Estimation and Prediction-Based Connection Admission Control in Broadband Satellite Systems

Yeong Min Jang

We apply a "sliding-window" Maximum Likelihood (ML) estimator to estimate traffic parameters of On-Off source and develop a method for estimating *stochastic predicted individual* cell arrival rates. Based on these results, we propose a simple Connection Admission Control (CAC) scheme for delay sensitive services in broadband onboard packet switching satellite systems. The algorithms are motivated by the limited onboard satellite buffer, the large propagation delay, and low computational capabilities inherent in satellite communication systems. We develop an algorithm using the *predicted individual* cell loss ratio instead of using steady state cell loss ratios. We demonstrate the CAC benefits of this approach over using steady state cell loss ratios as well as predicted total cell loss ratios. We also derive the predictive saturation probability and the predictive cell loss ratio and use them to control the total number of connections. Predictive congestion control mechanisms allow a satellite network to operate in the optimum region of low delay and high throughput. This is different from the traditional reactive congestion control mechanism that allows the network to recover from the congested state. Numerical and simulation results obtained suggest that the proposed predictive scheme is a promising approach for real time CAC.

Manuscript received December 23, 1999; revised November 10, 2000.

Yeong Min Jang is with School of Computer Science, Duksung Women's University, Seoul, Korea. (phone: +82 2 901 8348, e-mail: yjang@center.duksung.ac.kr)

I. INTRODUCTION

Future satellite-based communication systems will be expected to support a wide variety of multimedia services. To accommodate these diverse services with their inherent traffic fluctuations while efficiently utilizing the spectrum, onboard packet switching capabilities should be implemented at the satellite [1], [2].

To achieve both efficient utilization of the space-segment (or satellite) resources and acceptable QoS_m for each traffic class m, intelligent traffic management protocols that incorporate congestion/flow control must be provided to ensure *individual* quality of service objectives. Congestion control requires proper monitoring and control of traffic flow. Because congestion may be predicted, closer control at the cell burst and call-level can be placed on traffic flow. Before traffic is admitted to the network, the CAC should be applied to ensure that individual quality of service objectives can be satisfied. Although there are many QoS_m requirements, i.e., cell loss ratio, individual cell loss ratio, saturation probability, delay, and jitter etc., in this paper, individual cell loss ratio and saturation probability at the downlink are chosen as the QoS measures.

The congestion and resource allocation problem has been extensively studied for terrestrial broadband ATM networks [3]–[7], [26]. However, large propagation delays (typically 125 ms), small onboard buffers, and low computational capabilities characteristic of satellite systems make most of the proposed schemes inappropriate for satellite networks. Basically, there are two kinds of congestion control schemes: reactive and preventive. However, feedback-based reactive schemes may fail due to the propagation delay. Because of propagation delay and the desire to utilize the expensive bandwidth of satellite networks efficiently, it may be necessary to introduce larger buf-

fers than typically found in terrestrial ATM switches. Thus, we propose a predictive scheme that incorporates the ability to predict the future cell loss behavior and makes control decisions that prevent congestion rather than react to congestion that has already occurred.

To predict congestion and guarantee QoS_m, we need to estimate the individual cell loss ratio. For traffic modeled as a superposition of On-Off sources, many approaches for evaluating the cell loss ratio have been proposed. Because of its mathematical simplicity and tractability, the most attractive approach approximates the actual arrival process to the buffer by a continuous fluid flow [9]. To obtain the exact solution of the cell loss ratio by Matrix Geometric techniques [8], it is necessary to solve a set of matrix equations which is time consuming. Many other approximations, such as Markov Modulated Deterministic Process (MMDP) and Markov Modulated Poisson Process (MMPP), are also available. However, due to their analytical and numerical complexity, they are not appealing from the practical point of view. Since the connection setup time is constrained, these rather complex models are not practical.

Most real world traffic is bursty because most sources are transient. We desire a scheme which optimizes both the transient and the steady state performances. But the queuing analysis found in current fluid model approaches provides only *steady* state results [4]–[7], [10] due to the complexity of modeling transient behavior. To our knowledge, a general treatment of predictive or transient solutions to this single server queue has not been presented. Predictive solutions under more limited conditions are presented in [11]–[13]. In these papers, we present a general predictive scheme that outperforms steady state schemes. Information such as the measured network load may be used when performing the CAC function. This may allow a network to achieve higher network utilization while still meeting the network performance objectives. The CAC function is network specific [14].

Motivated by this work, this paper presents a simple and efficient traffic estimation and cell loss prediction approach to control congestion. Measurement and estimation-based connection control would provide better network utilization than using pre-specified traffic parameters. Instead of requiring the user to explicitly specify his traffic, the network attempts to "estimate" the traffic parameters of existing connections by making on-line measurements. This approach has several advantages. First, the user-specified traffic descriptor can be simple and rough (i.e., peak rate, mean rate, burstiness, and QoS requirement) because the traffic estimator will find exact traffic descriptor. Second, an overly conservative specification does not result in an overallocation of resources for the entire duration of connection.

The organization of the paper is as follows. In section II, we describe the network architecture. The traffic estimator is presented in section III. In section IV, we derive expressions for predicting cell loss ratio and individual cell loss ratio. A CAC algorithm is applied in section V. In section VI, we present numerical and simulation results. Finally, we conclude in section VII.

II. PROBLEM FORMULATION

1. Network Architecture

As technology advances, more onboard processing capability can be incorporated in satellites. It is desired to avoid computation-intensive procedures onboard. Thus, distributed techniques in flow and congestion control may be more appropriate. Figure 1 shows the network architecture. The up-links may use TDMA, MF-TDMA, FDMA, CDMA or MF-CDMA techniques. The downlinks use TDM. We assume that for MF-TDMA uplink, we may use Media Access Control (MAC) protocol using fixed-rate demand assignment. Earth stations are interconnected via a satellite switch with output buffering [2] and shared memory [15]. The shared memory approach provides more flexibility and better memory utilization. Recently, memory technology advanced rapidly, and the memory access speed is no longer a critical part of the whole structure, especially in small satellite switches. For example, as a result, the shared-buffering ATM switches seem to be the most promising architecture [16]. The satellite has a switching fabric capable of routing cells that arrive on the up-links to their destination downlink. The introduction of downlink queue improves the utility of downlink capacity by allowing for statistical multiplexing. Within each Virtual Path (VP), there are M classes of virtual circuit connections (VCCs), each with its own QoS_m requirements. Suppose that $N(= N_1 + \cdots + N_m + \cdots + N_M)$ independent heterogeneous On-Off sources (connections) are connected to a satellite downlink, where N_m denotes the number

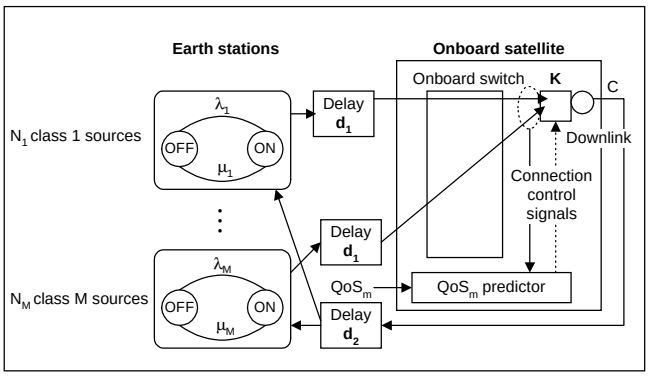

Fig. 1. Connection control architecture for broadband satellite communication.

of connections of class m. A VP therefore can be modeled as a bufferless single server system with output buffered downlink capacity of C bits/sec. A QoS_m predictor at the downlink predicts individual cell loss ratio for a time $t(= d_1 + d_2)$ ahead. The earth stations dynamically, yet optimally reject the new connections after receiving a choke signal from the satellite.

The symbols are represented in Table 1.

Table 1. Explanation of symbols.

symbols	meaning
QoS	QoS of total traffic
QoS_m	QoS of class m traffic
CLR(t)	transient total burst-level CLR
$P_{sat}(t)$	transient burst-level saturation probability
$ICLR_m(t)$	transient individual burst-level CLR of class m traffic
$QoS_{ICLR_m(onset)}$	required QoS_m for congestion onset
$QoS_{ICLR_m(abatement)}$	required QoS_m for congestion abatement

2. QoS Measures

Our performance measures are $P_{sat}(t)$ and $CLR(t)$. $P_{sat}(t)$ is the fraction of time that the aggregate demand for bandwidth from all sources exceeds the nodal bandwidth at time t when the number of active class-m sources at time 0 is $Y_m(0)$ (see Fig. 2). The number of active sources of class-m at time $t = 0$ ($Y_m(0)$) is measured using the Virtual Channel Identifier (VCI) of each VCC. We shall also investigate $CLR(t)$, the burst-level cell loss ratio.

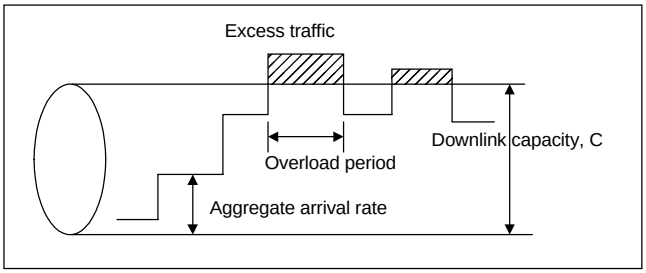

Fig. 2. Overload and underload periods.

3. Traffic Model

Although the traffic characteristics of future onboard satellite networks are hard to estimate with complete accuracy, there are a number of voice and video models reported as On-Off source models [17]–[19], and this On-Off model has been commonly used to model a voice source with speech activity detection. We assume that simple Markovian models, i.e., models based on the binary On-Off model, are sufficient to capture the short range correlation for telephone speech and video telephone. An aggregate of such binary Markov sources can be used to model a video source [19]. Each source operates independently. X *minisources* are used to model Y variable rate video sources ($X >> Y$). In [19], it was found that with $X = 20 * Y$, the analytical results using a discrete-state continuous-time Markov process model agree closely with simulations using a continuous-state Autoregressive Markov model. Heyman and Lakshman [20] suggest that long-range dependence is not an important property for most applications of a VBR-video-source model. They show that some Markov chain models will describe VBR video-conference traces. Lucantoni *et al*. [21] propose a Markov-renewal process model to describe a single source. Due to the statistical nature of multimedia traffic, the modeling of incoming traffic characteristics plays an important role.

Each source is modeled as an On-Off source. We assume that a series of cells arrive in the form of a continuous stream of bits to use a fluid model. We also assume that the "OFF" and "ON" periods for sources of class m are both exponentially distributed with parameters λ_m and μ_m, respectively. The rate of flow from the "ON" state to the "OFF" state is μ_m and from "OFF" to "ON" is λ_m. In this traffic model, when a source is in the "ON" state, it generates cells with a constant interarrival time, $\frac{1}{R_m}$ seconds/bit. When the source is in the "OFF" state, it does not generate any cells. See Fig. 3 for binary On-Off model for class m traffic. We assume that N_m class m sources of the N connections sharing a downlink have the same traffic parameters (λ_m, μ_m, R_m). The state of a source of class m is characterized by an underlying Markov process whose infinitesimal generator is given by

$$Q_m = \begin{bmatrix} -\lambda_m & \lambda_m \\ \mu_m & -\mu_m \end{bmatrix} \tag{1}$$

for $m = 1, 2, \cdots, M$. Let us define $\lambda_m = \frac{1}{\phi_m}$, $\mu_m = \frac{1}{\theta_m}$.

4. Outline of Predictive CAC Algorithm

A predictive connection control algorithm is implemented onboard the satellite and executed for each downlink. The algorithm is an on-line, per-connection (not per-QoS class) based estimation (measurement) algorithm that runs continuously using a sliding-window mechanism. The network management function resets the algorithm for new estimation. The algorithm is executed whenever we have a new connection request. The operation period of the algorithm is a trade-off between power

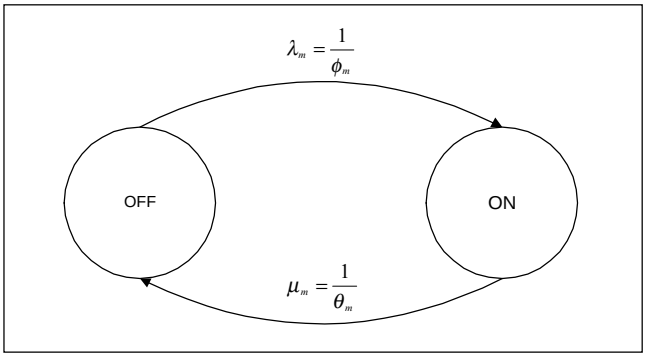

Fig. 3. On-Off model for class *m* traffic.

consumption onboard the satellite and estimation accuracy. We run the algorithm whenever we have a new connection or continuously for the entire duration of a connection or until the algorithm is reset by network management. If a connection is rejected, we throw away the estimation results for that connection.

The connection acceptance decision must be made within the required allowable connection set-up latency in accordance with the received connection set-up request. This means that the CAC must operate in real-time, even when various kinds of traffic are multiplexed. With the QoS_m requirements we can define the two thresholds, $QoS_{ICLR_m(onset)}$ and $QoS_{ICLR_m(abatement)}$. The selection of $QoS_{ICLR_m(onset)}$ and $QoS_{ICLR_m(abatement)}$ values depends on the trade-off between QoS_m requirement and satellite network efficiency. These normally depend on implementation and are beyond the scope of this paper.

An outline of the algorithm is as follows (see also Fig. 4):

1) *Characterize source parameters*: For each traffic source using the downlink, we collect sample data of the On-Off traffic for a number of active and idle periods and compute these estimates, ϕ_m and θ_m, using the ML method. Details are provided in section III.

2) *Analyze to get QoS measure*: Based on the estimated parameters, ϕ_m and θ_m, the number of connections N_m, and the number of active sources at the current time $t = 0$, i_m, predict the individual cell loss ratio, $ICLR_m(t)$ at time $t = 0.25\ sec$.

- If $ICLR_m(t) > QoS_{ICLR_m(onset)}$ under no congestion state, then send congestion onset messages to all class-*m* sources.
- Or, if the system is currently congested and $ICLR_m(t) < QoS_{ICLR_m(abatement)}$, then congestion has ended and a congestion abatement message is sent to all class-*m* sources.

3) *Find the number of connections*: Whenever the $ICLR_m(t)$ of a source is higher than the $QoS_{ICLR_m(onset)}$ requirement under no congestion state or the $QoS_{ICLR_m(abatement)}$ requirement under congestion state, the satellite controller will compute (using results obtained in 1 and 2 above) the optimum number of connections (N_m^*) of class-*m* sources.

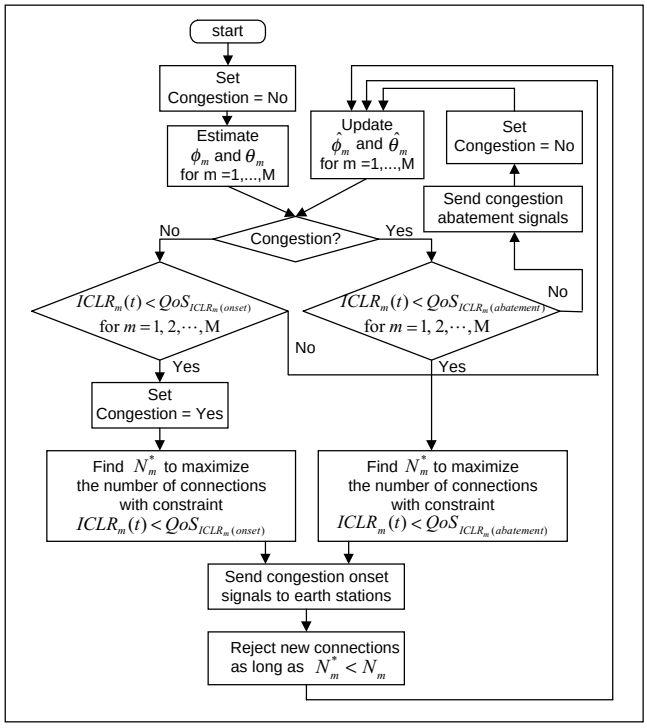

Fig. 4. Predictive CAC algorithm.

To satisfy the QoS_m, the satellite controller would accept any new connections as long as $N_m^* > N_m$, the number of current connections. Even though current N_m is measured, we have to find the N_m^* for determining whether the satellite controller can accept the new connection or not. Details are provided in sections V.

Some "memory" is required to record the current congestion state so that the connection control can react accordingly. We implement this memory element by adding a hysteresis loop as shown in Fig. 5.

At each earth station, the following control actions are taken:

- *Congestion onset message received*: Reject the new class-*m* connection requests.
- *Congestion abatement message received*: Allow new class-*m* equests to connect.

III. ESTIMATION OF ϕ_m AND θ_m USING ML ESTIMATOR

Most papers [4]–[7], [10] are based on the traffic parameters specified by a user prior to service, and do not consider the actual cell stream to check if it satisfies the traffic parameters. Therefore, the link utilization may become unnecessarily low. In order to overcome this effect, CAC using estimation/measurement/prediction of cell stream is considered. The traffic patterns are different depending on the nationality, language, culture, personal char-

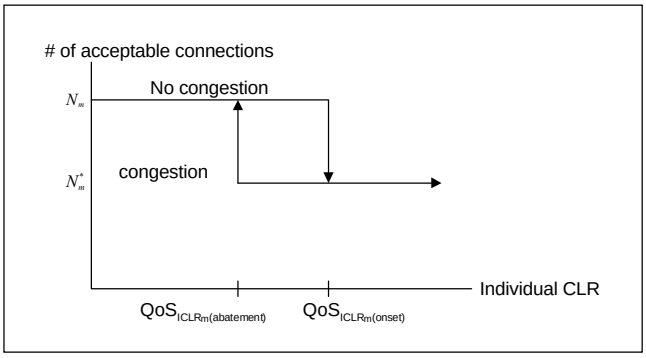

Fig. 5. Step function with hysteresis.

acteristics and so on. For On-Off speech, Table 1 of [18] gives a list of the traffic parameters for 32 speakers in 16 conversations. All traffic patterns are different. To predict the individual cell loss ratio, we have to accurately estimate three critical parameters: peak rate, burst length, and burst interarrival time. The traffic generated by individual connections must be monitored and controlled to ensure that it is consistent with traffic parameters agreed on at the connection establishment. We have to estimate these parameters accurately and quickly because source parameters will affect the $ICLR_m(t)$. Here, we propose an ML estimator to estimate On-Off traffic parameters. Such an estimator is useful because it yields the smallest estimation errors in the absence of a *priori* statistics.

Recall that we have assumed that the sequences of active and idle periods (observed at the downlink) are exponentially distributed random variables. Let the random point process of one source (the number of arriving cells per slot per connection) be denoted by 0 or 1 (see Fig. 6).

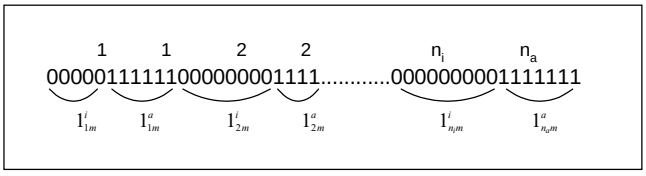

Fig. 6. ML diagram of class m source.

Let $1^a_{1m},\cdots,1^a_{km},\cdots,1^a_{n_am}$ and $1^i_{1m},\cdots,1^i_{km},\cdots,1^i_{n_im}$ be the length of successive active and idle periods, respectively. The estimation is based on the measurement of a *fixed* number n_a of active periods and a *fixed* number n_i of idle periods of the On-Off sources. It is assumed that all idle and active periods are *iid* random variables. After we formulate the likelihood function [12] of this sample, the ML estimate of θ_m is:

$$\hat{\theta}_m = \frac{1}{\hat{\mu}_m} = \frac{\sum_{k=1}^{n_a} l^a_{km}}{n_a}. \tag{2}$$

Similarly, we obtain the ML estimator of ϕ_m:

$$\hat{\phi}_m = \frac{1}{\hat{\lambda}_m} = \frac{\sum_{k=1}^{n_i} l^i_{km}}{n_i}. \tag{3}$$

The ML estimators defined above provide the best possible estimates when a *priori* statistics are not available, and as long as the statistics do not change with time (stationary). This extension leads to the following "sliding-window" estimator structure:

$$\hat{\theta}_m(k) = \frac{1}{\hat{\mu}_m(k)} = \frac{\sum_{l=k-n_a}^{k-1} l^a_{lm}}{n_a}. \tag{4}$$

For large numbers of samples and next short active period duration, the computational burden required by these estimators may become quite severe. A recursive form for the estimation is

$$\hat{\theta}_m(k) = \frac{1}{\hat{\mu}_m(k)} = \hat{\theta}_m(k-1) + \frac{1}{n_a}[l^a_{(k-1)m} - l^a_{(k-n_a-1)m}]. \tag{5}$$

Similarly, we obtain the recursive form for $\hat{\phi}_m(k)$:

$$\hat{\phi}_m(k) = \frac{1}{\hat{\lambda}_m(k)} = \hat{\phi}_m(k-1) + \frac{1}{n_i}[l^i_{(k-1)m} - l^i_{(k-n_i-1)m}]. \tag{6}$$

1. Quality Assessment of Sliding-Window ML Estimator

We now evaluate the quality of the ML estimator. For an *iid* sequence the sample mean is unbiased for all $\theta_m(k)$, i.e.,

$$E[\hat{\theta}_m(k)] = \frac{1}{n_a} E[\sum_{l=k-n_a}^{k-1} l^a_{lm}] = \frac{n_a\theta_m(k)}{n_a} = \theta_m(k). \tag{7}$$

The Cramer-Rao theorem [22] gives a bound on the minimum achievable variance of any unbiased estimator.

$$Var[\hat{\theta}_m(k)] \geq \frac{-1}{n_a E\left[\frac{\partial^2}{\partial\theta_m^2}[\ln f(l^a_{km};\theta_m(k))]\right]} = \frac{\theta_m^2(k)}{n_a}. \tag{8}$$

Because n_a is fixed in our estimation procedure, (2) has a Gamma distribution. The variance of estimated parameter value is therefore

$$Var[\hat{\theta}_m(k)] = Var[\frac{\sum_{k=1}^{n_a} l^a_{km}}{n_a}] = \frac{1}{n_a^2}[n_a\theta_m^2(k)] = \frac{\theta_m^2(k)}{n_a}. \tag{9}$$

Since this value is exactly the same as the Cramer-Rao lower bound, our estimator is *efficient*. We find that the larger n_a, the

more closely the values of the estimated about the true mean $\theta_m(k)$. Indeed by the Chebyshev inequality, the probability that an estimate $\hat{\theta}_m(k)$ deviates by more than ε from the true mean $\theta_m(k)$ is bounded by $P(\left|\hat{\theta}_m(k)-\theta_m(k)\right| \geq \varepsilon) \leq \frac{\theta_m^2(k)}{n_a \varepsilon^2}$ and hence go to zero as n_a goes to infinity. Therefore, our estimator is *consistent* since,

$$\lim_{n_a \to \infty} P[\left|\hat{\theta}_m(k)-\theta_m(k)\right| \geq \varepsilon] = 0 \tag{10}$$

for all $\varepsilon > 0$.

IV. PREDICTING QOS MEASURES

1. Transient Saturation Probability

In the previous section, we developed the estimator for (λ_m, μ_m) which we will use to predict the saturation probability and cell loss ratio. We will use a statistical bufferless fluid-flow model to predict the probability that congestion occurs at time t based on the traffic statistical behavior and the estimation at time 0. We assume that N On-Off sources share the capacity C of a downlink. We are interested in the number of active sources of real-time voice and video traffic on a downlink. Since we are interested in transient saturation probability, a formula involving the backward Kolmogorov equations of the process is used. The number of active class-m sources forms a birth-death process, with birth and death rates that depend on the state of the process $\lambda_{km} = (N_m - k)\lambda_m$ and $\mu_{km} = k\mu_m$ (see Fig. 7).

Without loss of generality, we assume that we have two traffic classes. Let $P_{k,j}(t)$ denote the probability that k class-2 and j class-1 sources are active. The transitions among states are expressed as a set of differential equations:

$$\frac{dP_{0,0}(t)}{dt} = -(N_1\lambda_1 + N_2\lambda_2)P_{0,0}(t) + \mu_1 P_{0,1}(t) + \mu_2 P_{1,0}(t) \tag{11}$$

$$\begin{aligned}\frac{dP_{k,j}(t)}{dt} &= (N_2 - k + 1)\lambda_2 P_{k-1,j}(t) + (N_1 - j + 1)\lambda_1 P_{k,j-1}(t) \\ &\quad -[(N_1 - j)\lambda_1 + (N_2 - k)\lambda_2 + k\mu_2 + j\mu_1]P_{k,j}(t) \\ &\quad +(j+1)\mu_1 P_{k,j+1}(t) + (k+1)\mu_2 P_{k+1,j}(t) \\ &\quad k = 2, 3, \cdots, N_2 - 1, \quad j = 2, 3, \cdots, N_1 - 1\end{aligned} \tag{12}$$

$$\begin{aligned}\frac{dP_{N_2,N_1}(t)}{dt} &= \lambda_2 P_{N_2-1,N_1}(t) + \lambda_1 P_{N_2,N_1-1}(t) \\ &\quad -(N_2\mu_2 + N_1\mu_1)P_{N_2,N_1}(t).\end{aligned} \tag{13}$$

We recognize the above (11)-(13) as the backward Chapman-Kolmogorov equations. In matrix form, they can be written as

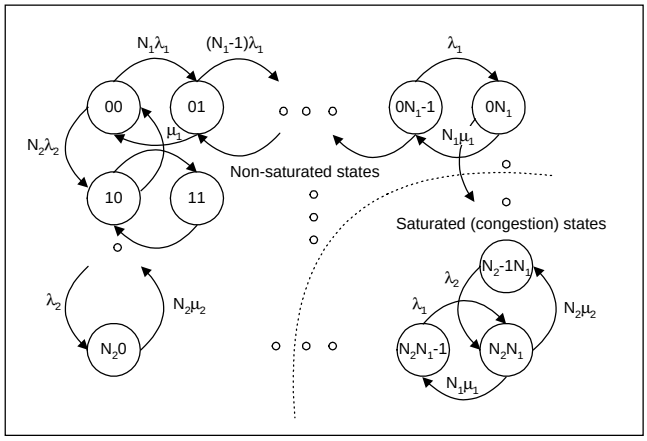

Fig. 7. State transition diagram for two traffic classes sources.

$$\frac{dP(t)}{dt} = QP(t), \tag{14}$$

where $P(t)$ is the column vector $(P_{0,0}(t), P_{0,1}(t), \cdots, P_{0,N_1}(t), P_{1,0}(t), \cdots, P_{N_2,N_1}(t))^T$ and Q is a $[(N_2+1)\times(N_1+1)]\times[(N_2+1)\times(N_1+1)]$ matrix. In order to solve (14) for the time-dependent behavior $P(t)$, we require initial conditions; that is, we must specify $P_{k,j}(0)$ for $k = 0, 1, \cdots, N_2$ and $j = 0, 1, \cdots, N_1$. In addition, we further require the following constraint:

$$\sum_{k=0}^{N_2}\sum_{j=0}^{N_1} P_{k,j}(t) = 1. \tag{15}$$

In (14), Q is a singular matrix with the rank $(N_1+1)(N_2+1) - 1$. Thus, we can find the predictive conditional state probability, $P(t)$, by using the eigenvalues of matrix Q:

$$P(t) = V\begin{bmatrix} e^{-s_1 t} & 0 & \cdots & 0 \\ 0 & e^{-s_2 t} & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & e^{-s_{(N_1+1)(N_2+1)} t} \end{bmatrix} V^{-1}P(0), \tag{16}$$

where $s_1, s_2, \cdots, s_{(N_1+1)(N_2+1)}$ are the eigenvalues of Q and s_1=0. V stands for the right eigenvectors of matrix Q. $P(0)$ is the column vector with $P_{Y_2(0),Y_1(0)}(0) = 1$ because the number of active class-2 and class-1 sources at time 0 are $Y_2(0)$ and $Y_1(0)$, respectively. The conditional transient saturation probability is then given by

$$P_{sat}(t) = P(\Lambda(t) > C \mid Y_1(0), Y_2(0)) = \sum_{k,j \in \{k,j \mid (R_2 k + R_1 j - C) \geq 0\}}^{R_2 N_2 + R_1 N_1} P_{k,j}(t), \tag{17}$$

where $\Lambda(t) = R_2 k + R_1 j$ denotes the aggregate arrival rate of the k active class-2 sources and the j active class-1 sources at time t.

2. Total Cell Loss Ratio

It is very important that the decision to accept or reject connections is made in real time. To do this, we need a simple and fast CAC scheme. Such a scheme should be able to predict the cell loss ratio rather than the steady state cell loss ratio. In the previous section, we developed the estimator for $(\hat{\lambda}_m, \hat{\mu}_m)$ which we will use to predict $CLR(t)$ *and* $ICLR_m(t)$.

Next we introduce the following definitions:

- $CLR(t)$: the predicted burst-level cell loss ratio at future time t.
- $A(t)$: the predicted and aggregated average arrival rate at future time t.
- $OV(t)$: predicted average excess traffic at future time t.

Let $\Lambda(t)$ denote the aggregate arrival rate from $Y_m(t)$ active sources. In a bufferless system, cell losses occur when $\Lambda(t)$ exceeds the link capacity C. The prediction of $CLR(t)$ is given by the ratio of mean excess traffic $(OV(t))$ and mean traffic load $(A(t))$ at time t. The $CLR(t)$ can be solved by using conditional expectations. The predictive $CLR(t)$ is given by:

$$CLR(t) = \frac{OV(t)}{A(t)} = \frac{E[(\Lambda(t)-C)^+]}{E[\Lambda(t)]} = \frac{E\left[E[(\Lambda(t)-C)^+ \mid Y(0)=I]\right]}{E\left[E[\Lambda(t) \mid Y(0)=I]\right]}. \quad (18)$$

Taking into consideration the fact that each of $N(=N_1+\cdots+N_M)$ existing connections belong to one of the M connection classes, given by an arbitrary initial condition $Y(0) = I = [Y_1(0)=i_1, Y_2(0)=i_2, \cdots, Y_M(0)=i_M]$, we obtain the conditional moment generating function of $\Lambda(t)$, $s \geq 0$:

$$\begin{aligned} G_{\Lambda(t)|Y(0)}(s) &= E\left[e^{s\Lambda(t)} \mid Y(0)=I\right] = \prod_{m=1}^{M} E\left[e^{sA_m(t)} \mid Y_m(0)=i_m\right] \\ &= \prod_{m=1}^{M} E\left[e^{sR_m Y_m(t)} \mid Y_m(0)=i_m\right] = \prod_{m=1}^{M} G_{\Lambda_m(t)|Y_m(0)=i_m}(s) \\ &= \prod_{m=1}^{M} \left[p_m(t)(e^{sR_m}-1)+1\right]^{N_m-i_m} \left[q_m(t)(e^{sR_m}-1)+1\right]^{i_m}. \end{aligned} \quad (19)$$

where $\Lambda(t) = \sum_{m=1}^{M} \Lambda_m(t)$. Thus,

$$A(t) = G'_{\Lambda(t)|Y(0)}(0) = \sum_{m=1}^{M} A_m(t) = \sum_{m=1}^{M} R_m\left[i_m q_m(t) + (N_m - i_m) p_m(t)\right]. \quad (20)$$

To derive $p_m(t)$ and $q_m(t)$, we can use the forward Chapman-Kolmogorov matrix differential equation:

$$\begin{bmatrix} \pi'_{00}(t) & \pi'_{01}(t) \\ \pi'_{10}(t) & \pi'_{11}(t) \end{bmatrix} = \begin{bmatrix} \pi_{00}(t) & \pi_{01}(t) \\ \pi_{10}(t) & \pi_{11}(t) \end{bmatrix} \begin{bmatrix} -\lambda_m & \lambda_m \\ \mu_m & -\mu_m \end{bmatrix}. \quad (21)$$

Let $A_{00}(t) = e^{(\lambda_m+\mu_m)t}\pi_{00}(t)$. Then, $\frac{dA_{00}(t)}{dt} = \mu_m e^{(\lambda_m+\mu_m)t}$ which can be integrated to yield $A_{00}(t) = \frac{\mu_m}{\lambda_m+\mu_m} e^{(\lambda_m+\mu_m)t} + B$. The initial condition $A_{00}(0) = 1$ determines the constant of integration to be $B = \frac{\lambda_m}{\lambda_m+\mu_m}$. Thus, $A_{00}(t) = \frac{\mu_m}{\lambda_m+\mu_m} e^{(\lambda_m+\mu_m)t} + \frac{\lambda_m}{\lambda_m+\mu_m}$. So $\pi_{00}(t) = \frac{\mu_m}{\lambda_m+\mu_m} + \frac{\lambda_m}{\lambda_m+\mu_m} e^{-(\lambda_m+\mu_m)t}$. Since $\pi_{01}(t) = 1 - \pi_{00}(t)$, we have

$$\pi_{01}(t) = p_m(t) = \frac{\lambda_m}{\lambda_m+\mu_m}\left[1 - e^{-(\lambda_m+\mu_m)t}\right], \quad (22)$$

and by symmetry,

$$\pi_{11}(t) = q_m(t) = \frac{\lambda_m}{\lambda_m+\mu_m} + \frac{\mu_m}{\lambda_m+\mu_m} e^{-(\lambda_m+\mu_m)t} \quad (23)$$

where $p_m(t)$ is the transition probability that a class-m source is active at future time t given the source is idle at time 0. $q_m(t)$ is the transition probability that a class-m source is active at future time t given the source is active at time 0.

Let $Y_m(t)$ and $Y_m(0)$ denote the numbers of active class-m sources at time t and 0. Then

$$OV(t) = \sum_{x_1=N_1}^{x_1} \cdots \sum_{x_m \in \{x_m \mid (\sum_{m=1}^{M} x_m R_m - C) \geq 0\}}^{x_m=N_m} \cdots \sum_{x_M}^{x_M=N_M} \left(\prod_{m=1}^{M} P(Y_m(t)=x_m \mid Y_m(0)=i_m)\right)\left(\sum_{m=1}^{M} x_m R_m - C\right) \quad (24)$$

where $\prod_{m=1}^{M} P(Y_m(t)=x_m \mid Y_m(0)=i_m)$, the predicted conditional state probability, can be derived as

$$\prod_{m=1}^{M} \sum_{k_m=0}^{x_m} \binom{i_m}{x_m-k_m} [q_m(t)]^{x_m-k_m} [1-q_m(t)]^{i_m-x_m+k_m} \binom{N_m-i_m}{k_m} [p_m(t)]^{k_m} [1-p_m(t)]^{N_m-i_m-k_m}. \quad (25)$$

3. Individual Cell Loss Ratio

We need a simple and dynamic CAC scheme. Also, we may

need to satisfy the individual QoS requirements. The predicted individual cell loss ratio, $ICLR_m(t)$, for class-m among M classes of traffic is given by:

$$ICLR_m(t) = \frac{OV_m(t)}{A_m(t)}. \quad (26)$$

Taking into consideration that each of N_m existing connections, given by an arbitrary initial condition $Y_m(0)$, we obtain the conditional moment generating function of $\Lambda_m(t)$, $s \geq 0$:

$$G_{\Lambda_m(t)|Y_m(0)}(s) = E\left[e^{s\Lambda_m(t)} \mid Y_m(0) = i_m\right]$$
$$= \left[p_m(t)(e^{sR_m} - 1) + 1\right]^{N_m - i_m} \left[q_m(t)(e^{sR_m} - 1) + 1\right]^{i_m}. \quad (27)$$

Thus,

$$A_m(t) = G'_{\Lambda_m(t)|Y_m(0)}(0) = R_m\left[i_m q_m(t) + (N_m - i_m)\, p_m(t)\right] \quad (28)$$

where $p_m(t)$ and $q_m(t)$ are given in (22) and (23), respectively.

Let $Y_m(t)$ and $Y_m(0)$ denote the number of active sources of class-m at future time t and 0. Then

$$OV_m(t) = E\left[\frac{(\Lambda(t) - C)^+ \Lambda_m(t)}{\Lambda(t)}\right]$$
$$= \sum_{x_1=N_1}^{x_1} \cdots \sum_{x_m \in \{x_m \mid (\sum_{m=1}^{M} x_m R_m - C) \geq 0\}}^{x_m = N_m} \cdots \sum_{x_M}^{x_M = N_M} \left(\prod_{m=1}^{M} P(Y_m(t) = x_m \mid Y_m(0) = i_m)\right) \frac{(\sum_{m=1}^{M} x_m R_m - C) x_m R_m}{\sum_{m=1}^{M} x_m R_m} \quad (29)$$

where $\prod_{m=1}^{M} P(Y_m(t) = x_m \mid Y_m(0) = i_m)$ is given in (25).

V. OPTIMAL CAC USING INDIVIDUAL CELL LOSS RATIO

When the traffic source is indirectly connected to the satellite system via another network, the satellite cannot control the traffic parameters directly. It can only control the connection admission or the number of connections on each downlink (N^*). When a subscriber requests to establish a connection with the satellite, the satellite management function which executes the bandwidth allocation, will first predict $ICLR_m(t)$ as described above. For example, for the $M = 2$ case, we determine the optimal number of connections, N_2^*, as follows:

$$N_2^* = \max\{n_2 \mid ICLR_2(i_1, i_2, t) < QoS_{ICLR_2(abatement\ or\ onset)}\}. \quad (30)$$

VI. NUMERICAL AND SIMULATION RESULTS

In this section, we present some results to verify the validity of our analysis. In Tables 2 and 3, we use $S_m = 10{,}000$ and $S_m = 1{,}000$ sample points, respectively. The percentage errors and variances from the real values of ϕ_m and θ_m in ML approach are also included. The number of active and idle periods, n_a and n_i, is found in order to compare the variance of the estimator. We observe that for high values of ϕ_m and θ_m (i.e., ϕ_m = 117.64 for n_a=113) the ML estimation method exhibits large errors and variances because n_a and n_i are small. However, as we decrease ϕ_m and θ_m, the percentage of error and variances dramatically decreases because n_a and n_i are large. Therefore, we find that the number of active and idle periods, n_a and n_i, affects the estimation quality more than S_m. The more active and idle periods we collect, the more accurate estimates.

Table 2. ML estimator results for $S_m = 10{,}000$.

θ_m	ϕ_m	$\hat{\theta}_m$ (%error)	$\hat{\phi}_m$ (%error)	n_a	$Var(\hat{\theta}_m)$
75.75	117.64	64.51 (17%)	108.69 (8%)	63	66
75.75	11.764	69.93 (8%)	12.07 (2.5%)	145	33
75.75	1.1765	69.93 (8%)	1.22 (2%)	199	24
7.575	117.64	8.30 (8%)	103.09 (14%)	113	0.6
7.575	11.764	7.15 (5.9%)	11.73 (0.2%)	687	0.07
7.575	1.1765	7.52 (0.7%)	1.17 (0.2%)	1506	0.03

Table 3. ML estimator results for $S_m = 1{,}000$.

θ_m	ϕ_m	$\hat{\theta}_m$ (%error)	$\hat{\phi}_m$ (%error)	n_a	$Var(\hat{\theta}_m)$
75.75	117.64	73.52 (3%)	81.30 (45%)	8	717
75.75	11.764	81.96 (7.5%)	9.90 (1.9%)	17	337
75.75	1.1765	70.42 (7.5%)	1.30 (10%)	20	286
7.575	117.64	5.41 (40%)	71.94 (63%)	5	11
7.575	11.764	7.62 (6%)	13.81 (15%)	68	0.84
7.575	1.1765	8.02 (5.6%)	1.1468 (2.5%)	145	0.39

For real time and low power hardware, we can easily implement a "sliding-window" ML CMOS module onboard the satellite. For any parameters of Table 2 and Table 3, the ML estimator takes less than 1 μ sec for all samples [23]. Because the number of VCs as well as the number of input and output ports of the onboard switch are small, VC-based estimation may not put large overhead for traffic management. Our results

suggest that the ML is a suitable estimator for this application.

For the numerical calculations of the predicted $CLR(t)$, we consider the following system: $N_1 = 25$ and $N_2 = 25$ PCM coded sources with $R_1 = 64$ kbps, $R_2 = 32$ kbps, $\hat{\lambda}_1 = \hat{\lambda}_2 = 0.5$ and $\hat{\mu}_1 = \hat{\mu}_2 = 0.833$ multiplexed onto a downlink of capacity $C = 1.544$ Mbps. We assume that the mean connection duration of a source is 180 seconds for voice. Thus, the mean interarrival time is 3.5 seconds for 50 On-Off sources. So we may assume that no new connection arrives during the prediction time (0.25 seconds).

In Fig. 8, we depict the predicted $CLR(t)$ as a function of the prediction time (in seconds) for various values of the initial conditions, $Y_1(0)$ and $Y_2(0)$.

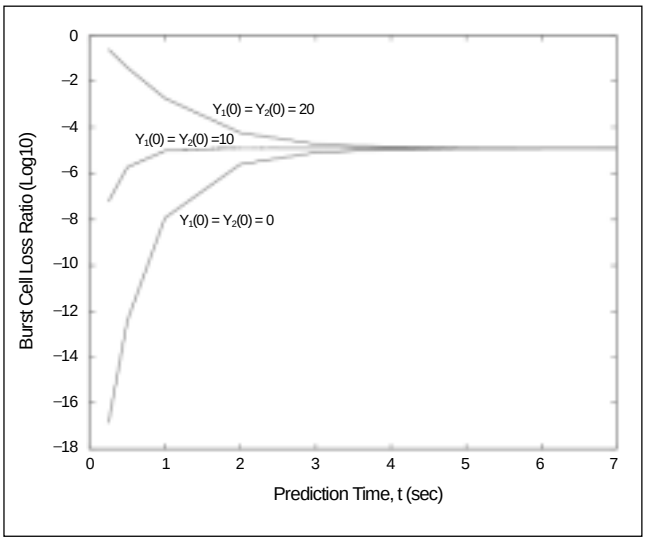

Fig. 8. Predicted *CLR*(*t*).

We observe that after approximately 4.0 seconds the predicted $CLR(t)$ will converge to the steady state value, $CLR(\infty)$ (obtained using the result presented in [6]). The round trip delay in a geostationary (GEO) communication satellite system is 0.25 seconds. We observe significant differences in the results obtained as a function of the different initial conditions (the number of active sources at the end of the estimation phase) of each traffic class. For example, at $t = 0.25$ *sec*, we observe that the predicted $CLR(t)$ given $Y_1(0) = 0$ and $Y_2(0) = 0$ is on the order of 10^{-16} while the steady state $CLR(t)$ is on the order of 10^{-4}. For another example, at $t = 0.25$ *sec*, we observe that the predicted $CLR(t)$ given $Y_1(0) = 20$ and $Y_2(0) = 20$ is approximately $10^{-0.6}$ while the steady state $CLR(t)$ is $10^{-4.9}$. We, therefore, conclude that the computation of predicted $CLR(t)$ is more accurate and can lead to significantly different values of the steady state $CLR(t)$ and consequently of the connection control decisions. We first compare the results obtained for the predicted $CLR(t)$ with the steady state $CLR(t)$ (which is independent of the number of active sources of each class at the beginning of the period).

In Fig. 9, we depict the predicted saturation probability as a function of the prediction time (in seconds) for various values of the initial conditions, $Y_1(0)$ and $Y_2(0)$. We observe that after approximately 4.0 seconds the predicted saturation probability converges to the steady-state saturation probability.

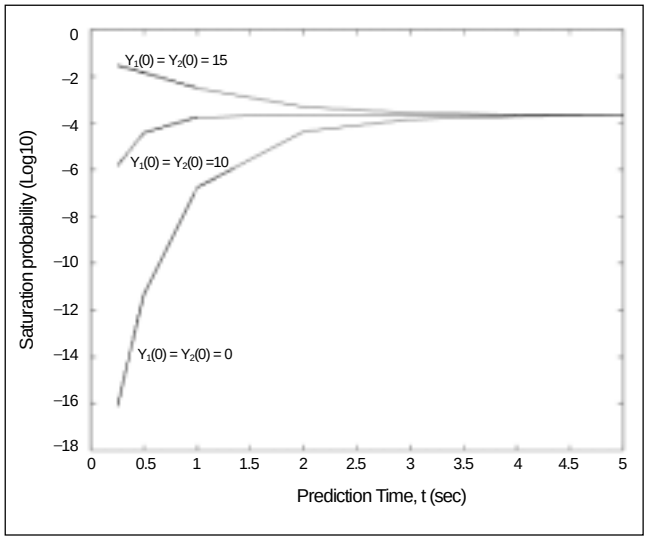

Fig. 9. Predicted saturation probability.

The predicted saturation probability is always larger than the predicted cell loss ratio. While the criterion $CLR(t)$ takes into account that not all cells are lost during overload, the criterion $P_{sat}(t)$ is based on the fraction of time during which is cell loss.

We have to satisfy the individual QoS requirements. We can find the optimal number of connections for each class under each class QoS_m requirements. In Fig. 10, we plot the predicted $ICLR_m(t)$ as a function of the prediction time (in seconds) for two classes of traffic. We observe that given values of $(\hat{\lambda}_m, \hat{\mu}_m)$ and traffic classes per connection, the predicted $ICLR_m(t)$ is different for each traffic class. This is due to the fact that the traffic class 1 with R_1= 64 kbps has a higher traffic arrival rate during "ON" period, so $ICLR_1(t)$ is larger than $CLR(t)$ and $ICLR_2(t)$. Thus, the difference between predicted $CLR(t)$ and predicted $ICLR_m(t)$ is so significant that it is necessary to make CAC, not based on average $CLR(t)$, but on individual $ICLR_m(t)$ in order to guarantee a specific QoS_m.

We also obtain the performance of the proposed CAC algorithm (as depicted in Fig. 4) to obtain the maximum number of connections, N_2^*. We assume the following network parameters: $t = 0.25$ sec, $R_1 = 64$ kbps, $R_2 = 32$ kbps, $C = 1.544$ Mbps, $i_1 = 10$, $i_2 =$

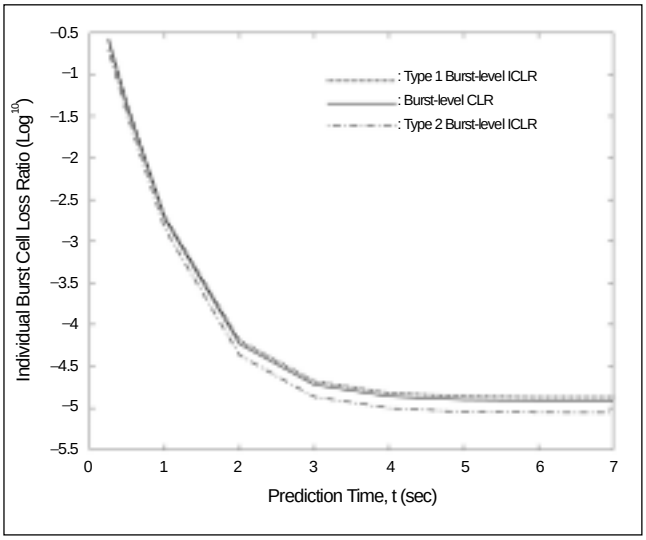

Fig. 10. Comparisons of $ICLR_m(t)$ and $CLR(t)$.

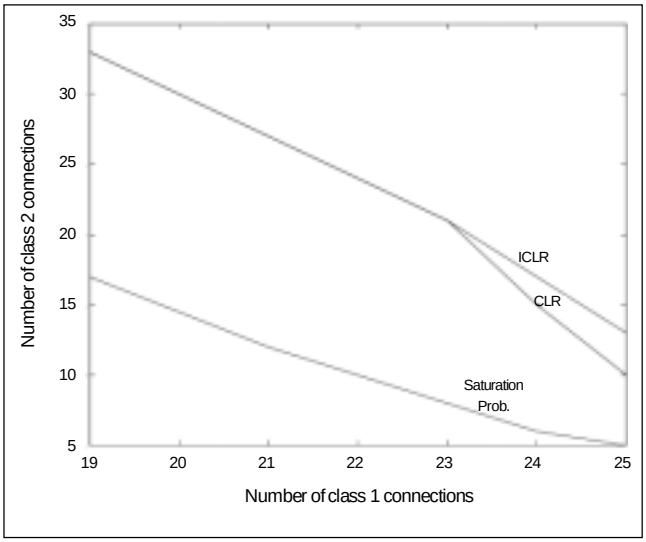

Fig. 11. The maximum number of class 2 connections versus the number of class 1 connections.

15, $\hat{\lambda}_1 = \hat{\lambda}_2 = 0.5$, $\hat{\mu}_1 = \hat{\mu}_2 = 0.833$ and $QoS_{ICLR_2(abatement)} = QoS_{P_{sat}} = 10^{-5}$. In Fig. 11, we observe that as the number of class-1 connections increases, the maximum number of class-2 connections will decrease. The peak rate strongly influences the cell loss rate and an increase in burst length cause significant increases in cell loss rate at high peak rates. There is a need to discourage long bursts of high intensity to effectively use network resources [24]. When the initial number of active sources is 20 and 20, the predictive approach determines a lower number of connections than the steady state approach. But whenever the initial number of active sources are 0 and 0, the predictive approach determines a larger number of connections than a steady state approach. We can see the difference between CLR-based CAC, saturation probability-based CAC and ICLR-based CAC. Therefore, we can clearly see a difference between predictive and steady state approaches for CAC. Generally, the number of connections based on the transient saturation probability fewer than those based on the transient cell loss ratio under the same QoS requirements ($QoS_{ICLR_2(abatement)} = QoS_{P_{sat}} = 10^{-5}$).

VII. CONCLUSIONS

In this paper, we introduced a predictive CAC architecture and algorithms for future onboard satellite systems that guarantee the QoS_m for multiple class traffic. The proposed connection control scheme considers the unique features of satellite communication; for example, buffered onboard satellite, large propagation delays, and a relatively low computational complexity. On-line measurements for established connections and traffic parameters specified by a user for new connections are used to determine the CAC based on the predicted individual cell loss ratio. We observed that after approximately 4.0 seconds the predicted $CLR(t)$ will converge to the steady state value, $CLR(\infty)$. We successfully derived predictive saturation probability, predictive burst-level cell loss ratio, and predictive individual burst-level cell loss ratio. For multiple classes of sources, numerical and simulation results show the importance of using a predictive analysis to compute the $ICLR_m(t)$ as opposed to a steady state analysis. An advantage of this predictive scheme is that the network may be more fully utilized because network states are almost the transient state, instead of the steady state.

For continuous resource management, we proposed a "sliding-window" ML estimator which can be easily implemented using a low power CMOS module for real time purposes. To avoid the low throughput due to difference between traffic parameters specified by a user prior to service and the actual cell stream, we applied the ML estimator to estimate traffic parameters. The "sliding-window" implementation effectively follows the variations in the traffic parameters. The model assumed constant parameters, but remains valid for time-varying parameters, provided the variations are slow compared to the time scale of the observations. A recursive form to save computational burden is also proposed. Therefore, the proposed ML estimator is the best estimator for a real time On-Off traffic model used for voice traffic and video telephone traffic.

Therefore, the proposed method provides an accurate and simple measurement-based congestion prediction and control scheme for execution onboard the satellite. The proposed predictive scheme is an excellent candidate for real time connection control.

REFERENCES

[1] Special issue, "Broadband via Satellite," *IEEE Comm. Magazine*, July 1997.

[2] W. D. Ivancic, M. J. Shalkhauser, and J. A. Quintana, "A Network Architecture for a Geostationary Communication Satellite," *IEEE Comm. Magazine*, July 1994.

[3] P. P. Chu, W. D. Ivancic, and H. Kim, "Onboard Closed-loop Congestion Control for Satellite Based Packet Switching Networks," *NASA Technical Memorandum 106446*, AIAA-94-1062, 1994.

[4] R. J. Gibbens, F. P. Kelly, and P. B. Key, "A Decision-theoretic Approach to Call Admission Control in ATM Networks," *IEEE JSACs*, Aug. 1995.

[5] J. W. Causey and H. S. Kim, "Comparisons of Admission Control Schemes in ATM Networks," *International Journal of Communication Systems*, Vol. 8, 1995. pp. 165–184.

[6] T. Murase and *et al*., "A Call Admission Control Scheme for ATM Networks Using a Simple Quality Estimate," *IEEE JSACs*, Dec. 1991.

[7] B. Jabbari and F. Yegenoglu, "An Efficient Method for Computing Cell Loss Probability for Heterogeneous Bursty Traffic in ATM Networks," *Int'l J. of Digital and Analog Comm. Systems*, 5:39-48, 1992.

[8] M. F. Neuts, *Matrix-Geometric solutions in stochastic models: An algorithmic approach*, Baltimore, MD: Johns Hopkins Univ. Press, 1981.

[9] D. Anick, D. Mitra, and M. M. Sondh, "Stochastic Theory of a Data-handling System with Multiple Sources," *Bell System Technical Journals*, Vol. 61, 1982, pp. 1871–1894.

[10] P. Wang and B. G. Kim, "An Estimation Technique for Cell Loss Ratio in ATM Networks with Bursty Multiclass Sources," *IEEE Globecom*, 1997.

[11] Y. M. Jang, "Connection Control and Buffer Dimensioning in Broadband Satellite Systems," *IEEE ICC*, June 1997.

[12] Y. M. Jang, A. Ganz and J. F. Hayes, "Predictive Congestion Control for Broadband Satellite System," *A Lecture Note in Computer Science Series No. 1044*, Springer-Verlag, Ed. by B. Plattner, Feb. 1996.

[13] Y. M. Jang, A. Ganz and J. F. Hayes, "A Predictive Congestion Control for Broadband Wireless LANs," *International Conference on Wireless Communications*, July 1995.

[14] H. G. Perros and K. M. Elsayed, "Call Admission Control Schemes: A Review," *IEEE Comm. Magazine*, 34(11):82-91, Nov. 1991.

[15] H. Yamanaka and et al., "Scalable Shared-buffering ATM Switch with a Versatile Searchable Queue," *IEEE JSACs*, June 1997.

[16] A.Jajszczyk, M. Roszkiewicz, and J. garcia-Haro, "Comparison of ATM Shared-memory Switches," *XV Int. Switching Symp*. (ISS '95), April. 1995, pp. 409–413.

[17] P. T. Brady, "A Statistical Analysis of On-Off patterns in 16 Conversations," *Bell System Technical Journal*, Vol. 47, Jan. 1968, pp. 73–91.

[18] P. T. Brady, "A Model for Generating On-Off Speech Patterns in Two-Way Conversation," *Bell System Technical Journals*, VOL. 48, Sept. 1969, pp. 2,445–2,472.

[19] B. Maglaris and et al., "Performance Models of Statistical Multiplexing in Packet Video Communications," *IEEE Tran. on Communications*, July 1988.

[20] D. P. Heyman and T. V. Lakshman, "What are the Implication of Long-range Dependent for VBR-video Traffic Engineering," *IEEE/ACM Tran. on Networking*, Vol. 4, pp. 301-317, June 1996.

[21] D. M. Lucantoni, M. F. Neuts, and A. R. Reibman, "Methods for Performance Evaluation of VBR Video Traffic Models," *IEEE/ACM Tran. on Networking*, Vol. 2, No.2, April 1994.

[22] Harry L. Van Trees, *Detection, Estimation, and Linear Modulation Theory*, Part I, John Wiley and Sons, 1968.

[23] N. H. E. Weste and K. Eshraghian, *Principles of CMOS VLSI Design: A System Perspective*, Addison-Wesley Publishing Company, 1993.

[24] J. Burgin and D. Dorman, "Broadband ISDN Resource Management: the Role of Virtual Paths," *IEEE Comm. Magazine*, Sept. 1991, pp. 44–48.

[25] R. Grunenfelder and L. Zubieta, "Measurement of ATM Traffic on the Cell, Burst and Activity Level by Traffic sampling," *IEEE INFOCOM*, 1992.

[26] Y. Kim, J. Kim, Y. Lee, and C.-H. Park, "Performance Analysis of the UPC/NPC Algorithm for Guaranteed QoS in ATM Networks," *ETRI Journal*, Vol. 20, No. 3, Sept. 1998.

Yeong Min Jang received the B.E. and M.E. degrees in Electronics Engineering from Kyungpook National University, Korea, in 1985 and 1987, respectively. He received the Doctor of Science degree in Computer Science from the University of Massachusetts, USA, in 1999. He worked for ETRI (Electronics and Telecommunications Research Institute), where he is a senior member of the technical staff between 1987 and 2000. Since Sept. 2000, he is with the School of Computer Science, Duksung Women's University, Seoul, Korea. He has been involved in the following research areas in which he made contribution published in over 30 papers: broadband radio access networks, satellite communications, wireless LANs, and CDMA radio networks. He is currently a member of the IEEE and KICS.