Policing Aggregates of Voice Traffic with the Token Bucket Algorithm

M.J.C. Büchli, D. De Vleeschauwer, J. Janssen, G.H. Petit
Alcatel Bell, Network Strategy Group
Francis Wellesplein 1
B-2018 Antwerpen, Belgium

***Abstract*- Estimating correct traffic descriptors is an important issue when deploying IP telephony with quality of service guarantees. Usually, a token bucket is used to police a traffic flow. Such a policer checks (and ensures) that the traffic flow fits inside a traffic envelope. For variable bit rate (VBR) sources normally two traffic envelopes are specified: one dealing with the peak rate and one dealing with the sustainable rate with an associated burst size. These parameters have to be specified when making a reservation with either the resource ReSerVation Protocol (RSVP) in case of IntServ or may be specified in a Service Level Specification (SLS) when using DiffServ. This paper shows how to choose the traffic descriptors for aggregate voice traffic for both constant bit rate sources and sources that use voice activity detection. Also the influence of variable packet size on the Packet Discard Ratio (PDR) is shown. A traditional policer discards large packets more frequent than smaller ones. To alleviate this, a modification to the token bucket algorithm is proposed in this paper.**

I. INTRODUCTION

A. Background

On the current public Internet only a best-effort transport service is offered. Hence, the performance of the network depends on the instantaneous load and no guarantees can be given. When real-time services are being deployed on IP networks there is a clear need for Quality of Service (QoS), i.e., guarantees on the performance of the network. Telephony is an interactive real-time application that puts stringent requirements on the network. Delay and packet loss have a (strong) impact on the voice quality, depending on the type of codec used and on the fact whether a Packet Loss Concealment (PLC) is implemented or not [8]. Introducing QoS in an IP network implies that extra functionality has to be added in order to be able to guarantee a bound on the delay and the packet loss. This functionality includes policers, shapers, markers, flow admission control etc. The IETF has introduced two architectures for QoS, the Integrated Services (Intserv) model [4] and the Differentiated Service (DiffServ) model [2]. In the IntServ model resources are reserved per flow by, for example, the resource ReSerVation Protocol (RSVP) [5]. The traffic specification is carried in the RSVP PATH message and each router on the path must admit the reservation, carried in the RESV message, in order for the call to be accepted. In the DiffServ model the routers in the network do not have per-flow awareness. The customer has an SLS with the provider in which the traffic profile of the aggregate traffic is specified. An example of a DiffServ SLS template can be found in [7].

In both the RSVP PATH message and the SLS the traffic profile is specified in terms of token bucket parameters. Choosing the appropriate traffic descriptors is an important issue. Overdimensioning the traffic descriptors will result in a waste of resources and possibly higher costs while underdimensioning will result in packet loss (when excess traffic is dropped) and hence degrade voice quality. When the reservation is admitted the traffic profile is enforced at the edge of the network. Discarding out-of-profile packets is called policing. The fraction of discarded packets is the Packet Discard Ratio (PDR).

B. Previous work

Research on traffic specifications and policing took off at the advent of ATM networks [10]. In ATM networks, traffic was described in terms of dual Leaky Bucket (LB) parameters: peak and sustainable cell rate and maximum burst size. In [5] a method was introduced to calculate the effective bandwidth (i.e. sustainable rate) of an on-off source. In IP networks, however, the packet size is variable. Therefore, the token bucket algorithm is used and the parameters are expressed in byte rates instead of cell rates.

The contribution of this paper is two-fold. First, it is shown how to choose the traffic descriptors for aggregated voice with and without Voice Activity Detection (VAD) for a given tolerated PDR at the policer. Secondly, the influence of the packet size on the PDR is investigated. Because large packets have a larger PDR than smaller packets a modification to the token bucket algorithm is proposed.

C. Contents of this paper

This paper is organized as follows. In section II, the model for voice sources is introduced. In section III, the token bucket algorithm, which is used for policing, is discussed. Section IV shows, based on simulations, how to choose the token bucket parameters for policing. Section V discusses the influence of the voice packet size on the PDR. Also a modification to the policing algorithm is proposed in order to make the PDR independent of the packet size. Finally, conclusions are drawn and directions for future research are given in section VI.

0-7803-7400-2/02/$17.00 (C) 2002 IEEE

II. VOICE SOURCE MODEL

The model for a constant bit rate source is quite simple. The source transmits packets of the same size with packet interdeparture time T_{pack}. The size of the packets is determined by the bit rate of the codec R_{cod} and the packetization delay T_{pack}. The packetization delay is the time needed to fill one IP packet. The packet size M consists of a 40-byte RTP/UDP/IP header and a payload. It can be expressed as $M=(R_{cod} \cdot T_{pack})/8+$ 40 byte.

When VAD is used the source only transmits packets when the user speaks. Hence, the voice source behaves as an on-off source. The characteristics of speech activity during a two-way conversation have extensively been studied in [9] and are specified in ITU-T recommendation P.59 [1]. Three states can be distinguished during a conversation:

- the speech state, the user is talking to the other party,
- the pauses between words and syllables (most of them are smaller than 200 ms),
- the silence state, the user listens to the other party.

In most VAD mechanisms a certain hangover time is used. Most VAD mechanisms are conservative in that they do not immediately switch from the speech to the silence state. This hangover time is the time the VAD mechanism remains in the speech state after detecting silence before switching to the silence state. The hangover time is usually set around 200 ms in order to filter out the pauses between words and syllables. The state of the conversation (speaking, listening), the characteristics of speech activity and the speech activity determined with VAD are shown as an example in Figure 1.

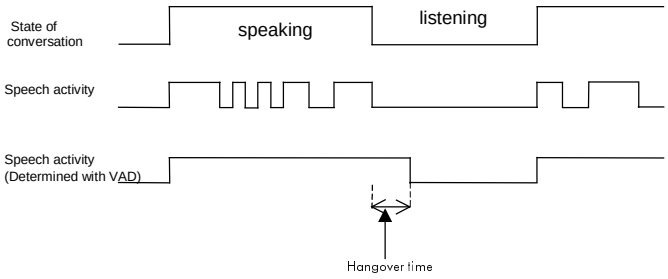

Figure 1: Speech activity in a two-way conversation

According to [1] the length of the on (speech) and off (silence) periods approximately have exponential tail distribution functions. The talkspurt duration can be modeled by a deterministic value (the hangover time) plus an exponentially distributed variable and the silence duration by an exponentially distributed variable. The average on period is 1.004 s and the average off period is 1.587 s. This results in an average speech activity of 39%.

III. TOKEN BUCKET ALGORITHM

The Token Bucket (TB) algorithm can be used for either policing (i.e. dropping non-conformant packets) or shaping (i.e. delaying non-conformant packets). The algorithm uses two parameters:

- token rate r in byte/s, this is the maximum sustainable bit rate,
- maximum burst size b in byte, this is the maximum size of a burst of packets that is conformant.

The traffic profile $A(t)$ that is defined with TB parameters (r, b) equals: $A(u,t) = r \cdot (t-u) + b$ byte $\forall\ t>0,\ u<t$. For bursty sources two traffic envelopes have to be specified, one for the peak rate and one for the sustainable rate. The token bucket algorithm is depicted in Figure 2. The parameter T_k denotes the total number of tokens generated over the interval $[0,t]$. ΔT_k is the number of tokens generated between the arrival instants t_k and t_{k-1} of packet k and $k-1$. $B(t_k)$ denotes the number of tokens in the bucket at the arrival time t_k of packet k. B_{th} is the decision threshold for discarding a packet. In the traditional TB algorithm it is 0, i.e., if there are less tokens than the packet size in byte the packet is discarded.

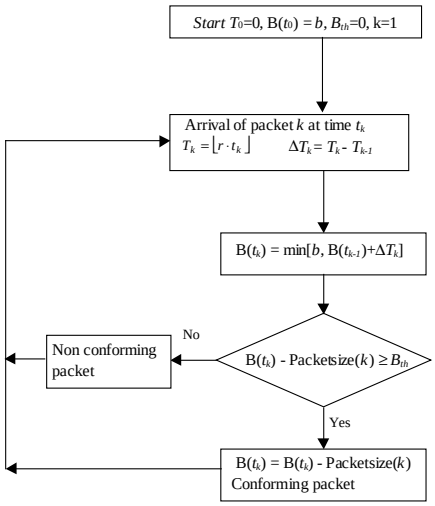

Figure 2: Token bucket algorithm

The algorithm can also be explained as shown in Figure 3. There is a token buffer, which is filled with tokens at rate r. When the buffer contains b tokens all new tokens are discarded. When a packet arrives and there is at least a number of tokens in the buffer equal to the packet size in bytes then the packet is conformant. Otherwise the packet is not conformant and will be discarded when the algorithm is used for policing purposes.

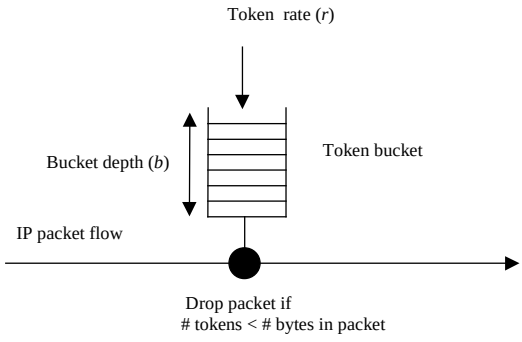

Figure 3: Token bucket policer

If a flow is conformant to a (*r*, *b*) traffic specification and is served by a FIFO system with queue length *b* and service rate *r* then no packet loss will occur. Hence, determining the PDR of a token bucket policer is equivalent to determine the overflow probability of a (FIFO) queuing system with a finite buffer of length *b*. To illustrate this we draw, as an example, the bucket contents and the buffer occupancy of a FIFO queue for the same arrival process in Figure 4. When the bucket gets empty the equivalent FIFO system (with buffer length *b*) is overflowing. Hence, the overflow probability of the FIFO queue is equal to the PDR of the token bucket algorithm.

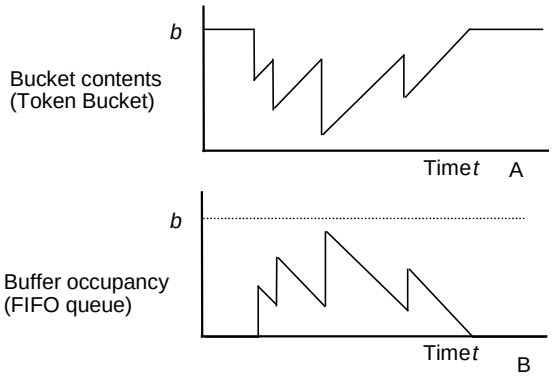

Figure 4: Similarity bucket contents (A) and FIFO buffer occupancy (B)

In the rest of the paper we define the load of a policer as the ratio between the mean rate *m* and *r*, as in a queuing system.

IV. ESTIMATING POLICING PARAMETERS

In DiffServ networks the traffic specification for an aggregate flow is defined in an SLS. In this section we discuss how to choose the token bucket parameters. This is done for both sources with VAD (i.e. on-off sources) and without VAD (i.e. constant bit rate sources).

When traffic of different constant bit rate sources is multiplexed the resulting traffic stream is not constant bit rate anymore. According to [3] the worst arrival pattern that can occur when multiplexing constant bit rate flows is a Poisson process. We simulated the token bucket algorithm using OPNET and used a Poisson process with a load equal to 1000, 2000 and 5000 voice sources. The aggregate of voice sources is a mix of sources, which is shown in Table 1 for an aggregate of 1000 sources. For 2000 or 5000 sources the ratio between the number of sources remains the same. The average bit rate of one source including overhead was 28 kb/s.

Table 1: Mix of 1000 sources used in simulation

Codec	G.711	G.723.1	G.723.1	G.729	GSM-FR
Bit rate [kb/s]	64	5.3	6.3	8	13
Tpack [ms]	10,20,..., 80	30, 60, 90	30, 60, 90	10,20,..., 80	20, 40, 60, 80
Nr. Sources with same codec	200	200	200	200	200
Nr. Sources with same codec and Tpack	25	67	67	25	50

In Figure 5 the curves are shown for different sustainable rates in the case of an aggregate of 5000 sources. The numbers in the legend denote the sustainable rate in Mb/s. In other words the sustainable rate is varied while the arrival rate, with mean 54 Mb/s, remains the same.

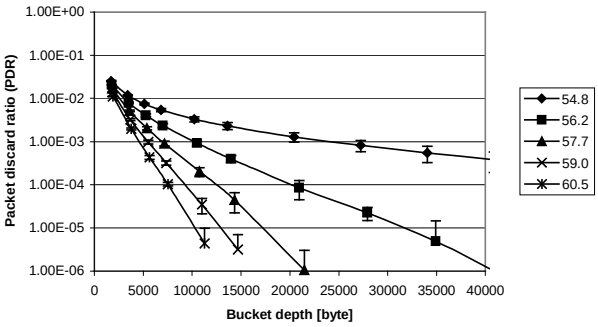

Figure 5: PDR for 5000 Poisson sources (*m*=54 Mb/s)

When on-off sources (i.e. voice sources with VAD) are multiplexed the behavior is different from the case of an aggregate of constant bit rate sources [11]. In Figure 6 to 8 the PDR curves are shown for aggregates of on-off sources. Also the curves for the Poisson process are included in the figures for comparison. The on-off sources are the curves that are denoted by 'a' and the Poisson source are denoted by 'p' in the legend. The numbers in the legend denote again the sustainable rate in Mb/s. The load for the three curves is respectively 0.95, 0.92 and 0.9 in all three figures.

From these curves we can conclude that an aggregate of on-off sources can be modeled as a Poisson source for small burst tolerances and are more bursty than Poisson for larger burst tolerances. Note that the point where the Poisson curve and the on-off aggregate start to differ depends on the burst size and not on the sustainable rate. This is because considered only over a small time scale there is no correlation between the arrivals. However, at a coarser time scale correlation will occur in the arrival process and the PDR will be influenced severely. Furthermore, the region where the on-off aggregate cannot be modeled as a Poisson source is not interesting for the purpose of dimensioning policers since increasing the maximum burst size results only in a modest decrease of the PDR. Hence, in this case the sustainable rate should be increased rather than the maximum burst size.

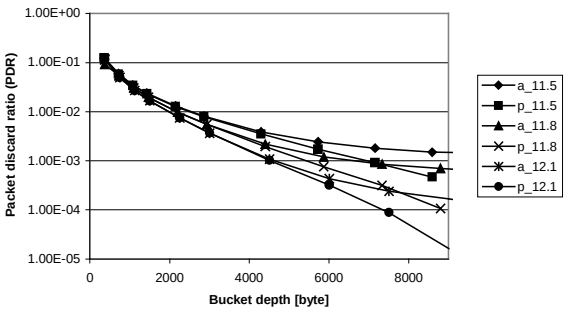

Figure 6: PDR for 1000 on-off sources (*m*=11 Mb/s)

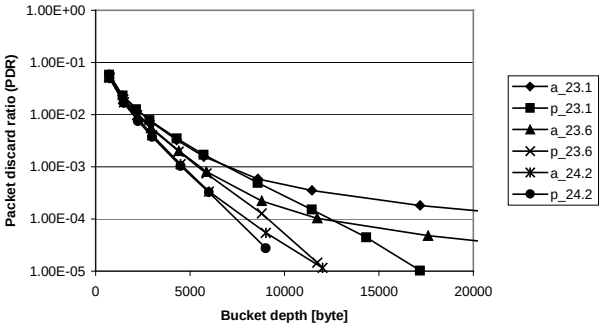

Figure 7: PDR for 2000 on-off sources (m=22 Mb/s)

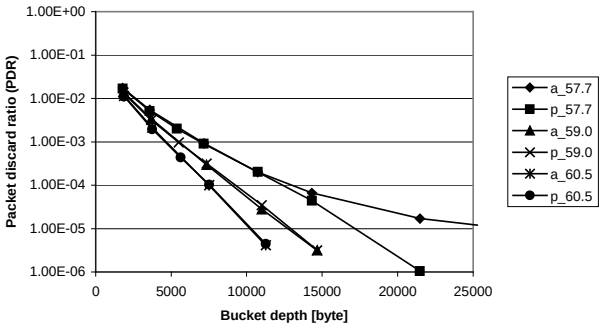

Figure 8: PDR for 5000 on-off sources (m=54 Mb/s)

In Figure 9 the load (m/r) is fixed at 0.92. The Poisson and on-off aggregates are shown again for 1000, 2000 and 5000 sources. The curves of the Poisson sources coincide because the overflow probability only depends on the load and not on the arrival intensity for Poisson arrivals. From this figure it becomes clear that the more sources are aggregated the better (i.e. for smaller PDRs) the Poisson approximation can be used for an aggregate of on-off sources.

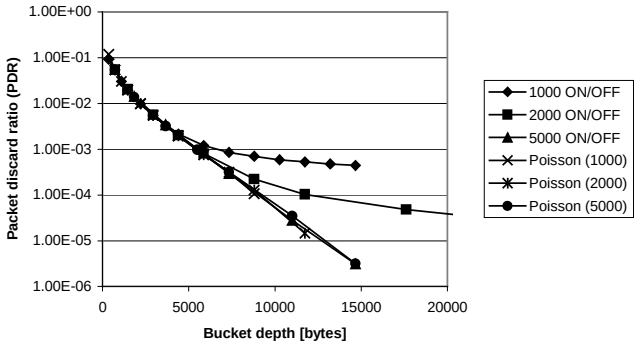

Figure 9: PDR curves for load 0.92

In Figure 10 the largest burst size is shown up to where the Poisson approximation for the aggregate of on-off sources is still accurate as a function of the number of on-off sources. The points are determined by judging in Figure 9 where the Poisson approximation statistically differs from the behavior of the real aggregate.

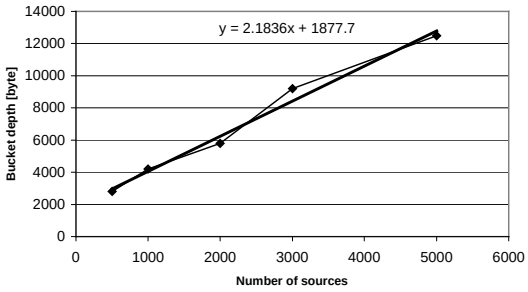

Figure 10: Burst size where the Poisson approximation can be used

V. INFLUENCE OF PACKET SIZE

In this section we investigate the influence of the packet size on the PDR. Therefore, we did a simulation with an aggregate of 1000 on-off sources with a mean packet rate of 2700 packet/s. Thirty seconds were simulated resulting in an average of 81000 sent packets. Seven different packet sizes (40, 70, 100, 130, 160, 190, 220 byte) were used. The packet size was uniformly distributed. The token bucket algorithm was simulated for three different maximum burst sizes. The number of discarded packets for each packet size was collected. This is shown in Figure 11. The numbers in the legend are the maximum burst sizes in byte. From the simulation results we can conclude that the PDR of the token bucket algorithm is dependent on the packet size. In other words, the larger the packet size the larger the PDR becomes.

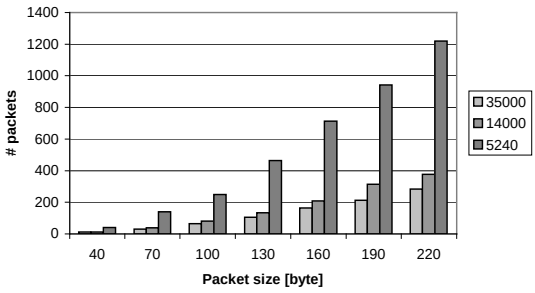

Figure 11: Discarded packets per packet size

In order to avoid this problem we propose to modify the token bucket algorithm slightly. This is done by setting a threshold such that all packets are rejected once the number of tokens is smaller than B_{th}. Hence, smaller packets are discarded while there may still be enough tokens available for it. B_{th} should be chosen equal to the maximum packet size M, as opposed to the standard token bucket algorithm where B_{th}=0. In this case the PDR for small packet sizes increases while it decreases for larger packet sizes. However, the total number of rejected bytes of the aggregate for both the TB algorithm and the modified version is nearly the same.

The simulation was run again but now with the modified token bucket algorithm. Again the packet discards for each packet size were collected, shown in Figure 12. From this figure we can conclude that the PDR is the same for all packet sizes. Hence, sources with higher bit rate codecs experience the same PDR at the policer as sources with low bit rate codecs when using the same packetization delay. The modification of the algorithm does not change how the traffic descriptors have to be chosen.

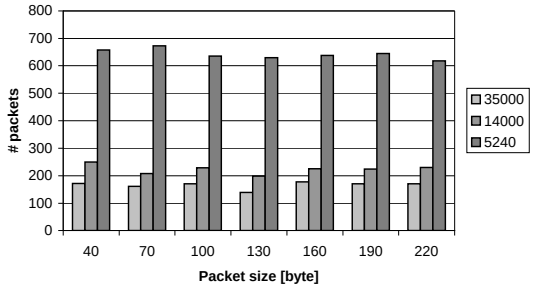

Figure 12: Discarded packets (modified TB algorithm)

VI. Conclusions and Future Work

In QoS enabled IP networks some form of resource reservation (e.g. SLS) is required in order to be able to guarantee the network performance in terms of delay and packet loss. Hence, the maximum amount of traffic has to be specified. This is done with token bucket parameters. In this paper we discussed how to choose these traffic descriptors for voice. When constant bit rate sources are aggregated the aggregate stream behaves at worst as a Poisson process. By determining the overflow probability of a FIFO queue with finite buffer length one can determine the possible combinations of the sustainable rate and maximum burst size. When voice sources with VAD (on-off sources) are aggregated the aggregate behaves as Poisson for small burst sizes and more bursty than Poisson for larger burst sizes.

The token bucket parameters are best chosen in the region where the aggregate of on-off sources behaves as Poisson because increasing the burst size beyond this region will results only in a small decrease of the PDR. In this case the sustainable rate should be increased rather than the maximum burst size.

The token bucket policer discards packets as soon as the number of bytes in the packet is larger than the number of tokens in the bucket. In this case the PDR is related to the packet size. The larger the packet the larger the discard probability. To solve this problem it is proposed to introduce a minimum number of tokens in the bucket below which all packets are discarded. This minimum number of tokens should be chosen equal to the maximum packet size of the aggregate.

This modification of the token bucket algorithm results in the fact that the PDR is independent of the packet size. A future research topic is policing of variable bit rate video (e.g. MPEG).

Acknowledgment

This work was carried out within the framework of the project CoDiNet, sponsored by the Flemish institute for the promotion of scientific and technological research in the industry (IWT).

References

[1] Artificial Conversational Speech, ITU-T recommendation P.59, March 1993.

[2] S. Blake, D. Black, M. Carlson, E. Davies, Z. Wang and W. Weiss, “An Architecture for Differentiated Service”, *IETF Request for Comment 2475*, December 1998.

[3] T. Bonald, A. Proutière and J.W. Roberts, “Statistical Performance Guarantees for Streaming Flows using Expedited Forwarding”, Proceedings of INFOCOM 2001, Volume 2, pp. 1104-1112, Anchorage (AK), USA, April 2001.

[4] R. Braden, D. Clark and S. Shenker, “Integrated Services in the Internet Architecture: an Overview”, *IETF Request for Comment 1633*, June 1994.

[5] M.J.C. Büchli, D. De Vleeschauwer, J. Janssen, A. Van Moffaert, G.H. Petit, “On the Efficiency of Voice over Integrated Services using Guaranteed Service”, *Proceedings of the 2nd IP-Telephony Workshop (IPTEL 2001)*, pp. 6-13, New York City (NY), USA, 2-3 April 2001.

[6] A.I. Elwalid and D. Mitra, “Effective Bandwidth of General Markovian Traffic Sources and Admission Control of High Speed Networks”, *Proceedings of INFOCOM 1993, IEEE Computer Society Press*, pp. 256-265, 1993.

[7] D. Goderis et al., “Service Level Specification Semantics, Parameters and negotiation requirements”, *IETF draft <draft-tequila-sls-01.txt>, work in progress*, June 2001.

[8] J. Janssen, D. De Vleeschauwer, G.H. Petit, “Delay and Distortion Bounds for Packetized Voice Calls of Traditional PSTN Quality”, *Proceedings of the 1st IP Telephony workshop (IPTEL 2000)*, GMD report 95, pp. 105-110, Berlin (Germany), April 2000.

[9] H.H. Lee and C.K. Un, “A Study of On-Off Characteristics of Conversational Speech”, *IEEE Transactions on Communications*, vol. Com-34, no. 6, June 1986.

[10] M. Ritter and P. Tran-Gia, “Performance Analysis of Cell Rate Monitoring Mechanisms in ATM Systems”, University of Würzburg, 1994.

[11] K. Sriram and W. Whitt, “Charaterizing Superposition Arrival Processes in Packet Multiplexers for Voice and Data”, *IEEE Journal of Selected Areas in Communication*, vol. 4, no. 6, pp. 833-846, 1986.

Revised: 4 January, 2002