Computer Communications 25 (2002) 1009–1017

computer communications

www.elsevier.com/locate/comcom

Token bucket characterization of long-range dependent traffic

Gregorio Procissi*, Anurag Garg, Mario Gerla, M.Y. Sanadidi

UCLA, Computer Science Department, Boelter Hall, Los Angeles, CA 90095-1596, USA

Received 1 August 2001; revised 30 October 2001; accepted 28 November 2001

Abstract

The token bucket characterization provides a deterministic yet concise representation of a traffic source. In this paper, we study the impact of the long-range dependence (LRD) property of traffic generated by today's multimedia applications on the optimal dimensioning of token bucket parameters. To this aim, we empirically illustrate the difference between the token bucket characteristics of traffic exhibiting different degrees of time dependence but with identical macroscopic properties (i.e. inter-arrival time and packet size distributions). In addition, we use a statistical model to analytically determine optimal token bucket parameters under various optimization criteria. The statistical model is based on fractional Brownian motion and takes LRD into account. We apply this model to several aggregated MPEG video sources. We then assess the validity of these analytic results by comparing them to empirical results. We conclude that the analytic approach presented here is effective in optimally sizing token buckets for LRD traffic, and promises to be applicable under different traffic conditions and for various optimization criteria. © 2002 Elsevier Science B.V. All rights reserved.

Keywords: Linear-bounded arrival process; Token buckets; Short- and long-memory process; Fractional Brownian motion

1. Introduction

Since the beginning of the 1990s, when the long-range dependent (LRD) nature of actual traffic was discovered [1], a large number of statistical models were developed for characterization of single and aggregated multimedia traffic. Also, the impact of the long-memory characteristic of actual traffic on network performance has been widely investigated since then.

At the same time, a different philosophy to address traffic characterization was first developed in Ref. [2], introducing the notion of linear-bounded arrival process (LBAP). An LBAP process is an arrival process constrained with a linear bound on the total amount of work produced by a traffic source over any time interval. The token bucket technique then permits a complete 'worst-case' characterization of the source using just two parameters (namely the slope of the straight line constraint—the token rate—and the additive constant—the bucket size). For a given source, the set of pairs (token rate, bucket size) that satisfy the LBAP constraint will clearly be infinite and the determination of the minimal set of LBAP pairs, that is, the closure of the aforementioned set, will require off-line processing of the whole trace. Furthermore, the choice of a suitable pair on this curve is arbitrary, though some heuristic criteria can be suitably followed as suggested in Ref. [3].

In this paper, we investigate the impact of long- and short-term-correlated traffic on the token bucket parameter selection. This analysis proves the substantial ineffectiveness of frequently adopted heuristics—rules of thumb—applied to token bucket design based solely on the knowledge of source first-order statistics such as average and peak rates. We then explore the effectiveness of an analytic model that takes into account the long-memory property when choosing the LBAP descriptors for the case of aggregated MPEG sources. It is important to point out that the proposed analysis, based on the use of the fractional Brownian motion (FBM) process, is not a statistical characterization of MPEG sources themselves. Instead, the main goal is to determine whether a model that includes the degree of LRD (measured by the Hurst parameter) in addition to the average rate and the variance of traffic can be effectively used in the token bucket characterization. The motivations for the use of FBM are manyfold. First, the use of FBM process for the description of aggregated sources is justified by the central limit theorem. Secondly, FBM is somewhat analytically tractable and there exist asymptotic results for the queueing behavior with FBM input traffic. Finally, the parsimonious structure of FBM allows the characterization of traffic by just three parameters: mean, variance and Hurst parameter. In particular, the Hurst

* Corresponding author. Present address: Department of Information Engineering, University of Pisa, Via Diotisalvi 2, 56126 Pisa, Italy. Tel.: +39-50-568-661; fax: +39-50-568-522.

E-mail address: gregorio@cs.ucla.edu (G. Procissi).

PII: S0140-3664(02)00015-4

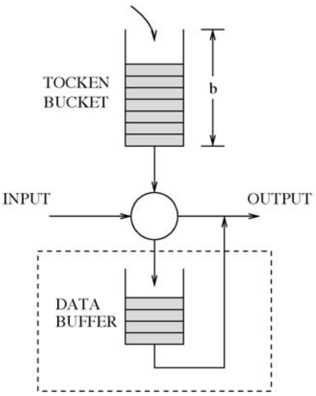


Fig. 1. General scheme of a Leaky-Bucket regulator.

parameter can be varied to achieve different degrees of traffic 'memory', including the case of short-range dependent (SRD) ($H \leq 0.5$). However, our approach does not strictly depend upon the use of FBM for the token bucket description of the input traffic. Different traffic models can be suitably adopted according to the specific kind of sources considered (e.g. SRD Markov-modulated Poisson processes for voice sources [4]).

The rest of the paper is organized as follows. In Section 2, a summary of LBAP processes and a brief description of the token bucket regulator are given. The problem of the optimal selection of token bucket descriptors is given in Section 3 where a number of optimal criteria are proposed in the form of suitable cost functions to be minimized, or predefined delay constraints to be obeyed. In Section 4, the basic concepts of short- and long-memory traffic are summarized. In Section 5, the impact of short- and long-memory property on the LBAP curve is analyzed by comparing the LBAP description of an MPEG video trace which exhibits LRD, to the LBAP description of many shuffled versions of the original trace which exhibit the desired correlation structure. Section 6 is dedicated to the definition of the analytic LRD traffic model for the aggregation of MPEG video sources and to the solution of the optimization problems suggested in Section 3. Finally, in Section 7, the validity of the analytic results is assessed by comparing them to empirical results.

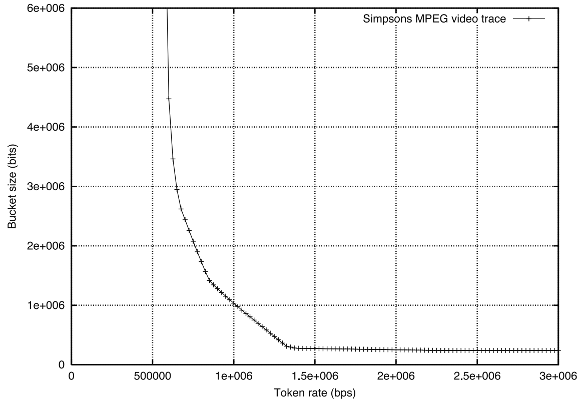


Fig. 2. LBAP curve for an MPEG video trace.

2. LBAP processes and the Leaky-Bucket regulator

An LBAP regulated source constrains the traffic it produces over any time interval τ by a linear function of the time interval. More specifically (for a complete overview, see Ref. [3]), if we denote with $A(\tau)$, the traffic the source transmits over the time interval τ, the traffic is said to be LBAP if there exist a pair (ρ, b) such that

$$A(\tau) \leq \rho\tau + b, \quad \text{for any } \tau > 0, \tag{1}$$

where ρ represents the long-term average rate of the source and b the maximum burst the source is allowed to send in any time interval of length τ. Note that b represents the maximum deviation a source may exhibit with respect to its long-term average behavior.

Operatively, an LBAP (ρ, b) source can be obtained by using a *Leaky-Bucket* regulator (Fig. 1). A leaky bucket builds up tokens of fixed size (typically 1 byte each) at a constant rate ρ in a *token bucket* of fixed size b. The token bucket size b is often referred to as the *token depth*. The source is allowed to send an application layer PDU (or packet if the leaky bucket operates at the network layer) if the bucket contains a number of tokens equal to or greater than the PDU size. If the above condition is not met, the source data can be discarded or delayed in a data buffer until the bucket has enough token to permit its transmission. In the first case, the leaky bucket acts as a *policer*, while in the latter case it acts as a *shaper*.

From a theoretical point of view, the pair (ρ, b) that satisfies the LBAP constraint is not unique. On the contrary, the solutions of Eq. (1) lie in a 2D region which is bounded from below by the curve of the infinite minimum (ρ, b) pairs that obey relation (1). Throughout the paper we will refer to this curve as the LBAP curve, and we will denote it with the equation $\Lambda(\rho, b) = 0$.

As ρ approaches the source average-rate from the right, simple queueing reasoning tells us that the corresponding value of b rapidly increases to infinity, while as ρ increases to reach the source peak-rate, b monotonically tends to zero. Fig. 2 shows the LBAP curve obtained from the MPEG video trace 'Simpson', one of the set of videos used by Rose and available from the web (see Ref. [5] for details).

While all the pairs (ρ, b) lying on the LBAP curve satisfy the LBAP description, the spare degree of freedom can be exploited to optimize some predefined criteria. This idea will be further explored in Section 3. In Ref. [3], Keshav pointed out that for many common real traces, the LBAP curve exhibits a well-defined 'knee area' outside of which, either ρ or b grows quickly for slight variations of the other parameter. This fact suggests as a first suitable choice a pair (ρ^*, b^*) that lies in the neighborhood of the knee region.

Finally, it is worth mentioning that when the peak-rate P

of the source is known, a slightly refined characterization can be obtained, realizing that the total amount of traffic produced by the source in this case is bounded as follows:

$$A(\tau) \leq \min\{\rho\tau + b, P\tau\}. \tag{2}$$

A source that satisfies such constraint is said to be dual leaky bucket regulated, with descriptors (ρ, b, P).

3. Optimal token bucket parameter selection

In Section 2, following Ref. [3] we have addressed the choice of the pair of token bucket descriptors pair (ρ^*, b^*) in a heuristic manner through qualitative arguments about the knee point of the LBAP curve. While this approach is valid, a more formal statement of the problem is developed below.

The first step is to find a proper cost function $\phi(\rho, b)$ to be minimized (or in the dual formulation, a satisfaction function $\Sigma(\rho, b)$ to be maximized) while satisfying the LBAP constraint $\Lambda(\rho, b) = 0$. The resulting point (ρ^*, b^*) is the optimal solution of this constrained min–max problem with respect to the criterion specified by the cost function. If the mathematical expression of $\Lambda(\rho, b) = 0$. were provided, the problem could be solved by the well-known method of Lagrange multipliers. In Section 6, we formulate and solve this problem when the statistical model of the traffic is given. Another important issue is the proper choice of the cost function $\phi(\rho, b)$. The heuristic idea of looking for the 'optimal' point in the vicinity of the knee point, and the decreasing nature of the LBAP constraint leads us to consider as a suitable cost function the distance (or its square) $d(\rho, b)$ of the pair (ρ, b) from the point (m, 0), where m is the average bit-rate of the source. Hence, the cost function to consider is given by

$$\phi(\rho, b) = \|(\rho, b) - (m, 0)\| = \sqrt{(\rho - m)^2 + b^2}. \tag{3}$$

The tight relationship between the token bucket descriptors and the bandwidth, and buffer size allocated in the network to the source suggests another cost function. Depending upon the condition and the topology of the network, either bandwidth or buffer size may have different marginal utility. Network operators might prefer to allocate additional amount of buffers to bandwidth or the other way around. This suggests a cost function defined as a linear combination with arbitrary weighting of the parameters ρ and b, that is

$$\phi(\rho, b) = \omega_\rho(\rho - m) + \omega_b b. \tag{4}$$

The geometric interpretation of this problem is straightforward and is represented by the point at which the LBAP curve is tangent to the constraint line $\omega_\rho(\rho - m) + \omega_b b = k$, where k is the cost to be minimized.

As suggested in Ref. [6], a third strategy to find the token bucket descriptor over the LBAP curve can be derived from analytical known results obtained for a network of *Latency-Rate* schedulers [7]. These schedulers constitute a wide class that include all the work conserving schedulers such as WFQ, WF^{2}Q + , virtual clock, SCFQ, WRR and so forth.

In particular, suppose an amount of bandwidth c and a amount of buffer space B is allocated to a dual leaky bucket regulated source with parameters (ρ, b, P) throughout all the schedulers in the network. In Ref. [7], Stiliadis and Varma proved that for a network of K Latency-Rate schedulers, the maximum end to end delay of packets belonging to the source is bounded by

$$D \leq \left(\frac{P - c}{P - \rho}\right)\left(\frac{B}{\rho}\right) + \sum_{j=1}^{K} \Theta^{s_j}, \tag{5}$$

where Θ^{s_j} is the latency of the jth node that is introduced to take into account the difference between real schedulers and the ideal reference fluid model represented by the generalized processor sharing (GPS) scheduler.

Result (5) is obtained using worst-case analysis and therefore is generally quite conservative. Moreover, the LBAP characterization itself is conservative with respect to statistical modeling of sources. Thus, as a first approximation, the contribution of the latency term can be neglected. Moreover, if the bandwidth and buffer allocated to the source in each node are ρ and b, respectively, condition (5) is met, provided:

$$D_{\max} \leq \frac{b}{\rho}. \tag{6}$$

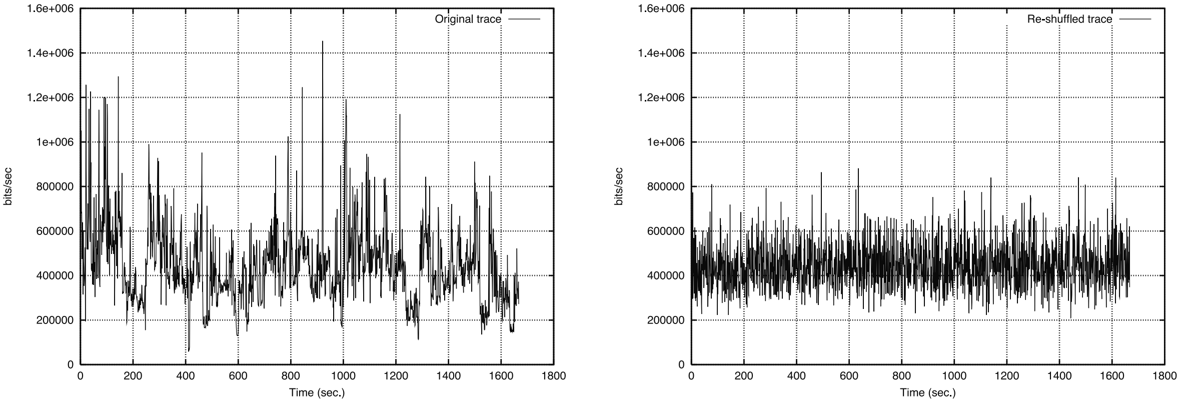
That gives us a new criterion for choosing the (ρ, b) pair as the intersection between the straight line $b = D_{\max}\rho$ and the LBAP curve. It is worth noting that this criterion applies even in the case in which the *peak-rate* is not known and the source is simply leaky bucket regulated [7].

It is also important to note that the trade-off between the two leaky bucket parameters ρ and b is also a trade-off between network utilization and the delay suffered by a flow. A value of ρ much higher than the average bit-rate m of a source will mean that the outgoing link will not be fully utilized as there will not always be a packet to send. On the other hand, a value of ρ closer to m implies a high value of buffer allocated b which will accordingly increase the delay suffered by the flow as the maximum delay $D_{\max}$ is bounded by b/ρ as shown in Eq. (6).

4. Basics of short- and long-memory processes

In this section, the terminology and fundamental concepts of SRD and LRD processes are summarized. For a detailed survey, the reader is referred to Refs. [1,8].

Let $A(t)$ be the total amount of traffic (in units of packets, bytes, or bits) produced by one or several sources over a time interval t. Let us divide the time interval t in non-overlapping time units of length T_u. Let X_n represents the amount of traffic (also called 'work') registered over the nth time unit, that is, X_n is the 'increments process' of $A(t)$. Suppose X_n is a wide sense stationary stochastic sequence with mean



(a) LRD

(b) SRD

Fig. 3. Typical LRD (a) and SRD (b) traffic patterns.

value m and autocovariance function

$$r(k) = E[(X_{t+k} - m)(X_t - m)]. \tag{7}$$

The process X_n is said to be:

- SRD, or with short memory, if

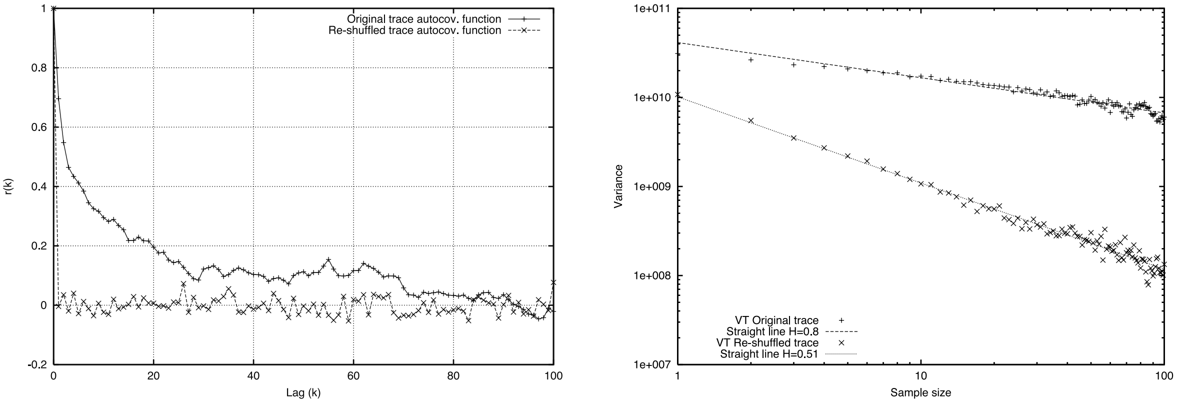
$$\sum_k^{\infty} r(k) < \infty. \tag{8}$$

- LRD, or with long memory, if

$$\sum_k^{\infty} r(k) = \infty. \tag{9}$$

Thus, in an LRD process, the series associated with the autocovariance function diverges. Also, it turns out that the autocovariance function itself decays to zero as k increases following the power form $k^{-\alpha}$, for some $0 < \alpha < 1$. This property determines several visible consequences allowing, through visual inspection or via quantitative analysis, the determination of whether a traffic process had LRD, and the 'degree' of LRD. For the sake of brevity, we only mention the most relevant aspects out of the many that could be outlined. Generally, a long-memory process exhibits relatively long periods where the observations X_n tend to stay at high level, and long periods where the X_n stay at low levels (Joseph effect) [1]. Moreover, a short period analysis reveals cycles or local trends, while the complete time series looks stationary. Also, cycles seem to occur at almost all frequencies. Quantitatively, one can further observe that the variance of the sample mean decays to zero as a power of $0 < \alpha < 1$ of the inverse of the sample size. This evidence is of particular interest as it is the basis of one of the most widely used tools for the detection and estimation of long-memory processes, the so-called variance–time plot (see Ref. [8]). For historical reasons, the Hurst parameter $H = 1 - \alpha/2$, for $1/2 < H < 1$, is used instead of α.

Fig. 3(a) and (b) shows the typical patterns (representing the number of bits per time units of 1 s) of an LRD trace (again the 'Simpsons' MPEG-coded movie) and of an SRD one, obtained by destroying the correlation structure of the 'Simpsons' trace via random shuffling.



(a) Autocovariance

(b) Variance-Time

Fig. 4. (a) Autocovariance functions and (b) variance–time plots for SRD and LRD traffics.

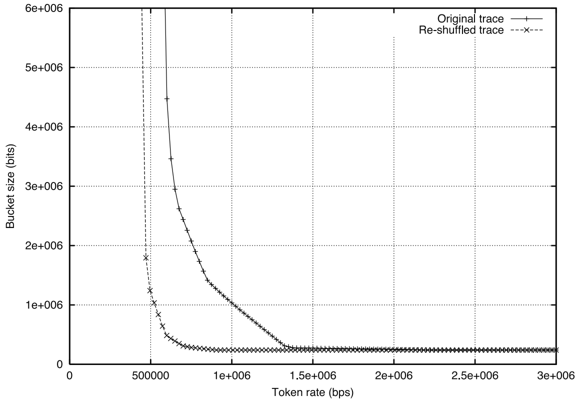


Fig. 5. LBAP curves for uncorrelated and correlated traffic sources.

Fig. 4(a) and (b) shows the autocovariance function and the variance–time plots for the two traces, respectively.

5. Leaky bucket performance with SRD and LRD traffic

The aforementioned characteristics of long-memory processes raise a number of issues about the LBAP description of LRD traffic. In particular, we are interested in the behavior of the LBAP curve in the presence of traffic that has the same macroscopic parameters, average bit-rate and peak bit-rate, but that exhibits correlation dependency over different time ranges.

In order to produce traffic with the desired correlation characteristics, we apply the technique of *shuffling* the traffic traces. This technique consists of randomly reordering the traffic in one of the following ways:

1. *Total shuffling*. The data of the entire trace is reordered randomly, producing a trace in which the packet arrivals are almost totally uncorrelated.
2. *External shuffling*. The original trace is divided into blocks of a given size. The blocks are then randomly reordered, while the order of data within the blocks is preserved. The aim of this method is to destroy the correlation in the packet arrivals beyond the time range determined by the block size.
3. *Internal shuffling*. The original trace is divided into blocks of a given size. Then, reordering of data is performed within each block, while the order of the blocks themselves is preserved. The aim of this method is to destroy the correlation in the packet arrivals within the time range determined by the block size.

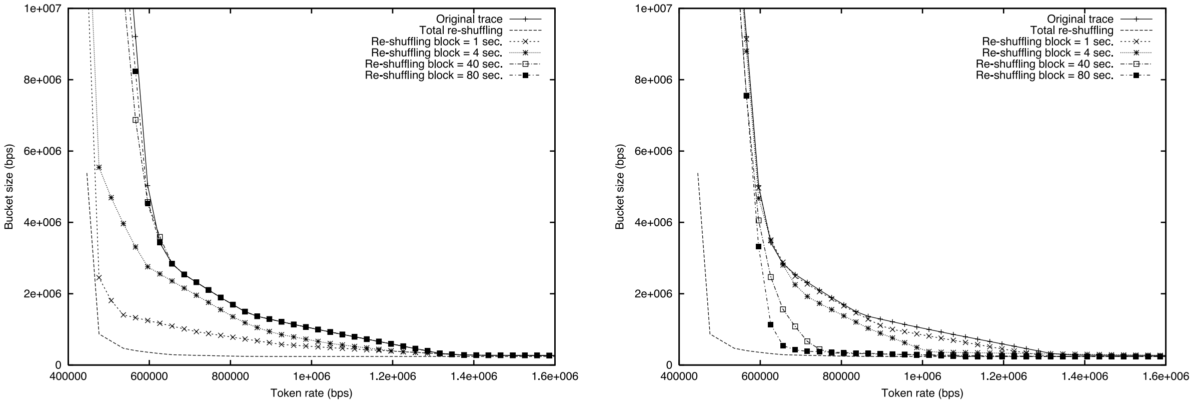
The shuffling procedure does not change the average and peak rates of the sources or the marginal distribution of the data size.

In this section, we apply the shuffling techniques to the MPEG video trace 'Simpsons.trc' whose average rate and the peak-rate are about 0.445 and 5.768 Mbps, respectively, while the observed Hurst parameter of the original trace is about 0.8.

Fig. 5 shows the LBAP curves of the original trace 'Simpsons' and its totally shuffled version. It is evident that the correlation structure has a dramatic impact on the shape of the curve. In particular, the curve associated with the uncorrelated traffic provides an inappropriate description of the original source and the resulting LBAP curve clearly underestimates the original traffic resource requirements.

Fig. 6(a) and (b) show the behavior of the LBAP curves when external and internal shuffling is performed. As a reference, the LBAP curves of the original trace and the totally shuffled trace are also drawn in both the figures.

It is widely accepted (see for example Ref. [9], and references therein) that the range of time for which traffic correlation on a buffered resource is significant is tightly related to the buffer size. Roughly speaking, the larger the buffer size, the larger is the time range to be considered. As an example, in Ref. [10], the authors showed that the most likely time at which a single server queue overflows for an exactly self-similar input process is proportional to the buffer size.



(a) External

(b) Internal

Fig. 6. LBAP curves for shuffled traces: (a) external; (b) internal.

Fig. 6(a) indeed shows that in the case of external shuffling, for small block sizes the LBAP curve diverges significantly from the original trace's LBAP curve, and this phenomenon is more evident for larger bucket sizes. On the other hand, when the block size increases (i.e. correlation is preserved over a longer time range), the LBAP curve is closer to the original trace's LBAP curve even when the bucket size is larger.

Fig. 6(b) proves the same argument from the opposite point of view. Internal shuffling preserves long-term correlation and weakens short-term correlation. As a result, the LBAP curves for the shuffled traces are very close to the original trace's LBAP curve for large bucket sizes, but they diverge increasingly from it as the shuffling block size increases.

6. An LRD model for aggregated video sources

All the techniques summarized in Section 5 require knowledge of the LBAP curve which, in fact, has to be determined off-line. In this section, we develop an analytic model for LRD traffic to be used in the framework of an a priori dimensioning of the token bucket for the aggregated video traffic streams. It is worth pointing out that the goal of this investigation is not the statistical modeling of aggregated video streams itself. The goal of this investigation is rather to develop a model that takes into account the long-memory property of the aggregated traffic to determine optimal bucket parameters for such LRD traffic. To this end, we consider a fluid model for the total amount of work produced by the sources based on the FBM and we derive the corresponding LBAP curve as well as the optimal token bucket pair according to the strategies summarized in Section 3. In Section 7, these results will be compared with results derived empirically aggregating actual video streams under three different scenarios.

Let us assume the following model for the aggregated amount of work produced by the sources in the time interval [0, t]:

$$A(0,t) = mt + B^H(t), \quad t \geq 0, \tag{10}$$

where mt is the mean value over the interval (0, t) and $B^H(t)$ the FBM process (see Ref. [11] for a rigorous definition).

The FBM process is a non-stationary, long-memory process, with Gaussian marginal distribution $\mathcal{N}(0, \sigma^2 t^{2H})$, stationary increments and $B^H(0) = 0$. The parameter H is *Hurst parameter* and takes value in (0, 1); if $H \in (1/2, 1)$ the process exhibits LRD. The amount of work produced over any time interval τ is then given by

$$\begin{aligned} &A(t, t+\tau) = m\tau + B^H(t+\tau) - B^H(t), \\ &A(t, t+\tau) \overset{d}{=} m\tau + B^H(\tau). \end{aligned} \tag{11}$$

Dealing with a stochastic process, the LBAP constraint can be expressed as a bound on the probability of non-conformance. The bound is denoted here by $e^{-\gamma}$ for convenience, as we will show below the solution to be also exponential. The constraint is expressed as

$$P(\rho, b, \tau) = \Pr\{A(\tau) \geq \rho\tau + b\} \leq e^{-\gamma} \text{ for all } \tau, \tag{12}$$

with the corresponding LBAP curve

$$\{(\rho, b) : P(\rho, b, \tau) = e^{-\gamma} \text{ for all } \tau\}. \tag{13}$$

Condition (13) can be reformulated as

$$P(\rho, b) = \sup_{\tau} P(\rho, b, \tau) \leq e^{-\gamma}, \tag{14}$$

where for $P(\rho,\ b)$ there exists an asymptotic expression firstly derived by Norros [12] and later on by means of large deviations techniques by Duffield and coworkers [10,13] given by

$$\log P(\rho, b) \asymp -\alpha(\sigma, H)(\rho - m)^{2H} b^{2-2H}, \tag{15}$$

with

$$\alpha(\sigma, H) = \frac{1}{2\sigma^2} \frac{1}{H^{2H}(1-H)^{2-2H}}. \tag{16}$$

Finally, the LBAP curve can be written as

$$\Lambda(\rho, b) = \alpha(\sigma, H)(\rho - m)^{2H} - \gamma b^{2H-2}. \tag{17}$$

6.1. Determining the optimal bucket parameters

In the following, the analytic expression of the optimal pair (ρ, b) according to the different optimal criteria outlined in Section 3 are derived. The solutions are obtained using the *Lagrange multipliers method*. Namely, we define the function

$$F(\rho, b, \lambda) = \phi(\rho, b) - \lambda\Lambda(\rho, b) \tag{18}$$

and solve the equations obtained by setting the partial derivatives of F with respect to ρ, b and λ equal to zero. We now consider each of the cost functions defined in Section 3, and solve for the optimal token bucket parameter for each function.

Minimum distance cost function. In this case, the cost function is given by

$$\phi(\rho, b) = \sqrt{(\rho - m)^2 + b^2}. \tag{19}$$

The optimal parameters in this case are given by

$$\rho^* = m + \sigma(1-H)\sqrt{2\gamma\left(\frac{H}{1-H}\right)^{H+1}} \tag{20}$$

and

$$b^* = \sigma(1-H)\sqrt{2\gamma\left(\frac{H}{1-H}\right)^{H}}. \tag{21}$$

Table 1
MPEG real traces description

Name	Average bit-rate, *m* (Mbps)	Peak-rate, *P* (Mbps)
Terminator	0.262	1.90
Bond	0.572	2.66
Simpsons	0.446	5.76
Lambs	0.175	3.22
Dino	0.314	2.87
Asterix	0.536	3.54
Starwars	0.223	2.995
Sbowl	0.564	3.38

Note that the optimal pair lies on the straight line

$$b = \sqrt{\left(\frac{1-H}{H}\right)}(\rho - m). \tag{22}$$

The corresponding minimum cost is given by

$$\phi(\rho^*, b^*) = \sigma\sqrt{2\gamma(1-H)\left(\frac{H}{1-H}\right)^H}. \tag{23}$$

Linear cost function. This function with arbitrary weighting of the two parameters is given by

$$\phi(\rho, b) = \omega_\rho(\rho - m) + \omega_b b. \tag{24}$$

The optimal pair is given by the following expressions:

$$\rho^* = m + H\left(\frac{\omega_\rho}{\omega_b}\right)^{H-1}\sqrt{2\gamma\sigma^2} \tag{25}$$

and

$$b^* = (1-H)\left(\frac{\omega_\rho}{\omega_b}\right)^H\sqrt{2\gamma\sigma^2}. \tag{26}$$

In this case, the optimal pair lies on the straight line

$$b = \left(\frac{\omega_\rho}{\omega_b}\right)\frac{1-H}{H}(\rho - m) \tag{27}$$

and the minimum cost is given by

$$\phi(\rho^*, b^*) = \omega_b\left(\frac{\omega_\rho}{\omega_b}\right)^H\sqrt{2\gamma\sigma^2}. \tag{28}$$

7. Numerical results

To test the effectiveness of this approach, we compare the results obtained from the model with those achieved using real traces. In particular, we consider the aggregation of MPEG traces in three different scenarios, varying the number of traces and the traces themselves. Table 1 summarizes the real traces used as well as their average bit-rate and peak-rate.

Using the estimated parameters in the table, we compute the optimal token bucket descriptors for each aggregated traffic, and compare them to the analytic model; the comparison criterion being the effectiveness of analytic and empirical parameters in matching the actual LBAP constraints.

Table 2
MPEG aggregate traces description

Name	Average bit-rate, *m* (Mbps)	Peak-rate, *P* (Mbps)	Hurst parameter, *H*
Aggregate 1	1.46	36.9	0.79
Aggregate 2	1.64	38.8	0.77
Aggregate 3	3.1	45.5	0.79

We produce three different traffic aggregations by multiplexing the first four traces ('Aggregate 1'), the last four ('Aggregate 2') and all eight traces together ('Aggregate 3'). For each aggregated trace, we evaluate all the parameters needed to compare with the theoretical model. Table 2 shows the average bit-rate and peak-rate of each aggregated traffic, as well as its estimated Hurst parameter.

In the following, we will assume unitary weights for the linear cost function (4). A critical aspect of the analysis is the choice of the maximum allowed non-conforming traffic probability: $e^{-\gamma}$. Since the real LBAP curve is drawn for the deterministic constraint (1), the γ parameter is chosen in a range of values close to the absolute value of the natural logarithm of the minimum non-conforming packet probability measurable from actual traces, that is, 1/(trace length). In our traces, as the trace length were given by 160,000 and 320,000 frames, we let γ assumed values from 11 to 13.

Aggregate 1. The first scenario we consider here is the one in which the traces Terminator, Bond, Simpsons and Lambs were multiplexed to get an aggregate trace. Fig. 7 represents the theoretical and empirical LBAP curves and the straight line used to identify the optimal pair under the three different criteria for $\gamma = 13$. The results show that the optimal pairs computed over the theoretical curve are substantially close to those obtained in the empirical LBAP curve. Moreover, they meet the initial condition we defined, that is, they lay on the knee region of the actual LBAP curve.

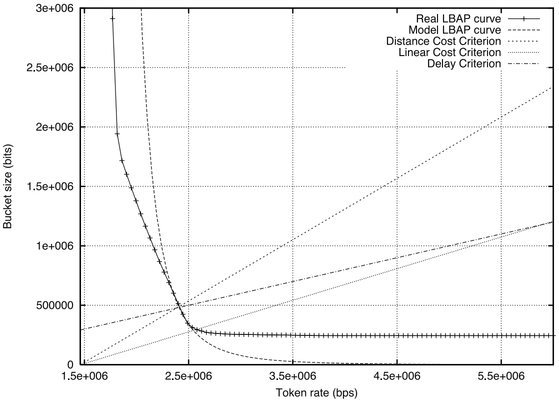


Fig. 7. Model vs. empirical curves for the Aggregate 1 trace.

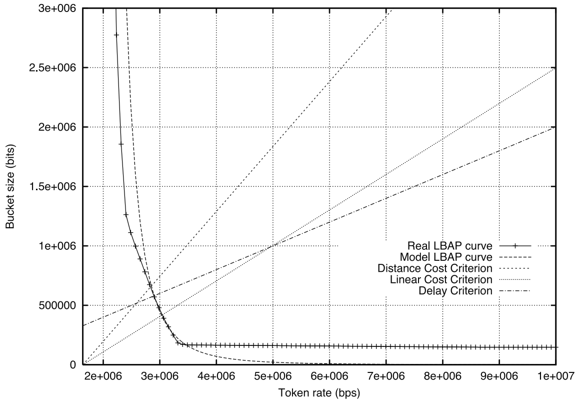


Fig. 8. Model vs. empirical curves for the Aggregate 2 trace.

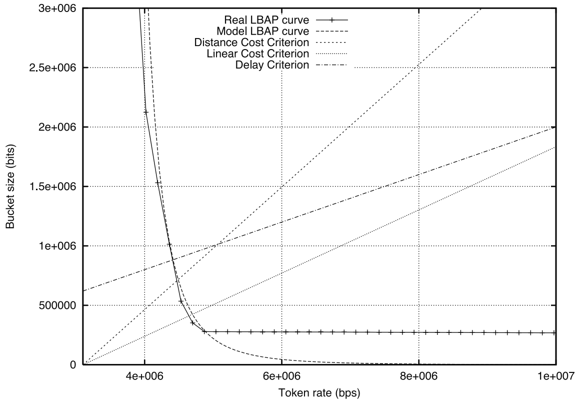


Fig. 9. Model vs. empirical curves for the Aggregate 3 trace.

Aggregate 2. In the second scenario, we multiplexed the real traces of Dino, Asterix, Starwars and Sbowl to get the second aggregated traffic. The result shown in Fig. 8 confirms the effectiveness of the analytical approach. The theoretical optimal points still lie on the knee point of the actual LBAP curve.

Aggregate 3. As a final case we consider the aggregation of all the traces listed in Table 1. As the number of sources increases, we expect that the Gaussian model would get closer to the real LBAP curve. Indeed that happens and the curves (Fig. 9) look much closer than in the first two cases over a larger area, not only in the vicinity of the knee point. Note that after the knee region, the actual LBAP curve stays constant for increasing values of the token rate up to the peak-rate. This is simply due to the packetized nature of actual traffic. The constant flat value of b corresponds to the maximum packet size of the trace. In contrast, in the analytic model, the traffic is treated as a fluid and thus the analytic curve monotonically decays to zero.

In each case, a visual inspection of the results confirms the effectiveness of the analytic model in the definition of the optimal token bucket descriptor. Also, it is interesting to note the beneficial impact of the higher level of multiplexing on the optimal pair computation: a good match between the analytic and the empirical curves is achieved for $\gamma = 11$, while for larger values of γ, the analytic model will produce slightly more conservative results.

8. Conclusions

In this paper, we have investigated the impact of short- and long-memory property of multimedia traffic on the choice of token bucket parameters. Starting from the definition of the LBAP, we compared the token bucket curves drawn for an LRD trace and for a number of modified traces exhibiting correlation over a number of different time ranges obtained by shuffling the original trace. We then developed and used a statistical LRD model based on the FBM process to determine optimal token bucket parameters for aggregated MPEG sources. A number of optimization criteria or cost functions were suggested. This was not done for the purpose of a statistical characterization of aggregated video traffic *tout court*, but with the goal of checking how an analytic model that takes into account the long-memory property could be effective in determining optimal LBAP descriptors within the set of LBAP pairs of interest. The analytic and empirical results show a good match and validate a promising analytic approach for network resource dimensioning.

Acknowledgements

The authors wish to thank Drs R.G. Garroppo and S. Giordano from the NetGroup of the Department of Information Engineering, University of Pisa, for the frequent and stimulating discussions during the preparation of this work. This research was supported by NSF under Grant ANI-9983138 on high-speed networks performance measurements and analysis.

References

[1] K. Park, W. Willinger, Self-similar Network Traffic and Performance Evaluation, Wiley, New York, 2000.

[2] R.L. Cruz, A calculus for network delay and a note on topologies of interconnection networks, PhD Thesis, University of Illinois, July 1987, Issued as report UILU-ENG-87-2246.

[3] S. Keshav, An Engineering Approach to Computer Networking, Addison-Wesley, Reading, MA, 1998.

[4] R.G. Garroppo, S. Giordano, M. Pagano, Estimation of token bucket parameters for aggregated VoIP sources, submitted for publication.

[5] O. Rose, Statistical properties of MPEG video traffic and their impact on traffic modeling in ATM systems, Technical Report No. 101, Institute of Computer Science, University of Wuerzburg, February 1995.

[6] R.G. Garroppo, S. Giordano, S. Niccolini, F. Russo, A simulation analysis of aggregation strategies in a WF^2Q+ schedulers network, IP Telephony, Columbia University, New York, 2–3 April 2001.

[7] D. Stiliadis, A. Varma, Latency-Rate schedulers: a general model for analysis of traffic scheduling algorithms, Technical Report No. UCSC-CRL-95-38, University of California, Santa Cruz, July 1995.

[8] J. Beran, Statistics for Long-memory Processes, Chapman & Hall, London, 1994.

[9] D. Heyman, T.V. Lakshman, Long-range dependence and queueing effects for VBR video, Self Similar Network Traffic Perform. Eval. (2000) 285–318.

[10] N. O'Connell, G. Procissi, On the build-up of large queues in a queuing model with fractional Brownian motion input, Technical Report No. HPL-BRIMS-98-18, HP Laboratories, BRIMS, Bristol, UK, August 1998.

[11] B. Mandelbrot, J. Van Ness, Fractional Brownian motion, fractional Gaussian noise and applications, SIAM Rev. 10 (4) (1968) 422–437.

[12] I. Norros, On the use of fractional Brownian motion in the theory of connectionless networks, IEEE J. Select. Areas Commun. 13 (6) (1995) 953–962.

[13] N.G. Duffield, N. O'Connell, Large deviations and overflow probability for the general single-server queue with applications, Math. Proc. Camb. Philos. Soc. 118 (2) (1995) 363–374.