USING TOKEN LEAKY BUCKETS FOR CONGESTION FEEDBACK CONTROL IN PACKETS SWITCHED NETWORKS WITH GUARANTEED BOUNDEDNESS OF BUFFER QUEUES

V. Guffens*, G. Bastin*, H. Mounier†

* Centre for Systems Engineering and Applied Mechanics (CESAME) Université Catholique de Louvain, Bâtiment Euler, 4-6, avenue G. Lemaitre, 1348 Louvain-la-Neuve, Belgium. Email: (guffens,bastin)@auto.ucl.ac.be

† Ecole Nationale Supérieure des Mines de Paris, Centre d'Automatique et de Robotique (CAOR), 60 Bd. Saint-Michel, 75006 Paris, France. Email: mounier@caor.ensmp.fr

Keywords: hop-by-hop control, pushback, traffic shaping, fluid flow network modeling, nonlinear control

Abstract

A fluid flow model of a FIFO queueing system is presented and extended to the so-called token leaky bucket case. A simple feedback strategy that guarantees the boundedness of packets buffer queue is then introduced. Some simulations are presented and confronted to experiments run on a network made of Linux machines.

1 Introduction

It is well known that the congestion avoidance features of TCP have served the Internet for years in preventing congestion collapse. However, the multiplication of TCP implementations accessing the network and an increasing use of "delay sensitive" applications are augmenting the number of flows that are not sufficiently responsive to congestion notification. This problem is discussed in RFC2309 [1] where the use of active queue management techniques such as Random Early Detection (RED) are recommended in order to maintain an average queue size sufficiently small. Nevertheless, these methods do not apply to flows that are not responsive or not responsive enough to congestion signals.

It is also widely accepted that TCP is not able to control the traffic at a time scale smaller than a few round-trip-times (RTT). This time scale, although suitable to ensure the global stability of the Internet, is not sufficient to provide a sufficient quality of service. The end-to-end nature of TCP does not seem appropriate to ensure these type of service that require some knowledge about the state of intermediary hops.

In this paper, we present a simple hop-by-hop feedback control method that is acting at the layer 3 of the OSI model which makes it immune to the "unresponsive flow" problem mentioned above. Basically, this approach ensures the conservation of packets at a hop-by-hop level. In practice the method is implemented by using standard token buckets connected in feedback. The simplicity and the feasibility of this approach is demonstrated by illustrative experiments performed with User Mode Linux (UML).

2 A fluid flow model of the FIFO queue

Consider a host computer connected to the network through a single-server queuing system with constant service rate as depicted in fig. 1

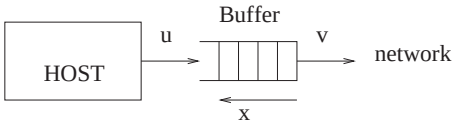

Figure 1: A single server queueing system.

The host is supposed to deliver packets to the buffer at a rate denoted $u(t)$ which can be highly bursty. The load of the buffer is denoted $x(t)$ (i.e. the number of packets in the buffer at time t). A continuous time fluid flow model of the buffer dynamics is as follows :

$$\dot{x} = -v + u \tag{1}$$

where $v(t)$ denotes the rate at which the packets are released to the network. Assuming that the buffer operates under a standard FIFO basis, the following model is proposed in [8]:

$$\dot{x} = -r(x) + u \tag{2}$$

with

$$r(x) = \frac{x}{\theta(x)}$$

$r(x)$ is referred to as the *processing rate function* and is defined as the ratio between the load x and the residence time $\theta(x)$. It should be pointed out that we don't make any a priori assumption on the probability distribution of the incoming traffic. Model (2) has to be interpreted as an averaged description of a wide class of network buffers depending on the particular form of the residence time function $\theta(x)$. For instance, if we select a linear $\theta(x)$ of the form :

$$\theta(x) = \frac{a + x}{\mu} \quad \text{with } a > 0, \mu > 0$$

then, for a constant inflow rate $\bar{u} = \lambda$, the corresponding steady state load $\bar{x}$ is given by :

$$\frac{\mu \bar{x}}{1 + \bar{x}} = \lambda \text{ or } \bar{x} = \frac{\lambda}{\mu - \lambda} \tag{3}$$

In that case, we observe that we recover the classical formula of queueing theory for M/M/1 systems with μ being the service rate of the buffer. Hence, the steady state behavior of an averaged model coincides with the steady state behavior of an M/M/1 queue. Furthermore our averaged model is also suitable for describing non steady-state situations such as the token-leaky bucket case represented in the next section.

3 A fluid flow model of the token leaky buffer

An improvement of this basic queueing strategy is the so-called "token-leaky buffer"(TBF) which allows the output to speed up when large bursts arrive. In this algorithm, the buffer is furnished with a "token bucket" which controls the service rate of the buffer as shown in fig. 2. In this algorithm, the bucket

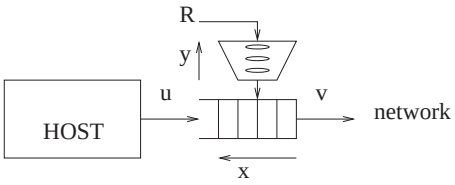

Figure 2: The token leaky buffer.

is filled by tokens at a constant rate $R > 0$, while a token is removed from the bucket each time a packet leaves the buffer. In addition, the service rate of the buffer is modulated by the level y of tokens in the bucket in such a way that $v = \mu$ when there are tokens in the bucket but $v = R < \mu$ when the bucket is nearly empty.

A continuous time fluid model of the server-bucket system is as follows:

$$\begin{aligned} \dot{x} &= -\frac{\mu x}{1+x}\frac{y}{\epsilon+y} + u \\ \dot{y} &= \begin{cases} -\frac{\mu x}{1+x}\frac{y}{\epsilon+y} + R & \text{if } 0 \leqslant y \leqslant \sigma \\ 0 & \text{if } y = \sigma \end{cases} \end{aligned} \tag{4}$$

with $v(t) = \mu x/(1+x).y/(\epsilon+y)$

In this model, the term $y/(\epsilon+y)$ is the modulation function mentioned above, with $0 < \epsilon \ll 1$. When $y \gg \epsilon$, it is clear that the bucket system is transparent and therefore operates as a standard single-server queueing system with service rate μ but when y is small ($y \ll \epsilon$), then the outflow rate of the buffer becomes close to R. σ is the size of the bucket which is initially full.

3.1 Burstiness Constraint

For the fluid flow model (4), we have the following positivity and boundedness property :

If $u(t) \geqslant 0\ \forall t$, $x(0) \geqslant 0$ and $0 \leqslant y(0) \leqslant \sigma$ then $0 \leqslant x(t)$ and $0 \leqslant y(t) \leqslant \sigma\ \forall t$

By integrating the second equation of the model (4), we get :

$$\int_{t_0}^{t_1} v(\tau)d\tau = y(t_0) - y(t_1) + R(t_1 - t_0) \tag{5}$$

which implies the following inequality :

$$\int_{t_0}^{t_1} v(\tau)\ d\tau \leqslant \sigma + R(t_1 - t_0) \qquad \forall t_0, t_1 | t_1 > t_0 \tag{6}$$

This inequality called "burstiness constraint" is well known and is discussed for instance in [3] and [2]. If R is a time varying function $R(t)$, the description given above is still valid. The inequality (6) is generalized as

$$\int_{t_0}^{t_1} v(t)\ dt \leqslant \sigma + \int_{t_0}^{t_1} R(t)\ dt \qquad \forall t_0, t_1 | t_1 > t_0 \tag{7}$$

This extension of the token leaky bucket is the core of the feedback strategy that is presented later in this paper as $R(t)$ is used as control variable.

4 A token leaky buffer with feedback

Let us now consider the interconnection of two buffers as shown in fig. 3. These buffers belong to two neighboring routers in a network. The first buffer is equipped with a token bucket as presented above. The second buffer is just a standard FIFO buffer. The point of interest here is the introduction of the feedback strategy: indeed, we can see that the token bucket is no longer fed at a constant rate R but rather at the rate at which its neighbor is sending its traffic. The fluid model corresponding to this system is :

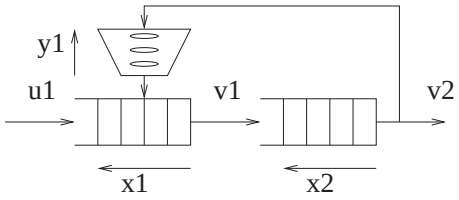

Figure 3: Interconnection of two buffers with feedback.

$$\begin{cases} v_1 = \phi(y_1)\psi(x_1)\mu_1 \\ \dot{y}_1 = v_2 - v_1 \\ \dot{x}_1 = u_1 - v_1 \\ \\ v_2 = \psi(x_2)\mu_2 \\ \dot{x}_2 = v_1 - v_2 \end{cases} \tag{8}$$

where $\phi(y) = y/(\epsilon+y)$ and $\psi(x) = x/(1+x)$

4.1 Property

If the fluid flow model is initialized as follows :

$$x_1(0) = 0 \qquad x_2(0) = 0 \qquad y_1(0) = \sigma_1 > 0$$

And if the inflow rate u_1 is non-negative :$u_1(t) \geqslant 0$ for all t then

$$\begin{array}{ll} \text{a)} & x_1(t) \geqslant 0 \quad x_2(t) \geqslant 0 \quad y_1(t) \geqslant 0 \quad \forall t \\ \text{b)} & y_1(t) + x_2(t) = \sigma_1 \quad \forall t \\ \text{c)} & y_1(t) \leqslant \sigma_1 \quad x_2(t) \leqslant \sigma_1 \quad \forall t \end{array}$$

From this property , we observe that the presence of the feedback loop guarantees that the buffer queue x_2 is naturally bounded by the size of the token bucket and therefore that the transmission is operated without packet loss. This is, at a hop-by-hop level, very similar to the principle of "conservation of packets" or "conservative flow" discussed in the famous paper from Jacobson [5].

4.2 Burstiness constraint

From the fluid flow model (8), the following inequality can be derived :

$$\int_{t_0}^{t_1} v_1 \, dt \leqslant \sigma_1 + \int_{t_0}^{t_1} v_2 \, dt \quad \forall t_0, t_1 | t_1 > t_0 \tag{9}$$

This inequality can be interpreted as a flow constraint, the left buffer is shaping its output according to the output of its neighbor. It will send at most a burst of σ_1 packets and will then send its traffic at a rate that can be sustained by its neighbor.

5 Interconnection with delay

Let's now consider again the system depicted in Fig. 3 with the addition of transmission delays τ, both in the link between the two buffers and in the feedback link. Although the addition of a delay doesn't destroy the boundedness property of the buffer queue length, it may intuitively be thought that a long delay will eventually cause the token bucket to be empty before the feedback information is received, setting $v_1(t)$ to zero. This situation can be analyzed by considering the flow of packets around the bucket and around the second buffer :

$$\dot{y}_1(t) = v_2(t-\tau) - v_1(t) \tag{10}$$
$$\dot{x}_2(t) = v_1(t-\tau) - v_2(t) \tag{11}$$

By time shifting equation (11)

$$\dot{x}_2(t-\tau) = v_1(t-2\tau) - v_2(t-\tau) \tag{12}$$

and eliminating $v_2(t-\tau)$ between (12) and (10)

$$\dot{y}_1(t) = v_1(t-\tau) - \dot{x}_2(t-\tau) - v_1(t) \tag{13}$$

If the system is in a quiescent state before time $t = 0$ and is initialized as in Section 4.1, integrating equation (13) from 0 to t gives :

$$y_1(t) - \sigma = -\int_{t-2\tau}^{t} v_1(\xi) \, d\xi - x_2(t-\tau)$$

Therefore, as $y_1(t) \geqslant 0 \ \forall t$, the following inequality is also true :

$$\frac{1}{2\tau} \int_{t-2\tau}^{t} v_1(\xi) \, d\xi \leqslant \frac{\sigma - x_2(t-\tau)}{2\tau} \tag{14}$$

The presence of a propagation delay limits the maximum achievable average throughput of the system. This problem is typical of systems with high bandwidth-delay product and is discussed, for instance in [7].

6 Implementation of the feedback loop

In practice, the feedback loop cannot be implemented on a per packet basis as it would generate too much overhead traffic. Instead, the number of outgoing packets are counted and this information is sent back at regular intervals, Δ, to the neighbor who originated these packets. As in the previous section, this modification does not destroy the boundedness properties discussed so far but puts some limits on the maximum average throughput of the system. The following equation can be written for $\dot{y}_1$ ($\delta(t)$ indicates the Dirac function and $1_+(t)$ is the step function) :

$$\dot{y}_1(t) = \sum_{k=1}^{\infty} \int_{(k-1)\Delta}^{k\Delta} v_2(\xi) \, d\xi \, \delta(t - k\Delta) - v_1(t)$$

After integration, it comes :

$$\begin{aligned} y_1(t) - \sigma &= \sum_{k=1}^{\infty} \int_{(k-1)\Delta}^{k\Delta} v_2(\xi) \, d\xi \, 1_+(t - k\Delta) \\ &\quad - x_2(t) - \int_0^t v_2(\xi) \, d\xi \end{aligned}$$

And finally, as $y_1(t) \geqslant 0 \ \forall \, t$,

$$\frac{1}{\Delta} \int_{(k-1)\Delta}^{k\Delta} v_2(\xi) \, d\xi \leqslant \frac{\sigma - x_2(t)}{\Delta} \tag{15}$$
$$(k-1)\Delta < t < k\Delta, \ k = 1, \ldots, \infty$$

7 Simulations and experimental results

In this section, fluid flow simulations of the TBF and the TBF with feedback (TBFFB) are presented and confronted with experimental results. The experimental setup is realized with User Mode Linux (UML)[4], a user mode port of the Linux kernel into itself. With the help of a backend switch daemon, virtual machines are connected together to form a virtual network. The network sniffer "tcpdump"[6] is used to collect network traces.

7.1 The simple token leaky bucket

We first consider the implementation of a simple TBF as traffic shaper. The parameters used for the TBF are fixed as follows :

$$\begin{array}{lcll} \mu & = & 50 \ \text{[pps]} & \text{(packets per seconds)} \\ R & = & 25 \ \text{[pps]} & \text{(packets per seconds)} \\ \sigma & = & 10 \ \text{[p]} & \text{(packets)} \end{array}$$

In order to observe the shaping action of the TBF, the input flow rate $u(t)$ is set to a constant value greater than R, namely

$$u(t) = 50\text{[pps]} > R = 25\text{[pps]}$$

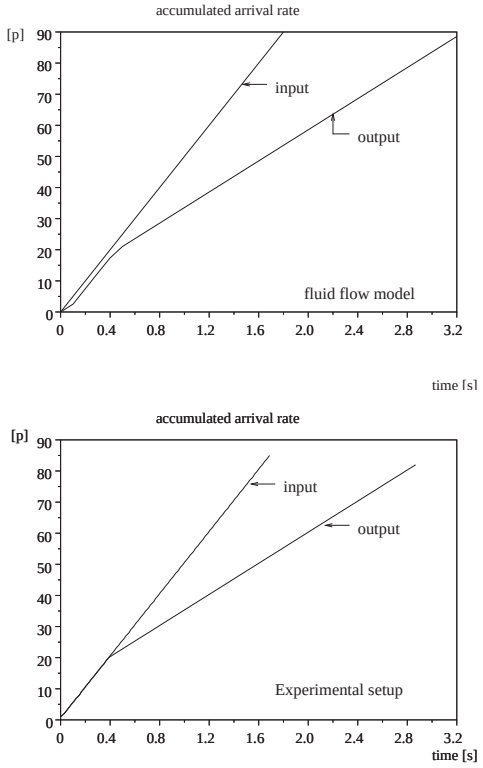

Figure 4: Fluid flow simulation and experimental result showing the typical shaping action of a token bucket filter.The dashed line is the output of TBF. The two graphics are identical.

The actual traffic is made of ICMP echo packets of 1024 bytes sent every 0.02 [s], yielding a rate of 50 [pps]. The Linux implementation of the TBF uses bytes counting but the packets being of equal size, the result are here presented in packets per second to make the comparison with the fluid flow model easier.

The result is shown in Fig. 4 where the cumulated number of packets transmitted on the time interval $[0, t]$ are presented. It can be seen that the model and the experiment give very similar results. After a time t_l that can be approximated by the following theoretical formula,

$$t_l = \frac{\sigma}{\mu - R} = 0,4[s]$$

the token bucket is empty and the output rate $v(t)$ is limited to 25 [pps]. This result is a clear experimental validation of the fluid model (4) of the TBF.

7.2 The token leaky buffer with feedback

Fig. 5 shows the experimental setup used to illustrate the properties of the token bucket with feedback. The source is configured with a token leaky bucket with feedback while the router is configured with a classical token leaky bucket with the following parameters :

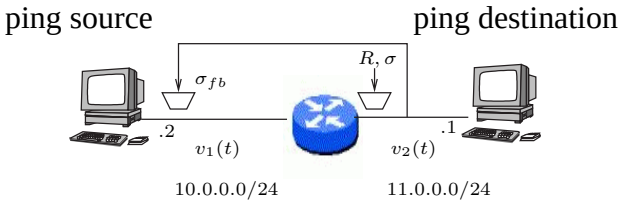

Figure 5: Experimental setup

$$\begin{array}{rcll} R & = & 25 & \text{[pps]} \quad \text{(packets per second)} \\ \sigma & = & 10 & \text{[p]} \quad \text{(packets)} \end{array}$$

In order to implement the feedback loop, A new protocol has been registered in the Linux kernel which is used to transmit the feedback information. As can be seen in Fig. 6 (see [6] for details about the trace format), the payload of this protocol is a single 4 bytes field holding a long integer carrying the following value (see Section 6) :

$$\int_{(k-1)\Delta}^{k\Delta} v_2(\xi)\, d\xi$$

The remaining part of the packet is filled with random data to reach the minimal Ethernet frame size.

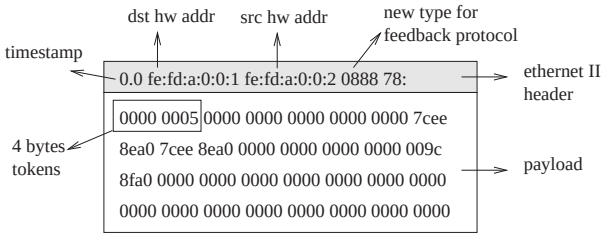

Figure 6: Sniffer trace showing a packet used in our experiment to carry the feedback information (five tokens are sent back).

Such a packet is sent every $\Delta = 200$[ms] in an Ethernet frame with type $0x0888$. The size of the feedback bucket is set to $\sigma_{fb} = 15$.

The router is in charge of updating this value as it forwards traffic on the network. Upon reception of such a frame, the source extracts the four bytes long integer (in the example of Fig. 6, the number five)and adds its value to the amount of tokens present in its bucket. Apart from this modification, the behavior of the token leaky bucket is left unchanged.

The source tries to emit a constant ICMP (echo request) stream at a rate of 50 [pps] but this transmission rate is modulated (controlled) by the amount of tokens present in the bucket. The feedback packets from the router to the source come as an overhead traffic.

As in the previous experiment, the router will limits its output rate $v_2(t)$ to 25 [pps] after 0.4 [s] but in contrast with the case

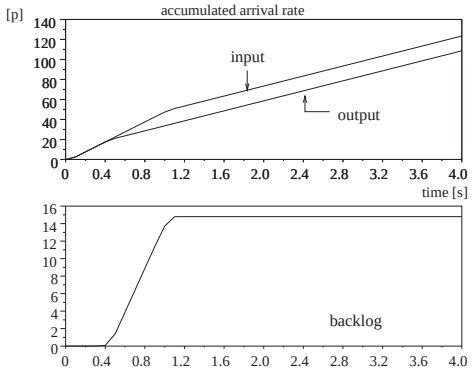

Figure 7: Fluid flow simulation showing at the top, the accumulated flow $V_1(t)$ and $V_2(t)$, at the bottom, the buffer length $x_2(t)$, clearly bounded by σ_{fb}

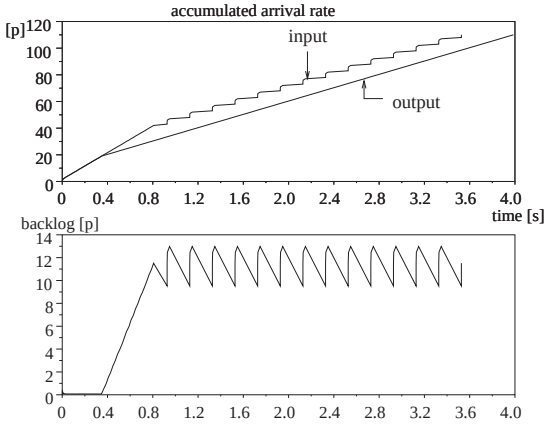

Figure 8: Experimental result showing the accumulated flow and the vertical deviation between the two curves, bounded by σ_{fb}

without feedback, the source is now adapting its sending rate to this new network condition. In effect, it can be verified in fig. 7 and 8 that the input flow v_1 is shaped so as to track the output v_2 as expressed by the inequality (9), ensuring the boundedness of the buffer queue length to $\sigma_{fb} = 15$.

7.3 A more complex setup

The hop-by-hop feedback strategy is now confronted with a more realistic topology made of five nodes, 2 sources and one destination, shown in Fig. 9. Source 2 is trying to emit at a constant rate of 50 [pps] during the time interval [0,14]. Source 1 will emit two small constant bursts of 10 [pps] during the two time intervals [8,10] and [18,24]. The bottleneck is realized with a classical token leaky bucket placed on .17 which is configured with the following parameters :

$$\begin{array}{lcll} \sigma_{tb} & = & 10 & [\text{p}] \\ R_{sust} & = & 25 & [\text{pps}] \\ \text{burst\_rate} & = & 50 & [\text{pps}] \end{array}$$

While this bucket is not empty, the burst rate is limited to 50 [pps]. The time needed to empty this bucket is therefore about 0.5 [s]. The parameters used to implement the token leaky buffer with feedback are :

$$\begin{array}{lcll} \sigma_{fb} & = & 15 & [\text{p}] \\ \Delta & = & 0.2 & [\text{s}] \end{array}$$

The results are displayed in Fig. 10 where the bandwidth limitation of the tbf placed on .17 can be readily seen on probe .18. After half a second, the throughput of the system is limited to 25 [pps] which causes the buckets placed on .13 and later, .1 to empty themselves and limit the throughput of their own link to 25 [pps] (See probe .2 and .14).

By looking successively at probe .18, .14 and .2, it can be seen that the initial burst becomes larger and larger. The burst seen on .18 is limited by σ_{tb}, the burst seen on .14 by $\sigma_{tb} + \sigma_{fb}$ and the burst on .2 by $\sigma_{tb} + 2\sigma_{fb}$. The output of source 2 is only reduced when the full buffer capacity of the network path is exhausted.

At time t=4 [s], Source 1 starts to transmit at a constant rate of 10 [pps](See probe .6). The TBFFB placed on .17 is now sharing its tokens between .9 and .13. The throughput of source 1 is automatically reduced to a value close to 15 [pps]. This result shows the ability of the feedback system to share the limited resources of a bottleneck link between concurrent sources. This point is the object of ongoing research and is still to be formally expressed.

The last small burst of source 1 shows that the feedback system is completely transparent in the absence of a bottleneck.

8 Conclusion

A fluid flow model that accurately describes the behavior of typical queueing systems such as the token leaky buffer has been presented. It has been shown that it is possible to derive mathematical properties from this model. In particular, a feedback control scheme that guarantees the boundedness of buffer queue length has been analyzed. Finally, the fluid flow model has been confronted with a real implementation and the properties of the token bucket with feedback have been validated.

References

[1] B. Braden, Clark, J. Crowcroft, et al. *RFC2309 - Recommendations on Queue Management and Congestion Avoidance in the Internet*, April 1988.

[2] C. Chang. Stability, queue length, and delay of deterministic and stochastic queueing networks. *IEEE Transactions on Automatic Control*, 39:913–931, May 1994.

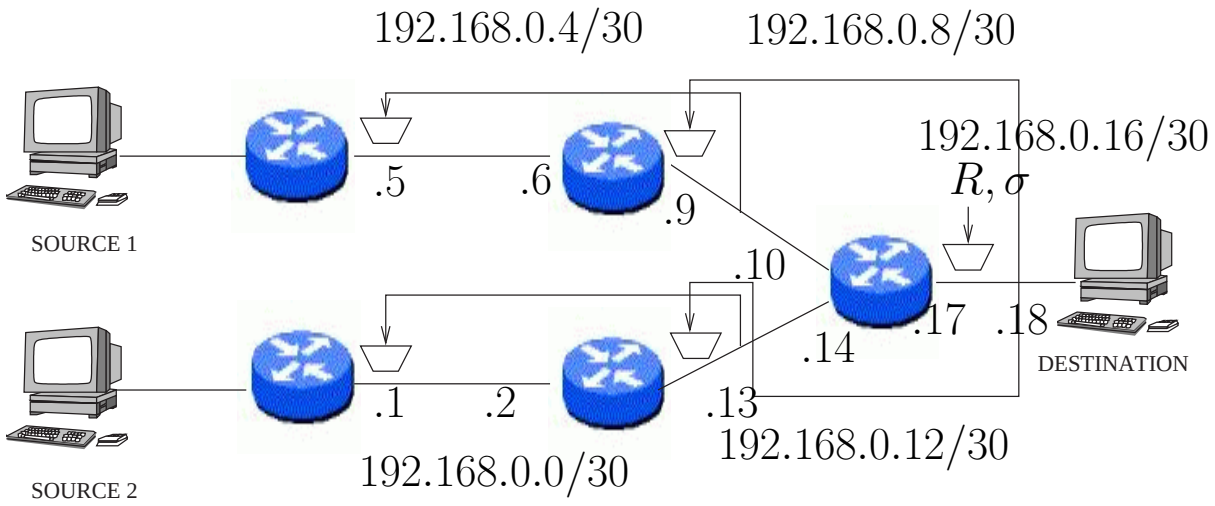

Figure 9: Setup used to test the feedback buffer on a more complex topology

[3] R. L. Cruz. A calculus for network delay, part i: Network element in isolation. *IEEE transactions on information Theory, vol.37, NO. 1*, January 1991.

[4] J. Dike. User mode linux kernel home page. http://user-mode-linux.sourceforge.net/.

[5] V. Jacobson. Congestion avoidance and control. *ACM Computer Communication Review; Proceedings of the Sigcomm'88 Symposium in Stanford, CA, 1988*, 18:314–329, 1988.

[6] V. Jacobson, C. Leres C., and S. McCanne. tcpdump/lipcap home page. http://www.tcpdump.org/.

[7] L. Kleinrock. The latency/bandwidth tradeoff in gigabit networks. *IEEE com. mag.*, 30:36–40, 1992.

[8] H. Mounier, G. Bastin, and V. Guffens. Compartmental modelling for traffic control in communication networks. *submitted to IEEE transactions on automatic control*, 2002.

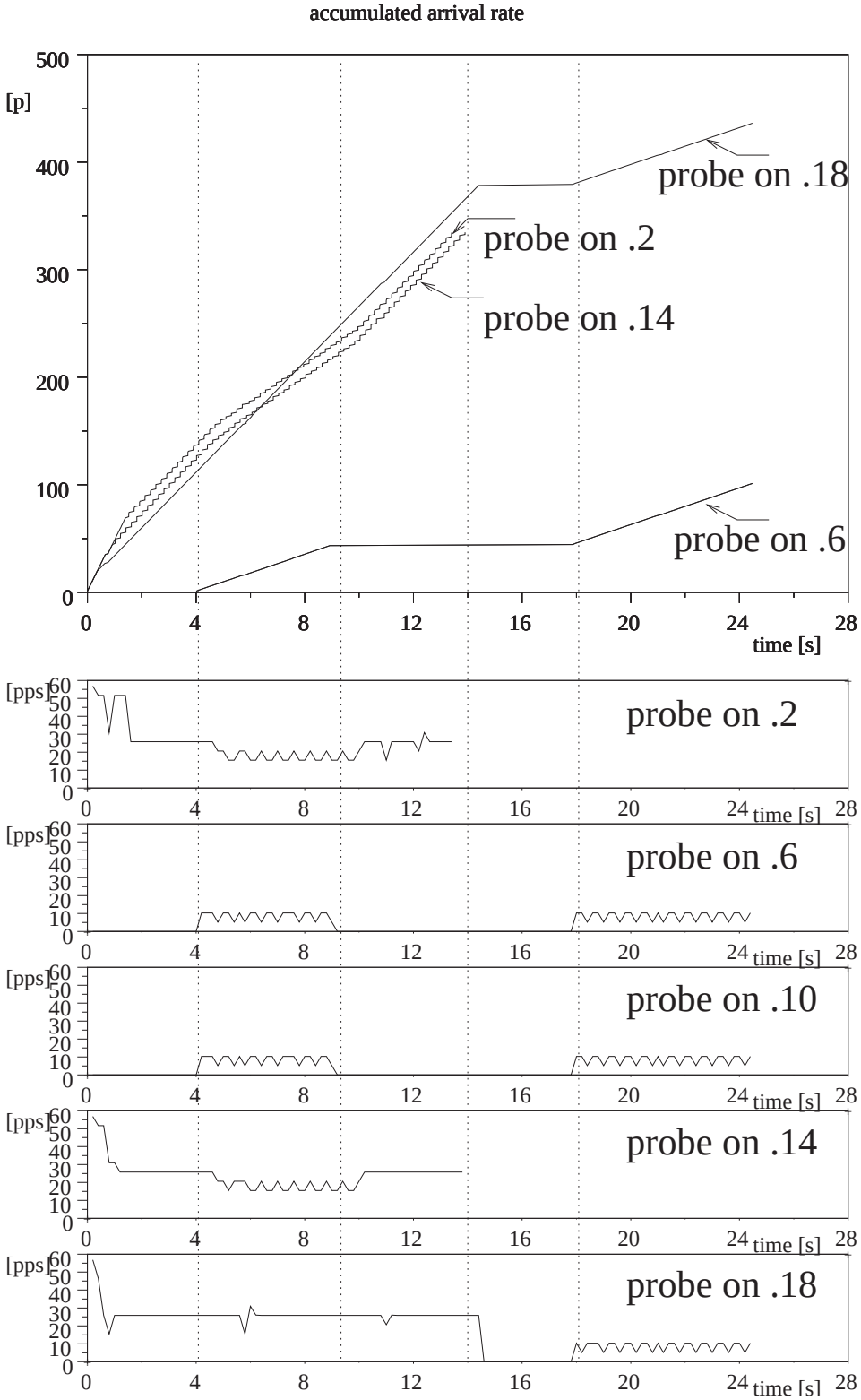

Figure 10: The rates measured by the different probes show the shaping action of the token leaky buffer with feedback that guarantees the boundedness of the buffer queues.