“Pay bursts only once” does not hold for non-FIFO Guaranteed Rate nodes

Gianluca Rizzo and Jean-Yves Le Boudec[*]

January 13, 2005

Abstract

We consider end-to-end delay bounds in a network of Guaranteed Rate (GR) nodes. We demonstrate that, contrary to what is generally believed, the existing end-to-end delay bounds apply only to GR nodes that are FIFO per flow. We show this by exhibiting a counter-example. Then we show that the proof of the existing bounds has a subtle, but important, dependency on the FIFO assumption, which was never noticed before. Finally, we give a tight delay bound that is valid in the non-FIFO case; it is noticeably higher that the existing one. In particular, the phenomenon known as “pay bursts only once” does not apply to non-FIFO nodes. These findings are important in the context of differentiated services. Indeed the existing bounds have been applied to cases where a flow (in the sense of the GR definition) is an aggregate of end-user microflows, and it is not generally true that a router is FIFO per aggregate; thus the GR node model of a differentiated services router cannot always be assumed to be FIFO per flow.

1 Introduction

In the differentiated services framework [3], end-to-end delay bounds may be obtained by assuming that sources satisfy leaky bucket [13] traffic specifications, and that routers can be modelled as *Guaranteed Rate* (GR) [9, 10] nodes. One of the main properties of a network of GR nodes is that a tight upper bound on end-to-end delay can be obtained, given the parameters of the leaky buckets at the source (burstiness and sustainable rate), and the parameters of the traversed GR nodes (delay and service rate). This end-to-end bound, derived in the original paper [9], is recalled in Equation (2); it has the remarkable property known as “pay bursts only once” [13], i.e. when a

[*] Author Affiliations: Gianluca Rizzo and Jean-Yves Le Boudec are with EPFL, CH-1015 Lausanne, Switzerland, `gianluca.rizzo@epfl.ch, jean-yves.leboudec@epfl.ch`

bursty flow traverses a number of GR nodes in sequence, the effect of the burstiness of the flow on the end-to-end delay bound is the same as if the flow traversed only one node. Another way to look at this property is that the end-to-end delay bound is much less than the sum of delay bounds at each node [12, 7].

Many scheduling disciplines have been shown to belong to the GR node model. Among these we have: Virtual Clock [11], Packet-by-Packet Generalized Processor Sharing (PGPS) [14], Self Clocked Fair Queuing [8], Bin Sort fair Queuing [6], and Leap Forward Virtual Clock [15].

The GR node model may be used in the differentiated services framework as follows [13]. End-user flows (called "microflows") are grouped into "aggregates" at the network edge; inside the network, each aggregate is handled as an individual flow, in other words, the "flow" that a GR node sees inside the network is in fact an aggregate. In practice, although this procedure usually preserves the ordering of packets within each microflow (in order to preserve sequence at the TCP or RTP layer), packet reordering can take place inside an aggregate between packets belonging to different microflows. In routers with multistage fabrics, this reordering is due to the presence of multiple parallel paths between input and output ports [5, 1]. Thus the GR node model of a differentiated services router cannot always be assumed to be FIFO per flow. More generally, the GR class encompasses a great variety of algorithms, which are not necessarily FIFO per flow [4]. In this paper we use the term "FIFO" to indicate a GR node that is FIFO per flow (since the definition of GR node is relative to the treatment it gives to a flow viewed as a single entity).

We address an issue that arises from the application of the end-to-end delay bounds in [9] when the GR nodes are not FIFO. In the original definition of GR node in [9], there is no mention of a FIFO assumption. Therefore, the end-to-end delay bound in [9] has silently been assumed to be valid whether the GR node is FIFO or not. It has formed the basis for delay computations in networks that perform aggregate scheduling [2].

However, and this is our first contribution, the end-to-end delay bound in [9] is *not* valid with non-FIFO GR nodes. We show this in Section 3.2, by exhibiting an example of a network with non-FIFO GR nodes, and which violates the delay bound. How can this happen given that the original derivation in [9] does not appear to make use of any FIFO assumption ? We have analyzed the proofs in [9], and indeed found a place where a hidden FIFO assumption is made; this assumption is subtle, but is essential and invalidates the results in [9] in the non-FIFO case (Section 3.3).

Our second contribution is an end-to-end delay bound that is valid in the general, possibly non-FIFO case (Section 4.1). The bound is valid for a network of GR nodes and for a leaky bucket constrained flow that traverses M of these nodes. We show that this bound is tight (Section 4.2). Unfortunately, the new bound has lost the attractive "pay bursts only once" property mentioned earlier: the burstiness of the flow appears in the bound with a factor M, and it is noticeably higher than the one valid for FIFO GR nodes. Our methodology to obtain this bound is based on the mapping between a GR node and a service curve element ([13], Section 2.1); we exploit the

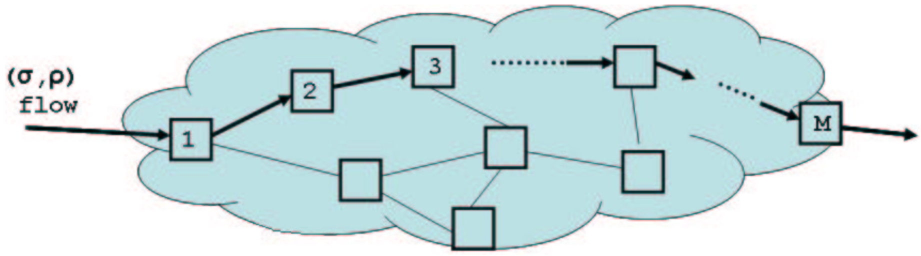

Figure 1: We consider a (σ, ρ) constrained flow which traverses a succession of M nodes in a network of GR servers. The flow can be interpreted as a differentiated services aggregate flow, in which case (σ, ρ) is the sum of the parameters of the constituent microflows.

fact that the concatenation result for service curve elements holds independently from the FIFO behavior of nodes, and we derive a bound on the burstiness increase due to traversing a (possibly non-FIFO) GR node. The burstiness increase result is also a bound of independent interest, and we show in Section 4.2 that it is tight as well.

In many practical cases of interest, it is possible to bound the non-FIFO behavior of a GR node by breaking down its latency into a fixed and a variable part. We give the improved bound for such cases in Section 4.3. We note in passing that there is a concatenation result available for FIFO GR nodes, that can be used to derive more simply the original end-to-end delay bound for FIFO GR nodes (Section 4.4).

2 Model and Assumptions

We define a *flow* as a sequence of packets travelling on a link in a network. To a data flow we associate the cumulative function $R(t)$, which counts the number of bits seen on the flow in the time interval $[0,\ t]$. A wide-sense increasing function $f(t)$ is said to be an arrival curve for a flow (which is then said to be *f(t) constrained*) with cumulative function $R(t)$ if it holds, for all $0 \leq \tau \leq t$:

$$R(t) - R(t - \tau) \leq f(\tau)$$

The arrival curve of a given flow upper bounds the number of packets of the flow that can be observed on a given time window. A $(\sigma,\ \rho)$ constrained flow is a flow whose arrival curve is of the form $f(t) = \sigma + \rho t$, where σ is the *burstiness* of the flow, and ρ its *sustainable rate*.

We consider a network of routers that can be modelled as *Guaranteed Rate* (GR) nodes [9, 10]. The definition of this model is based on the concept of *Guaranteed Rate clock value*:

Definition 2.1 *(*GR clock value *[9]) Consider a flow that is associated with a service rate r (in bit/s) at a given node. Let p^j denote the j-th packet of the flow, and l^j its length. Let $GRC(p^j)$ and $A(p^j)$ denote respectively the guaranteed rate clock value of packet p^j and its arrival time at the*

Table 1: Symbols used in formulas

$R(t)$	cumulative packet arrival function of the flow
σ	burstiness
ρ	sustainable rate
r	service rate for the flow
e^m	delay of the m-th GR node
M	nodes traversed by the flow
p^j	j-th packet of the flow
l^j	length of packet p^j
$l_{max}(l_{min})$	maximum (minimum) packet length for the flow
$GRC^m(p^j)$	GR clock value at node m for p^j (the j-th packet at the input of the m-th node)
$A^m(p^j)$	arrival time at node m of p^j (the j-th packet at the input of the m-th node)
$d^m(p^j)$	departure time from node m of p^j (the j-th packet at the input of the m-th node)
$\tau^{m,m+1}$	propagation delay between nodes m and $m+1$
α^m	$e^m + \tau^{m,m+1}$
$\beta_{r,e}$	service curve of the form $r[t-e]^+$
$f^m(t)$	arrival curve for the flow at the input to node m

node. The guaranteed rate clock value for packet p^j is given by:

$$GRC(p^j) = \begin{cases} 0 & j = 0 \\ \max\{A(p^j),\ GRC(p^{j-1})\} + \frac{l^j}{r} & j \geq 1 \end{cases}$$

The concept of GR clock value is used to define the *Guaranteed Rate* (GR) node, as follows:

Definition 2.2 (GR node *[9]) Consider a node that serves a flow. Packets are numbered in order of arrival. The node is a Guaranteed Rate node for the flow, with rate r and delay e, if it guarantees that packet p^j of the flow is transmitted by $GRC(p^j) + e$, where e depends on the scheduling algorithm and the server.*

Many practical implementations of the GPS scheduling algorithm, such as Virtual Clock scheduling, Packet-by-Packet Generalized Processor Sharing scheduling, and Self Clocked Fair Queuing have been shown [9] to belong to the GR category of servers.

We consider a flow that traverses a network of GR nodes, which are not necessarily FIFO (a node is FIFO when, for each flow the sequence of packets at the output of the node is identical to the sequence of packets at the input of the node).

We assume that the given flow traverses a succession of M nodes in the network. As they traverse the network, packets belonging to the flow experience a delay that accumulates along their path, and that can be different in principle at each node for each packet.

We assume that the arrival time of a packet at a node is the arrival time of the last bit of the packet, and the departure time of a packet from a node is the departure time of the last bit of the packet. This leads us to observe instantaneous packet arrivals and departures. In what follows we consider that each link between two nodes m and $m+1$ has a constant propagation delay $\tau^{m,m+1}$, and with $\tau^{M,M+1}$ we indicate the propagation delay of the link between the M-th node and the destination.

Finally, we consider that all the nodes traversed by the flow are stable. A node along the path of the flow is stable if $\rho \leq r$, where ρ is the sustainable rate of the arrival curve of the flow (at its source) and r is the reserved rate for the flow at the node [13].

3 The existing end-to-end delay bounds in GR nodes require FIFO assumption

3.1 The existing results

The main result about end-to-end delay bounds in a network of (not necessarily FIFO) GR servers which is presently available has been first derived in [9], and extended in [10]. In those papers a method is defined to derive an end-to-end delay bound, based on the following result:

Theorem 3.1 ([9]) *Consider a flow that traverses a succession of M nodes in a network. If the scheduling algorithm at each server $m \in \ (1,\ M)$ on the path of the flow belongs to GR for the given flow, with service rate r for the flow and delay e^m, then a bound to the end-to-end delay of the j-th packet of the flow, denoted with D^j, is given by*

$$D^j \leq GRC^1(p^j) - A^1(p^j) + (M-1) \max_{n \in \ 1,\ldots,j} \frac{l^n}{r} + \sum_{m=1}^{M} \alpha^m \tag{1}$$

where $\alpha^m = e^m + \tau^{m,m+1}$, and $\tau^{m,m+1}$ is the propagation delay between nodes m and $m+1$.

The difference $GRC^1(p^j) - A^1(p^j)$ in Equation (1) depends on source traffic specification. For a (σ, ρ) constrained flow, Equation (1) takes the form [9]

$$D^j \leq \frac{\sigma}{r} + (M-1) \max_{n \in \ 1,\ldots,j} \frac{l^n}{r} + \sum_{m=1}^{M} \alpha^m \tag{2}$$

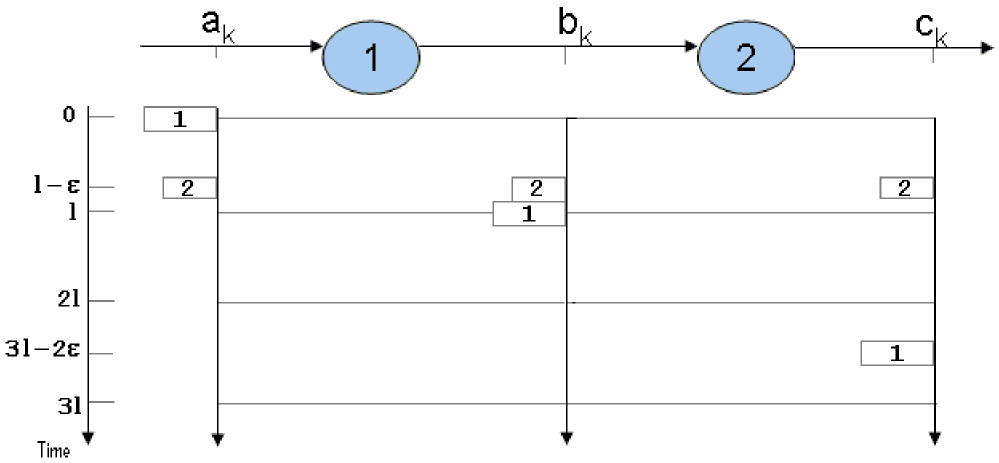

Figure 2: Example of non-FIFO behavior of two GR nodes, traversed by a (σ, ρ) constrained flow, with $\sigma = l, \ \rho = r = 1$, and where the propagation delay at all links is zero. a_k and b_k are respectively the arrival times of packets at node 1 and 2, and c_k are the departure times of packets from node 2. At all nodes only packet 1 takes its maximum possible delay at the node (equal to its GR clock value at the node), whereas all other packets get no delay from the nodes. The end-to-end delay of packet 1 is of $3l - 2\varepsilon$ time units: if $\varepsilon < \frac{l}{2}$, packet 1 gets a larger delay than the delay bound in [9], which is of $2l$ time units.

3.2 Counter example

However, from the analysis of even simple examples of non-FIFO behavior in GR nodes, we can verify that in a network of non-FIFO GR nodes the end-to-end delay for a packet can actually be higher than the bounds in Equation (1) and Equation (2).

As an example, we consider a sequence of two packets, belonging to a (σ, ρ) constrained flow that traverses two GR non-FIFO nodes (see Figure 3.2). To simplify the example, we took the propagation delay at all links equal to zero, $e = 0$ for both nodes, $\sigma = l$ bits, $\rho = r = 1$. We assume that packet 1 is of length l, and that packet 2 is of length $l - \varepsilon$. At the input to node 1, packet 1 arrives at time $t = 0$, packet 2 at time $l - \varepsilon$.

As there is no delay at the nodes, the maximum departure time for packet 1 is given by its GR clock value at node 1, equal to l time units: then we can assume that packet 1 leaves the first node at time $t = l$, and that all other packets get no delay at the node, so that their departure time equals their arrival time at the node. In this way at the input to node 2 we have that packet 2, arrived at time $l - \varepsilon$, precedes packet 1.

At node 2, we assume again that the departure time of packet 1 equals its maximum departure time (its GR clock value at the node) so that, as the GR clock value for packet 2 is $t = 2(l - \varepsilon)$, the GR clock value for packet 1 is $3l - 2\varepsilon$, and this is also the end-to-end delay for this packet.

The end-to-end delay bound from Equation (2) in this case would instead be of $2l$ time units, so that if $\varepsilon < \frac{l}{2}$ the delay of packet 1 in the example is larger than the delay bound in Equation (2). This simple example shows that the existing end-to-end delay bounds in a network of GR nodes are not valid for non-FIFO nodes.

3.3 The hidden FIFO assumption in [9]

In this section, we analyze the original derivation of the end-to-end delay bounds of Equation (1) and Equation (2) from [9], in order to put in evidence that those results are actually valid only for FIFO GR nodes.

The hidden assumption is in the proof of the following Lemma in [9], used to derive the bounds in Equation (1) and Equation (2):

Lemma 3.1 ([9]) *If the scheduling algorithm at servers m and $m+1$ along the path of the flow belongs to GR for that flow, then*

$$GRC^{m+1}(p^j) \leq GRC^m(p^j) + \max_{k\in[1,..,j]} \frac{l^k}{r} + \alpha^m, \;\; j \geq 1 \tag{3}$$

where p^j is the j-th packet of the flow, l^k is the length of the k-th packet of the flow, r is the guaranteed rate for the flow at nodes m and $m+1$.

The hidden FIFO assumption lies in the following inequality, between equations (23) and (24) of the proof of the lemma:

$$GRC^m(p^{j+1}) \geq GRC^m(p^j) + \frac{l^{j+1}}{r} \tag{4}$$

The hidden assumption is in the use of packet indices at two consecutive nodes m and $m+1$. In the proof, index j refers to the succession of packet arrivals at node $m+1$: indeed, from inequality (4) (which has been derived from the GR clock value definition at node m) we see that the same packet index j is used for the succession of packets at the input to node m. This implies that no packet reordering takes place at node m, and that node m is assumed to be FIFO for the flow.

Indeed, if we look at the example in Figure 3.2, we can clearly see that, if node 1 is non-FIFO, then Equation (4) is false for packet 1.

This can be shown by comparing the GR clock values of packet 1 at nodes 1 and 2. At node 1, the GR clock value for packet 1 is l. At node 2, the GR clock value for packet 2 (which is the first to arrive at the node, at time $l-\varepsilon$) is $2(l-\varepsilon)$, and the GR clock value for packet 1, arrived at the node at time l, is $3l-2\varepsilon$. Now, for packet 1 Equation (4) translates into the following inequality:

$$3l - 2\varepsilon \leq 2l \tag{5}$$

As we saw in the example, if $\varepsilon < \frac{l}{2}$ then $3l - 2\varepsilon > 2l$, and Equation (4) does not hold in this case. As in the non-FIFO case the lemma in [9] does not hold, the whole proofs of the delay bounds in Equation (1) and Equation (2) in [9], which rely on that lemma, are not valid in the non-FIFO case.

4 An end-to-end delay bound valid in the non-FIFO case

4.1 The delay bound

As we showed in the previous section, when nodes are not FIFO, as Theorem 3.1 and Equation (2) cannot hold, the question that arises is whether the end-to end delay in a network of generic, non-FIFO GR nodes is bounded, and which is the expression of the bound in this case. To our knowledge the issue is still open, as no result in the present literature addresses it in an exhaustive way. We answer to this with the following theorem, which is one of the main contributions of the present paper:

Theorem 4.1 *Consider a (σ, ρ) constrained flow that traverses a sequence of M GR nodes. If all nodes reserve the same service rate r to the flow, and if all nodes are stable, an end-to-end delay bound for a packet belonging to the flow is given by*

$$d = M\frac{\sigma}{r} + \frac{\rho}{r}\frac{l_{max}}{r}\frac{M(M-1)}{2} + \frac{\rho}{r}\sum_{m=1}^{M}\sum_{i=1}^{m-1} e^i + \sum_{m=1}^{M}\left(e^m + \tau^{m,m+1}\right) \qquad (6)$$

where l_{max} is the maximum packet length for the flow. Moreover, if we denote with σ_M the burstiness of the arrival curve for the flow at the output of node M, we have that

$$\sigma_M = \sigma + M\rho\frac{l_{max}}{r} + \rho\sum_{m=1}^{M} e^m \qquad (7)$$

By comparing the bound in Equation (6) with the one that can be obtained when we know that nodes are FIFO we can clearly see how, in the non-FIFO case, the contribution to the end-to-end delay which is due to the burstiness σ of the initial flow is multiplied by a factor M. Hence, in the non-FIFO case, the non-validity of the concatenation result for the computation of an end-to-end delay bound brings to "pay" burst M times, instead of only once [13].

We can also observe that another consequence of the non-FIFO behavior of GR nodes is an increment of the burstiness of the flow at the output of the last node by the quantity $\rho\frac{l_{max}}{r}$, with respect to the FIFO case.

Proof. (of Theorem 4.1) We first observe that the hypothesis of node stability implies that $\rho \leq r$ at all the M nodes. As GR nodes are not necessarily FIFO, for the end-to-end delay computation we exploit some properties of GR nodes that do not depend on their FIFO behavior. Among the Network Calculus results still valid in the non-FIFO case, we have the following:

Theorem 4.2 *(*Equivalence with service curve [13]*) A GR node with rate r and latency e, with L-packetized input, is the concatenation of a service curve element, with service curve equal to the rate-latency function $\beta_{r,\,e}$, and an L-packetizer. If the GR node is FIFO, then so is the service curve element.*

An important implication of the preceding theorem is the following corollary:

Corollary 4.1 ([13]) *A GR node (with rate r and latency e) offers a minimum service curve $\beta_{r,\,e+\frac{l_{max}}{r}}$.*

As the equivalence between a GR node and a rate-latency service curve element holds also for non-FIFO nodes, a sequence of M GR nodes can still be studied as the concatenation of service curve elements, each one of the form $\beta_{r,\,e+\frac{l_{max}}{r}}$. The link between two nodes on the path of the flow can be modelled as a FIFO constant delay element, with a minimum service curve of the form $\delta_{\tau^{m,m+1}}$, and a maximum service curve with the same expression [13].

As a consequence of the equivalence between GR nodes and service curve elements, in order to derive a delay bound at each of the M non-FIFO GR nodes we can exploit the following result [13]:

Theorem 4.3 *(*(Delay Bound) [13]*) For a flow with an arrival curve $f(t)$, served in a (possibly non-FIFO) GR node with rate r and latency e, the delay for any packet at the node is bounded by*

$$\sup_{t>0}\left[\frac{f(t)}{r} - t\right] + e$$

If we consider a (σ, ρ) constrained flow that traverses a sequence of M GR nodes, the sequence of GR servers offers to it a minimum service curve given by the min-plus convolution between the service curves of all the GR nodes and links in the sequence, and a maximum service curve given by the min-plus convolution between the maximum service curves of all links.

Proposition 4.1 ([13]) *Consider a flow that traverses a sequence of service curve elements in a network. In order to compute an output bound for the flow, fixed delay elements on the path of the flow can be ignored.*

As a consequence, if we indicate with $f^{m+1}(t)$ an arrival curve of the flow at the input to the $m+1$-th node in the sequence ($m \in [1, M]$), using the properties of the deconvolution operator [13] we have that

$$f^{m+1}(t) = \left[(\sigma + \rho t) \otimes \delta_{\sum_{i=0}^{m}\tau^{i,i+1}}\right] \oslash \beta_{r,\,\sum_{i=0}^{m}\left[\frac{l_{max}}{r}+e^{i}+\tau^{i,i+1}\right]} =$$

$$(\sigma + \rho t) \oslash \beta_{r,\,\sum_{i=0}^{m}\left[\frac{l_{max}}{r}+e^{i}\right]} =$$

$$= \sigma + \rho \left(t + \sum_{i=0}^{m} \left[\frac{l_{max}}{r} + e^i \right] \right) \tag{8}$$

where $\otimes$ is the convolution operator, $\oslash$ is the deconvolution operator [13], and $\beta_{r, \sum_{i=0}^{m}\left[\frac{l_{max}}{r}+e^i+\tau^{i,i+1}\right]}$ is the service curve of the concatenation of nodes $1, ..., m$ and of the links between them. From Equation (8), if we indicate with $f^{M+1}(t)$ the arrival curve at the output of node M, we can observe that its burstiness has the expression in Equation (7).

Using Theorem 4.3, a delay bound at the m-th node along the succession of the M GR nodes is given by

$$d_m = \frac{\sigma}{r} + (m-1)\frac{\rho}{r}\frac{l_{max}}{r} + \frac{\rho}{r}\sum_{i=1}^{m-1} e^i + e^m \tag{9}$$

and a delay bound for the concatenation of the m-th node and the link between nodes m and $m+1$ is given by $d_m + \tau^{m,m+1}$.

An end-to-end delay bound for the packets of the flow is obtained by summing the delay bounds in Equation (9) at each node along the path of the flow, and taking into account the propagation delays at all links:

$$d = \sum_{m=1}^{M} (d_m + \tau^{m,m+1}) =$$

$$= M\frac{\sigma}{r} + \frac{\rho}{r}\frac{l_{max}}{r}\sum_{m=1}^{M}(m-1) + \frac{\rho}{r}\sum_{m=1}^{M}\sum_{i=1}^{m-1} e^i + \sum_{m=1}^{M}\left(e^m + \tau^{m,m+1}\right) =$$

$$= M\frac{\sigma}{r} + \frac{\rho}{r}\frac{l_{max}}{r}\frac{M(M-1)}{2} + \frac{\rho}{r}\sum_{m=1}^{M}\sum_{i=1}^{m-1} e^i + \sum_{m=1}^{M}\left(e^m + \tau^{m,m+1}\right)$$

□

4.2 The delay bound in the non-FIFO case is tight

Theorem 4.4 *With the same assumptions as in Theorem 4.1, the bounds in Equation (6) and Equation (7) are tight. More precisely, we can always define a succession of packets and a series of scheduling behaviors of the chain of GR nodes such that the burstiness of the flow at the output of the M-th node achieves the bound in Equation (7), and that at least one packet from the given flow experiences an end-to-end delay equal to the bound in Equation (6).*

Proof. The proof of Theorem 4.4 is by example: let's take a (σ, ρ) constrained flow, that traverses a sequence of M non-FIFO GR nodes, all with the same delay e and the same service rate r for the flow.

We assume for simplicity that $e = k$ time units, with $k \in \mathbb{N}$, that $\forall m,\ \tau^{m,m+1} = \tau = \frac{l}{r}$ and we take $\sigma = nl > (k+1)l$. In order to simplify the notation, we assume that $\rho = r$, and that all packets are of the same length l.

The example can be built as follows:

The sequence of packets: we consider the following sequence of packets, at the input to the first of the M nodes:

- at $t = 0$, we have the arrival of a burst of dimension $\sigma = nl$;
- then, with a period $P(i) = \frac{\sigma}{r} + (i-1)\frac{l}{r} + ik\frac{l}{r},\ i \geq 1$, we have the arrival of a burst of dimension $\sigma = nl$. The arrival time of the i-th burst at the first of the M nodes is given by

$$t_i = \sum_{j=1}^{i} P(j) =$$

$$= i\frac{\sigma}{r} + \frac{l}{r}\sum_{j=1}^{i}(j-1) + k\frac{l}{r}\sum_{j=1}^{i} j$$

- For $i \geq 1$ we define the time instants t_i^* as

$$t_i^* = t_i - (i-1)\frac{l}{r} - ik\frac{l}{r}$$

  Then we assume that in the time intervals $[t_i^*,\ t_i),\ i \geq 1$ we have a packet arrival at time t_i^* and then the arrival of a packet each $\frac{l}{r}$ time units, so that a total of $i - 1 + ik$ packets arrive in each interval $[t_i^*,\ t_i)$.

We can verify that such a succession of packets is (σ, ρ) constrained. On Figure 3 we have an example of a succession of packets with these characteristics, with $\sigma = 4l$, $e = 2\frac{l}{r}$ and $\frac{l}{r} = 1$ time unit.

The scheduling behavior: given the initial burst of the sequence, of dimension $\sigma = nl$, which arrives at the first of the M nodes at time 0, we consider one of the packets that compose it, and we indicate it with p_n (in order to distinguish it from p^n, the n-th packet to get into a given node).

We assume that, at the input to the m-th node along the path of the flow:

- all packets that precede packet p_n (p_n included) at the input of the m-th node get the *maximum* delay at the node;
- if p_n is part of a burst of packets, arrived at a node in the same time instant as p_n, it is always the last to be served (non-FIFO behaviour);
- all packets p^j ([1]) that get into the m-th node after packet p_n, and in time intervals $[t_i^* + (m-1)\tau,\ t_i + (m-1)\tau),\ i \geq 1$, get a delay equal to $e + \frac{l}{r}$, but do not get out of the node after time $t_i + (m-1)\tau$. That is, their departure time is

$$d^m(p^j) = \min\left\{A(p^j) + e + \frac{l}{r},\ t_i + (m-1)\tau\right\}$$

- all packets that get into the m-th node after packet p_n, and at time instants $t_i + (m-1)\tau,\ i \geq 1$, get the *minimum* delay for that node.

At each GR node, the maximum departure time for each packet is equal to the sum of its Guaranteed Rate clock value and of the delay of the node, whereas its minimum departure time is equal to its arrival time at the node. As a consequence, given the structure of the sequence of packets, for all packets that precede or arrive at the same time as packet p_n at the input of a node (packet p_n included), we observe at the output, starting from the first packet, one packet departure each $\frac{l}{r}$ time units.

[1] We underline here that packet indices refer to the succession of packets at the input to a specific node, so that in general different packet indices are to be used for packets at the input to each node.

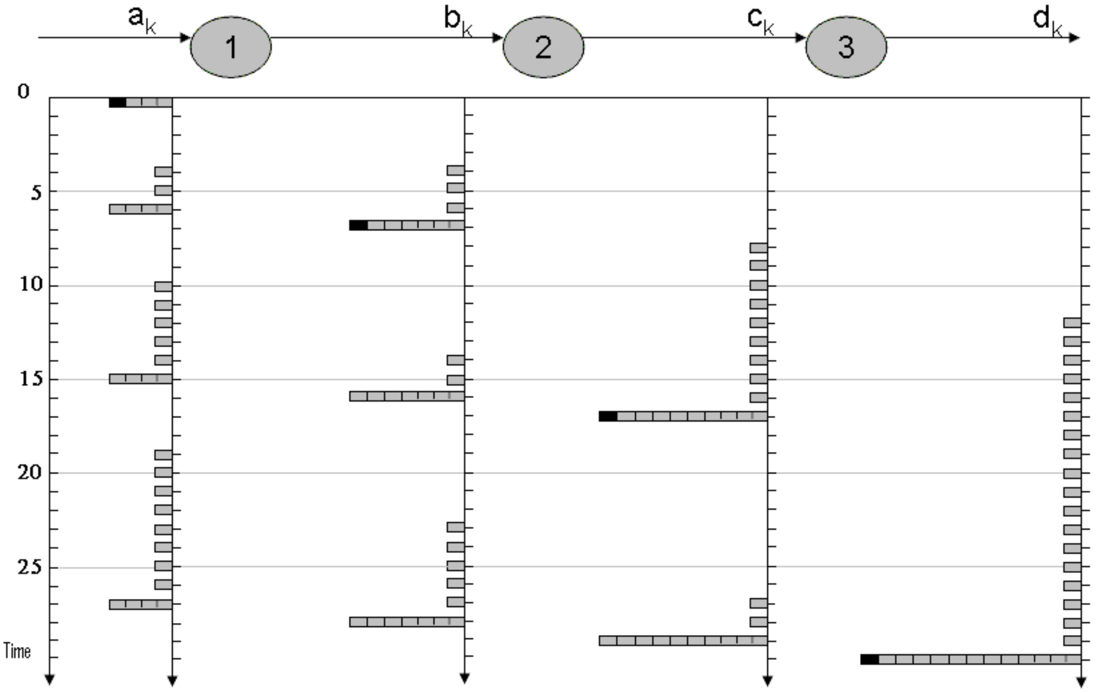

Figure 3: Evolution of a sequence of packets, at the input to each of three GR nodes on its path, and at the output of node 3. At the input to node 1, there is a sequence with the characteristics described in the proof of Theorem 4.4, and which is (σ, ρ) constrained, with $\sigma = 4l$, $e = 2\frac{l}{r}$ time units, $\frac{l}{r} = 1$ time unit. As the propagation delay at all links is of 1 time unit, the delay experimented by the packet marked in black at nodes 1, 2 and 3 (taking into account propagation delay of the link at the output of each node) is respectively of 7, 10 and 13 time units (for an end-to-end delay of 30 time units), and the burstiness of the output flow at each node is respectively of $7l$, $10l$ and $13l$, as predicted by Theorem 4.4.

In order to demonstrate the tightness of the bounds in Equation (6) and Equation (7), we use induction on the index m of the succession of the M nodes on the path of the flow.

Base case: $m = 1$, the first node of the path. The departure time of the first packet to be served at node 1 is

$$d^1(p^1) = A^1(p^1) + \frac{l}{r} + e = (k+1)\frac{l}{r}$$

since we assumed that the arrival time at node 1 is at $t = 0$. As all the n packets that arrived at $t = 0$ leave the node with the maximum delay, starting from time $t = d^1(p^1)$ we observe at the output of node 1 one packet leaving the node each $\frac{l}{r}$ time units, up to packet p_n, which then leaves the node at time

$$d^1(p_n) = A^1(p^1) + n\frac{l}{r} + e = \frac{\sigma}{r} + e$$

Taking into account propagation delay on the link between nodes 1 and 2, the delay of packet p_n at the input to node 2 is equal to $d^1(p_n) + \tau$, and in this case the delay bound in Equation (6) is verified.

In the time interval $[\frac{\sigma}{r},\ \frac{\sigma}{r} + k\frac{l}{r})$ at the input to node 1, we have one packet arrival each $\frac{l}{r}$ time units and a total of k packet arrivals in the whole interval. The first of this packets arrives at node 1 at time $t = \frac{\sigma}{r}$, and it leaves the node at time $t' = \frac{\sigma}{r} + k\frac{l}{r}$. This implies that all packets arrived in that time interval leave the node at time t'. Then, at time t' at the input of the node we have the arrival of a burst of dimension $\sigma = nl$, that is not delayed by node 1. As $d^1(p_n) = t'$, we have that at time t' at the output of the node the flow has a burst of dimension $(n + k + 1)l$, achieving the burstiness bound in Equation (7).

Iterative step: The inductive hypothesis is that Equation (6) and Equation (7) hold for the sequence of nodes from 1 to m. We want to demonstrate that they hold also for the sequence of nodes from 1 to $m + 1$ ($m + 1 \leq M$).

By the inductive hypothesis, the time at which packet p_n arrives at node m can be obtained by Equation (6):

$$A^m(p_n) = (m-1)\frac{\sigma}{r} + \frac{l}{r}\sum_{j=1}^{m-1}(j-1) + k\frac{l}{r}\sum_{j=1}^{m-1} j + (m-1)\tau$$

and the time at which it leaves node m is given by

$$d^{m+1}(p_n) = m\frac{\sigma}{r} + \frac{l}{r}\sum_{j=1}^{m}(j-1) + k\frac{l}{r}\sum_{j=1}^{m} j + (m-1)\tau$$

Due to the structure of the sequence of packets and the scheduling behavior of nodes, after the arrival of p_n at node m we have, in the time interval $[t^*_m + (m-1)\tau,\ t_m + (m-1)\tau]$, the arrival of $n + mk + m - 1$ packets, with

$$t_m = m\frac{\sigma}{r} + \frac{l}{r}\sum_{j=1}^{m}(j-1) + k\frac{l}{r}\sum_{j=1}^{m} j$$

Then, for the scheduling behavior of the sequence of nodes, the packet that arrives at node 1 at time t^*_m (and so, the first packet to arrive at node m in the time interval $[t^*_m + (m-1)\tau,\ t_m + (m-1)\tau]$) takes by each node in the succession $1, ..., m$ a delay equal to $(k+1)\frac{l}{r}$, and by each link a delay of τ. Therefore, it leaves node m at a time $t' + (m-1)\tau$, where t' is given by

$$t' = \min\left\{t^*_m + m(k+1)\frac{l}{r},\ t_m\right\}$$

As

$$t_m - t^*_m = mk\frac{l}{r} + (m-1)\frac{l}{r}$$

then we have that $t' = t_m$, and all the $n + mk + m - 1$ packets leave node m at time $t_m + (m-1)\tau$.

As $A^{m+1}(p_n) = t_m + m\tau$, at time $t_m + m\tau$ at the input to node $m+1$ we have the arrival of $n + mk + m$ packets, and a burst of dimension $n + m(k+1)$.

In general, at the output of each node x, $x \in 1, ..., m$, at time $t_x + (x-1)\tau$ we have the departure of packet p_n and of $n + xk + x - 1$ other packets. As at all nodes p_n is always the last packet to be served among those that arrived at the same time as p_n, and as all packets served before p_n get the maximum delay, we have that

Remark 4.1 *At the input of nodes $2, ..., m+1$, the arrival at time t of packet p_n is always preceded by the arrival, at time $t - \frac{l}{r}$, of another packet.*

Another result that is important for the rest of the proof, is the following:

Lemma 4.1 *At nodes $1, ..., m+1$, for all packets p that arrive at the node before packet p_n, we have that*

$$GRC(p) = A(p) + \frac{l}{r} \tag{10}$$

Proof.(of Lemma 4.1) At the node, the GRC of the first packet that arrives, is given by

$$GRC(p^1) = A(p^1) + \frac{l}{r}$$

as no packet precedes it. The GRC of the second packet is

$$GRC(p^2) = \max\left\{A(p^2), GRC(p^1)\right\} + \frac{l}{r} =$$

$$= \max\left\{A(p^2), A(p^1)0 + \frac{l}{r}\right\} + \frac{l}{r}$$

As all packets that precede p_n at the node get their maximum delay, packet interarrival times are at least of $\frac{l}{r}$ time units. So we have that $A(p^2) \geq A(p^1) + \frac{l}{r}$, and

$$GRC(p^2) = A(p^2) + \frac{l}{r}$$

For the same reason, in general (for all packets p^j that get at a node before p_n) we have that $A(p^j) \geq A(p^{j-1}) + \frac{l}{r}$, and Equation (10) holds.
□

At node $m+1$, the GR clock value of the first packet to be served among those arrived at time $t_m + m\tau$ (that we denote with p^j) is given by:

$$GRC^{m+1}(p^j) = \max\left(t_m + m\tau,\ GRC^{m+1}(p^{j-1}) + \frac{l}{r}\right) + \frac{l}{r}$$

where p^{j-1} is the packet that precedes packet p^j at the input to node $m+1$.

Using Lemma 4.1 and Remark 4.1, we have that $t_m + m\tau = GRC^{m+1}(p^{j-1}) + \frac{l}{r}$. So we have

$$GRC^{m+1}(p^j) = t_m + m\tau + \frac{l}{r}$$

Then packet p_n, that is the last packet to be served among the $n+mk+m$ packets arrived at time $t_m + m\tau$, will have a GR clock value at node $m+1$ given by

$$GRC^{m+1}(p_n) = t_m + m\tau + (n + mk + m)\frac{l}{r}$$

and the departure time from node $m+1$ for packet p_n is

$$d^{m+1}(p_n) = GRC^{m+1}(p_n) + k\frac{l}{r} =$$

$$= t_m + m\tau + (n + (m+1)k + m)\,\frac{l}{r} =$$

$$= (m+1)\frac{\sigma}{r} + \frac{l}{r}\sum_{j=1}^{m+1}(j-1) + k\frac{l}{r}\sum_{j=1}^{m+1} j + m\tau$$

and taking into account the propagation delay of the link at the output of node $m+1$, we have that the end-to end delay for packet p_n for the succession of nodes $1, ..., m+1$ is given by $d^{m+1}(p_n) + \tau$, and it equals the end-to-end delay bound in Equation (6).

Also, with a similar procedure to the one followed at node m, we have that, at time $t_{m+1} + m\tau$ $(= d^{m+1}(p_n))$ at the output of node $m+1$ we have a burst of $n + (m+1)(k+1)$ packets, so that the flow at the output of node $m+1$ achieves the burstiness bound in Equation (7). □

4.3 A Refined Result

We now introduce a new node model, more realistic than the GR node model. Specifically, this new model is composed by a FIFO GR node, with rate r and zero delay, followed by a FIFO constant delay element with delay e_a, and by a non-FIFO variable delay element, with maximum delay e_b.

Although the GR node model does not put a lower bound to a packet delay (which can even be equal to zero), real schedulers do not have a minimal delay equal to zero: they usually introduce a minimal delay for packets. That is, the departure time of a packet p (arrived at the node at time $A(p)$) is

$$A(p) + e_a \leq d'(p) \leq GRC(p) + e_a + e_b \tag{11}$$

As an example, this happens in input buffer switches, in which the minimum delay for a packet in the node is due to the minimum time necessary for a packet to traverse the fabric: in the presented model, the fabric is modelled by the succession of the constant delay element and of the variable delay element. In this sense, the model presented captures more closely and realistically the characteristics of network nodes.
In this case, we have the following proposition:

Proposition 4.2 *It is given a GR node, with rate r and delay $e_a + e_b$, at which the departure time for a generic packet falls inside the interval $[A(p) + e_a,\ GRC(p) + e_a + e_b]$. Such a node can be modelled as the succession of a FIFO GR server, with rate r and zero delay, followed by a FIFO constant delay element with delay e_a, and by a non-FIFO variable delay element, with maximum delay e_b.*

Proof. Let's analyze the delay of a packet at the output of such a succession of elements. By definition, the departure time of a packet p at the FIFO GR server is upper bounded by the GR clock value $GRC(p)$ for that packet at the GR node.

Now, for any $r' \geq 0$, a variable delay element is a GR node, with rate r' and delay $[e_b - \frac{l_{min}}{r'}]^+$([2]) [13], where l_{min} is the minimum packet size for the flow. If we indicate with $GRC'(p)$ the GR clock value of packet p at the variable delay element, and the arrival time of packet p at the variable delay element as $A'(p)$, the departure time $d'(p)$ of a generic packet p at this element is given by

$$A'(p) \leq d'(p) \leq GRC'(p) + \left[e_b - \frac{l_{min}}{r'}\right]^+$$

Now, letting $r' \longrightarrow \infty$, we have $GRC'(p) = A'(p)$, and the departure time $d'(p)$ falls in the interval

$$A'(p) \leq d'(p) \leq A'(p) + e_b$$

Then the total delay of the succession of the FIFO constant delay element and the variable delay element falls in the interval $[e_a,\ e_a + e_b]$. Taking into account also the delay of the FIFO GR element, we have that the departure time $d(p)$ of packet p from the succession of the three elements is

$$A(p) + e_a \leq d(p) \leq GRC(p) + e_a + e_b$$

□

We then have another version of Theorem 4.1:

Theorem 4.5 *Consider a (σ, ρ) constrained flow that traverses a sequence of M nodes. We assume that each node m, $m \in\ 1, ..., M$ is a GR node, with rate r and delay $e_a^m + e_b^m$, at which the departure time for a generic packet falls inside the interval $[A^m(p) + e_a^m,\ GRC^m(p) + e_a^m + e_b^m]$. If all nodes are stable, an end-to-end delay bound for a packet belonging to the flow is given by*

$$d = M\frac{\sigma}{r} + \frac{\rho}{r}\frac{l_{max}}{r}\frac{M(M-1)}{2} + \frac{\rho}{r}\sum_{m=1}^{M}\sum_{j=1}^{m-1} e_b^j + \sum_{m=1}^{M}\left(e_a^m + e_b^m + \tau^{m,m+1}\right) \tag{12}$$

Moreover, the burstiness of the arrival curve for the flow at the output of node M, σ_M, is given by

$$\sigma_M = \sigma + M\rho\frac{l_{max}}{r} + \rho\sum_{m=1}^{M} e_b^m \tag{13}$$

Proof. For each GR node we make use of Proposition 4.2. First of all, we can consider the FIFO constant delay element at the m-th GR node ($m \in\ 1, ..., M$) and the link between this node and node $m+1$ as a single FIFO constant delay element, with delay $\alpha^m = e_a^m + \tau^{m,\, m+1}$.

At the m-th GR node, the service curve of the succession of the FIFO GR element and of the non-FIFO variable delay element is given by the min-plus convolution between:

[2] The notation $[x]^+$ stands for $\max(x,\ 0)$.

Table 2: Comparison of the FIFO and non-FIFO delay bounds

σ	Delay bound (ms) (non-FIFO case, Equation (12))	Delay bound (ms) (FIFO case, Equation (2)[9])
512 bytes	117.49	31.47
1 kB	146.16	35.57
1.5 kB	174.83	39.66
2 kB	203.50	43.76

Values assumed by the delay bounds in Equation (2) from [9] and in Equation (12). We made the following assumptions: $M = 7$ nodes; all nodes have the same delay $e_a + e_b$ (and the same fixed part of delay e_a); all links introduce the same delay $\tau = 400\mu s$; all nodes guarantee a service rate $r = 1$ Mbit/s to the flow; $\rho = r$, $e_a = 100ns$, $e_b = 10ns$; all packets have the same length $l = 512$ bytes.

- the service curve of the FIFO GR element, equal to $\beta_{r,\,0}$;
- the service curve of the non-FIFO variable delay element that, by the equivalence with a GR node [13], is equal to $\beta_{r',\,[e_b^m - \frac{l_{min}}{r'}]^+}$, for any $r' \geq 0$. Letting $r' \longrightarrow \infty$, it becomes equal to $\delta_{e_b^m}$.

The resulting service curve is then given by $\beta_{r,\,e_b^m}$.

The proof then proceeds similarly to the one in Theorem 4.1, with e_b^m instead of e^m at each node, and substituting $\tau^{m,\,m+1}$ with α^m. □

In order to have an idea of the difference between the values assumed by delay bounds in Equation (2) and those given by Equation (12) of Theorem 4.5, in Table 2 we reported the values assumed by the two bounds in an example, for different values of the burstiness σ for the considered flow. We can observe that the actual delay bound in the non-FIFO case can be many times larger than the one holding for FIFO nodes.

4.4 The FIFO case

When GR nodes are FIFO per flow, the result in Theorem 1 [9] is valid. Another way to derive it in the FIFO case is to exploit Network Calculus results for the concatenation of FIFO GR nodes [13]:

Theorem 4.6 *(*Concatenation of FIFO GR nodes *[13]) The concatenation of M GR nodes (that are FIFO per flow) with rates r_m and latencies e^m is GR with rate $r = \min_m(r_m)$ and latency $e = \sum_{m=1}^{M} e^m + \sum_{m=1}^{M-1} \frac{l_{max}}{r_m}$.*

Using Theorem 4.4, a delay bound for the concatenation of the M FIFO GR nodes, when the flow that traverses them is ($\sigma,\ \rho$) constrained, and when all nodes reserve the same service rate r for the flow, is given by the expression in Equation (2), so that we find the result of Theorem 1.

5 Conclusion

In the present paper, we considered end-to-end delay bounds in a network of Guaranteed Rate nodes. We have demonstrated that, contrary to what is stated in the literature, the validity of the available methods to derive end-to-end delay bounds in a network of Guaranteed Rate servers is restricted to the case in which nodes are globally FIFO (that is, they are FIFO per flow and per microflow). We have proved with a counterexample that those delay bounds are not valid in the non-FIFO case. We have exhibited the implicit FIFO assumption in the original derivation of the bounds, and we have determined new bounds that are valid in the non-FIFO case. We have shown the tightness of the bounds derived in the non-FIFO case. We also gave evidence of how, in a realistic scenario, they can be sensibly higher than the ones valid in the FIFO case. We also showed how, in the FIFO case, the existing bounds derive from well-known Network Calculus results.

References

[1] Partridge C. Shectman N. Bennet, J.C.R. Packet reordering is not pathological network behavior. *IEEE/ACM Trans. on Networking*, 7(6):789–798, December 1999.

[2] B. Bensaou, K. Chan, and D. Tsang. Credit-based fair queueing (cbfq): A simple and feasible scheduling algorithm for packet networks. In *Proceedings of the IEEE ATM '97 Workshop, Lisbon, Portugal, 1997.*

[3] S. Blake, D. Black, M. Carlson, E. Davies, Z. Wang, and W. Weiss. An architecture for differentiated services, December 1998. RFC 2475, IETF.

[4] J. M. Blanquer and B. Ozden. Fair queuing for aggregated multiple links. In *Proc. Sigcomm 2001*, pages 189–197, September 2001.

[5] J-Y Le Boudec and A. Charny. Packet scale rate guarantee for non-fifo nodes. *IEEE/ACM Transactions on Networking*, 11(5):810–820, October 2003.

[6] S. Y. Cheung and C. S. Pencea. Bsfq: Bin sort fair queueing. In *Proceedings of Infocom 2002*, 2002.

[7] Markus Fidler and Volker Sander. A parameter based admission control for differentiated services networks. *Comput. Networks*, 44(4):463–479, 2004.

[8] S. J. Golestani. A self clocked fair queuing scheme for high speed applications. In *Proceedings of Infocom 94*, 1994.

[9] P. Goyal, S. S. Lam, and H. Vin. Determining end-to-end delay bounds in heterogeneous networks. In *5th Int Workshop on Network and Op. Sys support for Digital Audio and Video, Durham NH*, April 1995.

[10] P. Goyal and H. Vin. Generalized guaranteed rate scheduling algorithms: a framework. *IEEE/ACM Trans. Networking, vol 5-4*, pages 561–571, August 1997.

[11] Zhang L. Virtual clock: A new traffic control algorithm for packet switching networks. In *Proceedings of ACM SIGCOMM '90*, pages 19–29, August 1990.

[12] G. Stea L. Lenzini, E. Mingozzi. Delay bounds for fifo aggregates: A case study. In *Lecture Notes in Computer Science, Springer*, volume 2811, pages 31–40, 2003.

[13] J.-Y. Le Boudec and P. Thiran. *Network Calculus. A Theory of Deterministic Queueing Systems for the Internet*. Springer Verlag Lecture Notes in Computer Science volume 2050 (available online at http://lcawww.epfl.ch), July 2001.

[14] A. K. Parekh and R. G. Gallager. A generalized processor sharing approach to flow control in integrated services networks: The single node case. *IEEE/ACM Trans. Networking, vol 1-3*, pages 344–357, June 1993.

[15] G. Varghese S. Suri and G. Chandranmenon. Leap forward virtual clock: A new fair queueing scheme with guaranteed delays and throughput fairness. In *Proc. IEEE Infocom 97*, Kobe, Japan.