A GRACEFUL DEGRADATION METHOD FOR CONGESTION CONTROL IN WIRELESS PERSONAL COMMUNICATION NETWORKS

Nikos Passas and Lazaros Merakos

Department of Informatics, University of Athens
15784 Athens, Greece

Abstract — This paper presents an overload control method for the radio part of PCN to temporarily reduce the source rate requirements to a sustainable level, in order to avoid a sudden degradation in QoS. The control is activated when the aggregate rate crosses a predefined threshold that identifies congestion. To ensure fairness, the selection of the sources whose rate will be reduced is performed in co-operation with a priority-based scheduling technique. The performance of the system under the proposed method is analyzed and system parameter values are optimized. It is shown that the method attains considerable improvement in the loss probability performance.

I. INTRODUCTION

Wireless personal communications are considered as one of the most rapidly developing areas in telecommunications. There are no more than a few years where second-generation digital wireless systems were designed and introduced and now they are enjoying full market acceptance. The spreading of these systems shows how anxious the market is for new advanced "wireless" services. New systems, called "third-generation systems" are currently being developed to offer these services [1].

An important area of research for the support of multimedia applications is traffic scheduling. Different types of traffic, including variable-bit-rate and real-time traffic, necessitate the development of advanced scheduling techniques. Such a technique, based on the leaky bucket regulator, has been proposed in [2]. Nevertheless, bandwidth demanding sources, such as video, can readily overload the system leading to congestion. In this paper, we propose an overload control method to temporarily reduce the source rate requirements to a sustainable level, in order to avoid a sudden degradation in QoS. We call this method "graceful degradation method" because it actually acts for the benefit of the sources that reduce their requirements. The method can be easily applied to video sources resulting in a reduction of resolution or a color-to-black&white conversion, which can, in many cases, be acceptable by the end user, at least for a fraction of the time. Similar methods have been proposed for fixed ATM networks (e.g., [4], [5]).

The rest of the paper is organized as follows. Section II describes the proposed method and its interaction with the sources. In Section III, a performance analysis is presented, based on the construction of a two-dimensional Markov chain. In Section IV, numerical results, obtained from the previous analysis, are presented and analysed to quantify the effectiveness of the method. Finally, Section V contains our conclusions.

II. THE PROPOSED METHOD

The method is used in conjunction with the scheduling technique proposed in [2], and aims at improving its efficiency for better utilization of the radio medium. In brief, the leaky-bucket-based scheduling technique of [2] is applied in a TDMA-based cellular system, where each TDMA frame is subdivided into a number of request slots and a number of data slots. Request slots are used by the sources to inform the base station (BS) about the data slots they need in the next frame. The access method used for the request slots is slotted–ALOHA. The BS decides how to allocate the data slots, based on a priority mechanism similar to the well known leaky-bucket regulator proposed for fixed ATM networks. Every source is assigned a token pool, located at the BS, where tokens are generated with fixed rate equal to the mean packet rate declared by the source at connection setup. The priority of each source depends on the number of its tokens, compared to its requests. The more tokens a source has in its pool, the greater priority it has for allocation of slots, since it is below its declarations made at connection setup. For every slot allocated a token is removed from the appropriate pool. The method tries to allocate the slots of each time frame as fairly as possible, without affecting the system's performance.

During connection setup a source declares some parameters which characterize the kind of traffic that it is going to present to the network and the expected QoS. Usually they include the anticipated mean bit rate R_n and deviation σ_n, and the acceptable loss probability P_{loss}. For the graceful degradation method, a source must also declare the fraction of reduction r ($0 < r \leq 1$) that it is willing to accept (i.e., the sustainable rate), and the fraction of the time that it agrees to transmit with sustainable rate. This fraction is equal to the acceptable probability of finding

the source in sustainable state, referred to as sustainable state probability ($P_{sustain}$). Since the reduction is considered uniform, the anticipated mean bit rate in sustainable state is $R_s = (1-r)R_n$ and the deviation $\sigma_s = (1-r)\sigma_n$. When the sum of requests exceeds a predefined threshold some sources are forced to decrease their rates to sustainable. The number of these sources is the minimum needed to decrease the aggregate rate below the threshold. The inverse procedure of rate increase is performed when the system operates below the threshold.

Let $S_1, S_2, \cdots, S_k$ be the sources that request slots in a frame and let $Req(i)$ be the number of requests of source S_i. If $\sum_{i=1}^{k} Req(i) > threshold$, then some sources must reduce their rates from normal to sustainable, leading to reduction of the aggregate rate below the threshold; the question is how many and which sources will be affected. If, from the k sources, l are already in sustainable state, from previous calls of the method, then the minimum number of sources that must reduce their rates is found as follows: assumming that $S_1, S_2, \cdots, S_{k-l}$ are the sources transmitting with normal rate and $T_1 \leq T_2 \leq \cdots \leq T_{k-l}$ are the corresponding token variables, the reduction will be performed in sources S_1 through S_m where

$$m = \begin{cases} j & \text{if } \exists\ j \in [1, k-l] \text{ satisfying (1) and (2)} \\ k-l & \text{if } \exists\ j \in [1, k-l] \text{ satisfying (1) and (2)} \end{cases}$$

$$\sum_{i=1}^{j}(1-r_i)Req(i) + \sum_{i=j+1}^{k} Req(i) \leq threshold \qquad (1)$$

$$\sum_{i=1}^{j-1}(1-r_i)Req(i) + \sum_{i=j}^{k} Req(i) > threshold \qquad (2)$$

where r_i is the reduction fraction of source S_i.

The selection based on the state of the token pools is performed for two reasons. First it is fair, since at this moment the selected sources have consumed more bandwidth compared to their declarations. Second, according to the leaky-bucket scheduler, the selected sources have the lowest priority of getting slots allocated in the next frames. This means that with large probability these sources will experience a temporary degradation of the offered QoS, and that it is preferable for them to reduce their requirements until the traffic load is relaxed. This is a central point in the philosophy of the proposed method: it is preferred for a source to reduce its rate than experience more losses during overload.

As soon as the BS decides which sources must reduce their rates, it uses a special field in the downlink channel to send a control signal to them. To enforce the acceptance of its decision the BS also reduces the token generation rate of the selected sources from normal to sustainable.

The inverse procedure of increasing the bit rate of some sources is performed when it does not cause crossing of the threshold. Assume that $S_1, S_2, \cdots, S_k$ are the sources requesting slots in one frame and $\sum_{i=1}^{k} Req(i) < threshold$. If l of them are in sustainable state, then the maximum number of sources that can increase their rates without crossing the threshold is found as follows. Let $S_1, S_2, \cdots, S_l$ be the sources in sustainable state and $T_1 \geq T_2 \geq \cdots \geq T_l$ the corresponding token variables. The increase will be performed in sources S_1 through S_n, where

$$n = \begin{cases} j & \text{if } \exists\ j \in [1, l] \text{ satisfying (3) and (4)} \\ l & \text{if } \exists\ j \in [1, l] \text{ satisfying (3) and (4)} \end{cases}$$

$$\sum_{i=1}^{j} \frac{1}{1-r_i} Req(i) + \sum_{i=j+1}^{k} Req(i) \leq threshold \qquad (3)$$

$$\sum_{i=1}^{j+1} \frac{1}{1-r_i} Req(i) + \sum_{i=j+2}^{k} Req(i) > threshold \qquad (4)$$

Accordingly, the generation rate of the corresponding tokens is increased from sustainable to normal as well.

The selection procedure described above is fair and selects the sources with the highest priority of allocating slots in the next frames, which deserve a rate increase.

III. PERFORMANCE ANALYSIS

In this section, we analyze the performance of the method in a cellular environment where many sources communicate with the BS of their cell. For simplicity, the sources are considered identical, although the same analysis can be used for different sources. A well accepted model representing the aggregate bit rate of a number of independent sources is the birth-death Markov model described in [6]. In this model, state sojourn times are exponentially distributed and transmissions occur only between neighboring states, which represent rate quantization levels.

N sources are represented by $M \gg N$ minisources, where each minisource can be in one of two states: OFF, transmitting 0 packets/frame, or ON, transmitting A packets/frame. The transitions between the two states occur with rate α from OFF to ON and β from ON to OFF. According to [6],

$$\beta = a/(1 + \frac{E^2(\lambda_N)}{MC_N(0)}) \qquad \alpha = a - \beta$$
$$A = \frac{C_N(0)}{E(\lambda_N)} + \frac{E(\lambda_N)}{M}$$

$E(\lambda_N)$ is the aggregate mean rate and $C_N(t) = N\sigma^2 e^{-at}$ the aggregate autocovariance, where a is $3.9sec^{-1}$, a satisfactory value for many types of sources, especially video. The number M of minisources must be selected large enough for accurate results. Usually a value of $M = 20N$ is sufficient [6].

Assuming that feedback delay for notification of the sources is exponentially distributed [7], we develop a two-dimensional continuous-time Markov chain, such as the one shown in Fig.1. Let the mean feedback delay be $1/\gamma$. The number of rows, referred to as levels, the values of A_i, α_i, β_i for each level, and the direction of the transition arrows of γ depend on the value of the threshold.

Below we describe the construction of the two-dimensional Markov chain that models the proposed method. The chain is constructed gradually, starting from level 1, where all sources are transmitting with normal rate. The values of A_i, α_i, β_i are calculated and the directions of the arrows of γ are determined. In each level i, $E_i(\lambda_N)$ is the aggregate mean rate, $C_{Ni}(t)$ the aggregate autocovariance, N_{Ni} is the number of sources in normal state and N_{Si} the number of sources in sustainable state.

Level 1

Since all sources are in normal state, we have

$$E_1(\lambda_N) = N_{N1}R_n = NR_n$$

$$C_{N1}(0) = N_{N1}C(0) = N\sigma_n^2$$

$$A_1 = \frac{C_{N1}(0)}{E_1(\lambda_N)} + \frac{E_1(\lambda_N)}{M} = \frac{N\sigma_n^2}{NR_n} + \frac{NR_n}{M} = \frac{\sigma_n^2}{R_n} + \frac{NR_n}{M}$$

$$\beta_1 = a/(1 + \frac{E_1^2(\lambda_N)}{MC_N(0)}) = a/(1 + \frac{NR_n^2}{M\sigma_n^2})$$

$$\alpha_1 = a - \beta_1$$

If T is the threshold, the transition to the next level will be performed when the aggregate rate is above T. This means that, according to the model, the threshold is exceeded when κ_1 or more minisources are transmitting simultaneously (Fig.1), where κ_1 is the smallest integer satisfying $\kappa_1 A_1 > T$. This leads to $\kappa_1 = \lceil \frac{T}{A_1} \rceil$, where $\lceil x \rceil$ denotes the smallest integer greater than x.

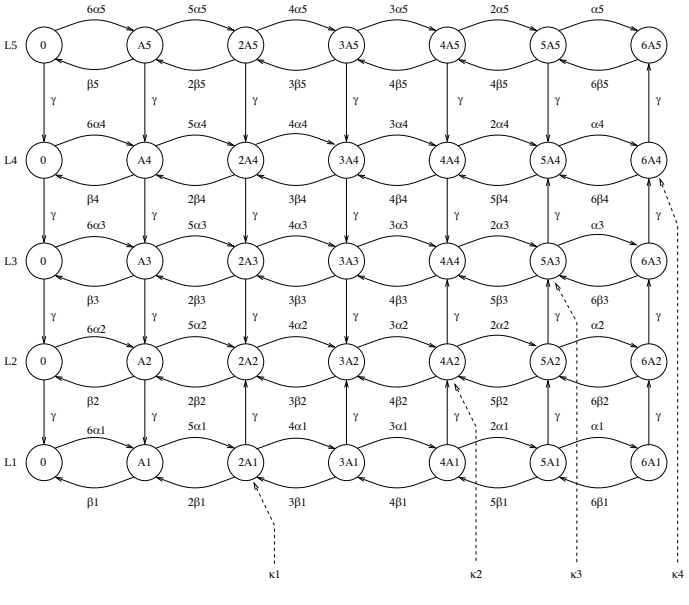

Fig.1: Two-dimensional Markov chain representing the proposed method

If $\kappa_1 \leq M$, then level 2 exists and we must proceed to the next step. If $\kappa_1 > M$ then, according to this model, the threshold cannot be crossed, even if all M minisources are transmitting simultaneously, i.e., there is no need for rate reduction and the Markov chain is actually one-dimensional.

Level i

Assuming $\kappa_{i-1} \leq M$, level i exists and we must compute the values of A_i, α_i, β_i and κ_i. To do so we must determine the number of sources that are in normal and sustainable state. Let, as in level 1,

$$E_i(\lambda_N) = N_{Ni}R_n + N_{Si}R_s$$

$$C_{Ni}(0) = N_{Ni}\sigma_n^2 + N_{Si}\sigma_s^2$$

$$A_i = \frac{N_{Ni}\sigma_n^2 + N_{Si}\sigma_s^2}{N_{Ni}R_n + N_{Si}R_s} + \frac{N_{Ni}R_n + N_{Si}R_s}{M} \quad (5)$$

Using $N_{Ni} = N - N_{Si}$ in (5) yields

$$A_i = f(N_{Si}) \doteq$$

$$\frac{N\sigma_n^2 - (\sigma_n^2 - \sigma_s^2)N_{Si}}{NR_n - (R_n - R_s)N_{Si}} + \frac{NR_n - (R_n - R_s)N_{Si}}{M} \quad (6)$$

Additionally,

$$\beta_i = \frac{a}{1 + \frac{E_i^2(\lambda_N)}{MC_{Ni}(0)}} = \frac{a}{1 + \frac{(NR_n - (R_n - R_s)N_{Si})^2}{M(N\sigma_n^2 - (\sigma_n^2 - \sigma_s^2)N_{Si})}} \quad (7)$$

$$\alpha_i = a - \beta_i \quad (8)$$

We must calculate the minimum value of N_{Si} that reduces the total rate below the threshold T. In level i the aggregate rate, when κ_{i-1} minisources are transmitting, must not exceed T, meaning: $A_i \leq \frac{T}{\kappa_{i-1}}$. It can be shown that the function $f(x)$, defined in (6), is strictly decreasing for $x \in [0, N]$. We distinguish two cases.

Case 1: $\frac{T}{\kappa_{i-1}} < f(N)$

In this case, it is impossible for the product $\kappa_{i-1}A_i$ to fall below the threshold T, even if all sources are transmitting with sustainable rate, since N is the maximum value that N_{Si} can take. Nevertheless, all sources are transferred to sustainable state to relax the heavy traffic load. From equations (6), (7), (8) we can compute the parameters of level i by setting $N_{Si} = N$.

Case 2: $f(N) \leq \frac{T}{\kappa_{i-1}} \leq f(0)$

There is only one solution of $f(x) = T/\kappa_{i-1}$ in the interval $[0, N]$, referred as x_s. Thus, the minimum acceptable N_{Si} is $N_{Si} = \lceil x_s \rceil$. Having N_{Si} we can compute, according to (6), (7), (8), the parameters of level-i. The number κ_i is found as in level 1: $\kappa_i = \lceil \frac{T}{A_i} \rceil$. If $\kappa_i \leq M$ and $N_{Si} < N$, we continue the procedure with level $i + 1$. If $\kappa_i > M$, threshold T cannot be crossed, and if $N_{Si} = N$ then all sources are in sustainable state. In both cases, level i is the last level of the chain.

Each level in the above analysis is equivalent to a model in which a buffer receives packets from a finite number of statistically independent and identical information sources that asynchronously alternate between exponentially distributed periods in the ON and OFF states. The buffer is

connected to an output channel with known capacity [8]. The output rate is equal to the number of slots per frame, referred to as B. Assuming time-of-expiry for each packet between two and three frames, there is no time for retransmissions [2]. Thus, the buffer size in the equivalent model is equal to 0.

According to the analysis in [8] and [9], we can compute the expected overflow probability of the buffer in the equivalent model, which is equal to the loss probability in our model, as follows. Let $\mathbf{T}(p,q)$ be the (p,q) element of rate transition matrix $\mathbf{T}$ of the system Markov chain; we have

$$\mathbf{T}(p,q) = \begin{cases} t_{nA_i,mA_j} & p \neq q, \quad p \leftrightarrow nA_i, \quad q \leftrightarrow mA_j \\ -\sum_{mA_j} t_{nA_i,mA_j} & p \neq q, \quad p \leftrightarrow nA_i, \quad mA_j \in V_{nA_i} \end{cases}$$

where V_{nA_i} is the set of states from where state nA_i can be reached in one transition.

The drift matrix $\mathbf{D}$ is

$$\mathbf{D}(p,q) = \begin{cases} n \cdot A_i - B & p = q, \quad p \leftrightarrow nA_i \\ 0 & p \neq q \end{cases}$$

Then the loss probability is equal to [9]:

$$P_{loss} = \sum_{nA_i} \sum_{j:z_j<0} a_j \cdot \Phi_j^{nA_i}$$

where the pairs $(z_j, \mathbf{\Phi}_j)$ are solutions to the eigenvalue problem: $\mathbf{D}z\mathbf{\Phi} = \mathbf{T}^T\mathbf{\Phi}$ and the coefficients a_j are obtained from boundary conditions.

The sustainable state probability is given by

$$P_{sustain} = \sum_{i=1}^{M} \sum_{j=1}^{L} P(iA_j) \cdot \frac{N_{Sj}}{N}$$

where L is the number of levels of the Markov chain, M the number of minisources, N_{Sj} the number of sources transmitting with sustainable rate in level j and $P(iA_j)$ the equilibrium probability of state iA_j.

Finally the channel utilization is given by

$$U_{CH} = \frac{E[v] \cdot (1 - P_{loss})}{B} \qquad (9)$$

where $E[v] = \sum_{nA_i} nA_i P(nA_i)$ is the mean aggregate rate.

IV. NUMERICAL RESULTS AND DISCUSSION

To show the improvement that the proposed method can attain, we consider a TDMA-based cellular environment with channel capacity of 1.92 Mb/sec and frame size of 12msec [10]. For compatibility with fixed ATM networks we use a slot size of 53 bytes to fit an ATM cell [2]. In this system we consider variable bit rate sources with mean bit rate 128 Kb/sec and deviation 64 Kb/sec. These characteristics can fit to low-resolution compressed video sources. The acceptable reduction fraction for every source is 0.5.

From the analysis we can compute the interval of threshold values that satisfy the agreements for both P_{loss} and $P_{sustain}$. As shown in Fig.2, the interval depends on the number of sources n, and it is between the value that corresponds to the maximum acceptable sustainable state probability, referred to as $t_s(n)$, and the value that corresponds to the maximum acceptable loss probability, referred to as $t_l(n)$.

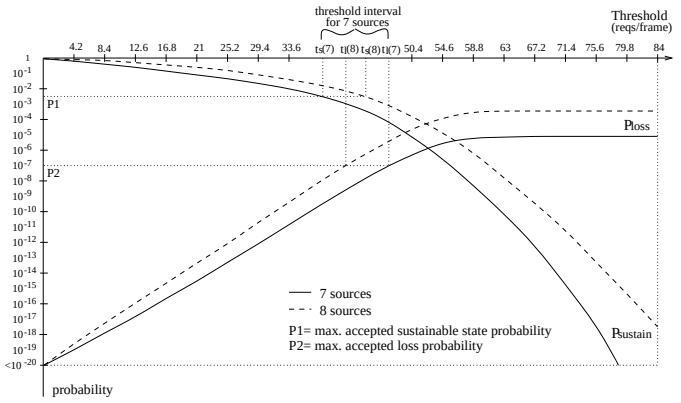

Fig.2: Loss and sustainable state probability for different values of the threshold

Clearly, we must have $t_s(n) \leq t_l(n)$. If $t_s(n) > t_l(n)$, then there is no value for the threshold that can simultaneously satisfy the agreed values for P_{loss} and $P_{sustain}$. For a small number of sources the interval is quite large but, as more sources are entering the system, the interval gets smaller until $t_s(n) > t_l(n)$. In this way, we can compute the maximum number of sources, referred to as N_{max}, that can be supported by the system. As shown in Fig.2, the system under study can support up to seven sources, since $t_s(7) < t_l(7)$ and $t_s(8) > t_l(8)$ and the optimal threshold is in the interval $[t_s(7), t_l(7)]$.

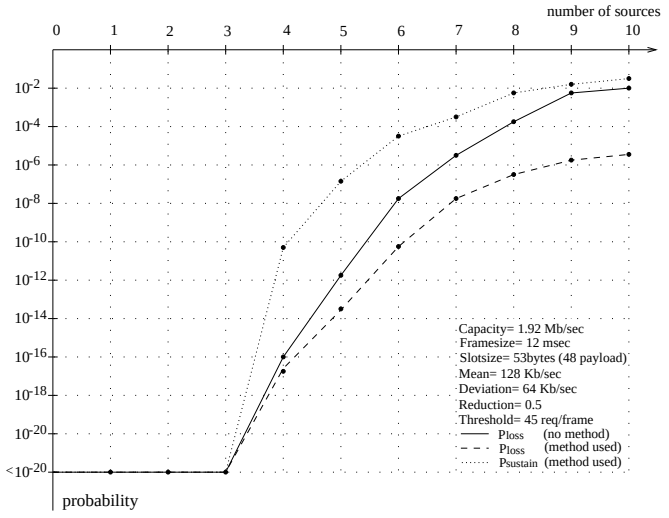

Fig.3: Loss and sustainable state probability for constant threshold

Fig.3 shows the gain attained in loss probability when a constant threshold of 45 requests/frame is used for d-

ifferent numbers of sources. As we can, see the gain increases as more sources are entering the system. This is explained as follows. When the number of sources is small (≤ 3) the particular threshold is hard or impossible to be reached, but, as the load increases the threshold is more often reached and a selected number of sources each time are forced to transmit with sustainable rate.

In Fig.4, we can observe the maximum gain attained in loss probability, provided that the sustainable state probability has a guaranteed constant value equal to the maximum acceptable value agreed at connection setup. To do this we use different thresholds for different numbers of sources. For a given number of sources the threshold used is the one that leads to the guaranteed value of sustainable state probability. As shown in Fig.2, this threshold leads to the minimum obtainable loss probability, thus to the maximun attainable gain for a given number of sources. Note also in Fig.4 that the largest improvement occurs at medium loads. For light loads (number of sources ≤ 3) there is very little space for improvement in loss probability. When the load is very heavy, the network is congested, regardless of the bit rate of the sources (normal or sustainable).

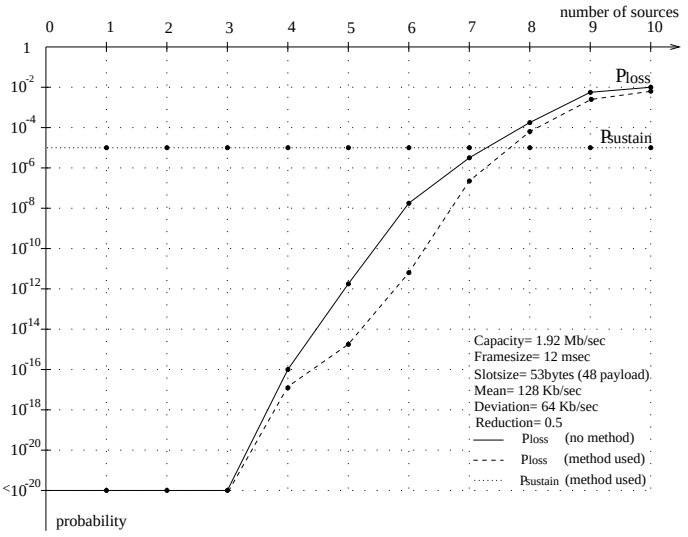

Fig.4: Loss and sustainable state probability for variable threshold

V. CONCLUSIONS

We have presented a method for congestion control in the radio interface of wireless personal communication networks (PCN) where multimedia traffic is considered. The method monitors the total bit rate of all active sources and, when this rate exceeds a predefined threshold, forces some sources to reduce their bandwidth requirements in order to reduce the total rate below the threshold.

With an extensive performance analysis we have quantified the improvement that the method can attain in loss probability, at the expense of a controlled quality reduction for an acceptable fraction of time. We have also found the optimal threshold values in order to maximize system performance. The obtained results show that the proposed method is an effective tool for controlling congestion, especially when variable bit rate, loss sensitive sources (e.g., video) are involved.

References

[1] J.S. DaSilva, B.E. Fernandes, "The European Research Program for Advanced Mobile Systems", IEEE Personal Commun., Feb. 1995.

[2] N. Passas, N. Loukas, L. Merakos, "A Leaky-Bucket-Based Scheduling Technique for Wireless Personal Communication Networks", Proc. ICT'96, Istanbul, Apr. 1996.

[3] M. Schwartz, "Network Management and Control Issues in Multimedia Wireless Networks", IEEE Personal Commun., June 1995.

[4] Belkacem Kraimeche, Mischa Schwartz, "Analysis of Traffic Access Control Strategies in Integrated Service Networks", IEEE Trans. on Communications, vol.COM-33, no.10, Oct. 1985.

[5] Nanying Yin, Michael G. Hlychyj, "On Closed-Loop Rate Control for ATM Cell Relay Networks", IEEE INFOCOM'94.

[6] B. Maglaris et al., "Performance Models of Statistical Multiplexing in Packet Video Communications", IEEE Trans. on Commun., vol.36, no.7, July 1988.

[7] N. Yin, M.G. Hlychyj, "A Dynamic Rate Control Mechanism for Source Coded Traffic in a Fast Packet Network", Proc. IEEE INFOCOM'91.

[8] D.Anick, D.Mitra, M.M. Sondhi, "Stochastic Theory of a Data-Handling System with Multiple Sources", Bell Sys. Tech. Journal, vol.61, no.2, Oct. 1982.

[9] L.K. Reiss, L.F. Merakos, "Adaptive Bandwidth Reservation for Traffic Streams Sharing an ATM Virtual Path", Proc. IEEE GLOBECOM'93.

[10] D. Raychaudhuri, N.D. Wilson, "ATM-Based Architecture for Multiservices Wireless Personal Communication Networks", IEEE JSAC, vol.12, no.8, Oct. 1994.