A new estimation method of video traffic specification in QoS-guaranteed networks

T.C. Thang and Y.M. Ro

Abstract: Traffic specification plays a crucial role in the resource reservation for video services over the packet-switching networks. The current development of QoS-guaranteed service still leaves a wide space for the selection of traffic specification. We propose a new method to estimate the traffic specification of variable-bit-rate (VBR) video for deterministic service. The method is based on the concept of empirical envelope and the delay bound. The solution shows to be simple yet it provides excellent network utilization.

1. Introduction

Integrated-service networks need to accommodate diverse traffic characteristics and QoS requirements. Among various classes of traffic, video transmission over packet-switching networks has been an interesting topic for a decade. Two important features of VBR video traffic are the burstiness and the delay-sensitiveness, which make it difficult to stream video content on packet-switching networks.

To provide QoS guarantee for video services, the networks must reserve resources for each connection. The amount of resources reserved for a connection is dependent on the QoS requirements (e.g. delay, loss) and the traffic specification that is some characterization of the traffic from the source of the connection. The traffic specification is also referred as traffic descriptor (in ATM) or Tspec (in guaranteed service of IETF) [1].

Normally, traffic specification is represented in the form of one or a series of token buckets ((σ,ρ) and $(\vec{\sigma},\vec{\rho})$ models) where σ is the bucket size and ρ is the bucket rate [2]. Specifically, traffic specifications of guaranteed services of ATM and IETF consist of a token bucket (σ,ρ) plus a peak rate p, which is actually in the form of two token buckets: $(0,p)$ and (σ,ρ).

However, the mapping from the real traffic to some specification parameters is still an open issue [8]. Traffic specification can be estimated by various methods. A large number of methods are based on stochastic models of video sequence [9][10][11]. These methods have the advantage that higher network utilization can be achieved using statistical multiplexing. However, they have some significant disadvantages as well [2]. First, most stochastic models are either not powerful enough to capture the burstiness of video source, or they are too complicated for practical implementation of call admission control.

As opposed to stochastic models, a number of methods have been proposed based on the deterministic traffic model [2][3][4], which was initiated by Cruz [6]. The traffic specification is estimated as an approximation of the so-called *empirical envelope*, resulting in a worst-case characterization of the video traffic. The services developed in this track are often called deterministic services. In practice, the guaranteed services of ATM and IETF belong to this category.

The traffic specification can be determined further by various extensions such as the trial and error approach [8], the method based on some constraints of σ and ρ [12], the method based on deterministic model but also exploiting the periodic pattern of video frame sizes for statistical multiplexing [4][5] etc.

The usefulness of a traffic specification is ultimately evaluated by the network utilization, which is essentially proportional to the maximum number of concurrent connections accepted by the call admission control. Normally, a more complicated traffic specification gives a more accurate characterization of the source, thus resulting in higher network utilization. However, the estimation of the traffic specification is usually independent of the QoS requirements, for example, the delay bound in deterministic services, that is, the traffic specification is fixed for all sessions and may be computed and stored in advance.

In this paper, we propose a new method to estimate the traffic specification of deterministic service. The method is based on the point-of-view of the call admission control, which takes the delay bound requirement into the estimation process. The method is efficient in the sense that the resulting traffic specification is very simple (just one pair of (σ,ρ)) while still providing excellent network utilization. As the traffic specification of the proposed method is only computed right before the transmission, we will show a procedure by which the estimation process can be performed in a very short time.

The remainder of the paper is organized as follows. In section 2, we review the concept of empirical envelope, which is fundamental in building a traffic model for deterministic service. In section 3, we present the delay bound tests of call admission control, from which a new method to estimate the traffic specification is derived. To evaluate the performance of the proposed method, various simulations are presented in section 4. Finally, section 5 will have the conclusion of the paper.

2. Traffic specification based on empirical envelope

In this section, the concept of empirical envelope is described. Knowing the empirical envelope, we can find some traffic specification that is capable of characterizing the worst-case traffic of a connection.

Let's denote the actual traffic of a connection by a function A where $A[\tau, \tau+t]$ represents the cumulative traffic arrivals in the time interval $[\tau, \tau+t]$. An upper bound on A can be given by a function $A^*(t)$ if for all times $\tau \geq 0$ and all interval lengths $t \geq 0$, the following holds [2] [6]: $A[\tau, \tau+t] \leq A^*(t)$

Any $A^*(t)$ satisfying this property is called a traffic constraint function.

The empirical envelope is defined as:

$$E^*(t) = \max_{\tau \geq 0} A[\tau, \tau + t]$$

This expression shows that $E^*(t)$ is the tightest the time-invariant bound on A at any interval of length t, and every traffic constraint function $A^*(t)$ satisfies $A^*(t) \geq E^*(t)$ for all times t. So, the empirical envelope is the most accurate traffic constraint function for an arrival function A. Figures 1a and 1b show an example of a MPEG video sequence and the corresponding cumulative arrivals and empirical envelope.

If the empirical envelope is employed as the traffic specification, the highest network utilization will be achieved. However, the empirical envelope is usually too complicated for a practical implementation. The solution is to find some parameterized approximation of the empirical envelope.

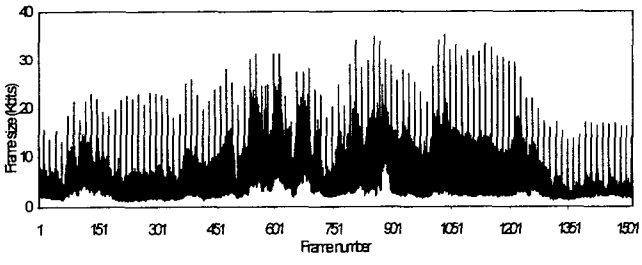


Fig. 1a: An example of MPEG video sequence

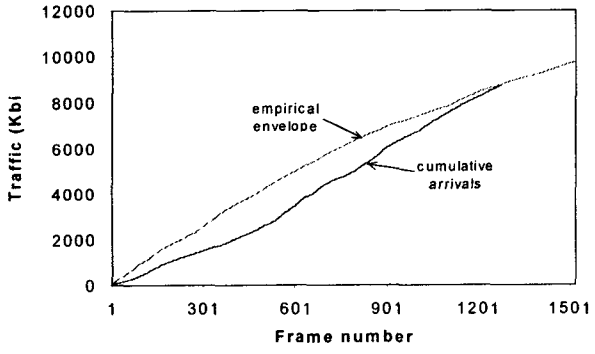


Fig. 1b: The cumulative arrivals, empirical envelope of the above sequence

The empirical envelope may be approximated by a concave upper poly-line (abbreviated as poly-line hereafter) that can be represented by the $(\vec{\sigma}, \vec{\rho})$ model. An algorithm to find the poly-line is given in [2]. Essentially, the poly-line covers all convex regions of empirical envelope, but all vertices of the poly-line still lie on the empirical envelope. It should be noted that the empirical envelope and its resulting poly-line are unique for each arrival function.

For deterministic service, the traffic specification can be found by selecting directly one or several segments of the poly-line [2]. In [3], the traffic specification is also determined further by some heuristic approximations of the poly-line based on user-defined criteria. In both cases, the traffic specification of four or five (σ,ρ) pairs can give performances close to that of the empirical envelope. As mentioned above, in practice some parameterized traffic constraint functions, usually consisting of one or two pairs of (σ,ρ), are employed as traffic specifications for the purpose of simplicity. While the solutions consisting of four or five (σ,ρ) pairs have excellent performance, an acceptable solution with two, especially one, pairs of (σ,ρ) is rather difficult to find.

In our method, we will focus on the simple model of just one (σ,ρ) pair, and we will show how it can be related to the $(\vec{\sigma}, \vec{\rho})$ model. Traffic specification of (σ,ρ) model is simply a straight line $l(t)=\sigma+\rho t$ with $l(t)=0$ for $t<0$. Figure 2 illustrates an arrival function, its corresponding empirical envelope and poly-line together with a constraint line $l(t)$ which can be a candidate for the traffic specification. As a traffic constraint function, the straight line $l(t)$ is always above the empirical envelope.

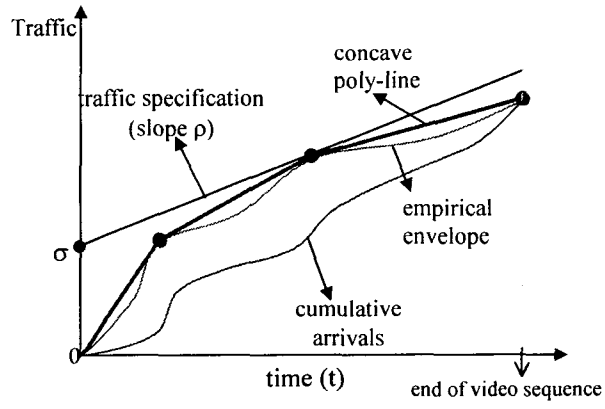


Fig. 2 Illustration of arrival function, empirical envelope, poly-line and traffic specification

3. Call admission control and the estimation of the traffic specification

In order to decide whether a connection request is acceptable or not, the call admission control has to perform a number of tests such as availabilities of the

link capacity, buffer size, CPU usage, etc. Among these tests, the delay bound test is the most important for a network providing deterministic services [2]. The delay bound test verifies that, for all connections, the delay of each packet is less than a required delay bound.

The delay bound tests of the call admission control are different from packet scheduler to packet scheduler. A large number of packet schedulers have been proposed to support integrated services in the packet-switching networks. Some typical packet schedulers are first-come-first-served (FCFS), static priority (SP), rotating-priority-queues (RPQ) and earliest-deadline-first (EDF). The FCFS scheduler is the simplest one, which transmits all packets in the order of arrival. All connections in an FCFS scheduler have identical delays. Other types of schedulers are designed to support a large number of connections with diverse delay requirements. The EDF scheduler was shown to be the optimal one [2]. The formula of delay bound test for FCFS scheduler is as follows [2][7]:

$$d \geq \frac{1}{C}\sum_{j=1}^{N} b_j(t) - t + \max_{k \in N} s_k \qquad \text{for all } t \geq 0 \qquad (1)$$

and for EDF scheduler:

$$t \geq \frac{1}{C}\sum_{j=1}^{N} b_j(t-d_j) + \max_{d_k > t} s_k \qquad \text{for all } t \geq \min_{j \in N} d_j \qquad (2)$$

where

d_j is the accepted delay of connection j; for FCFS, $d_j = d, \forall j$

$b_j(t)$ is the traffic specification of connection j

C is the link speed

N is the number of connections

s_k is the maximum packet size for connection k

As mentioned above, the usefulness of a traffic specification can be measured as the maximum number of concurrent connections accepted by the call admission control. This measurement is often examined in the homogeneous case, that is, the same streams with the same delay requirements. In this case, those different schedulers produce the same schedule [2], and the delay tests can be reduced to the simple formulae of FCFS as follows [4]:

$$N(d) = \max\left\{ n \;\middle|\; \frac{1}{C}\max_{t \geq 0}\left\{\sum_{j=1}^{n} b_j(t) - Ct\right\} \leq d \right\} \qquad (3)$$

From (3) we see that, given a value of d, if $b_j(t)$ is reduced, then N will be increased. In other words, more streams will be accepted. The problem here is how to determine $b_j(t)$.

The test (3) can be changed into finding the maximum value of N that satisfies:

$$\sum_{j=1}^{N} b_j(t) - Ct \leq Cd \qquad \text{for all } t \geq 0 \qquad (4)$$

$$\text{or} \quad \sum_{j=1}^{N} b_j(t) \leq C(t+d) \qquad \text{for all } t \geq 0 \qquad (5)$$

In case of homogenous streams, $b_j(t) = b(t)$ for all j, (5) is equivalent to

$$N*b(t) \leq C(t+d) \qquad \text{for all } t \geq \qquad (6)$$

$$\text{or} \quad b(t) \leq (C/N)(t+d) \qquad \text{for all } t \geq 0 \qquad (7)$$

From (7) we see that, given a certain value of d, if N is maximum then (C/N) is minimum, and vice versa (at this point we suppose N can be a non-integer). Also $b(t) \geq E^*(t)$ for all $t \geq 0$. So, the minimum (C/N) is the slope of the line $(C/N)(t+d)$ that goes through $(-d,0)$ and touches $E^*(t)$ at only one point.

In turn, *let b(t) be selected as that tangent line*, then (C/N) will be able to achieve the minimum value. Because from a point $(-d,0)$, there is only one tangent line to $E^*(t)$, the selected $b(t)$ is unique for a given d. This is illustrated in Figure 3.

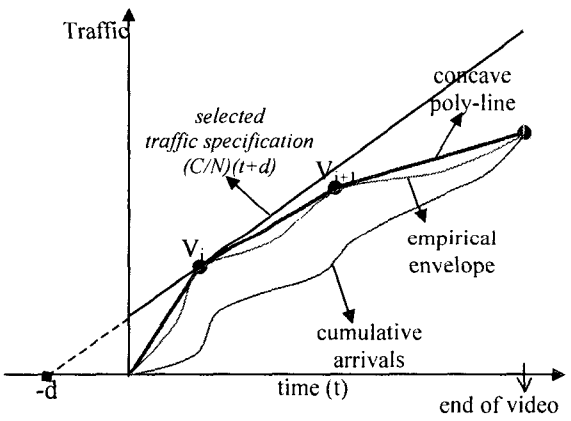


Fig. 3 Illustration of the selected traffic specification

From point $(-d,0)$, finding the tangent line to $E^*(t)$ could result in high computational complexity. However, taking advantage of the parameterized poly-line, the finding process would be very short time. We have the following lemma:

Lemma: Among all lines that start from $(-d,0)$ and go through the vertices of the poly-line, the line having the highest slope is the tangent line to E*(t).

Proof: Assume that the line l_i, going through vertex V_i (Figure 3), is the line having the highest slope. If l_i is not the tangent line to $E^*(t)$, it will cut the poly-line at vertex V_i. Because the poly-line is concave, line l_i will be below the vertex V_{i+1}, which means that the line l_{i+1} going through V_{i+1} will have a slope higher than that of l_i. That is, l_i does not have the highest slope, which

contradicts with the first assumption. So the line having the highest slope is the tangent line to $E^*(t)$.

From the above presentation, we can see that in both cases: $b(i)$ equal to the selected traffic specification or $b(i)$ equal to the empirical envelope, the slopes of tangent lines are the same, so the performance of the selected traffic specification is the same as that of the empirical envelope. This traffic specification is very simple since it is just a straight line, or just one pair of (σ,ρ). Meanwhile, its slope (C/N) is minimum, so that the number of accepted connections, N, is maximum. This means that no other traffic specifications, even the more complicated ones of $(\vec{\sigma},\vec{\rho})$ model, can have better performance.

As a result, the estimation procedure is performed as follows:

Step 1: compute empirical envelope.

Step 2: compute the poly-line (as in [2]).

Step 3: determine the acceptable delay bound d.

Step 4: draw the lines between point $(-d,0)$ and vertices of poly-line.

Step 5: find the line with the highest slope, which is the selected traffic specification.

In essence, this proposed method differs from other methods in a way that its traffic specification is not fixed for all sessions. Instead, the traffic specification is determined just before the transmission by taking into account the predefined delay bound.

The overhead of the just-in-time computation is very short because the empirical envelope and the poly-line can be computed in advance and the number of vertices of the poly-line is usually small (about several dozen points). In addition, the duration of estimation process, that is the time for searching the tangent line, can be further reduced by some heuristic searching algorithms. In the above procedure, steps 1 and 2 may be performed off-line for each video stream. Steps 3, 4 and 5 will be carried out on-line and in real time for each transmission session.

4. Results and discussion

Simulations with the typical MPEG video sequences are performed for the purpose of comparing the network utilizations of the proposed method and those of some other traffic specifications having different complexities, namely the peak rate, the first two and four (σ,ρ) pairs [2]. We carried out two kinds of experiments: the homogeneous (same sequences with same delay requirements) and the heterogeneous (different sequences with different delay requirements). In these experiments, all video sequences have the length of approximately 50 seconds and the rate of 30 frames per second. We consider a single multiplexer whose output capacity is chosen to be 45Mbps. It should be noted that most of traffic specifications in the literature so far have been tested only in the homogeneous case.

For homogeneous case, the video sequence in Figure 1a is used. Figure 4 shows the maximum number of concurrent connections for the four types of traffic specifications. As mentioned before, it is obvious that the performance of the proposed traffic specification is always the highest, which is as good as that of the empirical envelope. This again emphasizes the advantage of the proposed method. That is, while the solution is very simple with only one (σ,ρ) pair, its performance is the best in homogeneous case.

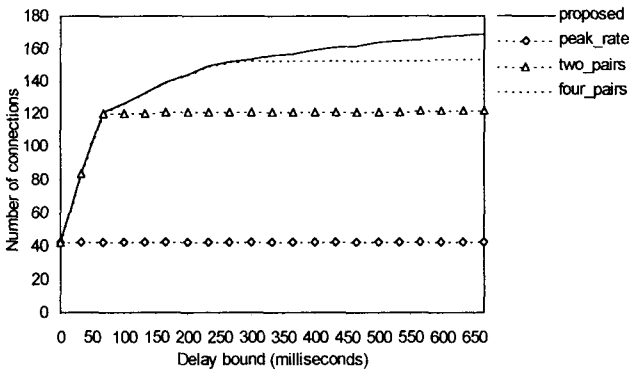


Fig. 4: Comparison of traffic specifications in the homogeneous case

The result in Figure 4 also proves the notion that the traffic specification of four (σ,ρ) pairs provides the performance very close to that of the empirical envelope. We can see that when the delay bounds are within 300ms, the performances of the empirical envelope and the four (σ,ρ) pairs are the same.

For the heterogeneous case, two video sequences, namely the above sequence (called sequence A) and the one in Figure 5 (called sequence B), are employed. We can see that the two sequences are different in the point that sequence A is very bursty whereas sequence B is rather smooth. In addition, the bit rate of sequence A is higher than that of sequence B. In these experiments, the selected scheduler is EDF as it has the best performance compared to the other schedulers.

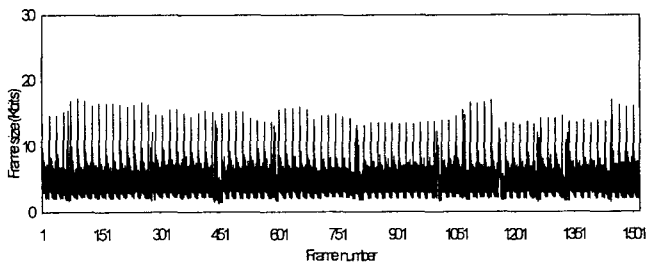
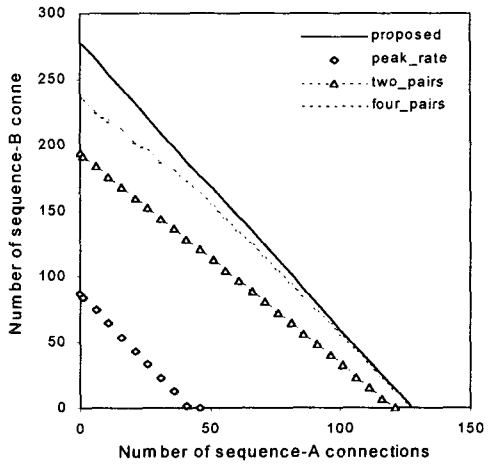


Fig. 5: The second video sequence (sequence B)

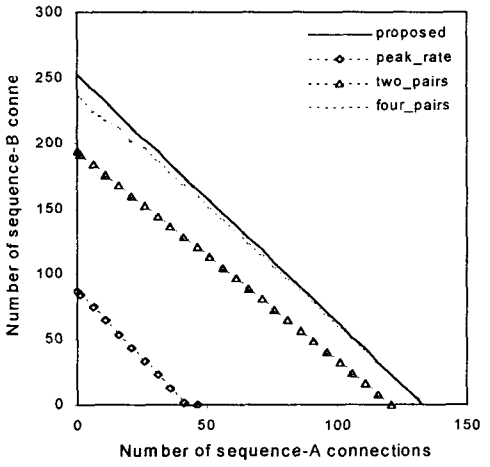
The two sequences are assigned with different delay bound requirements, however the connections carrying

the same video sequence will have identical delay bound requirements.

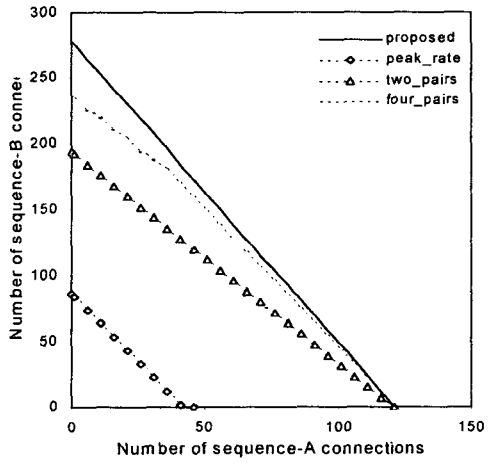
Figures from 6a to 6f show the maximum number of sequence-A and sequence-B connections with different traffic specifications for various contexts of delay bound requirements. In these figures, d_A and d_B are delay bounds for sequence A and sequence B respectively.

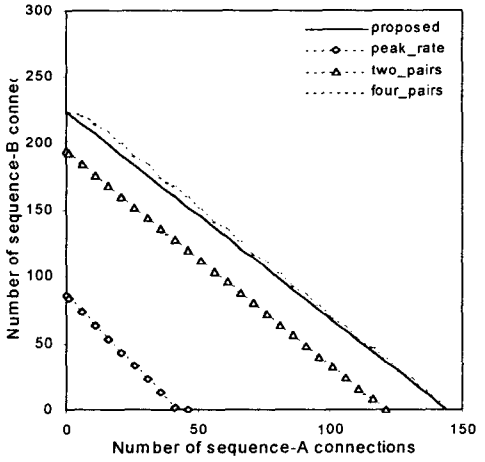


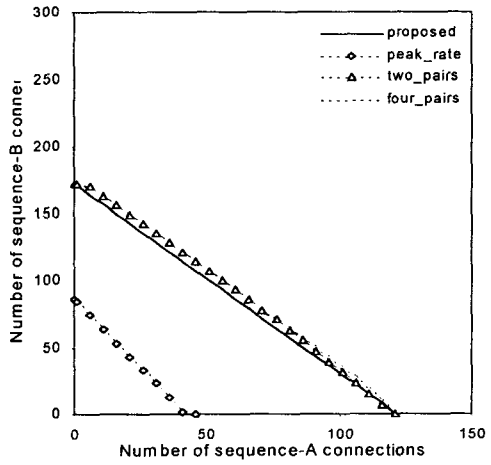
(a) d_A = 99ms, d_B = 132ms



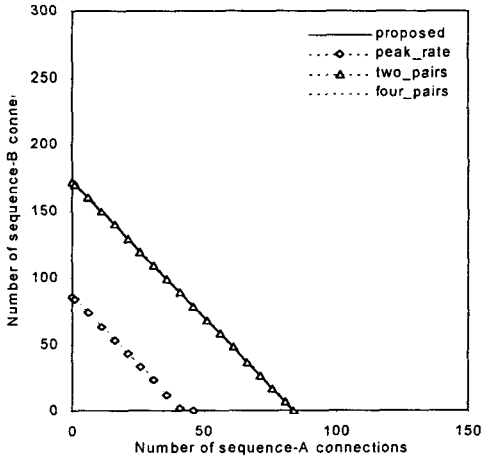
(b) d_A = 132ms, d_B = 99ms







(e) d_A = 66ms, d_B = 33ms



(f) d_A = 33ms, d_B = 33ms

Fig. 6: Comparison of traffic specifications in the heterogeneous case

The results show that generally the proposed traffic specification provides a better performance than the other traffic specifications. The advantage is very clear when compared with the specification of two (σ,ρ) pairs. Only in the cases where the delay bounds are very small (about 33 ms with sequence A and about 66ms with sequence B), the proposed traffic specification may have a somewhat worse performance than the specification of four (σ,ρ) pairs, and even the two (σ,ρ) pairs (Figures 6d, 6e, 6f). This phenomenon can be explained by the fact that the proposed traffic specification is essentially just one straight line whose slope changes according to the delay bound d. When the delay d is too small, the slope of that straight line may be much higher than those of the second and the third segments of the poly-line (although it is still smaller the peak rate). That is the "distance" between the proposed traffic specification and the poly-line could be large, resulting in the degradation of network utilization in the heterogeneous case.

The problem may be solved by constraining the maximum value of the slope (ρ value) or augmenting the proposed specification with one or more segments. These issues are reserved for future study.

Nevertheless, even with very small delay bounds, the performance of the proposed method is still close to that of the four (σ,ρ) pairs, which is in fact nearly the same as that of the empirical envelope. So the performance of the proposed method is still acceptable given the simplicity of the traffic specification.

5. Conclusion

The selection of a proper traffic specification from the video data is a hard issue due to the bursty characteristic of video content. For deterministic service, the empirical envelope is the most accurate traffic constraint function for an arrival function. However, the empirical envelope is too complex to be used as a practical traffic specification. Therefore, various solutions were proposed based on some approximations of the empirical envelope.

In this paper, we proposed a new method to estimate the traffic specification. The method takes into account the predefined delay bound to estimate the traffic specification just before the transmission. An estimation procedure is presented for real-time computation of traffic specification. The solution is simple yet it is shown to achieve the highest network utilization in the homogeneous case and have a good performance in the heterogeneous case. This method is efficient, straightforward and can be easily applied to existing applications such as the guaranteed services of ATM and IETF.

References:

[1] R. Guerin and V. Peris, "Quality-of-Service in Packet Networks: Basic Mechanisms and Directions," Computer Networks, vol.31, no.3, pp.169-179, 1999.

[2] D.E. Wrege, E.W. Knightly, H. Zhang, and J. Liebeherr, "Deterministic delay bounds for VBR video in packet-switching networks: fundamental limits and practical trade-offs," IEEE/ACM Trans. Netw., vol.4, no.3, pp.352-362, 1996.

[3] J. Liebeherr, and D.E. Wrege, "Traffic characterization algorithm for VBR video in multimedia networks," ACM/Springer-Verlag Multimedia Systems J., vol.6, no.4, pp.271-283, 1998.

[4] E.W. Knightly and H. Zhang, "D-BIND: An accurate traffic model for providing QoS guarantees to VBR traffic," IEEE/ACM Trans. Netw., vol.5, no.2, pp.219-231, 1997.

[5] E.W. Knightly, "H-BIND: a new approach to providing statistical performance guarantees to VBR traffic," INFOCOM '96, vol.3, pp.1091-1099, 1996.

[6] R.L. Cruz, "A calculus for network delay, Part I; network elements in isolation," IEEE Trans. Inform. Theory, vol.37, no.1, pp.114-131, 1991.

[7] J. Liebeherr, D.E. Wrege and D. Ferrari, "Exact admission control for networks with a bounded delay service," IEEE/ACM Trans. Networking, vol.4, no.6, pp.885-901, 1996.

[8] R. Vannithamby and A. Leon-Garcia, "Efficient estimation techniques and a new description of variable-bit-rate video traffic parameters," IEEE Canadian Con. Electrical and Computer Engineering, vol.1, pp.180-185, 1999.

[9] A. Lombardo, G. Schembra and G. Morabito, "Traffic specifications for the transmission of stored MPEG video on the Internet," IEEE Trans. Mult., vol.3, no.1, pp.5-17, 2001.

[10] G. Sisodia, S. De, M. Hedley and L. Guan, "New statistical model for VBR video traffic in ATM networks," 7th International Conference on Computer Communications and Networks, pp.448-453, 1998.

[11] H. Liu, N. Ansari and Y.Q. Shi, "The tale of a simple accurate MPEG video traffic model," IEEE International Conference on Communications, vol.4, pp.1042-1046, 2001.

[12] F. Guillemin, C. Rosenberg and J. Mignault, "On characterizing an ATM source via the sustainable cell rate traffic descriptor," IEEE Inforcom'95, vol.3, pp.1129-1136, 1995.