UDC 004.73(042.3)

Chang Shu, Ph. D. (Computer Sciences) (Tel.: +86 029 88182667. E-mail: changshucn@126.com)
(Northwest University of Politics & Law, People Republic of China)

ADAPTIVE CONTROL TRAFFIC RATE by MULTIVARIABLE REGULATING TOKEN SHAPER

Чанг Шу. Адаптивна система управління мережевимтрафіком з багатопараметричним непрямим зворотним зв'язком. У роботі представлений метод адаптивного формування потоків мережного трафіку і метод настройки систем управління з непрямим зворотним зв'язком. За рахунок наявності зворотного зв'язку можливе підстроювання параметрів і структури формувача під зміни інтенсивності і статистичні характеристики трафіку. Розроблений алгоритм і пристрій багатопараметричної адаптації формувача до зміни безлічі параметрів вхідного трафіку. Показано, що згаданий алгоритм має набагато більше степенів свободи, ніж традиційні алгоритми формування трафіку. Завдяки цьому забезпечується хороша гнучкість алгоритму і ефективне функціонування мережі в широкому діапазоні умов.

Ключові слова: мережевий трафік, система управління, багатопараметрична адаптація, непрямий зворотний зв'язок

Чанг Шу. Адаптивная система управления сетевым трафиком с многопараметрической косвенной обратной связью. В работе представлен метод адаптивного формирования потоков сетевого трафика и метод настройки систем управления с косвенной обратной связью. За счет наличия обратной связи возможна подстройка параметров и структуры формирователя под изменения интенсивности и статистические характеристики трафика. Разработан алгоритм и устройство многопараметрической адаптации формирователя к изменению множества параметров входящего трафика. Показано, что упомянутый алгоритм имеет гораздо больше степеней свободы, чем традиционные алгоритмы формирования трафика. Благодаря этому обеспечивается хорошая гибкость алгоритма и эффективное функционирование сети в широком диапазоне условий.

Ключевые слова: сетевой трафик, система управления, многопараметрическая адаптация, косвенная обратная связь

Chang Shu. Adaptive control traffic rate by multivariable regulating token shaper. There are on the presentation a method of the adaptive shaping of flows of network traffic and method of tuning of the control systems with an indirect feedback. Due to availability feedback the adjustment parameters and structure of shaper to variations of traffic intensity and statistical distributions is possible. The algorithm and device of multivariable adaptation of shaper to changing the set of parameters of input traffic is developed. It was shown that mentioned algorithm has much more ranges of freedom. So it has good flexibility and is able to provide efficient functioning of network in wide range of conditions.

Keywords: network traffic, control e system, multivariable adaptation, indirect feedback

I. Introduction

For bursty traffic flows with random time-varying rates, it makes sense to allocate bandwidth less than the peak rates and assume that some capacity can be saved by statistical multiplexing [1]. When a large number of flows are multiplexed, it is unlikely that all flows will be bursting at their peak rates simultaneously. Hence, an amount of bandwidth somewhat less than the sum of peak rates is needed.

In any case, the delay through the queue and probability of buffer overflow are calculated for the hypothetical traffic. If the delay and buffer overflow probability meet the desired QoS, then the new traffic burst is accepted [2].

Algorithms of leaky and/or token bucket are using for policing/shaping network traffic [3]. Those algorithms can be deployed throughout network to ensure that a packet, or data source, adheres to a stipulated contract and to determine the QoS to render the packet. Both policing and shaping mechanisms use the traffic descriptor for a packet – indicated by the classification of the packet – to ensure adherence and service.

Policers and shapers usually identify traffic descriptor violations in an identical manner. They usually differ, however, in the way they respond to violations, for example:

– a policer typically drops traffic. (For example, the CAR rate-limiting policer will either drop the packet or rewrite its IP precedence, resetting the type of service bits in the packet header.);

– a shaper typically delays excess traffic using a buffer, or queuing mechanism, to hold packets and shape the flow when the data rate of the source is higher than expected. (For example, GTS and Class-Based Shaping use a weighted fair queue to delay packets in order to shape the flow, and DTS and FRTS use either a priority queue, a custom queue, or a FIFO queue for the same, depending on how you configure it.).

Traffic shaping and policing can work in tandem. For example, a good traffic-shaping scheme should make it easy for nodes inside the network to detect misbehaving flows. This activity is sometimes called policing the traffic of the flow.

II. Terms and definitions

The rate-limiting features of committed access rate (CAR) and the Traffic Policing feature provide the functionality for policing traffic. The features of Generic Traffic Shaping (GTS), Class-Based Shaping, Distributed Traffic Shaping (DTS), and Frame Relay Traffic Shaping (FRTS) provide the functionality for shaping traffic.

A token bucket is a formal definition of a rate of transfer. It has three components: a burst size, a mean rate, and a time interval (T_c). Although the mean rate M_r is generally represented as bits per second, the relation shown as follows may derive any two values from the third: $M_r = B_c / T_c$, where B_c is burst size.

Here are some definitions of these terms:

– *mean rat*e – also called the committed information rate (CIR), it specifies how much data can be sent or forwarded per unit time on average;

– *burst size* – also called the Committed Burst (B_c) size, it specifies in bits (or bytes) per burst how much traffic can be sent within a given unit of time to not create scheduling concerns. (For a shaper, such as GTS, it specifies bits per burst; for a policer, such as, it specifies bytes per burst.);

– time interval – also called the measurement interval, it specifies the time quantum in seconds per burst.

A token bucket is used to manage a device that regulates the data in a flow. For example, the regulator might be a traffic policer, such as CAR, or a traffic shaper, such as FRTS or GTS. A token bucket itself has no discard or priority policy. Rather, a token bucket discards tokens and leaves to the flow the problem of managing its transmission queue if the flow overdrives the regulator.

If not enough tokens are in the bucket to send a packet, the packet either waits until the bucket has enough tokens (in the case of GTS) or the packet is discarded or marked down (in the case of CAR). If the bucket is already full of tokens, incoming tokens overflow and are not available to future packets.

Thus, at any time, the largest burst a source can send into the network is roughly proportional to the size of the bucket.

Note that the token bucket mechanism used for traffic shaping has both a token bucket and a data buffer, or queue; if it did not have a data buffer, it would be a policer. For traffic shaping, packets that arrive that cannot be sent immediately are delayed in the data buffer.

For traffic shaping, a token bucket permits burstiness but bounds it. It guarantees that the burstiness is bounded so that the flow will never send faster than the token bucket's capacity, divided by the time interval, plus the established rate at which tokens are placed in the token bucket.

See the following formula: $S_{fl\max} = C_{tb}/T_c + E_r$, where C_{tb} is token bucket capacity in bits, and E_r is established rate in bps, and $S_{fl\max}$ is maximum flow speed in bps.

This method of bounding burstiness also guarantees that the long-term transmission rate will not exceed the established rate at which tokens are placed in the bucket.

A single CAR rate policy includes information about the rate limit, conform actions, and exceed actions.

Each interface can have multiple CAR rate policies corresponding to different types of traffic. For example, low priority traffic may be limited to a lower rate than high priority traffic. When there are multiple rate policies, the router examines each policy in the order entered until the packet matches. If no match is found, the default action is to send.

Rate policies can be independent: each rate policy deals with a different type of traffic. Alternatively, rate policies can be cascading: a packet may be compared to multiple different rate policies in succession.

Cascading of rate policies allows a series of rate limits to be applied to packets to specify more granular policies (for example, you could rate limit total traffic on an access link to a specified sub rate bandwidth and then rate limit World Wide Web traffic on the same link to a given proportion of the sub rate limit) or to match packets against an ordered sequence of policies until an applicable rate limit is encountered (for example, rate limiting several MAC addresses with different bandwidth allocations at an exchange point). You can configure up to a 100 rate policies on a sub interface.

Traffic policing allows to control the maximum rate of traffic sent or received on an interface, and to partition a network into multiple priority levels or class of service (CoS). The Traffic Policing feature manages the maximum rate of traffic through a token bucket algorithm. The token bucket algorithm can use the user-configured values to determine the maximum rate of traffic allowed on an interface at a given moment in time. The token bucket algorithm is affected by all traffic entering or leaving (depending on where the traffic policy with Traffic Policing configured) and is useful in managing network bandwidth in cases where several large packets are sent in the same traffic stream.

The token bucket algorithm provides users with three actions for each packet: a conform action, an exceed action, and an optional violate action. Traffic entering the interface with Traffic Policing configured is placed in to one of these categories. Within these three categories, users can decide packet treatments. For instance, packets that conform can be configured to be transmitted, packets that exceed can be configured to be sent with a decreased priority, and packets that violate can be configured to be dropped.

Traffic Policing is often configured on interfaces at the edge of a network to limit the rate of traffic entering or leaving the network. In the most common Traffic Policing configurations, traffic that conforms is transmitted and traffic that exceeds is sent with a decreased priority or is dropped. Users can change these configuration options to suit their network needs.

The primary reasons you would use traffic shaping are to control access to available bandwidth, to ensure that traffic conforms to the policies established for it, and to regulate the flow of traffic in order to avoid congestion that can occur when the sent traffic exceeds the access speed of its remote, target interface.

III. Traditional Mechanism of Traffic Shaping

Traffic shaping smoothes traffic by storing traffic above the configured rate in a queue. When a packet arrives at the interface for transmission, the following sequence happens:

1. If the queue is empty, the traffic shaper processes arriving packet at once:

– if possible, the traffic shaper sends the packet;

– otherwise, the packet is placed in the queue.

2. If the queue is not empty, the packet is placed in the queue.

When packets are in the queue, the traffic shaper removes the number of packets it can send from the queue every time interval.

Generic Traffic Shaping (GTS) shapes traffic by reducing outbound traffic flow to avoid congestion by constraining traffic to a particular bit rate using the token bucket mechanism (see Fig. 1).

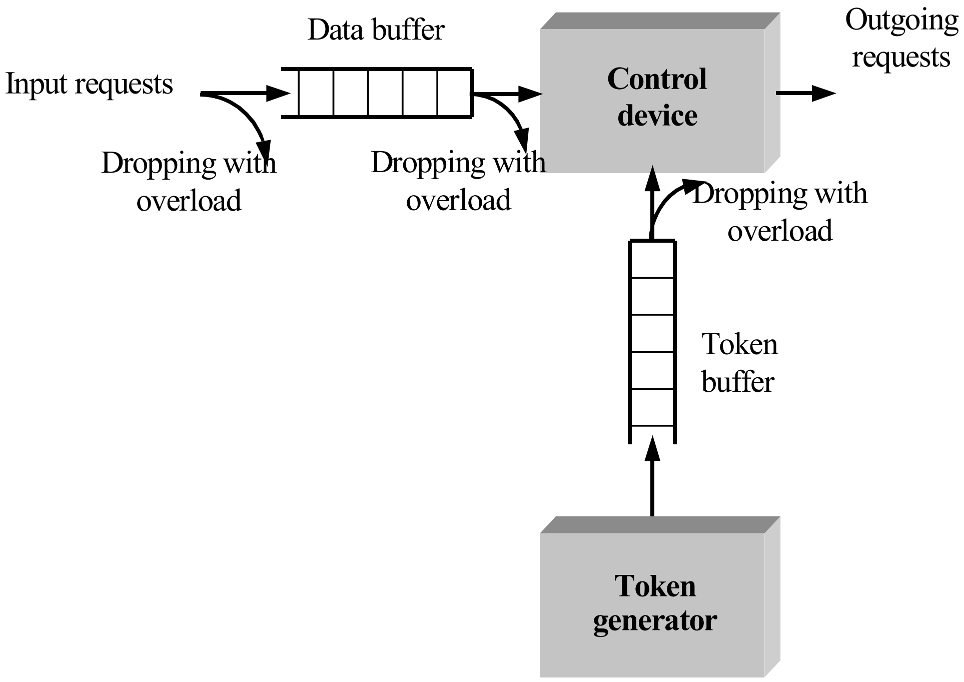

Fig. 1. Traditional token bucket mechanism

The description of token bucket algorithm

1. Token generator of the TG_E throws down in E bucket the tokens at a speed of E_{IR} per second. If a bucket is filled, excess tokens are dropped.

Time of filling $t_{fE} = E_{BS} / E_{IR}$.

2. The token generator throws down in a C bucket tokens at a speed of C_{IR} per second. If a bucket is filled, excess tokens dropped.

Time of filling $t_{fC} = C_{BS} / C_{IR}$.

3. Tokens are in the E and C buckets. Total length of the time interval occupied by token in the E bucket is equal $\tau_e = \tau_{te} + \tau_{ge}$, where τ_{te} is length of token in the E bucket; τ_{ge} it is length of guard interval.

Total length of the time interval occupied by a token in a C bucket is equal

$$\tau_c = \tau_{tc} + \tau_{gc},$$

where τ_{tc} is length of token in a C bucket; τ_{gc} it is length of guard interval.

The number of tokens in the E bucket is equal n_e, in a C bucket equal n_c.

Then the total size of tokens in the E bucket is equal $T_e = n_e \cdot \tau_e$, in a C bucket equal $T_c = n_c \cdot \tau_c$.

IV. Adaptive Mechanisms of Traffic Shaping

Adaptation to the change of length and instantaneous intensity of entering packets can be carried out as follows:

- by changing length of token at permanent length of guard interval;
- by changing length of protective interval at permanent length of token;
- by changing size of "yellow range" [3];
- by changing size of data and token buffer memory.

A counter counts up the number of packets coming on the entrance of shaper. A store, essentially, is discrete integrator, scalar or vector. Frequency of token generator is regulated depending on the number of the accumulated packets, speed of accumulation (and in theory – and higher derivative). At devastation of token buffer (bucket) the growth of speed can be limited, based the parameters of incoming traffic and potential possibilities of destination node.

Anyway speeds of E_{IR} and C_{IR} will change to the limits which depend on the maximal carrying capacity of switch node. It's expedient adapting to the change of middle intensity of packets by the change of the speeds of E_{IR} and C_{IR} and changing size of "yellow range".

The chart of multivariable adaptive token bucket mechanism is shown on Fig. 2 (changing size of "yellow range" isn't shown).

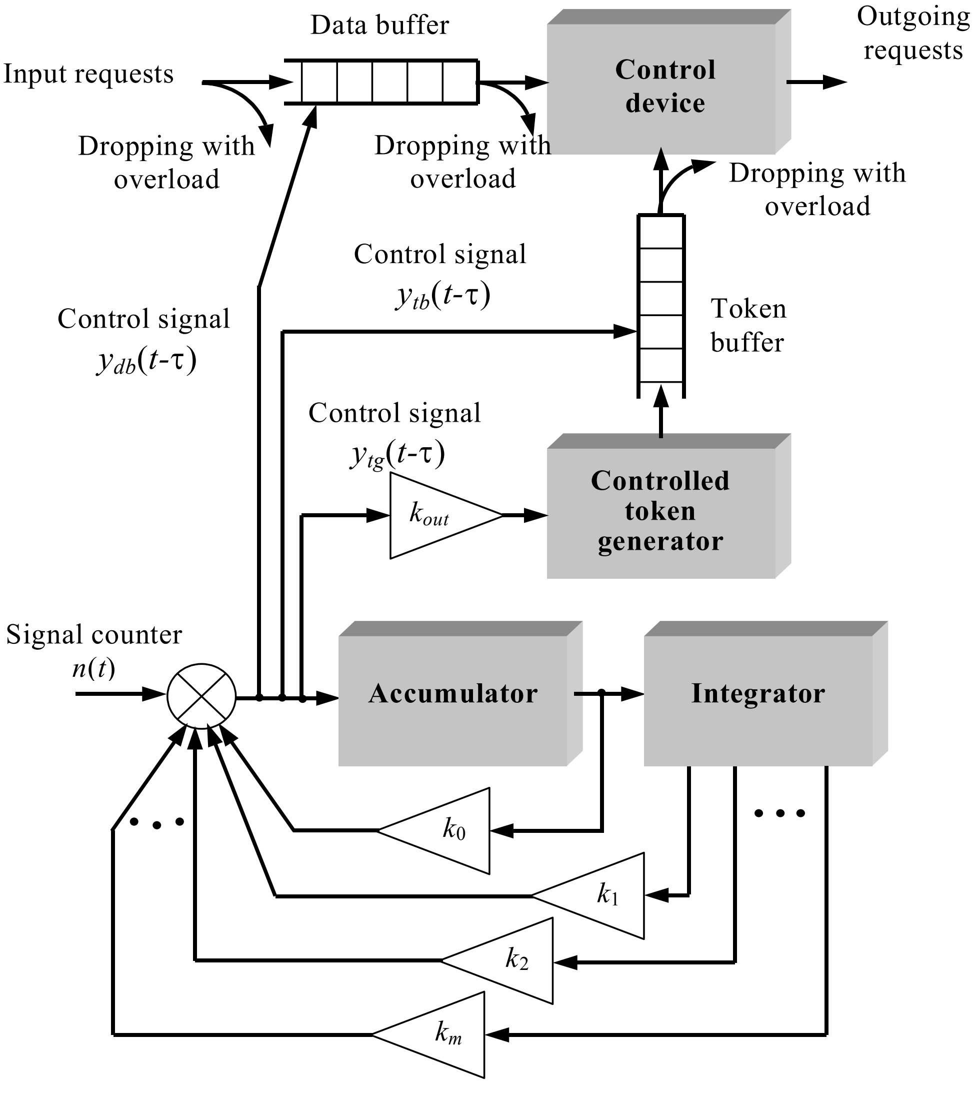

Fig. 2. The chart of multivariable adaptive token bucket mechanism (changing size of "yellow range" isn't shown)

M-range integrator with weight coefficients $k_1 = k_1(t), k_2 = k_2(t), \ldots, k_m = k_m(t)$ estimates speed, acceleration and higher derivatives of packets flow. Control signals $y_{db}(t-\tau)$, $y_{tb}(t-\tau)$, $y_{tg}(t-\tau)$ regulate size of data and token buffers and frequency of tokens series. Those signals are defined through traffic parameters including instant intensity and kind of statistic distribution (for instance heavy-tail distribution if traffic is self-similar).

In practice it's inappropriate to calculate derivatives of accumulation process higher than 2nd (speed and acceleration) due to quick deterioration of precision of results. So those results will be without effect on resulting efficiency of control process.

Besides the multivariable adaptive mechanisms of traffic shaping have a lot of ranges of freedom in contrast with traditional approaches. As a result we get auxiliary options for storing of arriving traffic and excluding packets loss and reduce the quantity of retransmission.

V. Conclusions

Traffic regulation mechanisms (referred to as policers and shapers) throughout network to ensure that a packet, or data source, adheres to determine the QoS to render the packet. Both policing and shaping mechanisms use the traffic descriptor for a packet indicated by the classification of the packet to ensure adherence and service.

Proposed procedure of traffic shaping is rather simple and efficient. The results of modelling shows that it is possible to limit the frequency of token generator till such value, when all input traffic would been received and then transferred without losses and retransmissions.

In future work we are going to research sensibility of developed devices to deviations traffic flows with various statistical distribution, including heavy-tail distributions adhere for self-similar traffic.

VI. References

1. Chang Shu. The method of adaptive shaping of the traffic flows of calculating networks / Chang Shu, Nick A. Vinogradov. // Proceedings of the fourth world congress "Aviation in the XXI-st century", Sept. 21-23, 2010, Kiev, Ukraine. – Vol.1. – PP. 18.13-18.17.

2. Chang Shu. Shaping of traffic parameters and elimination of overload in aeronautical telecommunication networks / Chang Shu // Наукові записки Українського науково-дослідного інституту зв'язку. – 2011. – №2(18)ю – С. 52-57.

3. Tanenbaum, A.S. Computer Networks, 5th Ed. / Andrew S. Tanenbaum, David J. Wetherall. – Prentice Hall, Cloth, 2011. – 960 pp.

4. Савченко А. С. Повышение качества сервиса в сетях доступа с использованием адаптивных алгоритмов формирования и упорядочения трафика / А. С. Савченко, Чанг Шу // Проблемы информатизации и управления: зб. наук. пр. – 2008. – № 2 (24). – С. 161-169.