Bandwidth Allocation for self-similar traffic consisting of multiple traffic classes with distinct characteristics

P. Orenstein, H. Kim and C.L. Lau

Telcordia Technologies
Applied Research
Red Bank
NJ 07701, USA

***Abstract* – It is now generally accepted that modern network traffic (for example, Ethernet, VBR and Web) is 'self-similar' in nature, i.e., the statistical characteristics of the traffic stream remains largely unchanged when examined at several different scales. One appropriate model for self-similar traffic is Fractional Brownian Motion (FBM).**

The main contribution of this paper is to formulate and verify a set of general rules that can be applied to calculate the FBM parameters of an aggregate traffic profile consisting of multiple input streams, each with its own characteristic burstiness profile.

The parameters of the aggregated FBM traffic can then be used to calculate the effective bandwidth of an aggregate traffic pattern formed by multiplexing a mixture of heterogeneous traffic streams.

I. Introduction

Statistical analysis of high-resolution traffic measurements from a wide range of working packet networks (e.g. Ethernet LANs [1], compressed video streams [2], wide area TCP/IP traffic [3] and World Wide Web (WWW) traffic loads [4] have convincingly shown the presence of fractal or self-similar properties in both local area and wide area traffic traces. This means that similar statistical patterns may occur over different time scales that can vary by many orders of magnitude (i.e. ranging from milliseconds to minutes and even hours). The fact that network traffic is inherently fractal or long-range dependent (LRD) poses a problem for many teletraffic related traffic engineering problems, e.g. traffic measurements and performance, buffer sizing, admission control and congestion control. While most of the work done in science and engineering has almost exclusively focused on the practical features of fractal models (e.g. data analysis and mathematical modeling), their engineering impacts on performance and analysis is not well understood. This is mainly because of the difficulties related to analysis and the ability to use these models in control.

One such model is the bandwidth allocation formula, which is based on Fractional Brownian Motion (FBM), [5]. It is attractive because it relates the fractal characteristics of the traffic stream with quality of service (loss) and the physical network constraints. The model can be used for planning. However, While the model can be used to guarantee bandwidth for different types of fractal traffic (Web, Video, etc.), it can only be used on a per class basis. Traffic in modern packet networks is often aggregated from multiple traffic types and the resultant stream is typically a heterogeneous mix. In designing SLAs, network planners are interested in determining the bandwidth required to guarantee service for traffic aggregated from a number of distinct traffic classes.

II. Overview of self-similar traffic

The mathematics of self-similarity is described in detail in [6]. We plan to employ two statistical tests to measure the degree of self-similarity of the aggregate traffic stream. These are (a) the variance-time plot and (b) the R/S Statistic. For a rigorous definition, see [2,6].

A. *Mathematics*

Let $X = (X_t : t = 0,1,2,....)$ be a stochastic process with constant mean $m = E[X_t]$ and finite variance $s^2 = E[X_t - m]^2$. For each m=1,2,3,…., let $X^{(m)} = (X^{(m)}{}_k :$ k=0,1,..) denote the time series obtained by averaging the original time series X over non-overlapping blocks of size m. That is, each m=1,2,3,…, $X^{(m)}$, is given by $X_k{}^{(m)} = \frac{1}{m}(X_{km-m+1} + \cdots + X_{km}), \quad (k \geq 1)$.

The "variance-time" plot is based upon the fact that for self-similar processes, the variances of the aggregated processes $X^{(m)}$ decrease linearly (for large m) in log-log plots against *m* with $Var(X^{(m)}) \approx m^{-b} s^2$, where $0 < b < 1$. The "variance-time plot" is a graphical method for distinguishing between SRD (β=1) and LRD (0<β<1) for a given empirical record. To estimate β, which is related to *H* by $H=1-b/2$, we plot $\log(\mathrm{var}(X^{(m)}))$ against *log(m)*, yielding β as the limiting slope as m→∞.

0-7803-7206-9/01/$17.00 © 2001 IEEE

Another graphical method is called the R/S analysis method. It is based on the rescaled adjusted range statistics (R/S), see [2] for a formal definition. For a self-similar process, a plot of the R/S statistic versus the sample size results in a slope of (0.5 < H <1).

An example of a self-similar process with self-similarity parameter H is Fractional Gaussian Noise (FGN) with parameter ½ < H < 1 introduced by [7]. Associated with FGN is Fractional Brownian Motion (FBM), which is the integrated version of FGN (that is, an FBM process is simply the sum of FGN increments). Later, we will use a technique described by [8], to generate synthetic traces of FGN.

B. *Effective Bandwidth*

When traffic is composed of flows from many independent sources, then, providing each source is heavy-tailed, the resulting aggregation is self-similar [9]. The author in [10] provides a mathematical characterization of the effective bandwidth *r,* required by a self-similar traffic source, as

$$r \geq m + \left[\frac{-2am\ln e}{b^{2(1-H)} f(H)} \right]^{1/2H}$$

$$f(H) = \frac{1}{(1-H)^{2(1-H)} H^{2H}} \qquad \textbf{(1)}$$

where, m is the mean traffic arrival rate of the traffic stream (say, in bps), a is the 'variance coefficient' of the traffic stream (say in bit-sec), H is the Hurst parameter in the range $0.5 \leq H \leq 1$, b is the buffer size (in bits) and , ε is the target loss rate for the traffic stream. The variance coefficient *a,* is calculated as the variance to mean ratio of the traffic stream. We refer to *(m,a,H)* as the three parameters which characterize an FBM traffic source. Note that the two parameters *a* and *H* characterize the "quality" of the traffic, whereas the long run mean rate *m,* characterizes the "quantity".

Two assumptions in the theoretical derivation of the formula should be observed. First, the formula depends on the traffic being sufficiently Gaussian, which is more likely when traffic is aggregated from a large number of sources. Second, the derivation of the formula uses the Weibull distribution to approximate the tail of the queue length distribution. This approximation is logarithmically accurate for large buffers, so a sufficiently large buffer must be available for the theory to apply.

The applicability of the Norros formulae in practice has been demonstrated through simulation, [11]. Furthermore, the authors justify the large buffer and Gaussian assumptions.

III. *Rules for Combining Non-Homogeneous Traffic Streams*

The aim of this paper is to formulate a set of general rules that can be applied to calculate the FBM parameters of an aggregate traffic profile consisting of multiple input streams, each with its own characteristic burstiness profile.

A. *Background*

Consider finding a reasonable FBM approximation to an arbitrarily given stochastic process $A(.)$ with stationary increments (so that $A(t) - A(s)$ might occur as a model for the amount of traffic arriving in the interval *(s,t]* of time).

Assume that $A(.)$ is defined such that *E[A(t)]=mt* for all m in (-∞,+∞), $m\,\hat{I}\,\hat{A}$, (where, m is the mean arrival rate for A), and there is a function *V(t)=var[A(t)]* which is defined and finite for all *t.* Since we are dealing with a model of traffic arrivals, we restrict our attention to cases where *m>0.*

In general, $V(.)$ is not particularly elementary, but one can hope to approximate it with a relatively simple formula, such as that of an FBM with Hurst parameter *H* and variance parameter *v (=a*m, m>0)* expressed as

$$V_{fbm}(t;H,v) = vt^{2H}$$

The true arrival process *A* is well approximated by an FBM with parameters *(H,v,m)* if (a) the arrival process *A* is Gaussian, and (b) $V(.)$ and $V_{fbm}(;,H,v)$ are sufficiently close. Since the main question here is the choice of parameters, the discussion here can ignore (a), so that attention can be concentrated on (b).

Next, consider an aggregate stream $A_g(.) = \sum_s A_s(.)$ which is formed from the superposition of independent arrival processes $A_s(.)$, where each $A_s(.)$ is assumed to be well approximated by an FBM with parameters m_s ,H_s and v_s.

Naturally, the mean rate m_g, of the aggregate stream A_g is just the sum of the mean rates m_s of the individual streams. But we need to find values for H_g and v_g for which $V_{fbm}(t;H,v)$ » $V_g(t)$. Since we assume that the individual streams are well approximated by FBMs, we can write

$$V_g(t) = \mathrm{var}(A_g(t)) = \sum_s \mathrm{var}(A_s(t) \cong \sum_s V_{fbm}(t;H_s,v_s)$$

On this basis, our aim is to find values H_g and v_g so that

$$V_{fbm}(t;H_g,v_g) \cong \sum_s V_{fbm}(t;H_s,v_s)\ .$$

Now the FBM formulas are straight lines in log-log coordinates, so we can consider functions of the type $W_{fbm}(t;H,v) = \log(V_{fbm}(e^t;H,v)$, which is a straight-line graph, since,

$$\log(V_{fbm}(e^t;H,v) = \log(ve^{t^{2H}}) = 2Ht + \log(v)$$

where, the independent variable τ is the log of the time scale at which fluctuations are observed, and the dependent variable W_{fbm} is the log of the variance). Correspondingly, on the same graph, one would be trying to approximate the curve

$$W_g(t) = \log(V_g(e^t) \cong \log(\sum_s v_s e^{2H_s t}) = W_{g0}(t)$$

For the homogeneous case (where H_s=H " H_s), $W_g(t)$ reduces to a linear expression of the form $2Ht + \log(\sum_s v_s)$ which is just $W_{fbm}(t;H_g = H, v_g = \sum_s v_s)$.

In general, however, the H_s are not all the same, in which case W_g is not linear. Assuming that W_{g0} is a good representation of W_g, we can state that W_{g0} is a convex function with a slope rising to $2\max(H_s)$ as $t \to +\infty$. From this perspective, our aim is to find parameter values, H_g and v_g for which the straight line $W_{fbm}(.;H_g,v_g)$ is close to the curve $W_{g0}(.)$

To do this, we consider two approaches. First, motivated by the desire to match the LONG-RANGE dependence as accurately as possible, one could choose $H_g = H_\infty$ and $v_g = v_\infty$ with

$$H_\infty = \max(H_s) \text{ and } v_\infty = \sum_s \{v_s \mid H_s = H_\infty\},$$

so that the line $W_{fbm}(.;H_\infty,v_\infty)$ would be the asymptote of W_{g0} as $t \to +\infty$ (i.e. at large time scales). This will be referred to as Method A.

A second approach is to fix an arbitrary logarithmic time scale x_i, and find $H_g = H(x_i)$ and $v_g = v(x_i)$ so that the line $W_{fbm}(.;H(x_i),v(x_i))$ is tangent to $W_{g0}(.)$ at $t = x_i$. This constraint yields a different set of conditions

$$H(x_i) = \frac{\sum_s H_s v_s e^{2H_s x_i}}{\sum_s v_s e^{2H_s x_i}} \text{ and } v_s = \sum_s v_s e^{2x_i(H_s - H_g)}$$

Notice that, as one might expect, $H(x_i)$ is an average of the (with weights $v_s e^{2H_s x_i}$). In particular, $H(x_i) < H_\infty$. For the special case of the distinguished logarithmic time scale being $x_i = 0$, corresponding to a physical time scale of $e^{x_i} = 1$ time unit, the expressions simplify to being

$$H(0) = \frac{\sum_s H_s v_s}{\sum_s v_s} \text{ and } v_s = \sum_s v_s$$

These will be referred to as Method B. The approximations for v_g and H_g for both Method A and Method B (assuming unit time scale) are summarized in TABLE I. Note also that $a_g = v_g/m_g$.

TABLE I

VARIANCE/ HURST APPROXIMATIONS FOR AGGREGATE STREAM

	Variance	Hurst
A	$v_g = \sum(v_s \mid H_s = H_{\max})$	$H_g = H_{\max}$
B	$v_g = \sum v_s$	$H_g = \sum \frac{v_s H_s}{\Sigma v_s}$

B. *Numerical Examples*

We used the Fast Fourier Transform (FFT) method described in detail in [8] to generate approximately self-similar traffic traces. Each represented a single source of traffic, for example Web, Video, Ftp and Email. Each number in the trace corresponded to the number of bytes allocated to bins of fixed size duration (1 sec). The results of the analysis are summarized in TABLE II.

For the analysis, we produced two traffic traces, each of which was characterized by separate FBM parameters. These could represent the characteristics of two traffic classes, for example Web and Email. The traces were then combined and the FBM parameters of the aggregated trace were estimated using the methods described in Section IIA. In total, six parameter sets were investigated, labeled (A-F). Each simulation case (A-F) was replicated with 20 traces. Thus, the reported entries correspond to the mean over all traces along with the standard deviation.

Case A is the homogeneous case and the simulation results track the analytic predictions as expected. For cases (B-F), the simulation results for variance-to-mean ratio of the aggregated stream are close to Method B. By contrast, the Hurst parameter estimate falls in between the values predicted by the analytic methods A and B. TABLE III compares the relative error between the simulation and analytic for the different methods for both the variance and the Hurst parameter predictions. It illustrates the point that while Method B provides a tighter estimate for the variance parameter, for the Hurst parameter, both methods produce estimates with reasonably sufficient accuracy.

We then applied the bandwidth allocation formula (1), assuming *SLA Loss = 0.001* and buffer size varying from 0.5Mb to 10 Mb. The mean, variance and Hurst parameters were obtained from TABLE I.

Fig. I portrays the computed effective bandwidth using the parameters from the different methods. It confirms that for case C, Method B is preferred. This trend was noticed for the other cases too, but the degree of discrepancy between the simulation results and Method B was sometimes more noticeable. We were not able to quantify this discrepancy within the limited scope of these examples, but expect to explore this phenomenon in future numerical experiments.

TABLE II
APPROXIMATION OF FBM PARAMETERS: ANALYTIC VS SIMULATION

	Input 1	Input 2	Analytic A	Analytic B	Simulation
A	0.20, 0.56, 0.85	0.33, 0.48, 0.85	0.53, 0.51, 0.85	0.53, 0.51, 0.85	0.53, 0.51 ± 0.0133, 0.85 ± 0.02
B	0.33, 0.68, 0.65	0.33, 0.48, 0.85	0.66, 0.24, 0.85	0.66, 0.58, 0.73	0.66, 0.58 ± 0.0016, 0.80 ± 0.01
C	0.33, 0.48,0.85	0.20, 1.20, 0.80	0.53, 0.30, 0.85	0.53, 0.75, 0.82	0.53, 0.76 ± 0.0099, 0.83 ± 0.02
D	0.20, 0.56, 0.85	0.33, 0.68, 0.65	0.53, 0.34, 0.85	0.53, 0.63, 0.72	0.53, 0.64 ± 0.0021, 0.80 ± 0.02
E	0.20, 0.56, 0.85	0.20, 1.20, 0.80	0.40, 0.56, 0.85	0.40, 0.88, 0.82	0.40, 0.88 ± 0.0143, 0.82 ± 0.02
F	0.20, 1.20, 0.80	0.33, 0.68, 0.65	0.53, 0.73, 0.80	0.53, 0.88, 0.73	0.53, 0.88 ± 0.0023, 0.77 ± 0.01

TABLE III
RELATIVE ERROR FOR EACH METHOD

	Var(A)	Var(B)	H(A)	H(B)
A	0.0208± 0.016	0.0208± 0.016	0.0242± 0.015	0.0242± 0.015
B	1.417± 0.0065	0.0021±0.002	0.0551± 0.016	0.096± 0.0182
C	1.525± 0.033	0.0113± 0.075	0.026± 0.0178	0.0202± 0.0139
D	0.871± 0.0061	0.0025± 0.002	0.064± 0.0263	0.1100± 0.0312
E	0.5761± 0.025	0.0139±0.008	0.041± 0.0265	0.0257± 0.0136
F	0.205± 0.0032	0.0022± 0.002	0.0910± 0.011	0.0620± 0.0129

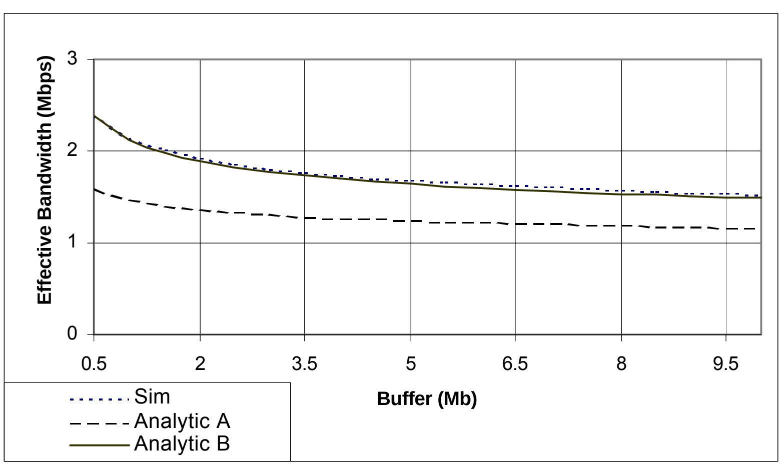

Fig. I: EFFECTIVE BANDWIDTH COMPUTED USING DIFFERENT APPROXIMATIONS FOR FBM PARAMETERS

IV. CONCLUSIONS

The FBM approximation for an aggregated stream with parameters (m_g,a_g,H_g) was computed using two methods, Method A and Method B. In general, while Method A and Method B could be used to track the estimated Hurst parameter, H_g, for the variance, Method A was found to be inadequate. The bandwidth allocation for the aggregated traffic stream was also computed using the different techniques. In some cases, Method B was found to be more appropriate in calculating the effective bandwidth but there were exceptions, highlighting the need for further analysis.

The scope for this work was limited to SLA Planning. We expect that the rules derived in this study could be used to design buffer and bandwidth requirements for aggregate traffic formed from the superposition of individual self-similar streams.

V. FUTURE WORK

While the analysis has been limited to two sources, in principle the methodology can be applied to any number of sources, or groups of sources (classes). The next step would be to conduct the numerical experiments with an increased numbers of inputs. This would illustrate the effectiveness of FBM to model the superposition of in-homogenous inputs, and test the limitations of the assumptions.

In addition, we would like to use simulation to observe the loss experienced when multiplexing heterogeneous sources through a single buffer. This would provide a more reliable estimate for the effective bandwidth.

Ultimately, it is envisaged that the methodologies described in this paper will be incorporated into a tool for SLA planning of self-similar traffic in modern high-speed networks.

ACKNOWLEDGMENT

The authors would like to thank their colleague, Arnie Neidhardt, for his insightful comments and contribution to the material outlined in IIIA.

REFERENCES

[1] W. Leland, Taqqu M., W. Willinger and D. Wilson, “On the self-similar nature of Ethernet Traffic (Extended Version)”, *IEEE/ACM Transactions on Networking,* vol. 2, no. 1, pp. 1-15, Feb 1994.

[2] M. Garrett and W. Willinger, “Analysis, Modeling and Generation of Self-Similar and VBR Video Traffic”, *Proceedings of 1994 ACM Sigmetrics Conference*, London, UK, pp. 160-169, Aug 1994.

[3] V. Paxson and S. Floyd , “Wide Area Traffic: The failure of Poisson Modeling, *IEEE/ACM Transactions on Networking,* vol. 3, no. 3, June 1995.

[4] M. Crovella and A. Bestavros, “Self Similarity in World Wide Web Traffic: Evidence and Possible Causes” Proceedings of 1996 *ACM Sigmetrics Conference*, Philadelphia, PA, pp. 160-169, May 1996.

[5] I. Norros, A storage model with self-similar input, *Queuing Systems*, vol. 16, pp. 387-396, 1994.

[6] W. Leland, Taqqu M., W. Willinger and D. Wilson, "Self-Similarity in High-Speed Packet Traffic: Analysis and Modeling of Ethernet Traffic Measurements," *Statistical Science*, vol. 10, no. 1, pp 67-85, 1995.

[7] B. Mandelbrot and J. W. Van Ness, Fractional Brownian Motions, Fractional Noises and Applications, *SIAM Review*, vol. 10, pp. 422-437, 1968.

[8] V. Paxson, Fast Approximation of Self-Similar Traffic, *Technical Report LBL-36750, Lawrence Berkeley Laboratory and EECS Division*, University of California, Berkeley, April 1995.

[9] M. Taqqu, W. Willinger and R. Sherman, “Proof of a fundamental result in self-similar traffic modeling”, *Comput. Commun. Rev.* vol. 26., pp. 5-23, 1997.

[10] I. Norros, “On the use of Fractional Brownian Motion in the theory of connectionless networks”, *IEEE Journal on Selected Areas in Commun.,* vol. 13(6), pp. 1028-1038, Aug 1995.

[11] A. Patel and C. Williamson, “Effective Bandwidth of self-similar traffic sources: theoretical and simulation results”, *Proc. of the IASTED Conf. on Applied Modeling and Simulation (AMS'97)*, Banff, AB, pp. 298-302, July 1997.