Estimation of Aggregate Effective Bandwidth for Traffic Admission in ATM Networks *

Zbigniew Dziong, Boris Shukhman and Lorne G. Mason

INRS-Telecommunications
16, place du Commerce, Verdun, Quebec H3E 1H6, Canada
e-mail: dziong@inrs-telecom.uquebec.ca

Abstract

A framework for adaptive bandwidth management in ATM based networks is presented. The central concept of this approach is an adaptive estimation of the aggregate effective bandwidth required by connections carried on each link of the network. To achieve reliable results the estimation process takes into account both the traffic source declarations and the connection superposition process measurements on the network links. This is done in an optimization framework provided by estimation theory. In the paper we concentrate on evaluation of a bandwidth reserved for possible estimation error. For this purpose we use the error covariance matrix from a linear two state Kalman filter applied for the system state estimation. The bandwidth reserved for the estimation error provides that the source parameter declarations can be more relaxed and that the source policing can be less stringent.

1 Introduction

The central objective of bandwidth management in ATM based networks is to provide fair and efficient access to network resources for all services while meeting Quality of Service (QoS) requirements. Several approaches were proposed to achieve this objective. They range from effective bandwidth allocation concepts (e.g. [1, 2, 3, 4]) via buffer reservation algorithms (e.g. [5]) and permit schemes (e.g. [6]) to strategies based on measurements (e.g. [7]). All of them have certain strengths and weaknesses (see e.g. [5]). In the paper we propose an approach which, to a large extent, eliminates basic weaknesses of the effective bandwidth allocation concepts.

The main inconvenience with the effective (or equivalent) bandwidth allocation is that the QoS can be guaranteed only if the real source parameters conform to the source declarations or if the source declaration parameters can be enforced by a policing mechanism. Note that it can be difficult for some sources to predict their traffic parameters in advance. Also it might be hard to enforce declarations of some statistical parameters (e.g. mean rate). In [10] it is argued that the drawbacks of the bandwidth management scheme based on effective bandwidth allocation can be avoided if an adaptive algorithm is employed. The proposed adaptive algorithm is based on both the source parameter declarations and measurements of link connection superposition process parameters. This information is used to estimate the aggregate effective bandwidth required by all connections carried on the link. The aggregate effective bandwidth is estimated by means of a Kalman filter and is utilized by the traffic admission control to decide whether a new connection can be accepted.

Estimation of the aggregate effective bandwidth, instead of updating effective bandwidth allocated to each connection based on the source process measurement (as proposed in [9]), is rationalized by several factors. Firstly, it is usually the connection superposition process which determines directly the QOS. Secondly, in general the measurement of the superposition process is more accurate than the sum of the individual source measurements. Lastly, this approach fits very well into connection admission procedure. The results presented in [10] show that the estimated effective bandwidth adapts very well to the real effective bandwidth even under very stressful and non-stationary conditions.

In the paper we extend the model from [10] to provide certain guarantee for quality of service. This is done by evaluation of a bandwidth reserved for possible estimation error. The bandwidth reserved for the estimation error provides that the source parameter declarations can be more relaxed and that the source policing can be less stringent compared to the effective bandwidth allocation based solely on the source parameters declarations. It is important to under-

*The research was supported by a grant from the Canadian Institute for Telecommunications Research under the NCE program of the Government of Canada.

line that this flexibility is achieved not at the expense of the network throughput. Quite opposite, the throughput in the case of adaptive mechanism can be significantly higher. This is explained by the fact that in the non-adaptive scheme many sources which cannot predict well its parameters would declare higher bandwidth requirements to avoid policing mechanism interference.

In the first part of the paper (Section 2) the proposed framework for network bandwidth management is presented. Central to the framework is an estimation algorithm which evaluates the aggregate effective bandwidth required by already admitted connections. The algorithm is executed at each network switch control processor for each outgoing link. The measurement of the connections superposition cell process on the link is performed at the switch output port. Since both the declarations and measurements have errors, the framework includes the bandwidth reserved for protection against these errors. The general model for the estimation of the effective bandwidth and the bandwidth reserved for the estimation error is presented in Section 3. The main idea is to decompose the process into two parts. First the algorithm estimates certain parameters of the link connection superposition process. Then, based on these parameters, the control procedure evaluates the corresponding bandwidths required for connection admission decisions. The estimation of the chosen parameters (mean and variance of the instant rate process) is based on a two-state linear Kalman filter (other applications of Kalman filter for network management can be found in [13, 14]). This model provides also the basis for analysis and evaluation of the parameter estimation errors (described in Section 4). The details of the connection admission procedure, including evaluation of the aggregate effective bandwidth and the bandwidth reserved for the estimation error, are given in Section 5. The numerical results (reported in Section 6) demonstrate that the algorithm copes very well with unpredicted changes in source parameters by providing good bandwidth efficiency and required quality of service. Several further possible studies based on the proposed model are indicated in the concluding remarks (Section 7).

2 Bandwidth Management Framework

In the following we describe a framework for the proposed adaptive bandwidth management scheme which is based on a three layer structure of traffic entities illustrated in Fig.1.

The network bandwidth is allocated logically to the accepted connections by means of the traffic admission and routing algorithms (see e.g. [15]). The objective of this allocation is to provide, for all services, fair and efficient access to the network resources which is quantified by the

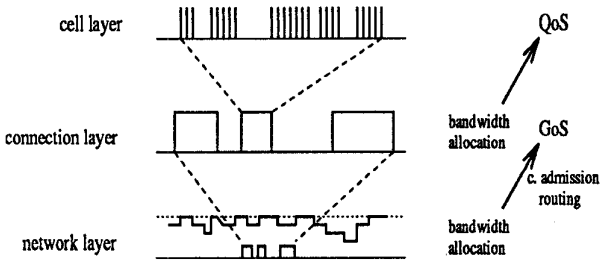

Figure 1: Three layer structure.

Grade of Service (GoS) measure (e.g. connection rejection probability). The amount of bandwidth allocated logically to each connection determines the Quality of Service (QoS) on the cell layer (e.g. cell loss probability).

In general two concepts for logical bandwidth allocation to connections are being considered in the ATM trials. In the first one each connection is allocated bandwidth equal to the connection peak rate, P. The advantage of this approach is that the QoS can be easily guaranteed and the implementation of the algorithm is simple. On the other hand, in the case of bursty sources, the bandwidth utilization can be quite low. The second option is so called effective (or equivalent) bandwidth allocation which has been widely investigated in the literature (e.g.[1, 3, 4, 8]). In this case the gain from statistical multiplexing of bursty sources is taken into account so the bandwidth allocated to a connection, g, is between the peak rate and the average rate, A, of the connection, $A \leq g \leq P$. The optimal allocation provides that when all the link bandwidth is allocated to connections, the QoS constraint is tight, thus maximum link utilization is achieved. Several algorithms are proposed (e.g. [3, 4, 11, 12]) to evaluate effective bandwidth as a function of the connection declared parameters, h:

$$g = f_g(h)$$

for given link speed, L. Throughout the paper we assume that the function f_g is available.

Note that the effective bandwidth concept works optimally only if the connection declared parameters (describing the connection cell process) are in agreement with the real connection parameters which also implies that the connection cell process is stationary. These conditions are seldom met for many services and in reality the effective bandwidth required by a source can be different from the declared value and moreover it can be time dependent. We describe this feature by introducing the notion of connection declaration error, $c(t)$, which defines the source parameters at time t as $h^r(t) = h + c(t)$.

To cope with unpredictability of source parameters and malicious users many algorithms incorporate a source policing mechanism which provides that the connection parameters cannot exceed certain values, h^p, defined by policing mechanism parameters. In this case the effective bandwidth allocation based on policer parameters will guarantee the required QoS. There are two drawbacks of this approach. Firstly, since there is no perfect mechanism which could police statistical parameters, the effective bandwidth based on policer parameters can be significantly higher than the one based on source declarations. Secondly, in many cases the source cannot predict well its statistical parameters so this approach can force source to declare overestimated parameters which will reduce bandwidth utilization.

In the paper we propose and investigate an alternative adaptive concept where the effective bandwidth allocation is based mainly on the declarations and measurements while the policing mechanism is assumed to have a secondary role. The key element of this approach is incorporation of a measurement process which allows to correct the initial bandwidth allocation based on declarations and to adapt to non-stationary changes in the cell process. Concerning the measured process, the natural choice would seem to be measurement of the parameters of individual source process since these parameters define the effective bandwidth required by the connection ([9]). Nevertheless we chose the measurement of the connection superposition process on each network link. There are many reasons for this approach. Firstly, it is the connection superposition process on the link which determines directly the QOS. Secondly, in general the measurement of the superposition process is more accurate than the sum of the individual source measurements. Finally this approach fits very well into connection admission procedures which are based on verification of the residual bandwidth on each link on the chosen path.

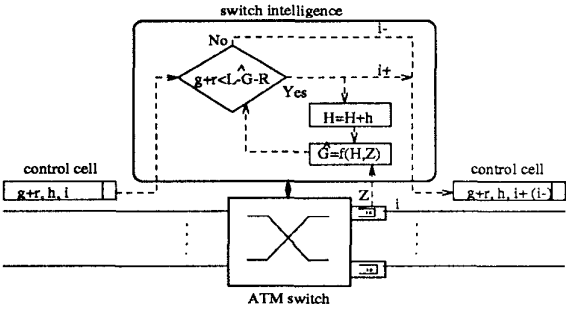

Figure 2: Adaptive bandwidth allocation concept.

The investigated adaptive bandwidth allocation concept is illustrated in Fig.2. where a control packet tries to reserve bandwidth g for a new connection on link i. The switch intelligence checks whether the requested bandwidth is smaller than (or equal to) the current residual bandwidth, C. A natural definition of the current residual bandwidth would be the difference between the link speed and the effective bandwidth of already accepted connections, G^r. Since the exact value of G^r is not known we use an estimate of this value, $\hat{G}$. The estimated effective bandwidth is a function of the declared parameters of accepted calls, $H = \sum h_j$, and measured parameters, Z, of the superposition process on the link.

Observe that both the requested effective bandwidth g and estimated effective bandwidth of accepted calls $\hat{G}$ may be different from the real values due to the declaration and estimation errors. Thus there exist certain probability that after acceptance of the new connection the effective bandwidth G^r will exceed link capacity resulting in violating QoS constraints. In order to keep this probability on acceptable level we introduce bandwidths reserved for errors so the acceptance rule is defined as

$$g + r \leq L - \hat{G} - R \qquad (1)$$

where R and r denote the bandwidths reserved for estimation and declaration errors, respectively. Assuming close to optimal estimation, this approach should provide higher bandwidth utilization under less stringent source policing and more relaxed source parameters declarations, compared to non adaptive approach (under given QoS constraints).

3 Control Model

In the following we limit our considerations to one output port of an ATM switch. The structure of the considered system is illustrated in Fig.3a where the set of sources is served by the ATM link. Before elaborating further details of the control procedure let us note that from the connection admission point of view we are interested in discrete points of time. Thus the system will be modeled in the discrete time domain where t_k,

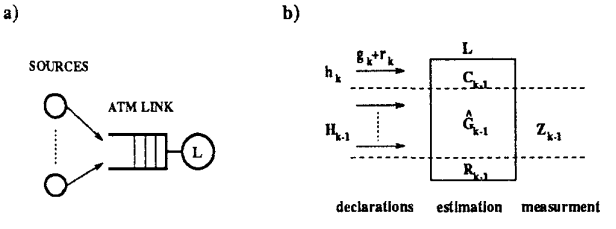

Figure 3: System and bandwidth allocation structures.

$k = 1, 2..$ denotes the instant of the system state change, $X_{k-1} \rightarrow X_k$, caused either by a new connection, or by a connection release. In general the system state can be described by a vector of the source superposition process parameters. The choice of the state description will be discussed later.

The bandwidth management scheme for these particular setting is illustrated in Fig.3b where the ATM link bandwidth, L (represented by the rectangle), is divided into three regions (in state X_{k-1}): estimated aggregate effective bandwidth required for the superposition process, $\hat{G}_{k-1}$, bandwidth reserved for estimation error, R_{k-1}, and bandwidth available to new connections, $C_{k-1} = L - \hat{G}_{k-1} - R_{k-1}$. It is assumed that the superposition process, $S(t)$, is measured in each state resulting in the vector of measured superposition process parameters, Z_{k-1}. The superposition process is also characterized by the vector of parameters, H_{k-1}, being a function of the parameters declared by each source, h_i (where i denotes the instant of the connection admission). A new connection is admitted if its declared effective bandwidth, g_k, increased by the bandwidth reserved for declaration error, r_k, is smaller than C_{k-1}. Throughout the paper we assume that the function $g_k = f_g(h_k)$ is available.

The issue of estimation of $\hat{G}_{k-1}$ and R_{k-1} fits very well into the framework of optimal estimation theory. In general it can be shown that, by applying a recursive discrete filter, the state of our system X_{k-1} could be estimated as a function of the following parameters: Z_{k-1}, h_{k-1} (the declared parameters of the connection added or released in the transition $X_{k-2} \rightarrow X_{k-1}$), $\hat{G}_{k-2}$ (in addition the knowledge of the declaration error and measurement error distributions, assumed to be gaussian, is required). This conclusion suggests that the natural choice for the state description would be the aggregate effective bandwidth, G^r_{k-1}, so the required estimate, $\hat{G}_{k-1}$, would be achieved directly. While this approach is possible, in the following we propose another state description which simplifies estimation process and provides that it is not limited to a particular model for effective bandwidth evaluation.

There are two difficulties with direct estimation of $\hat{G}_{k-1}$. Firstly, the relation between the effective bandwidth and the parameters which can be measured, $\hat{G}_{k-1} = f_{\hat{G}}(Z_{k-1})$, is in general non linear. This feature implies that a nonlinear filter should be applied which is typically more complex and less accurate than a linear filter. Secondly, although one can find in the literature several models for evaluation of the effective bandwidth, these algorithms are relatively complex and do not provide a universal closed form solution. These characteristics significantly increase the complexity of the problem since the function $f_{\hat{G}}$ should be inverted in the estimation algorithm.

To avoid the mentioned problems we chose as the state description a vector of the superposition process parameters which can be directly measured. Thus a linear filter can be applied and the inverse function problem becomes trivial. Having the estimate of the system state, $\hat{X}_{k-1}$, the estimate of aggregate effective bandwidth can be evaluated from a differential approach:

$$\hat{G}_{k-1} = f_{\Delta\hat{G}}(\hat{X}_{k-1} - X^d_{k-1}, G^d_{k-1}) \qquad (2)$$

where X^d, G^d denote the values evaluated from the declared parameters. In this manner the full evaluation of the effective bandwidth function, $f_{\hat{G}}$, can be avoided in the estimation process.

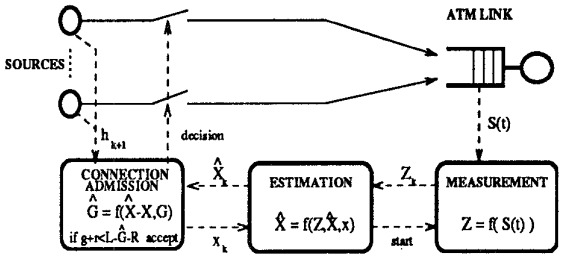

Figure 4: Structure of the control system.

The general structure of the control system with the decomposed estimation process is given in Fig.4. The control model consists of three distinct blocks: connection admission, estimation and measurement. In the remainder of the paper we describe and investigate a particular version of this scheme where mean and variance of the instant rate of the superposition process are chosen as the system state description and measured parameters. While at first glance it might be viewed as a bold simplification, observe that the aggregate effective bandwidth can be evaluated from the full set of declared parameters which take into account all important source features while the estimated parameters serve only to correct this bandwidth allocation. The implied assumption is that in general the sources do not change their basic features associated with the type of the connection (otherwise a more sophisticated set of estimated and measured parameters should be chosen).

4 Estimation of the Mean and Variance

The objective of the estimator is to provide the best estimate of the mean, M_k, and variance, V_k,

of the instant rate of the connection superposition process. We start by defining a mathematical model of the considered system which constitutes the basis for optimal estimation. The state of the system is defined as $X_k = [M_k, V_k]^T$. The dynamics of the system model are illustrated in Fig.5 and are defined by

$$X_k = X_{k-1} + x_k + e_k \tag{3}$$

where e_k denotes the model error and $x_k = [\delta_m m_k, \delta_v v_k]^T$ denotes either the declared mean and variance of the accepted connection ($\delta_m = \delta_v = 1$) or the normalized declared mean and variance of the relased connection ($\delta_m = -\hat{M}_k/M_k^d$, $\delta_v = -\hat{V}_k/V_k^d$ where the index d denotes parameters evaluated from declarations). It is assumed that the model error, $e_k = [\tilde{m}_k, \tilde{v}_k]^T$, is a gaussian random variable with zero mean and covariance matrix, Q_k. The evaluation and interpretation of the model error parameters is discussed in the next section.

The system model in Fig.5 is complemented by the measurement model which provides the measure of the parameters in state k, $Z_k = [\bar{M}_k, \bar{V}_k]^T$. The delay corresponds to the fact that the result of measurement of the parameters is required at the time of the next state change. The measurement model is defined as follows

$$Z_k = X_k + u_k \tag{4}$$

where $u_k = [\tilde{M}_k, \tilde{V}_k]^T$ is the measurement error. Like the model error, the measurement error is assumed to be a gaussian random variable with zero mean and known covariance matrix, Y_k, (for details see [10]).

The system and measurement model fit very well into the framework of linear recursive filters. In the following we demonstrate application of the linear discrete Kalman filter (see e.g. [16]) to the considered system. In general the Kalman filter provides an optimal least-squares estimate of a system state on condition the system is linear and the model and measurement errors are gaussian random variables. In this paper we assume that these conditions are fulfilled although in reality the model error can have different distribution especially in the case of the system with a policing mechanism (this issue will be discussed in a separate publication). The dynamics of the applied filter is shown in Fig.5. The system state estimate update is given by

$$\hat{X}_k = \hat{X}_k^e + K_k[Z_k - \hat{X}_k^e] \tag{5}$$

where $\hat{X}_k^e = \hat{X}_{k-1} + x_k$ denotes the state estimate extrapolation and K_k is the Kalman filter gain.

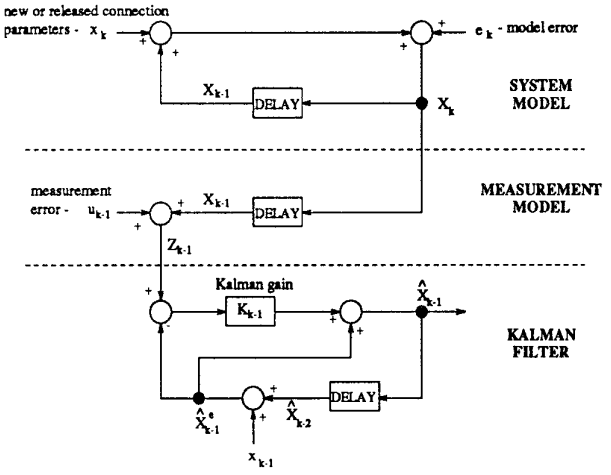

Figure 5: System model and Kalman filter.

From the form of (5) it is clear that the gain is a weight which decides how much confidence should be given to the measurement vs. the declaration.

To simplify further considerations let us introduce the transition matrix, F_k, with the following elements: $a_{11} = \frac{M_{k-1}+\delta_m m_k}{M_{k-1}}$, $a_{22} = \frac{V_{k-1}+\delta_v v_k}{V_{k-1}}$, and $a_{12} = a_{21} = 0$. Thus the system model can be described by

$$X_k = F_k X_{k-1} + e_k \tag{6}$$

Then the Kalman gain can be evaluated from

$$K_k = P_k^e[P_k^e + Y_k]^{-1} \tag{7}$$

where P_k^e denotes the estimate error covariance matrix extrapolation given by

$$P_k^e = F_{k-1} P_{k-1} F_{k-1}^T + Q_{k-1} \tag{8}$$

where P_k is updated estimate error covariance matrix evaluated from

$$P_k = [I - K_k]P_k^e \tag{9}$$

Concerning the initial values we assume that the system is empty at the time $k = 0$ so $X_0 = 0$ and $P_0 = 0$.

4.1 Estimation error

From the connection admission point of view it is crucial that the estimation procedure provides also an information about the possible estimation error. In particular this information is necessary to evaluate the bandwidth R_k reserved for the error in evaluation of $\hat{G}_k$. The obvious basis for the estimation error analysis is the error covariance matrix, P_k. In the following we assume that the

estimation errors for mean and variance of the instant rate process are independent.

Let us define the estimation errors for mean and variance as

$$\delta_k^M = \hat{M}_k - M_k \quad (10)$$

$$\delta_k^V = \hat{V}_k - V_k \quad (11)$$

respectively. Under the gaussian assumption of the model and measurement errors, the estimation error distributions are also gaussian with zero mean and variance defined by diagonal elements of the error covariance matrix. Let p_k^M, p_k^V denote the variances of δ_k^M and δ_k^V, respectively. These variances can be used to evaluate the bandwidth R_k reserved for the error in evaluation of $\hat{G}_k$. The details of the procedure are given in Section 4.2.

The error covariance matrix is a function of the measurement and model errors. While the assessment of the measurement error is rather straightforward (see [10]), the model error requires more thorough discussion. In general the model error is a function of the declaration errors which are defined as a difference between the real and declared connection parameters

$$c_j(t) = h_j^r(t) - h_j \quad (12)$$

where time t indicates that in general the error can be nonstationary. The transformation of the error into the discrete time domain can be done in several ways. One possibility is to assume $c_{k+i} = c(t_{k+i})$; $i = 0, 1, ...$ where t_k is the time of connection acceptance. Another possibility is to use an average error over the discretization interval

$$c_{j,k+i} = \frac{\int_{k+i}^{k+i+1} c_j(t)}{t_{k+i+1} - t_{k+i}} \quad (13)$$

In the context of our problem the second option seems to be more appropriate since we are interested in QoS in the whole period between the state transition. Then the model error is defined by

$$e_k = c_{j,k} + \sum_i [c_{i,k} - c_{i,k-1}] \quad (14)$$

where index j corresponds to the new or departing connection and index i corresponds to existing connections. Observe that in general the periods between the state transitions are significantly shorter than the call duration. This feature indicates that although the connection parameters might be different from declarations, one can expect that there will be very large autocorrelation between the values in the subsequent system states so the terms under summation in (14) will be small. Moreover, under the assumption of statistical independence of the connection instant rate processes, these terms will be positive and negative. These premisses lead to the conclusion that the second term in (14) can be neglected so the model error can be approximated by the declaration error of the new or departing connection

$$e_k = c_{j,k} \quad (15)$$

In this case the covariance matrix of the model error is defined by the predicted variances of declaration errors for mean and variance, v_k^m and v_k^v respectively. These values can be evaluated from statistics. In our model it is assumed that the declaration errors have gaussian distribution with zero mean. Note that this approximation is exact when each connection process is stationary.

5 Connection Admission

5.1 Estimation of aggregate effective bandwidth

To simplify presentation we omit the time index in this section. In general the estimated parameters $\hat{X} = [\hat{M}, \hat{V}]^T$ are not sufficient to evaluate accurately the aggregate effective bandwidth. In the proposed model we use these parameters only to evaluate the deviation in the effective bandwidth allocation compared to the reference point which is defined by the declared parameters, X^d. Thus we are looking for the following function

$$\hat{G} = f_{\Delta\hat{G}}(\hat{X} - X^d, G^d) \quad (16)$$

To estimate this function one can use approximations based on theoretical models presented in [17, 18, 19]. For a link with small buffers the suggested admissible region is defined by

$$M + \alpha_1 \cdot \sqrt{V} < \beta_1 \quad (17)$$

where α and β depend on buffer capacity, link speed and QoS constraint. For relatively large buffers, an analysis using large deviations [19] implies the following approximation

$$M + \alpha_2 \cdot V < \beta_2 \quad (18)$$

This relation is also confirmed by empirical results reported in [12].

In the following we will use the relation (18) although the procedure is equally applicable to (17). Based on (18), for particular state of accepted connections, we will have

$$G^d = \gamma \cdot M^d + \theta \cdot V^d \quad (19)$$

We assume that the coefficient γ is independent from the current state and can be evaluated off line based on link speed and QoS constraint. Then the coefficient θ, for a given state, is evaluated from

$$\theta = \frac{G^d - \gamma \cdot M^d}{V^d} \quad (20)$$

Finally the estimated effective bandwidth of the connection superposition process is given by

$$\hat{G} = \gamma \cdot \hat{M} + \theta \cdot \hat{V} \qquad (21)$$

5.2 Bandwidth reserved for errors

Before discussing the details of evaluation of the bandwidths reserved for errors, R, r, it is important to define more precisely what is the design objective of the connection admission procedure and how the quality of this procedure should be judged? The answers to these questions are not straightforward. Note that the main criterion of connection admission is to provide that the QoS constraint is met (on the cell level). In our model this requirement is fulfilled when the real bandwidth required by the admitted connections does not exceed the link capacity, $G_k^r \leq L$. Obviously strict execution of this condition might cause that the reserved bandwidth will be large. Thus to improve bandwidth utilization we allow that $G_k^r > L$ with certain small probability:

$$P\{G_k^r > L\} \leq \epsilon \qquad (22)$$

There is one drawback of this formulation. Namely this probability depends strongly on the connection arrival process and connection bandwidth requirements. In particular the smaller traffic level the smaller $P\{G_k^r > L\}$. This feature indicates that the condition (22) is not convenient for the connection admission algorithm design.

To cope with the issue we propose another definition of connection admission procedure quality which is independent from the connection arrival process and connection bandwidth requirements. It is based on the following conditional probability

$$P\{G_k^r > L \mid g_k + r_k = L - (\hat{G}_{k-1} + R_{k-1})\} \leq \epsilon \qquad (23)$$

In this case the quality is defined for the critical case where the residual capacity is equal to the one required by a new connection. Observe that condition (23) is also fulfilled when

$$P\{G_k^r > g_k + r_k + \hat{G}_{k-1} + R_{k-1}\} \leq \epsilon \qquad (24)$$

The latter condition constitutes the basis for design and evaluation of the proposed connection admission procedure.

The sum $r_k + R_{k-1}$ could be evaluated from the superposition of distributions of g_k^r and G_{k-1}^r. Nevertheless from the connection admission viewpoint it is more convenient to separate evaluation of r_k from evaluation of R_{k-1}. The main reason behind this approach is that we would like to have a very simple decision algorithm.

For the bandwidth reserved for the declaration error we apply a conservative approach where the parameter r_k is defined by the connection peak rate, P_k:

$$r_k = P_k - g_k \qquad (25)$$

Concerning the bandwidth reserved for the estimation error, under the gaussian assumptions of the estimated mean and variance errors, the estimated bandwidth error, $\delta_G = \hat{G}_k - G_k^r$, has also gaussian distribution. Thus under the assumption of the mean and variance independence we have

$$R_{k-1} = U(\epsilon)\sqrt{\gamma^2 p_{k-1}^M + \theta^2 p_{k-1}^V} \qquad (26)$$

where $U(\epsilon)$ denotes coefficient derived from the normalized gaussian distribution which provides that

$$P\{G_{k-1}^r > \hat{G}_{k-1} + R_{k-1}\} \leq \epsilon \qquad (27)$$

6 Numerical Results

We start from description of a simulation model used for assessment of the proposed algorithms. To avoid excessive complexity of the system we simplified the model as much as possible to concentrate on the main issues. In particular only the instant rate layer is modeled and the link buffer has zero length. Also the investigation is limited to homogeneous sources.

The requests for connections are generated with intensity λ (Poissonian distribution) and mean holding time μ^{-1} (exponential distribution). The connections are of on-off type and are described by the peak rate, P, fixed burst length, B, and the average silence length, S (exponential distribution). This parameters together with QoS constraint and link capacity are used for computation of the effective bandwidth. The declaration error can be generated in many ways. For our study we have chosen the burst length as the parameter influenced by the error. Concerning the error distribution we applied a model which is significantly different from the Kalman filter assumption. Namely the error generator has two cyclic states ("on" and "off") with the same period T. When the generator is in the state "on" all connections accepted in this state are generated with the burst length larger than the declared one ($B' = B + \Delta B$). In the "off" state all new connections are generated with the declared burst length (for entire duration of the connection). Note that in this case the error has no zero mean. The reason for this model is to evaluate the adaptation scheme under more stressing conditions. Since there is no buffer in the system the QoS is given by the probability of packet loss defined as

$$\bar{M}_k = \frac{\int_{t_0}^{t_k} [d(t) - L]^+}{\int_{t_0}^{t_k} d(t)} \qquad (28)$$

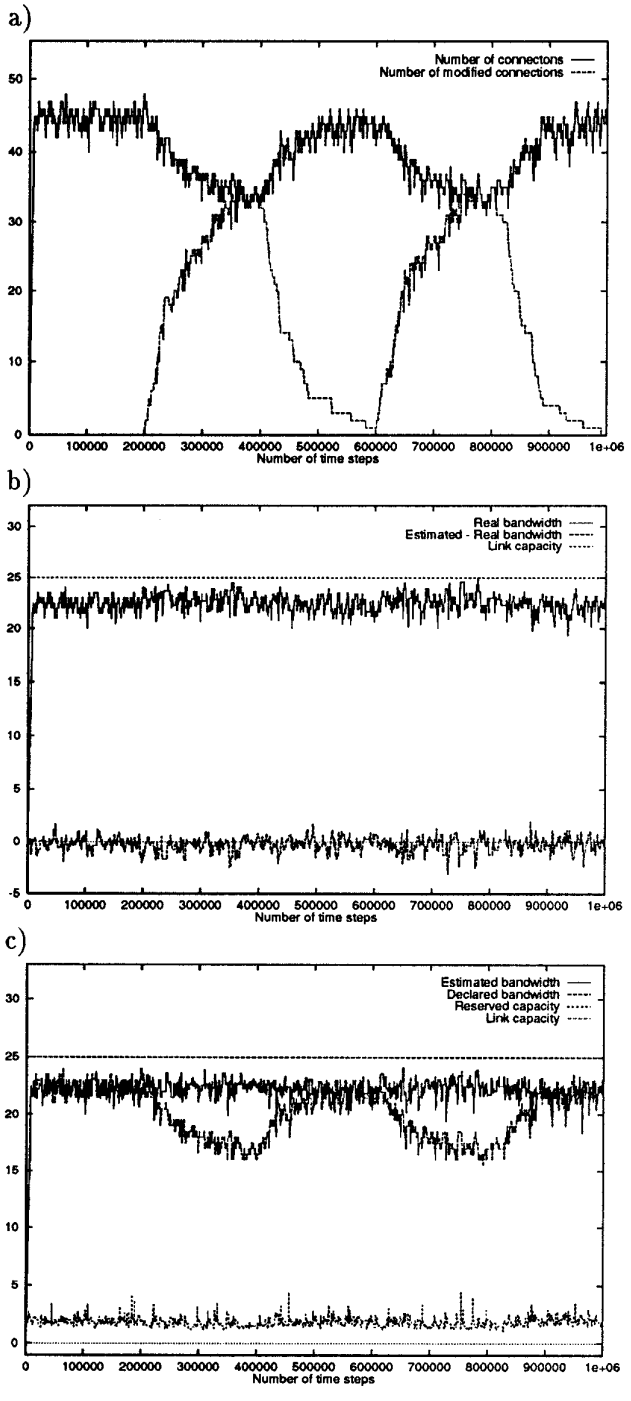

Figure 6: Trajectories of the system variables vs. simulation time.

where $d(t)$ denotes the instant rate and $[x]^+ = x$ for $x \geq 0$ and $[x]^+ = 0$ for $x < 0$. Observe that, due to the source model and lack of the buffer, the QoS for a given number of sources can be evaluated analytically from the binomial distribution. In the simulation package this model is used to evaluate both the effective bandwidth from the source declarations, $g_k = f_g(h_k)$, and the aggregate effective bandwidth based on the real parameters, G^r_{k-1} (to assess the accuracy of the adaptation scheme).

The example selected for presentation is defined by the following parameters: link and connection parameters - $L = 25$, $\lambda = 200^{-1}$, $\mu^{-1} = 50000$; source parameters - $P = 1$, $B = 50$, $S = 70$, $\sqrt{v^m} = 0.1$, $\sqrt{v^v} = 0.1$; error generator parameters: $T = 200000$, $B' = 100$

A sample of the effective bandwidth allocation dynamics, during a part of the simulation run, is presented in Fig.6. In Fig.6a the total number of connections and the number of connections with modified burst length are given. As we can see during the "on" period of the error generator the parameters of almost all connections are modified while at the end of the "off" period almost all connections have the declared parameters.

The trajectory of the real aggregate effective bandwidth, G^r, and the difference between the estimated and real aggregate effective bandwidth, $\hat{G} - G^r$, are presented in Fig.6b. The estimated aggregate effective bandwidth tracks well the real value. It can be noticed that the highest underestimation of the effective bandwidth occurs during the "on" period of the error generator. Nevertheless the bandwidth reserved for estimation error provides that the real aggregate effective bandwidth does not exceed the link capacity.

The efficiency of the proposed adaptive algorithm is underlined in Fig.6c where the estimated aggregate effective bandwidth allocation, $\hat{G}$, and the declared aggregate effective bandwidth, G^d are plotted. The large gap between the two trajectories, in the periods when the number of modified calls is significant, indicates that the connection admission scheme based solely on the declared parameters would allow too many calls and the real aggregate effective bandwidth would significantly exceed the allocated capacity. Additionally the bandwidth reserved for the estimation error R is depicted in Fig.6c.

7 Conclusions

We have described a general framework for adaptive bandwidth management in ATM based networks. The central concept of this framework is adaptive estimation of the aggregate effective bandwidth required by accepted connections on each network link. In order to take advantage of all available informations the estimation process

takes into account both the traffic source declarations and the connection superposition process measurements. This is done in an optimization framework provided by the estimation theory. To provide required QoS the estimation error is analyzed. The analysis is based on the error covariance matrix from the applied linear two state Kalman filter. The covariance matrix is used to evaluate bandwidth reserved for the estimation error.

The numerical study of the proposed scheme demonstrated that the approach copes well with undeclared changes in traffic parameters. In particular it was shown that the bandwidth reserved for estimation error provides that the quality of service requirements are met even though the sources exceed the declared average rate by 50%. At the same time it was indicated that the scheme based solely on declarations can cause significant deterioration of the QoS.

There are several further model extensions which could be investigated. For example, the influence of the policing mechanism on the error analysis can be verified. Also the measurement process constitutes an important area for study where declared and measured autocorrelation functions can play an important role. More complex filters can be tried to check whether it is possible to estimate some characteristics corresponding more directly to the QoS. Finally the problem of correlated sources can be investigated since the model can cope with such cases.

References

[1] Turner J.S., "The challenge of multipoint communication", 5th ITC Seminar, Lake Como, Italy, May 1987.

[2] Woodruff G., Kositpaiboon R., Fitzpatrick G., Richards P., "Control of ATM Statistical Multiplexing Performance" ITC Specialist Seminar, Adelaide 1989.

[3] Dziong Z., Choquette J., Liao K.-Q., Mason L., "Admission Control and Routing in ATM Networks", Proc. of ITC Specialist Seminar, Adelaide, September 1989. Also published in "Computer Networks and ISDN-Systems" Vol. 20, December 1990, 189-196.

[4] Guerin R., Ahmadi H., Naghshineh M., "Equivalent Capacity and its Application to Bandwidth Allocation in High-Speed Networks" IEEE J.S.A.C No 9, 1991.

[5] Turner J.S., "Managing Bandwidth in ATM Networks with Bursty Traffic", IEEE Network, September 1992.

[6] Mitra D., "Asymptotically optimal design of congestion control for high speed data networks" *IEEE Trans. on Commun.*, vol. COM-40, pp. 301-311, Feb. 1992.

[7] Saito H., "Dynamic Call Admission Control in ATM Networks" IEEE J.S.A.C, vol.9, No 7, 1991.

[8] Kelly F.P., "Effective Bandwidths at Multiclass Queues" Queuing Systems No 9, 1991.

[9] Tedijanto T.E., Gun L., "Effectiveness of Dynamic Bandwidth Management Mechanisms in ATM Networks", Proc. of INFOCOM'93, San Francisco 1993.

[10] Dziong Z., Montanuy O., Mason L., "Adaptive Traffic Admission in ATM Networks - Optimal Estimation Framework ", ITC-14, France, June 1994.

[11] Gallassi G., Rigolio G., Fratta L., "Bandwidth Assignment and Bandwidth Enforcement Policies", Proc. of Globecom'89, Dallas 1989.

[12] Roberts J.W. "Performance evaluation and design of multiservice networks" COST 224 final report October 1992.

[13] Pack C.D., Whitaker B.A., "Kalman Filter Models for Network Forecasting", BSTJ, Vol 61, No 1, January 1982.

[14] Chemouil P., Filipiak J., "Kalman Filtering of Traffic Fluctuations for Real-Time Network Management" Annales des Telecommunications, Tome 44, No.11-12, pages 633-640, 1989.

[15] Dziong Z., Mason L., " Call Admission and Routing in Multi-Service Loss Networks", IEEE Transactions on Communications, Vol.42, No.2, Part 3, April 1994.

[16] Gelb A., "Applied Optimal Estimation" The MIT Press, 1974.

[17] Gach A., Mialaret C., Allard P.E., "An Experimental Evaluation of Call Acceptance Management Algorithms in ATM Based Networks" Proc. of Canadian Conference on Electrical and Computer Engineering CCECE-92, Toronto, Ontario, Canada, Sep. 1992.

[18] Kas B., Kleinewillinghofer-Kopp R., "The Use of the Two Moment Allocation Scheme" FIDBP 123 0006 CD CC, March 1990.

[19] Courcoubetis C., Walrand J., "A Note on Effective Bandwidth" (to be published).