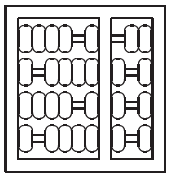


INSTITUTO DE COMPUTAÇÃO
UNIVERSIDADE ESTADUAL DE CAMPINAS

Statistical Multiplexing of Multifractal Flows

Cesar Augusto Viana Melo
Nelson Luis Saldanha da Fonseca

Technical Report - IC-03-22 - Relatório Técnico

November - 2003 - Novembro

Statistical Multiplexing of Multifractal Flows

Cesar A. V. Melo and Nelson L. S. da Fonseca

November 12, 2003

Abstract

This paper introduces the computation of an expression for the time at which the length of a queue fed by several multifractal flows reaches its maximum. Expressions for the equivalent bandwidth of an aggregate of multifractal flows is also presented. Moreover, it is shown that modelling based on monofractal process rather than based on multifractal processes leads to overprovisioning of resources.

1 Introduction

Since the seminal work of Leland et al [1], several studies have shown that network traffic presents scale invariance, or "scaling", which is the absence of any specific time scale at which the "burstiness" of a traffic stream can be characterized. Instead, it is necessary to describe the traffic across different time scales. Self-similar or (mono) fractal processes have been used for modeling network traffic since then.

Scaling of fractal traffic is defined by a single constant value: the Hurst parameter, H. One of the most popular fractal processes for traffic modeling is the Fractal Brownian Motion process (fBm) due to its parsimonious representation of the modeled traffic. fBm is an accurate model when: i) the traffic results from the aggregation of several sources streams with low activity compared to the link bandwidth, ii) the impact of flow control is not relevant and iii) the time scale of interest is within the scaling region. The multifractal Brownian motion (mBm) is the multifractal generalization of the fractal Brownian motion. mBm has the nice property that at small time scales (locally) its realization can be described by an fBm.

Both Internet Protocol (IP) and Variable Bit Rate (VBR) video traffic present non-trivial scaling structure at small scales in addition to long memory [2][3]. At small scale, traffic is highly variable, more complex and follows less definitive scaling laws. For these traffics the marginal distribution of counts is non-Gaussian, calling for a representation beyond second-order statistics. Moreover, the scaling exponent of the variance on time scale shorter than a typical (cut-off) one is smaller than an asymptotic exponent.

If on one hand, at the network core long term correlations are more important than the variability at small scales due to traffic aggregation (additive property) [4]. On the other hand, at the network edge, where admission control is performed, variability at small time scale (multiplicative property) plays a major role [5]. These patterns can be modelled by multifractal processes which capture both long memory and high variability at small scales.

In networks employing statistical multiplexing, traffic streams are merged at the multiplexers and transferred to the outgoing link. Solving queueing systems with statistical multiplexing under (multi/mono) fractal input is of paramount importance for admission control. Nonetheless, this is not a trivial task. Large Deviation theory can be employed to overcome such difficulty. However, solutions based on this theory imply in making non-realistic assumption about the buffer size.

An envelope process is an upper bound for the accumulated amount of work (traffic) arrived from a process up to a certain time. Envelope processes are parsimonious representations of stochastic processes and allow simple solutions of queueing systems fed by (mono/multi) fractal processes which do not incorporate any unrealistic assumption about the buffer size.

In [6], an envelope for multifractal traffic modeling was introduced. This envelope process was extensively validated using both synthetic and real network traces. The envelope process is an upper bound for the accumulated amount of traffic arrived up to a certain time from a multifractal Brownian motion process (mBm)[7]. It has been shown that although mBm is a steady state Gaussian process, the envelope process is a tight bound for the amount of traffic arrived in Internet streams.

The major contribution of this paper is a method to compute the time instant at which a queue fed by several multifractal flows reaches its maximum. This computation can be used to calculate the loss probability as well as to determine the equivalent bandwidth of an aggregation of several multifractal flows. Such computation can be employed in admission control policies at the ingress of network domains, such as DiffServ domains. Moreover, the expressions derived here can be used in measurement based frameworks.

This paper is organized as follows. In Section II, the definition of the multifractal Brownian motion is given, and in Section III an envelope process based on mBm is presented. In Section IV, the method to compute the time scale at which overflow occurs in queueing system fed by a single stream is shown whereas in Section V the generalization of the method for multiple streams is introduced. In Section VI, numerical examples are provided. Finally, in Section VII conclusions are drawn.

2 The Multifractional Brownian Motion Process

Multifractal processes exhibits highly irregular patterns as a function of time. Local Holder exponents describes the local regularity of the sample path of a process. It is a measure of scaling and can be regarded as a generalization of the Hurst parameter [4].

The local Holder regularity is related to scaling at small time scales since it expresses the regularity of the sample path of a process by comparing it to a power-law function[4]. The exponent of this power law, $h(t)$, is called Holder exponent and depends both on time and on the sample path. The Holder exponent is the largest value of h, $0 \leq h \leq 1$, such that

$$|X(t+\gamma) - X(t)| \leq k|\gamma|^h \quad \text{for } \gamma \to 0 \tag{1}$$

For monofractal processes the Holder function (Hurst parameter) is a constant value

whereas for multifractal processes the Holder function changes randomly with time. Let $H : (0,\infty) \rightarrow (0,1)$ be a Holder function. The multifractional Brownian motion is a continuous Gaussian process with non-stationary increments defined on $(0,\infty)$ as:

$$\begin{aligned} W_{H(t)} &= \frac{1}{\Gamma(H(t)+1/2)} \\ &\left\{ \int_{-\infty}^{0} [(t-s)^{H(t)-1/2} - (-s)^{H(t)-1/2}] dB(s) + \right. \\ &\left. \int_{0}^{t} (t-s)^{H(t)-1/2} dB(s) \right\} \end{aligned} \tag{2}$$

where $B(s)$ is the Brownian motion.

The multifractal brownian motion process is a generalization of the fractal brownian motion process and exhibits the property that locally it is asymptotically self-similar (lass), i.e.

$$\lim_{\rho \rightarrow 0^+} \left\{ \frac{W(t+\rho u) - W(t)}{\rho^{H(t)}} \right\}_{u \in R^+} = \{B_{H(t)}(u)\}_{u \in R^+} \tag{3}$$

where $W(.)$ is an mBm and $B_{H(t)}(u)$ is an fBm process with Hurst parameter H, given by $H(t)$.

3 An Envelope Process for the Multifractal Brownian Motion Process

To solve a queueing system fed by an input process, it is necessary to know both the amount of work arrived to the system as well as the service rate. Envelope processes are upper bounds for the amount of arrivals, and allow less complex solutions than the ones that consider the accumulated work arrived by an exact process. Envelope processes can be either deterministic or probabilistic. In deterministic envelopes, the amount of work arrived never surpasses the envelope value whereas in probabilistic envelopes it may surpass with a certain pre-defined probability. Probabilistic envelope processes are tighter bounds than deterministic envelopes. Dimensioning based on deterministic envelope processes may lead to waste of resources, since the provision of resource needs to take into account the maximum amount of work arrived at any time. When probabilistic envelopes are used, there is no need to consider spikes of work up to a certain amount defined by the probability of violation. However, loss of packets may occur.

An upper bound for the accumulated amount of work arrived can be computed as the mean amount of work plus an upper bound for the accumulated increments. An upper bound for mBm increments can be computed by using the upper bounds for the local fBm increments, since in the neighborhood of time t, an mBm can be approximated by an fBm with Hurst parameter $H(t)$. It is known that [8]:

$$Z_H(m) \leq \kappa H t^{H-1} \tag{4}$$

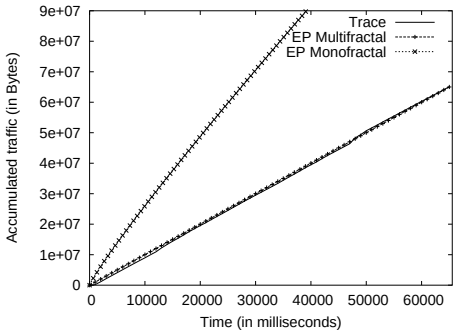


Figure 1: The EP monofractal and EP multifractal for the trace MEM-1053844177

As the size of local infinitesimal neighborhood of t goes to zero, the envelope process, $\hat{A}(t)$, of an mBm with mean $\bar{a}$, standard deviation σ and Holder function $H(.)$ can be expressed as :

$$\hat{A}(t) = \int_0^t \bar{a} + \kappa\sigma H(x) x^{H(x)-1} dx \tag{5}$$

which is called mBm envelope process.

This envelope reduces to the fBm envelope previously derived in [8] when $H(.)$ is a constant value, i.e.,

$$\hat{A}(t) = at + \kappa\sigma t^H \tag{6}$$

Extensive simulation experiments using both synthetic traffic and real network traffic were conducted in order to assess the accuracy of the proposed envelope. It was shown that the mBm envelope process is an accurate model [6]. Figures 1 shows the accumulated traffic arrived from a real network trace, the mBm envelope process and the monofractal fBm envelope process. The trace used was collected at the University of Memphis on May 24, 2003 (please, see Table 2). The monofractal envelope process for this trace is defined by the following parameters $\bar{a} = 1788.9$, $\sigma^2 = 8546650.0$ and $H = 0.94$. It can be observed that the multifractal envelope process is a tight bound for the multifractal stream whereas the monofractal envelope overestimates the amount of traffic arrived. Thus, dimensioning networks with multifractal streams based on monofractal models leads to overprovisioning of resources.

4 Time Scale of Interest of a Queue Fed by a Single Stream

In this section, the time at which a queue reaches its maximum occupancy in a probabilistic sense is derived. The queue size at this time provides a simple delay bound. Consider a continuous-time queueing system, with deterministic service given by C. The cumulative arrival process is represented by $A(t)$ (for $A(0) = 0$). Let $\hat{A}(t)$, a continuous and differentiable function, be the probabilistic envelope process of $A(t)$, such that $P(A(t) > \hat{A}(t)) \leq \epsilon$.

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, $q(t) = A(t) - ct \geq 0$.

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

$$\hat{q}(t) = \hat{A}(t) - Ct \geq 0, \tag{7}$$

we can see that $P(q(t) > \hat{q}(t)) = P(A(t) > \hat{A}(t)) \leq \epsilon$.

The maximum delay in a FIFO queue is given by the maximum number of cells in the queue during the busy period, which can be defined as

$$\hat{q}_{max} = max(\hat{q}(t)) \qquad t \geq 0 \tag{8}$$

Therefore,

$$P(q(t) > \hat{q}_{max}) \leq P(q(t) > \hat{q}(t)) \leq \epsilon \tag{9}$$

$$P(q(t) > \hat{q}_{max}) \approx \epsilon. \tag{10}$$

The queue length at time t, $q(t)$, will only exceed the maximum queue length $\hat{q}_{max}$ 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 the estimated value. Intuitively, by bounding the behavior of the arrival process, it is possible to transform the problem of obtaining a probabilistic bound for the stochastic system defined by $q(t) = A(t) - Ct$, into an easier problem of finding the maximum of a deterministic system, described by $\hat{q}(t) = \hat{A}(t) - Ct$.

Inserting the mBm envelope process into Equation (7) gives:

$$\begin{aligned} \hat{q}(t) &= \hat{A}(t) - Ct \\ &= \int_0^t \bar{a} + \kappa\sigma H(x) x^{H(x)-1} dx - Ct \end{aligned} \tag{11}$$

In order to compute $\hat{q}_{max}$ it is necessary to find $t^\star$ such that

$$\frac{d\hat{q}(t)}{dt} = 0 \tag{12}$$

or equivalently,

$$\frac{d\hat{A}(t)}{dt} = C \tag{13}$$

The time-scale of interest, $t^\star$, is the time at which the queue size reaches its peak, called the Maximum Time-Scale (MaxTS) and $t^\star$ defines the point in time at which the unfinished work in the queue achieves its maximum in a probabilistic sense. Hence, $t^\star$ can be computed from Equation (13) as:

$$t^\star = \left[\frac{\kappa\sigma H(t^\star)}{(C-\bar{a})}\right]^{\frac{1}{1-H(t^\star)}} \tag{14}$$

Substituting $t^\star$ back into Equation (11) , it can be concluded that:

$$\hat{q}_{max} = \hat{A}(t^\star) - Ct^\star \tag{15}$$

$$\begin{aligned}\hat{q}_{max} &= \int_0^{t^\star} \kappa\sigma H(x)x^{H(x)-1}dx - \\ &\quad (C-\bar{a})^{\frac{H(t^\star)}{1-H(t^\star)}}(\kappa\sigma H(t^\star))^{\frac{1}{1-H(t^\star)}}\end{aligned}$$

5 Time Scale of Interest of a Queue Fed by Several Multifractal Flows

In this section, MaxTS computed in the previous section is used to derive expressions for predicting the equivalent bandwidth and buffer requirements for an aggregate of multifractal flows. Essentially, a method for computing the bandwidth necessary to support the requirements of buffer overflows is proposed, as well as for determining the maximum probabilistic delay for an aggregate of heterogeneous flows. The problem in this section can be stated as follows:

Given a set of flows with mean $\bar{a}_i$, standard deviation σ_i^2 and Holder exponents given by $H_i(t)$, what is the link capacity needed so that the maximum queue size will be bounded by $\hat{q}_{max}^N$ with probability ϵ?

To answer this question the expression of the envelope process resulting from the aggregation of several flows is needed. To compute the amount of traffic aggregate the *local asymptotically self-similar (lass)* property is used. In [9] it was proved that the aggregate of N fBm processes with mean $\bar{a}_i$ and varincia σ_i^2 is an fBm process with mean $\bar{a}$ and σ^2, given by the $\sum_{i=1}^N \bar{a}_i$ and by $\sum_{i=1}^N \sigma_i^2$, respectively. Thus, locally the mBm process can be represented by an fBm resulting from the aggregation of fBm processes. Similarly, the mBm envelope process can be locally approximated by an fBm envelope process which results from the aggregation of the N fBm envelope processes.

Assume N independent flows defined by the following parameters: mean $\bar{a}_i$, variance σ_i^2 and Holder exponents $H_i(t)$. Let the aggregate process be denoted by $W(.)$, and the envelope process of each source given by $\hat{A}_i(t)$. The aggregate envelope process $\hat{A}^N()$ for the cumulative work of $W(.)$ in the interval $[0,t]$ is given by:

$$\hat{A}^N(t) = \sum_{i=1}^N \hat{A}_i(t)$$

$$\hat{A}^N(t) = \int_0^t \sum_{i=1}^N \bar{a}_i + \kappa\left(\sum_{i=1}^N \sigma_i^2 H_i(x)x^{2H_i(x)-1}\right)$$

$$\left(\sum_{i=1}^{N}\sigma_i^2 x^{2H_i(x)}\right)^{-\frac{1}{2}} dx \tag{16}$$

where $\hat{A}_i(t)$ is the envelope process for the i^{th} flow.

Replacing $\hat{A}(t)$ in Equation (7) by the aggregate envelope process $\hat{A}^N(t)$, gives the following:

$$\kappa\left(\sum_{i=1}^{N}\sigma_i^2 H_i(t)t^{2H_i(t)-1}\right)\left(\sum_{i=1}^{N}\sigma_i^2 t^{2H_i(t)}\right)^{-\frac{1}{2}} = C - \sum_{i=1}^{N}\bar{a}_i \tag{17}$$

Equation (17) can be solved numerically to find the maximum time scale of a queue fed by several streams, $t^{\star\star}$, which is then inserted in Equation (11) to compute $\hat{q}^N_{max}$. Moreover, combining Equations (13) and (15) results in the following:

$$\kappa\int_0^t\left(\sum_{i=1}^{N}\sigma_i^2 H_i(x)x^{2H_i(x)-1}\right)\left(\sum_{i=1}^{N}\sigma_i^2 x^{2H_i(x)}\right)^{-\frac{1}{2}} dx - \kappa\left(\sum_{i=1}^{N}\sigma_i^2 H_i(t)t^{2H_i(t)}\right)\left(\sum_{i=1}^{N}\sigma_i^2 t^{2H_i(t)}\right)^{-\frac{1}{2}} - \hat{q}^N_{max} = 0 \tag{18}$$

Computing $t^{\star\star}$ from Equation (18) and inserting it in Equation (13) makes it possible to answer the fundamental question posed at the beginning of this section, i.e., what is the equivalent bandwidth of an aggregate of multifractal traffic streams, which is given by:

$$\hat{C} = \sum_{i=1}^{N}\bar{a}_i + \kappa t^{\star\star-1}\int_0^{t^{\star\star}}\left(\sum_{i=1}^{N}\sigma_i^2 H_i(x)x^{2H_i(x)-1}\right)\left(\sum_{i=1}^{N}\sigma_i^2 x^{2H_i(x)}\right)^{-\frac{1}{2}} dx - \frac{\hat{q}^N_{max}}{t^{\star\star}} \tag{19}$$

or equivalently

$$\hat{q}^N_{max} = \int_0^{t^{\star\star}}\sum_{i=1}^{N}\bar{a}_i + \kappa\left(\sum_{i=1}^{N}\sigma_i^2 H_i(x)x^{2H_i(x)-1}\right)\left(\sum_{i=1}^{N}\sigma_i^2 x^{2H_i(x)}\right)^{-\frac{1}{2}} dx - Ct^{\star\star} \tag{20}$$

Note that Equations (17) and (18) do not require previous knowledge of the whole stream and can be used in a measurement-based framework. In such a framework the mean, the variance and the Holder exponent values can be measured and inserted in Equation 19 to estimate on-line the equivalent bandwidth of an aggregate stream.

For the special case of multiplexing N identical sources, the envelope process, $\hat{A}^N(.)$, is given by:

$$\begin{aligned}\hat{A}^N(t) &= \sum_{i=1}^{N} \hat{A}_i(t) \\ &= \int_0^t N\bar{a} + \kappa\left(N\sigma^2 H(x)x^{2H(x)-1}\right)\left(N\sigma^2 x^{2H(x)}\right)^{-\frac{1}{2}} dx \end{aligned} \quad (21)$$

In this case, Equation 17 is reduced to

$$\frac{\kappa(N\sigma^2 H(t)t^{2H(t)-1})}{\sqrt{N}\sigma t^{H(t)}} = N(c-\bar{a}) \quad (22)$$

Using the same approach as above, it is possible to obtain $t^{\star\star}$ and $\hat{q}^N_{max}$:

$$t^{\star\star} = N^{\frac{1}{2(H(t^{\star\star})-1)}} t_i^\star \quad (23)$$

$$\hat{q}^N_{max} = \sqrt{N}\int_{t_i^\star}^{t^{\star\star}} \kappa\sigma H(x)x^{H(x)-1}dx + N^{\frac{H(t^{\star\star})-1/2}{H(t^{\star\star})-1}}\hat{q}_{max} \quad (24)$$

$$t_i^\star = \left[\frac{\kappa\sigma H(t^\star)}{(c-\bar{a})}\right]^{\frac{1}{1-H(t^\star)}} \quad (25)$$

$$\hat{q}_{max} = \hat{A}(t_i^\star) - ct_i^\star \quad (26)$$

where $t_i^\star$ and $\hat{q}_{max}$ are derived from a queueing system fed by a single source and $c = C/N$.

6 Numerical Examples

In order to evaluate the accuracy of the expressions defined in Section 5, simulation experiments using both synthetic and real network data were pursued. In the experiments a queue is fed by several multifractal streams; the service rate and the buffer size were varied. In the numerical examples presented here the service rate is 5% higher than the total mean rate. The queue length is recorded and the maximum queue length is compared against the the maximum queue length estimated by Equation 20. Table 6 shows the characteristics of the synthetic traces utilized. Different combinations of the process in Table 6 were employed to produce the aggregate traffic. Figure 2 shows the evolution of the queue length for the aggregation of all five traces. It can be seen that the predicted maximum time scale by Equation 17 matches exactly the one obtained via simulation. Such pattern were observed in all experiments with synthetic traces.

Real network traffic were obtained from the NLANR site (www.nlanr.net). Traces used in previous investigations of others were also utilized for comparison purpose [10][11]. Table 2 shows the characteristics of the traces used. Experiments using real network traffic were also carried out. Figure 3 shows the maximum time scale predicted by Equation 17 and by the one obtained in the simulation using the four traces *dec-pkt* in Table 2. It can be seen

Fluxo	$\bar{a}$	σ^2	$H(.)$
1	14.22	82.09	$\frac{19}{10}t^2 - \frac{19}{10}t + 0.985$
2	12.96	88.18	$\frac{49}{10}t^3 - \frac{79}{10}t^2 + \frac{33}{10}t + 0.51$
3	13.03	200.93	$-\frac{21}{10}t^4 + \frac{11}{10}t^3 - \frac{1}{10}t^2 + \frac{8}{10}t + 0.51$
4	14.43	87.93	$\frac{sin(t)}{10} + 0.61$
5	14.43	187.93	$\frac{49}{100}t + 0.5$

Table 1: the synthetic network traffic

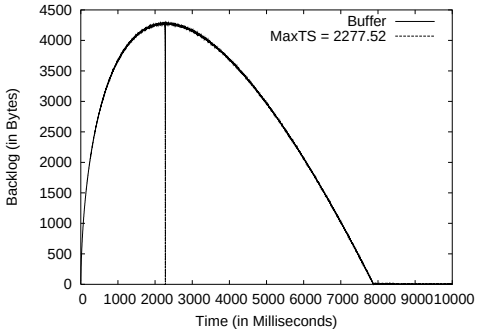


Figure 2: MaxTS for synthetic network traffic

that the predicted value for the maximum time scale is very close to the one obtained via simulation. Actually, the deviation is within the interval defined by the pre-defined error margin.

The whole advantage of statistical multiplexing is the efficient use of resources achieved by interleaving packets of different streams which allows the support of a higher number of users when compared to circuit switching. To evaluate the benefits of using the expression for the equivalent bandwidth of an aggregate of multifractal flows a gain measure, $G(n)$, was defined as the ratio between n times the equivalent bandwidth of a trace and the equivalent

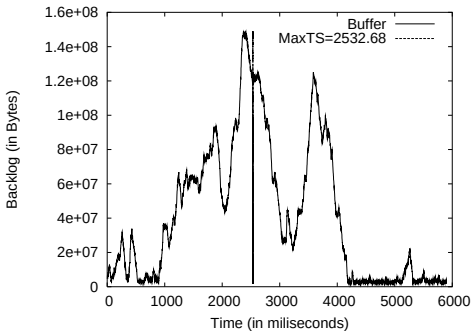


Figure 3: MaxTS for real network traffic

Trace	Date	# of packets	aggregation point
ANL-1050127417	04/11/03	121998	Agonne NL to STARTAP
ANL-1050225668	04/13/03	105641	Agonne NL to STARTAP
MEM-1053844177	05/24/03	220904	University of Memphis
MEM-1054459191	06/01/03	266708	University of Memphis
COS-1057970154	07/12/03	1247518	Colorado State University
BWY-1058086940	07/13/03	1168143	Columbia University
dec-pkt-1	03/08/95	3300000	Digital IAP
dec-pkt-2	03/09/95	3900000	Digital IAP
dec-pkt-3	03/09/95	4300000	Digital IAP
dec-pkt-4	03/09/95	5700000	Digital IAP

Table 2: Real trace with Traffic gather at Internet 2 ABILENE and Digital's Intenet Access Point

bandwidth for the aggregate of n traces with equal statistical characteristics. $G(n)$ is given by:

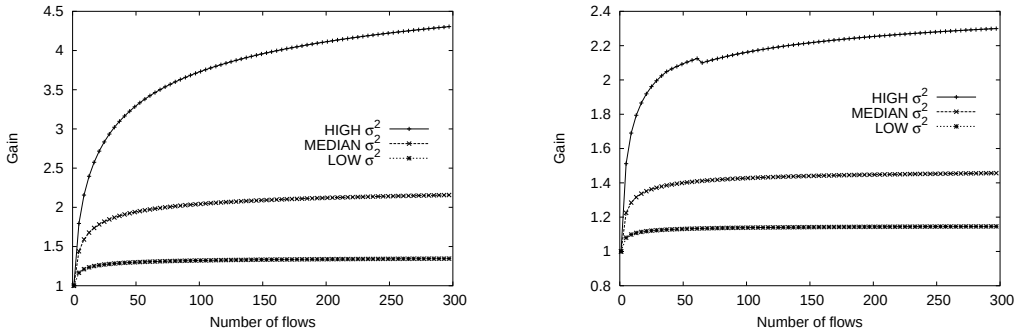
$$G(n) = \frac{nEB(1)}{EB(n)} =$$

$$\frac{\int_0^{t^\star} \bar{a} + \kappa\left(\sigma^2 H(x) x^{2H(x)-1}\right)\left(\sigma^2 x^{2H(x)}\right)^{-\frac{1}{2}} dx - K}{\int_0^{t^{\star\star}} \bar{a} + n^{\frac{-1}{2}}\kappa\left(\sigma^2 H(x) x^{2H(x)-1}\right)\left(\sigma^2 x^{2H(x)}\right)^{-\frac{1}{2}} dx - K'} \tag{27}$$

where EB_1 is the equivalent bandwidth for a single flow and EB_n is the equivalent bandwidth of an aggregate of n flows. $t^\star$ e $t^{\star\star}$ are the time scales given by Equation 14 and by Equation 23, respectively. K is the buffer size at the multiplexer and $K' = \frac{K}{n}$.

The gain $G(.)$ was evaluated as a function of the number of aggregated flows for different traffic characteristics. Figure 4 shows the gain for traces with Holder exponents, $H(.)$, given by the quadratic and by the cubic functions defined in Table 6 for different variance values. The mean arrival rate is $\bar{a} = 125.09$ and the variance for the flow called "low" is $\sigma^2 = 290.00$. For the curves named "median" and "high" the variance values are $10\sigma^2$ qand $100\sigma^2$, respectively.

It can be observed in Figure 4 that the gain increases with the variance. For instance, for Holder exponents given by a quadratic function, the maximum gain is 1.35 for streams with low variance whereas it is 4.0 for streams with high variance. The gain is also influenced by the Holder exponent values. Actually, what is relevant is the mean value of the Holder exponent up to the maximum time scale, i.e., $\frac{\int_0^{t^{\star\star}} H(x)dx}{t^{\star\star}}$, which can be noticed by comparing Figures 4.a and 4.b. For the same mean and variance value, the gain is higher for the traces with Holder exponents given by a quadratic function than for the traces with exponents given by a cubic function. For instance, considering flows with median variance, the maximum gain is 2.2 for the traces with exponents given by a quadratic function whereas it is 1.45 for traces with exponents given by a cubic function.



(a) Quadratic H(.) (b) Cubic H(.)

Figure 4: Multiplexing gain with homogeneous flows for synthetic traffic

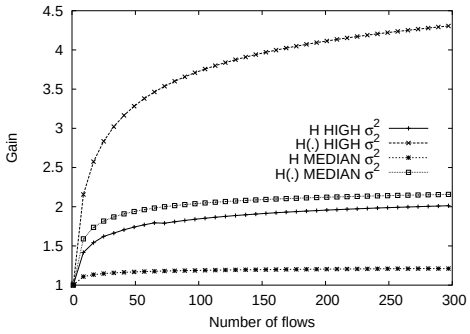
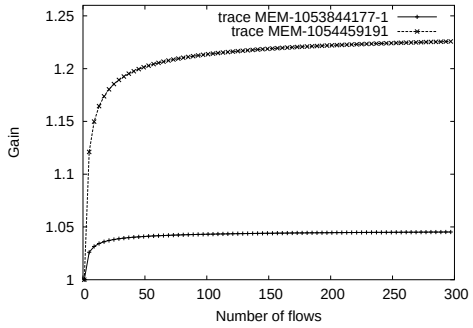


Figure 5: Multiplexing gain with homogeneous flows for H(.) versus mean H

To answer the question of whether a monofractal approach could be used for resource dimensioning in networks with multifractal flows, a monofractal envelope (Equation 6) was derived for the trace with Holder exponents given by a cubic function. Figure 5 shows the gain considering the mbm envelope process, denoted, $H(.)$, and considering the monofractal envelope process, H, for both high and median variance. It can be seen that the gain obtained by using the mBm process is double the value produced by using the monofractal envelope for median values of the variance whereas it is more than the double for high variance values. This happens since the Hurst parameter overestimates the local behavior of the Holder exponents which leads to overprovisioning of resources and, consequently, a lower gain.

Evaluation of the gain was also pursued using real network traces. The evaluation with the traces MEM-1053844177 and MEM-1054459191 (Table 2) are displayed in Figure 6. The traffic descriptors of these traces are $\bar{a} = 1014.88$, $\sigma^2 = 3.2e6$ and $\bar{a} = 1810.98$, $\sigma^2 = 8.6e6$, respectively. The gain for the trace MEM-1054459191 is higher than for the trace MEM-1053844177. Again, this happens because the trace MEM-1054459191 has higher variance value and higher mean $H(.)$ value up to the maximum time scale.



(a)

Figure 6: Multiplexing gain with homogeneous flows for real traffic

7 Related work

A. Erramilli, O. Narayan, A. Neidhardt and I Saniee [3] proposed that traffic should be modeled by random cascades at time scales smaller than a cutoff value and be represented by an fBm at larger scales. They show that for IP traffic, the cutoff scale is of the order of one Round Trip Time (RTT), while for VBR video it is typically of the order of a frame duration. Erramili et. al. showed that much more accurate results can be obtained by using their model rather than using purely monofractal models.

Other models based on multiplicative cascade have been proposed. These models map a given sample into a binary multiscale tree [11]. Each node in the tree corresponds to the aggregation of the traffic mapped into its descendents. Thus, nodes at higher levels of the tree correspond to coarser time scale whereas nodes at lower levels correspond to finer time scales. The multipliers (weights) assigned to each descendent of a node can be set to represent a specific marginal distribution and scaling. In the Multifractal Wavelet Model (MWM)[12], multipliers are multiplicative innovations, generating approximately a log-normal marginal distribution. Both models require the setting of $2 + log_2 N$ parameters where N is the sample size. The major drawback of MWM, is the number of parameters to be fitted and the requirement of having to construct a multiscaling binary tree which is not suitable for on-line characterization. A maximum time scale for queues fed by MWM has also been derived.

Recent investigation [13] on small time scales of Internet traffic points out that monofractal behavior is observed at these scales. It is claimed that correlations at small time scales are caused mainly by flows with bursts of densely clustered packets and not by the acknowledgement mechanism of TCP. However, in our investigations using publicly available traces we found clear multifractal behavior at these scales.

Another thread of work related to this paper is traffic modeling based on measurements. In [14], the time scale at which losses most probably occur, the (dominant) times scale of interest is computed using a hybrid measurement/analytical approach. The time scale is

determined by observing a virtual queue with smaller capacity than the one under study given that losses at the former are more frequent. In [15], a maximum rate envelope process is used to caracterize the arrival as well as the service rate for these envelopes are derived via measurement. The maximum time scale can be derived by using these two measures.

Conclusions

The provisioning of Quality of Service for applications in the Internet is a major challenge yet to be overcome. Central to such provisioning is the ability to compute the ammount of bandwidth demanded by a flow so that the QoS requirements of that flow are supported. Moreover, it has been shown that Internet traffic presents multi-scaling characteristics which can be accurately captured by multifractal processes.

This paper introduced expressions for the time at which the length of a queue fed by several multifractal flows reaches its maximum. The equivalent bandwidth of an aggregate of multifractal flows was also furnished. These expressions can be used in measurement based admission control in DiffServ networks as well as for dimensioning LSP's in MPLS networks.

Acknowledgments

This work was partially sponsored by FAPESP (grant 00/09772-6) and by CNPq (grant 300064/95-0).

References

[1] W. Leland, M. Taqqu, W. Willinger, and D. Wilson, "On the self-similar nature of ethernet traffic (extended version)," *IEEE/ACM Trans. Networking*, pp. 1–15, February 1994.

[2] R. Riedi and J. Levy-Vehel, "Tcp traffic is multifractal: A numerical study," INRIA Rocquencourt, Tech. Rep. 3129, March 1997.

[3] A.Erramilli, O.Narayan, A.Neidhart, and I.Saniee, "Performance impacts of multi-scaling in area TCP/IP traffic," in *INFOCOM 2000*, April 2000, pp. 352–359.

[4] P. Abry, R. Baraniuk, P. Flandrin, R. Riedi, and D. Veitch, "The multiscale nature of network traffic: Discovery, analysis and modelling," *IEEE Signal Processing Mag.*, vol. 19, no. 3, pp. 28–46, 2002.

[5] A. Feldmann, A. C. Gilbert, P. Huang, and W. Willinger, "Dynamics of ip traffic: a study of the role of variability and the impact of control," in *Proceedings of the conference on Applications, technologies, architectures, and protocols for computer communication.* ACM Press, 1999, pp. 301–313.

[6] C. A. Melo and N. Fonseca, "An Envelope Process for Multifractal Traffic Modeling," Institute of computing State University of Campinas, Tech. Rep. 10, April 2003, on-line http://www.ic.unicamp.br/reltec-ftp/2003/Titles.htm.

[7] R. Peltier and J.L.Vehel, "Multifractional brownian motion: definition and preliminary results," INRIA, Tech. Rep., 1995.

[8] N.L.Fonseca, G. Mayor, and C. Neto, "On the equivalent bandwidth of self-similar source," *ACM Transactions on Modeling and Computer Simulation*, vol. 10, no. 2, pp. 104–124, April 2000.

[9] I. Norros, "A storage model with self-similar input," *Queueing Systems*, vol. 18, pp. 387–396, 1994.

[10] A.Erramilli, O.Narayan, A.Neidhart, and I.Saniee, "Multi-scaling models of TCP/IP and sub-frame VBR video traffic," *Journal of Communications and Networks*, vol. 3, no. 4, pp. 383–395, December 2001.

[11] R. H. Riedi, M. S. Crouse, V. J. Ribeiro, and R. G. Baraniuk, "A multifractal wavelet model with application to network traffic," *IEEE Trans. Inform. Theory*, vol. 45, no. 4, pp. 992–1018, 1999.

[12] V. J. Ribeiro, R. H. Riedi, M. S. Crouse, and R. G. Baraniuk, "Multiscale queuing analysis of long-range-dependent network traffic," in *INFOCOM (2)*, 2000, pp. 1026–1035.

[13] Z. Zhang, V. Ribeiro, S. Moon, and C. Diot, "Small-Time Scaling behaviors of internet backbone traffic: An Empirical Study," in *IEEE Infocom*, San Francisco, March 2003.

[14] D. Y. Eun and N. B. Shroff, "A measurement-analytic approach for qos estimation in a network based on the dominant time scale," *IEEE/ACM Trans. Networking*, vol. 11, no. 2, pp. 222–235, 2003.

[15] J. Qiu and E. Knightly, "Measurement-based admission control using aggregate traffic envelopes," *IEEE/ACM Trans. Networking*, vol. 9, no. 2, pp. 199–210, April 2001.