Journal of Mobile Communication 1 (3): 85-96, 2007
ISSN: 1990-794X

Local- and System-Level QoS-Aware Traffic Conditioning in 3G Radio Access Networks

Frank Y. Li
Center of UniK-University Graduate, University of Oslo, N-2027 Kjeller, Norway

Abstract: Based on the framework of employing traffic shaping at the Mobile Stations (MSs) and policing at the Radio Network Controller (RNC), we propose a heuristic local and system-level Quality of Service (QoS)-aware token bucket parameter determination technique for traffic conditioning in Code Division Multiple Access (CDMA)-based radio access networks. The local QoS-awareness is achieved by bounding either shaping delay or out-of-profile probability at the MS and the system-level QoS-awareness is obtained by joint consideration of the local attributes and the system-level attribute, packet loss ratio for conformed packets, at the radio access system level. By tuning the system operating on the obtained system-level 'optimal' token rate and bucket size, the requirements for all concerned QoS attributes are assured. Thus a QoS-aware Service Level Agreement (SLA) can be reached between the end users and the network. Numerical results with respect to the system model considered are also presented in this study.

Key words: Traffic conditioning, local and system-level QoS-aware, token bucket, parameter determination, CDMA-based radio access networks

INTRODUCTION

In order to provide negotiated Quality of Service (QoS) in Code Division Multiple Access (CDMA)-based cellular systems, for example 3rd Generation (3G) networks, a traffic conditioner may be deployed optionally at the Mobile Station (MS) and obligatorily at the gateway nodes (3 GPP, 2004), regardless of whether Integrated Service (IntServ) (Braden *et al*., 1994), Differentiated Service (DiffServ) (Black *et al*., 1998) or a mixed IntServ/ DiffServ architecture is employed.

Traffic conditioning is performed by traffic shaping or/and policing. The traffic generated by the application is regulated by a traffic shaper, e.g., in the form of the Token Bucket (TB) algorithm. The MS sends out the specific QoS requirements of a service to the network, e.g., via Flow Specification (FlowSpec) (Black *et al*., 1998). The network and the user then negotiate, based on available network resources, to reach a Service Level Agreement (SLA). The FlowSpec specifies a regulated traffic flow in the form of a token bucket specification (i.e., token rate r and bucket size b) plus a peak rate p, a minimum policed unit m and a maximum packet size M (Skenker and Wroclawski, 1997) referred to as 5-tuples (r, b, p, m, M). After shaping, the regulated traffic flow is classified into two categories: conformed (i.e., in-profile) or non-conformed (i.e., out-of-profile) packets. The negotiated QoS is guaranteed only for conformed traffic.

Although the token bucket has been recognized as a recommended algorithm for traffic shaping, determining the TB parameters of a bursty traffic source is not a trivial task. We categorize the criteria appearing in the literature into two types, lossless (i.e., no out-of-profile packets allowed) criterion and loss/delay-bounded criterion (Li, 2002). The first criterion identifies a set of TB parameters for any given flow so that all arriving packets will be delivered immediately without incurring any delay or loss. In other words, with the appropriate token bucket parameters (r, b), all packets in the flow will be kept in-profile. This criterion does not really shape the traffic but identifies a pair of TB parameters to deterministic bounds so that there are no packets out-of-profile. It is thus more suitable for traffic characterization. The second criterion uses indeed the concept of traffic shaping and allows a small loss (out-of-profile) probability or a short shaping delay at the shaper. This criterion provides the flexibility of tradeoffs among various QoS parameters and is therefore more beneficial for resource reservation for a negotiated SLA.

In order to acquire appropriate TB parameters, one can either use a measurement-based technique which does not need any prior knowledge of the traffic pattern (Tang and Tai, 1999), or estimate the (r, b) quantitatively by a pre-defined rule based on known traffic characteristics [1, Annex C]. The former technique can capture the dynamic characteristics of a traffic flow, but the obtained TB parameters are comparatively large in order to keep any unexpected bursty traffic in-profile and moreover the parameters are usually time variable. For example, the TB parameters determined by the

measurement-based traffic specification may vary periodically every 30 sec (Tang, 1997). The latter technique is based on the prerequisite that the statistical characteristics of the traffic flow are known. It gives more static TB parameters, but the obtained parameters are very case-sensitive. Measurement-based techniques have also widely been used in admission control schemes in various kinds of communication systems (Jiang *et al*., 2004; Breslau and Jamin, 2000). Related work on TB parameter determination can be found in other literature as well. For example, (Alam *et al*., 2000) derived a relationship among r, b and p for Motion Picture Expert Group (MPEG) application for Guaranteed Service (GS) in the IntServ architecture, using a token bucket with leaky bucket rate control (Partridge) as the traffic shaper. Procissi *et al.* (2002) used a statistical model to analytically determine optimal token bucket parameters under various optimization criteria and applied the approach to several aggregated MPEG video sources. Through an example in GS (Glasmann *et al*., 2000) calculated TB parameters by using bounded end-to-end delay as the criterion, for several real-time video and audio applications. Another study (Garroppo *et al*., 2002) proposed a stochastic TB parameters estimation approach for aggregated traffic flows in a DiffServ architecture which may apply to fixed packet size applications for example Voice over Internet Protocol (VoIP). Tang and Tai (1999) derived mathematical expressions for measurement-based TB parameters determination, for different cases. Shan and Yang (1999) identified the bucket size of the dual-leaky-bucket traffic shaper with lossless criterion, for aggregated traffic. Lombardo *et al.* (2000) proposed using the tradeoff between loss and delay for TB parameters determination, where the TB functioned as an independent network entity. However in general, it is difficult to determine TB parameters analytically, for arbitrary traffic patterns.

From a system perspective, we propose a *heuristic* TB parameter determination procedure in this study, where the (r, b) parameters are obtained by searching the 'optimal' values, at the system level. We refer to this technique as QoS-aware in a sense that the QoS to be achieved is transparently known to the system administrator at the connection setup phase. The approach is based on the framework of imposing traffic shaping at the MS and traffic policing at the Radio Network Controller (RNC). The approach consists of two steps. With the first step, we obtain a set of local QoS-aware (r, b) pairs by bounding the shaping delay or the out-of-profile probability at the MS. The local QoS-aware (r, b) pairs constitute the candidates for system level performance. With the second step, we further consider the packet loss ratio for conformed packets as a factor in our optimality and achieve the system-level QoS-aware TB parameters by compromising the local and system-level QoS attributes. The obtained system-level QoS-aware (r, b) pair provides assured QoS (in terms of out-of-profile probability, traffic shaping delay and packet loss ratio) for the considered radio access networks. With all 5-tuples in a FlowSpec determined, a QoS-aware SLA can be formed between the mobile stations and the radio access network.

SYSTEM DESCRIPTION

The system model used in this study is based on a traffic conditioning-enabled radio access network. Related background information is also provided in this study.

Traffic conditioner and token bucket algorithm: According to Black *et al.* (1998) a traffic conditioner is an entity which performs traffic conditioning functions and it may contain meters, markers, droppers and shapers. A typical scenario for traffic conditioning is to apply traffic shaping and marking at the edge node(s) and traffic policing at the aggregate node(s).

For traffic shaping, the token bucket algorithm has been recognized as a *de facto* algorithm in the Internet community. There exist several variants of the TB algorithm in the literature. The 'standard' version of the TB algorithm can be found in [13 Annex B]. Our implementation of the TB algorithm, referred to as Simple Token Bucket (STB) (Glasmann *et al*., 2000) in the context, is shown in Fig. 1, where Token Bucket Counter (TBC) is an internal variable representing the number of remaining tokens at any time.

Different from the standard TB algorithm which performs only metering and marking to a traffic flow, the STB algorithm shapes the traffic when necessary. As a consequence of the STB, only packets with size larger than bucket size b will be judged as non-conformant. Other packets would be shaped to be conformed, if there are not enough tokens upon packet arrival. This would correspondingly introduce shaping delays, as illustrated in Fig. 1.

System model

A traffic conditioned radio access network: We are mainly interested in the radio access subnetwork of a communication system in this study. A traffic conditioning-enabled Radio Network Subsystem (RNS) which may consist of several Base Stations (BSs) with surrounding MSs is depicted in Fig. 2. Within this framework, the traffic conditioning is performed by traffic

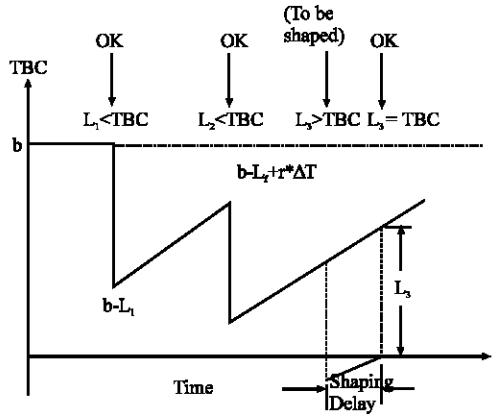

Fig. 1: Simple token bucket algorithm

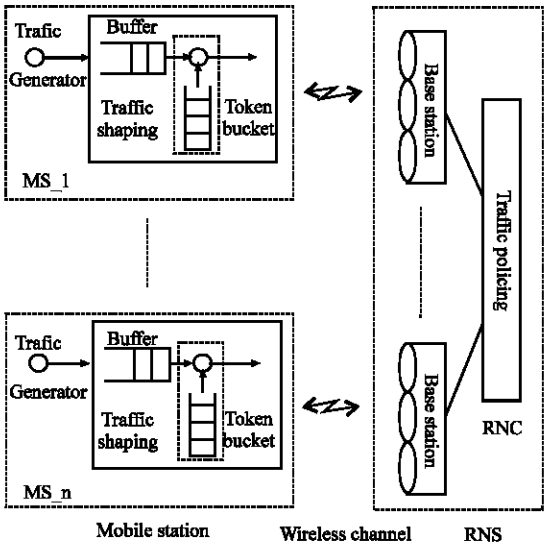

Fig. 2: System model: A traffic conditioning-enabled radio access network

shaping at each individual MS and traffic policing at the RNC, which may control several BSs. The traffic shaping scheme applies only to uplink traffic flows in this study.

As shown in the figure, the traffic generator at each MS is connected to a TB traffic shaper. The packets generated at the application layer have to pass through a First in First Out (FIFO) queue. The conformance of a packet is verified by the STB algorithm embedded at each MS. The shaped packets sent over the channel have either one of the two possible conformance status markings, i.e., conformed or non-conformed. A packet does not have to wait in the buffer if there are enough tokens available when it arrives. Only one packet is allowed to transmit from each MS at the same time. The regulated packets from each MS will be transmitted immediately, regardless of the conformance marking and channel status. The reason that a packet might be dropped after transmission is that each mobile station does not explicitly know the channel condition and wishes all its packets, no matter compliant or non-compliant, to be successfully received. Another alternative is to let some mobile stations drop their packets before transmission to make the channel less congested. However, how to make a fair decision among all mobile stations would be a challenging task for the RNC.

In our model, the traffic policing function is accomplished at the RNC, where packets received at the BSs are aggregated. The policing policy adopted here, which is not a recommended policing scheme described in (Shenker *et al*., 1997) but a specific means used in this study, is based on the channel congestion status and the system load.

Moreover, a fundamental assumption on packet generation in this model is that no segmentation happens. The packets are generated and forwarded to the traffic shaper as they are according to their length distributions. Furthermore, we do not make any pre-assumption on the size of a bucket. That is, the bucket size could be either larger or smaller than the maximum packet size M, depending on what the QoS-aware algorithm determines. Finally, at each MS, the total delay for each packet consists of two parts, the queueing delay and the shaping delay. In this study, we are solely interested in shaping delay, which is one of the concerned QoS attributes defined in Subsection II-C and the queueing time is thus neglected.

QoS attributes considered in the model: Delay and packet loss are two major QoS attributes in many communication systems. Throughout this paper, we use the following three self-defined QoS attributes which are delay and packet loss relevant as the major performance observations. They are defined as follows:

Shaping delay: Denoted by Ds, the average time period a packet spends in the buffer while waiting for tokens.

Out-of-profile probability: Denoted by πo, the probability that a packet is judged as non-conformed with respect to the specified TB traffic shaper (r, b). In the presence of the STB algorithm, πo is equal to the probability that a packet is larger than the bucket size, i.e., πo = Prob(Lj>b), given a large enough buffer.

Packet loss ratio: Denoted by P_{conf} for compliant packets and P_{non} for non-compliant packets, the ratio between the number of the discarded packets and the total number of the received packets at the RNC with the same

conformance status marking. For example, P_{conf} is the number of the conformed packets discarded divided by the total number of the conformed packets received at the RNC.

Furthermore in the context, the first two attributes defined above, Ds and πo, are regarded as local QoS attributes since they are associated locally at the MS. The third attribute, P_{conf}, is regarded as a system-level QoS attribute since it can only be obtained at the RNS system level. The values of these QoS attributes are obtained as follows.

Ds and πo. Given packet length and interarrival time distributions of a traffic flow, the local QoS attributes Ds and πo are decided by the values of (r, b) in the TBé algorithm. Generally speaking, it is not easy to find a closed form expression for Ds and πo for a given arrival process. However, when STB is applied, an explicit expression for πo may be obtained. In any case, an expression of the shaping delay Ds is not easily obtainable. Anyhow, in our simulation, this value is obtained by using a counter to log the sum of the individual shaping delay, divided by the total number of packets.

Figure 3 shows how the packet loss ratio for both compliant and non-compliant packets are calculated, by taking all concurrent connections into consideration and comparing with the channel capacity. When a packet is received, the RNC first classifies the packet by checking its conformance status in the header, as depicted in F ig. 3. No matter the arriving packet is compliant or not, all concurrent compliant and non-compliant packets are summed up in the load calculation at first. In the case where a non-compliant packet arrives, it is discarded if the channel is overloaded. Channel here means the wireless CDMA channel between MSs and one BS. However if the arriving packet is conformed, once congestion happens, another packet which is non-compliant will be preferentially discarded. A compliant packet may also be dropped only if all packets contributing to load calculation are compliant. Note the difference between these two policies. For non-compliant packets, the arriving packet itself will be discarded if congested. For compliant packets, another 'unlucky' non-compliant packet will be discarded preferentially instead, not the arriving packet itself.

Furthermore, although both P_{conf} and P_{non} are simulated in our study, only P_{conf} is considered as the third QoS attribute in our system-level scenario due to the fact that the contracted QoS is only guaranteed for compliant traffic.

The following system load calculation, known as Pole capacity (Heras *et al*., 2000) for a CDMA-based radio

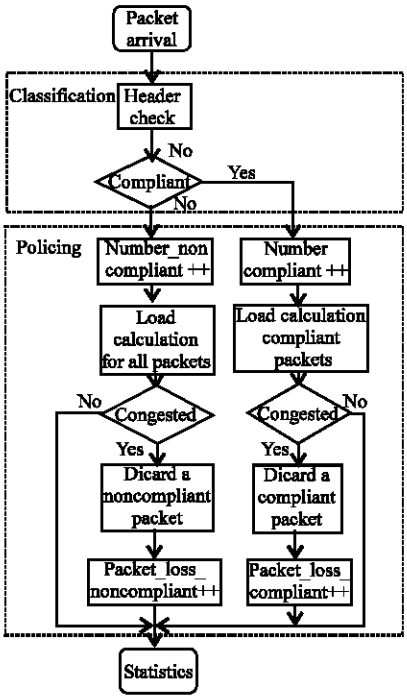

Fig. 3: Flow chart: Packet loss ratio calculation

network, is adopted for deciding whether the CDMA channel is 'congested' or not, when we calculate P_{conf} and P_{non} according to Fig. 3. If the total load calculated based on individual load given in Eq. 1 exceeds the pre-defined threshold, packet drop would happen.

Assuming perfect power control and negligible background thermal noise, the individual load of service i, denoted as ηi, can be calculated by

$$\eta i = \frac{(Eb/No)}{W/Ri} \times \nu i \times (1+f) \quad (1)$$

Where $(E_b/N_o)i$ is bit energy to noise ratio required for desired BER of service i, W is the chip rate, Ri is the information bitrate of service i, ii is the activity factor of service i, f is the interference factor from adjacent cells (i.e., ratio between adjacent cell interference and own cell interference).

The total load ηtotal in a cell is the sum of the individual loads with different services from all existing connections, i.e., $\eta \text{ total} = \sum_{i=1}^{Nu} ni$, where Nu denotes the total number of connections, including all traffic classes, currently in service.

The (Eb/No)i values used in this study are adopted from (Li and Stol, 2001). The channel is regarded as 'congested' if the total load ηtotal exceeds a pre-defined threshold, empirically ranging from 0.4-0.9 (Prasad *et al*., 2000). Value 0.9 is used in our simulation.

QoS-AWARE DETERMINATION OF THE TOKEN BUCKET PARAMETERS

As already mentioned earlier, deciding the TB parameters is very service-specific or application-dependent. On the one hand, large r and b could guarantee that all packets fall in-profile (Black *et al.*, 1998) but at a price of low network utilization and excessive resource requirement. On the other hand, small r and b may lead too many packets out-of-profile, thus violating the SLA of a service.

The objective of this study is to determine an optimal (r, b) pair so that the requirements for Ds, πo and P_{conf} are guaranteed, by tuning the TB shaper into operation at the obtained system-level QoS-aware TB parameters. By terminology QoS-aware in this context it is meant that the application is explicitly aware of the QoS to be achieved regarding the concerned QoS attributes before the connection is established, i.e., a QoS-aware SLA* has been reached. Furthermore, the local QoS-awareness is achieved by taking only local attributes into account at the MS and the system-level QoS- awareness is obtained by the tradeoff between the local and system-level attributes at the RNS system level. Throughout the context, the local and system-level QoS-aware

*An SLA covers of course much wider aspects than what the 5-tuples FlowSpec specifies, but an agreement between the end users and the network is mainly represented by FlowSpec in this study.

TB parameters are denoted by $(\bar{r},\bar{b})$ and (r*,b*), respectively.

The relationship between the local and system-level QoS-awareness is explained as follows. By local awareness, the MS knows explicitly how much shaping delay is introduced by the traffic shaper, or how much percent of the generated packets are classified as non-conformed. At this stage, the values for $(\bar{r},\bar{b})$ is not uniquely determined yet, since a set of $(\bar{r},\bar{b})$ pairs can provide this awareness. The MS does not know how much percent of these packets might be discarded, unless the system-level awareness is included. By system-level awareness, the RNS knows explicitly the packet loss ratio for each traffic class. The system-level QoS-aware TB pair (r*, b*) is selected based on the available set of $(\bar{r},\bar{b})$ pairs.

The proposed local and system-level QoS-aware TB parameters acquisition approach is presented in 2 steps in the following subsections:

Local QoS-aware TB parameters: Comparing the 2 techniques for TB parameter determination described in the study the advantage of the measurement-based technique is that no prior knowledge on traffic pattern is needed and the time-varying feature of the traffic is dynamically characterized. However, the disadvantages are also obvious. The values of the obtained TB parameters may be quite large especially for very bursty traffic and the fluctuation of the (r, b) values over time makes bandwidth reservation more difficult, as well as under-utilization of the resources. To overcome this problem, quantitative estimation technique uses the loss/delay-bounded criterion which allows a small loss (i.e., out-of-profile) probability or a small shaping delay at the TB. The prerequisite for utilizing this technique is that the traffic flow pattern must be known already. However, none of these 2 techniques provides the awareness of the QoS to be achieved for the end user.

The proposed local QoS-aware search procedure adopts somewhat similar approach as loss/delay-bounded criterion, which allows a certain amount of out-of-profile packets or a tolerable small shaping delay. The search process starts from a pre-defined token rate, which is referred to as reference token rate in the context. Usually the reference token rate is set slightly higher than the average bitrate of a traffic flow. By letting bucket size b be variable, we first figure out the out-of-profile probability πo and the shaping delay Ds against a range of r and b values and then based on these resulting πo and Ds values, we obtain a set of (r, b) pairs which give a bounded performance of the traffic shaper, by using either delay-bounded, out-of-profile-bounded, or other criteria. This set of (r, b) pairs is regarded as the local QoS-aware TB parameters $(\bar{r},\bar{b})$.

As different traffic classes have different QoS requirements (3GPP, 2004), we can employ separate criterion as the rule for deciding local QoS-aware (r, b) pairs. For example, a bounded maximum out-of-profile probability requirement may be suitable for loss-sensitive type of applications such as data transfer and a bounded maximum shaping delay criterion might be suitable for delay-sensitive applications like video streaming traffic.

When we observe the relationship between πo, Ds and (r, b), we find that the shaping delay Ds decreases monotonically as r becomes larger, or as b becomes smaller, i.e., Ds and when r % or b and; on the other hand, the out-of-profile probability πo has a reverse tendency as (r, b) changes, i.e., πo % when r % or b and[2] (Note that πo is independent of r only for STB}. Depending on the criterion applied, we formulate the solution to local QoS-aware TB parameter determination problem as follows.

Ds-bounded: With respect to maximum expected shaping delay D_s^{max}:

- Given the maximum acceptable shaping delay D_s^{max},
- Search for a set of TB parameters (r, b),
- So that for the set of $(\bar{r},\bar{b})$ parameters, we always have $D\_s(\bar{r},\bar{b}) = D_s^{max}$.

At this stage the exact values of πo $(\bar{r},\bar{b}$ are not determined yet locally, but fall into a restricted range by available $(\bar{r},\bar{b})$ pairs.

πo-bounded: With respect to maximum out-of-profile probability πo

- Given the maximum acceptable out-of-profile probability π_o^{max},
- Search for a set of TB parameters (r, b),
- So that} for the set of $(\bar{r},\bar{b})$ parameters, we always have $\pi o(\bar{r},\bar{b}) = \pi_o^{max}$.

At this stage the exact values of Ds $(\bar{r},\bar{b})$ are not determined yet locally, but fall into a restricted range by available $(\bar{r},\bar{b})$ pairs.

In summary, no matter which bound to apply, only 1 of the 2 local QoS attributes is set as the target at first. The other attribute can only be decided later by the tradeoff with the system-level attribute P_{conf}. Next, even though it is possible to bound both π_o^{max} and D_c^{max} simultaneously at the MS, the corresponding $(\bar{r},\bar{b})$ pair is not 'optimized' at the system-level, at this step. The (r*, b*) can only be obtained when P_{conf} is available.

Furthermore, in addition to the above two bounds, one may also apply other criteria based on different considerations, e.g., a static token rate r close to the average bitrate of a flow is more beneficial to Resource reSerVation Protocol (RSVP). If so, both Ds and πo are not strictly bounded. They may vary within an allowed range in exchange for a static r. We indeed adopt this idea as one of the alternative criteria for web browsing traffic in the next section.

System-level QoS-aware TB parameters: Based on the set of obtained local QoS-aware TB pairs $(\bar{r},\bar{b})$, we further develop the idea into a system-level QoS-aware TB parameter acquisition procedure. The already obtained local QoS-aware $(\bar{r},\bar{b})$ pairs are regarded as the candidates for system-level QoS-aware parameters (r*, b*). The searching process for (r*, b*) is based on the assumption that the MSs are able to justify their $(\bar{r},\bar{b})$ values according to the downlink feedback from the RNC.

Let the traffic shaper operate on the local QoS-aware $(\bar{r},\bar{b})$ parameters. We continue to investigate the system performance by the third QoS attribute, i.e., packet loss ratio at the RNC. Now we face an RNS with multiple users from various classes. Even though heterogeneous traffic classes are allowed to co-exist at the same time, we concentrate only on one specific traffic subclass each time when we talk about system-level awareness. In other words, we focus on a homogeneous traffic subclass and other classes in the system are treated as background traffic each time when designing the (r*, b*) pair for the corresponding traffic class. Furthermore, we assume that all MSs in this subclass have identical traffic pattern so that the available local QoS-aware pairs $(\bar{r},\bar{b})$ are identical for all those MSs. That is, $\bar{r}_i,\bar{b}_i = (\bar{r},\bar{b})$ for all i in N_u^k, where N_u^k is the number of users from the concerned traffic subclass k with $1 < N_u^k < Nu$ and Nu is the total number of users in the system.

To obtain the system-level QoS-aware TB parameters for heterogeneous traffic, we further classify the flows into various subclasses where each subclass has identical pattern. The following formulation on system-level QoS-awareness is mainly targeted to each homogeneous traffic subclass. However, it applies to heterogeneous traffic as well.

As already mentioned, the compliant packets could also be dropped if there are too many concurrent compliant packets on the channel. When the observation of P_{conf} values are available, we have an overall control of all three QoS attributes. Again, P_{conf} exhibits an upward or downward tendency against (r, b) value variations which is different from that of πo or Ds's. For example, as b becomes larger, Pconf increases monotonically while πo decreases monotonically. On the other hand, when r increases, Ds decreases but Pconf increases due to more concurrent compliant packets on the channel. The tradeoff among them provides us with the system-level 'optimized' TB parameters which guarantee the desired πo, Ds and Pconf.

We formulate the system-level QoS-aware solution as:

Ds-bounded:

$$\text{minimize } d = |\, Pconf(\bar{r},\bar{b}) - \pi o(\bar{r},\bar{b}) \,| \quad \text{subject to } Ds(\bar{r},\bar{b}) = D_s^{max} \qquad (2)$$

Where d is the absolute value of the numerical difference between Pconf $(\bar{r},\bar{b})$ and πo $(\bar{r},\bar{b})$. Formula (2) means that the system-level QoS-aware pair (r*,b*) is a

selected $(\bar{r},\bar{b})$ pair which gives the minimal difference between Pconf $(\bar{r},\bar{b})$ and πo $(\bar{r},\bar{b})$ while guaranteeing an earlier-designed shaping delay D_s^{max}.

πo -bounded:

minimize Ds $(\bar{r},\bar{b})$
subject to

$$\pi o\,(\bar{r},\bar{b}) \le \pi_o^{max} \text{ and } Pconf\,(\bar{r},\bar{b}) \le P_{conf}^{opt} \qquad (3)$$

There are 2 aspects of Formula (3). On the one hand, due to the nature of STB, πo $(\bar{r},\bar{b})$ is independent of r. So b* is selected as the minimal $\bar{b}$ among all $\bar{bs}$ which guarantee $\pi o(\bar{r},\bar{b}) \le \pi_o^{max}$. On the other hand, given a targeted P_{conf}^{opt}, we must have $\bar{b}$ smaller than or equal to certain value for guaranteeing P_{conf} $(\bar{r},\bar{b}) \le P_{conf}^{opt}$. The maximal possible value of $\bar{b}$ which guarantee P_{conf} $(\bar{r},\bar{b}) \le P_{conf}^{opt}$ is then selected as r*. In the meantime, this r* ensures a minimized D_s $(\bar{r},\bar{b}) = D_s^{opt}$ $(\bar{r},\bar{b})$ since D_s and while r %.

In summary, the system-level QoS-aware solution can be summarized as follows:

- Given a set of local QoS-aware parameters $(\bar{r},\bar{b})$ as the references for system level performance evaluation,
- Determine a unique system-level QoS-aware pair (r*, b*),
- So that for this (r*, b*) pair, we always have Ds $(r^*,b^*) \le D_s^{max}$ and $\pi o\,(r^*,b^*) = \pi_0^{opt}$, $Pconf(r^*,b^*) = P_{conf}^{opt}$ for Ds-bounded criterion, or πo $(r^*,b^*) = \pi_0^{max}$, $Pconf(r^*, b^*) \le P_{conf}^{opt}$ and Ds $(r^*, b^*) = D_s^{opt}$ for πo-bounded criterion, where π_0^{opt}, D_s^{opt} and P_{conf}^{opt} are obtained or targeted 'optimal' values for πo, D_s and P_{conf} by Formulae (2) and (3).

This (r*, b*) pair is referred to as the system-level QoS-aware TB parameters. Note here (r*, b*) are finally obtained parameters for TB traffic shaper at the MS, for the corresponding traffic subclass. To obtain the (r*, b*) pair for another traffic subclass, one needs to run the same procedure once again specifically for that subclass.

RESULTS

To assess the performance of the proposed technique, a simplified simulation model regarding a traffic conditioning-enabled RNS is implemented using the OPNET network simulator, with only one base station in the system. Three traffic classes, namely voice, steaming video and web browsing, are considered in our simulation. The total number of MSs distributed in the system is 70, with 50 voice users, 10 streaming video users and 10 web browsing users, respectively. Only streaming video and web browsing flows are regarded as traffic conditioning-applicable in the model. The voice traffic has a constant bitrate (12.2 Kbps) and is injected into the network without shaping as 'background' traffic, at a moderate traffic load. The moderate load is meant that the channel will never be congested by pure voice traffic. The streaming video and web browsing packets have to pass through the traffic shapers before they are sent out over the CDMA channel. A confidence level of 95% is targeted for all numerical results illustrated.

Traffic Models: As the two-state, On-Off traffic model has been widely used for bursty traffic source (Schwartz, 1996), we apply this model as well for all three traffic classes in this study. The packet is transmitted at peak rate Rp during On period and the average rate Ra is obtained by

$$R_a = \frac{T_{on}}{T_{on} + T_{off}} \bullet R_P = \frac{T_{on}}{\Delta T} \bullet R_P$$

Where $\Delta T = T_{on} + T_{off}$ is the interarrival time (Gu *et al.*, 1995). The two traffic shaping applicable classes with their characteristics are tabulated in Table 1. The reference token rate is calculated by

$$R_a + \frac{L_h}{\delta_y}$$

Where L_h is the length of the protocol header and ΔT is the sampling interval of the source streams (the inverse of ΔT is the number of frames per second). Regarding to the traffic models shown in Table 1, a 40 bytes Real-time Transport Protocol/User Datagram Protocol/IP (RTP/UDP/IP) header is assumed for streaming video traffic and a 4 bytes compressed Transmission Control Protocol/IP (TCP/IP) header is assumed for web browsing traffic.

Pareto distribution with cut-off is well suited to describe packet length distribution of web browsing traffic according to ETSI (1998). The frame length of some streaming video streams, for example, a Joint Photographic Experts Group (JPEG) flow, may also obey Pareto distribution (Stallings, 1998). We therefore apply Pareto distribution to both considered traffic classes. The distinction between these 2 classes in our model is that

Table 1: Traffic models used in simulation

Traffic class	Streaming video	Web browsing
Average bitrate	28 Kbps	60.8 Kbps
Peak rate	40 Kbps	144 Kbps
Interarrival time	Constant	Exponential
Packet length	Pareto with cut-off (1.7, 1864 bits, 12000 bits)	Pareto with cut-off (1.1, 652 bits, 12000 bits)
Reference token rate	32.76 Kbps	62.4 Kbps

the streaming video traffic has a constant interarrival time while the interarrival time for web browsing traffic is exponentially distributed. Furthermore, similar to the Maximum Transmission Unit (MTU) of 1500 bytes in Internet, or the maximum Service Data Unit (SDU) size of 1502 octets in Universal Mobile Telecommunication System (UMTS), we set the packet length cut-off for both classes in our traffic model as 1500 octets (12000 bits).

Note from Table 1, we have defined up to now the last three elements (p, m, M) of the 5-tuples (r, b, p, m, M) of a FlowSpec which will be used for SLA establishment. We are going to decide the first 2 elements r and b in the following subsections, with QoS-aware values.

Local QoS-aware TB pairs ($\bar{r}, \bar{b}$): Based on the packet size assumption in Subsection IV-A, the out-of-profile probability π_o can be calculated explicitly according to STB. Given Pareto (α, k) distribution with cut-off C, the probability that a packet with length L>b, which is equivalent to π_o in our case, is decided by

$$\pi_o = \Pr(L > B) = \int_b^{\infty} \frac{\alpha \cdot k^{\alpha}}{x^{\alpha+1}} dx = k / b^{\alpha} \quad (4)$$

Where α is the shape factor with α>1 and k is the minimum packet size.

The theoretical curves for π_o by Eq. 4 are plotted in Fig. 4. The simulated results, even though not plotted in the figure, are identical to the analytical results, for both web browsing and streaming video streams. The figure shows that the larger the bucket size b, the smaller the π_o. Note here that the result that π_o is independent of r is not generally true. When other traffic shaping scheme, e.g., marking a packet as non-conformed if Lj > TBC upon arrival, as described in (3 GPP, 2004), is employed, π_o is a function of both r and b.

The out-of-profile probability π_o is one of our main observations for QoS-aware TB parameter determination. Together with the shaping delay Ds, we are going to determine the local QoS-aware (r, b) pairs for web browsing and streaming video using different criteria in this subsection. Before proceeding, we have to state that

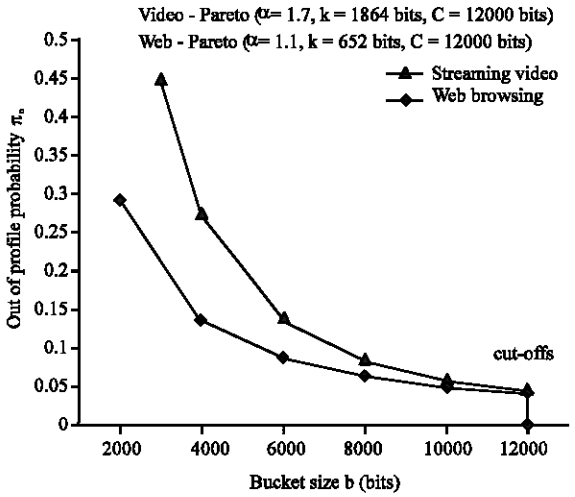

Fig. 4: Out-of-profile probability π_o for Pareto distributed packets using STB

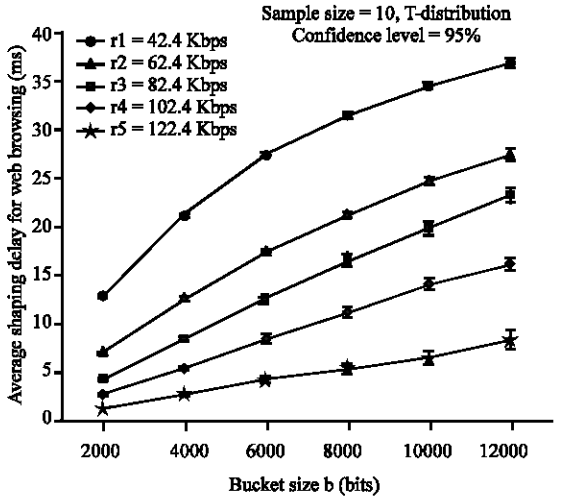

Fig. 5: Determining local QoS-aware ($\bar{r}, \bar{b}$) pairs for web browsing

Formula (2) is used only for searching streaming video (r*, b*) in this study. For web browsing traffic, in addition to Formula (3), we choose yet another alternative in order to show that it is also sensible to apply other criterion when deciding ($\bar{r}, \bar{b}$).

Local QoS-aware TB Pairs ($\bar{r}, \bar{b}$) for Web Browsing: Figure 5 depicts the simulation results of the average shaping delay Ds as a function of b, for several token rates. The average shaping delay depicted in the context is calculated as the total amount of time all packets in a flow spend in the buffer while waiting for tokens, divided by the number of packets which have to wait. The Ds will

be much smaller if we use the total number of packets generated as the denominator. Here the concrete milliseconds values for Ds do not imply any correspondence to the standard delay values defined for UMTS in (3 GPP, 2004).

Now we can observe the results from two perspectives. First, for a given token rate r, the shaping delay Ds increases as b increases. Second, for a given bucket size (which corresponds to a deterministic πo as shown in Fig. 4, the packets suffer longer shaping delay as r decreases.

There are two methods for designing local web browsing $(\bar{r},\bar{b})$. As web browsing traffic has more elastic delay tolerance and very stringent loss requirement, a natural option is to use the π_0^{max}-bounded criterion described in Subsection III-A. For example, giving the requirement as π_0^{max} = 5% corresponds to $\bar{b}$ ~ 10000 bits. As πo is independent of r in this example, all r values, from r_1 = 42.4 Kbps to r_5 = 122.4 Kbps give the same π_0^{max} for given $\bar{b}$. The local $(\bar{r},\bar{b})$ could thus be ($\bar{r}$ = 42.4 Kbps, $\bar{b}$ = 10000 bits), ($\bar{r}$ = 62.4 Kbps, $\bar{b}$ = 10000 bits), ... and ($\bar{r}$ = 122.4 Kbps, $\bar{b}$ = 10000 bits.

As already mentioned, we would also like to introduce another method here by imposing a static token rate $\bar{r}$ for web browsing. Directly, we take the reference token rate 62.4 Kbps as the local QoS-aware token rate $\bar{r}$ in this study. However, this straightforward solution is achieved by giving up the strictly targeted π_0^{max} requirement. Instead, a restricted range of πo is targeted in this case.

With this approach, both Ds and πo are allowed to be variable within a certain range which corresponds to a range of controllable Ds and πo values. More specifically, the set of local QoS-aware $(\bar{r},\bar{b})$ pairs in the studied example is given as ($\bar{r}$ = 62.4 Kbps, $\bar{b}$ = 3000 bits), ($\bar{r}$ = 62.4 Kbps, $\bar{b}$ = 4000 bits), ..., ($\bar{r}$ = 62.4 Kbps, $\bar{b}$ = 8000 bits), which corresponds to a range of restricted Ds ≈ 10 ~ 22 ms and πo ≈ 20 ~ 6%.

Local QoS-aware TB pairs $(\bar{r},\bar{b})$ for streaming video: Now we apply the bounded maximum shaping delay D_s^{max} criterion for local streaming video $(\bar{r},\bar{b})$ determination. The approach selects a set of $(\bar{r},\bar{b})$ which provides a bounded maximum shaping delay, i.e., Ds $(\bar{r},\bar{b})$ = D_s^{max}.

The simulation results are depicted in Fig. 6, with regard to three Ds requirements Ds = 10, 15, 20 ms, respectively. We plot the curves with b as X-axis and r as Y-axis. The curves decide a set of $(\bar{r},\bar{b})$ pairs which has a guaranteed shaping delay of 10, 15 or 20 ms, respectively. The guaranteed shaping delay here is meant that with the corresponding $(\bar{r},\bar{b})$ values, Ds will never exceed targeted D_s^{max}. For example, the area with dashed lines corresponds to a set of (r, b) values which guarantees a shaping delay of Ds = 15 ms. As the candidates for system-level QoS-aware TB parameters determination, the resulting local $(\bar{r},\bar{b})$ pairs with respect to Ds = 15 ms are tabulated in the first two columns of Table 2. It means that all the possible $(\bar{r},\bar{b})$ pairs, ($\bar{r}$ = 29.76 Kbps, $\bar{b}$ = 3550 bits), ($\bar{r}$ = 30.76 Kbps, $\bar{b}$ = 4000 bits), ..., ($\bar{r}$ = 34.76 Kbps, $\bar{b}$ = 6600 bits) provide a guaranteed Ds of 15 ms.

The curves shown in Fig. 6 are obtained under the assumption that an MS is able to adjust its (r, b) value in an allowable range in order to achieve a guaranteed shaping delay. The curves are a collection of (r, b) pairs which guarantee a pre-defined Ds value and are obtained through extensive simulations conducted as follows. To obtain one point in each curve in the figure, we first set the token rate b to a given value and then carry out the same simulation several times with different r values until we find one r which meets the corresponding Ds requirement. The corresponding (r, b) pair is one element of the local QoS-aware $(\bar{r},\bar{b})$ set.

System-level QoS-aware TB pairs (r*,b*): Using the local TB $(\bar{r},\bar{b})$ pairs obtained above, we are ready to decide the system-level QoS-aware TB pair (r*, b*), by joint consideration of the out-of-profile probability πo, the shaping delay Ds and the packet loss ratio for compliant packets P_{conf}. The packet loss ratio is simulated according to the flow chart shown in Fig. 3. The Pnon is only shown as a reference.

System-level QoS-aware TB pair (r*,b*) for web browsing: As stated earlier, the out-of-profile probability πo decreases monotonically as b increases, while the packet loss ratio for compliant packets Pconf increases monotonically. This is because with larger b, there are more conformed packets on the channel, thus leading to a higher Pconf.

The simulated results for πo, P_{conf} and P_{non} are plotted together in Fig. 7. One may notice that P_{non} keeps comparatively constant as b varies. This is because all packets, no matter conformed or not, are taken into account for channel load calculation. Within the same simulation period, the total numbers of packets generated by the same amount of sources are nearly the same, regardless of how many of them are judged as non-conformed.

Continuing our study in Subsection IV-B, there are also two alternatives for obtaining system-level (r*, b*) of web browsing traffic. Let us continue with the static token

Table 2: Local/System-Level QoS-aware (r, b) determination for streaming video flow

r(Kbps)	b(bits)	po (%)	Pconf (%)	Pnon (%)	Ds (ms)
29.76	3550	33.4±0.0009	4.2±0.0064	49.0±0.0083	15±0.0068
30.76	4000	27.3±0.0007	6.5±0.0073	50.8±0.0079	15±0.0045
31.76	4500	22.4±0.0006	9.1±0.0085	52.4±0.0069	15±0.0037
32.76	5600	15.4±0.0005	13.9±0.0057	53.8±0.0064	15±0.0063
33.76	6200	12.9±0.0005	16.1±0.0073	54.3±0.0053	15±0.0059
34.76	6600	11.6±0.0003	17.5±0.0053	54.5±0.0046	15±0.0075

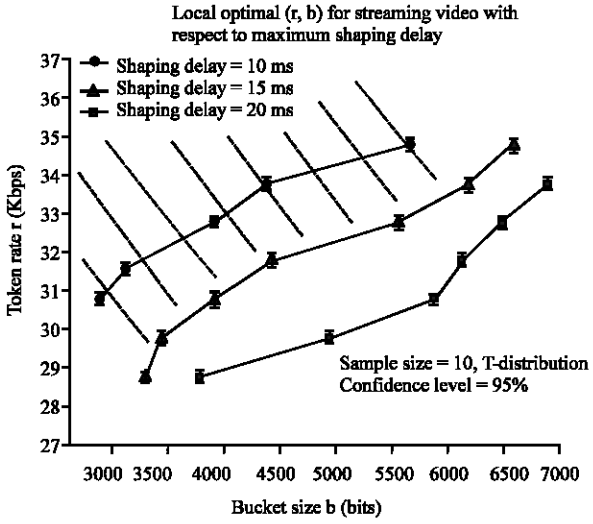

Fig. 6: Local QoS-aware $(\bar{r},\bar{b})$ pairs for streaming video with respect to a fixed shaping delay

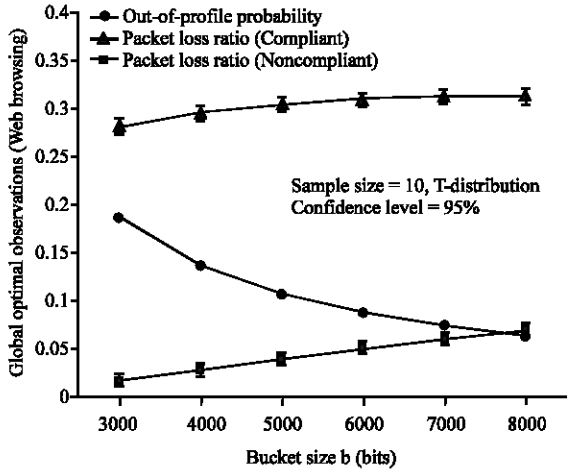

Fig. 7: System-Level QoS-aware pair (r*, b*) for web browsing with tradeoff between πo and Pconf

rate idea first. With this option, we adopt $\bar{r}$ directly as r* and then use Formula (2) for deciding for b*. As shown in Fig. 7, we find that this system-level 'optimal' bucket size b* lies at the intersection between πo and Pconf curves, at b = 7660 bits. This is a $\bar{b}$ which gives a minimized value of

$$\left|P_{conf}^{opt}-\pi_o^{opt}\right|\sim 0$$

Going back to Fig. 5, we find that the corresponding shaping delay is about Ds = 20.12 ms. With this alternative, the system-level QoS-aware pair is (r* = 62.4 Kbps, b* = 7660 bits), which gives an ensured performance as Ds = 20.12±0.0035 ms, Pconf = 6.69±0.0055 and πo = 6.69±0.0062 %.

Another alternative, which is probably more sensible for web browsing, is using Formula (3) instead. Recall in Subsection IV-B π_o^{max} the is set as 5%. We directly obtain b* = P_{conf}^{opt} = 10000 bits. By setting the targeted P_{conf}^{opt} also as 5%, we must have $\bar{r}$ ≤72.6 Kbps. The system-level r* is therefore selected as r* = 72.6 Kbps, which gives a minimal shaping delay as D_s^{max} = 23.5 ms. With this system-level QoS-aware pair (r* = 72.6 Kbps, b* = 10000 bits), we have Ds = 23.5 ms, Pconf = 5% and πo = 5 %. Comparing these two results, it basically reflects the fact that to achieve a smaller packet loss, more delay must be suffered.

System-level QoS-aware TB Pair (r*, b*) for streaming video: The searching process for (r*, b*) of the streaming video flow is carried out using Formula (2). Even though all possible $(\bar{r},\bar{b})$ values within the first two columns of Table 2 meet the shaping delay requirement of Ds = 15 ms, only one pair can be regarded as the system-level QoS-aware pair (r*,b*). The system-level performance of the studied system for streaming video flow are simulated with results tabulated in columns 3-6 of the same table.

The ultimate system-level QoS-aware pair (r*, b*), which minimizes the difference between πo and Pconf, lies somewhere within the highlighted area in the table. By running the simulation exhaustively with finer (r, b) granularity within the highlighted area, we reach finally the system-level QoS-aware (r*, b*) pair as (r* = 32.9 Kbps, b* = 5780 bits), which ensures Ds = 15.00±0.0068 ms, Pconf = 14.59±0.0073% and πo = 14.51±0.0052 % for the studied streaming video flow.

DISCUSSION

The numerical results for designing (r*, b*) presented in this section are conducted on a homogeneous traffic class each time. However, as we mentioned earlier, the proposed technique also applies to heterogeneous traffic by subdividing the flows into multiple homogeneous subclasses. In this case, the threshold for traffic load and congestion calculation of each subclass needs be adjusted accordingly.

To implement our heuristic technique in a real system, the MSs must have embedded software package to justify their TB parameters according to the feedback from the RNC which is also updated with the corresponding software. The 'handshake' process is done at the connection setup phase. It will of course introduce new 'burden' for signalling function of the system. However, how much extra overhead is needed is beyond the major interest of this study.

Finally, it is worth mentioning that, similar to our idea of iteratively selecting optimal TB parameters, a token bucket parameter renegotiation scheme has been proposed in a recent study (Song and Lee, 2004). By applying the scheme to video traffic, the authors demonstrated that better QoS performance has been achieved by using negotiated TB parameters between the traffic source and the network.

CONCLUSION

Based on the framework of applying traffic shaping at the MS and traffic policing at the RNC for QoS provisioning in radio access networks, we have presented a heuristic local and system-level QoS-aware TB parameter searching technique in this paper. The approach provides us with a local and system-level awareness of the QoS to be achieved for reaching an SLA between the mobile users and the radio access network. A major contribution of this work is a general scheme for deciding optimal TB parameters that may be applied to various kinds of applications, in contrast with many other solutions which apply usually only to a specific type of application. However, the QoS-awareness in our approach is achieved as a result of exhaustive simulations. For a large-scale system with multiple network nodes and much more QoS attributes, the searching process could be a tedious task. Finally, even though the radio access system-level QoS awareness does not provide end-to-end QoS awareness directly, it has its significance since it constitutes part of the end-to-end QoS provisioning.

ACKNOWLEDGMENT

This research is supported by the Research Council of Norway (NFR) under the FUCS (Future Communication Systems) program.

REFERENCES

3GPP TS, 2004. 23.107v6.1.0, QoS Concept and Architecture, http://www.3gpp.org.

3GPP, 2004. TS 23.207v6.4.0, End-to-End QoS Concept and Architecture, http://www.3gpp.org.

Alam, M.F., M. Atiquzzaman and M.A. Karin, 2000. Traffic Shaping for MPEG Video Transmission over the Next Generation Internet, Computer Communications, Elsevier., 23: 1336-1348.

Blake, S., D. Black, M. Carlson, E. Davies, Z. Wang and W. Weiss, 1998. An Architecture for Differentiated Services, RFC 2475, IETF.

Braden, R., D. Clark and S. Shenker, 1994. Integrated Services in the Internet Architecture: an Overview, RFC 1633, IETF.

Breslau, L. and S. Jamin, 2000. Comments on the Performance of Measurement-Based Admission Control Algorithms, in Proc. IEEE INFOCOM, Tel-Aviv Israel, pp: 1233-1242.

ETSI SMG, 1998. Selection Procedures for the Choice of Radio Transmission Technologies of the UMTS, http://www.etsi.fr, UMTS 30.03v3.2.0.

Garroppo, R.G., S. Giordano and M. Pagano, 2002. Estimation of Token Bucket Parameters for Aggregated VoIP Sources. Int. J. Commu. Sys., WILEY Int. Sci., 15: 851-866.

Glasmann, J., M. Czermin and A. Reidl, 2000. Estimation of Token Bucket Parameters for Videoconferencing System in Corporate Networks, in Proc. Soft COM, Split, Croatia.

Gu, X., K. Sohraby and D.R. Vaman, 1995. Control and Performance in Packet, Circuit and ATM Networks, Kluwer Academic Publishers.

Heras-Brandin, A., P. Bartolome-Pascual, D. Gomez-Mateo and J. Izquierdo-Arce, 2000. A Multiservice Dimensioning Procedure for 3G CDMA, In: Proc. IEE Int. Conf. 3G Mobile Commun. Tech. London, UK, pp: 406-410 .

Jiang, Y., P. Emstad, V.F. Nicola and A. Nevin, 2004. Measurement-Based Admission Control: A Revisit, In: Proc. Seventeenth Nordic Teletraffic Seminar, Fornebu, Norway, pp: 99-112.

Li, F.Y. and N. Stol, 2001. A Priority-oriented Call Admission Control Paradigm with QoS Re-negotiation for Multimedia Services in UMTS, in Proc. IEEE VTC, Rhodes, Greece, pp: 2021-2025.

Li, F.Y., 2002. Local and Global QoS-aware Token Bucket Parameters Determination for Traffic Conditioning in 3rd Generation Wireless Networks, In: Proc. Eur. Wireless (EW) Florence, Italy, pp: 362-368.

Lombardo, A., G. Morabito and G. Schembra, Traffic Specification for MPEG Video Transmission over the Internet, In: Proc. IEEE ICC, New Orleans, USA, pp: 853-857.

Partridge, C., 1994. Gigabit Networking, Addison-Wesley, (Token Bucket with Leaky Bucket Rate Control), pp: 262-263.

Prasad, R., W. Mohr and W. Konhauser, 2000. Third Generation Mobile Communication System, Artech House.

Procissi, G., A. Garg, M. Gerla and M.Y. Sanadidi, 2002. Token Bucket Characterization of Long-range Dependent Traffic, Computer Commun. Elsevier., 25: 1009-1017.

Schwartz, M., 1996. Broadband Integrated Networks, Prentice-Hall, Inc.

Shan, T. and O.W.W. Yang, 1999. Improving Resource Utilization for the Rate-Controlled Traffic Flows in High Speed Networks, in Proc. IEEE ICC, Vancouver, BC, Canada, pp: 864-868.

Shenker, S. and C. Patridge and R. Guerin, 1997. Specification of Guaranteed Quality of Service, RFC 2212, IETF.

Shenker, S. and J. Wroclawski, 1997. General Characterization Parameters for Integrated Service Network Elements, RFC 2215, IETF.

Song, H. and D.-B. Lee, 2004. Effective Quality-of-Service Renegotiating Schemes for Streaming Video, EURASIP J. Applied Signal Proce., 2: 280-289.

Stallings, W., 1998. High-Speed Networks: TCP/IP and ATM Design Principles, Prentice-Hall, Inc.

Tang, P., 1997. The Interface Specification for Measurement-Based Traffic Specifier (MBTS), Int. Corporation, unpublished.

Tang, P.P. and T-Y. C. Tai, 1999. Network Traffic Characterization Using Token Bucket Model, In: Proc. IEEE INFOCOM, New York, USA, pp: 51-62.