CALL ADMISSION: DIMENSIONING A SINGLE VBR VIDEO SOURCE USING TRANSITION PROBABILITY MATRIX

Yon Jun Chung, Antonio Ortega, and C.-C. Jay Kuo

Integrated Media Systems Center
University of Southern California
Los Angeles, California 90089-2564

ABSTRACT

This paper describes a novel and simple comparison method for variable bit rate (VBR) video sources according to their average excess rate/frame. We present the parameter average excess rate/frame (α) as an alternative to the traditional parameters used in call admission, e.g. long term rate, variance, peak rate. We propose a way of calculating α through the use of transition probability matrices (TPM) of VBR sources, specifically using their holding times and asymptotic mean occupancies. A useful property of the proposed algorithm is that it allows comparisons between TPMs of different sizes and VBR video sequences of different durations. This algorithm can serve as an aid in call admission for bandwidth allocation purposes.

1. INTRODUCTION

In asynchronous tranfer mode (ATM), every time a source requests a connection, the network performs call admission. The source negotiates with the network for a connection at a certain quality of service (QoS). The source requesting the connection has to share ATM channels with other sources already connected. The network estimates the amount of bandwidth the new connection will consume to decide whether granting this new connection will degrade the service of already connected sources to the point of violating their respective negotiated QoS. This process is referred to as dimensioning a single VBR video source.

The standard specified by the User Network Interface (UNI) [1, 2] envisioned call admission being performed with a set of source parameters, mean, burst (peak rate or (peak rate)/mean), burst duration and others which may be defined by application. However, multiple VBR sources can share common values for parameters such as mean, μ, or peak rate and yet have very different delay behavior. The fault lies in the fact that traditional statistical parameters are poor at revealing information related to delay.

Previous works with VBR sources include an approach where "equivalent" capacity, a function of its long term rate, is used to predict their congestive behavior [3]. Alternatively, dynamic bandwidth allocation was proposed by [5], where the allocation is performed with a TPM whose states are multiples of the VBR source's standard deviation, σ.

While delay is a function of network traffic, the reference to delay in this paper is limited to a single source entering one buffer with a constant output rate. This limited perspective is acceptable for the purpose of dimensioning a single VBR source, as sources that experience greater delay are more likely to increase network traffic than sources experiencing less delay in a single queue. The algorithm presented in this paper computes α from the TPM of the VBR source by focusing on the specific entries of the TPM that are relevant to portions of the VBR source that exceed the output rate.

2. DESCRIPTION OF TPM BASED DIMENSIONING ALGORITHM

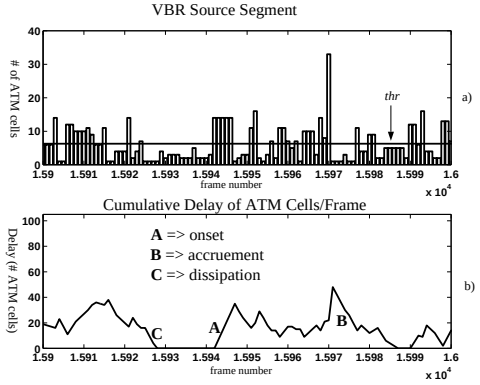

Figure 1: (a) Sample VBR source (b) The delay corresponding to the VBR source in Fig. 1(a).

To compute α for a VBR source, let us examine the actual occurrence of delay. Fig. 1(a) shows a segment of a VBR source trace in time. The horizontal line cutting through the VBR trace represents the buffer output rate, *thr*, for a buffer of infinite size. This is the very same buffer that we will use to measure the delay behavior of this VBR source. It is reasonable to assume that VBR sources with greater buffer occupancy will consume more bandwidth inside the network if placed there unbuffered. Assume that *thr* has been set to the rate negotiated during call admission. Fig. 1(b) shows the delay respective to the VBR source in Fig. 1(a).

In Fig. 1(b), delay starts whenever the VBR trace exceeds *thr* as in point **A**. Notice how this plot continues to increase monitonically until a change in the transmission rate takes place. The VBR trace can now change to a rate above or below *thr*. If the new transmission rate is above *thr*, the plot in Fig. 1(b) will continue to increase albeit at a different slope. This is the case at point **B**. However, if

the VBR trace changes to a transmission rate below thr as in point **C**, the area under the plot will begin to dissipate. Three scenarios (pts. **A**, **B** & **C**) determine the size of the area under the curve. The area under the curve corresponds to the cumulative delay for the VBR source. Points **A**, **B** & **C** are referred to respectively as onset, accruement and dissipation of delay.

We are interested in caculating the area under the curve. We now define the cumulative sum of these areas divided by the total number of frames in the VBR sequence as α, the average excess rate/frame. Our intention is to calculate α using these three scenarios of delay. In all three scenarios, there is a change in the transmission rate. We choose to model the VBR video sources as Markov processes because they are good at representing changes (state transitions) and the holding times of states are easily calculable.

A process is Markov if its current state is solely dependent on its previous state. Researchers in [4, 5, 6] have observed this Markov property in their dealings with VBR video. A non birth-death, first order Markov process can be characterized by a TPM **P**, a square matrix whose elements consist of $p_{i,j}$ defined as

$$p_{i,j} = \frac{\text{number of transitions } i \text{ to } j}{\text{number of transitions out of } i}$$

If we can match a VBR source to a TPM, then the TPM can be used to calculate α. Previously, Heeke used a VBR video TPM in [4] to perform policing function. He used the TPM along with empirical holding and recurrence time to modulate the durations a VBR source is allowed to transmit at predetermined tramission rates. Our algorithm could be used to perform admission control for a VBR source that is then policed as in [4].

For VBR video sources the states are represented by transmission rates, $r = \begin{bmatrix} r_1 & r_2 & \cdots & r_k \end{bmatrix}$. The equilibrium state probability $\Pi = \begin{bmatrix} \pi_1 & \pi_2 & \cdots & \pi_k \end{bmatrix}$ can be calculated by solving Eqns. (1,2) below,

$$\Pi = \Pi * \mathbf{P} \quad (1)$$

$$\sum_i \pi_i = 1. \quad (2)$$

Eqns. (1,2) and the matrix **P** provide a wealth of information. For example, given that the VBR sequence is N frames long, statistically it may be assumed that there are $N\pi_i$ frames transmitted at rate r_i. Furthermore of those $N\pi_i$ frames, there are $N\pi_i p_{i,j}$ transitions from transmission rate r_i to r_j.

Let us now focus on those elements in the TPM that correspond to the three regions of delay. For a 5x5 TPM with a given buffer output rate thr, where $r_3 < thr < r_4$, the three regions correspond to the elements in **P** shown in Fig. 2, where $a_{i,j}, b_{i,j}$, and $c_{i,j}$ are the elements in **P** which correspond to the regions marked by points **A**, **B**, and **C** respectively in Fig. 1(b). The diagonal elements $p_{i,i}$ marked "x" are excluded because they play a role in calculating the average holding time in Eqn. (4). Including those diagonal elements $p_{i,i}$ in matrix **D** would introduce redundancies. Denote the version of **P** in Fig. 2 with the elements "x" and "%" being zero as **D**.

The average holding time for state i expresses the average time the VBR source spends transmitting at r_i once it

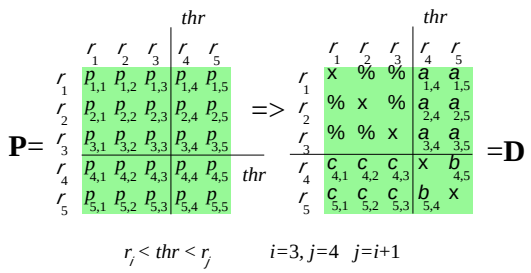

Figure 2: Matrix **D** represents the elements of TPM **P** that correspond to the three regions of delay, **A**, **B** and **C** in Fig. 1(b) for a 5x5 matrix.

enters state i from a different state. [7] shows the holding time for state i, h_i, to be

$$h_i = \frac{1}{1 - p_{i,i}} . \quad (3)$$

We can now combine $a_{i,j}$, $b_{i,j}$ and $c_{i,j}$ in **D** with their respective holding times to extract the delay behavior of a VBR source. For example, given a transition from any state other than 4 to state 4, the sequence will remain in state 4 for an average duration of h_4. In our 5x5 TPM example, $a_{i,4}$ and $b_{i,4}$ correspond to transitions into state 4. Now, r_4 is greater than thr by the amount $r_4 - thr$, thus the delay caused by transitions into state 4 corresponding to the elements $a_{i,4}$ can be expressed as $N\pi_i a_{i,4} (r_4 - thr) h_4$. Similarly for transitions into a state with a transmission rate lower than thr, say r_j, the amount of delay dissipation can be expressed as $N\pi_i c_{i,j} (thr - r_j) h_j$. The three regions of delay can be summed up by matrix multiplication to get an expression for the net delay:

$$\xi = N\pi * \mathbf{D} * \mathbf{q}^t \quad (4)$$
$$\text{where } \mathbf{q}^t = [(r_i - thr)h_i]\,|_{i=\{1,\ldots,5\}} .$$

Note that the terms from the delay dissipation region **C** subtract from ξ. Eqn. (4) accounts for all three regions **A**, **B** and **C**, but there are some correction terms left to consider.

If the equilibrium state probability of, say, state f equals zero, it does not mean that the sequence never enters state f. A zero equilibrium state probability implies only that the number of frames that transmit at rate r_f is statistically insignificant. However from delay's point of view, if the state following state f causes a significant amount of onset, accruement, or dissipation of delay then it cannot be ignored. For example, a two hour movie formatted at 30 frames/sec contains more than 200K frames. At such a length, it is reasonable to approximate the number of VBR video frames as infinite. Under this assumption, $\vec{t}$, the asymptotic mean occupancy can be calculated. Given an infinitely long sequence whose states f and m have zero equilibrium state probability, then there are $\vec{t}_{f,m}$ transitions from state f to state m, where $\vec{t}_{f,m}$ is defined below,

$$\left[\vec{t}_{f,m}\right] = \left([I - \mathbf{P}^*(\mathbf{z})]|_{z=1} \right)^{-1} = [I - \mathbf{P}^*]^{-1} . \quad (5)$$

$\vec{t}_{f,m}$ is the mean number of times state m will be entered in

an infinite number of transitions from state f [7]. $\mathbf{P}^*$ is the result of eliminating the rows and columns in $\mathbf{P}$ corresponding to recurrent states, i.e., those with non-zero equilibrium state probability.

Finally, the correction factor from asymptotic mean occupancies equals,

$$correc = \sum_{\vec{t}_{f,m} \in \mathbf{A},\mathbf{B},\mathbf{C}} \vec{t}_{f,m}\, q_m \quad , \tag{6}$$

where the $\vec{t}_{f,m}$ terms only apply if states f and m are part of the three regions of delay.

We now propose our new parameter, α, which represents the average excess delay per frame.

$$\alpha = \frac{\xi + correc}{\mathrm{N}} \quad . \tag{7}$$

As shown, α equals the net delay added to the correction factor normalized by N, the total number of frames. Note that when using α to compare two VBR sources, their respective TPMs do not have to be of the same dimension nor use the same transmission rates $r = \left[\begin{array}{cccc} r_1 & r_2 & \cdots & r_k \end{array} \right]$.

3. EXPERIMENTAL RESULTS

We envision an alternative QoS negotiation being performed with TPM rather than those parameters specified in [1, 2]. In this experiment we demonstrate that αs both from model and simulation possess a greater discriminating ability over other parameters such as μ, $\mu + \sigma$, and peak rate. These five parameters are assigned to discriminate two hundred different test VBR simulations according to their 99% delay threshold. The 99% delay threshold is a good way of quantifying the delay behavior for a single source. This means if the test VBR sources to be compared are placed into their individual buffers with output rates uniformly set to thr, then the 99% delay threshold states that 99% percent of the time the content of the buffer will be less than this threshold. In other words, if delay curves similar to Fig. 1(b) are plotted for each test VBR simulation then 99% of the curve would lie under its respective 99% delay threshold.

	mod. α	sim. α	μ	99% delay	$\mu + \sigma$	peak
θ	0.2522	0.3578	5.39	110	10.77	36

Table 1: Parameter values for the sample sequence θ in units of ATM cells/frame.

The test VBR simulations are generated from Markov modulated poisson process (MMPP) models. First, we generate a VBR MPEG video sequence at 0.5 Mbps and then quantize the resulting rates/frame to five levels. These five quantization levels are chosen to represent the transmission rates of the MPEG video. To better match the characteristics of the MPEG video, we choose the quantized rates with an algorithm akin to the Lloyd-Max algorithm used in optimal quantization. The quantization levels are selected to minimize the mean squared error (MSE) between the quantized and the original MPEG transmission rate values. These quantization levels now serve as the states of the MMPP model. MMPP are coupled to one TPM and used to generate one test VBR simulation. The model α are calculated from this TPM. This process is repeated with two hundred different TPMs and the end result is our two hundred test VBR simulations. Each TPM is a 5x5 matrix whose individual entry values range from 0 to 1 with stepsize 0.1, i.e. 0, 0.1, 0.2, ... Each test VBR simulation contains 20000 frames (about 11 min). Their long-term mean, μ, varies between 2.0 and 7.0 cells/frame.

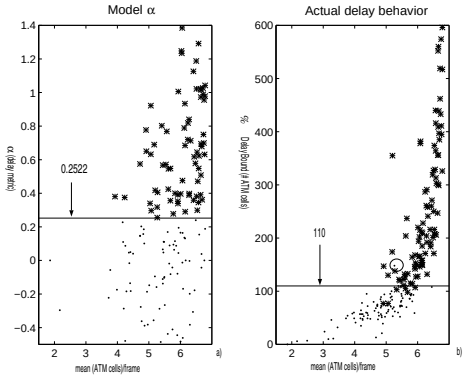

Figure 3: Comparison between dimensioning using α vs. the delay behavior actually measured on a 99% threshold.

The 99% delay threshold for each test VBR simulation at buffer output rate of $thr = 7.0$ cells/frame is then recorded. These recorded values serve as the control in our experiment. The main purpose of this experiment is to see if the correlation between α (from both the model and the simulation) and the 99% delay threshold for a given set of test VBR simulations is stronger than that of the other three parameters ($\mu, \mu + \sigma$, peak). A single test VBR simulation θ, whose parameter values are listed in Tab. 1, is picked to serve as the cutoff for the whole set of test VBR simulations.

Each point in Fig. 3 corresponds to one VBR simulation 20000 frames long. These test VBR simulations are first separated along their α value in Fig. 3(a). Points marked by "." correspond to VBR simulations whose α value is lower than that of our cutoff, θ. Those marked by "*" correspond to VBR simulations that are greater. Using the same markings, the 99% delay threshold corresponding to each test VBR simulation is plotted in Fig. 3(b). The horizontal line match the parameter values for θ listed in Tab 1. Ideally, the right side plots should have all the "." points lying below the horizontal lines and the "*" points above the same line. This would validate that the dimensioning is accurate. The points circled represent the worst dimensioning error.

The α values in Fig. 3(a) are from the two hundred TPMs used in the MMPP model. For each VBR test simulation, the α is recalculated to see if it corroborates the performance of its model α. However instead of re-using the same TPM from its MMPP model, we recalculated a new TPM based on the observed values of the MMPP simulations. First the transmission rates, $r = \left[\begin{array}{cccc} r_1 & r_2 & \cdots & r_k \end{array} \right]$, which represent the states of the new TPM are obtained in the exact same manner as described earlier for the MPEG sequence. These newly quantized rates are then used to

generate the new TPM which ultimately yields the simulation α.

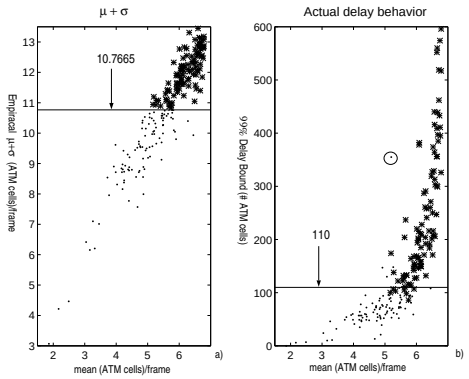

Figure 4: Comparison between dimensioning using σ plus μ vs. the delay behavior actually measured on a 99% threshold.

Observe the right-hand plot of Fig. 3, notice the "." points lying above the 110 cutoff and the "*" points lying below. These points are dimensioning errors and the absolute difference between these error points and the 110 cutoff are tabulated in Tab. 2. The first row lists the number of dimensioning errors followed by the maximum dimensioning error in the second row. The last row lists the average error. Notice that our proposed α has by far the lowest average and maximum error. Both our model and simulation αs, while not free of errors, show a performance improvement over the other three parameters. The parameter $\mu + \sigma$ is closest in performance to our α and its result is shown in Fig. (4).

dim error (ATM cells)	mod. α	sim. α	μ	$\mu + \sigma$	peak
# occur.	19	15	30	15	57
maximum	44	90	245	245	450
average	16.21	25.13	33.77	29.2	58.95

Table 2: Dimensioning error for the various parameters: the number, maximum and average of the absolute difference between the error points and the threshold.

This brings us to a key point. Delay behavior of a VBR source should not only depend on the source itself but on the magnitude of the negotiated rate. If the 200 VBR simulations are ranked using the parameters μ, $\mu + \sigma$ or peak, the ranking of these VBR simulations does not change with the rate negotiated between the network and each source. This implies that once a source is deemed to be more congestive than another, it must also be so at all negotiated rates. However if we use the definition of delay as the area under a cumulative delay plot such as Fig. 1, then there exist two hypothetical VBR sources, VBR1 and VBR2, where the delay for VBR1 is greater than VBR2 at one rate where the reverse is true at a another rate. α, on the other hand, is a function of both the VBR source and the negotiated rate thr. All 200 VBR simulations are compared using α at an uniform negotiated rate thr and the ranking among the VBR sources would change if compared at another rate.

To strengthen earlier results, the experiments illustrated in Tab. 2 were repeated for 200 different VBR simulations, narrower in range $6.5 > \mu > 6.75$ and tighter in bound 99% vs. 99.9%. Tabs. (2, 3) illustrate the need not to single out one specific transmission rate, but to couple the magnitude of those transmission rates higher than thr along with their holding times because this is where delay onsets, accrues and dissipates. There is no absolute measure that accurately reflects delay behavior independent of bandwidth. For this reason, VBR sources must be compared with a common value for the bandwidth thr.

$6.5 > \mu > 6.75$ and 99.9% delay threshold

dim error (ATM cells)	mod. α	sim. α	μ	$\mu + \sigma$	peak
# occur.	84	12	73	65	66
maximum	481	167	686	765	809
average	80.19	72.66	84.07	113.2	134.3

Table 3: Same as Table 2 with 200 different VBR sources with narrower μ range and tighter delay bounds

4. CONCLUSION

In this paper, we introduce the notion of average excess rate/frame, α and demonstrate its improvement over other parameters traditionally used in dimensioning VBR sources. The proposed algorithm makes use of strategic elements of transition probability matrices that reflect excess rate/frame. The work presented in this paper can be extended to multiple buffers and eventually incorporated with the multiple leaky bucket method presented in [8].

5. REFERENCES

[1] "The ATM Forum." Prentice-Hall, 1993. ATM user-network interface specification, V 3.0.

[2] "Traffic control and congestion control in B-ISDN." Intl. Telecomm. Union, March 1993.

[3] R. Guerin, H. Ahmadi and M. Naghshineh, "Equivalent capacity and its applications to bandwidth allocation in high-speed networks," *IEEE JSAC*, vol. 9, pp. 968–981, Sept. 1991.

[4] H. Heeke, "A traffic-control algorithm for ATM networks," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 3, pp. 182–189, June 1993.

[5] P. Pancha and M. El Zarki, "Bandwidth-allocation schemes for variable-bit-rate MPEG sources in ATM networks," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 3, pp. 190–198, June 1993.

[6] P. Skelly, M. Schwartz, and S. Dixit, "A histogram-based model for video traffic behavior in an ATM multiplexer," *IEEE/ACM Transactions on Networking*, vol. 1, pp. 446–458, August 1993.

[7] R. Howard, *Dynamic Probabilistic Systems vol.I: Markov Models*, John Wiley & Sons, 1971.

[8] A. Ortega and M. Vetterli, "Multiple leaky buckets for increased statistical multiplexing of ATM video," *Packet Video Workshop, Portland, OR*, September 1994.