A Policy for fast Connection Admission Management of Video Streams

Nelson L.S. Fonseca & Cesar A. V. Neto
State University of Campinas
Institute of Computing
13083-970 Campinas SP
Brazil

Abstract

In this paper, we introduce a simple greedy policy for connection admission of video streams. We show that this simple policy produces, for moderate to high loads the same revenue as the revenue produced by a policy based on a Knapsack solution which maximizes the total network revenue. This simple policy is a potential good candidate for fast connection admission management in ATM networks.

1. Introduction

Several studies [1]-[4] have claimed that different types of network traffic, *e.g*. video streams, 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], [6] 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. Furthermore, this queueing system suffers from the buffer inefficacy phenomenon. By just increasing the buffer size we are not able to significantly decrease the buffer overflow probability.

Connection admission is a traffic management mechanism which aims at controlling the acceptance of incoming connections so that the required QoS of all connections are provided. One of the key ideas behind Asynchronous Transfer Mode (ATM) is the statistical multiplexing of heterogeneous packetized streams. The concept of Effective Bandwidth is intimately connected with admission control and associated service requirement [7]. The equivalent bandwidth of a connection (source) is a characterization of the connection demanded bandwidth such that its QoS requirements are supported 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 [8]-[9] and on spectral expansion for Markov fluid models [7]) very few results are available for traffic with long-range dependencies [8].

In [10] 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 [11]. In [12], we showed how to use the previously defined framework to derive connection admission management based on realistic assumption.

In this paper, we introduce a simple connection admission policy for video streams. We show that this simple policy produces the same revenue, for moderate to high loads, as the revenue produced by a policy target at maximizing the total revenue. Such policy is a potential good candidate for fast connection admission management.

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 show how to analyse the statistical multiplexing of videio streams, and, in section V, we introduce a policy for fast connection admission management of video streams. Finally, conclusions are drawn in section VI.

2. 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 [13]

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

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 $\Phi(\psi)$ 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 $\Phi(k) \le \varepsilon$. Hence, *k* is given by $k = \sqrt{-2\ln\varepsilon}$

We claim that $(A(t) > \hat{A}(t)) \approx \varepsilon$, 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(t), t = t - s* defines the maximum number of cell arrivals in an interval of size *t*. Therefore, we can choose the fBm envelope process's parameters $\hat{A}_H(.)$ so that it matches the behaviour of *Y*(.).

We extensively validated the effectiveness of the fractal Brownian motion envelope process by utilizing several traces of true network data as well as synthetic traces generated by Mandelbort's procedure [14]. 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.

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.

3. 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 [10]. A rigorous mathematical derivation of the delay bound can be found in [11]. 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

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

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,

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

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

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

We can see that

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

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

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

Therefore,

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

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

We can say that the queue length at time *t* $q(t)$ will only exceed the maximum queue length q_{max} with probability e. 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 $q(t) = A_H(t) - ct \ge 0$ into an easier problem of finding the maximum of a deterministic system described by $\hat{q}(t) = \hat{A}_H(t) - ct \ge 0$.

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 [15] and Duffield [16]. 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.

4. 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:

$$k\frac{1}{2}\left(\sum_{i=1}^{N} \sigma_i^2 t^{2H_i}\right)^{-1/2}\left(\sum_{i=1}^{N} \sigma_i^2\ 2H_i\ t^{2H_i-1}\right)c = c - \sum_{i=1}^{N} \bar{a}_i \quad (6)$$

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:

$$k\frac{1}{2}\left(\sum_{i=1}^{N} \sigma_i^2 t^{2H_i}\right)^{-1/2}\left(\sum_{i=1}^{N} \sigma_i^2 2H_i t^{2H_i-1}\right) - k\left(\sum_{i=1}^{N} \sigma_i^2 t^{2H_i-2}\right)^{1/2} + \frac{q_{max}}{t} = 0 \quad (7)$$

By using Equations 6 and 7 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 6 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}:

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

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

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

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

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

5. A Greedy Policy for Connection Admission Management

Whenever a request for connection establishment arrives, the connection admission controller needs to check if it will be possible to support required QoS, as well as to continue to provide QoS to the existing connection. In order to maximize the revenue, the connection admission controller may collect requests during small time windows, to decide to which request admission should be granted.

In this section, we focus on connection admission management subject to revenue maximization. In our study, we use a common function which takes into account the traffic volume and the duration of a connection. The used revenue function is *Revenue = aT + bV*, where *T* is time duration of a connection and *V* is the traffic volume [17].

One simple approach for connection admission management is to use a greedy policy which accepts connections according to the decreasing order of revenue. This is a simple policy which leads to fast implementation. However, this simple policy may not maximize the total revenue, since a set of connections, each with low revenue, may give a higher total revenue, and consume the same network resources as a single connection with a high revenue. In other words, we need to find a combination of connections that gives the highest revenue subject to available network resources. This is a classic knapsack problem where the knapsack is the available network resource, and the size of each object is a connection network resource demand (bandwidth). This knapsack problem can be state as:

$$max \sum_{i=1}^{M} R_i x_i$$

$$\gamma + \sum_{i=1}^{M} \bar{a}_i t^* x_i + k\left(\sum_{i=1}^{M} \sigma_i^2 t^{*2H_i} x_i\right)^{1/2} \le q_{max} \quad (8)$$

$$x_i \in \{0, 1\}$$

where: R_i - connection *i* revenue;
H_i - connection *i* Hurst parameter;
$\bar{a}_i$ - connection *i* mean arrival rate;
σ_i - connection *i* standard deviation;
t^* - Maximum Time-Scale (MaxTS);
γ - is the network resource already allocated to existing connections, which is given by:

$$\gamma = \sum_{i=1}^{N} \bar{a}_i t^* x_i + k\left(\sum_{i=1}^{N} \sigma_i^2 t^{*2H_i} x_i\right)^{1/2} - ct^*$$

Equation 8 as well as γ are derived from Equations 6 and 7 and express the resources requirements of a set of connections, subject to the required QoS.

To verify the extend to which the simple greedy policy leads to the same result of an exact approach we ran extensive simulations using video traces. Figures 1 and 2 show typical findings. In the numerical examples, we used traces with the following traffic parameters:

Table 1: Traffic parameters of video streams.

	$\bar{a}$	σ	H
A	0.16	1.01	0.67
B	0.12	1.12	0.78
C	0.2	0.9	0.85
D	0.22	0.84	0.91

Note that the mean arrival rate are normalized to the channel capacity. The mean connection duration varies between 20 and 500 seconds, and the buffer size is 1000 ATM cells.

In Figure 1, we show the total revenue as a function of the mean connection inter arrival time. The total revenue is computed by adding the revenue of all accepted connections during the simulation experiment. The exact value of the revenue is immaterial in the sense that it changes according to the choice of *a* and *b*. The lowest the interarrival time, the higher the arrival rate is, and consequently the higher the load. Note that for moderate to high loads the revenue produced by the greedy policy is the same revenue produced by the Knapsack problem. Under low load conditions, this trend cannot be observed.

In Figure 2, we show the total revenue as a function of the mean connection duration, for a normalized utilization of 0.7. We can see that the greedy policy gives the same revenue than the knapsack problem irrespective of the video duration (for moderate and high loads).

6. Conclusions

In video streams, there are the so called long-range dependencies. Such dependencies may cause massive cell loss in ATM multiplexers. Therefore, it is of paramount importance to adopt effective connection admission procedures. In this paper, we introduced a simple policy which can be used for fast connection admission management. This policy accepts connections according to their decreasing order of revenue. We showed that it produces the same revenue, for moderate to high loads, asa policy based on the Knapsack problem which is target to revenue maximization.

Acknowledgements

This work was partially sponsored by CNPq, CAPES and PRONEX SAI.

7. 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] 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.

[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] 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

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

[9] 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

[10] 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

[11] 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.

[12] N.L.S. Fonseca, G. S. Mayo e C. A. V. Neto, Connection Admission Management of Self Similar Traffic, IEEE Latin America Network Operations and Management Symposium, pg 320-332, 1999.

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

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

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

[16] 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

[17] C. Courcoubetis, V. A. Siris and G.D. Stamoulis, “Comparing Usage-Based Pricing Schemes for Broadband Networks”, IEE Colloquium on “Changing for ATM - the Reaity Arrives”, London, UK, Nov 1997.

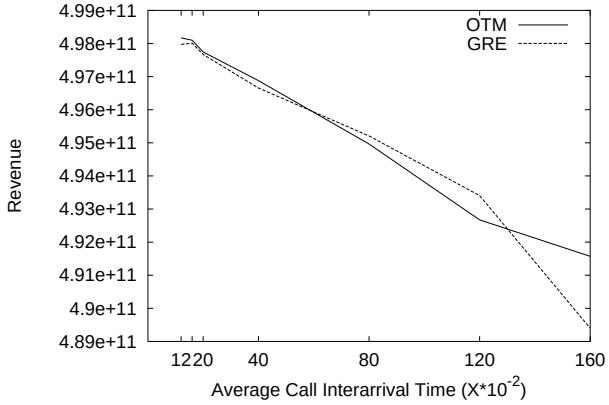

Figure 1: Revenue x Average call Interarrival Time

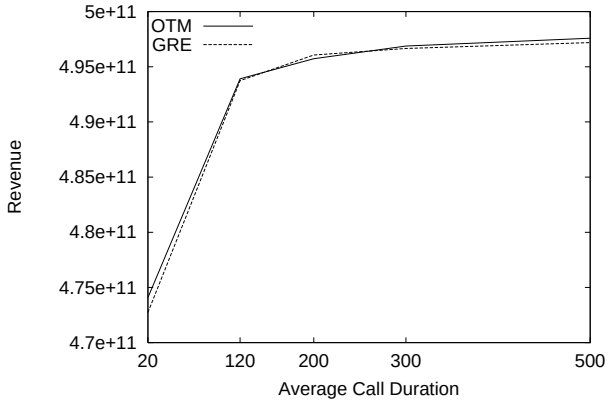

Figure 2: Revenue x Average Call Duration