ELSEVIER

Computer Communications 23 (2000) 942–949

computer communications

www.elsevier.com/locate/comcom

An adaptive approach to manage traffic in CDMA ATM networks

M.S. Obaidat[a,*], C.B. Ahmed[b], N. Boudriga[b]

[a]*Department of Computer Science, Monmouth University, W. Long Branch, NJ 07764, USA*
[b]*School of Telecommunications, University of Tunis, Tunis, Tunisia*

Abstract

Wireless Asynchronous Transfer Mode (*WATM*) systems are becoming an important field in recent days due to their increasing use and applications. A WATM system is the result of synergy between *wireless networks and asynchronous transfer mode* (*ATM*) *technologies*. *Designing an* efficient mobile communication system is considered as a challenging task, given its complexity and operation environments. A major task activity related to these systems addresses the call admission control and resource management.

In this paper, we consider these issues in wireless ATM networks integrating the Code Division Multiple Access (*CDMA*) technology. An approach for a dynamic resource management is proposed using heterogeneous traffic descriptors as well as the concept of signal-to-interference rate, computed at the base station receivers. An adaptive monitoring scheme that is based on an estimation algorithm and driven by a measurement of the signal-to-interference and predicted traffic parameters of the admitted connections is established. The Dynamic control that we propose at the user network interface, UNI, provides information about the instantaneous bit rate of a source allowing more effective flow control and achieves a good match in terms of predicting the congestion of any switch. It is able to police implicit resource management, which is used for both Constant Bit Rate (*CBR*) and Variable Bit Rate (*VBR*) traffics as well as explicit resource management that is used for Available Bit Rate (*ABR*) traffic.

Keywords: Traffic engineering; Effective bandwidth; Energy function; Entropy function; Code Division Multiple Access

1. Introduction

In recent days, there is an increasing interest in integrating mobility to Asynchronous Transfer Mode (*ATM*) networks. However, researchers and developers are facing several problems that complicate the task of integration. One of these problems is handoff and rerouting of connection with Quality of Service (*QoS*) guarantees. This problem has become critical in the inherently connection-oriented ATM systems and is the major concern of this paper.

Recent advancement in communications and computers technologies have enabled high-speed networks to support real-time constraints and have encouraged the implementation of sophisticated applications. Next generation wireless networks will support not only data and voice services but also multimedia applications that will need sophisticated error control, efficient synchronisation control, and flexible resource allocation and management. Existing ATM networks are designed to support wireline-users with fixed locations, and offer no capability of connection setup, handoff, cell forwarding, and re-routing functions that the wireless users might need during their mobility. Moreover, they tend to be susceptible to congestion with mobile environment. Congestion may, in that case, reduce the ATM's high QoS by increasing the cell loss and delays to unacceptable levels. Effective policing of traffic can prevent congestion from occurring; and therefore, a policing function that can control traffic to the necessary reliability level (i.e. a congestion probability close to zero) is crucial requirement. One such policing requirement, known as the leaky bucket policing function, has the potential to control the critical demand for traffic at the UNI interface [1–5]. There are a variety of leaky bucket algorithms, each is suited to a particular traffic type of those depicted by the ATM Adaptation Layers AAL1-5, [1,2,5]. The most interesting schemes are those algorithms that can be used for bursty traffics. The present challenge for any one of these algorithms is the ability to be adapted to the varying characteristics of bursty traffic in wireless environment, particularly when the Code Division Multiple Access (*CDMA*) technology is used for wireless systems.

The growth of wireless communications paired with the rapid developments in ATM networking technology foresees a new era in telecommunications. Additionally, demand for cellular communications has placed a heavy demand on the capacity of wireless/air interfaces and the

* Corresponding author. Tel.: + 1-732-571-4482; fax: + 1-732-263-5202.

E-mail address: obaidat@monmouth.edu (M.S. Obaidat).

0140-3664/00/$ - see front matter
PII: S0140-3664(00)00176-6

network resources available. The success of cellular mobile communications has helped the implementation of Personal Communications Services (*PCS*). PCS will provide voice, text, video and data, and as a consequence, demand for higher transmission speed and mobility is even greater. Congestion in Wireless ATM systems is becoming an important issue, particularly when the CDMA technology is used.

In this paper, we propose to develop a dynamic monitoring scheme of traffic that provides information about the instantaneous bit rate of a source, allowing more effective flow control and achieving a good match in terms of predicting the congestion state of the connection source. We show that our scheme is able to police implicit resource management (which is used for both CBR and VBR traffic types) as well as explicit resource management for Available Bit Rate (*ABR*) traffic and Unspecified Bit Rate (*UBR*) traffic. Our approach is based on the large deviation techniques that have been developed in a general mathematical study of the tail of the steady-state queue length distribution. The general framework of the theory that we develop in this paper includes the computation of the effective bandwidth, which was first proposed in Refs. [6–8]. Our prime objective in this work is to develop a theory for dynamic and predictable traffic regulation.

2. The request for additional resources

Many publications have dealt with bandwidth utilization in CDMA ATM networks [9–12]. Most of them have focused on monitoring traffic without considering effectively the ATM resources. In our paper, we develop a control monitoring scheme that takes into consideration the information about channel characteristics pertinent to CDMA, and the information about connection traffic parameters pertinent to ATM resources (bandwidth and buffer storage).

Consider a set of base stations providing access to a network. When a certain connection needs more resources for a certain period of time, the Dynamic Leaky Bucket (*DLB*) *algorithm* has to determine whether this request can be accepted or not. The decision criterion is to accept the request if, in the new network state, the expected information Bit Error Rate (*BER*), across all existing connections will not exceed the QoS threshold, and if the amount of requested resources (bandwidth and buffer) are available. Two categories of information flow can be considered for the DLB algorithm. The first category is concerned with information about connection traffic parameters. We will show later, how it is possible to establish relationships between the traffic descriptors and the BER. The second category concerns information about channel characteristics pertinent to BER. Such information is available in the local base station. One possible information is the interference to signal ratio value. The admission decision is based on both the source traffic declarations and the channel characteristics measured at the local base station. The main objective of this paper is to design both optimal Call Admission Control and DLB algorithms, with respect to physical resources utilization, and channel characteristics (BER, and interference magnitude) under QoS constraints.

To make an admission decision based on the estimated interference parameters, one needs to establish the relation between the interference and the information of BER for which the constraint is imposed. In our approach, we use the *adaptive* signal-to-interference threshold $\Psi_{\min}$, that corresponds to the bit information, which ensures that the BER constraint is met if signal-to-interference samples is smaller than the threshold. We assume that if the signal-to-interference ratio is greater than its threshold, then the receiver will not demodulate exactly the bit. According to the QoS requirements. We can assume that each base station is allowed to loose a number of bits b_s, for each connection that is handled, and depending on the traffic class s, $1 \leq s \leq S$. The admission decision of requesting additional resources is based on the computing of the effect of additional resources on the signal-to-interference ratio, for the connections that have the number of lost bits close to their threshold, as shown below.

- If $\Psi \geq \Psi_{\min}$, then the request for additional resources is authorized (Ψ is the new value of the signal-to-interference ratio when the additional resource is considered).
- If $\Psi < \Psi_{\min}$, and $b \leq b_s$, where b denotes the number of lost bits corresponding to the most critical connection (i.e. the connection that has the number of lost bits close to its threshold), when the request of additional resources is considered then the request for additional resource is authorized.
- If $\Psi < \Psi_{\min}$, and $b > b_s$ then the request for additional resources is rejected or partially accepted (later we will show how the second decision can be made).

3. Traffic description and resource allocation

We assume that we allow statistically identical traffic types with the same QoS requirements to share buffers and bandwidths. Each traffic type, characterized by some QoS requirements, belongs to some traffic class s, $1 \leq s \leq S$, where S denotes the total number of classes. We denote by $N_s(t)$ the number of traffic connections, which belong to the sth class of traffic at time t. At each base station, we assume that the network supervisor attributes to the sth aggregate traffic a bandwidth equal to b_s, and a capacity buffer equal to m_s. Later, we will show how to compute both resources.

Now, consider a connection requesting a QoS requirement of class s. Our preliminary QoS goal is to limit large delays or ensure that cell loss probability is quite small. In

order to do so, we require that

$$P(X_s(t) > m_s) \leq \xi,$$

where $X_s(t)$ is a stochastic process distributed as the buffer's stationary workload of the sth aggregate traffic, and ξ is a positive small real number that is fixed by the network supervisor. When $N_s(t)$ connections are handled by a base station at some time t, each one is requesting a buffer storage equal to $m_{i,s}$. The constraint $1 - P(X_s(t) > m_s) \geq 1 - \xi$, is equivalent to $1 - \prod_{i=1}^{N_s(t)} P(X_{i,s}(t) > m_{i,s}) \geq 1 - \xi$, where $X_{i,s}(t)$ is a stochastic process distributed as the buffer's stationary workload of the ith connection of sth aggregate. Assuming that the set of stochastic processes $\{X_{i,s}(t)\}_{1 \leq i \leq N_s(t)}$, have similar statistical characteristics, we can easily deduce that:

$$P(X_{i,s}(t) > m_{i,s}) \leq \frac{\xi}{N_s(t)} \equiv \mathrm{e}^{-m_{i,s}\lambda_{i,s}(N_s(t))}$$

where the parameter $\lambda_{i,s}(N_s(t))$ depends on time t and $N_s(t)$, determines the instantaneous stringency (strictness) of the QoS requirement corresponding to the ith connection of the sth aggregate.

We propose to view each source traffic i, like a particle in statistical mechanics, [9–11]. We assume that this source traffic behaves as a constant rate flow with a rate r for a period of time t with probability $p_{i,s}(r,t)$. Using the results of the large deviation theory, the probability density function $p_{i,s}(r,t)$ can be shown to have the form of Gib's distribution, say $\mathrm{e}^{-t\beta^*_{i,s}(r)}$, where $\beta^*_{i,s}(r)$ is the "*entropy function*", [9], of the ith connection belonging to the sth aggregate traffic. This *entropy function* is obtained from the Legendre transform of the energy function $\beta_{i,s}(\lambda_{i,s}(N_s(t)))$:

$$\beta^*_{i,s}(r) = \max_x(rx - \beta_{i,s}(x)).$$

The instantaneous energy function $\beta_{i,s}(\lambda_{i,s}(N_s(t)))$ can be determined by using the principle of large deviation of Gartner–Ellis, [9–11]. Using the results mentioned in [13], we can easily prove that:

$$\beta_{i,s}(\lambda_{i,s}(N_s(t))) = \log\left[\sum_{k=0}^{\infty} \frac{[\lambda_{i,s}(N_s(t))]^k \mu^k_{i,s}}{k!}\right],$$

where

$$\mu^k_{i,s} = \lim_{t\to\infty} \frac{\{E[X_{i,s}(t)]\}^k}{t},$$

and $X_{i,s}(t)$ denotes the number of cell arrivals of the ith connection belonging to the sth class, in the interval of time $[0,t]$. Note that $\mu^1_{i,s}$ corresponds to the bandwidth needed for the *CBR* traffic, which is indicated in the QoS contract of the ith connection. For $k \geq 2$, we can easily prove that $\mu^k_{i,s}$ can be deduced in terms of $\mu^1_{i,s}$ as shown below:

$$\mu^k_{i,s} = \sum_{j=0}^{k} \binom{k}{j} (\mu^1_{i,s})^{k-j} \delta^k_{i,s},$$

where

$$\delta^k_{i,s} = \lim_{t\to\infty} \frac{E\{[x_{i,s}(t) - E[x_{i,s}(t)]]^k\}}{t}.$$

We denote by $b_{i,s}(\lambda_{i,s}(N_s(t)))$ the instantaneous effective bandwidth of ith connection belonging to the sth aggregate traffic. It is given by:

$$b_{i,s}(\lambda_{i,s}(N_s(t))) = \frac{\beta_{i,s}(\lambda_{i,s}(N_s(t)))}{\lambda_{i,s}(N_s(t))}.$$

The probability density function of the ith connection belonging to the sth aggregate traffic is given by:

$$p_{i,s}(r,t) = \mathrm{e}^{-tB^*_{i,s}(r)}.$$

In order to investigate the performance evaluation of the system, we need to characterize the number of cells related to the ith connection belonging to the sth aggregate traffic, and present at time t. The following relation explains this.

$$x_{i,s}(\lambda_{i,s}(N_s(t))) = \max(0, (r - b_{i,s}(\lambda_{i,s}(N_s(t))))t),$$

Once $p_{i,s}(r,t)$ is computed, we can deduce the probability that there are at least $\alpha_{i,s}$ cells from the sth aggregate traffic, present at time t in the buffer of the ith base station:

$$P(x_{i,s}(\lambda_{i,s}(N_s(t)))) \geq \alpha_{i,s}) = \int_{r=\frac{\alpha_{i,s}}{t} + b_{i,s}(\lambda_{i,s}(N_s(t)))}^{\infty} \mathrm{e}^{-tB^*_{i,s}(r)}\,\mathrm{d}r$$

There are three parameters to specify the quality of service. It turns out that all of them are related to the probability distribution of the queue length $P(x_{i,s}(\lambda_{i,s}(N_s(t)))) \geq \alpha_{i,s})$. A certain quality of service is maintained if all three parameters are below the bounds specified by the user of the network. The three parameters are:

- Cell loss ratio (*CLR*): This is the fraction of cells that are lost due to buffer overflow. This obviously relates to the probability distribution of the queue length.
- Mean cell delay (*MCD*): This is the average time each cell spends in the queue. The mean delay is therefore directly related to the expected value of the queue length $E\{x_{i,s}(\lambda_{i,s}(N_s(t)))\}$.
- Cell delay variance (*CDV*): This is the square of the standard deviation of the cell delays from their mean. It is related to the variance of the queue length.

The instantaneous CLR corresponding to the ith connection that belongs to the sth aggregate, is defined as the expected number of cell losses due to buffer overflow divided by the expected number of the arriving cells:

$$\mathrm{CLR}_{i,s}(\lambda_{i,s}(N_s(t))) = \int_{r=\frac{m_s}{t} + b_{i,s}[\lambda_{i,s}(N_s(t))]}^{\infty} \mathrm{e}^{-tB^*_{i,s}(r)}\,\mathrm{d}r.$$

The instantaneous MCD corresponding to the ith connection and belonging to the sth aggregate is simply the length of the queue divided by the service rate. The MCD, is the expected

value of this delay:

$$\mathrm{MCD}_{i,s}(\lambda_{i,s}(N_s(t))) = \frac{E[x(\lambda_{i,s}(N_s(t)))]}{b_{i,s}(\lambda_{i,s}(N_s(t)))},$$

where

$$E[x(\lambda_{i,s}(N_s(t)))] = \sum_{q=0}^{m_s-1} qP[x(\lambda_{i,s}(N_s(t))) = q],$$

and $P[x(\lambda_{i,s}(N_s(t))) = q]$ is given by $P[x(\lambda_{i,s}(N_s(t))) \geq q] - P[x(\lambda_{i,s}(N_s(t))) \geq q-1]$. We can easily prove that:

$$\mathrm{MCD}_{i,s}(\lambda_{i,s}(N_s(t))) \approx \frac{\sum_{q=0}^{m_{i,s}-1} P[x(\lambda_{i,s}(N_s(t))) \geq q]}{b_{i,s}(\lambda_{i,s}(N_s(t)))}.$$

By the same procedure we can find that the CDV, is given by:

$$\mathrm{CDV}_{i,s}(\lambda_{i,s}(N_s(t))) = \frac{E\{[x(\lambda_{i,s})]^2\} - \{E[x(\lambda_{i,s})]\}^2}{[b_{i,s}(\lambda_{i,s})]^2},$$

$$\approx \frac{\sum_{q=0}^{m_s-1}(2q+1)P[x(\lambda_{i,s}) \geq q] - \left\{\sum_{q=0}^{m_s-1} P[x(\lambda_{i,s}) \geq q]\right\}^2}{[b_{i,s}(\lambda_{i,s})]^2}.$$

We now compute the stringency parameter, $\lambda_{i,s} = \lambda_{i,s}(N_s(t))$. The relationship between $m_{i,s}(\lambda_{i,s}(N_s(t)))$ and $b_{i,s}(\lambda_{i,s}(N_s(t)))$ is simply given by:

$$m_{i,s}(\lambda_{i,s}(N_s(t))) = b_{i,s}(\lambda_{i,s}(N_s(t)))\mathrm{MCD}_{i,s}(\lambda_{i,s}(N_s(t))).$$

Using the constraint given earlier, we can easily prove that:

$$\lambda_{i,s}(N_s(t)) = \log\left[\left(\frac{N_s(t)}{\xi}\right)^{m_{i,s}(\lambda_{i,s}(N_s(t)))^{-1}}\right].$$

Therefore, the following procedure can be used.

Algorithm 1:. while $|x_n - x_{n-1}| > \epsilon$,

$$\text{compute } x_n = -\frac{\log\left[\left(\frac{\xi}{N_s(t)}\right)\right]}{m_{i,s}(x_{n-1})}.$$

4. CDMA system description

We will consider a CDMA system where the received signal in the channel is the sum of the transmission due to all users and additive white Gaussian noise. The received signal at time t, due to the uth user which belongs to the sth class, is given by; $R_{us}(t) = \sqrt{2P_{us}(t)}\cos(\omega_C + \phi_{us})\sum_{j=1}^{+\infty} b_{us}(j)w_{us}(t - jT)$, where:

- $P_{us}(t)$ is the instantaneously power signal of uth user,
- $b_{us}^i \in \{-1, 1\}$ is the ith bit of uth user,
- T is the bit period,
- $w_{us}(t)$ is the instantaneously spreading waveform of uth user,
- ω_C is the carrier frequency, and
- ϕ_{us} is the carrier phase of uth user.

The spreading waveform of the uth user is given by:

$$w_{us}(t) = \sum_{n=1}^{N} w_{us}(n)\tau(t - nT_C),$$

where $w_{us}(n)$ is the nth element of the sequence for uth user, $\tau(t)$ is a unit pulse of length T_C, and N is the processing gain given by $N = T/T_C$.

Due to the fading phenomena, the received signal in the channel at the base station is given by:

$$r(t) = \sum_{s=1}^{S}\sum_{u=1}^{N_s} \alpha_{us}R_{us}(t) + \varpi(t),$$

where α_{us} is the attenuation coefficient due to the Rayleigh fading, and $\varpi(t)$ is the white Gaussian noise. The nth sample, $1 \leq n \leq N$, of the uth user belonging to the sth class, at the output of the filter is given by:

$$r_{us}(n) = \sqrt{2}\int_{nT_C}^{(n+1)T_C} r(t)\tau(t - nT_C)\cos(\omega_C t + \phi_{us})\,\mathrm{d}t,$$

where the above signal has been converted to baseband via filters.

In a conventional matched filter detector, these samples are multiplied by the spreading sequence for the desired user, and are given by:

$$x_{us}[n] = w_{us}[n]\left\{\alpha_{us}\sqrt{P_{us}}w_{us}[n]T_C\sum_{j=1}^{+\infty} b_{us}(j) + \varpi[n]\right\} + \sqrt{2}w_{us}[n]$$

$$\times\left\{\sum_{j=1}^{+\infty}\sum_{s=1}^{S}\sum_{\substack{z=1\\ z\neq u}}^{N_s} \alpha_{zs}b_{zs}(j)\sqrt{P_{zs}}\alpha_{zs}\cos(\phi_{zs} - \phi_{zs})\right.$$

$$\left.\times\int_{nT_C}^{(n+1)T_C} w_{us}(t - jT)\cos(2\omega_C t + \phi_{us} + \phi_{us})\,\mathrm{d}t\right\}.$$

Thus, we can deduce the expression of the signal-to-interference ratio related to the bit information, of the uth user belonging to the sth traffic class:

$$\Psi_{us} = \frac{\alpha_{us}\sqrt{P_{us}}\tilde{w}_{us}T_C}{\sqrt{2}\left\{\sum_{s=1}^{S}\sum_{\substack{z=1\\ z\neq u}}^{N_s} \alpha_{zs}b_{zs}(j)\sqrt{P_{zs}}\alpha_{zs}\cos(\phi_{zs} - \phi_{us})I_{zs}\right\}},$$

where

$$\tilde{w}_{us} = \frac{\sum_{n=1}^{N} w_{us}[n]}{N},$$

and

$$I_{zs} = \int_{nT_C}^{(n+1)T_C} w_{zs}(t - jT)\cos(2\omega_C t + \phi_{us} + \phi_{zs})\,dt.$$

We need now to express the instantaneous power of the signal of a connection i, belonging to the sth class, in terms of the instantaneous stringency of the QoS requirement $\lambda_{i,s}(N_s(t))$. First, we need to compute the autocorrelation function of the stochastic process $x_{i,s}(\lambda_{i,s}(N_s(t)))$ given by:

$$R_{is}^{x}(\lambda_{i,s}(N_s(t_1)), \lambda_{i,s}(N_s(t_2)))$$

$$= E(x_{i,s}(\lambda_{i,s}(N_s(t_1)))x_{i,s}(\lambda_{i,s}(N_s(t_2)))),$$

Let $R_r(t_1, t_2) = E(r(t_1)r(t_2))$, be the autocorrelation function of the process $r(t)$, and $\mu_r(t) = E(r(t))$ be its instantaneous mean. Therefore, we can easily prove that:

$$R_{is}^{x}(\lambda_{i,s}(N_s(t_1)), \lambda_{i,s}(N_s(t_2)))$$

$$= t_1 t_2\{R_r(t_1, t_2) - b_{is}(\lambda_{i,s}(N_s(t_1)))b_{is}(\lambda_{i,s}(N_s(t_2)))$$

$$\times \mu_r(t_1)b_{is}(\lambda_{i,s}(N_s(t_1)))\mu_r(t_2)b_{is}(\lambda_{i,s}(N_s(t_2)))\},$$

The Fourier integral of the stochastic process $x_{i,s}(\lambda_{i,s}(N_s(t)))$, is a stochastic process in the variable ω given by:

$$N_{is}(\omega, -\omega) = \int\int_{t_1,t_2} R_{is}^{x}(\lambda_{is}(N_s(t_1)), \lambda_{is}(N_s(t_2)))\, e^{-j\omega(t_1 - t_2)}\,dt_1\,dt_2,$$

Thus, the instantaneously power signal of the ith user, $P_{is}^{u}(t) = P_{is}^{u}(b_{is}(\lambda_{is}(N_s(t))))$, when it is requesting bandwidth $b_{is}(\lambda_{is}(N_s(t)))$ at time t, is given by:

$$P_{is}^{u}(t) = \frac{1}{2\pi}\int_0^{b_{is}(\lambda_{is}(N_s(t)))} N_{i,s}(\omega, -\omega)\,d\omega.$$

5. Usage parameter control

In order to prevent gross misuse of network resources, it is reasonable to include a mechanism for peak rate policing at all access points for a public ATM network. However, as ATM provides bandwidth on demand, peak rate policing will not suffice to ensure QoS and fairness to other users sharing buffers and bandwidth. Consequently, connections violating agreed upon traffic descriptors must also be throttled into compliance or penalized accordingly. We now describe an approach to accomplish this task.

Let $b_{is}(\lambda_{i,s}^{j}(N_s^{j}(0)))$ be the user specified bandwidth of a connection i, $1 \le i \le N_s^{j}(0)$, that belongs to traffic class s, and passing through base station j, at time 0. Let $N_s^{j}(0)$ be the number of connections belonging to the sth class and handled by the jth base station at time $t = 0$), and $b_{is}(\lambda_{is}^{j}(N_s^{j}(t)))$ the effective bandwidth of the departure process from the leaky bucket policy at time t. A connection j is said to have violated its contract at time t, if:

$$b_{is}(\lambda_{i,s}^{j}(N_s^{j}(t))) > b_{is}(\lambda_{is}^{j}(N_s^{j}(0)))$$

which is equivalent to $\lambda_{is}^{j}(N_s^{j}(t)) < \lambda_{is}^{j}(N_s^{j}(0))$. When a violation is detected, we begin to adaptively determine the token rate, $J_{is}^{j}(t)$ such as $b_{i,s}(\lambda_{i,s}^{j}(N_s(t))) = J_{i,s}^{j}(t)$.

This dynamic mechanism may take place, by identifying all other connections belonging to the same traffic class s, and satisfying the following condition:

$$b_{is}(\lambda_{i,s}^{j}(N_s^{j}(t))) < b_{is}(\lambda_{is}^{j}(N_s^{j}(0)))$$

which is equivalent to $\lambda_{i,s}^{j}(N_s^{j}(t)) > \lambda_{is}^{j}(N_s^{j}(0))$.

For each connection i, requesting additional amount of resources, the DLB proceeds as explained by algorithm 2 discussed below.

Algorithm 2:. *Step 1:* Compute the following parameter:

$$\chi_{is}(t) = \mathrm{Min}_l\left\{\sum_{l\in \mathrm{Route}(i,s)} (\lambda_{k,s}^{l}(N_s^{l}(t)) - \lambda_{i,s}^{j}(N_s^{j}(t)))\right\}$$

which is the smallest policed *positive* leaky bucket that can be offered to the connection for which a violation has been detected at time t, where Route (i,s) designates the route of connection i which belongs to traffic class s.

Step 2: Repeat the following:

- Compute

$$\Delta\Psi_{i,k,s}^{j} = \min_{\substack{k,s \\ k\neq i}} \frac{\partial\Psi_{ks}^{j}\lambda_{is}^{j}(N_s^{j}(t))}{\partial\lambda_{is}^{j}(N_s^{j}(t))}\Delta\lambda_{is}^{j},$$

  where $\Delta\Psi_{i,k,s}^{j}$ designates the interference caused by connection i for the connection k, and $\Delta\lambda_{is}^{j} = \lambda_{is}^{j}(N_s^{j}(t)) - \lambda_{is}^{j}(N_s^{j}(0))$.
- If $\max_{k,j,s}(\Psi_{i,k,s}^{j} + \Delta\Psi_{i,k,s}^{j}) \ge \Psi_{\min}$ or $\max_{k,j,s}(\Psi_{i,k,s}^{j} + \Delta\Psi_{i,k,s}^{j}) < \Psi_{\min}$ and the total number of lost bits will not exceed its threshold for the most critical connection, then the request is accepted.
- If $\max_{k,j,s}(\Psi_{i,k,s}^{j} + \Delta\Psi_{i,k,s}^{j}) < \Psi_{\min}$ and the total number of lost bits exceeds its threshold for the most critical connection, then replace $\Delta\lambda_{is}^{j}$ by $\Delta\lambda_{is}^{j}/(\eta + 1)$, ($\eta$ is an arbitrary positive real number in]0,1[, fixed by the administrator).

  When reaching the exact value of $\lambda_{is}^{j}(N_s^{j}(t))$, the value of $J_{i,s}^{j}$ could be set equal $b_{i,s}(\lambda_{is}^{j}(N_s(t)) - \chi_{is}(t))$.

This adaptive leaky bucket policy can be used to manage dynamically both CBR and VBR traffic. For *ABR* flow control that has a mechanism for explicit allocating resources, we suggest to use this policy that can achieve a good match in terms of predicting the congestion state of the network.

Let us assume that an *ABR* source requests some amount of resources (bandwidth, and buffer) for a certain period of time P. We choose to decompose this period to D small interval of time $[0, P[= \bigcup_{l=0}^{D-1}[t_l, t_{l+1}[$ such that we can predict on each interval of time during which the amount of

resource is guaranteed with a certain accuracy. This dynamic policy enables us to maximise the resource utilisation.

To do this, we assume that a connection *js* that belongs to class *s* is transiting over *M* base stations. Each base station $m, 1 \le m \le M$, is assumed to hold $N_s^m(t)$ connections of class *s* at time *t*. Also, we assume that this connection formulates the need for additional resources for a certain period of time *P*. The leaky bucket algorithm at the access point (1st base station), will compute its stringency parameter $\lambda_{is}^j(N_s^j(t_k))$ at each time $t_k, 0 \le k \le D$, using algorithm 1 explained earlier. Given $\lambda_{is}^j(N_s^j(t_k))$ then using algorithm 2, we can readjust the cell rate during the period $[t_{k+1}, t_{k+2}[$.

6. Real time measurement and implementation of the dynamic leaky bucket

Cell Delay Variation Tolerance is a problem in statistical multiplexing. It greatly impacts bandwidth determination, and it is difficult to specify accurately. In addition CDV generated in the network may affect the amount of bandwidth needed. The existence of CDV makes it difficult to derive an accurate reference model, and so the control of the respect of the terms of the QoS contracts.

Now, we need to characterize the aggregated traffic of class *s* transiting over base station *m*, $1 \le m \le M$, without considering the connection *j*. Roughly speaking, we need to compute the probability that this aggregated traffic at node *m*, behaves as a constant rate fluid with rate *r* for a period of time *t*. Let $P_{\mathrm{ms}}(r,t)$ be this probability, then:

$$P_{m,s}(r,t) = \oint\limits_{\Sigma r_i = r} \mathrm{e}^{-t \sum\limits_{i,i\neq j} \breve{B}_{i,s}^*(r_i)} \, \mathrm{d}r_1 \ldots \mathrm{d}r_{n_{m,s}}.$$

One way to overcome these difficulties is to implement a DLB algorithm that uses the observation of cell streams and counts the number of cells arriving to the input buffers. Let x_h be this number counted during the *i*th measurement period. We have:

$$q(w,t) = \frac{\sum\limits_{h=1}^{L} 1(x_h = w, t)}{L}, \; w = 0, 1, 2, \ldots$$

During each *L* measurement period, $1(x_h = w, t) = 1$ if $x_h = w$, otherwise, it is 0.

Within the interval of time $[(t-1)L, tL]$, the DLB algorithm predicts the probability distribution $\{q(w,t)\}$ based on $\{q(w,t-1)\}$. Let this prediction be $\{\hat{q}(w,t)\}$. The prediction problem can be stated as follows. Given the random variable $q(w,t\text{-}1)$, predict

$$\hat{q}(w,t) = aq(w,t-1) + b\frac{\mathrm{d}q(w,t-1)}{\mathrm{d}t},$$

such that the mean squared value $E|\hat{q}(w,t) - q(w,t)|^2$ of the resulting error $\epsilon = \hat{q}(w,t) - q(w,t)$ is minimum. The solution is based on the result known as the orthogonal principle, which states that the mean squared error is minimum if the constants *a* and *b* are such that:

$$E\{[\hat{q}(t) - q(t)]q(t-1)\} = 0.$$

Using this principle, we can prove easily that:

$$a = \frac{R_q(1)}{R_q(0)}, \quad b = \frac{R_q'(1)}{R_q''(0)},$$

where $R_q(1) = E(q(w,t)q(w,t-1))$ and $R_q(0) = E((q(w,t))^2)$.

Now, we describe a real time approach that estimates the true value of $\lambda_{is}^j(N_s^j(t))$, which represents the true supported QoS at time *t*. A good estimation can be given using Kulback–Leibler formula [4], that enables us to reduce the distance between the exact value of $\lambda_{is}^j(N_s^j(t))$ and its estimation:

$$\varphi_{i,s}^j(t) = \log\left(1 + \frac{1 - \sum\limits_{w=0}^{m_{is}-1} \hat{\imath}(w,t)}{\sum\limits_{w \ge m_{is}} w\hat{\imath}(w,t) - m_{is}\sum\limits_{w=0}^{m_{is}-1} \hat{\imath}(w,t)}\right).$$

The method we propose is as follows. At time t_k, we have the set $\{\hat{I}(w,t_k)$ according to the aggregated values. We predict $\{\hat{I}(w,t), t_k \le t \le t_{k+1}$, using the following formula:

$$\hat{I}_{ms}(w,t) = \sum_r \hat{I}_{ms}\left(w - r\left(t - t_k - \frac{mT_s}{n}\right), t_k\right) P_{ms}(r,t).$$

Then we compute the estimation $\varphi_{i,s}^j(t)$ at time *t* and use algorithm 2 in order to readjust the cell rate during the period $[t_{k+1}, t_{k+2}[$.

7. Guaranteed QoS continuity

In this section, we develop a strategy for addressing the problem of guaranteed QoS continuity in ATM CDMA network. We define the function $N_s^j(t) = H(\vec{Q}_s^j(t))$, where $\vec{Q}_s(t) = (q_{is}^j(t))_{1 \le i \le N_s(t)}$, indicates the instantaneous guaranteed QoS vector for the connections handled by the *j*th base station, at time *t*, in terms of the number of mobiles $N_s^j(t)$. Our approach is based on the development of an instantaneous indicator that first predicts the algebraic value of additional number of mobiles $\Delta N_s^j(t)$, which will be handled by each base corresponding to each traffic class station. This is equal to the number of mobiles leaving minus the number of mobiles arriving to the base station. Then it establishes the relationships between this algebraic value and the instantaneous QoS of different connections.

We can now determine the new value $N_s^j(t) + \Delta N_s^j(t) = H(\vec{Q}_s^j(t) + \Delta\vec{Q}_s^j(t))$, when the number of mobiles at time *t*, change from $N_s^j(t)$ to $N_s^j(t) + \Delta N_s^j(t)$. Using the results we developed in Ref. [3], we can prove easily that:

$$\frac{\Delta N_s^j(t)}{N_s^j(t)} \le \Delta\vec{\Xi}_s^j(t)\vec{\Theta}(\vec{Q}_s^j(t)),$$

where

$$\Delta\vec{\Xi}_s^j(t) = \left(\frac{\Delta q_{is}^j(t)}{q_{is}^j(t)}\right)_{1\le i\le N_s^j(t)} \text{ is a row vector and}$$

$$\vec{\Theta}_s^j(\vec{Q}_s^j(t)) = \frac{1}{N_s^j(t)}\frac{\mathrm{d}H(\vec{Q}_s^j(t))}{\mathrm{d}(N_s^j(t))}$$

is a column vector and $\vec{\Theta}_s^j(\vec{Q}_s^j(t))$ is called the Relative Increase Ratio of Mobiles (RIRM).

7.1. Computing the Relative Increase Ratio of Mobiles (RIRM) metric

The QoS criterion is governed by the CLR, MCD, and CDV. All these parameters depend on the instantaneous stringency of the QoS requirement, $\lambda_{is}^j(N_s^j(t))$. Since

$$\vec{\Theta}_s^j(\vec{Q}_s^j(t)) = \frac{1}{N_s^j(t)}\frac{\mathrm{d}H(\vec{Q}_s^j(t))}{\mathrm{d}(N_s^j(t))},$$

we can easily deduce that:

$$\frac{\mathrm{d}(\vec{Q}_s^j(t))}{\mathrm{d}(N_s^j(t))} = \left[\frac{\mathrm{d}(q_{is}^j(t))}{\mathrm{d}(\lambda_{is}^j(t))}\frac{\mathrm{d}(\lambda_{is}^j(t))}{\mathrm{d}(N_s^j(t))}\right]_{1\le i\le N_s^j(t)}$$

where $q_{is}^j(t)$ describes the CLR, MCD, or CDV. The computation of

$$\Delta\vec{\Xi}_s^j(t) = \left(\frac{\Delta q_{is}^j(t)}{q_{is}^j(t)}\right)_{1\le i\le N_s^j(t)}$$

is easily given by:

$$\Delta\vec{\Xi}_s^j(t) = \left(\frac{\dfrac{\mathrm{d}q_{is}^j(t)}{\mathrm{d}\lambda_{is}^j(N_s^j(t))}}{q_{is}^j(t)}\Delta\lambda_{is}^j(N_s^j(t))\right)_{1\le i\le N_s^j(t)}$$

7.2. Computing $(\Delta N_s^j(t)/N_s^j(t))$

In order to guarantee the continuity of QoS, it seems very natural to predict periodically the number of handled mobiles given their present values and past behaviour. The problem can be stated as follows. Given the set of random variables $\{N_s^j(t_k)\}_{0\le k\le n}$, estimate $N_s^j(t_{n+r})$ such that each mean is given by;

$$\hat{N}_s^j(t_{n+r}) = \sum_{k=0}^{n} h(k)N_s^j(t_{n-k}),$$

where r is the prediction period. This predictor is the output of a linear time invariant causal system with input $\{N_s^j(t_k)\}_{0\le k\le n}$ and delta $\{h(k)\}_{0\le k\le n}$. To do this, we need to have the mean squared value $E\{|\hat{N}_s^j(t_{r+n}) - N_s^j(t_{r+n})|^2\}$ of the resulting error $\epsilon = \hat{N}_s^j(t_{r+n}) - N_s^j(t_{r+n})$ to be minimum. The solution is based on the result known as the orthogonal principle, which states that the mean squared error is minimum if the constant $\{h(k)\}_{0\le k\le n}$ are such that:

$$E\{[\hat{N}_s^j(t_{r+n}) - N_s^j(t_{r+n})]N_s^j(t_i)\} = 0, \qquad \forall 0 \le i \le n.$$

Using this principle, we can prove easily that $\{h(k)\}_{0\le k\le n}$ is a solution of:

$$R_{N_s^j}(m+r) = \sum_{k=0}^{n} h(k)R_{N_s^j}(m-k), \qquad \forall 0 \le m \le n.$$

where $R_{N_s^j}(m)$ is the autocorrelation function of the stochastic process $N_s^j(t)$. Given the set $\{h(k)\}_{0\le k\le n}$, we can deduce the expression of $(\Delta N_s^j(t)/N_s^j(t))$ as $((\hat{N}_s^j(t+r) - N_s^j(t))/(N_s^j(t)))$.With $(\Delta N_s^j(t))/(N_s^j(t))$ in hand, the network administrator can predict the value of the stringency $\hat{\lambda}_{is}^j(N_s^j(t+r))$ that designates the QoS contracted by the existing connections at time $t+r$. Then:

$$\begin{cases} \Delta N_s^j(t) \le \displaystyle\sum_{i=1}^{N_s^j(t)} \left\{\left[\frac{\mathrm{d}q_{is}^j(t)}{\mathrm{d}\lambda_{is}^j(N_s^j(t))}\right]^2 \frac{\mathrm{d}\lambda_{is}^j}{\mathrm{d}N_s^j(t)}\frac{\Delta\lambda_{is}^j}{q_{is}^j(t)}\right\}, \\ (\lambda_{is}^j)_{1\le i\le N_s^j(t)} \in \textit{QoS Contract} \end{cases}$$

8. Conclusion

To conclude, we defined dynamic monitoring schemes of traffic that provide information about the instantaneous bit rate of a source. This allows more effective flow control and can achieve a better match in terms of predicting the congestion state of the connection source. We also proposed an adaptive approach for the call admission and request for additional resource management, as well as a predictive scheme that enables us to maintain the continuity of the QoS guarantee. Our approach is based on the determination of a signal-to-interference rate and uses the large deviation theory. Both DLB, and guaranteed QoS continuity algorithms can be considered to achieve a good match in terms of predicting the performance indices to monitor and control connections dynamically. Our approach requires monitoring the parameter that helps determining the stringency of the QoS.

References

[1] A.W. Berger, Performance analysis of a rate-control throttle where tokens and jobs queue, IEEE Journal on Selected Areas on Communications 9 (1991) 165–170.
[2] A.W. Berger, W. Whitt, The impact of a job buffer in a token-bank rate-control throttle, Stochastic Models 8 (4) (1992) 685–717.
[3] C. Ben Ahmed, N. Boudriga, M.S. Obaidat, Predictive traffic engineering in CDMA ATM networks, in: Proceedings of the Symposium on Performance Evaluation of Computer Telecommunication Systems, SPECTS'99, Chicago, July 1999, p. 26–32.
[4] S. Shioda, H. Saito, Real time cell loss ratio estimation and its application to ATM traffic control, Proc. Infocom'97 12 (1997) 1401–1414.
[5] J.W. Roberts, Virtual spacing for flexible traffic control, International Journal of Communication Systems 7 (1994) 307–318.

[6] R. Gherin, H. Ahmadi, M. Naghshinech, Equivalent capacity and its application to bandwidth allocation in high speed networks, IEEE Journal on Selected Areas on Communications 9 (1991) 968–981.
[7] R.J. Gibbens, P.J. Hunt, Effective bandwidths for multi-type UAS channel, Queueing Systems 9 (1991) 17–28.
[8] F.P. Kelly, Effective bandwidths at multi-class queues, Queueing Systems 9 (1991) 5–16.
[9] D. Makrakis, K.M. Sundara Murthy, Spread slotted ALOHA techniques for mobile and personal satellite communication systems, IEEE Journal on Selected Areas on Communications 10 (6) (1992) 985–1002.
[10] A. Jamalipour, M. Katayama, T. Yamazato, A. Ogawa, Performance of integrated voice/data system in non uniform traffic low earth-orbit satellite communication systems, IEEE Journal on Selected Areas on Communications 13 (2) (1995) 465–473.
[11] A. Jamalipour, M. Katayama, T. Yamazato, A. Ogawa, Throughput analysis of spread-slotted ALOHA in LEO communication systems with non uniform traffic distribution, IEICE Transactions on Communications E78-B (12) (1995) 1657–1665.
[12] Z. Dziong, M. Jia, P. Mermelstein, Adaptive traffic admission for integrated services in CDMA wireless access networks, IEEE Journal on Selected Areas on Communications 14 (9) (1996) 1737–1757.
[13] V. Dijk, E. Aanen, H. Van Den Berg, Extrapolating ATM simulation results using extreme value theory, Queueing Performance and Control in ATM networks (ITC-13), Elsevier, Amsterdam, 1991 (p. 97–104).