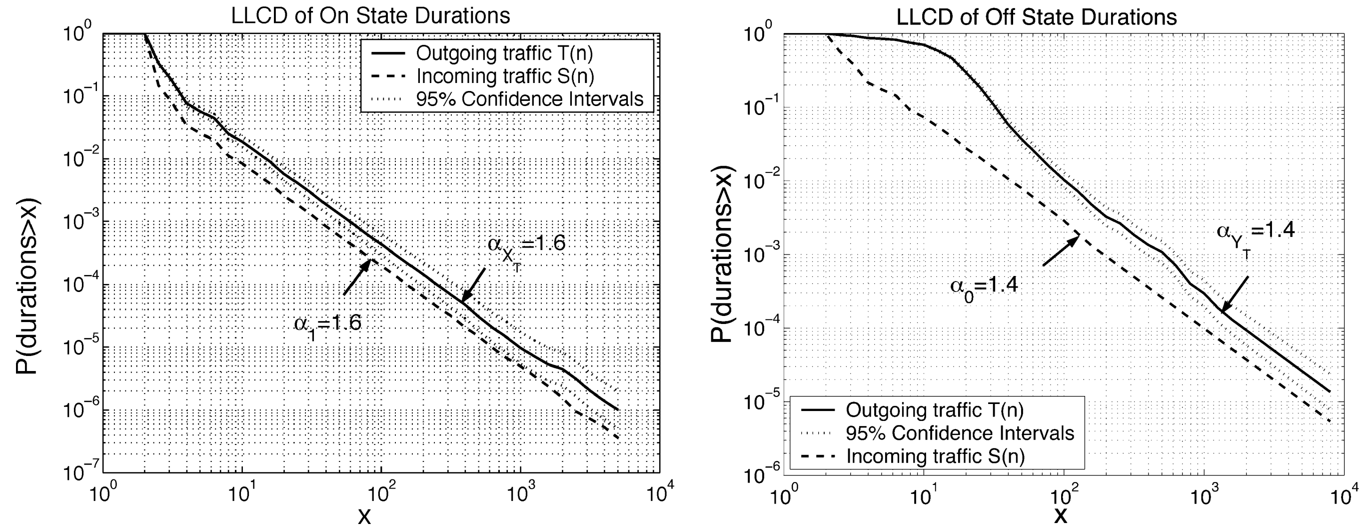

Fig. 3. Small buffer system under SM1: LLCD of ON durations (left) and OFF durations (right) of incoming process $S(n)$ (dashed line) and outgoing process $T(n)$ (solid line) with $\alpha_0 < \alpha_1$.

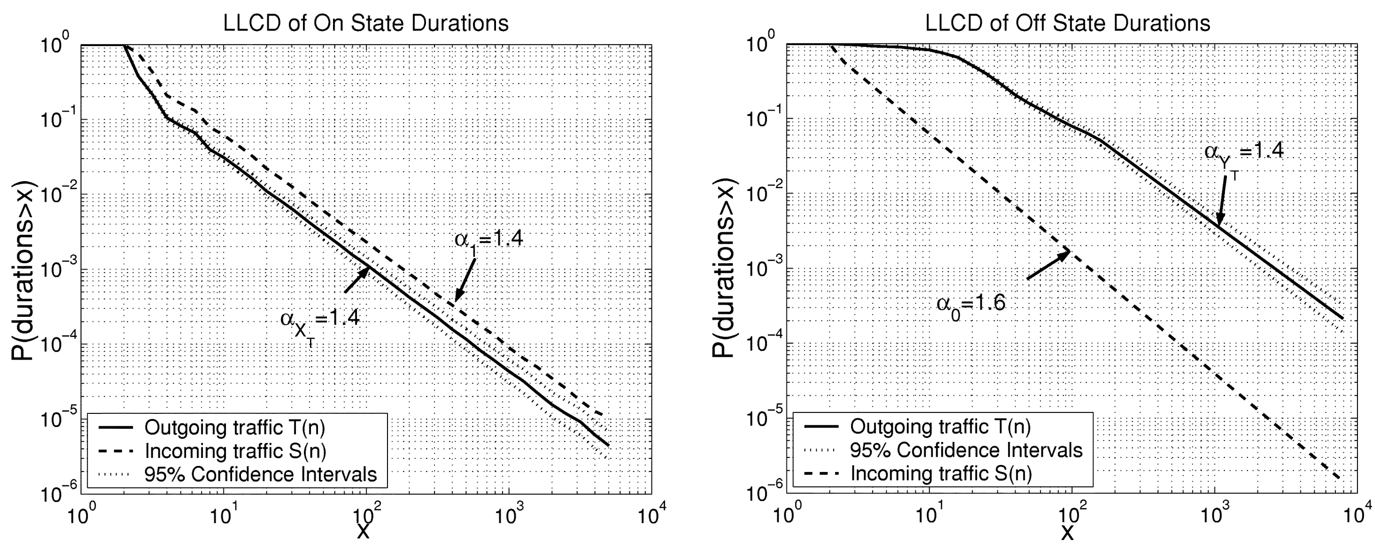

Fig. 4. Small buffer system under SM1: LLCD of ON periods (left) and OFF durations (right) of incoming process $S(n)$ (dotted line) and outgoing process $T(n)$ (solid line) with $\alpha_0 > \alpha_1$.

We also generated incoming traffic $S'(n)$ as before but with $\alpha_0 = 2.2$ (i.e., ON durations are heavy-tailed and OFF durations with finite variance). We set $B = 6000$ bits and choose BCH(127,120) and BCH(127,110) for FEC in good and bad channel state respectively, so we get $c_g = 1.2$ Mb/s and $c_b = 1.1$ Mb/s via (10). Thus, we have $c_g\tau = 1200$ bits, $c_b\tau = 1100$ bits, i.e., gateway always operates under large buffer model. For this case, our theoretical results suggest that the outgoing traffic will not be self-similar. The normalized variance-time plots of the outgoing traffic, $T'(n)$, along with that of $S'(n)$ are given in Fig. 7 (right). The slope of $T'(n)$ is -1, or equivalently, $H = 0.5$, indicating loss of self-similarity.

3) Service Model 1—Alternating Buffer Systems: Here we set $B = 1200$ bits. We choose BCH(127,120) and BCH(127,29) for FEC in good and bad channel state respectively. Via (10), we get $c_g = 1.2$ Mb/s and $c_b = 290$ Kb/s, so $c_g\tau = 1200$ bits and $c_b\tau = 290$ bits.

Therefore, during good channel states the buffering system satisfies $B = c\tau$ (small buffer system), while during bad states of channel the system satisfies $B \gg c\tau$ (large buffer system).

We use the same incoming traffic $S(n)$ generated according to the description of Section VI-A, and model the channel as discussed in the same Section.

The LLCDs of ON/OFF durations of $T(n)$ (with their 95% confidence intervals) and $S(n)$ are shown in Fig. 6. The tail indexes of ON and OFF states of $T(n)$ were estimated to be 1.6 and 1.4, respectively, which equal α_1 and α_0, as expected by our results of Section IV-C. The Hurst parameter of $T(n)$ is $H = 0.8$, which implies self-similarity. This can also be confirmed by looking at the normalized variance-time plot of $T(n)$ in Fig. 7 (left); its slope of $2H - 2 = -0.4$ indicates that the output traffic is self-similar.

4) Service Model 2: The analysis in Section IV-B concluded that the tail indexes of X^T and Y^T are nearly the same with

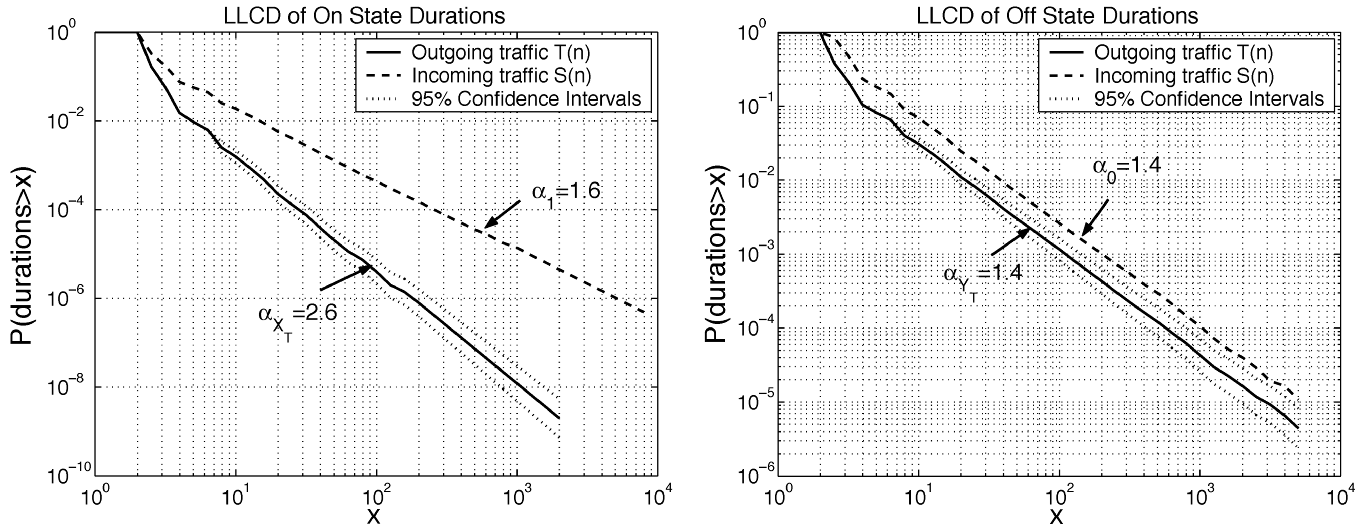

Fig. 5. Large buffer system under SM1: LLCD of ON durations (left) and OFF durations (right) of incoming process $S(n)$ (dashed line) and outgoing process $T(n)$ (solid line).

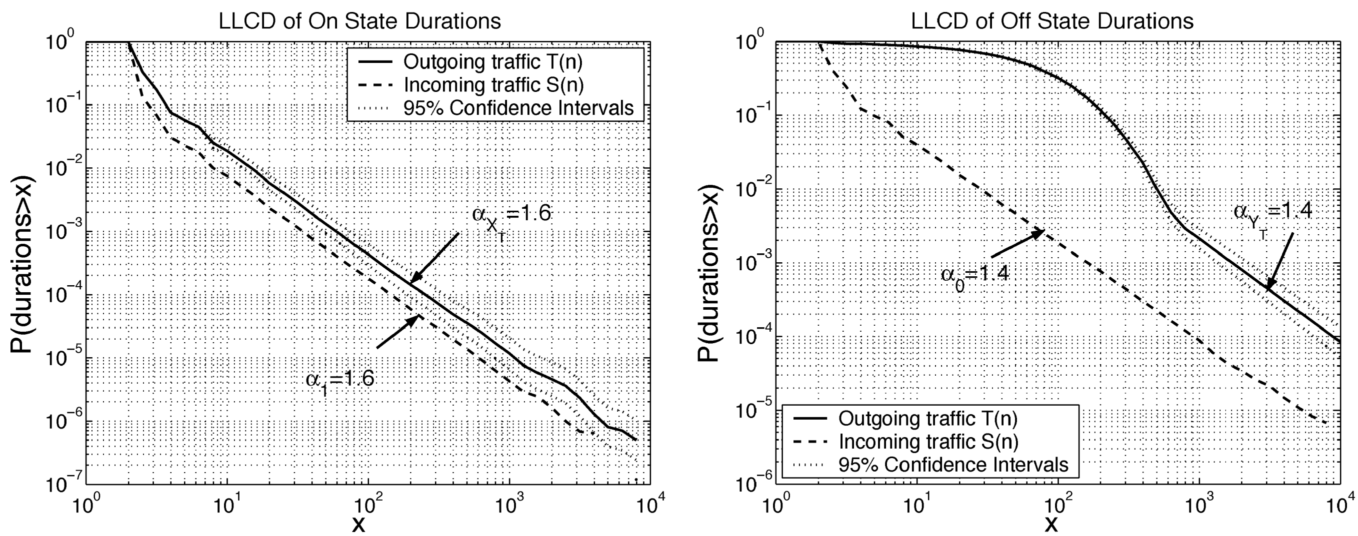

Fig. 6. Alternating buffer system under SM1: LLCD of ON periods (left) and OFF periods (right) of incoming process $S(n)$ (dashed line) and outgoing process $T(n)$ (solid line).

those of X^S and Y^S. The LLCDs of the ON/OFF durations of $S(n)$ and $T(n)$ (with 95% confidence intervals) are plotted in Fig. 8. According to our theoretical result, $T(n)$ will have a little longer ON durations and shorter OFF durations as compared to $S(n)$, i.e., $P(X^T > x) > P(X^S > x)$ and $P(Y^T > x) < P(Y^S > x)$, but the tail indexes of X^T, Y^T will be the same with those of X^S, Y^S respectively. This is confirmed in Fig. 8.

C. Delay Bounds

We assumed an infinite buffer to hold incoming packets. The size of the j_1j_2th packet (the j_2th packet in the j_1th ON period) of input traffic $S(n)$ is the amount of data bits coming into buffer during the $n^1_{j_1j_2}$th time slot. For the j_1j_2th packet, we recorded the time instants $n^1_{j_1j_2}$ and $n^2_{j_1j_2}$ as defined in Section V, and computed the delay of that packet as $\tau^S(j_1j_2) = n^2_{j_1j_2} - n^1_{j_1j_2}$.

Since our delay analysis is not given in closed form, we used the semiexperiment approach described in Section V-A to compute the survival function of the delays. We performed 10^6 Monte Carlo runs, in each run generating N, τ^T, A, X, Y, $\mathrm{Res\_}X^S$ according to their distributions given by (54), (28), $\bar{F}_A$ (x;$\alpha_I = 1.19$, $K_I = 48$, $\alpha_{II} = 6.5$, $L = 10^{4.64}$, $R = 1.25 \times 10^5$), $\bar{F}$ $(x; \alpha_1 = 1.6, K_1 = 1)$, $\bar{F}$ $(x; \alpha_0 = 1.4, K_0 = 1)$, (56), respectively. Based on the generated samples we computed the excess waiting time W via (35), and then computed τ^S via (31). Fig. 9 (left) shows the survival function based on simulations (dashed lines) and the semiexperimental approach (solid line—also denoted as "Analytic result"). The two curves are in close agreement, which suggests the analytical result is accurate.

Fig. 9 (left) indicates that the probabilities $P(\tau^S > x)$ for SM1 and SM2 are the same when x is larger than $10^{2.3} \approx$ 200 slots.

Also from Fig. 9 we can see that when x takes large values, the curves corresponding to SM1 and SM2 show a power-law

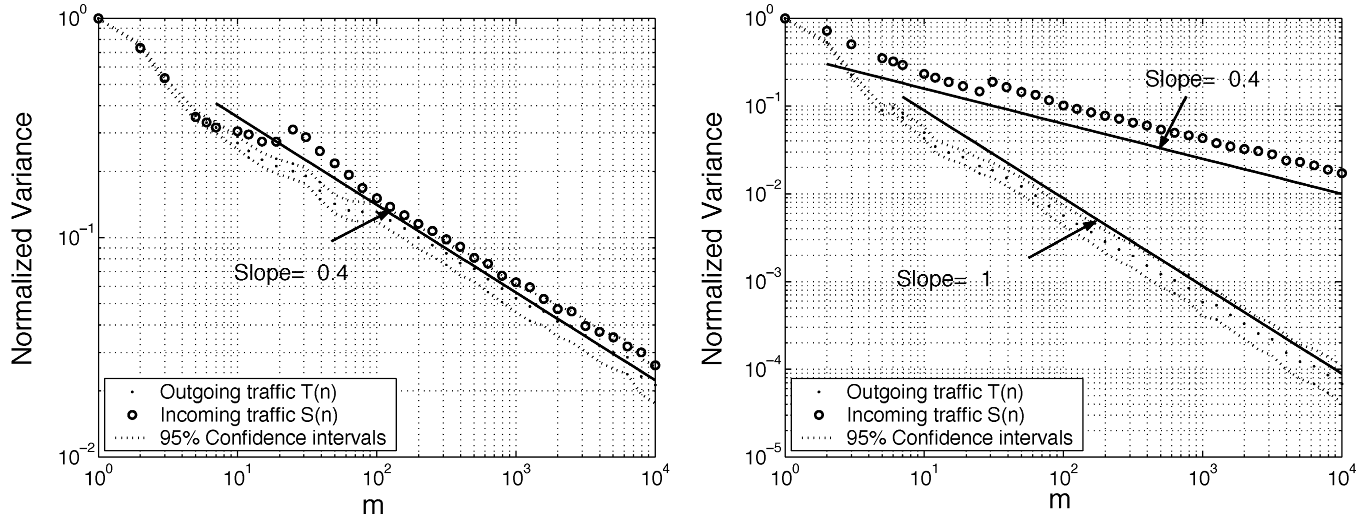

Fig. 7. Normalized variance-time plot of outgoing traffic for a large buffer system under SM1 (right) and an alternating buffer system under SM1 (left).

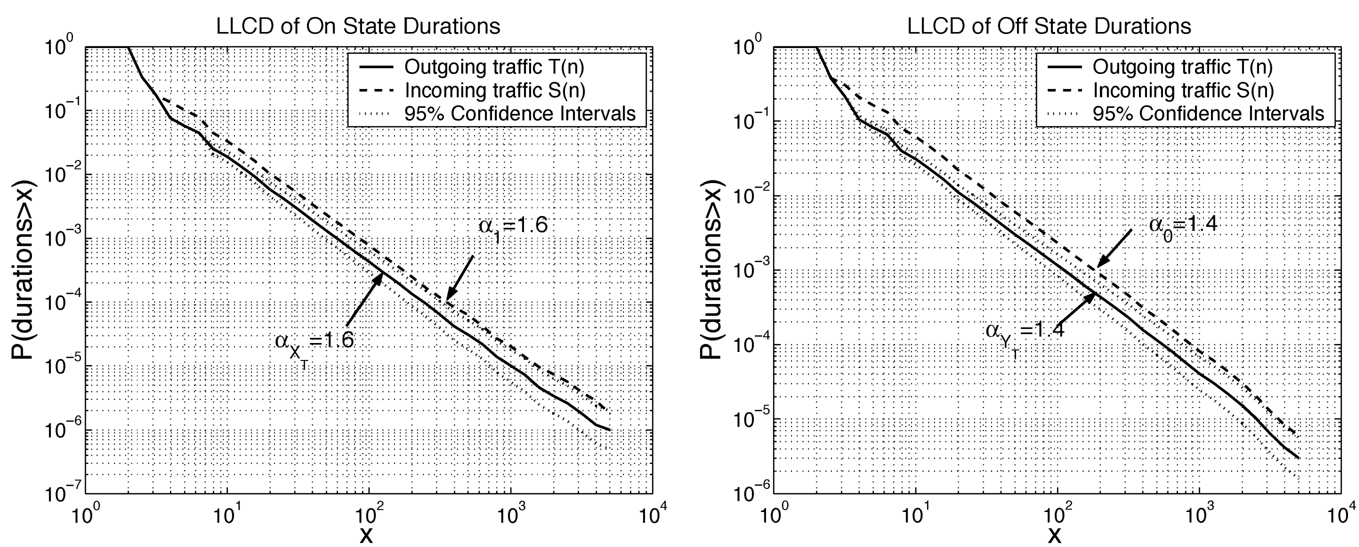

Fig. 8. LLCD of ON durations (left) and OFF durations (right) of incoming process $S(n)$ (dashed line) and outgoing process $T(n)$ (solid line). $T(n)$ was obtained by passing $S(n)$ through the buffering system operating under SM2.

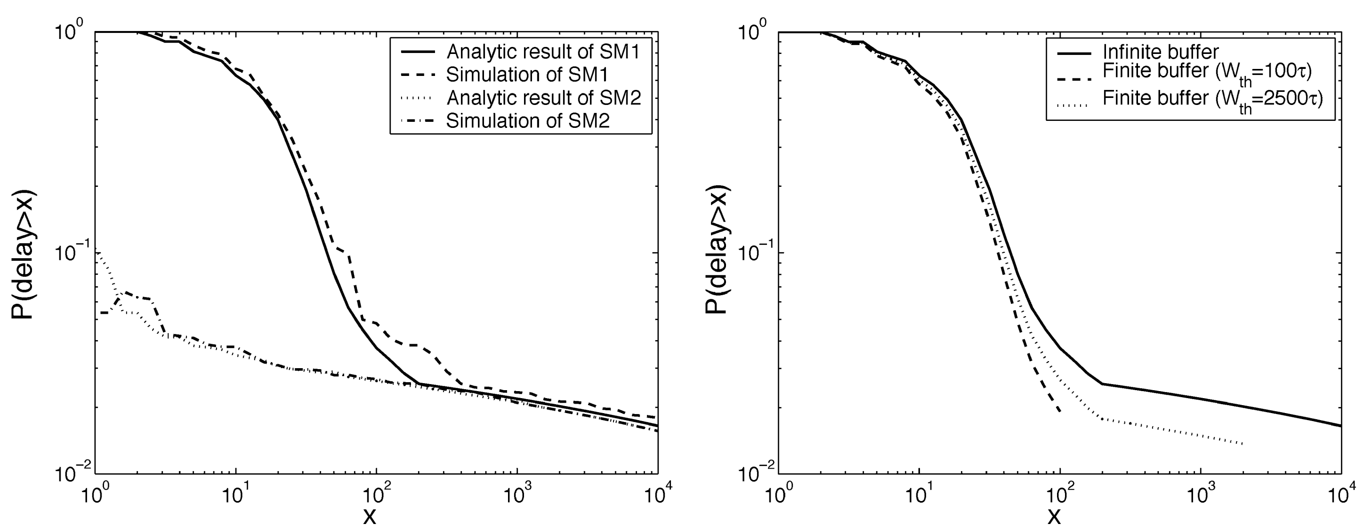

Fig. 9. Cumulative distribution function of delay.

decaying tail with exponent 0.18. This implies that the packets of self-similar inputs have longer delays than those of Poisson traffic input. For Poisson input, $P(\tau^S > x)$ decays with x in an exponential fashion.

Next, we provide some simulations results on the delay distribution for a practical system where there is a timeout mechanism and for packets whose waiting times exceed a certain threshold are discarded. We repeated the above simulation by imposing a timeout threshold W_{th}. If a packet stays in the buffer longer than that threshold, i.e., $W_j \geq W_{\mathrm{th}}$, that packet is discarded. The survival function results are shown in Fig. 9 (right) for the simulation with infinite buffer (equivalently, $W_{\mathrm{th}} = \infty$) and finite buffers of $W_{\mathrm{th}} = 100\tau$, $W_{\mathrm{th}} = 2500\tau$ ($\tau = 0.001$ s). One can see the probability of delays exceeding a certain value of x match our analysis as the W_{th} increases.

VII. Conclusion

We studied the output of a gateway fed by self-similar traffic of low to moderate rate. We conclude that a gateway operating under an instant transferring protocol (SM2) will produce self-similar traffic. On the other hand, a gateway operating under an energy conserving protocol (SM1) can, under certain circumstances, generate traffic that is not self-similar. In particular, if the gateway is equipped with a buffer that can store a large number of packets (large buffer model), the statistics of the ON/OFF durations of the output traffic can change in comparison to the corresponding states of the incoming traffic. For non-real-time variable-bit-rate multimedia traffic [13], where the OFF states have finite variance (i.e., $\alpha_0 = 2$), the change in the statistics will be such that it will result in loss of self-similarity.

Identifying the cases where wireless traffic is self-similar, or understanding what is causing self-similarity or the lack there of, is important in network engineering. The self-similar nature of traffic implies that the traffic rate fluctuates far away from its mean value (if existed), with nonnegligible probability. Thus, the performance in any system with fixed rate may degrade significantly if the system is either overloaded (the rate is too large compared with the allocated resources), or overprovisioned (the rate is small). In addition, a static resource scheduling/allocation scheme is not efficient for data traffic. The self-similarity of traffic indicates a nontrivial predictive structure at coarse time scales, which can be exploited for the design of dynamic resource management to improve system performance.

We also conclude that the survival functions of delays experienced by packets at a gateway under SM1 or SM2 become identical as the delay gets large, both exhibiting a power-law decay. Since the delay of SM1 represents an upper bound and that of SM2 a lower bound, all gateway models will also behave in a similar fashion for large delays. These delay results imply that packets may stay in the buffer for a very long time with nontrivial probability. Thus, the buffering system can become ineffective in handling self-similar input traffic in the sense that increasing the buffer size cannot achieve proportional reduction of packet loss rate occurring due to the buffer overflow. The theoretic results on delay analysis given in this paper can be used for the user admission control according to its quality of service (QoS), i.e., a new user is admitted if its QoS requirements can be fulfilled, and, *vice versa*, denied if not.

Appendix A
Proof of Proposition 1

Via assumptions A1) and A2), we approximate the complimentary distribution function (CDF) of X^T and Y^T (see (13) and (14)) by $P(X^T > x|A \geq c)P(A \geq c)$ and $P(Y^T > x|A < c)$, respectively. We next calculate these two probabilities by considering the following cases.

Case 1) $A_j \geq c$. In this case, we have $Q(n-1) + S(n)\tau = Q(n-1) + A_j\tau > c\tau$, and thus $T(n)$ will also be in the ON state. Therefore, we have $P(X^T > x) \approx P(X^T > x|A > c) \approx P(X^S > x)$.

Case 2) $A_j < c$. The buffer system virtually converts the jth ON period of $S(n)$ into part of an OFF period of $T(n)$, which results in $T(n)$ staying in OFF state for a larger duration that is equal to $Y_{j-1}^S + X_j^S + Y_j^S$. Such *combining* action can be extended to more ON/OFF periods of $S(n)$ to produce an even bigger OFF period for $T(n)$, of duration $Y_k^T = Y_{j-1}^S + \sum_{i=j}^{j+M-1}(X_j^S + Y_j^S)$ given that $A_j, A_{j+1}, \ldots, A_{j+M-1} < c$. Thus, M consecutive ON periods of $S(n)$, which satisfy $A_i < c$ $(i = j, \ldots, j+M-1)$, are combined together. The probability for a such combination is $P(M = m) = P(A < c)^m P(A \geq c)^2$.

Based on the above, we get

$$\begin{aligned} P(Y^T > y) &\approx P(Y^T > y|A < c) \\ &\approx \sum_{m=1}^{\infty} P\left(\sum_{i=1}^{m} X_i^S + \sum_{i=1}^{m+1} Y_i^S > y\right) \\ &\quad \times P(A < c)^m P(A \geq c)^2 \\ &= \sum_{m=1}^{\infty} P\left(\sum_{i=1}^{m} X_i^S + \sum_{i=1}^{m+1} Y_i^S > y\right) \\ &\quad \times \left[1 - \left(\frac{K_I}{c}\right)^{\alpha_I}\right]^m \left(\frac{K_I}{c}\right)^{2\alpha_I} \end{aligned} \tag{38}$$

where $1 - P(A < c) = P(A \geq c) = \bar{F}_A(c; \alpha_I, K_I, \alpha_{II}, L, R)$ (see (7)) and via A3) it holds $c < L$.

If $\alpha_0 = 2$, via the central limit theorem arguments, the term $\sum_{i=1}^{m} X_i^S$ will be asymptotically Gaussian; thus, as $y \to \infty$, its contribution in (38) can be ignored.

If $\alpha_0 < 2$, according to properties of Pareto distributions, for $Z \triangleq \sum_{i=1}^{m} X_i^S + \sum_{i=1}^{m+1} Y_i^S$, it holds as $y \to \infty$ [2]

$$P(Z \geq y) = \bar{F}_Z(y) \sim y^{-\alpha} L_m(y) \tag{39}$$

where $K_m = mK_1 + (m+1)K_0$, $\alpha = \min\{\alpha_1, \alpha_0\}$ and $L_m(y)$ is a slow-varying function.

And thus we have

$$\begin{aligned} P(X^T > x) &\approx P(X^S > x) \\ P(Y^T > y) &\overset{y\to\infty}{\sim} \left\{\sum_{m=1}^{\infty}(2m+1)K_m^{\alpha} L_m(y) \right. \\ &\quad \left. \times \left[1 - \left(\frac{K_I}{c}\right)^{\alpha_I}\right]^m \left(\frac{K_I}{c}\right)^{2\alpha_I}\right\} y^{-\alpha} \\ &= L(y) y^{-\alpha} \end{aligned} \tag{40}$$

where $L(y)$ can be easily inferred from the above equation. It is easy to see that $\lim_{\tau\to\infty}(L(\tau y)/L(y)) = 1$, which means that $L(y)$ is a slowly varying function.

If Y_i^S is exponentially distributed, $P(Z \geq y)$ it holds that

$$P(Z\geq y) \overset{y\to\infty}{\sim} P\left(\sum_{i=1}^{m} X_i^S \geq y\right) \sim m\left(\frac{y}{MK_1}\right)^{-\alpha_1} L_m(y). \tag{41}$$

Therefore, we have

$$\begin{aligned} P(Y^T > y) &\overset{y\to\infty}{\sim} \left\{\sum_{m=1}^{\infty} m(MK_1)^{\alpha_1} L_m(y)\right. \\ &\qquad \left.\times \left[1-\left(\frac{K_I}{c}\right)^{\alpha_I}\right]^m \left(\frac{K_I}{c}\right)^{2\alpha_I}\right\} y^{-\alpha_1} \\ &= L'(y) y^{-\alpha_1}. \end{aligned} \tag{42}$$

It is easy to see that $L'(y)$ is also a slowly varying function.

Appendix B
Proof of Proposition 2

Based on the assumption A1), i.e., in a low-to-moderate-load scenario, the buffer system is always overprovisioned. In other words, the buffer content at the regenerative points (i.e., the beginning of ON periods) is less than $c\tau$ with high probability; thus, we can approximate it as $P(Q_e < c\tau) \approx 1$. Therefore, $P(X^T > x) \approx P(X^T > x | A < c, Q_e < c\tau)$.

The probability $P(X^T > x | A < c, Q_e < c\tau)$ can be calculated as the probability of the intersection of the following three conditions.

- E1) There is a period of time with length l ($l = 0, 1, 2, \ldots$), when $T(n) = 0$ even if $S(n) > 0$.
- E2) The sum of the previous buffer content Q and the accumulating amount of incoming bits during the mean time is enough to form at least x packets.
- E3) The corresponding ON period of $S(n)$, X^S, should be larger than $x + l$.

The following observations will also be used in the derivations.

- F1) The buffer content Q can only take nonnegative values.
- F2) According to A3), we have $\min(c, L) = c$.
- F3) According to A1), i.e., $K_A \ll c$, we have $K_A < (x-1)c/x$.
- F4) The steady-state distribution of Q can be found in (33).
- F5) For the Pareto-distributed ON durations of $S(n)$, it holds $P(X^S > x + l) = (K_1/(x+l))^{\alpha_1}$.
- F6) We find that $0 \leq l \leq x/(x-1) < 2$, or simply$l = 0, 1$.

By combining E1)–E3) and F1)–F6), we get

$$\begin{aligned} &P(X^T > x | A < c, Q_e < c\tau) \\ &\quad \overset{x\to\infty}{\sim} \sum_{l=0}^{1} \int_{\frac{x-1}{x}c}^{c} \left\{C_Q(c\tau - la\tau)^{(1-\alpha_1)}\right. \\ &\qquad\qquad \left. - C_Q\left[xc\tau - (x+l)a\tau\right]^{(1-\alpha_1)}\right\} \\ &\qquad \times \left(\frac{K_1}{x+l}\right)^{\alpha_1} f_A(a)da \end{aligned} \tag{43}$$

where $f_A(a)$ is the pdf of ON state rates of $S(n)$ (see (8)) and C_Q is a constant related to the distribution of Q (see (33)).

When x takes large values, $x + l \approx x$ ($l = 0, 1$), and (43) can be approximated as

$$\begin{aligned} &P(X^T > x | A < c, Q_e < c\tau) \\ &\overset{x\to\infty}{\sim} K_1^{\alpha_1}\tau^{(1-\alpha_1)} \\ &\quad \times \sum_{l=0}^{1} \int_{\frac{x-1}{x}c}^{c} \left\{\left[C_Q(c-la)^{(1-\alpha_1)}\right] x^{-\alpha_1}\right. \\ &\qquad\qquad \left. - \left[C_Q(c-a)^{(1-\alpha_1)}\right] x^{-(2\alpha_1-1)}\right\} f_A(a)da \\ &\approx K_1^{\alpha_1}\tau^{(1-\alpha_1)} \left\{\sum_{l=0}^{1} \int_{\frac{x-1}{x}c}^{c} \left[C_Q(c-la)^{(1-\alpha_1)}\right] f_A(a)da\right\} \\ &\quad \times x^{-\alpha_1} - 2K_1^{\alpha_1}\tau^{(1-\alpha_1)} \\ &\quad \times \left\{\int_{\frac{x-1}{x}c}^{c} \left[C_Q(c-a)^{(1-\alpha_1)}\right] f_A(a)da\right\} x^{-(2\alpha_1-1)}. \end{aligned} \tag{44}$$

To simplify the above equation, let us define $I(l) = \int_{(x-1)c/x}^{c} [C_Q(c-la)^{(1-\alpha_1)}] f_A(a)da$ for $l = 0,1$. When $l = 0$, we have

$$\begin{aligned} I(0) &= C_Q c^{(1-\alpha_1)} \int_{\frac{x-1}{x}c}^{c} dF_A(a) \\ &= C_Q c^{(1-\alpha_1)} K_I^{\alpha_I} \left[\left(\frac{x-1}{x}c\right)^{-\alpha_I} - c^{-\alpha_I}\right] \\ &= C_Q c^{(1-\alpha_1-\alpha_I)} K_I^{\alpha_I} \left[\left(1-\frac{1}{x}\right)^{-\alpha_I} - 1\right]. \end{aligned} \tag{45}$$

For large x, we can use the approximation $(1 - 1/x)^{-\alpha_I} \sim 1 + \alpha_I/x$, based on which (45) becomes

$$I(0) \sim \alpha_I C_Q c^{(1-\alpha_1-\alpha_I)} K_I^{\alpha_I} x^{-1}. \tag{46}$$

Similarly, for $l = 1$, we have

$$\begin{aligned} I(1) &= \int_{\frac{x-1}{x}c}^{c} C_Q(c-a)^{(1-\alpha_1)} \alpha_I K_I^{\alpha_I} a^{-(1+\alpha_I)} da \\ &= \alpha_I K_I^{\alpha_I} C_Q c^{(1-\alpha_1)} \\ &\quad \times \int_{\frac{x-1}{x}c}^{c} \left(1-\frac{a}{c}\right)^{(1-\alpha_1)} a^{-(1+\alpha_I)} da. \end{aligned} \tag{47}$$

Using Taylor series expansion, we have

$$\begin{aligned} I(1) =& \alpha_I K_I^{\alpha_I} C_Q c^{(1-\alpha_1)} \\ &\times \int_{\frac{x-1}{x}c}^{c} \left[1 + m\frac{a}{c} + \sum_{n=2}^{\infty} \frac{m(m+1)\cdots(m+n-1)}{n!} \times \left(\frac{a}{c}\right)^n\right] a^{-(1+\alpha_I)} da \\ =& \alpha_I K_I^{\alpha_I} C_Q c^{(1-\alpha_1-\alpha_I)} \\ &\times \left\{ \frac{1}{\alpha_I}\left[\left(1-\frac{1}{x}\right)^{-\alpha_I} - 1\right] + \frac{m}{1-\alpha_I}\left[1-\left(1-\frac{1}{x}\right)^{(1-\alpha_I)}\right] \right. \\ &\left. + \sum_{n=2}^{\infty} \frac{m(m+1)\cdots(m+n-1)}{n!}\frac{1}{n-\alpha_I} \times \left[1-\left(1-\frac{1}{x}\right)^{(n-\alpha_I)}\right]\right\} \end{aligned}$$

where $m = -(1-\alpha_1)$.

Using the approximating $(1-1/x)^{(n-\alpha_I)} - 1 \overset{x\to\infty}{\sim} -(n-\alpha_I)x^{-1}$, we get

$$I(1) \overset{x\to\infty}{\sim} \alpha_I K_I^{\alpha_I} C_Q Z c^{(1-\alpha_1-\alpha_I)} x^{-1} \tag{48}$$

where Z is a constant. Finally, for $1 < \alpha_1 < 2$, we have

$$\begin{aligned} P(X^T > x) \approx& P(X^T > x | A < c, Q_e < c) \\ =& \alpha_I C_Q K_I^{\alpha_I} K_1^{\alpha_1} \tau^{(1-\alpha_1)} c^{(1-\alpha_1-\alpha_I)} \times [1+Z] x^{-(\alpha_1+1)} \\ =& C_1 x^{-(\alpha_1+1)}. \end{aligned} \tag{49}$$

Appendix C
Distribution Functions

(D1): The Distribution of N:

Since $N = \lfloor (Q + A\tau)/c\tau \rfloor$, the distribution function of N can be calculated as

$$P(N = i | \text{service rate} = \text{c}) = \int_{ic\tau}^{(i+1)c\tau} f_Q(x) \otimes \left[\frac{1}{\tau} f_A\left(\frac{x}{\tau}\right)\right] dx \qquad i = 1, 2, \ldots \tag{50}$$

where $f_Q(x)$, $f_A(x)$ are the pdf of Q and A (see (33) and (8), respectively).

Therefore, (50) can be written, for large i, as

$$\begin{aligned} P(N = i | &\text{service rate} = \text{c}) \\ \sim& \int_{ic}^{(i+1)c} \left\{ \left(\frac{K_{II}}{R}\right)^{\alpha_{II}} (x-R)^{-\alpha_1} + \int_{K_I}^{\min(x,L)} \alpha_I K_I^{\alpha_I} \upsilon^{-(\alpha_I+1)} (x-\upsilon)^{-\alpha_1} d\upsilon \right. \\ &\left. + \left[\int_{L}^{\min(x,R)} \alpha_{II} K_{II}^{\alpha_{II}} \upsilon^{-(\alpha_{II}+1)} (x-\upsilon)^{-\alpha_1} d\upsilon\right] \times u(x-L) \right\} dx \end{aligned} \tag{51}$$

where $u(.)$ is a step function. When i takes large values, x is also large due to $x \in [ic, (i+1)c)$. Thus, we can get: $\min(x, L) \approx L$, $\min(x, R) \approx R$, $u(x-L) = 1$ and

$$\begin{aligned} P(N = i | &\text{service rate} = \text{c}) \\ \sim& \int_{ic}^{(i+1)c} \left\{ \left(\frac{K_{II}}{R}\right)^{\alpha_{II}} (x-R)^{-\alpha_1} + \int_{K_I}^{L} \alpha_I K_I^{\alpha_I} \upsilon^{-(\alpha_I+1)} (x-\upsilon)^{-\alpha_1} d\upsilon \right. \\ &\left. + \int_{L}^{R} \alpha_{II} K_{II}^{\alpha_{II}} \upsilon^{-(\alpha_{II}+1)} (x-\upsilon)^{-\alpha_1} d\upsilon \right\} dx. \end{aligned} \tag{52}$$

Moreover, we can have the following approximations: $(x-\upsilon)^{-\alpha_1} \sim x^{-\alpha_1}$ and $(x-R)^{-\alpha_1} \sim x^{-\alpha_1}$ because $\upsilon \le R \ll x$. Substituting these two approximations into (52), we have

$$\begin{aligned} P(N = i | &\text{service rate} = \text{c}) \\ =& \left[\left(\frac{K_{II}}{R}\right)^{\alpha_{II}} + K_I^{\alpha_I}\left(K_I^{-\alpha_I} - L^{-\alpha_I}\right) + K_{II}^{\alpha_{II}}(L^{-\alpha_{II}} - R^{-\alpha_{II}})\right] \int_{ic}^{ic+c} x^{-\alpha_1} dx \\ =& C_N \frac{1}{1-\alpha_1}\left[ic^{(1-\alpha_1)} - (ic+c)^{(1-\alpha_1)}\right] \end{aligned} \tag{53}$$

where $C_N \triangleq (K_{II}/R)^{\alpha_{II}} + K_I^{\alpha_I}(K_I^{-\alpha_I} - L^{-\alpha_I}) + K_{II}^{\alpha_{II}}(L^{-\alpha_{II}} - R^{-\alpha_{II}})$. After Taylor series expansion and using the approximation $(ic+c)^{(1-\alpha_1)} \overset{i\to\infty}{\sim} (ic)^{-m}[1 - m(1/i)^n]$, where $m = \alpha_1 - 1$, we have

$$P(N = i | \text{service rate} = \text{c}) \sim C_N c^{1-\alpha_1} i^{-\alpha_1}, \qquad \text{as } i \to \infty. \tag{54}$$

(D2): The Distribution of $N\tau^T$:

We assume that N is independent of τ^T. We validate this assumption through the simulation in Section V-B. Thus, we get

$$\begin{aligned} P(N\tau^T = k) =& \sum_{i_1 i_2 = k, i_1, i_2 = 1, 2, \ldots} \left[P(N = i_1, \tau^T = i_2 | c = c_g) P(c = c_g) + P(N = i_1, \tau^T = i_2 | c = c_b) P(c = c_b)\right] \\ =& \sum_{i_1 i_2 = k, i_1, i_2 = 1, 2, \ldots} \left[P(N = i_1 | c = c_g) P(\tau^T = i_2 | c = c_g) \times P(c = c_g) + P(N = i_1 | c = c_b) \times P(\tau^T = i_2 | c = c_b) P(c = c_b)\right] \\ =& \sum_{i_1 i_2 = k, i_1, i_2 = 1, 2, \ldots} \left\{ c_g^{1-\alpha_1} i_1^{-\alpha_1} (1 - P_{\text{fail,good}}) P_{\text{fail,good}}^{i_2 - 1} \times \frac{\gamma}{\gamma + \beta} + c_b^{1-\alpha_1} i_1^{-\alpha_1} (1 - P_{\text{fail,bad}}) \times P_{\text{fail,bad}}^{i_2 - 1} \frac{\beta}{\gamma + \beta} \right\} \end{aligned} \tag{55}$$

where $c_1 = c_g$, $c_2 = c_b$, and $P_{\text{fail,state}}$ (state = {good,bad}).

(D3): The Survival Function of $\mathrm{Res\_}X^S$*:*

The stationary complimentary distribution function of the residue life of an ON period, $\bar{F}_{\mathrm{Res\_}X^S}(x)$ was given in [38] for $\alpha_1 > 1$ as

$$\begin{aligned}\bar{F}_{\mathrm{Res\_}X^S}(x) &= P(\mathrm{Res\_}X^S > x)\\ &= \frac{1}{\mu_1}\int_x^{\infty}\left(\frac{K_1}{u}\right)^{\alpha_1} du\\ &= \frac{\alpha_1 K_1^{\alpha_1} x^{1-\alpha_1}}{\mu_1(\alpha_1 - 1)}.\end{aligned} \tag{56}$$

ACKNOWLEDGMENT

The authors would like to thank the associate editor and the anonymous reviewers for several suggestions that improved the final paper.

REFERENCES

[1] G. Almes, S. Kalidindi, and M. Zekauskas, *A One-Way Delay Metric for IPPM* RFC 2679, 1999.
[2] B. C. Armold, *Pareto Distributions*. Baltimore, MD: International, 1983.
[3] S. Asmussen, *Applied Probability and Queues*. New York: Wiley, 1987.
[4] S. Banerjee, A. Misra, J. Yeo, and A. Agrawala, "Energy-efficient broadcast and multicast trees for reliable wireless communication," in *Proc. IEEE Wireless Communications Networking Conf. (WCNC 2003)*, Mar. 16–20, 2003, vol. 1, pp. 660–667.
[5] J. Beran, *Statistics for Long-Memory Process*. New York: Chapman & Hall, 1994.
[6] F. Brichet, J. Roberts, A. Simonian, and D. Veitch, "Heavy traffic analysis of a storage model with long range dependent ON/OFF sources," *Queueing Syst.*, vol. 23, pp. 197–215, 1996.
[7] O. Cappe, E. Moulines, J.-C. Pesquet, A. Petropulu, and X. Yang, "Traffic modeling for high-speed communication networks," *IEEE Signal Process. Mag. (Special Issue on Network Traffic Modeling)*, vol. 19, no. 3, pp. 14–27, May 2002.
[8] D. Mitra, "Stochastic theory of a fluid model of products and consumers couples by a buffer," *Adv. Appl. Probability*, vol. 20, pp. 646–676, 1988.
[9] Y. Eisenberg, C. E. Luna, T. N. Pappas, R. Berry, and A. K. Katsaggelos, "Joint source coding and transmission power management for energy efficient wireless video communications," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 12, no. 6, pp. 411–424, Jun. 2002.
[10] Z. Harpantidou and M. Paterakis, "Random multiple access of broadcast channels with pareto distributed packet interarrival times," *IEEE Pers. Commu.*, vol. 5, no. 2, pp. 48–55, Apr. 1998.
[11] E. P. Harris, S. W. Depp, E. Pence, S. Kirkpatrick, M. Sri-Jayantha, and R. R. Troutman, "Technology directions for portable computers," *Proc. IEEE*, vol. 83, no. 4, pp. 636–658, Apr. 1995.
[12] D. Heath, S. Resnick, and G. Samorodnitsky, "Patterns of buffer overflow in a class of queues with long memory in the input stream," *Ann. Appl. Probab.*, vol. 5, pp. 1021–1037, 1997.
[13] E. Hossain and V. K. Bhargava, "Link-level traffic scheduling for providing predictive QoS in wireless multimedia networks," *IEEE Trans. Multimedia*, vol. 6, no. 1, pp. 199–217, Feb. 2004.
[14] R. Jain and S. A. Routhier, "Packet trains: measurements and a new model for computer network traffic," *IEEE J. Sel. Areas Commun.*, vol. 4, pp. 986–995, 1986.
[15] P. R. Jelenkovic and A. A. Lazar, "General applied probability—Asymptotic results for multiplexing subexponential ON-OFF sources," *Adv. Appl. Probab.*, vol. 31, no. 2, pp. 394–421, 1979.
[16] J. G. Kim and M. M. Krunz, "Bandwidth allocation in wireless networks with guaranteed packet-loss performance," *IEEE/ACM Trans. Netw.*, vol. 8, no. 3, pp. 337–349, Jun. 2000.
[17] W. E. Leland, M. S. Taqqu, W. Willinger, and D. V. Wilson, "On the self-similar nature of Ethernet traffic (extended version)," *IEEE/ACM Trans. Netw.*, vol. 2, no. 1, pp. 1–15, Feb. 1994.
[18] S. B. Lowen and M. C. Teich, "Fractal renewal processes generate 1/f noise," *Phys. Rev. E*, vol. 47, pp. 992–1001, Feb. 1993.
[19] C. L. Nikias and M. Shao, *Signal Processing With Alpha-Stable Distributions and Applications*. New York: Wiley, 1995.
[20] I. Norros, "On the use of fractional Brownian motion in the theory of connectionless networks," *IEEE J. Sel. Areas Commun.*, vol. 13, no. 6, pp. 953–962, 1995.
[21] K. Park, G. Kim, and M. Crovella, "On the effect of traffic self-similarity on network performance," in *Proc. 1997 SPIE Int. Conf. Performance Control Network Systems*, Nov. 1997, pp. 296–310.
[22] K. Park and W. Willinger, Eds., *Self-Similar Network Traffic and Performance Evaluation*. New York: Wiley, 2000.
[23] L. L. Peterson and B. S. Davie, *Computer Networks-A Systems Approach*, 2nd ed. San Mateo, CA: Morgan Kaufmann, 2000.
[24] J. G. Proakis, *Digital Communications*, 4th ed. New York: McGraw-Hill, 2000.
[25] B. V. Rao, K. R. Krishnan, and D. P. Heyman, "Performance of finite-buffer queues under traffic with long-range dependence," in *Proc. IEEE Global Telecommunications Conf. (IEEE GLOBECOM): Communications: The Key to Global Prosperity*, Nov. 18–22, 1996, vol. 1, pp. 607–611.
[26] J. Redi and D. Avresky, "Performance of energy-conserving access protocols under self-similar traffic," in *Proc. IEEE Wireless Communications Networking Conf. (WCNC 1999)*, Sep. 1999, vol. 2, pp. 626–630.
[27] Z. Sahinoglu and S. Tekinay, "On multimedia networks: Self-similar traffic and network performance," *IEEE Commun. Mag.*, vol. 37, no. 1, pp. 48–52, Jan. 1999.
[28] G. Samorodnitsky and M. S. Taqqu, *Stable Non-Gaussian Random Processes: Stochastic Models With infinite Variance*. New York: Chapman & Hall, 1994.
[29] Y. Zhou and H. Sethu, "A simulation study of the impact of switching systems on self-similar properties of traffic," in *Proc. IEEE Workshop Statistical Signal Array Processing (SSAP)*, Pocono Manor, PA, Aug. 2000, pp. 500–504.
[30] B. Sikdar and K. S. Vastola, "The effect of TCP on the self-similarity of network traffic," presented at the Conf. Information Science Systems, Baltimore, MD, Mar. 21–23, 2001.
[31] B. Sikdar and K. S. Vastola, "Issues in TCP and Self-Similarity on Network Traffic," Networks Laboratory, ECSEm Dept., Rensselaer Polytechnic Institute, Troy, NY, Tech. Rep. ECSE-NET-2001-2, 2001.
[32] B. C. Sowden and K. W. Sowerby, "The impact of long-range dependent traffic in a CDMA system supporting real-time services," in *Proc. IEEE Global Telecommunications Conf. (IEEE GLOBECOM)*, Nov. 25–29, 2001, vol. 6, pp. 3509–3513.
[33] A. Veres and M. Boda, "The chaotic nature of TCP congestion control," in *Proc. IEEE INFOCOM*, Tel-Aviv, Israel, 2000, pp. 1715–1723.
[34] A. Veres, Z. Kenesi, S. Molnar, and G. Vattay, "On the propagation of long-range dependence in the internet," in *Proc. AGM SIGCOMM*, Stockholm, Sweden, Sep. 2000, pp. 243–254.
[35] L. Wang and W. Zhuang, "Call admission control for self-similar data traffic in cellular communications," in *IEEE Global Telecommunications Conf. (IEEE GLOBECOM)*, Dec. 1–5, 2003, vol. 2, pp. 982–986.
[36] W. Willinger, V. Paxson, and M. S. Taqqu, "Self-similarity and heavy tails: structurel modeling of network traffic," in *A Practical Guide to Heavy Tails: Staticstical Techniques and Applications*, R. J. Adler, R. E. Feldman, and M. S. Taqqu, Eds. Cambridge, MA: Birkhäuser, 1998, pp. 27–53.
[37] W. Willinger, M. S. Taqqu, R. Sherman, and D. V. Wilson, "Self-similarity through high-variability: Statistical analysis of Ethernet LAN traffic at the source level," *IEEE/ACM Trans. Netw.*, vol. 5, pp. 71–86, Feb. 1997.
[38] X. Yang and A. P. Petropulu, "The extended alternating fractal renewal process for modeling traffic in high-speed communication networks," *IEEE Trans. Signal Process.*, vol. 49, no. 7, pp. 1349–1363, Jul. 2001.
[39] K. G. Yoon and P. S. Min, "On the prediction of average queueing delay with self-similar traffic," in *Proc. IEEE Global Telecommunications Conf. (IEEE GLOBECOM)*, Dec. 2003, vol. 22, no. 1, pp. 2987–2991.
[40] J. Yu, A. P. Petropulu, and H. Sethu, "Rate-limited EAFRP: a new improved model for high-speed network traffic," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, Hong Kong, Apr. 2003, pp. VI-25–28.
[41] ——, "Rate-limited EAFRP: A new improved model for high-speed network traffic," *IEEE Trans. Signal Process.*, vol. 53, no. 2, pp. 505–522, Feb. 2005.
[42] J. Zhang, M. Hu, and N. B. Shroff, "Bursty data over CDMA: MAI self similarity, rate control and admission control," in *Proc. 21st IEEE Annu. Joint Conf. IEEE Computer Communications Societies (IEEE INFOCOM)*, 2002, vol. 1, pp. 391–399.

Jie Yu (S'99–A'02–M'04) received the B.Eng. and M.Sc. degrees, both in electronics engineering, from Tsinghua University, Beijing, China, in 1999 and 2001, respectively, and the Ph.D. degree in electrical and computer engineering from Drexel University, Philadelphia, PA, in 2005.

In July 2005, he joined US&S, Inc., Pittsburgh, PA, where he is now a System Engineer. His research interests are in the area of blind channel estimation and equalization, modeling and analysis of high-speed network traffic.

Athina P. Petropulu (S'87–M'87–SM'01) received the Diploma degree in electrical engineering from the National Technical University of Athens, Greece, in 1986, the M.Sc. and Ph.D. degrees in electrical and computer engineering, both from Northeastern University, Boston, MA, in 1988 and 1991, respectively.

In 1992, she joined the Department of Electrical and Computer Engineering at Drexel University, Philadelphia, PA, where she is now a Professor. During the academic year 1999–2000, she was an Associate Professor at Université Paris Sud, École Supérieure d'Electrícité, France. During the academic year 2006–2007, she was a visiting Fellow with the Department of Electrical Engineering at Princeton University, Princeton, NJ. Her research interests span the area of statistical signal processing, wireless communications, and networking and ultrasound imaging. She is the coauthor (with C. L. Nikias) of the textbook entitled *Higher-Order Spectra Analysis: A Nonlinear Signal Processing Framework* (Englewood Cliffs, NJ: Prentice-Hall, 1993).

Dr. Petropulu is the recipient of the 1995 Presidential Faculty Fellow Award in Electrical Engineering given by the National Science Foundation and the White House. She has served as an Associate Editor for the IEEE TRANSACTIONS ON SIGNAL PROCESSING and the IEEE SIGNAL PROCESSING LETTERS, and is a member of the editorial board of the *IEEE Signal Processing Magazine* and the EURASIP *Journal on Wireless Communications and Networking*. She is IEEE Signal Processing Society (SPS) Vice President—Conferences and member of the IEEE SPS Board of Governors and the SPS Executive Committee. She was the General Chair of the 2005 International Conference on Acoustics Speech and Signal Processing (ICASSP), Philadelphia, PA.