Predicting Properties of Congestion Events for a Queueing System With fBm Traffic

Yasong Jin, Soshant Bali, Tyrone E. Duncan, *Fellow, IEEE*, and Victor S. Frost, *Fellow, IEEE*

***Abstract*—In packet networks, congestion events tend to persist, producing large delays and long bursts of consecutive packet loss resulting in perceived performance degradations. The length and rate of these events have a significant effect on network quality of service (QoS). The packet delay resulting from these congestion events also influences QoS. In this paper a technique for predicting these properties of congestion events in the presence of fractional Brownian motion (fBm) traffic is developed.**

***Index Terms*—Internet, network congestion, networks, quality of service.**

I. Introduction

CONGESTION events in communication networks cause packet losses, and it is well known that these losses occur in bursts [1], [2]. Furthermore the frequency and the duration of these congestion events significantly influence the perceived network performance [3], [4]. The Internet Engineering Task Force has defined measurement-based QoS metrics [5] aimed at characterizing packet loss patterns. Measured packet traces [1], [6] have been used to create models for the temporal dependence of packet loss. These models assume a specific packet loss process, e.g., one that transits between different states, such as a no-loss state and a loss state. However, transforming network traffic parameters directly into predictions of the properties of congestion events will aid network design and provide a useful indication of QoS. The properties to be considered here include the rate, the duration, and the magnitude of the delay induced by congestion events. An approach for determining the rate of congestion events for some standard traffic models was presented in [7]. In this paper the approach is extended in two directions: 1) to a fluid queueing model with a self-similar input and 2) to include additional properties of congestion events.

In the early 1990s, researchers with Bellcore observed the phenomena of self-similarity and long-range dependence in LAN traffic [8], which roughly means that the traffic "looks" similar under different time scales and the correlation between packets decays very slowly. This observation is inconsistent with the short dependence assumption in traditional traffic models, such as the Poisson process and other Markov models. Subsequent studies [9], [10] showed that the traditional models seem inadequate for data networks. Since then, many other traffic models have been proposed, such as fractal point processes [11] and multifractal models [12]. In 1994, Norros [13] proposed a fluid queueing model with a fractional Brownian motion (fBm) as input. A fluid model whose input is not packetized is suitable for modeling high speed networks. For example, Hohn *et al.* [14] used a fluid model to analyze high-precision router measurement. A fBm for suitable values of the Hurst parameter process has the properties of self-similarity and long-range dependence. By analyzing the origin of self-similarity and long-range dependence in network traffic, it was shown in [15] that the superposition of a family of homogeneous ON/OFF traffic sources with heavy tailed ON and OFF periods, with proper scaling, converges in distribution to a fBm plus a linear component. The superposition of traffic sources is well-suited to the network core, which has thousands of simultaneous traffic flows. It has been observed that long-range dependence is a property of the backbone traffic [16]. Recent network measurements [17] also justify the applicability of a fBm, which is a Gaussian process, as a traffic model for aggregated network traffic. Thus we focus on the Norros model to study the characteristics of congestion events.

The primary contribution of this paper is the development of methodologies for predicting the expected values for the properties of congestion events in a queue with a fBm input. The structure of this paper is as follows: In Section II, a congestion event is defined, some preliminaries on fBm, the Norros model, conditioned fractional Brownian motion, and the Poisson clumping approximation are given. In Section III, an approximation for a congestion event is proposed to simplify the analysis. The properties of congestion events and approximation methods are discussed in Sections IV and V. Comparisons between the predictions made by the proposed methodologies and simulations are presented in Section VI. Finally, some conclusions are drawn in Section VII.

II. Preliminaries

In this section, a congestion event is defined and some preliminaries on the Norros model, a fBm and a conditioned fBm are given for the future analysis.

Manuscript received May 12, 2005; revised April 14, 2006; approved by IEEE/ACM TRANSACTIONS ON NETWORKING Editor M. Roughan. This work was supported in part by the National Science Foundation under Grant ANI-0125410. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the author and do not necessarily reflect the views of the National Science Foundation.

Y. Jin and T. E. Duncan are with the Mathematics Department, University of Kansas, Lawrence, KS 66045 USA (e-mail: jinyasong@math.ku.edu; duncan@math.ku.edu).

S. Bali is with the Electrical Engineering and Computer Science Department, University of Kansas, Lawrence, KS 66046 USA (e-mail: sbali@ittc.ku.edu).

V. S. Frost is with the Information and Telecommunication Technology Center, University of Kansas, Lawrence, KS 66045 USA (e-mail: frost@eecs.ku.edu).

Digital Object Identifier 10.1109/TNET.2007.896538

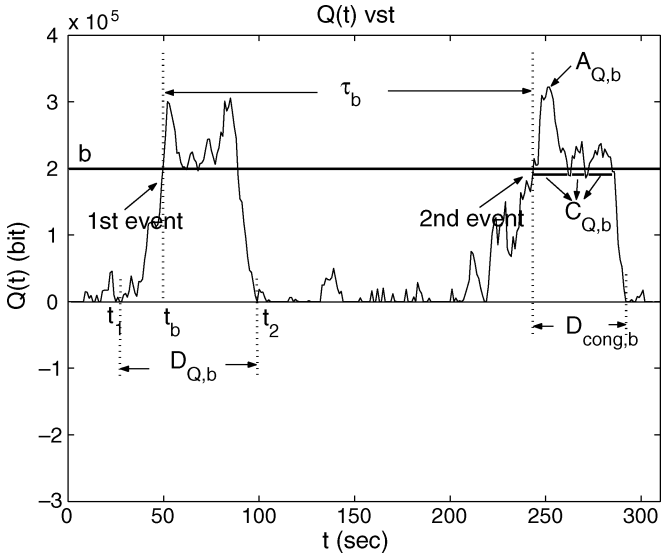


Fig. 1. Example of workload process and definitions of random variables of interest.

A. Congestion Events

Let $\{Q(t), t \in \mathbb{R}\}$ be a queue length process. A busy period from t_1 to t_2 is a period such that $Q(t_1) = Q(t_2) = 0$ but $Q(t) > 0$ for $\forall t \in (t_1, t_2)$. In a busy period from t_1 to t_2, a congestion event with a level b is defined to occur at time t_b if t_b is the first time that the process $Q(t)$ reaches a fixed level b. The congestion event ends at time t_2, i.e. the first time the queue becomes empty after t_b. Two congestion events are shown in Fig. 1. Given this definition of congestion the process $Q(t)$ can reenter the level b multiple times during one congestion event. The premise of this work is that, for a large b, a congestion event as defined here results in a burst of packet losses.

Formally as in [18], let $(\Omega, \mathcal{F}, P)$ be a probability space and θ_t be a measurable flow on $(\Omega, \mathcal{F})$ which is invariant under P. Let $t_b^{(i)}$ denote the beginning of the ith congestion event, such that, $-\infty \leq \ldots < t_b^{(-1)} < t_b^{(0)} \leq 0 < t_b^{(1)} < t_b^{(2)} < \ldots \leq \infty$. Let $t_1^{(i)}$ and $t_2^{(i)}$ be the corresponding beginning and end of the busy period in which the ith congestion event occurs. Let $N = \{t_b^{(i)}, i \in \mathbb{Z}\}$ denote the set of the beginning times of congestion events, then $\{N, \theta_t, P\}$ forms a stationary marked point process, in which the paths of congestion events are viewed as marks. Let P_N^0 be the associated Palm probability defined as

$$P_N^0(A) = \frac{1}{E[N(C)]} E\left[\int_C (1_A \circ \theta_s) N(ds)\right]$$

where $A \in \mathcal{F}$, $N(C)$ denotes the number of points in a Borel set C and 1_A is an indicator function. We use E to represent the expectation with respect to P, and use E^0 to represent the expectation with respect to P_N^0.

The inter-congestion event time between the i^{th} and the $(i+1)^{\text{th}}$ congestion events is denoted by $\tau_b^{(i+1)} = t_b^{(i+1)} - t_b^{(i)}$. We are interested in the properties of an arbitrary congestion event. To simplify the notation, we omit the superscripts. Then the mean inter-congestion event time is $E^0[\tau_b]$. As shown in [7], $E^0[\tau_b]$ (or the rate $1/E^0[\tau_b]$) is a useful QoS metric. The other metrics of an arbitrary congestion event are $E^0[C_{Q,b}]$, the mean sojourn time that $Q(t)$ spends above threshold b in a congestion event; $E^0[D_{cong,b}]$, the mean duration of a congestion event, i.e., the time from t_b to t_2; $E^0[D_{Q,b}]$, the mean duration of a busy period containing a congestion event, i.e., the time from t_1 to t_2; and $E^0[A_{Q,b}]$ which is the mean peak queue length of a congestion event. In a study of high precision router measurements Hohn *et al.* [14] demonstrated that the $(D_{Q,b}, A_{Q,b})$ pairs can be used to describe a busy period in which the queue length exceeds a congestion threshold b. The set of metrics, $E^0[\tau_b]$, $E^0[C_{Q,b}]$, $E^0[D_{cong,b}]$, $E^0[D_{Q,b}]$, $E^0[A_{Q,b}]$ can be used to characterize the nature of congestion events.

Fig. 2. A queue with fractional Brownian input, $A(t) = mt + \sqrt{a}B^H(t)$.

B. A Queueing Model With Fractional Brownian Traffic

As in [13], we use a fBm, which is a self-similar Gaussian process with stationary increments, to model network traffic. The definition of a fBm is given as follows.

Definition 1: [19] A standard fractional Brownian motion (fBm) with Hurst parameter $H \in (0,1)$, $\{B^H(t), t \in \mathbb{R}\}$, is a real-valued Gaussian process such that for $s, t \in \mathbb{R}$, $E[B^H(t)] = 0$ and $E[B^H(s)B^H(t)] = (1/2)[|s|^{2H} + |t|^{2H} - |s-t|^{2H}]$.

In this paper, it is assumed that $H \in [1/2, 1)$. For $H > 1/2$, $\{B^H(t), t \in \mathbb{R}\}$ has the property of long range dependence, that is, if

$$r(n) = E\left[B^H(1)\left(B^H(n+1) - B^H(n)\right)\right]$$

for $n = 1, 2, \ldots$, then $\sum_{n=1}^{\infty} r(n) = \infty$. A fluid queue with a fBm as input was proposed by Norros [13], Fig. 2. A fBm, $B^H(t)$, is used to capture the self-similarity and the long-range dependence in the input network traffic. Let $A(t) = mt + \sqrt{a}B^H(t)$ be the cumulated arrivals up to time t, where m is the mean input rate (bps), a stands for the variance (bit^2), and $B^H(t)$ is a standard fBm with parameter H. For an input traffic modeled by $A(t)$, we say that the input is determined by (m, a, H). At time t, the queue length $Q_o(t)$ can be expressed as, see [13] and the references therein, $Q_o(t) = A(t) - \mu t - \inf_{s \leq t}(A(s) - \mu s)$, where μ is a fixed service rate in (bps). Now $Q_o(t)$ can be written as

$$Q_o(t) = \sqrt{a}B^H(t) - (\mu - m)t - \inf_{s \leq t}\left(\sqrt{a}B^H(s) - (\mu - m)s\right) \tag{1}$$

in which $\mu - m$ is the surplus rate. For the stability of the queue, it is assumed that $\mu - m > 0$.

Consider a scaled $Q_o(t)$, which is defined as $Q(t) = Q_o(t)/\sqrt{a}$. We can observe that the temporal properties of the congestion events of $Q_o(t)$ with a level b_o are the same as those

of the congestion events of $Q(t)$ with a level $b = b_o/\sqrt{a}$. Therefore to study the properties of congestion events of a queue with an input (m, a, H) and a service rate μ, it is equivalent to study the corresponding scaled queue length process $Q(t)$

$$Q(t) = B^H(t) - ct - \inf_{s \le t} \left(B^H(s) - cs\right) \tag{2}$$

in which $c = (\mu - m)/\sqrt{a}$ stands for the scaled surplus rate.

C. Conditioned fBm

We will use a conditioned fBm with a negative drift to study congestion events, here a conditioned fBm is defined and some properties are discussed.

Definition 2.2: A conditioned fBm $\{\tilde{B}^H(t; r, d), t \ge 0\}$ with parameters H, r and d is defined as $B^H(t+r) - B^H(r)$, given that $B^H(0) = 0$ and $B^H(r) = d$, in which $r > 0$ and $d \in \mathbb{R}$ are fixed constants.

Proposition 2.3: Let $r > 0$, $d \in \mathbb{R}$ and $H \in [1/2, 1)$. A conditioned fBm $\{\tilde{B}^H(t; r, d), t \ge 0\}$ is a Gaussian process. For $t, s \ge 0$, let $\mu_{r,d}(t) = E[\tilde{B}^H(t; r, d)]$, $\sigma_r(s,t) = cov(\tilde{B}^H(s; r, d), \tilde{B}^H(t; r, d))$, and let $C_H(r,t) = (1/2)[(t+r)^{2H} - r^{2H} - t^{2H}]$, $D_H(s,t) = (1/2)[s^{2H} + t^{2H} - |t-s|^{2H}]$, then

$$\mu_{r,d}(t) = \frac{C_H(r,t)}{r^{2H}} d \tag{3}$$

$$\sigma_r(s,t) = D_H(s,t) - \frac{C_H(r,t) C_H(r,s)}{r^{2H}}. \tag{4}$$

Proof: The covariance matrix of $B^H(r)$, $B^H(s+r) - B^H(r)$, $B^H(t+r) - B^H(r)$, for $t, s \ge 0$, is

$$\begin{bmatrix} r^{2H} & C_H(r,s) & C_H(r,t) \\ C_H(r,s) & s^{2H} & D_H(s,t) \\ C_H(r,t) & D_H(s,t) & t^{2H} \end{bmatrix}.$$

Conditioned on $B^H(r) = d$, and using the properties of multinormal random variables [20], $E[\tilde{B}^H(t; r, d)] = E[B^H(t+r) - B^H(r) | B^H(r) = d]$ and $cov(\tilde{B}^H(s; r, d), \tilde{B}^H(t; r, d))$ are as given in (3) and (4), respectively. Thus, for fixed t, $\tilde{B}^H(t; r, d)$ is a normal random variable with mean $\mu_{r,d}(t)$, variance $\sigma_r^2(t) = \sigma_r(t,t)$. ∎

Remark 2.4: For $H = 1/2$, a conditioned fBm $\tilde{B}^H(t; r, d)$ reduces to a standard Brownian motion.

D. Poisson Clumping Approximation

Following [7], the Poisson clumping approximation [21] is used to find the inter-congestion event time. For a threshold b, the queue tail distribution, $P(Q(0) \ge b)$, and the mean sojourn time of $Q(t)$ above the threshold in a congestion event, $E^0[C_{Q,b}]$, are applied to evaluate the mean inter-congestion event time as

$$E^0[\tau_b] \approx \frac{E^0[C_{Q,b}]}{P\left(Q(0) \ge b\right)}. \tag{5}$$

Note that for a fBm traffic $P(Q(0) \ge b)$ can be approximated using the results in [22]. Thus the problem reduces to finding $E^0[C_{Q,b}]$. By applying the Poisson clumping approximation, we assume that the congestion events are rare and the dependence among the events are small. These assumptions are reasonable for the case studied here. When b is large, the congestion events are rare and far apart. Although $B^H(t)$ has long range dependence, the dependence among congestion events are small. We validated the Poisson clumping approximation with simulations, some of which are shown in Fig. 7(a)–(c). These results indicate that the Poisson clumping approximation can be used to evaluate the average inter-congestion event time.

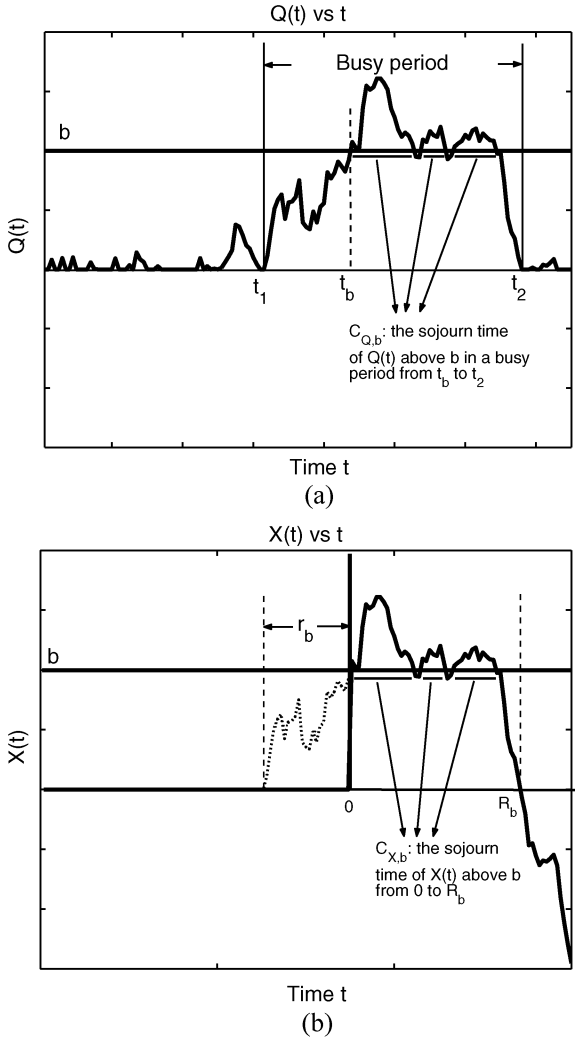


Fig. 3. A busy period of $Q(t)$ from t_1 to t_2, and the approximation process $X(t)$. (a) A busy period of $Q(t)$. (b) Approximation process $X(t)$.

III. Busy Periods Containing Congestion Events

The busy periods of a queue with a fBm input have been discussed in [23], and recently in [24], in which the busy periods are defined as the periods that the queue is not empty. In this paper we are interested in the periods in which congestion events occur. Note that a busy period hereafter always means a busy period containing a congestion event. A busy period from t_1 to t_2 is shown in Fig. 3(a), where t_b is the first time that the queue reaches a level b in the busy period, and t_2 is the first time that the queue returns to 0 after t_b. The time t_b separates one busy period

into two parts, $[t_1, t_b]$ and $[t_b, t_2]$. In order to apply the Poisson clumping approximation, it is necessary to find $E^0[C_{Q,b}]$, which is the mean time that the queue spends above the level b in a congestion event, Fig. 3(a). It will be demonstrated next that the problem can be simplified by approximating $Q(t)$ in $[t_b, t_2]$ with a process $X(t)$, which is a conditioned fBm with a negative drift [Fig. 3(b)].

Proposition 3.1: Let t_1 and t_2 be the end points of a busy period. Let $t_b \in [t_1, t_2]$ be the first time that $Q(t)$ reaches a level b. Then for $t \in [t_1, t_2]$, $Q(t)$ can be rewritten as

$$Q(t) = B^H(t) - B^H(t_1) - c(t - t_1), \qquad t \in [t_1, t_b] \tag{6}$$

$$Q(t) = b + B^H(t) - B^H(t_b) - c(t - t_b), \qquad t \in [t_b, t_2]. \tag{7}$$

Proof: From the conditions, we have that $Q(t_1) = 0$, $Q(t_b) = b$ and $Q(s) > 0$ for $s \in (t_1, t_2)$. From (2), it can be verified that for $\forall t \in (t_1, t_2)$,

$$B^H(t_1) - ct_1 = \inf_{s \le t} \left(B^H(s) - cs \right).$$

Then based on (2), for $t \in [t_1, t_b]$,

$$\begin{aligned} Q(t) &= B^H(t) - ct - \inf_{s \le t} \left(B^H(s) - cs \right) \\ &= B^H(t) - B^H(t_1) - c(t - t_1). \end{aligned}$$

Similarly, $Q(t) = b + [B^H(t) - B^H(t_b)] - c(t - t_b)$, for $t \in [t_b, t_2]$. ∎

Remark 3.2: In $[t_1, t_b]$, $Q(t)$ increases from 0 to the level b. Since $Q(t_b) = b$, from (6), the increment of the fBm in $[t_1, t_b]$ is $B^H(t_b) - B^H(t_1) = b + c(t_b - t_1)$. For the period of $[t_b, t_2]$, recall that if t_b is a constant, $\{B^H(t) - B^H(t_b), t \in [t_b, \infty)\}$ is equivalent to $\{B^H(t), t \in [0, \infty)\}$ in distribution. This is the motivation for approximating the period $[t_b, t_2]$ of $Q(t)$ with a conditioned fBm with a negative drift.

Define a process $\{X(t; r_b, d_b) = b + \tilde{B}^H(t; r_b, d_b) - ct, t \in [0, \infty)\}$, where $\{\tilde{B}^H(t; r_b, d_b), t \ge 0\}$ is a conditioned fBm with parameters H, r_b, d_b, and

$$r_b = \frac{bH}{c(1 - H)} \tag{8}$$

$$d_b = b + cr_b. \tag{9}$$

For a large b, the congestion events are rare. Since "rare events occur in the most likely way" and the most probable sample path of $Q(t)$ found in [23] spends time $bH/c(1-H)$ increasing from 0 to a large fixed level b, we use the constants r_b and d_b to represent the time $t_b - t_1$ and the increment of the fBm in $[t_1, t_b]$, respectively. At $t = 0$, the process $X(t; r_b, d_b)$ starts at b, i.e., $X(0; r_b, d_b) = b$. Let R_b be the first time that $X(t; r_b, d_b)$ returns to 0, that is, $R_b = \inf\{t \ge 0 : X(t; r_b, d_b) \le 0\}$. To simplify the exposition, we sometimes denote $X(t; r_b, d_b)$ with $X(t)$.

The part $[t_b, t_2]$ of a busy period of $Q(t)$ is approximated by $[0, R_b]$ of $X(t)$, Fig. 3. Let $C_{X,b}$ denote the sojourn time that $X(t)$ spends above the level b in the period of $[0, R_b]$. The idea is to approximate $E^0[C_{Q,b}]$ with $E[C_{X,b}]$, that is,

$$E^0[C_{Q,b}] \approx E[C_{X,b}]. \tag{10}$$

Since the process $X(t)$ is not related to the point process N, defined in Section II-A, we use $E[C_{X,b}]$, which is the expectation with respect to P, to represent the expectation of $C_{X,b}$.

Remark 3.3: We use $[0, R_b]$ of $X(t)$ to approximate the part $[t_b, t_2]$ of a busy period. This approximation has some inherent shortcomings. The parameters r_b and d_b of $X(t)$ are used to represent $t_b - t_1$ and the corresponding increment of the fBm, respectively. However, they cannot capture the property that $Q(t)$ is less than b and strictly positive in (t_1, t_b), i.e., $0 < Q(t) < b, \forall t \in (t_1, t_b)$. And for a fixed b, r_b is a constant, but $t_b - t_1$ is obviously a random variable. Thus $\{Q(t), t \in (t_b, t_2)\}$ is not equivalent to a conditioned fBm. As an approximation, $X(t)$ cannot exactly capture all the characteristics of a congestion event. However its use simplifies the analysis and produces useful results.

IV. Mean Sojourn Time

The sojourn time $C_{X,b}$, that is, the time that $X(t)$ spends above b in $[0, R_b]$, can be written as

$$C_{X,b} = \int_0^{R_b} 1_{[b,\infty)} \left(X(t; r_b, d_b) \right) dt.$$

Let $\bar{\Phi}(\cdot) = (1/\sqrt{2\pi}) \int_{\cdot}^{\infty} e^{-\xi^2/2} d\xi$ denote the complement of a standard normal distribution. Let

$$U_{C_b} = \int_0^{\infty} \bar{\Phi} \left(\frac{ct - \mu_{r_b, d_b}(t)}{\sqrt{\sigma^2_{r_b}(t)}} \right) dt. \tag{11}$$

Then U_{C_b} is an upper bound of $E[C_{X,b}]$, since

$$\begin{aligned} E[C_{X,b}] &\le E \int_0^{\infty} 1_{[b,\infty)} \left(b + \tilde{B}^H(t; r_b, d_b) - ct \right) dt \\ &= U_{C_b}. \end{aligned}$$

For $H = 1/2$, the process $X(t)$ reduces to a standard Brownian motion with a negative drift. By the dominated convergence theorem, it can be verified that $\lim_{b \to \infty} E[C_{X,b}] = U_{C_b}$ (for $H = 1/2$, U_{C_b} is a constant which is independent of b). Then for a large b, $E[C_{X,b}] \approx U_{C_b}$. Although the limiting result can only be shown for $H = 1/2$, the simulation results in Section VI demonstrate that U_{C_b} is a good approximation for $E[C_{X,b}]$ ($\approx E^0[C_{Q,b}]$) when $H \neq 1/2$. Thus we approximate $E[C_{X,b}]$ with U_{C_b}, i.e.,

$$E[C_{X,b}] \approx U_{C_b}. \tag{12}$$

Combining (5), (10) and (12), the mean inter-congestion event time, $E^0[\tau_b]$, can be expressed as

$$E^0[\tau_b] \approx \frac{E[C_{X,b}]}{P(Q(0) \ge b)} \approx \frac{U_{C_b}}{P(Q(0) \ge b)}. \tag{13}$$

Even though several approximations were applied to obtain (13), the above analysis successfully predicts trends observed from simulations. The method provides better predictions for the inter-congestion event time than directly using the

reciprocal of the tail probability, $1/P(Q(0) \geq b)$, as will be discussed in Section VI.

In the following we obtain an upper bound for $E[R_b]$, which will be used to approximate the mean duration time of a congestion event, $E^0[D_{cong,b}]$. Let

$$U_{R_b} = \int_0^\infty \bar{\Phi}\left(\frac{cu - b - \mu_{r_b,d_b}(u)}{\sqrt{\sigma^2_{r_b}(u)}}\right) du. \tag{14}$$

Since $E[R_b] = \int_0^\infty P(R_b > u)du$ and $\{R_b > u\} = \{\inf_{0\leq s\leq u}(b + \tilde{B}^H(s; r_b, d_b) - cs) > 0\}$, we have

$$\begin{aligned} E[R_b] &= \int_0^\infty P\left(R_b > u\right) du \\ &\leq \int_0^\infty P\left(b + \tilde{B}^H(u; r_b, d_b) - cu > 0\right) \\ &= U_{R_b}. \end{aligned}$$

Remark 4.1: By deriving a lower bound for $E[R_b]$, it can be illustrated that the relative difference between U_{R_b} and $E[R_b]$, defined as $(U_{R_b} - E[R_b])/(E[R_b])$, approaches 0 as the level b increases. Therefore for a large b, U_{R_b} can be used as an approximation for $E[R_b]$, i.e.,

$$E[R_b] \approx U_{R_b}. \tag{15}$$

The lower bound for $E[R_b]$ is not presented here, but as can be seen from the simulation results, U_{R_b} provides a good approximation for $E[R_b]$ ($\approx E^0[D_{cong,b}]$).

V. Mean Duration Time and Mean Amplitude

A. Mean Duration of Congestion Events

As shown in Fig. 1, a congestion event starts at time t_b and ends at t_2. Let $D_{cong,b} = t_2 - t_b$ denote the duration time of a congestion event. Since the period $[t_b, t_2]$ of $Q(t)$ is approximated by $[0, R_b]$ of $X(t)$ and from (15), $E^0[D_{cong,b}]$ can be expressed as

$$E^0[D_{cong,b}] = E^0[t_2 - t_b] \approx E[R_b] \approx U_{R_b}. \tag{16}$$

B. Mean Duration of Busy Periods

Let $D_{Q,b}$ denote the duration of a busy period in which a congestion event occurs. The mean duration is $E^0[D_{Q,b}] = E^0[t_2 - t_1]$. From Fig. 1, $E^0[D_{Q,b}]$ can be written as $E^0[D_{Q,b}] = E^0[D_{cong,b}] + E^0[t_b - t_1]$. Recall that $t_b - t_1$ is approximated with a constant r_b, which can be evaluated with (8). Thus, combining (16), we have

$$E^0[D_{Q,b}] \approx E[R_b] + r_b \approx U_{R_b} + r_b. \tag{17}$$

C. Mean Amplitude

The busy periods in a network router have been previously modeled by triangles in [14], so we use a triangle to approximate a busy period in Fig. 4. The triangle has a base of B_b, crosses the level b at s_1 and s_2. The mean amplitude of a busy period, $E^0[A_{Q,b}]$, can be approximated with the height of the triangle A_b. Let L_b denote $s_2 - s_1$. Note that L_b is the length that the triangle stays above the level b, we use $E^0[C_{Q,b}]$, the mean sojourn time of a congestion event, to approximate L_b, that is, $L_b \approx E^0[C_{Q,b}]$. The base B_b is approximated with the mean duration time of a busy period $E^0[D_{Q,b}]$. With simple geometry, it can be derived that $A_b = b(B_b/(B_b - L_b)) \approx b((E^0[D_{Q,b}])/(E^0[D_{Q,b}] - E^0[C_{Q,b}]))$. Combining (12) and (17), we have

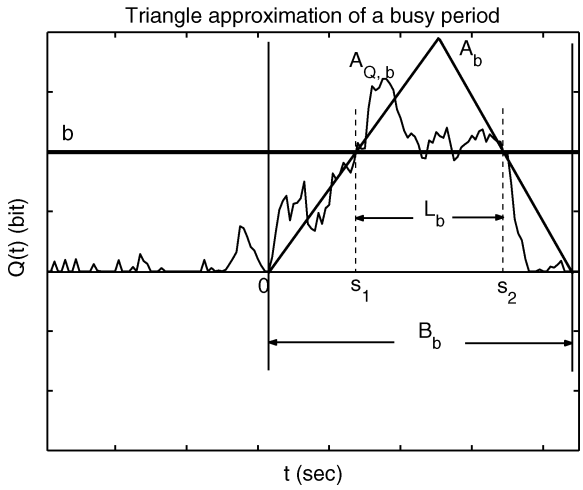


Fig. 4. Triangle approximation of a busy period.

$$E^0[A_{Q,b}] \approx A_b \approx b\frac{U_{R_b} + r_b}{U_{R_b} + r_b - U_{C_b}}. \tag{18}$$

VI. Evaluation

So far we have focused on a scaled queue length process $Q(t)$, given by (2). Let $Q_o(t)$ denote the queue length which has an input (m, a, H) and a service rate μ, the load of the queue is $\rho = m/\mu$. Recall that the properties of a congestion event of $Q_o(t)$ with a level b_o are equivalent to the congestion event of $Q(t)$ with a level $b_o/\sqrt{a}$. On the basis of (2), given a scaled service rate c and the input parameters (m, a, H), we can conveniently transform the characteristics of congestion events of $Q(t)$ to those of $Q_o(t)$.

Now we can evaluate the temporal characteristics of congestion events and the corresponding busy periods. Evaluations based on the above analysis are compared with simulation results. Fractional Brownian motions are generated with the algorithm proposed in [25]. For $H \in [0.5, 0.79]$, 20 traces of fBm are generated, each trace has 2^{24} samples; for $H = 0.85$, 80 traces are generated, each has 2^{22} samples.[1] The parameters H and c are varied to modify the long-range intensity and the scaled surplus rate. The relative error of the approximations is reported, which is defined as $\|x - \hat{x}\|/\|x\|$, where x is the simulation result, $\hat{x}$ is the corresponding approximation and $\|\cdot\|$ is the Euclidean norm.

[1]The simulations were performed on a computer with two Intel Xeon Processors running at 2.8 GHz with 2 GB RAM. The memory capacity combined with the numerical limitation of the algorithm in [25] limited the sample size to 2^{24} for $H \in [0.5, 0.79]$, and 2^{22} for $H = 0.85$.

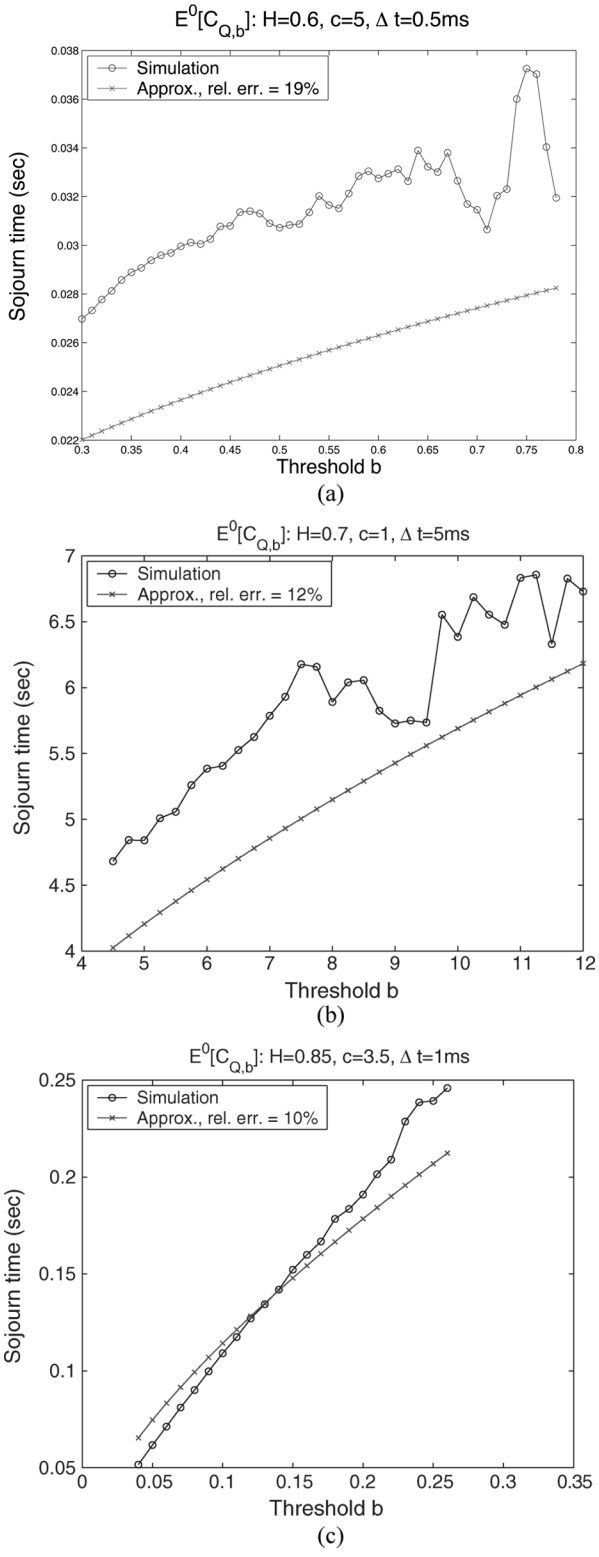


Fig. 5. Comparison of mean sojourn time versus b. (a) $H = 0.6$. (b) $H = 0.7$. (c) $H = 0.85$.

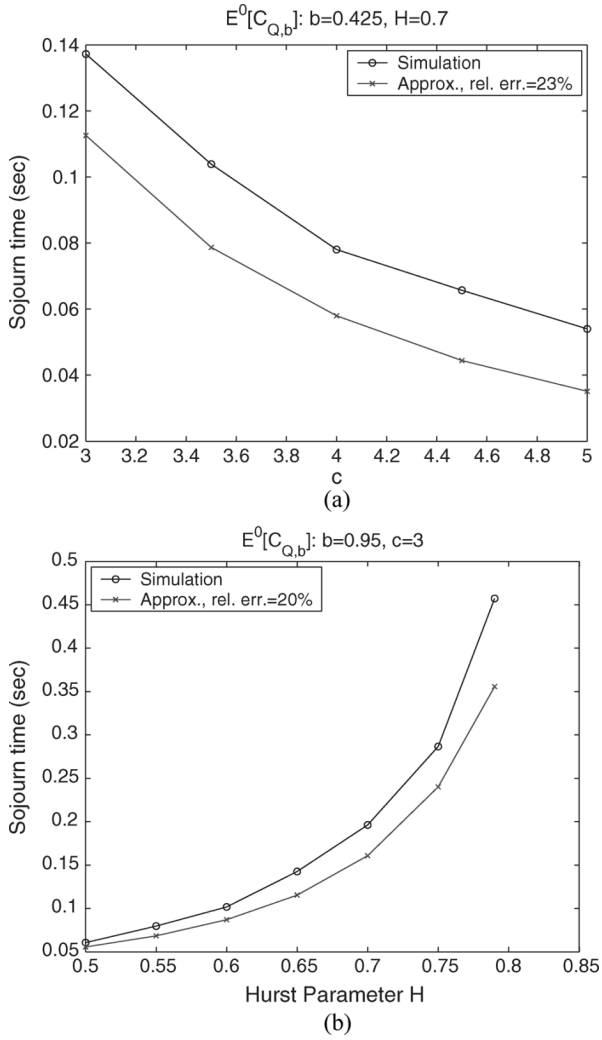


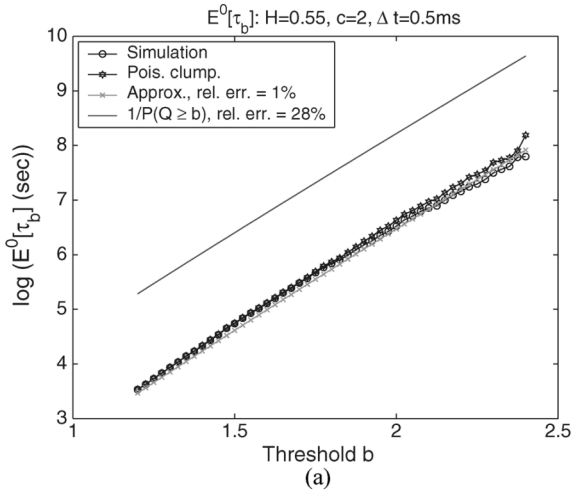
Fig. 6. Comparison of mean sojourn time. (a) Sojourn time versus c. (b) Sojourn time versus H.

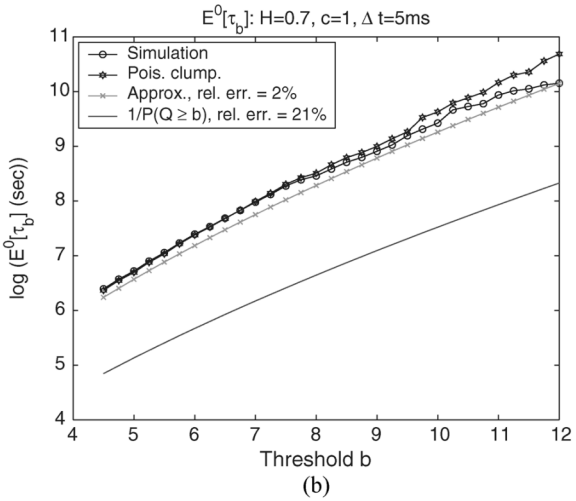
In simulations, we use different values for Δt, the time between consecutive samples. On one hand, we need to let Δt be small so that we can measure the sojourn and duration times accurately; on the other hand, to collect enough congestion events, we want the whole trace (2^{24} or 2^{22} samples) to represent a time series which is in the order of hours.

For a fixed simulation length when the threshold b increases, fewer and fewer congestion events occur (the events become rare). For example, under the conditions $H = 0.85$, $c = 3.5$, for $b = 0.05$, there are over 50,000 congestion events, but for $b = 0.25$, we can only collect 600 events over 80 traces. Consequently, fluctuations can be noticed for large b in the simulation results, see Figs. 5(a)–(c) and 9(a).

A. Mean Sojourn Time $E^0[C_{Q,b}]$

The comparisons between the predicted and simulated $E^0[C_{Q,b}]$ are shown in Figs. 5 and 6. The approximation results follow the trends as a function of the surplus rate c and the Hurst parameter H, the relative errors range from 10% to 20%. The errors are partly caused by r_b. It is observed that r_b overestimates $E^0[t_b - t_1]$, i.e., the time that the queue builds up from 0 to b in a busy period. Fluctuations, which are caused by small sample sizes, can be observed for large b [Fig. 5(a)–(c)]. Similar phenomena have been observed in network router performance measurements, e.g., Fig. 13 in [14].





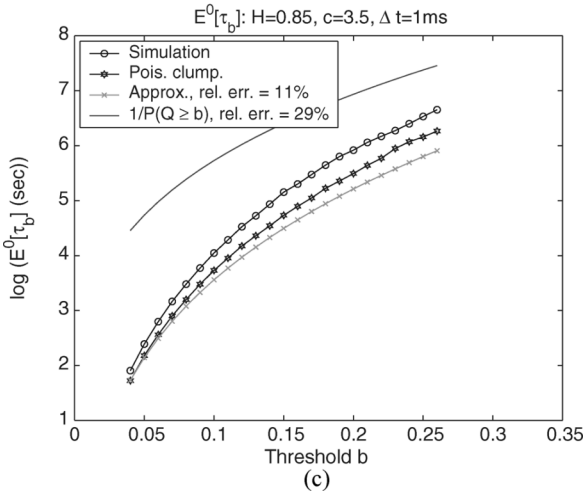
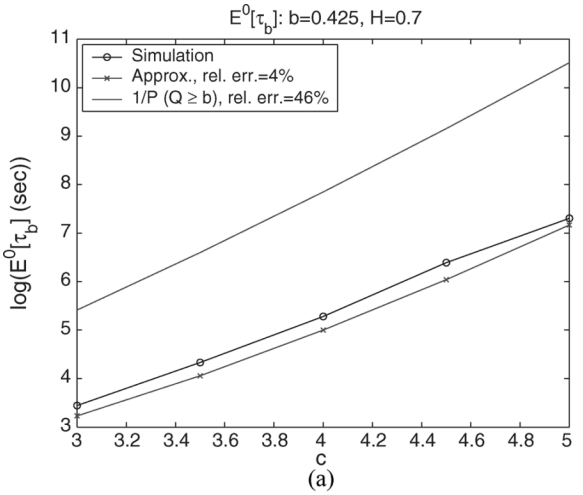
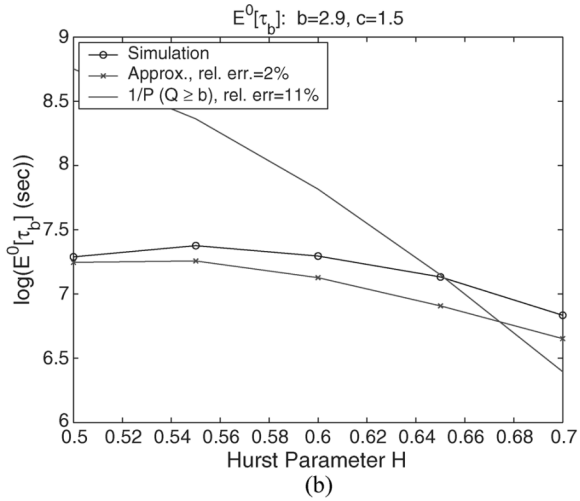


Fig. 7. Comparison of mean inter-congestion time versus b. (a) $H = 0.55$. (b) $H = 0.7$. (c) $H = 0.85$.





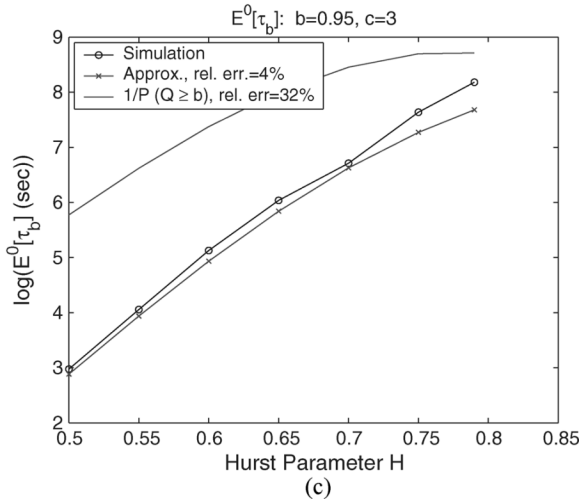


Fig. 8. Comparison of mean inter-congestion time. (a) Inter-congestion time versus c. (b) Inter-congestion time versus H, $b = 2.9$. (c) Inter-congestion time versus H, $b = 0.95$.

B. Inter-Congestion Event Time $E^0[\tau_b]$

The approximation given in (13) is compared with the simulation results and another approximation method, $1/P(Q(0) \geq b)$, the reciprocal of the tail of the queue fill probability. As shown in Figs. 7 and 8, the approximation (13) outperforms $1/P(Q(0) \geq b)$ in most cases. We notice that for different parameter sets, $E^0[\tau_b]$ may increase or decrease with respect to H. For example, when $b = 2.9$, $c = 1.5$, $E^0[\tau_b]$ decreases versus H as shown in Fig. 8(b); but for $b = 0.95$, $c = 3$, $E^0[\tau_b]$ increases in Fig. 8(c). In both cases, our approximation results can follow the observed trends. In Fig. 7(a)–(c), the Poisson clumping approximation, given in (5), is validated; both

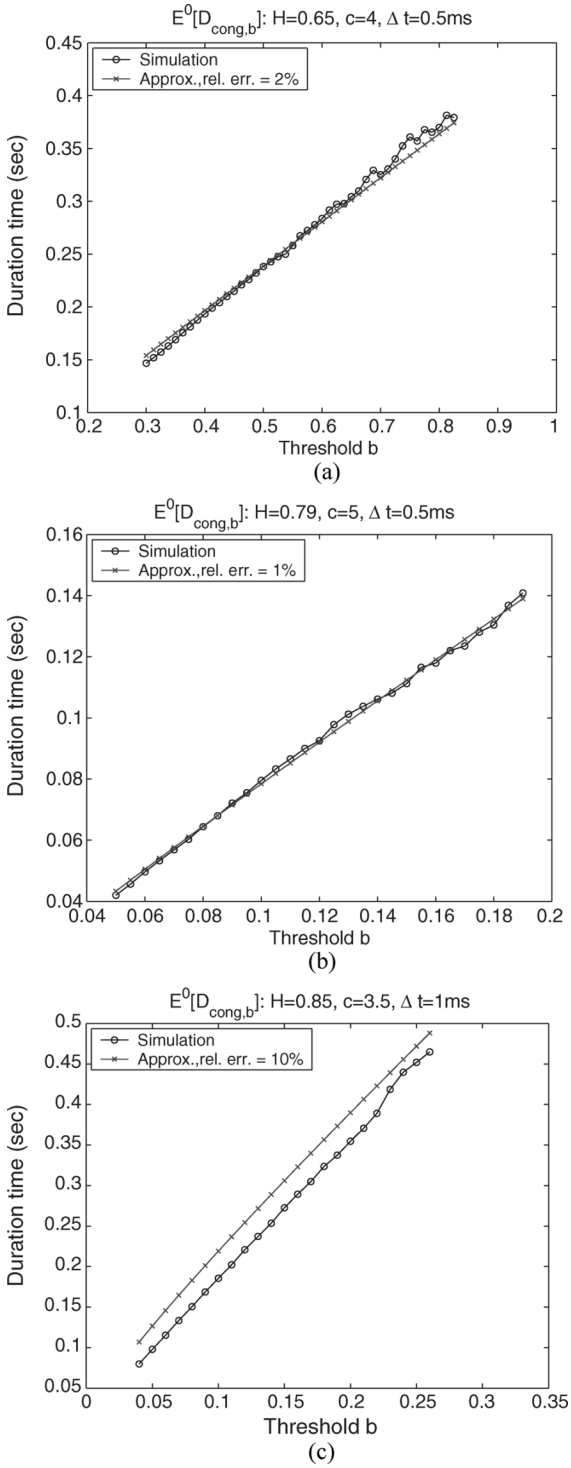


Fig. 9. Comparison of mean duration of congestion events versus b. (a) $H = 0.65$. (b) $H = 0.79$. (c) $H = 0.85$.

$E^0[C_{Q,b}]$ and $P(Q(0) \geq b)$ in (5) are measured from the simulations.

C. Mean Duration Time of Congestions $E^0[D_{cong,b}]$

It is shown in Figs. 9 and 10 that the approximation, given in (16), is close to the simulation results of $E^0[D_{cong,b}]$, the relative errors are around 10%. In all situations, as shown in Fig. 10(a)–(b), the approximations follow the trends of the simulation results.

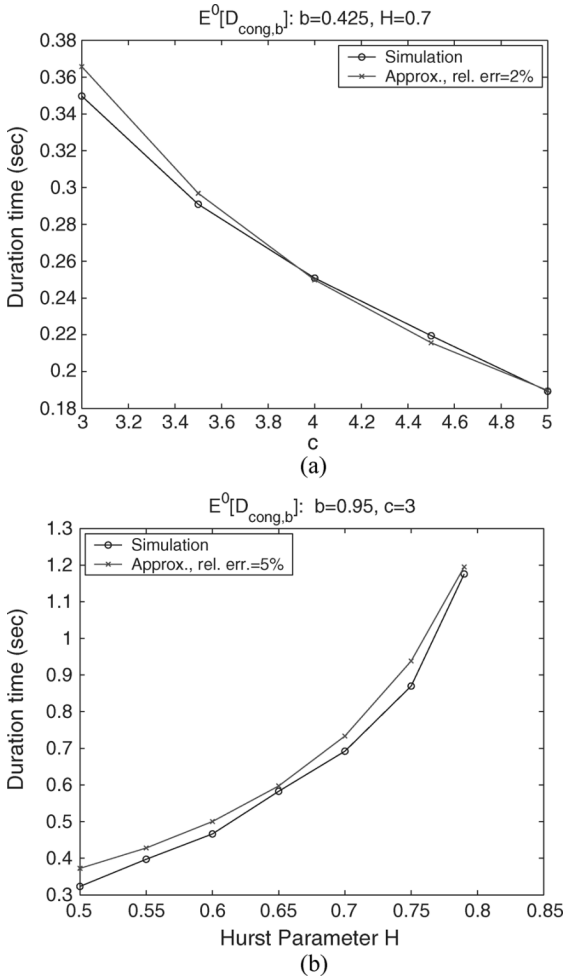


Fig. 10. Comparison of mean duration of congestion events. (a) Duration of congestions versus c. (b) Duration of congestions versus H.

D. Mean Duration Time of Busy Periods $E^0[D_{Q,b}]$

In Figs. 11 and 12, the mean durations of busy periods observed from simulations are compared with the approximation (17). We noticed that r_b, given in (8), overestimates the mean time that the queue increases from 0 to b. Thus $E^0[D_{Q,b}]$ is overestimated by the approximation. However, the approximation results follow the observed trends, the relative errors are from 10% to 30%.

E. Mean Amplitude $E^0[A_{Q,b}]$

From the simulation results, it is observed that the mean amplitude follows a linear trend as a function of the threshold b. As shown in Figs. 13 and 14, the approximations underestimate $E^0[A_{Q,b}]$. Based on (18), the underestimation is caused by the overestimation of $E^0[D_{Q,b}]$. But again the approximations follow the simulation trends, the relative errors are around 10%.

The errors in the approximations are partly from r_b, which overestimates the time that $Q(t)$ increases from 0 to b. If we

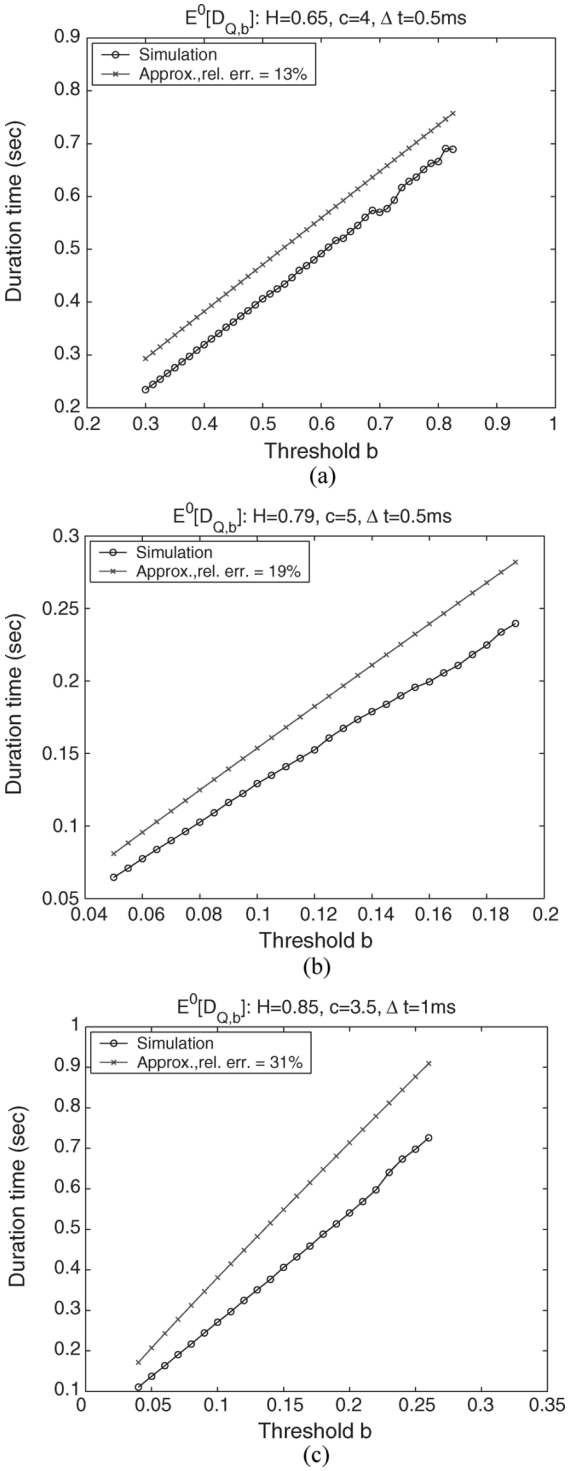


Fig. 11. Comparison of mean duration of busy periods versus b. (a) $H = 0.65$. (b) $H = 0.79$. (c) $H = 0.85$.

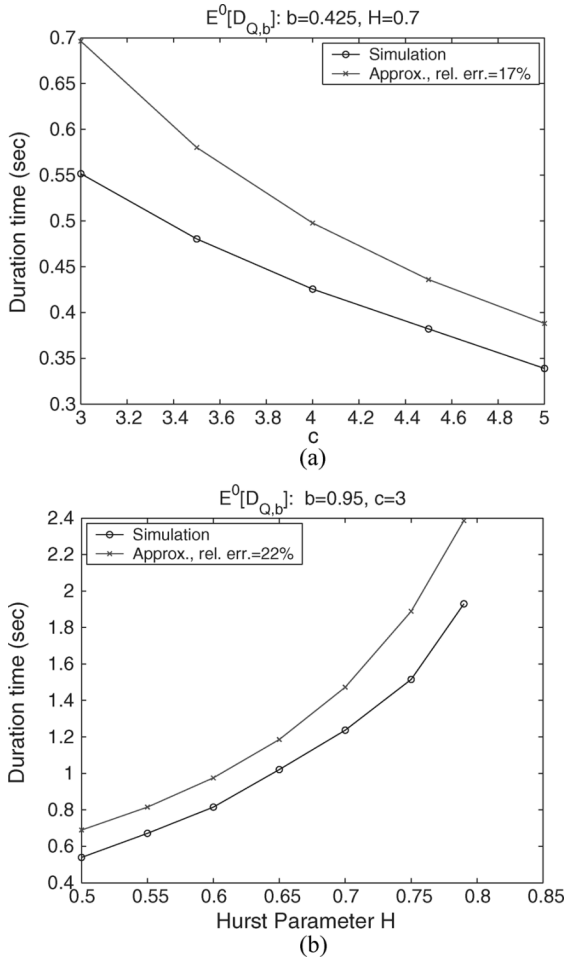


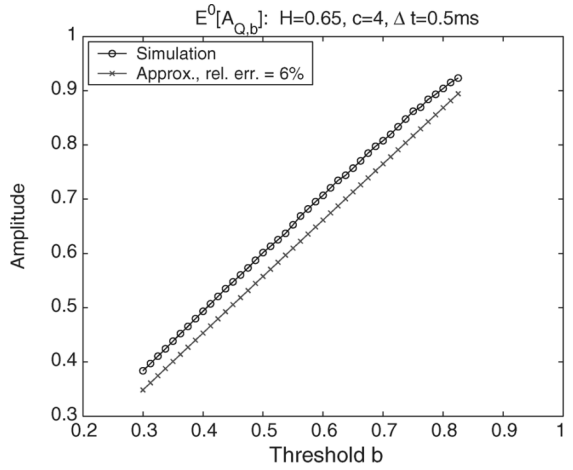
Fig. 12. Comparison of mean duration of busy periods. (a) Duration of busy periods versus c. (b) Duration of busy periods versus H.

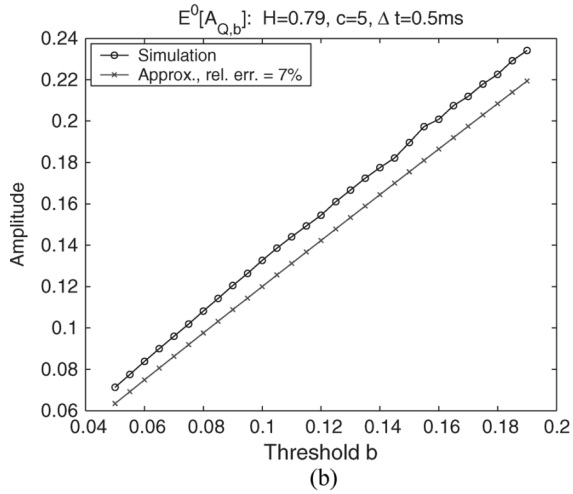
have better knowledge of r_b, the approximation results can be improved.

To illustrate an application of the proposed methodology, suppose that we need to choose a link capacity for a conferencing teleservice. The requirement of an error-free interval for audio and video multimedia conferencing teleservices is given as 30 minutes [26], i.e., the average inter-congestion event time $E^0[\tau_b]$ is 1800 seconds, $\log(E^0[\tau_b]) \approx 7.5$. Then for a fBm traffic characterized by $m = 100$ Mbps, $a = 10^{14}$ bit^2, $H = 0.75$, the proposed method indicates that for a congestion level of $b = 5.5$ Mb, a link capacity of 140 Mbps (traffic load $\rho \approx 0.7$) would be required to ensure the average congestion free interval of 30 minutes, and in this case, $E^0[C_{Q,b}] \approx 90$ ms, $E^0[D_{cong,b}] \approx 395$ ms, $E^0[D_{Q,b}] \approx 800$ ms, $E^0[A_{Q,b}] \approx 6.2$ Mb.

VII. Conclusion

It has been recognized that the frequency and the duration of congestion events significantly impact user-perceived performance. Previous efforts have focused on measurement-based approaches to determine the frequency and duration of these events. However, for network design, techniques are needed to predict the congestion events given the nature of traffic. This paper provides new techniques to approximate several properties of congestion events, their rate, duration, and amplitude given a fBm traffic. The technique to approximate the rate outperforms the reciprocal of the tail of the queue fill probability, i.e., $1/P(Q(0) \geq b)$, and follows the trends observed from simulations. As in [7], the approach for predicting the rate of congestion events can be directly extended to determine the expected rate of congestion events for an end-to-end flow that passes through several queues. Congestion events at each queue





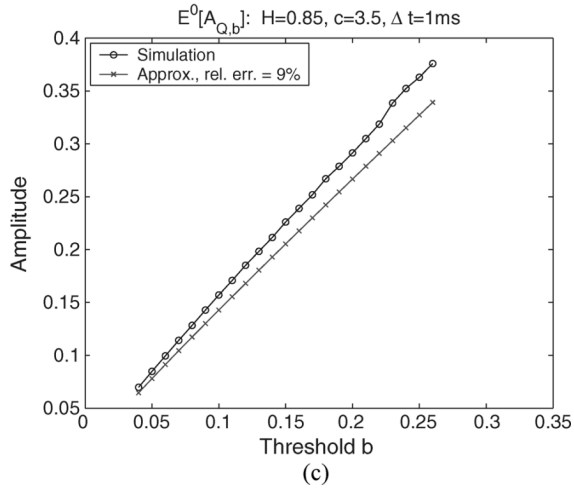
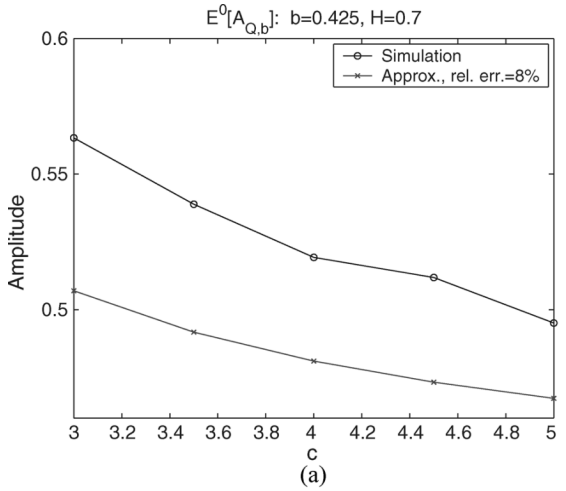


Fig. 13. Comparison of mean amplitude of congestion events versus b. (a) $H = 0.65$. (b) $H = 0.79$. (c) $H = 0.85$.

along a path can be assumed to be independent and rare, so an end-to-end flow will experience the sum of the congestion events along the path. The inter-congestion event time $E^0[\tau_b]$ (or its rate $1/E^0[\tau_b]$), which can be easily understood by network users, is a useful QoS metric for network design. The other metrics of congestion events, such as the sojourn time above a threshold, the duration, and the amplitude, give additional insights into the nature of congestion events.



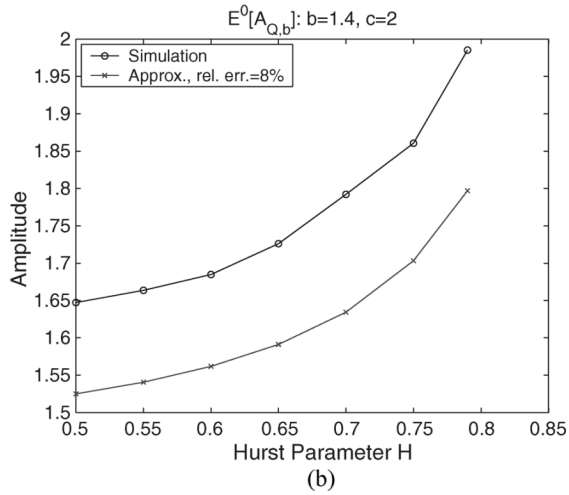


Fig. 14. Comparison of mean amplitude of congestion events. (a) Amplitude versus c. (b) Amplitude versus H.

These results can be extended in several areas. The accuracy of the techniques developed here can be improved. The properties of busy periods whose durations are larger than a fixed T, discussed in [24], [27], are interesting problems for further study. Other self-similar traffic models need to be considered, such as the Levy processes. To understand fully the impacts of self-similar traffic on networks, these processes need to be analyzed and additional methodologies developed.

REFERENCES

[1] W. Jiang and H. Schulzrinne, "Modeling of packet loss and delay and their effect on real-time multimedia service quality," in *ACM Int. Workshop on Network and Operating Systems Support for Digital Audio and Video (NOSSDAV)*, Chapel Hill, NC, Jun. 2000 [Online]. Available: http://www.nossdav.org/2000/abstracts/27.html

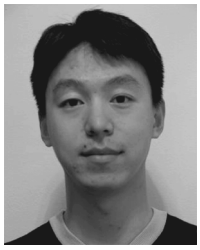
[2] Y. Zhang, N. Duffield, V. Paxson, and S. Shenker, "On the constancy of Internet path properties," in *Proc. ACM SIGCOMM Internet Measurement Workshop*, Nov. 2001, pp. 197–211.

[3] M. S. Borella, D. Swider, S. Uludag, and G. B. Brewster, "Internet packet loss: Measurement and implications for end-to-end QoS," in *Proc. 1998 Int. Conf. Parallel Processing Workshops (ICPPW'98)*, Aug. 1998, pp. 3–12.

[4] M. Yajnik, J. Kurose, and D. Towsley, "Packet loss correlation in the MBONE multicast network," in *Proc. IEEE Global Internet*, London, U.K., Nov. 1996, pp. 94–99.

[5] R. Koodli and R. Ravikanth, "One-way loss pattern sample metrics," RFC 3357, Aug. 2002.

[6] M. Yajnik, S. Moon, J. Kurose, and D. Towsley, "Measurement and modeling of the temporal dependence in packet loss," in *Proc. IEEE INFOCOM'99*, Mar. 1999, pp. 345–352.
[7] V. S. Frost, "Quantifying the temporal characteristics of network congestion events for multimedia services," *IEEE Trans. Multimedia*, vol. 5, no. 3, pp. 458–465, 2003.
[8] W. Leland, M. Taqqu, W. Willinger, and D. Wilson, "On the self-similar nature of ethernet traffic (extended version)," *IEEE/ACM Trans. Networking*, vol. 2, no. 1, pp. 1–15, Feb. 1994.
[9] M. E. Crovella and A. Bestavros, "Self-similarity in world wide web traffic: Evidence and possible causes," *IEEE/ACM Trans. Networking*, vol. 5, no. 6, pp. 835–846, Dec. 1997.
[10] V. Paxson and S. Floyd, "Wide-area traffic: The failure of Poisson modeling," *IEEE/ACM Trans. Networking*, vol. 3, no. 3, pp. 226–244, Jun. 1995.
[11] B. K. Ryu and S. B. Lowen, "Point-process approaches to the modeling and analysis of self-similar traffic," in *Proc. IEEE INFOCOM*, Mar. 1996, vol. 3, pp. 1468–1475.
[12] A. Feldmann, A. C. Gilbert, and W. Willinger, "Data networks as cascades: Investigating the multifractal nature of Internet WAN traffic," *ACM Comput. Commun. Rev.*, vol. 28, pp. 42–45, 1998.
[13] I. Norros, "A storage model with self-similar input," *Queueing Systems*, vol. 16, pp. 387–396, 1994.
[14] N. Hohn, D. Veitch, K. Papagiannaki, and C. Diot, "Bridging router performance and queuing theory," in *Proc. Joint Int. Conf. Measurement and Modeling of Computer Systems*, New York, Jun. 2004, pp. 355–366, ACM Press.
[15] M. S. Taqqu, W. Willinger, and R. Sherman, "Proof of a fundamental result in self-similar traffic modeling," *ACM Comput. Commun. Rev.*, vol. 27, pp. 5–23, 1997.
[16] T. Karagiannis, M. Molle, and M. Faloutsos, "Long-range dependence, ten years of Internet traffic modeling," *IEEE Internet Computing*, pp. 57–64, Sep., Oct. 2004.
[17] R. van de Meent, M. R. H. Mandjes, and A. Pras, "Gaussian Traffic Everywhere?," CWI, PNA-E0602, ISSN 1386-3711, Tech. Rep., 2006.
[18] F. Baccelli and P. Brémaud, *Elements of Queueing Theory*, 2nd ed. New York: Springer, 2003.
[19] T. E. Duncan, "Some aspects of fractional Brownian motion," *Nonlinear Analysis*, vol. 47, pp. 4475–4782, 2001.
[20] Y. L. Tong, *The Multivariate Normal Distribution*. New York: Springer-Verlag, 1990.
[21] D. Aldous, *Probability Approximations via the Poisson Clumping Heuristic*. New York: Springler-Verlang, 1989.
[22] J. Hüsler and V. Piterbarg, "Extremes of a certain class of Gaussian processes," *Stochastic Processes and Their Applications*, vol. 83, pp. 257–271, 1999.
[23] I. Norros, "Busy periods of fractional Brownian storage: A large deviations approach," *Adv. Perf. Anal.*, vol. 2, no. 1, pp. 1–19, 1999.
[24] M. R. H. Mandjes, P. Mannersalo, I. Norros, and M. J. G. van Uitert, "Large deviations of infinite intersections of events in Gaussian processes," *Stochastic Analysis*, to be published.
[25] R. B. Davies and D. S. Harte, "Tests for hurst effect," *Biometrika*, vol. 74, no. 1, pp. 95–101, 1987.
[26] "Multimedia communications quality of service," Multimedia Communications Forum, Sep. 1995.
[27] I. Norros, "Queueing behavior under fractional Brownian traffic," in *Self-Similar Network Traffic and Performance Evaluation*. New York: Wiley-Interscience, 2000, pp. 101–114.

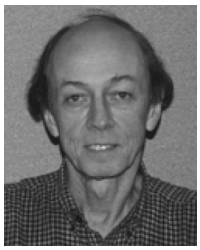


Yasong Jin (S'04) received the B.E. degree in automation from Beijing Polytechnic University, China, in 1998 and the M.A. degree in mathematics from the University of Kansas in 2003. He is a Ph.D. candidate in the Mathematics Department of the University of Kansas.

His current research interests are in stochastic analysis and queueing theory. He is a member of AMS and SIAM.

Soshant Bali received the B.E. degree in electronics engineering from Mumbai University, India, in July 2000 and the M.S. degree in electrical engineering from Virginia Tech in December 2002. He is a Ph.D. candidate in the Electrical Engineering and Computer Science Department at the University of Kansas. His advisor is Dr. Victor S. Frost.

His research interests include Internet measurements and quality of service.



Tyrone E. Duncan (M'92–SM'96–F'99) received the B.E.E. degree from Rensselaer Polytechnic Institute in 1963 and the M.S. and Ph.D. degrees from Stanford University in 1964 and 1967, respectively.

He has held regular positions at the University of Michigan (1967–1971), the State University of New York, Stony Brook (1971–1974) and the University of Kansas (1974-present), where he is Professor of Mathematics. He has held visiting positions at the University of California, Berkeley (1969–1970), the University of Bonn, Germany (1978–1979), and Harvard University (1979–1980) and shorter visiting positions at numerous other institutions.

Dr. Duncan is a Corresponding Editor of *SIAM Journal on Control and Optimization* and is a member of AMS, MAA, and SIAM.

Victor S. Frost (S'75–M'82–SM'90–F'98) received the B.S., M.S., and Ph.D. degrees from the University of Kansas, Lawrence, in 1977, 1978, and 1982, respectively.

In 1982, he joined the faculty of the University of Kansas. He is currently the Dan F. Servey Distinguished Professor of Electrical Engineering and Computer Science and Director of the University of Kansas Telecommunications and Information Technology Center (ITTC). His current research interest is in the areas of Internet quality of service, traffic management, and integrated broadband communication networks. He has been involved in research on several national scale high speed wide area testbeds. Government agencies, including, NSF, DARPA, Rome Labs, and NASA have sponsored his research. He has been involved in research for numerous corporations, including Harris, Sprint, NCR, BNR, Telesat Canada, AT&T, McDonnell Douglas, DEC, and COMDISCO Systems. He has published over 100 journal articles and conference papers.

Dr. Frost is a Fellow of the IEEE and received a Presidential Young Investigator Award from the National Science Foundation in 1984.