Policing and Statistical Multiplexing of Self-Similar Sources

N.L.S Fonseca, G. S. Mayor e C.A.V. Neto

Relatório Técnico IC 99-01

Fevereiro de 1999

Policing and Statistical Multiplexing of Self-Similar Sources

Nelson L. S. Fonseca, Gilberto S. Mayor and Cesar A. V. Neto

State University of Campinas
Institute of Computing
P.O.Box 6176
13083-970 Campinas SP
Brazil
Phone: +55+19+7885878
Fax:+55+19+7885847
e-mail: nfonseca@dcc.unicamp.br

Keyword: **Modeling and Simulation of Communications Systems**

Abstract

In this paper, we describe an envelope process for the fractal Brownian motion (fBm) process. We investigate the time scale of interest of queuing systems fed by a fBm process. We also introduce the fractal leaky bucket, a novel policing mechanism which is able to accurately monitor self-similar sources. Moreover, we show expressions for computing the equivalent bandwidth of an aggregate of heterogeneous self-similar sources.

I) Introduction

Several studies [1]-[4] have claimed that different types of network traffic, *e.g*. local area network traffic (LAN), can be accurately modeled by a self-similar process. A

self-similar process is able to capture the long-range dependence (LRD) phenomenon exhibited by this traffic. Moreover, series of simulation and analytical studies [5]-[8] have demonstrated that this phenomenon might have a pervasive effect on queueing performance. In fact, there is clear evidence that it can potentially cause massive cell losses in ATM networks. Norros [7] and Duffield [9] showed that the buffer overflow probability for an ATM queueing system with fractional Brownian arrivals follows a Weibull distribution. Furthermore, this queueing system suffers from the buffer inefficacy phenomenon [6], [7]. By just increasing the buffer size we are not able to significantly decrease the buffer overflow probability. Although several work have analyzed the self-similar nature of traffic, control mechanisms for self-similar traffic have not yet been fully investigated.

One of the key ideas behind Asynchronous Transfer Mode is the statistical multiplexing of heterogeneous packetized streams. The concept of Effective Bandwidth is intimately connected with admission control and associated service requirement [11]. The equivalent bandwidth of a connection (source) is a characterization of the demanded bandwidth of the connection such that its QoS requirements are provided in a network based on statistical multiplexing. Designers have gravitated towards the concept of equivalent bandwidth because it promises to bridge to familiar circuit-switched network design. Although there is a remarkable collection of equivalent bandwidth results (mainly based on the theory of large deviation [12]-[15] and on spectral expansion for Markov fluid models [11]) very few results are available for traffic with long-range dependencies [12]-[14].

Once a connection is admitted into the network, we need to police it to make sure that the generated stream is in accordance with the traffic descriptors declared at the connection admission time. The leaky bucket, the most popular policing mechanism, suffers from severe drawbacks for monitoring bursty sources due to the reduced number of parameters which can be set [16]-[18]. In the case of LRD processes, setting the parameters of a LB is a complex task since the variance of a stream increases with t^{2H} where t denotes time and H the Hurst parameter.

In [19] we proposed a new traffic model called a fractional Brownian motion (fBm) envelope process which characterizes a LRD source. We also derived a new framework for computing probabilistic delay bounds for a deterministic queuing system, as a model of an ATM network, driven by this source. We showed that the

delay bounds agree with known results obtained by large deviation theory. This new traffic characterization made possible a more intuitive understanding of the dynamics of the queuing system, and we derive three time-scales that completely characterize the queuing system behaviour in [20]. We analyzed different buffer management policies for providing diverse loss requirements for self-similar sources in overflow situations in [21]-[23].

In this paper, we introduce a novel policing mechanism called the fractal leaky bucket (FLB) which is able to monitor sources with long-range dependencies. We compare the effectiveness of the fractal leaky bucket with the effectiveness of the LB for monitoring LRD sources. Furthermore, we extend our previously defined framework to model statistical multiplexing of heterogeneous LRD sources. Moreover, we show expressions to compute the equivalent bandwidth of an aggregate of heterogeneous LRD sources. Our approach allow us to compute bound with little computational effort yet achieve the same accuracy of results predicted by the Large Deviation Theory.

This paper is organized as follow. In section II we show an envelope process for a fractal Brownian motion process. In section III we derive the time scale of interest for a queuing system fed by a self-similar process. In section IV we introduce the fractal leaky bucket and in section V we study statistical multiplexing of heterogeneous self-similar sources. Finally, conclusions are drawn in section VI

II) A Fractal Brownian Motion Envelope Process

It is well known that for a Brownian motion (Bm) process *A(t)* with mean $\bar{a}$ and variance σ^2, the envelope process $\hat{A}(t)$ can be defined by [24]

$$\hat{A}(t) \stackrel{def}{=} \bar{a}t + k\sqrt{\sigma^2 t} = \bar{a}t + k\sigma t^{1/2}$$

The parameter *k* determines the probability that *A(t)* will exceed $\hat{A}(t)$ at time *t.* Since *A(t)* is a Brownian motion process we can write:

$$P\left(\frac{A(t)-\bar{a}t}{\sigma t^{H}} > k\right) = \Phi(k)$$

where $\Phi(y)$ is the residual distribution function of the standard Gaussian distribution. Using the approximation $\Phi(y) \approx (2\pi)^{-1/2}(1+y)^{-1}\exp((-y^2/2) \approx \exp(-y^2/2))$

we find k such that $\Phi(k) \leq \varepsilon$. Hence, k is given by $k = \sqrt{-2\ln\varepsilon}$

We claim that $P(A(t) > \hat{A}(t)) \approx \varepsilon$, where $k = \sqrt{-2\ln\varepsilon}$. This approach can be extended to deal with LRD traffic. Let $A_H(t)$ be a fractional Brownian motion process with mean $\bar{a}$. Hurst's law states that the variance of the increment of this process is given by $Var[A_H(t+s) - A_H(t)] = \sigma^2 s^{2H}$ where $H \in [1/2, 1)$ is the Hurst parameter. Thus, we can also define a fBm envelope process by:

$$\hat{A}_H(t) \stackrel{def}{=} \bar{a}t + k\sqrt{\sigma^2 t^{2H}} = \bar{a}t + k\sigma t^H \quad (1)$$

The Brownian motion envelope process is just the special case of H = 1/2. Similarly, k determines the probability that $A_H(t)$ will exceed $\hat{A}_H(t)$. In addition, since the process exhibits LRD, if $A_H(t)$ exceeds $\hat{A}_H(t)$ at time t, it is possible that it will stay above it for a long period of time.

We should note that the source does not necessarily need to be self-similar in order to match this characterization, as long as it matches the behaviour of the envelope process over the time-scale of interest. We investigate the accuracy of the fBm envelope process representation by inspecting how well it can model the worst-case behaviour of real network traffic. Assume that the input traffic is characterized by trace with N sample points, defined by $A(t)$, where $A(t)$ represents the cumulative number of cell arrivals up to time t, $t \in [1,2\ldots,N]$. We propose a very simple method for computing the fBm envelope process parameters for this trace, by computing the trace's optimal envelope process. The advantage of this approach relies on the fact that we do not need to accurately estimate the trace's Hurst parameter. The optimal envelope process (the worst-case sample path) for this trace is defined by $Y(t-s) = max_{s<t}(A(t) - A(s))$. We assume that the process is stationary so that $Y(\tau)$, $\tau = t - s$ defines the maximum number of cell arrivals in an interval of size τ. Therefore, we can choose the fBm envelope process's parameters $\hat{A}_H(.)$ so that it matches the behaviour of $Y(.)$.

We compare the envelope process representation to Bellcore's LAN trace. We compute the sample average arrival rate and the sample variance for this trace and substitute for $\bar{a}$ and σ^2 in Equation 1. We compute the optimal envelope process, i.e

$Y(.)$, and choose H so that $\hat{A}_H(.)$ matches the behaviour of $Y(.)$. In Figure 1, the upper curve corresponds to the fBm envelope process with $\varepsilon = 10^{-3}$. The lower curve represents the Brownian motion envelope process with same ε. The middle curve corresponds to $Y(\tau)$. We can see that the fBm envelope process matches closely the behaviour of the LRD trace. Moreover, we also note that the ordinary Brownian motion envelope process is unable to bound the behaviour of the LRD source even if we choose ε large.

We extensively validated the effectiveness of the fractal Brownian motion envelope process by utilizing synthetic traces generated by Mandelbort's procedure [25]-[26]. For every trace used, we verified if the mean, the variance and the Hurst parameter were in agreement with the specified values. We investigate the accuracy of the envelope process by varying the traffic parameter in the following range: $\bar{a} \in [0.5,1.0]$, $\sigma^2 \in [0.01,0.7]$, $H \in [0.5,1.0]$, $\varepsilon \in [10^{-3},10^{-9}]$, where the mean and the variance are normalized to the channel capacity. Results indicate that the fBm envelope process is a close upperbound for a fBm process. Moreover, the fBm envelope process is highly accurate in all the mentioned ranges. Figure 2 illustrates the fBm envelope precision for a process with $\bar{a} = 0.8$, $\sigma^2 = 1.0$, $\varepsilon = 10^{-6}$ and different H values.

The fBm envelope presents several advantages:

• It is parsimonious, i.e. only three parameters are required in order to completely characterize a source;

• It can represent SRD and LRD, i.e, the source does not necessarily need to be LRD. We need only to choose the parameters for the fBm envelope process so that it matches the source's optimal envelope process over the appropriate time-scale;

• The input parameters $\bar{a}$, σ, and H can be provided by the source or estimated in real-time from the incoming traffic sample by estimating its optimal envelope process;

• It provides very accurate delay bounds with minimal computational complexity.

III) Time Scale of Interest

In this section, we show the time until a queue reaches its maximum occupancy, in a probabilistic sense. The queue size at this time gives us a simple delay bound [19].

A rigorous mathematical derivation of the delay bound can be found in [20]. Here, we introduce an heuristic derivation in order to preserve the intuition behind the framework presented in this paper. Consider a continuous-time queuing system, with deterministic service given by *c*. The cumulative arrival process is given by $A_H(t)$ ($A_H(0) = 0$). Let $\hat{A}_H(t)$, continuous and differentiable, be the probabilistic envelope process of $A(t)$ such that

$$P(A_H(t) > \hat{A}_H(t)) \leq \varepsilon$$

During a busy period which starts at time 0, the number of cells in the system at time *t* is given by $q(t)$. Thus,

$$q(t) = A_H(t) - ct \geq 0 \quad .$$

By defining $\hat{q}(t)$ as

$$\hat{q}(t) = \hat{A}_H(t) - ct \geq 0 \quad (2)$$

We can see that

$$P(q(t) > \hat{q}(t)) = P(A_H(t) > \hat{A}_H(t)) \leq \varepsilon$$

The maximum delay in a FIFO queuing system is given by the maximum number of cells in the queue during the busy period. We define

$$q_{max} \stackrel{def}{=} max(\hat{q}(t)) \qquad t \geq 0$$

Therefore,

$$P(q(t) > q_{max}) \leq P(q(t) > \hat{q}(t)) \leq \varepsilon$$

$$P(q(t) > q_{max}) \approx \varepsilon$$

We can say that the queue length at time *t* $q(t)$ will only exceed the maximum queue length q_{max} with probability ε. In other words, only when the arrival process exceeds the envelope process, will the maximum number of cells in the system exceed its estimated value. Intuitively, by bounding the behaviour of the arrival process we are able to transform the problem of obtaining a probabilistic bound of the stochastic system defined by $q(t) = A_H(t) - ct \geq 0$ into an easier problem of finding the maximum of a deterministic system described by $\hat{q}(t) = \hat{A}_H(t) - ct \geq 0$.

For the case of the fBm process, we substitute the envelope process defined

previously into Equation 2 which gives

$$\hat{q}(t) = \hat{A}_H(t) - ct = \bar{a}t + k\sigma t^H - ct \quad (3)$$

In order to compute q_{max} we need to find t^* such that

$$\frac{d\hat{q}(t^*)}{dt} = 0$$

or equivalently,

$$\frac{d\hat{A}_H(t^*)}{dt} = c \quad (4)$$

Hence, t^* is given by

$$t^* = \left[\frac{k\sigma H}{(c-\bar{a})}\right]^{\frac{1}{1-H}}$$

The time-scale of interest is defined by the time until a queue size reaches its peak, i.e., t^*. We call it the Maximum Time-Scale (MaxTS), and it defines the point in time where the unfinished work in the queuing system achieves its maximum in a probabilistic sense. It means that the average arrival rate just dropped below the link capacity so that the queue size starts decreasing. The average arrival rate converges to the source's mean arrival rate by the law of large numbers. Consequently, we only need to worry only about the time scale for which the source's rate still exceeds the link capacity, in a probabilistic sense. In other words, after a period of time, the probability that the average arrival rate exceeds the link capacity is negligible, so that the arrival model does not need to reproduce the source's behaviour for those time-scales. This is the most important time-scale in terms of traffic modelling. As a rule of thumb to choose the parameters of an input source in order to match the fBm envelope process, we need to find MaxTS analytically, and to choose the parameters of the fBm process, so that it matches the source's optimal envelope process at MaxTS.

Substituting t^* back into Equation 2, we conclude that:

$$q_{max} = \hat{A}_H(t^*) - ct^* \quad (5)$$

$$q_{max} = (c-\bar{a})^{\frac{H}{H-1}} \; (k\sigma)^{\frac{1}{1-H}} \; H^{\frac{H}{1-H}} \; (1-H)$$

Since the fBm process does not exceed $\hat{A}_H(t)$ with probability 1 - ε, the maximum

number of cells will be bounded by q_{max} with the same probability. We find $\hat{c}$ so that q_{max} is equal to K. In other words, a buffer of size K will overflow with probability ε if the link capacity is $\hat{c}$. Therefore, $\hat{c}$ is given by

$$\hat{c} = a + K^{\frac{H-1}{H}} (k\sigma)^{1/H} H (1-H)^{\frac{H-1}{H}}$$

This result was also obtained by Norros [7] [27] and Duffield [9]. In summary, our framework allow us to compute delay bounds with little computational effort yet achieve the same accuracy of the results predicted by large deviation theory. We have also reduced the sensitivity of the estimation process by using a bound rather than attempting to directly estimate the parameters from the full trace.

In [20] a comparison between the MaxTS and the so called Critical Time Scale (CTS) [28]-[29] is provided. It is shown that CTS is the most likely time scale when buffer overflow occurs, but it does not furnish any information about the marginal distribution of the random variable which describes the overflow process.

IV) The Fractal Leaky Bucket

Once a connection is admitted, we need to police it in order to make sure that the generated stream of cells is in accordance with the declared traffic descriptors at the connection admission time. An ideal policing mechanisms allows cells (packets) into the network if and only if the connection is well-behaved. Otherwise it should drop incoming cells or mark them as low priority.

The leaky bucket (LB) uses two parameters to control a connection transmission: the leaky rate and the bucket size. It has extensively been shown that it is very hard to police bursty sources using only these two parameters. If we set the leaky rate close to the source mean rate we may drop (or mark) well-behaved sources. In addition, if we increase the bucket size we may allow long bursts into the network [16]-[18]. One way to set the leaky bucket parameters is to solve a G/D/1/k queue. Depending on the arrival process, choosing the leaky bucket parameters can be a complex problem. To overcome this difficulty, we propose a very simple calculus based on the fBm envelope process to set the LB parameters without having to solve a queuing system.

The leaky bucket can be seen as a traffic regulator with output given by the process $L(t)$ so that $L(t) \leq Rt + S$. In other words, in an interval of length t the leaky

bucket accepts up to ($Rt + S$) cells. $L(t)$ can also be seen as a deterministic envelope process, i.e. it defines the maximum number of cells that a source can send in any time interval.

Let $A_H(t)$ and $\hat{A}_H(t)$ be the cumulative number of cell arrivals up to time t (cumulative arrival process), and its probabilistic envelope process at time t, respectively. In order to minimize the probability of incorrectly dropping cells, we should have $L(t) \geq A_H(t), \quad \forall t > 0$. Furthermore, as long as the source is not misbehaving, the LB mechanism should not mark any incoming cell. If we assume that the probability of a source exceeding its probabilistic envelope process is negligible, we can write $A_H(t) \leq \hat{A}_H(t), \quad \forall t > 0$. In addition, we can assume that

$$L(t) \geq \hat{A}_H(t), \quad \forall t > 0 \quad (6)$$

since the probabilistic envelope process is a tighter bound than the deterministic envelope process $L(t)$. Hence, by substituting the fBm envelope process formula into Equation 6, we can write:

$$\bar{a}t + k\sigma t^H \leq Rt + S$$

Moreover, we have

$$t(\bar{a} - R) + k\sigma t^H - S \leq 0 \quad (7)$$

In order to choose the leaky bucket parameters, we find t^* which maximizes Equation 7:

$$t^* = \left\lceil \frac{k\sigma H}{R - a} \right\rceil^{\frac{1}{1-H}}$$

Substituting t^* back into Equation 7 we get

$$(\bar{a} - R)\left\lceil \frac{k\sigma H}{R - \bar{a}} \right\rceil^{\frac{1}{(1-H)}} + k\sigma \left\lceil \frac{k\sigma H}{R - \bar{a}} \right\rceil^{\frac{H}{(1-H)}} - S \leq 0 \quad (8)$$

Therefore, by using Equation 8 we can compute R given S, or vice versa. In the case of a Brownian motion process which has independent and identically distributed arrivals, Equation 8 degenerates in a simple quadratic equation. For the general case, we can solve it numerically.

In Figure 3 we show the required bucket size as a function of the ratio between the source mean arrival rate and the leaky rate. The dotted line corresponds to $H = 0.5$

whereas the solid curve corresponds to $H = 0.9$. Note that for high values of the Hurst parameter if R is close to the mean rate, S is prohibitively large. Actually, even for a leaky rate twice the value of the source mean arrival rate the required bucket size is considerably large (10^{+4}).

The problem with the leaky bucket, is that it assumes that the aggregate arrival process behaves as a linear function of time. This is not true even in the simplistic case of Brownian motion process. For example, the envelope process of a Brownian motion process with mean $\bar{a}$ and standard deviation σ is given by $\hat{A}(t) = \bar{a}t + k\sigma t^{1/2}$.

We can see that the aggregate traffic is *not* a linear function of time since its envelope process has the additional term $t^{1/2}$. For this reason, it is extremely hard to set-up the leaky bucket's parameters for bursty sources, i.e. when the standard deviation is large. Indeed, we have to choose R large enough in order to account for the term $t^{1/2}$. Otherwise, we have to choose a high leaky rate so that it includes both terms $\bar{a}t + k\sigma t^{1/2}$. In the case of LRD sources, the problem is even more complex, since the variance increases with t^{2H}, where $2H > 1$. Therefore, we claim that the leaky bucket is not very tight envelope process when the source has large variability.

We propose a new traffic regulator based on a better characterization of the input traffic named the fractal leaky bucket (FLB). The amount of work accepted by the fractal leaky bucket is given by:

$$\hat{L}(t) = \bar{a}t + \Psi\sigma t^{H} + S \quad (9)$$

where $\bar{a}$ is the average arrival rate of the source. Ψ is given by $k\sigma$, where k is a constant and σ is the source's standard deviation.

The fractal leaky bucket works as follows. We define a time window with duration of τ time units. We verify if the number of arrivals during this time window exceeds the declared mean value ($\bar{a} \times \tau$). If it does exceed, we compare the cumulative number of arrivals during this time window against the allowed number cells by the FLB envelope process during the same period (Equation 9). If the number of arrivals surpasses the allowed number of cells by the FLB envelope process, we mark all cells in excess. We then increase the time window of τ time units, i.e., we consider a time window of duration 2τ. This new window begins at the time the arrival process violated the declared mean arrival rate. We compare again the cumulative number of arrivals during

this time window of duration 2τ against the allowed mean number of arrivals. If the number of cells surpasses the allowed mean number of arrivals we compare it to the number of arrivals allowed by the FLB envelope process and mark the exceeding number of cells less the number of cells already marked in the previous window. While the mean number of arrivals surpasses the declared mean value, we continue increasing the sampling interval (time window) in units of τ time units. Whenever the mean number of arrivals drops below the declared value we shrink back the time window to τ time units and continue checking the mean arrival rate.

A mathematical description of the FLB dynamics is as follows. Let $C(\tilde{t} + n\tau)$ define the cumulative number of cell arrivals during the interval $[\tilde{t}, \tilde{t} + n\tau]$

$$C(\tilde{t} + n\tau) = A(\tilde{t} + n\tau) - A(\tilde{t})$$

where *A* (t) is the number of arrivals up to time *t*

We check if $C(\tilde{t} + n\tau)$ exceeds the allowed mean number of arrivals during the interval $n\tau$, i.e, $\bar{a} \times n\tau$. If it does exceed, we verify if the number of arrivals exceeds the number of arrivals allowed by the envelope process $\lambda(\tilde{t} + n\tau)$ where $\lambda(\tilde{t} + n\tau) = \hat{L}(\tilde{t} + n\tau) - \hat{L}(\tilde{t})$. If it does surpasses we mark the exceeding number of cells not marked in previous windows, i.e., we mark $C(\tilde{t} + n\tau) - \lambda(\tilde{t} + n\tau) - C(\tilde{t} + (n-1)\tau) + \lambda(\tilde{t} + (n-1)\tau)$. We then increase the time window ($n = n + 1$) and repeat the whole process. Whenever, the mean arrival rate drops below the declared value $\bar{a}$, we shrink back the time window to τ units (Figure 4).

We simulate the FLB mechanism in order to compare it to the LB mechanism. In this example, we use a 20 minute sequence of the video STAR WARS as the input traffic. This MPEG sequence has LRD and can be characterized by a fBm envelope process. The FLB parameters are given by the mean and standard deviation of this sequence. We choose $k = \sqrt{-2\ln(10^{-6})} = 5.25$ and $H = 0.90$ so that the FLB matches the optimal envelope process of this sequence. We increase the mean arrival rate *k* times its nominal value and compute the new VP for $k \in [1.5 \ldots 3]$. In this case, the source is misbehaving so that flow control mechanism should achieve a high VP. We choose the leaky rate and the LB's threshold, so that both mechanisms have the same VP at nominal mean rate. Figure 5 shows that the FLB achieves a higher VP than

the LB's VP, when the source misbehaves. We can see that although the VP is very low at nominal rate, whenever the source's achieves 1.5 its mean rate, the FLB achieves a VP above 10%.

Figures 6 and 7 compare the ability of the fractal leaky bucket to the ability of the leaky bucket in monitoring well-behaved sources. In our simulation experiments we used Mandelbort's procedure [26] to generate a fBm trace and the method of independent replication to derive confidence intervals with 99% confidence level. In Figure 6 we display the violation probability as a function of the mean arrival rate for a fBm with $H = 0.8$, $\sigma^2 = 1.0$. To derive the FLB we use $\varepsilon = 10^{-3}$, i.e., the fractal leaky bucket is based on a probabilistic envelope process which allows violation of the predicted number of arrivals of at most 10^{-3}. We can see that the leaky bucket gives high violation probabilities even for high values of the leaky rate. On the one hand, we can observe that the leaky bucket with parameters set according to the procedure previously described in this section provides violation probabilities which are roughly in the same order of magnitude of the envelope process violation probability. On the other hand, the fractal leaky bucket gives violation probabilities which are several order of magnitude lower that the LB. In other words, the FLB is able to accurately monitor a fBm process. It is worth mentioning that for SRD processes and large bucket sizes the violation probability is sensitive to high leaky rate values. However, this is not true for LRD processes since the LB is not able to couple with long bursts existing in LRD processes. In Figure 7 we plot the violation probability as a function of the Hurst parameter. The observations made about Figure 6 are carried over to Figure 7. The LB gives high violation probability even for high leaky rates.

In Figure 8 we increase the mean arrival rate k times the nominal value. Under an ideal policing mechanism the violation probability jumps from a very low value to 1 as soon as the source starts transmitting with mean arrival rate above the declared value. We notice that the violation probability given by the fractal leaky bucket follows a pattern which is similar to an ideal mechanism behaviour except that it reacts to violation at a rate 1.15 the nominal value. This reaction delay is quite acceptable for a non-ideal mechanism. Conversely, the leaky bucket makes no significant distinction between a violating source and a non-violating source. Furthermore, the violation probability given by the leaky bucket does not tend to 1 for reasonably high arrival rates. It is worth commenting that such delay is in the same range of the reaction delay

given by LB based mechanisms for SRD processes.

In Figure 9 we keep $\bar{a}$ = 0.8, H = 0.8 and we increase the variance from its nominal value. We realize again that the LB does not differentiate violating sources from non-violating sources. In addition, the LB violation probability does not tend to 1 even for high values of the variance. On the contrary, the fractal leaky bucket does penalize violating sources.

The fractal leaky bucket mechanism compares from time to time the cumulative number of arrivals with the number of arrivals allowed by the FLB envelope process. We investigate whether results are sensitive to the time window duration. We consider time windows 10, 100 and 1000 longer than the window size used in the previous examples. We observe that the outcomes are not dependent on the duration of the time window. Figures 10 and 11 display respectively the violation probability as a function of both the arrival rate and the variance for different window durations. Such findings reinforce the robustness of the fractal leaky bucket as policing mechanism for fBm sources.

V) Statistical Multiplexing of Self-Similar Sources

In this section, we use MaxTS to derive expressions for predicting the equivalent bandwidth and buffer requirements of an aggregate of self-similar sources. Essentially, we propose a way to compute the demanded bandwidth to support requirements of buffer overflow as well as a maximum probabilistic delay for an aggregate of sources with diverse traffic parameters. The problem we study in this section can be stated as:

Given a set of sources with mean $\bar{a}_i$, standard deviation σ_i and Hurst parameter H_i, what is the link capacity needed so that the maximum queue size will be bounded by q_{max} with probability ε?

Assume that we have N independent sources $A_H^i(t)$ defined by the following parameters: mean $\bar{a}_i$, standard deviation σ_i and Hurst parameter H_i for $i \in [1,N]$. Let the aggregate traffic be denoted by $A_H(t) = \sum_{i=1}^{N} A_H^i(t)$. The envelope process of each source is given by $\hat{A}_H^i(t)$, and the envelope process of the aggregate traffic is provided

by $\hat{A}_H(t)$. We can compute q_{max} of a queue with heterogeneous sources by finding t^* for the envelope process of the aggregate stream.

The mean of the aggregate traffic is given by the sum of the mean of individual sources. Similarly, since the sources are independent, the variance of the aggregate traffic is also given by the sum of the variance of individual sources. Hence, the envelope process of the aggregate traffic is defined by

$$\hat{A}_H(t) = \sum_{i=1}^{N} a_i t + k\left(\sum_{i=1}^{N} \sigma_i^2 t^{2H_i}\right)^{1/2}$$

By substituting $\hat{A}_H(t)$ in equation 4, we have:

$$k\frac{1}{2}\left(\sum_{i=1}^{N} \sigma_i^2 t^{2H_i}\right)^{-1/2}\left(\sum_{i=1}^{N} \sigma_i^2 \; 2H_i \; t^{2H_i-1}\right) = c - \sum_{i=1}^{N} \bar{a}_i \quad (10)$$

We can solve equation 10 numerically in order to find t^* and then substitute t^* into Equation 5 to compute q_{max}

Moreover, by combining Equations 4 and 5, we have:

$$k\frac{1}{2}\left(\sum_{i=1}^{N} \sigma_i^2 t^{2H_i}\right)^{-1/2}\left(\sum_{i=1}^{N} \sigma_i^2 2H_i t^{2H_i-1}\right) - k\left(\sum_{i=1}^{N} \sigma_i^2 t^{2H_i-2}\right)^{1/2} + \frac{q_{max}}{t} = 0 \quad (11)$$

By using Equations 10 and 11 we can answer the fundamental question posed in the beginning of this section.

For the special case of multiplexing *N* identical sources, the envelope process is given by:

$$\hat{A}_H(t) = N\bar{a}t + \sqrt{N}k\sigma t^H$$

insofar as the Hurst parameter is preserved when aggregating *N* identical sources.

In this case Equation 10 is, reduced to:

$$\frac{k}{2}\frac{(N\sigma^2 2Ht^{2H-1})}{(\sqrt{N}\sigma t^H)} = N(c-\bar{a})$$

Using the previous approach, we can find t^* and $q_{max:}$

$$t^* = \left[\frac{\sqrt{N}k\sigma H}{N(c-\bar{a})}\right]^{\frac{1}{1-H}} = N^{\frac{1}{2(H-1)}} t_i^*$$

$$q_{max} = N(\bar{a}-c)N^{\frac{1}{2(H-1)}} t_i^* + N^{\frac{H}{2(H-1)}} N^{1/2} k\sigma (t_i^*)^H = N^{\frac{(H-1/2)}{H-1}} \hat{q}_{max}$$

$$t_i^* = \left[\frac{k\sigma H}{(c-\bar{a})}\right]^{\frac{1}{1-H}}$$

$$\hat{q}_{max} = \hat{A}_H(t^*) - c\,t_i^*$$

where t_i^* and $\hat{q}_{max}$ corresponds to a queueing system fed by just one source.

We first analyze the specific case of a single source. Figure 12 displays the overflow probability as a function of the maximum buffer size for a link utilization of 60% and for different values of the Hurst parameter. We compare the overflow probability given by the analytical models with the overflow probability observed in the simulation experiments. It can be noted that the analytical model is in fair agreement with the simulation model. We observe that the higher the Hurst parameter the more precise are the analytical results. Furthermore, we verified that our equivalent bandwidth expressions increase with the link utilization.

To evaluate the effectiveness of the equivalent bandwidth expressions (Equations 10/11), we define multiplexing gain as the ratio between *N* times the equivalent bandwidth of a single source and the equivalent bandwidth of *N* identical sources. We realize that a significant multiplexing gain can be achieved when multiplexing homogeneous sources. In Figure 13 we plot the gain for a link capacity of 150 Mbits and sources with mean arrival rate 1.1Mbps for different Hurst parameter. Figure 13.a displays the multiplexing gain for sources with σ^2 = 0.01 whereas Figure 13.b considers sources with σ^2 = 0.3. We observe that the gain for streams with moderate to high variance (σ^2 = 0.3) is significantly higher than for streams with low variance (σ^2 = 0.01). While for streams with *H* = 0.9 and low variance the gain is 1.25, it is almost 2.5 for streams with moderate to high variance. The multiplexing gain also increases with the Hurst parameter, specially for streams with moderate to high variance. This can be understood by the fact that Equations 10/11 take into consideration the existence of long periods with no arrivals in streams with high Hurst parameters. As a consequence they demand less bandwidth when multiplexing

several sources than a no-multiplexing approach.

The multiplexing gain increases not only with the variance and Hurst parameter but also with the number of multiplexed sources. Figure 14 illustrates the multiplexing gain as a function of the Hurst parameter for 10 and 100 sources. It is worth mentioning the benefit of taking into account MaxTS when computing the demanded bandwidth since the gain increases with the number of sources. For instance, the multiplexing gain for H = 0.85 is 1.7 for 10 sources while it is 2.5 for 100 sources. We also observe that the higher the Hurst parameter the larger is the difference between the equivalent bandwidth predicted by Equations 10/11 and a no-multiplexing approximation. Indeed, for 100 sources we note that for H = 0.6 the multiplexing gain is 1.5 whereas it is 3.0 for H = 0.95.

We also compare the equivalent bandwidth expressions derived in this paper with the equivalent bandwidth expressions derived by Kelly [12] and by Stahis and Maglaris [30]. To compare these three approaches, we use Equation 3.35 in [12] and Equation 6 in [30]. We utilize real source traffic parameters as described in [30]. In Figure 15 we display the number of accepted connections as a function of the overflow probability considering sources with 1.4 Mbits mean rate σ = 0.28 Mbits and H = 0.85. The buffer size was set to 1000 ATM cells. We can see that the number of admitted sources predicted by our work is in fair agreement with the predicted number by Kelly's work. Moreover, we reached the same result predicted by the large deviation theory with much less computational effort.

In Figure 16 we display the accuracy of the overflow probability computed by Equations 10/11 as a function of the buffer size. In Figure 16.a we consider five different sources with mean and variance given in Table 1. All sources have the same Hurst parameter (H = 0.86). In Figure 16.b we consider an aggregate of sources with diverse traffic parameters (Table 2). As it can be seen, Equations 10/11 predict overflow probability which is in fair agreement with simulation outcomes. The larger the buffer the more accurate are our results. For small buffer sizes the difference between the overflow probability computed via our analytical model and via simulation is less than an order of magnitude.

We can use Equations 10/11 to derive admissible regions for scenarios with heterogeneous sources. In Figure 17 we illustrate the admission region for two classes of sources for overflow probability of 10^{-6} and different buffer sizes. Note that as the

variance increases we considerably decrease the number of accepted sources. Moreover, the impact of the variance on the number of admitted sources is stronger for smaller buffer sizes.

VI) Conclusions

In this paper, we introduced the fractal leaky bucket a novel policing mechanism target at monitoring self-similar sources. We showed that the LB is not able to monitor these types of sources. We compare our new mechanism to the LB and show that the FLB not only accurately police self-similar sources but also it approximates reasonably an ideal mechanism. We also showed that the FLB is insensitive to the duration of sampling time window.

Moreover, we propose a simple way to derive the equivalent bandwidth of a aggregate of heterogeneous self-similar sources. The concept of the Maximum time scale (MaxTS) is of paramount importance to our works. Based on the MaxTS, we compute both performance metrics with little computational effort and with the same accuracy of results predicted by the large deviation theory. In addition, we have also reduced the sensitivity of the estimation process by using bounds rather than attempting to directly estimate the parameters from the full trace.

VII) References

[1] W. Leland, M. Taqqu, W. Willinger and D. Wilson, “On the Self-Similar Nature of Ethernet Traffic (Extended Version)”, *IEEE/ACM Transaction on Networking*, vol 2, no 1, pp. 1-15, February 1994.

[2] M. Garrett and W. Willinger , “Analysis Modeling and Generation of Self-Similar VBR Video Traffic”. In *Proc. of ACM SIGCOMM*, 1994.

[3] J.Beran et al., “Long-Range Dependence in Variable-Bit-Rate Video Traffic”. *IEEE Transactions on Communications*, 1995.

[4] V. Paxson and S. Floyd, “Wide-Area Traffic: The Failure of Poisson Modeling”, In *Proc. of ACM SIGCOMM*, 1994.

[5] G. Mayor and J.Silvester, “A Trace-Driven Simulation of an ATM Queueing System with Real Network Traffic”, *Proc. of IEEE ICCCN*, 1996.

[6] G. Mayor and J. Silvester, “An ATM Queueing System with a Fractional Brownian Noise Arrival Process”, *Proc. of IEEE ICC*, pp. 1607-1611, 1996.

[7] I. Norros, “A Storage Model with Self-Similar Input”, Queueing Systems 16, 1994.

[8] N. Likhanov and B. Tsybakov, “Analysis of an ATM Buffer with Self-similar (“Fractal”) Input Traffic, in *Proc of SIGCOMM*, 1995.

[9] N. Duffield, J. T. Lewis, N. O’Connell, R. Russel, F. Toomey, “Predicting Quality of Service with Long-range Fluctuations”, In *Proc of IEEE ICC’95*, pp. 473-477, 1995.

[10] H. D. Sheng and S. Q. Li,``Spectral Analysis of Packet Loss Rate at a Statistical Multiplexer for Multimedia Services'', *IEEE/ACM Transactions on Networking*, January 1994.

[11] A. I. Elwalid and D. Mitra, “Effective Bandwidth of general Markovian Traffic sources and Admission Control of High Speed Networks”, *IEEE/ACM Trans on Networking*, 1(4), pp 329-343, 1993

[12] F. Kelly, “Notes on Effective Bandwidth”, in Stochastic Networks: Theory and Applications, F. Kelly, S. Zachary and I. Ziednis ed., Oxford Press, 1996

[13] G de Veciana and J. Walrand, “Effective Bandwidth: Call Admission, Traffic Policing and Filtering for ATM Networks, 1994.

[14] N. G. Duffield and N. O’Connell, Large Deviatiion and Overflow probabilies for the general single-server with application, Tech Rep DIAS-APG-93-30, Dublin Institute for Advanced Studies, 1993.

[15] G. Kesidis, J. Walrand and C.S. Chang, Effective bandwidth for Multiple Class markov Fluid and other ATM Sources, *IEEE/ACM Transaction on Networking*, Aug, 1993

[16] E. Rathgeb, “Modeling and Performance Comparison of Policing Mechanisms for ATM Nets”, *IEEE JSAC*, Apr,1991.

[17] K.Sohraby and M.Sidi, ``On the Performance of Bursty Modulated Sources Subject to Leaky Bucket Rate-Based Access Control Schemes”, *IEEE Transaction on Communications*, Feb/Mar/Apr, 1994.

[18] J.A. Silvester, N.L.S. Fonseca, G. S. Mayor and S. P. S. Sobral, “The Effecitiveness of Multi-level Policing Mechenisms in ATM Traffic Control, In *Proc. of IEEE International Telecommunications Symposium 96*, pg 98-102, 1996.

[19] G. S. Mayor and J.A. Silvester, “Time Scale Analysis of an ATM Queueing System with Long-range Dependent Traffic”, in *Proc of Infocom’97*, pp 205-212, 1997

[20] G.S. Mayor and J. A. Silvester, “Providing QoS for Long-Range Dependent Traffic”, the *7th IEEE Computer-aided Modeling Analysis and Design of Communications Links and Networks*, pp 19-28, 1998.

[21] N.L.S. Fonseca e M.J. Ferreira, “On the Effectiveness of the Longest-Packet-In Packet Discard Policy”, in *Proc of IEEE GLOBECOM98*, pp. 1747-1753, 1998.

[22] N.L.S. Fonseca e M.J. Ferreira, “Multiple Class Selective Discard Under a Long-Range Dependent Process”, *Proc of IEEE International Telecommunications Symposium 98*, pag 95-101, 1998

[23] N.L.S. Fonseca e M.J. Ferreira, “Maximizing the Cell Goodput During Overload in ATM Multiplexers”, in *Proc of IEEE International Telecommunications Symposium 98*, pag 189-194, 1998.

[24] A. Papoulis, Probability, Random Variables, and Stochastic Processes, McGraw-Hill, 1991

[25] M. Chi, E. Neal and G. Young, “Practical Application of Fractional Brownian Motion and Noise to Synthetic Hydrology” , *Water Resources Research*, Vol. 9, 1523-1533, December 1973.

[26] B.B. Mandelbrot, “Long-run Linearity, Locally Gaussian Process, h-spectra and Infinitive Variance, International Economic Review, 10:82-113, 1969.

[27] I. Norros, “The management of large Flows of Connectionless traffic on the basis of Self-similar, in *Proc of IEEE ICC*, 1995

[28] R. G. Addie, M. Zukerman and T. Neame, “Fractal Traffic: Measurements, Modelling and Performance Evaluation", In *Proc. of IEEE INFOCOM* 1995.

[29] B. Ryu and A. Elwalid,``The Importance of Long-Range Dependence of VBR Video Traffic in ATM Traffic Engineering: Myths and Realities", *In Proc. of ACM SIGCOMM*, 1996.

[30] C.Stathis and B. Maglaris,"Modeling the self-similar behaviour of network traffic" , IFIP 6th Workshop on Performance Modeling and Evaluation of ATM Networks, 1998.

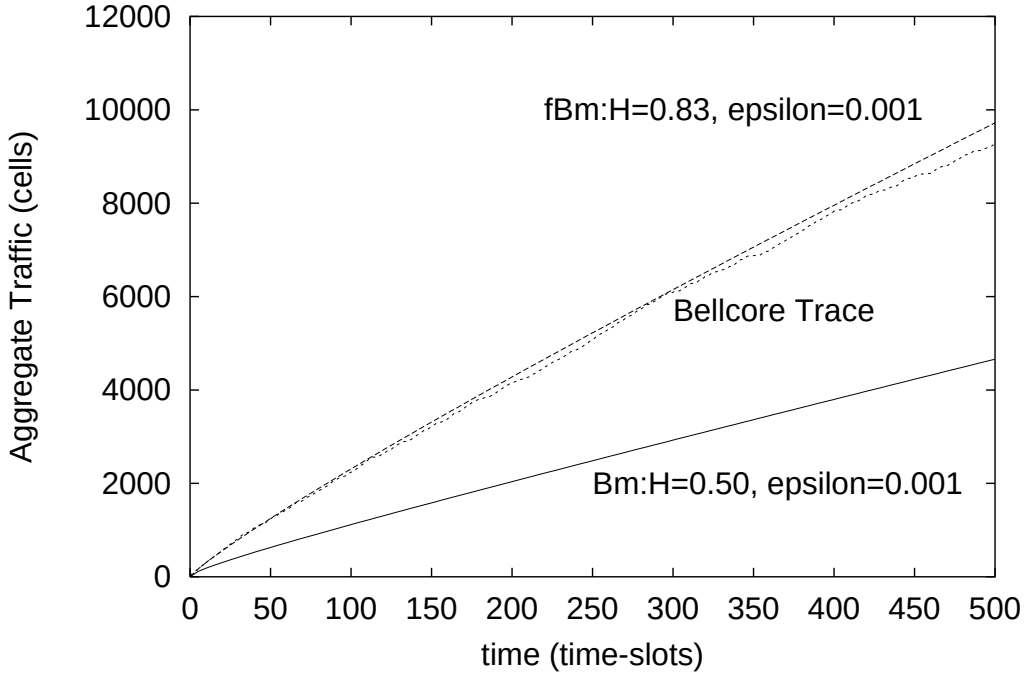

Figure 1: The Accuracy of the fBm Envelope Process Y (τ) (middle curve) and fBm Envelope process *H* = 0.5 (lower curve) and *H* = 0.83 (upper curve)

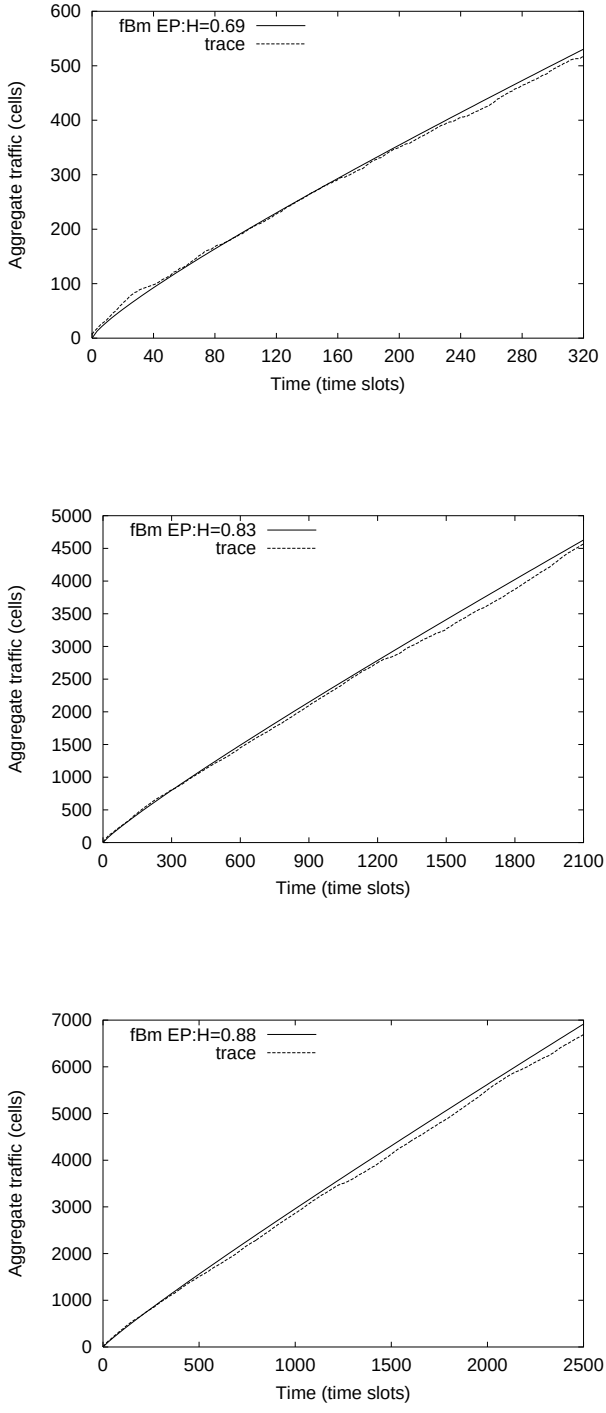

Figure 2: Accuracy of the fBm Envelope Process for Different Values of the Hurst Parameter

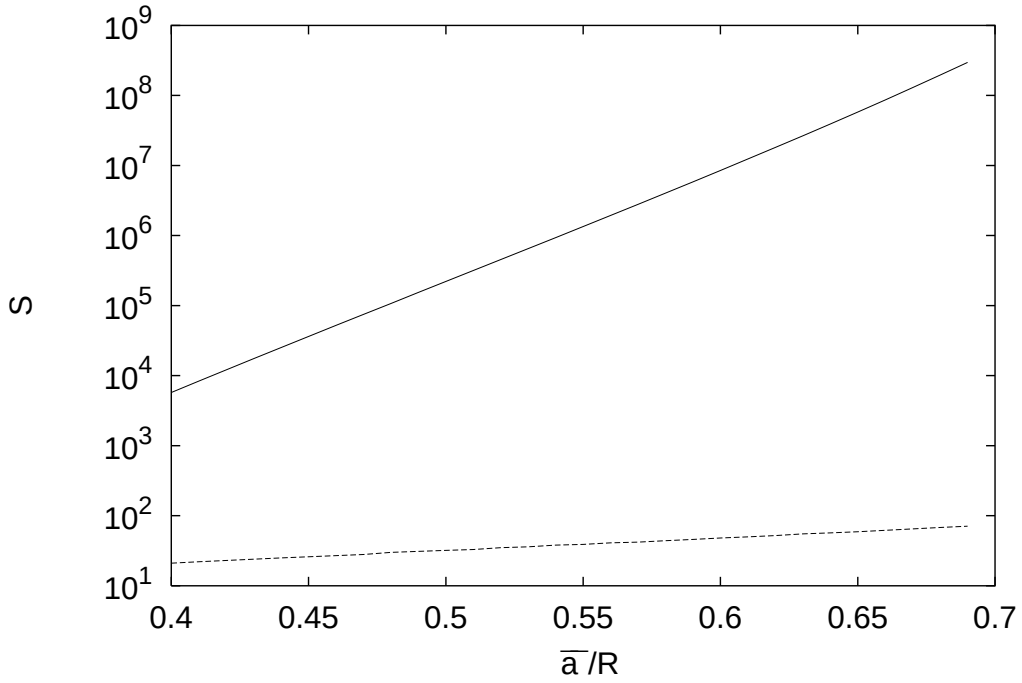

Figure 3: The Bucket Size x the Ratio Between the Mean Arrival Rate and the Leaky Rate

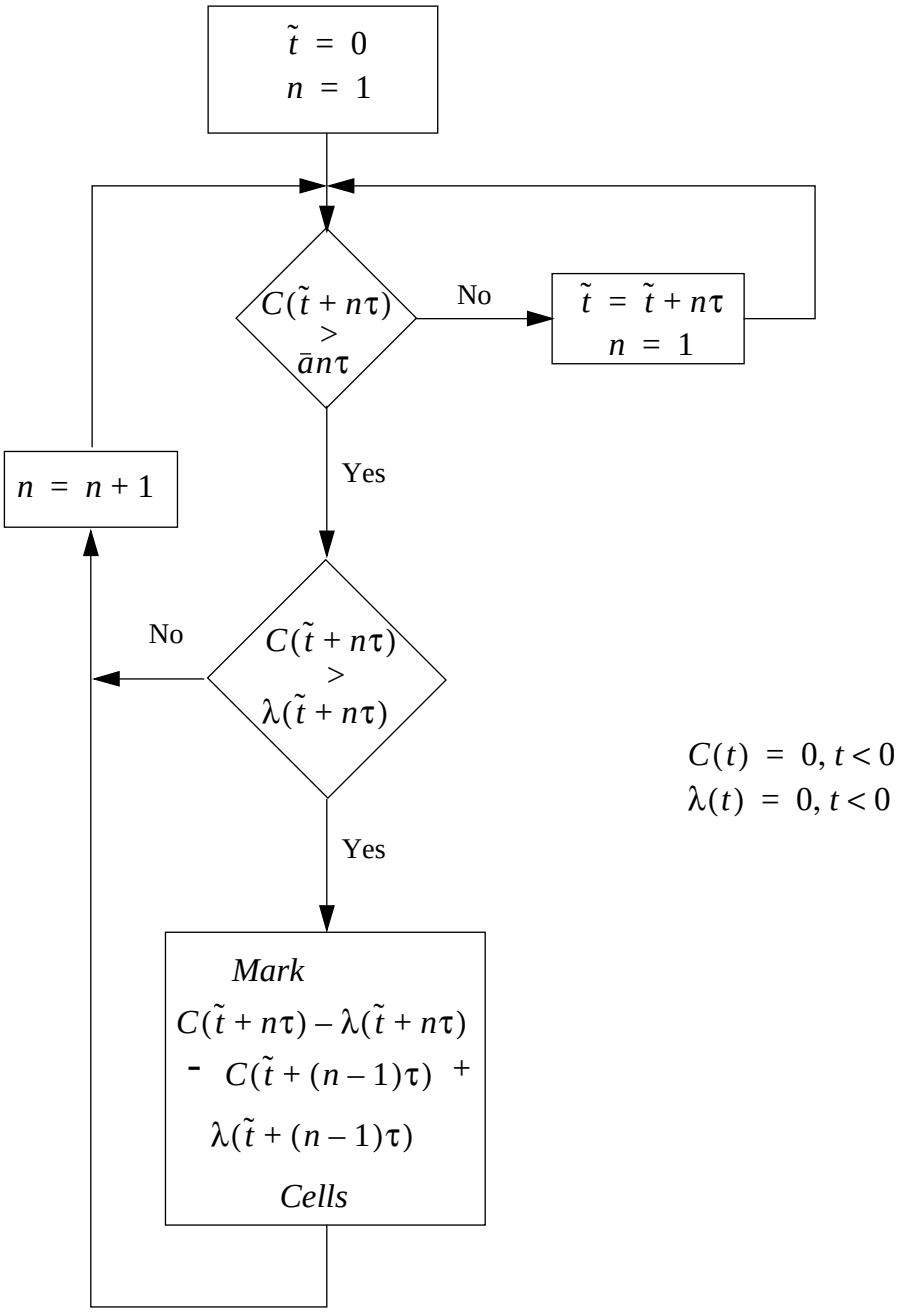

Figure 4: the Fractal Leaky Bucket

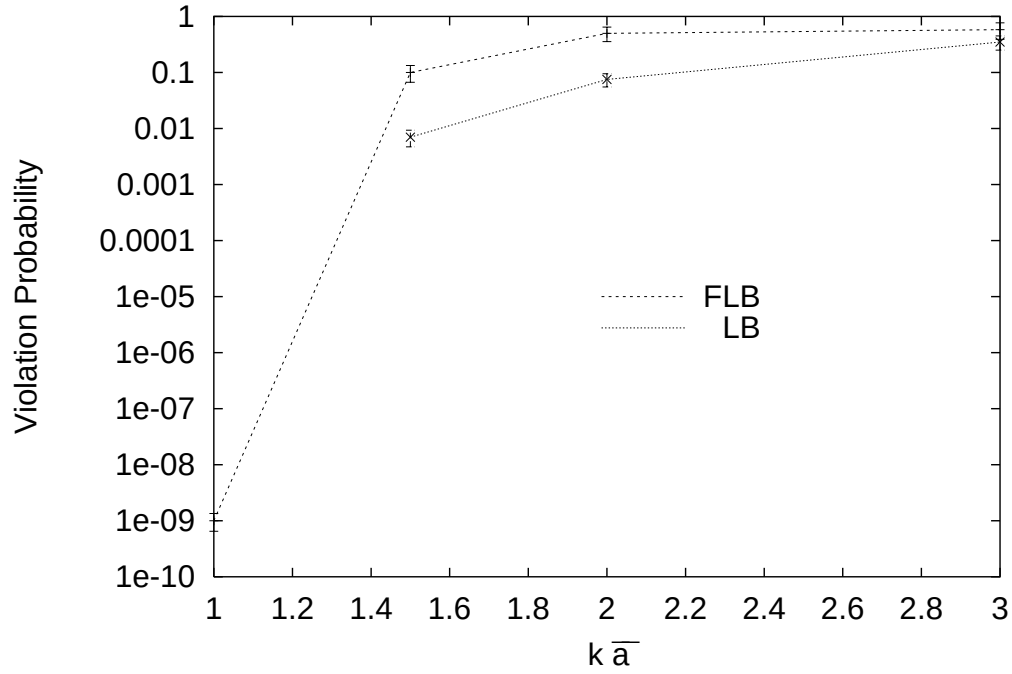

Figure 5: A Comparison Between the FLB and the LB for Violating Sources based on a Sequence of the STAR WARS Video Stream

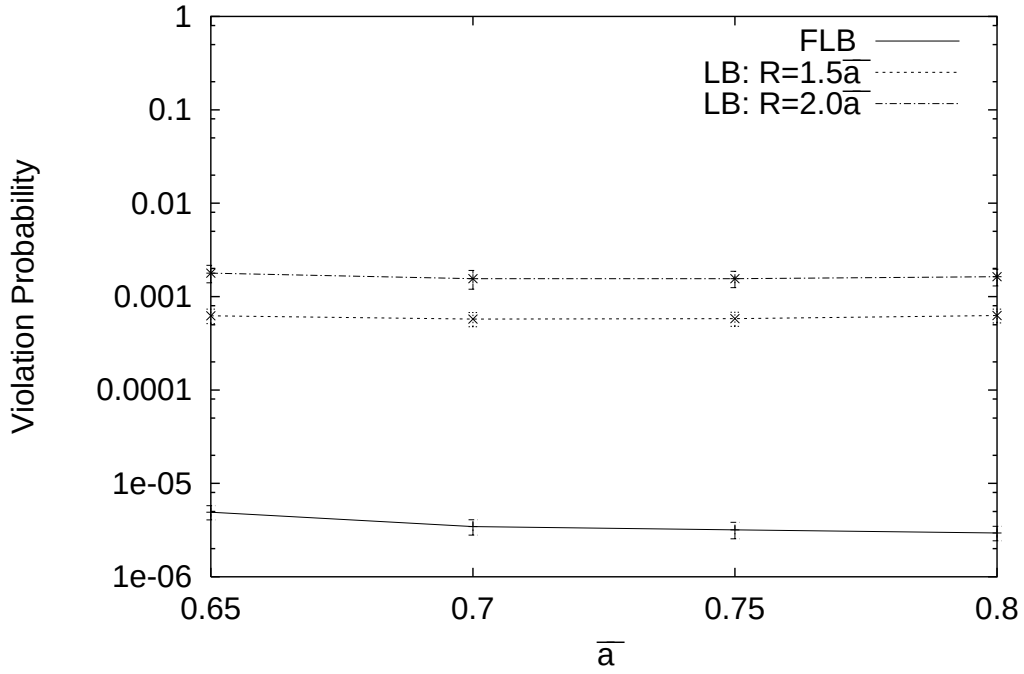

Figure 6: A Comparison Between the Violation probability given by the FLB and the Violation Probability given by the LB for Different Leaky Rates for Well-behaved Sources

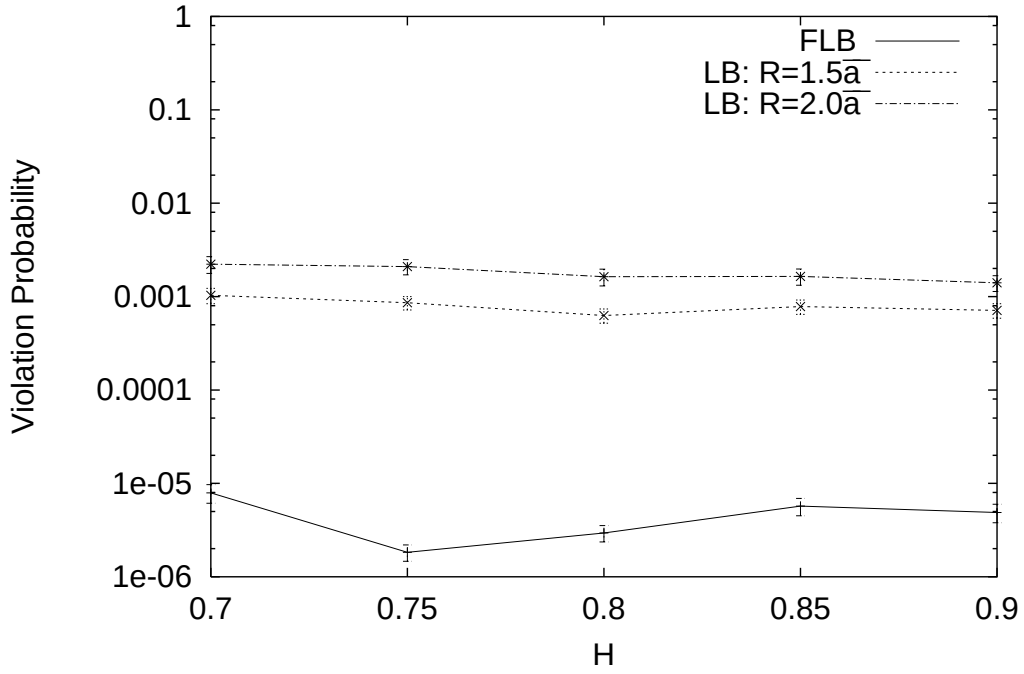

Figure 7: A Comparison Between the Violation Probability given by the FLB and the Violation Probability given by the LB for Well-behaved Sources with Different Hurst Parameters

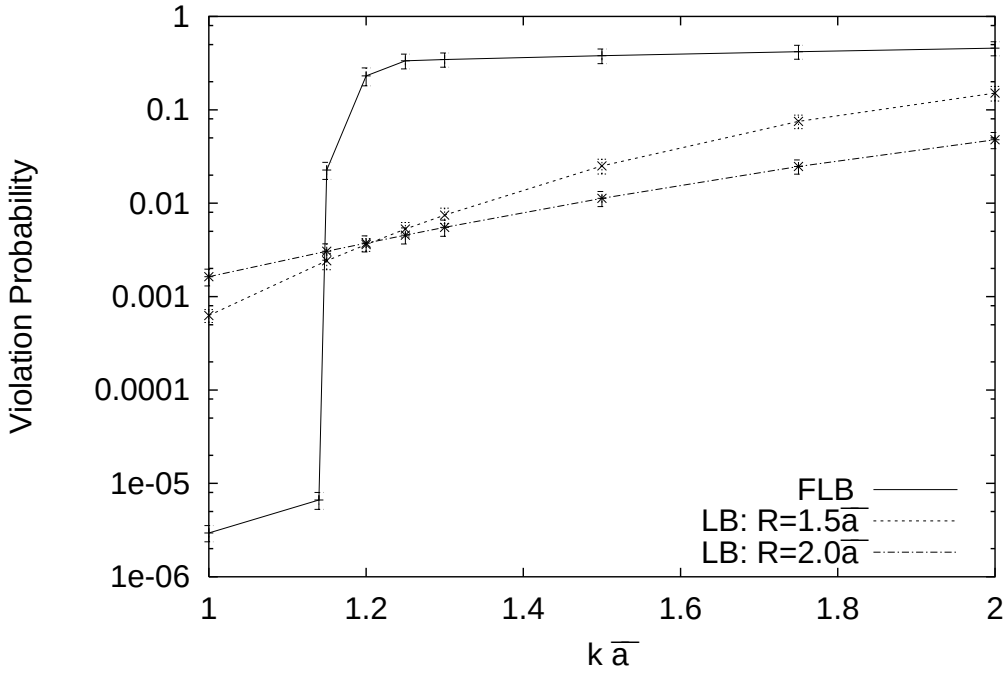

Figure 8: Violation Probability x Violating Mean Arrival Rate

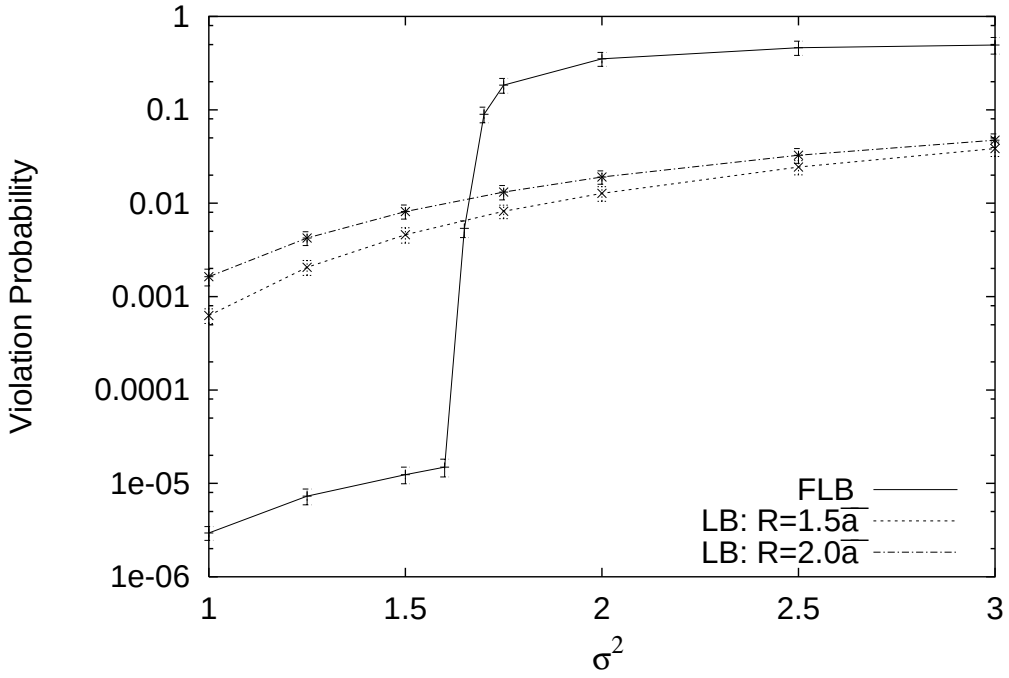

Figure 9: Violation Probability x Violating Variance

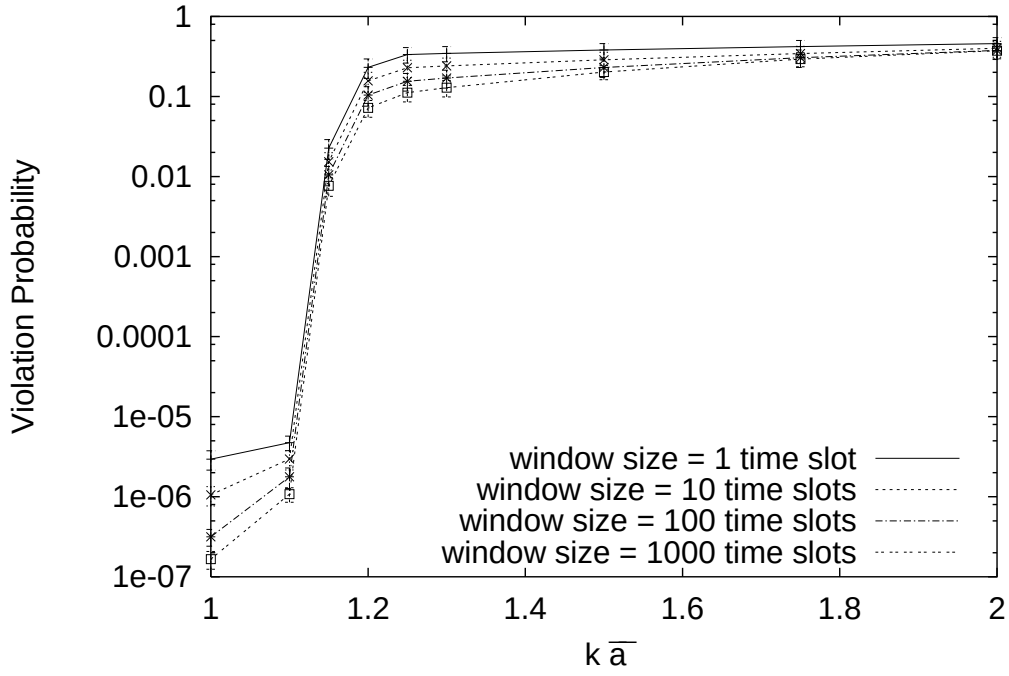

Figure 10: Sensitivity of the FLB Violation Probability to the Sampling Time Window Duration x Violating Mean Arrival Rate

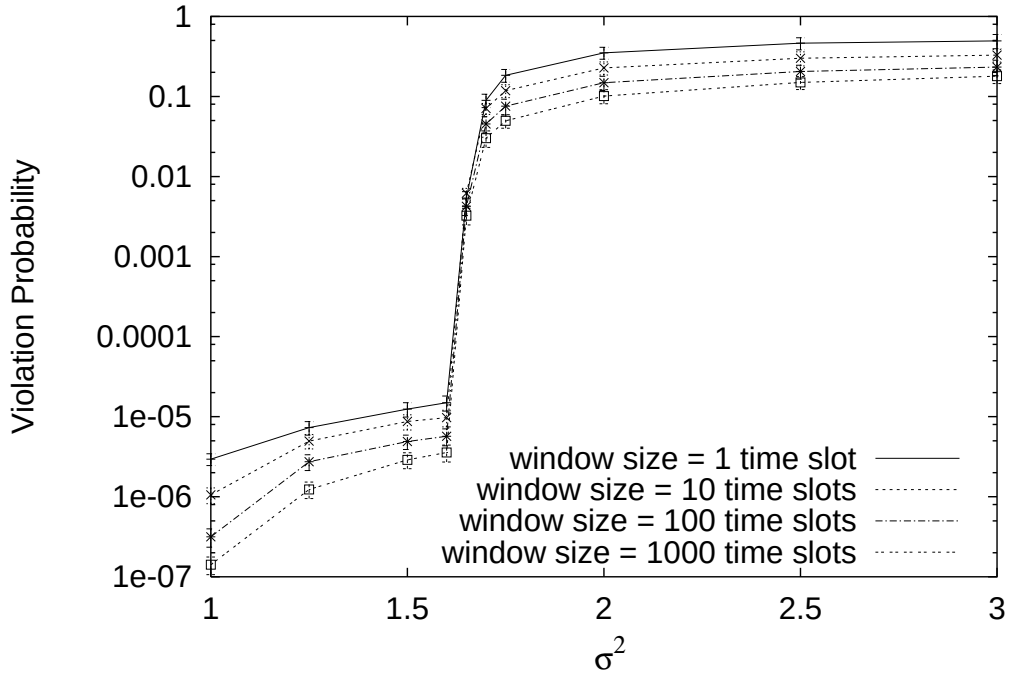

Figure 11: Sensitivity of the FLB Violation Probability to the Sampling Time Window Duration x Violating Variance

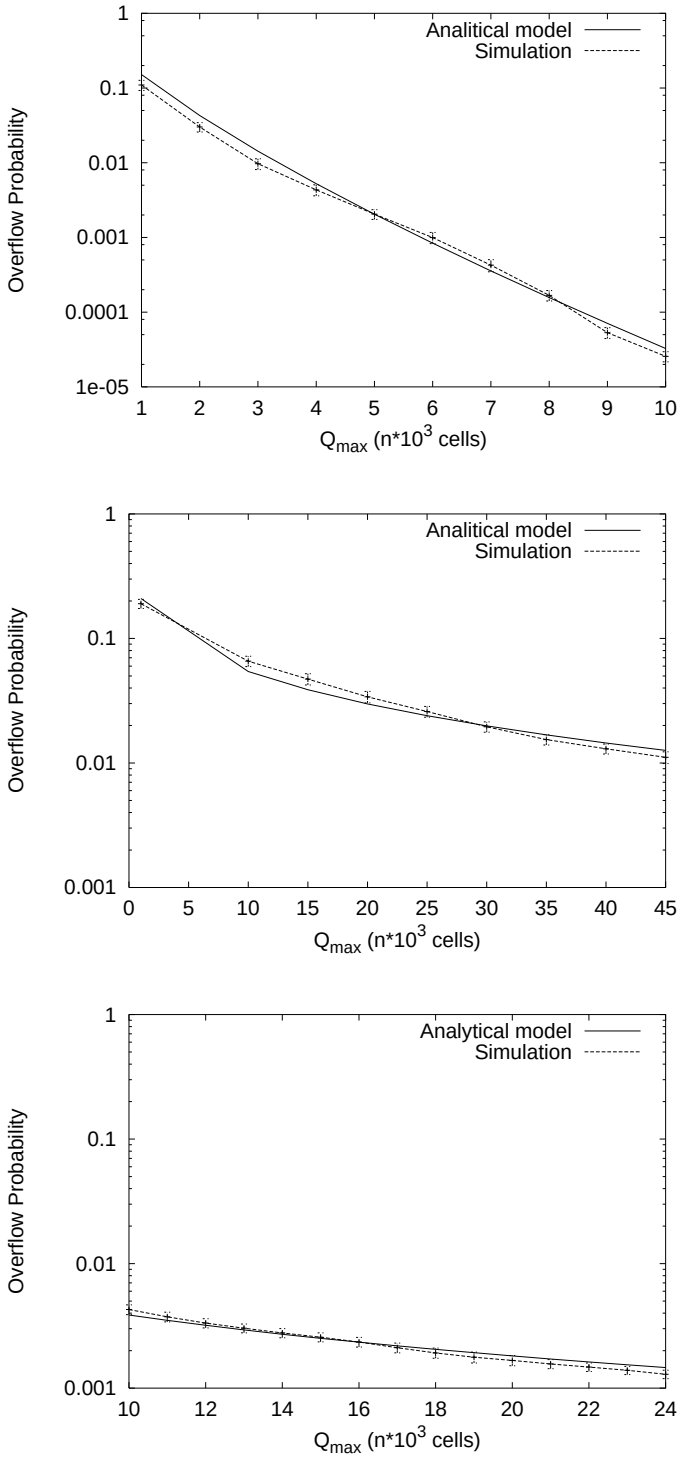

Figure 12: Accuracy of the Predicted Overflow Probability by Equations 10/11 for a Single Source as a Function of the Buffer Size

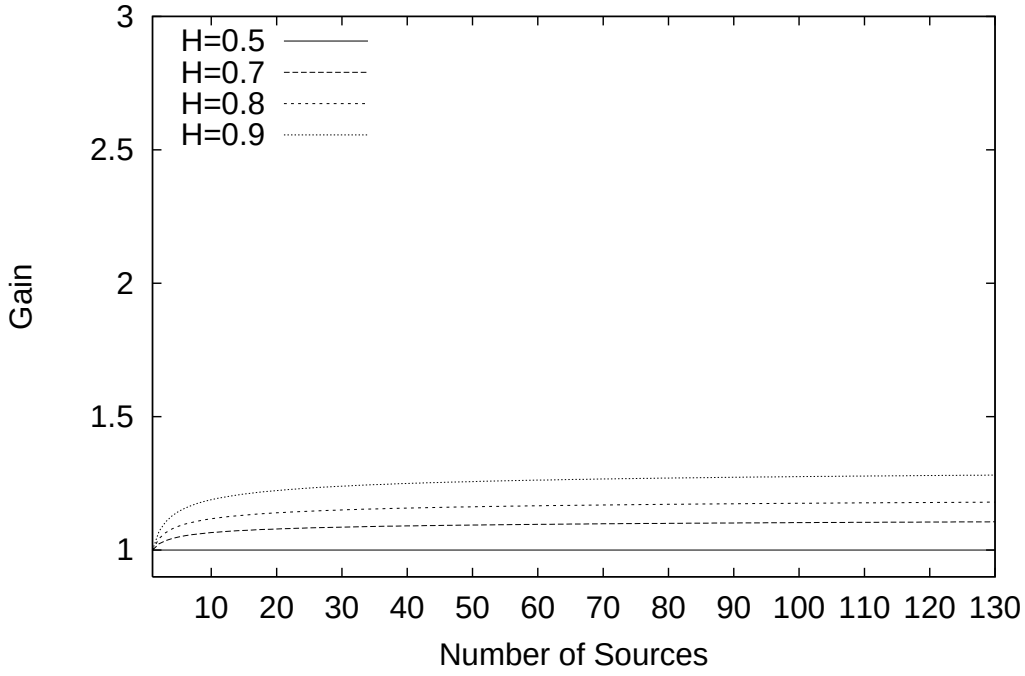

Figure 13.a: Streams with Low Variance

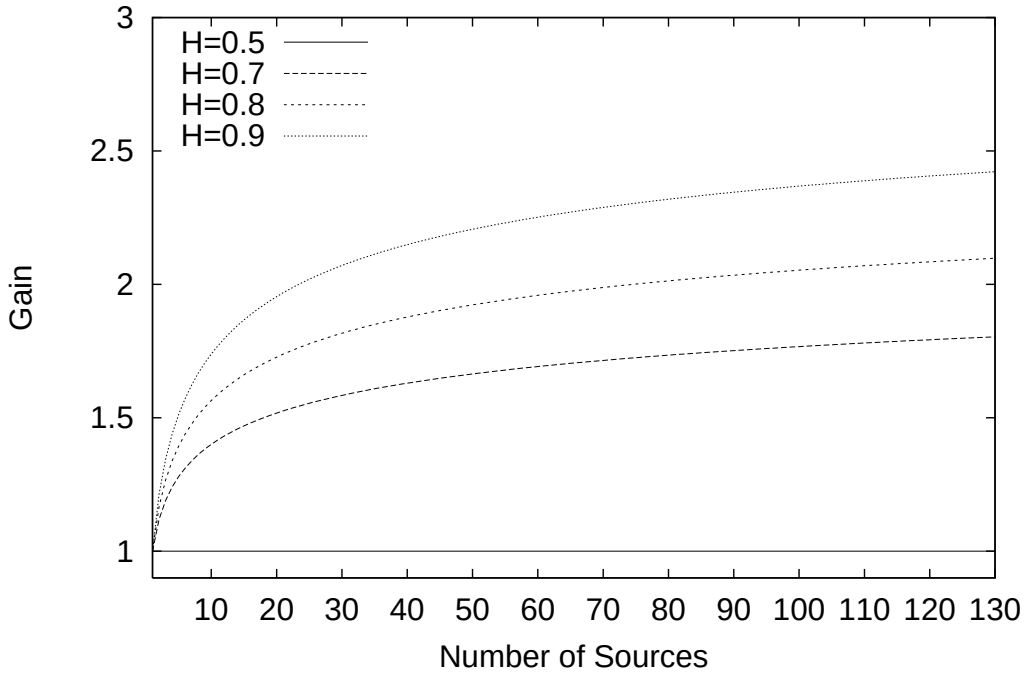

Figure 13.b: Stream with Moderate to High Variance
Figure 13: Multiplexing Gain for Streams with Different Hurst Parameter

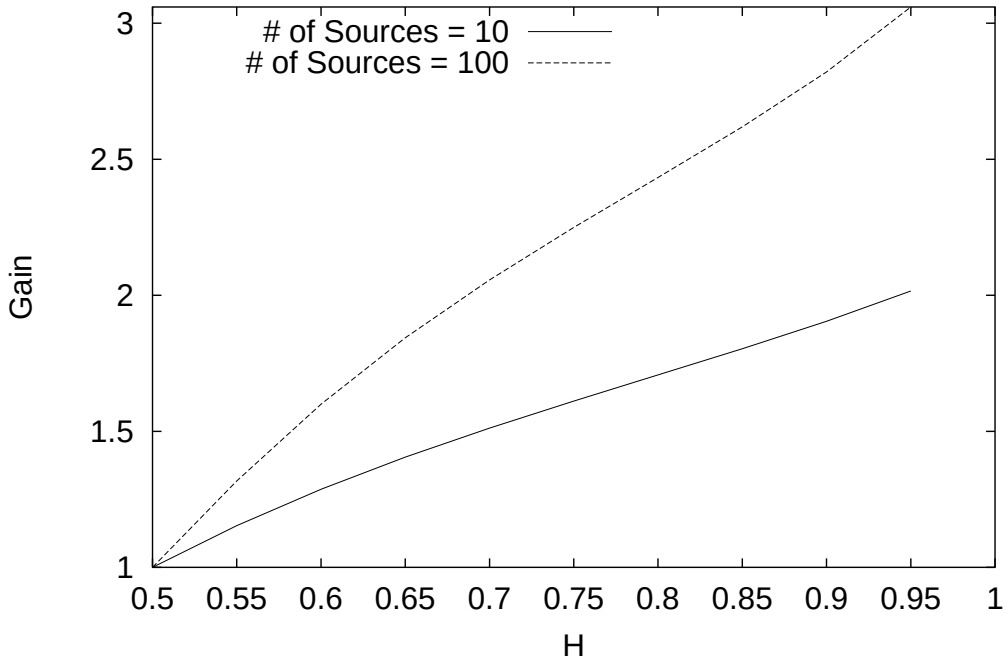

Figure 14: Multiplexing Gain as a Function of the Hurst Parameter
for Different Number of Sources

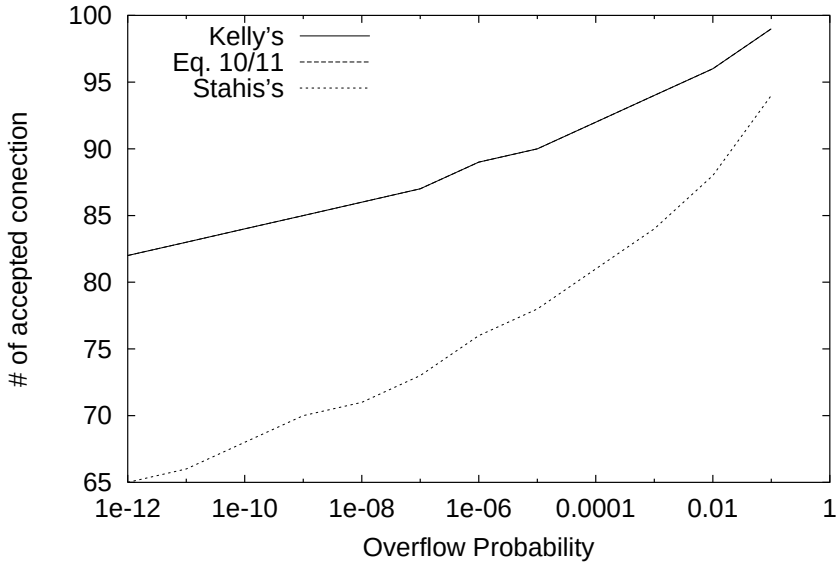

Figure 15: Comparison of the Number of Admitted Sources given Equations 10/11 and the work in [12] and [30] for Different Overflow Probability Requirements

Table 1: Traffic Parameters for Figure 16.a

Source	$\bar{a}$	σ	H
A	0.11	0.08	0.86
B	0.11	0.07	0.86
C	0.13	0.07	0.86
D	0.15	0.10	0.86
E	0.10	0.06	0.86

Table 2: Traffic Parameters for Figure 16.a

Sources	$\bar{a}$	σ	H
A	0.13	0.10	0.63
B	0.11	0.07	0.72
C	0.13	0.07	0.78
D	0.12	0.07	0.86
E	0.11	0.04	0.90

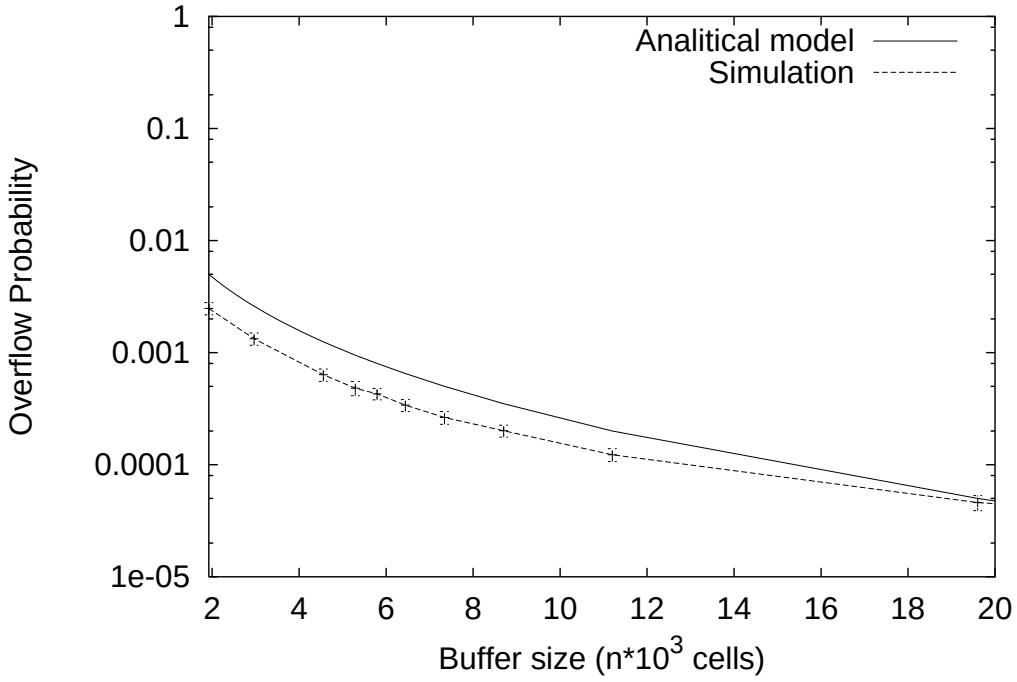

Figure 16a: Heterogeneous Sources with Same Hurst Parameter

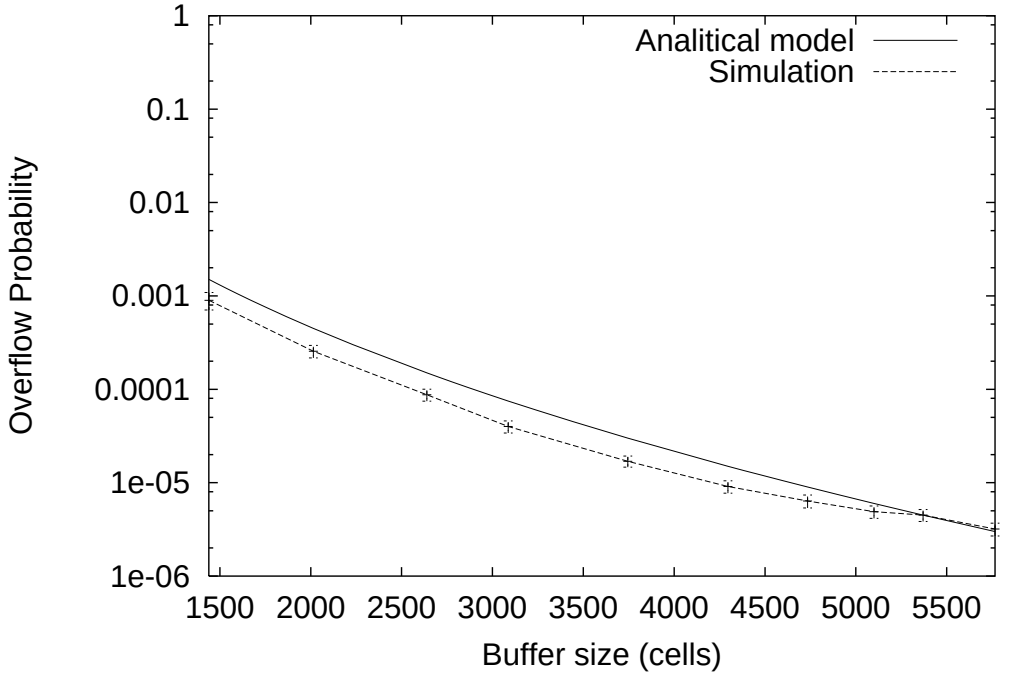

Figure 16.b: Heterogeneous Sources with Different Hurst Parameter
Figure 16: Overflow probability x Buffer Size for Heterogeneos Sources

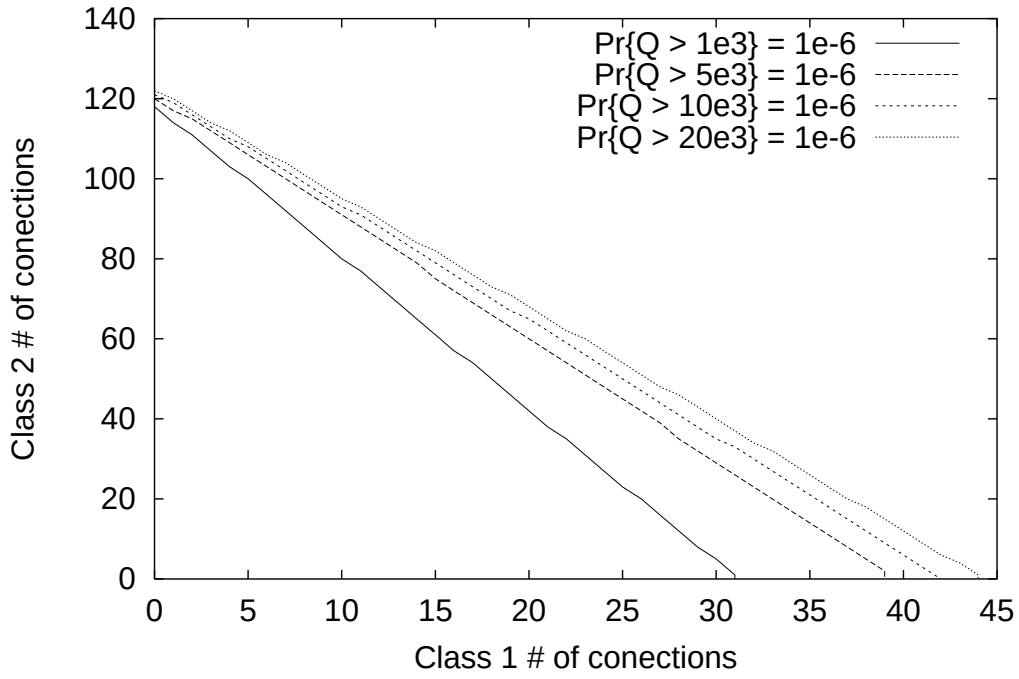

Figure 17.a: Streams with Low variance

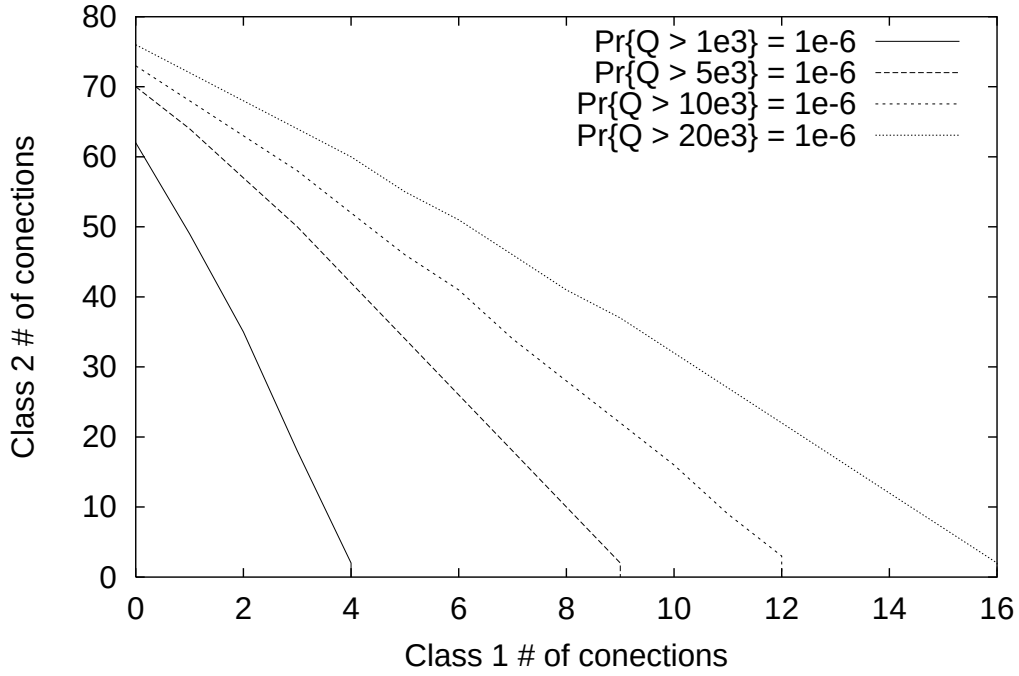

Figure 17.b: Streamd with High Variance
Figure 17: Admission Regions for 2 Different Classes of Sources