Efficient Network QoS Provisioning Based on per Node Traffic Shaping

L. Georgiadis, R. Guérin, V. Peris
Advanced Networking Laboratory
IBM T. J. Watson Research Center
P.O. Box 704, Yorktown Heights, NY 10598
{leonid,guerin,vperis}@watson.ibm.com

K. N. Sivarajan*
Indian Institute of Science
Bangalore 560-012, India
kumar@ece.iisc.ernet.in

Abstract

This paper addresses the problem of providing per-connection end-to-end delay guarantees in a high-speed network. We assume that the network is connection oriented and enforces some admission control which ensures that the source traffic conforms to specified traffic characteristics. We concentrate on the class of Rate-Controlled Service Disciplines, in which traffic from each connection is reshaped at every hop, and develop end-to-end delay bounds for the general case where different reshapers are used at each hop. In addition, we establish that these bounds can also be achieved when the shapers at each hop have the same "minimal" envelope.

The main disadvantage of this class of service disciplines is that the end-to-end delay guarantees are obtained as the sum of the worst case delays at each node, but we show that this problem can be alleviated through "proper" reshaping of the traffic to an envelope, which is in general different from the original envelope of the source traffic. We illustrate the impact of this reshaping by demonstrating its use in designing Rate-Controlled Service disciplines that outperform GPS-based service disciplines. Furthermore, we show that we can restrict the space of "good" shapers to a family which is characterized by only one parameter. We also describe extensions to the service discipline that make it work conserving, and as a result reduce the average end-to-end delays.

Keywords: QoS Provisioning, Real-time Traffic, Traffic Shaping, ATM, Scheduling, End-to-end Delay Guarantees.

1 Introduction

In this paper, we consider the problem of providing per connection end-to-end delay (and throughput) guarantees in high speed networks. Various scheduling policies have been suggested in the literature for this purpose. Among them, policies based on Fair Queueing, alternatively known as Generalized Processor Sharing (GPS) [7, 13, 11, 12], have attracted special attention since they guarantee throughput to individual connections and provide smaller end-to-end delay bounds than other policies, for connections that cross several nodes. A key factor in obtaining these smaller delay bounds is the ability to take into account (delay) dependencies in the successive nodes that a connection has to cross, which is in general very difficult to do with other policies.

*And IBM T.J. Watson Research Center.

One notable attempt at addressing this general problem is that of [6] which introduced the concept of service burstiness, and used it to provide a framework to characterize service disciplines and evaluate their end-to-end delay performance. However, the generality of the framework in [6] did not result in as tight end-to-end delay bounds as those obtained by focusing on a specific policy. For example, the bounds available based on the techniques of [6] are no better than the looser bounds found in [12].

In this paper we concentrate on Rate-Controlled Service (RCS) disciplines, which have also been proposed in the literature [16] to provide performance guarantees to individual connections. In this class of service disciplines, the traffic of each connection is reshaped at every node to ensure that the traffic offered to the scheduler arbitrating local packet transmissions conforms to specific characteristics. In particular, it is typically used to enforce, at a node inside the network, the same traffic parameter control as the one performed at the network access point, which is based on the parameters negotiated during connection establishment. Reshaping makes the traffic at each node more predictable and, therefore, simplifies the task of guaranteeing performance to individual connections; when used with a particular scheduling policy, it allows the specification of worst case delay bounds at each node [16]. End-to-end delay bounds can then be computed as the sum of the worst case delay bounds at each node along the path.

The main advantages of an RCS discipline, especially when compared to GPS, are simplicity of implementation and flexibility. Also, in the single node case the RCS discipline that uses the Non Preemptive Earliest Deadline First (NPEDF) scheduling policy, is known to be optimal [8]. However, for the more interesting case of general networks with many nodes, optimality does not hold. In section 4.1 we show with simple examples that when a connection has to cross many nodes, GPS outperforms the "naive" rate-controlled NPEDF discipline. As a result, it has been argued that despite its potentially greater complexity, a GPS-based service discipline should be the solution of choice to provide performance guarantees to individual connections (see for example [3]).

A key result of this paper is to establish that RCS disciplines can be designed so as to outperform GPS-based ones, even in a network environment. This is achieved by proper selection of the traffic reshaping performed at each node. Specifically, any end-to-end delay bounds that can be guaranteed by the GPS discipline, can also be achieved by an RCS discipline by using a simple algorithm to determine how to reshape the traffic, and then specify worst case delay bounds at each node. The sum of the worst case delay bounds of this RCS discipline is then no larger than the delay guarantees provided by the GPS discipline. We also show that RCS disciplines have the additional flexibility of providing end-to-end delay bounds that cannot be guaranteed by the GPS discipline. Furthermore, because of traffic reshaping, the network buffer requirements of RCS disciplines are in general significantly smaller than those of the GPS discipline (see [6] for related discussions). Based on these advantages and their implementation simplicity, we believe that RCS disciplines are very effective candidates for providing end-to-end performance guarantees to individual connections in integrated services networks.

The paper is structured as follows. In Section 2 we introduce our traffic model, and in particular our assumptions concerning properties of the envelope of the input traffic, and the general structure of our shapers. Section 3 is dedicated to the description of RCS disciplines and to the derivation of several results concerning the delay guarantees they can provide given the traffic and shaper models of Section 2. Section 4 is devoted to a comparison with the GPS service discipline. Section 4.1 considers first the simpler

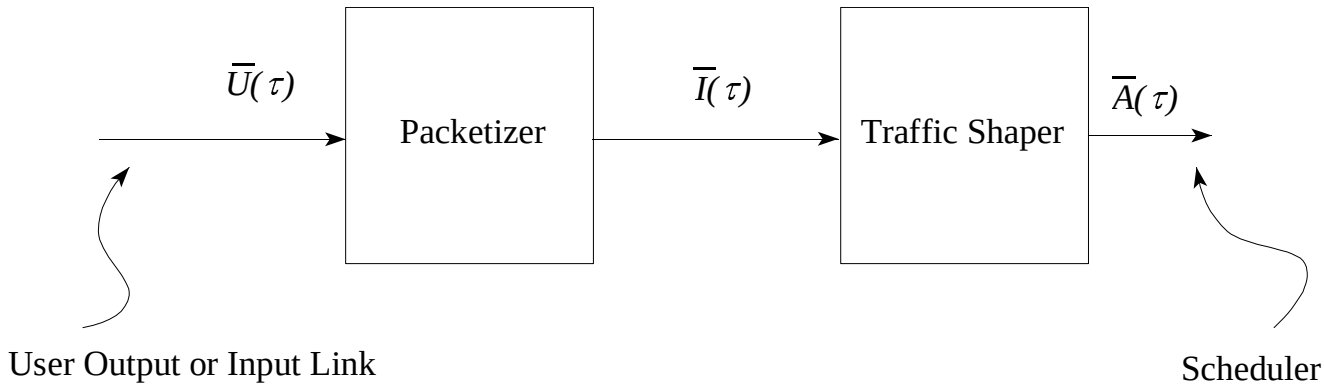

Figure 1: Connection Traffic Flow

version of GPS, i.e., Rate Proportional Processor Sharing (RPPS), as it is of greater practical significance. Section 4.2 considers the more complicated case of general GPS for which similar results are established. Various properties of traffic shapers are investigated in Section 5, and used to establish that the reshaping needed for RCS disciplines to perform well can be achieved using "simple" shapers. Finally, the important extension demonstrating that the results of the paper hold when reshaping is performed only in case of congestion, is the topic of Section 6. A brief conclusion summarizes the main findings of the paper. The Appendices contain proofs of the Lemmas, as well as an extension to the more general case of subadditive traffic envelopes.

2 System Model and Definitions

We consider a store-and-forward network comprised of packet switches in which a packet scheduler is available at each output link. Traffic from a particular connection entering the switch passes through a packetizer and a traffic shaper before being delivered to the scheduler, as indicated in Figure 1. The traffic shaper regulates traffic, so that the output of the shaper satisfies certain pre-specified traffic characteristics.

In this paper, we use a deterministic approach to specify the traffic characteristics of a connection. Traffic arriving on the link is modeled as a fluid, with $U[t, t+\tau]$ denoting the amount of traffic arriving at the network ingress in the interval $[t, t+\tau]$. However,the network element typically operates on packets and so there is a packetizer (see Figure 1) that splits the input traffic into packets. These packets are then regulated by the traffic shaper before reaching the link scheduler which arbitrates the transmission of packets on the link.

We assume that $U(\tau) = U[0, \tau]$ is right-continuous and that there is a nonnegative function $\overline{U}(\tau)$ called *envelope* of $U[t, t+\tau]$, such that

$$U[t, t+\tau] \leq \overline{U}(\tau), \qquad t \geq 0, \tau \geq 0.$$

The envelope function is not unique; without loss of generality (see [2]) we can assume that $\overline{U}(\tau)$ is

right-continuous, nondecreasing, and subadditive.

The packetizer splits the input traffic into packets of maximum length L, that are instantaneously delivered to the shaper when the last bit of the packet is received. We denote the traffic at the output of the packetizer in the interval $[t, t+\tau]$ as $I[t, t+\tau]$. It is easy to see that, for any non-negative t and τ,

$$I[t, t+\tau] \leq U[t, t+\tau] + L \leq \overline{U}(\tau) + L =: \overline{I}(\tau) \tag{1}$$

The traffic shaper reshapes the incoming traffic by delaying the packets according to the rules described next, and then delivers them to the scheduler. The traffic shaper is characterized by a traffic envelope, $\overline{A}(\tau)$, which upper bounds the amount of traffic that is output by the shaper in any interval of length τ. If $A[t, t+\tau]$ denotes the traffic that is output from the shaper in the interval $[t, t+\tau]$, then $A[t, t+\tau] \leq \overline{A}(\tau)$.

More precisely, the traffic shaper outputs packets in order with each packet being released at the smallest time t such that

$$A[t-\tau, t] \leq \overline{A}(\tau), \qquad 0 \leq \tau \leq t. \tag{2}$$

The traffic shapers that we use in this paper can be constructed from simple (σ, ρ) traffic shapers that can alternatively be described in terms of the backlog traffic in a hypothetical queue with a server of rate ρ [4]. Assume that traffic $I[0, t]$ is queued for transmission at a link of speed $\rho > 0$ and define $W_\rho(I)(t)$ as the amount of traffic queued at time t at this link, including the packet that may have arrived at time t. It is known [4] that $W_\rho(I)(t)$ is given by

$$W_\rho(I)(t) := \max_{0 \leq s \leq t} \{I[s, t] - \rho(t-s)\}. \tag{3}$$

The (σ, ρ) traffic shaper, operates on the ith packet arriving at time s_i, according to the following rule. The packet is released to the scheduler at the earliest time $t_i \geq s_i$, such that the shaper output traffic, $A[t, t+\tau]$, satisfies the condition

$$W_\rho(A)(t_i) \leq \sigma = L + \delta, \qquad \delta \geq 0,$$

where $W_\rho(A)(t_i)$ is defined as in (3). Note that the condition $\delta \geq 0$ is necessary in order to allow packets of size L to pass through the shaper. This shaper corresponds to the operation of a leaky bucket in a store-and-forward network [1], which differs from the one defined in [4] in two minor respects: i) packets are entering and exiting the shaper instantaneously and not at a constant rate C, and ii) the length of the packet that exits the traffic shaper at time t_i is taken into account in the calculation of $W_\rho(A)(t_i)$. Note that s_i and t_i are defined as the times when the *last* (not the *first* as in [4]) bit of the ith packet enters and exits the shaper respectively. However, with $d_i := t_i - s_i$ denoting the delay that the ith packet experiences in a shaper, the analysis in [4] can be repeated with minor modifications to show that

$$\begin{aligned} d_i &= \frac{1}{\rho}\left(W_\rho(I)(s_i) - \sigma\right)^+, \text{ and} \\ A[t, t+\tau] &\leq \sigma + \rho\tau. \end{aligned}$$

The (σ, ρ) shaper has also been described in the literature in terms of a token bucket (leaky bucket), with ρ being the rate of token accumulation, and σ being the bucket depth. In general, we will be using shapers

whose output is a concave, increasing (i.e., $f(t_1) < f(t_2)$ whenever $t_1 < t_2$), piecewise linear function with finite number of slopes, K. We are interested in these types of shapers because they are a generalization of the shapers adopted by the the Internet [15] and ATM standards [1]. These shapers can also be easily implemented by passing the traffic through a series of K (σ_m, ρ_m)-shapers [5]. Let $\mathcal{A}$ be such a shaper and for the input traffic model described earlier, the delay of packet i through the shaper is [5, Theorem 5.1]

$$d_i = \max_{m=1,2,\ldots,K} \left\{ \frac{1}{\rho_m} \left(W_{\rho_m}(I)(s_i) - \sigma_m\right)^+ \right\} \text{ and} \tag{4}$$

$$A[t, t+\tau] \leq \overline{A}(\tau) := \min_{m=1,2,\ldots,K} \left\{\sigma_m + \rho_m \tau\right\}, \tag{5}$$

where $(x)^+ = \max(0, x)$. From (4) we can develop a worst case delay bound that depends on the envelope of the input process to the shaper. Taking into account (3) we have that

$$\begin{aligned} d_i &\leq \max_{m=1,2,\ldots,K} \left\{ \frac{1}{\rho_m} \left(\max_{0 \leq s \leq s_i} \left\{ \overline{I}(s_i - s) - \rho_m (s_i - s) - \sigma_m \right\} \right)^+ \right\} \\ &\leq \max_{m=1,2,\ldots,K} \left\{ \left(\max_{\tau \geq 0} \left\{ \frac{\overline{I}(\tau) - \sigma_m - \rho_m \tau}{\rho_m} \right\} \right)^+ \right\} \qquad (6) \\ &= \max_{\tau \geq 0} \left\{ \left(\max_{m=1,2,\ldots,K} \left\{ \frac{\overline{I}(\tau) - \sigma_m}{\rho_m} \right\} - \tau \right)^+ \right\}. \qquad (7) \end{aligned}$$

We will denote the bound on the delay of traffic with envelope $\overline{I}(\tau)$ through shaper $\mathcal{A}$ as,

$$D\left(\overline{I}, \mathcal{A}\right) := \max_{\tau \geq 0} \left\{ \left(\max_{m=1,2,\ldots,K} \left\{ \frac{\overline{I}(\tau) - \sigma_m}{\rho_m} \right\} - \tau \right)^+ \right\}. \tag{8}$$

We can write (8) in another form that will be useful in the sequel. The range of $\overline{A}(\tau)$ is $[\min_m \sigma_m, \infty)$ and the inverse of $\overline{A}(\tau)$ is given by

$$\overline{A}^{(-1)}(y) = \max_{m=1,\cdots,K} \left\{ \frac{y - \sigma_m}{\rho_m} \right\}, \quad \min_m \sigma_m \leq y < \infty. \tag{9}$$

Extending the definition of $\overline{A}^{(-1)}(y)$ by setting $\overline{A}^{(-1)}(y) = 0$ whenever $0 \leq y < \min_m \sigma_m$, it can be seen from (8) and (9) that

$$D(\overline{I}, \mathcal{A}) = \max_{\tau \geq 0} \left\{ \left(\overline{A}^{(-1)} \left(\overline{I}(\tau) \right) - \tau \right)^+ \right\}. \tag{10}$$

When the traffic entering shaper $\mathcal{A}$ is the output of a shaper $\mathcal{A}_1$ with envelope $\overline{A}_1(\tau)$, we will also use the notation $D\left(\mathcal{A}_1, \mathcal{A}\right) := D\left(\overline{A}_1, \mathcal{A}\right)$. Notice that if $\overline{I}(\tau) \leq \overline{A}(\tau)$, $\tau \geq 0$, then from (6) we have that $D\left(\overline{I}, \mathcal{A}\right) = 0$ which implies that no packet is delayed in shaper $\mathcal{A}$. In particular, $D\left(\mathcal{A}, \mathcal{A}\right) = 0$.

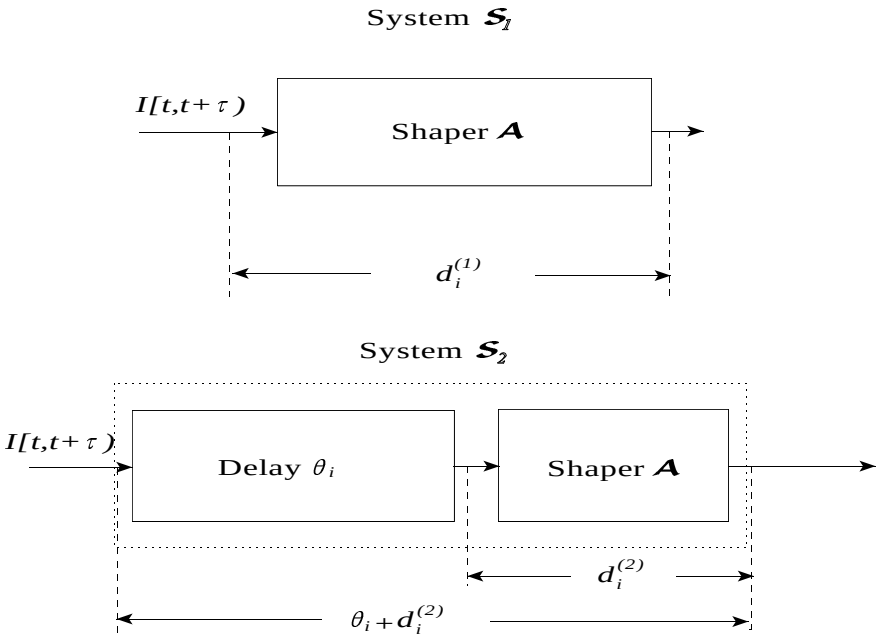

Figure 2: The Systems $\mathcal{S}_1$ and $\mathcal{S}_2$

Consider next two shapers $\mathcal{A}_1$, $\mathcal{A}_2$ in series. Equations (4) and (5) imply that this arrangement is equivalent to a traffic shaper $\mathcal{A}_3$ with envelope

$$\overline{A}_3(\tau) = \min\left\{\overline{A}_1(\tau), \overline{A}_2(\tau)\right\}. \tag{11}$$

Equivalence here means that for any input traffic pattern, the delay of every packet from the time it enters $\mathcal{A}_1$ to the time it exits $\mathcal{A}_2$ is identical to the delay of the packet in $\mathcal{A}_3$.

Next, we state a useful lemma, that relates the packet delays in the two systems $\mathcal{S}_1$ and $\mathcal{S}_2$ of Figure 2. System $\mathcal{S}_1$ consists of a traffic shaper, $\mathcal{A}$. System $\mathcal{S}_2$ consists of a "delay" subsystem and an identical shaper $\mathcal{A}$ connected in series. The delay subsystem delays the ith arriving packet by $\theta_i \geq 0$, and then delivers it to $\mathcal{A}$.

Lemma 1. *Assume that packets arrive to systems $\mathcal{S}_1$, $\mathcal{S}_2$ according to the same arrival process $I[t, t+\tau]$. If $d_i^{(1)}$ and $d_i^{(2)}$ are the delays of packet i in the traffic shaper in systems $\mathcal{S}_1$ and $\mathcal{S}_2$ respectively, then for all $i = 1, 2, \ldots$,*

$$d_i^{(1)} \leq d_i^{(2)} + \theta_i,$$

that is, the delays of all packets in system $\mathcal{S}_1$ are smaller than their corresponding delay in system $\mathcal{S}_2$.

The proof of the above lemma is given in Appendix A. Lemma 1 identifies the monotonicity property of the shaper with respect to the arrival process. This is an important property of the traffic shapers considered in this paper and is key to establishing the general end-to-end delay bounds for RCS disciplines.

3 Rate-Controlled Service Disciplines

We are interested in a generalized form of the class of Rate-Controlled Service disciplines introduced in [16]. In that work, it is assumed that connections whose traffic satisfies certain burstiness constraints enter the network at various nodes. A node can have several output links, each of which contains a scheduler that decides the order in which packets are transmitted. At each node along the path of a connection, traffic is reshaped to conform to its original envelope before it enters the appropriate scheduler. Based on the traffic envelope of the connection, upper bounds on the scheduling delays at each node can be guaranteed. It is also shown in [16], that for the traffic shapers considered there, reshaping the traffic to its original envelope does not introduce extra delays. Therefore, an upper bound on the end-to-end packet delay is simply the sum of the scheduling and propagation delays.

In this paper, we study the following general class of service disciplines. The traffic of connection n entering the network has an envelope function $\overline{U}_n(\tau)$. At node m, the traffic of connection n is shaped by a traffic shaper $\mathcal{A}_n^m$. Traffic shapers $\mathcal{A}_n^m$ are of the general type considered in Section 2, and different traffic shapers can be used for the same connection at different nodes. The connection traffic exiting $\mathcal{A}_n^m$ enters a scheduler at the appropriate output link at node m, and it is scheduled for transmission to the next node or to its destination. We develop end-to-end delay bounds based on the scheduling policies at each node as well as the form of the traffic shapers $\mathcal{A}_n^m$. These bounds will then be used to provide delay guarantees to each connection. In the rest of this paper, we use the term *service discipline* to denote the operation of the system consisting of the traffic shaper as well as the scheduler. We are interested in designing service disciplines of the type described above, so that end-to-end delay guarantees can be provided as efficiently as possible.

We assume that the nodes are output queueing switches, and without loss of generality, that there is no delay inside the switch. In other words, the only delay that a packet incurs at a switch is due to queueing at the output link. Let $C^{m,l}$ be the set of connections passing through output link l of node m. Given $\mathcal{A}_n^m$, $n \in C^{m,l}$, and the scheduling policy employed at link l at node m, we assume that a delay bound on the scheduling delay D_n^m is known for each connection $n \in C^{m,l}$. The scheduling delay includes both queueing and transmission time of a packet. For example, bounds of this form can be developed for the general traffic shapers of this paper, when the Earliest Deadline First (EDF) scheduling policy is employed, by a straightforward extension of the method in [17, 8] (see also Theorem 1 in Section 4 in this paper). We also assume that an upper bound on the propagation delay of link l is T^l. Knowledge of the bounds D_n^m and T^l alone are not enough to provide bounds on the end-to-end packet delays. We still have to account for any additional delays incurred in the traffic shapers and this is done based on the following proposition.

Proposition 1. *Assume that the output of traffic shaper $\mathcal{A}_1$ enters a system $\mathcal{S}$ where it is known that the delay experienced by these packets is upper bounded by $D_{\mathcal{S}}$. The output of system $\mathcal{S}$ enters shaper $\mathcal{A}_2$. The total delay, $\widehat{d}_i$, that packet i experiences from the time it exits $\mathcal{A}_1$ to the time it exits $\mathcal{A}_2$ is upper bounded by*

$$\widehat{d}_i \leq D_{\mathcal{S}} + D(\mathcal{A}_1, \mathcal{A}_2).$$

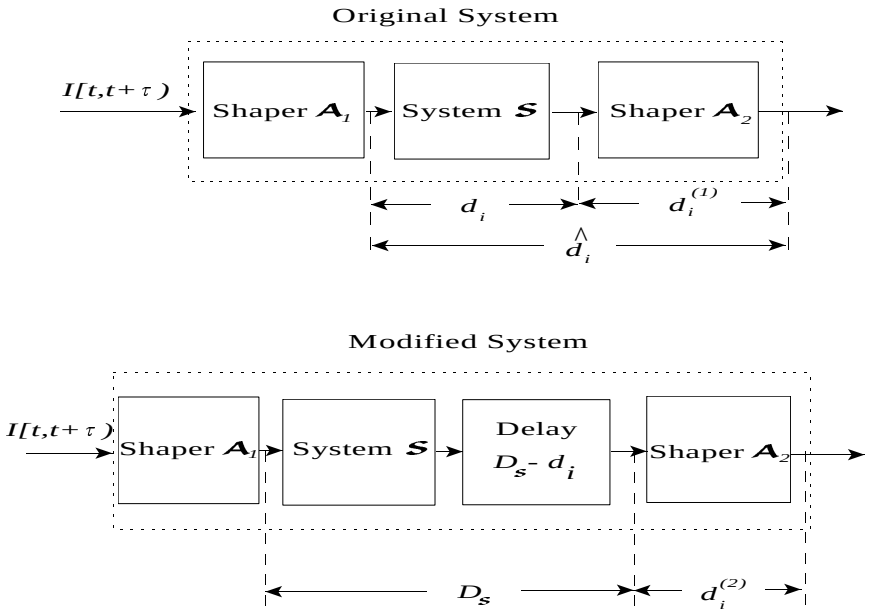

Figure 3: Original and Modified System

Proof. Let d_i be the delay of packet i in system $\mathcal{S}$, and let $d_i^{(1)}$ be its delay in $\mathcal{A}_2$. By definition, $\widehat{d}_i = d_i + d_i^{(1)}$. Consider next a modified system where a delay system that delays the ith packet by $\theta_i = D_{\mathcal{S}} - d_i$, is inserted between $\mathcal{S}$ and $\mathcal{A}_2$ (see Figure 3). Now let $d_i^{(2)}$ denote the delay of packet i in $\mathcal{A}_2$ under this new arrangement. Note that $\theta_i \geq 0$ by the definition of $D_{\mathcal{S}}$. Applying Lemma 1 we conclude that

$$d_i^{(1)} \leq D_{\mathcal{S}} - d_i + d_i^{(2)}$$

and therefore,

$$\widehat{d}_i \leq D_{\mathcal{S}} + d_i^{(2)}.$$

Observe now that since the delay of every packet between its entrance time to $\mathcal{S}$ and its exit from the delay system is $d_i + \theta_i = D_{\mathcal{S}}$, the traffic entering shaper $\mathcal{A}_2$ when the delay system is inserted, is a time-shifted version of the traffic exiting $\mathcal{A}_1$, and therefore it has envelope $\overline{A}_1(\tau)$. Hence, $d_i^{(2)} \leq D(\mathcal{A}_1, \mathcal{A}_2)$. ■

Note: As explained in Section 2, when the shapers $\mathcal{A}_1, \mathcal{A}_2$ are identical, $D(\mathcal{A}_1, \mathcal{A}_2) = 0$, i.e., in this case reshaping does not introduce extra delays. Also, from the proof we see that any shaper that has the property described in Lemma 1 satisfies Proposition 1 as well. In particular, the shaper of [16] can easily be seen to satisfy Lemma 1.

Assume now, that connection n passes through M network nodes, numbered from 1 to M, and let $M+1$ denote the destination. We then, apply Proposition 1 with the system $\mathcal{S}$ consisting of both the

scheduler at node m and the link $l = (m, m+1)$, and the shapers $\mathcal{A}_1 \equiv \mathcal{A}_n^m$, and $\mathcal{A}_2 \equiv \mathcal{A}_n^{m+1}$. We can conclude that the delay that a packet from connection n experiences between the time it exits shaper $\mathcal{A}_n^m$ and the time it exits $\mathcal{A}_n^{m+1}$ is upper bounded by,

$$D_n(m, m+1) = D_n^m + T^{(m,m+1)} + D\left(\mathcal{A}_n^m, \mathcal{A}_n^{m+1}\right). \tag{12}$$

Taking (12) into account we then have the following guaranteed upper bound on the end-to-end delay

$$\begin{aligned} \overline{D}_n &= D(\overline{I}_n, \mathcal{A}_n^1) + \sum_{m=1}^{M-1} D_n(m, m+1) + D_n^M + T^{(M,M+1)} \\ &= D(\overline{I}_n, \mathcal{A}_n^1) + \sum_{m=1}^{M-1} D\left(\mathcal{A}_n^m, \mathcal{A}_n^{m+1}\right) + \sum_{m=1}^{M} D_n^m + \sum_{m=1}^{M} T^{(m,m+1)}. \end{aligned} \tag{13}$$

It is important to note that the delay bounds D_n^m depend on the choice of the traffic shapers $\mathcal{A}_n^m$. Therefore, one should not conclude from (13) that the end-to-end delay guarantees are minimized by choosing $\overline{I}_n(\tau)$ as the envelope for all the traffic shapers so that $D(\overline{I}_n, \mathcal{A}_n^1) = D\left(\mathcal{A}_n^m, \mathcal{A}_n^{m+1}\right) = 0$. In fact, as we will see in the next section, this choice may be quite inappropriate.

As in the policies proposed in [16], the delay bounds in (13) are basically a sum of the worst case delays at *each* node along the path of a connection. However, an individual packet may not encounter the worst case delays at each node. Therefore, one may suspect that these bounds are overly pessimistic and lead to inefficient allocations compared to bounds for other disciplines, that take into account delay dependencies between nodes along the path. As mentioned earlier, the impact of delay dependencies is in general difficult to evaluate but can be accounted for in some instances. In particular, it can be done for the GPS discipline [7, 11, 12], which is one of the reasons that tight bounds can be obtained for this discipline. This argument about the inefficiency of worst-case delay assignment relative to GPS was also mentioned in [16].

In the next section we address this issue, by demonstrating that with a suitable choice of shaper envelopes the RCS discipline can provide the same end-to-end delay guarantees that the best delay bounds of GPS can provide. More specifically, we show that for a given set of connections, and their associated paths, the RCS discipline can provide the same end-to-end delay bounds as the GPS discipline. In addition, we show that the RCS discipline can accept a set of connections with associated delay requirements, that cannot be accepted by GPS. This demonstrates the advantage of RCS over GPS in providing efficient end-to-end delay guarantees.

4 Comparison with GPS

In this section, we compare the performance of the GPS service discipline to the performance of the RCS disciplines introduced in the previous section. In order to compare two service disciplines, we need to define the performance measure which is of interest to us. The ability of a discipline to provide efficient end-to-end delay guarantees to a given set of connections, is best quantified by the notion of *schedulable region*. Assume that we have N_T connections in a communication network, with the same scheduling discipline, π, operating at all the links in the network. The input traffic of connection n has envelope

function $\overline{I}_n(\tau)$, and traverses path P_n of the network, $1 \leq n \leq N_T$. Under these assumptions, we require that the packets of connection n have an upper bound on their end-to-end delay (delay guarantee), $\overline{D}_n$, $1 \leq n \leq N_T$. The vector $\overline{\mathbf{D}} = \left(\overline{D}_1, \ldots, \overline{D}_{N_T}\right)$ is *schedulable* under discipline π if the delay bound $\overline{D}_n$ can be guaranteed under π for all packets of connection n, $1 \leq n \leq N_T$. The *schedulable region* of discipline π is the set of all vectors $\overline{\mathbf{D}}$ that are schedulable under π. Note that the schedulable region of a service discipline depends on the envelope functions $\overline{I}_n(\tau)$ and the paths P_n. We say that service discipline π_1 is *at least as good as* the discipline π_2, if the schedulable region of π_1 is a superset of π_2, for any given set of connections and paths. If, in addition, there is a set of connections, paths and associated delay bounds that can be guaranteed by π_1, but not by π_2, we say that π_1 is *better* than π_2.

Note that the schedulable region is defined in terms of delay bounds that can be guaranteed *a priori*. These bounds are an integral part of the service discipline and may in fact be significantly worse than the delays actually experienced by packets. Their choice may be due either to their simplicity or to the fact that these are the only bounds that can be guaranteed and no method is known to derive lower bounds. From the point of view of admission control, it is irrelevant if in the actual operation of a policy smaller delays are observed, since what is required at the time of connection establishment, is to know whether the delay bounds can be *guaranteed* or not.

Before we proceed with the comparison of RCS and GPS disciplines, we need to recall some preliminary results regarding the NPEDF scheduling policy. This policy has the largest schedulable region among the class of non-preemptive policies in the single-node case [8], and is therefore the most efficient to use when considering RCS disciplines. The schedulable region is defined here with respect to scheduler delays only. The schedulable region for N connections that are entering the scheduler through traffic shapers with envelopes $\overline{A}_n(\tau) = L + \delta_n + \rho_n \tau$, $1 \leq n \leq N$, and contending for an output link of speed r, is given by Theorem 4 in [8], which we repeat here for convenience, slightly rephrased to conform to our definitions and notation.

Theorem 1. *The NPEDF policy is optimal among the class of non-preemptive scheduling policies when the connection n traffic entering the scheduler has envelope $\overline{A}_n(\tau) = L + \delta_n + \rho_n \tau$, $1 \leq n \leq N$. Under the stability condition $\sum_{n=1}^{N} \rho_n \leq r$, the schedulable region of NPEDF consists of the set of vectors $(D_1, \ldots, D_N)$ that satisfy the constraints*

$$\min\{k+1, N\} L + \sum_{n=1}^{k} \delta_{i_n} \leq D_{i_k} \left(r - \sum_{n=1}^{k-1} \rho_{i_n} \right) + \sum_{n=1}^{k-1} \rho_{i_n} D_{i_n}, \quad 1 \leq k \leq N.$$

whenever $D_{i_1} \leq \ldots \leq D_{i_N}$.

We note that while the optimality of NPEDF was established in [8] for envelopes of the form $\overline{I}_n(\tau) = L + \delta_n + \rho_n \tau$, it is straightforward to see that all the arguments used to derive Theorem 1 in [8], go through by simply replacing $L + \delta_n + \rho_n \tau$ with a general envelope $\overline{A}_n(\tau)$ of the type considered here. For these general envelopes, the appropriate analogue of Theorem 1 can be easily derived by simply rephrasing Lemmas 1 and 2 in [8].

4.1 Achieving RPPS Delay Guarantees

In this and the next section, we assume for comparison purposes that the traffic of connection n, entering the first node packetizer has envelope $\overline{U}_n(\tau) = \delta_n + \rho_n\tau$. Therefore, the envelope of the traffic that enters the first traffic shaper is $\overline{I}_n(\tau) = L + \delta_n + \rho_n\tau$. We also assume that propagation delays are zero. For definitions and notations relating to GPS the reader is referred to [11], [12]. Recall from Section 3, that $C^{m,l}$ is the set of connections that pass through the output link l of node m. Denoting the speed of this link as $r^{m,l}$, we will assume throughout the rest of this section the stability condition

$$\sum_{n \in C^{m,l}} \rho_n \leq r^{m,l}.$$

The GPS policy operates by allocating weight ϕ_n^m for connection n whose traffic crosses node m. These weights are used to determine the rate at which traffic from connection n is served when a set B^m of connections is backlogged at the output link l of node m through which connection n passes. Specifically, the service rate of connection n is given by

$$g_n^m = \frac{\phi_n^m}{\sum_{k \in B^m} \phi_k^m} r^m,$$

where for simplicity in notation we denote $r^{m,l}$ as r^m when there is no possibility of confusion. PGPS is a non-preemptive policy that tracks GPS. In general the procedure developed in [11] to obtain delay bounds given the weights, ϕ_n^m, is complicated and imposes certain restrictions on the ϕ_n^m. Moreover, it becomes even more cumbersome in the practically more important inverse procedure of specifying appropriate weights, in order to satisfy predetermined delay bounds. However, a simple bound can be obtained in the special case of non-preemptive RPPS, where $\phi_n^m = \rho_n$ for all nodes through which the connection passes. Specifically, the end-to-end delay bound, $\overline{D}_n^*$, obtained under non-preemptive RPPS for connection n with envelope $\overline{I}_n(\tau) = L + \delta_n + \rho_n\tau$, that crosses nodes $1, \ldots, M$, is given by,

$$\overline{D}_n^* = \frac{\delta_n + ML}{\rho_n} + \sum_{m=1}^{M} \frac{L}{r^m}, \tag{14}$$

where we have replaced σ_n with $\delta_n + L$ to conform to our input model (see [12, 9]).

From formula (14) we can already see the weakness of the RCS disciplines relative to RPPS, if the traffic shapers for connection n at every node have envelopes identical to the input envelope $\overline{I}_n(\tau)$. In this case the delays $D(\overline{I}, \mathcal{A}_n^1) = D\left(\mathcal{A}_n^m, \mathcal{A}_n^{m+1}\right) = 0$, $1 \leq m \leq M-1$. Since propagation delays are assumed zero, we therefore have that

$$\overline{D}_n = \sum_{m=1}^{M} D_n^m.$$

It is easy to see that since the connection n shaper at node m has envelope $\overline{I}_n(\tau)$, the delay bound D_n^m is at least $(\delta_n + L)/r^m$. Therefore, the end-to-end delay bound guaranteed by the RCS discipline verifies,

$$\overline{D}_n \geq \sum_{m=1}^{M} \frac{\delta_n}{r^m} + \sum_{m=1}^{M} \frac{L}{r^m}.$$

Since δ_n can be much larger than L, the bounds provided by the RCS discipline under the scenario considered here can be much worse than those obtained under RPPS. For example, if $\delta_n = 50L$ and $r^m = 1.25\rho_n$, $1 \leq m \leq M$, we have

$$\frac{\overline{D}_n}{\overline{D}_n^*} \geq \frac{40.8 \times M}{50 + 1.8 \times M}.$$

Therefore, when $M = 2$ we already have $\overline{D}_n/\overline{D}_n^* \geq 1.52$, and for large M, $\overline{D}_n/\overline{D}_n^* \geq 22.67$. As was mentioned in Section 3, this discrepancy is due to the fact that the bounds for RPPS take into account delay dependencies at the various nodes, while the bounds for the RCS disciplines are based on independently summing the worst case bounds at each node.

The previous example notwithstanding, we show next that we can design RCS disciplines that provide the same delay guarantees as RPPS, by employing traffic shapers with envelopes that are, in general, different from that of the input traffic.

We design the RCS discipline π as follows. We choose NPEDF as the scheduling policy at the output link of each node. The traffic shaper for connection n at each node along its path has envelope

$$\overline{A}_n^m(\tau) = L + \rho_n \tau, \quad 1 \leq m \leq M.$$

Assume that connection n is routed through output link l at node m and let r^m denote the speed of this link. For connection n, we specify the delay bounds for the NPEDF scheduling policy, at node m as

$$D_n^m = L/\rho_n + L/r^m. \tag{15}$$

Let us first show that these bounds can be guaranteed by the NPEDF policy at every node. Consider output link l at node m. Denote by N the total number of connections multiplexed on this link, and index the connections by $i_1, i_2, \ldots, i_N$ such that $D_{i_1}^m \leq D_{i_2}^m \leq \ldots \leq D_{i_N}^m$. Using Theorem 1, and noting that $\delta_n^m = 0$ for all traffic shapers by design, it suffices to show that

$$\min\{k+1, N\} L \leq D_{i_k}^m \left(r^m - \sum_{n=1}^{k-1} \rho_{i_n} \right) + \sum_{n=1}^{k-1} \rho_{i_n} D_{i_n}^m, \; 1 \leq k \leq N. \tag{16}$$

Using (15) we have

$$\begin{aligned} D_{i_k}^m \left(r^m - \sum_{n=1}^{k-1} \rho_{i_n} \right) + \sum_{n=1}^{k-1} \rho_{i_n} D_{i_n}^m &= L \frac{r_m - \sum_{n=1}^{k-1} \rho_{i_n}}{\rho_{i_k}} + L \frac{r^m - \sum_{n=1}^{k-1} \rho_{i_n}}{r^m} \\ &\quad + (k-1)L + L \frac{\sum_{n=1}^{k-1} \rho_{i_n}}{r^m} \\ &= L \frac{r^m - \sum_{n=1}^{k-1} \rho_{i_n}}{\rho_{i_k}} + kL \\ &\geq (k+1)L, \end{aligned}$$

where the last inequality follows from the stability condition, $r^m \geq \sum_{n=1}^{N} \rho_i$. This shows (16).

We now proceed to derive the end-to-end delay bounds for the connections. Since the traffic shapers are identical, we have that $D\left(\mathcal{A}_n^m, \mathcal{A}_n^{m+1}\right) = 0$, $1 \leq m \leq M-1$. Therefore, from (13) we have

$$\overline{D}_n = D\left(\overline{I}_n, \mathcal{A}_n^1\right) + \sum_{m=1}^{M} D_n^m.$$

For the delay $D\left(\overline{I}_n, \mathcal{A}_n^1\right)$, using (6), we have

$$D\left(\overline{I}_n, \mathcal{A}_n^1\right) = \max_{\tau \geq 0}\left\{\frac{\overline{I}_n(\tau) - L - \rho_n \tau}{\rho_n}\right\} = \frac{\delta_n}{\rho_n}.$$

Therefore, taking into account (15) obtain,

$$\begin{aligned}\overline{D}_n &= \frac{\delta_n}{\rho_n} + \sum_{m=1}^{M}\frac{L}{\rho_n} + \sum_{m=1}^{M}\frac{L}{r^m} \\ &= \frac{\delta_n + ML}{\rho_n} + \sum_{m=1}^{M}\frac{L}{r^m}. \qquad (17)\end{aligned}$$

Since (14) is identical to (17) we see that the proposed RCS discipline π can guarantee the same delays as RPPS.

From the above argument we see that if the delay bounds in (14) are required by the connections in the network, then the RCS discipline π, proposed above can be used. Its implementation is simpler than that of RPPS. In addition, it provides the flexibility of easily specifying other delay bounds, whereas the bounds in RPPS are tied to the rate ρ_n of a connection.

If the end-to-end delay requirements of a connection are smaller than (14), a slightly more general version of RPPS can be used. Rather than providing a rate of ρ_n to connection n, better delay performance can be obtained by giving it a rate of $g_n \geq \rho_n$, at each node. The end-to-end delay bound is then given by,

$$\overline{D}_n^* = \frac{\delta_n + ML}{g_n} + \sum_{m=1}^{M}\frac{L}{r^m}. \qquad (18)$$

The previous analysis still applies with very little modification and can be used to specify an RCS discipline that guarantees the bounds in (18). In this case, all traffic shapers have envelopes $\overline{A}_n^m(\tau) = L + g_n\tau$ and the delay guarantees at the scheduler of node m are,

$$D_n^m = L/g_n + L/r^m.$$

Observe that, the schedulability check for RPPS is now $\sum_{l \in C_n^m} g_l \leq r^m$, $m = 1, \ldots, M$, where C_n^m denotes the set of connections that are multiplexed on the same link as connection n at node m. This implies that some amount of bandwidth *viz.* $r^m - \sum_{l \in C_n^m} g_l$, cannot be utilized by RPPS. This bandwidth can be used by an RCS discipline to accept additional connections that require relatively larger end-to-end delay guarantees. At the end of this section we provide a specific example of this benefit of RCS disciplines over the more general GPS disciplines.

4.2 Achieving GPS Delay Guarantees

In [12, Section VIII], tight bounds on per connection packet delays are developed for GPS under a fairly general assignment of weights, ϕ_i^m, called Consistent Relative Session Treatment (CRST). These bounds are achieved in certain node configurations, and even in the special case of RPPS they can be much tighter than those provided by (14). However, the calculation of the bounds is much more cumbersome as they take into account the effect of all the other connections along a connection's path. We will show that even with these tight bounds, an RCS discipline can be designed that guarantees the same delay bounds.

To simplify the discussion and to avoid obscuring the main idea of the argument, we assume a continuous flow model, i.e., packetization is not taken into account. Therefore, we consider the GPS policy (instead of PGPS), and assume that the RCS discipline uses the EDF scheduling policy (instead of NPEDF). As far as the design of traffic shapers is concerned, this assumption basically amounts to setting $L = 0$.

Before proceeding with the design of the RCS discipline, we need some preliminary results. Consider a single link on which N connections are multiplexed, and assume that all of them are "greedy", i.e., the amount of connection v traffic, $1 \leq v \leq N$, arriving in the interval $[0, t]$ is $\widetilde{\delta}_v + \rho_v t$. Then, the service function of connection n, $S_n(t)$, is the amount of connection n traffic that is served in the interval $[0, t]$. In [11, page 355] $S_n(t)$ is used to derive delay bounds of connection n traffic whose envelope is $\widetilde{\delta}_n + \rho_n \tau$. The next lemma improves these delay bounds for connection n, when it has a smaller envelope $\overline{I}_n(\tau)$ such that $\overline{I}_n(\tau) \leq \widetilde{\delta}_n + \rho_n \tau$, $\tau \geq 0$. Let $S_n[t_1, t_2]$ be the connection n traffic served under GPS in the interval $[t_1, t_2]$. Assuming that the system starts empty, the backlog of connection n traffic at time t is defined as the difference $I_n[0, t] - S_n[0, t]$.

Lemma 2. *Assume that the connection n traffic satisfies* $I_n[t, t+\tau] \leq \overline{I}_n(\tau) \leq \widetilde{\delta}_n + \rho_n \tau, \quad t, \tau \geq 0$, *for every connection n that is multiplexed on a given link. If the system starts empty, then an upper bound on connection n delay under GPS is*

$$D_n^* = \max_{\tau \geq 0} \left\{ \min_{t \geq \tau} \left\{ t : S_n(t) \geq \overline{I}_n(\tau) \right\} - \tau \right\}.$$

The proof can be found in Appendix A. For our purposes, the case where $\overline{I}_n(\tau) = \min\{c_n \tau, \ \delta_n + \rho_n \tau\}$, $c_n \geq \rho_n$, $\delta_n \leq \widetilde{\delta}_n$ will be of interest. For convenience, we summarize in the next corollary two specific cases of Lemma 2 that will be useful in the rest of this section. recall from [11] that $S_n(t)$ is a piece-wise linear function, convex in the range $[0, t^B]$ where t^B is the end of the first busy period of connection n, when all the N connections are greedy. In this range, $S_n(t)$ is characterized by the pairs $(s_k, b_k)_{k=1}^{k_n}$ where s_k is the slope of the kth segment and b_k is its duration. Because of the convexity of $S_n(t)$ we have that

$$s_1 \leq s_2 \leq \ldots \leq s_{k_n}.$$

Corollary 1. *Assume that the conditions of Lemma 2 hold, so that* $\overline{I}_l(\tau) \leq \widetilde{\delta}_l + \rho_l \tau, \ \tau \geq 0, \ 1 \leq l \leq N$, *and furthermore let* $\overline{I}_n(\tau) = \min\{c_n \tau, \ \delta_n + \rho_n \tau\} \leq \widetilde{\delta}_n + \rho_n \tau, \ c_n \geq \rho_n, \ \tau \geq 0$.

1. *If* $s_1 \geq c_n$, *then* $D_n^* = 0$.

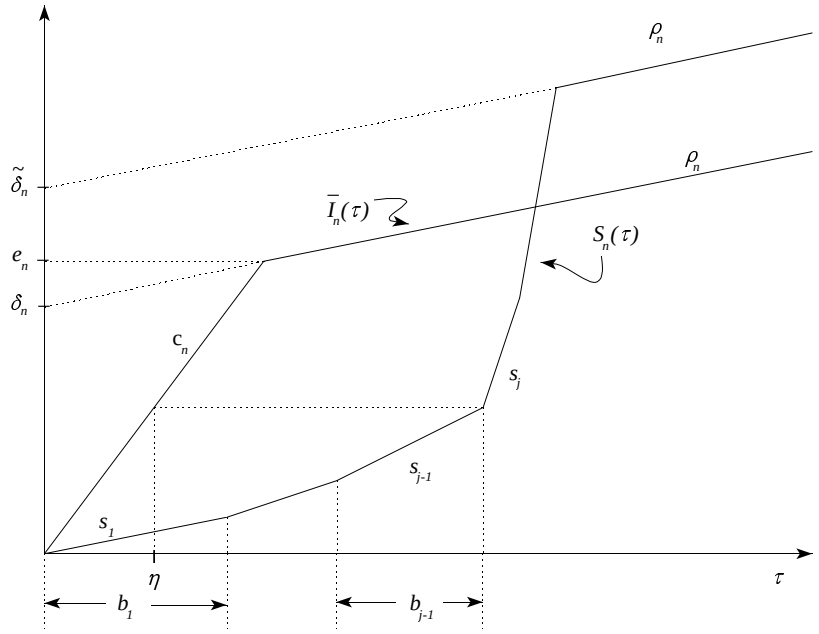

Figure 4: Delays Under GPS

2. *If* $s_k < c_n$, $k = 1, \ldots, j-1$, $s_j \geq c_n$, *and* $S_n\left(\sum_{k=1}^{j-1} b_k\right) \leq e_n := (c_n\delta_n)/(c_n - \rho_n)$, *then* $D_n^* = \sum_{k=1}^{j-1} b_k - \eta$, *where* $\eta = S_n\left(\sum_{k=1}^{j-1} b_k\right) \Big/ c_n$.

The first part of the corollary follows by observing that $s_1 \geq c_n$ implies that $\overline{I}_n(\tau) \leq S_n(\tau)$ and therefore

$$\min_{t \geq \tau} \left\{ t : S_n(0,t) \geq \overline{I}_n(\tau) \right\} = \tau.$$

A geometric interpretation of the second part is given in Figure 4.

The development of GPS bounds for connection n is based on the Universal Service Curve (USC) for that connection [12, Section VIII]. Just as $S_n(t)$ characterizes the service that connection n receives at a single node, the USC of a connection characterizes the end to end service that it receives. We summarize here the method by which the USC is obtained when all the nodes use a GPS discipline. [12].

1. Under a CRST weight assignment, an algorithm is developed by which an envelope function, $\delta_n^m + \rho_n \tau$, is guaranteed for every connection n traffic entering node m [12, page 142]. For our purposes, it is important to note that,
$$\delta_n^1 = \delta_n, \quad \delta_n^m \geq \delta_n, \quad 2 \leq m \leq M. \tag{19}$$

2. Given the envelope functions $\delta_l^m + \rho_l \tau$, for all connections that are multiplexed with connection n on the same output link at node m, the service function for connection n, $S_n^m(\tau)$, is calculated. Let (s_k^m, b_k^m), $k = 1, \ldots, k_n^m$ be the set of slopes that characterize $S_n^m(\tau)$.

3. The USC, $\widehat{S}_n(\tau)$, for connection n is given by the formula

$$\widehat{S}_n(\tau) = \min\left\{G_n^M(\tau), \overline{I}(\tau)\right\},$$

where $G_n^M(\tau)$ is defined as infinity for $\tau > \sum_{m=1}^{M}\sum_{k=1}^{k_n^m} b_k^m$, and for $\tau \leq \sum_{m=1}^{M}\sum_{k=1}^{k_n^m} b_k^m$ it is composed of the segments (s_k^m, b_k^m), $m = 1, \ldots, M$, $k = 1, \ldots, k_n^m$ of $S_n^m(\tau)$, arranged in a nondecreasing order of slopes [12, page 144]. We denote by $(\widehat{s}_k, \widehat{b}_k)$, $k = 1, \ldots, \sum_{m=1}^{M} k_n^m$ this nondecreasing order.

Let k_q be such that

$$\widehat{s}_k < \rho_n,\ k = 1, \ldots, k_q - 1,\ \widehat{s}_{k_q} \geq \rho_n. \tag{20}$$

We are now ready to design an RCS discipline that is at least as good as GPS. Consider first the design of traffic shapers. Recall from the beginning of Section 4.1, that for the purpose of comparison with GPS we assume that the envelope of connection n traffic entering the first traffic shaper is of the form $\overline{I}_n(\tau) = \delta_n + \rho_n\tau$ (Recall, that we have assumed $L = 0$). For connection n, at each node m on the path, we choose traffic shapers that have the same envelope i.e. $\overline{A}_n^m(\tau) = \min\{c_n\tau,\ \delta_n + \rho_n\tau\}$. To specify how the parameter c_n is picked, we need to distinguish between two classes of connections.

1. **Class (a).** Connection n belongs to this class when,

$$\widehat{S}_n\left(\sum_{k=1}^{k_q-1}\widehat{b}_k\right) < \delta_n, \tag{21}$$

where the USC, $\widehat{S}_n$ is defined as above. In this case, the delay bound for connection n traffic under GPS is given by the solution of the equation [13, p. 136] (see Figure 5.i),

$$\overline{D}_n^* : \widehat{S}(\overline{D}_n^*) = \delta_n.$$

Let $k^* \geq k_q$, be the index of the slope of the USC at time $\overline{D}_n^*$. If at time $\overline{D}_n^*$ there is a change in slope, then define k^* as the index of the smaller of the two slopes (in fact either slope would work). We set $c_n = \widehat{s}_{k^*}$.

2. **Class (b).** Connection n belongs to this class when ,

$$\widehat{S}_n\left(\sum_{k=1}^{k_q-1}\widehat{b}_k\right) \geq \delta_n.$$

In this case, the delay bound for connection n traffic under GPS is [13, p. 136] (see Figure 6.i),

$$\overline{D}_n^* = \sum_{k=1}^{k_q-1}\widehat{b}_k - \frac{\widehat{S}_n\left(\sum_{k=1}^{k_q-1}\widehat{b}_k\right) - \delta_n}{\rho_n}.$$

We then set $c_n = \rho_n$.

For connection n, we assign the scheduler delay at node m, D_n^m, to be equal to the maximum delay that would be experienced by the connection under the GPS scheduling policy at that node, when the conditions of Corollary 1 are satisfied. This amounts to the following assignment.

- If at node m, $s_1^m \geq c_n$, then set $D_n^m = 0$.
- If at node m, $s_k^m < c_n$, $k = 1, \ldots, j_m - 1$, $s_{j_m}^m \geq c_n$, then assign

$$D_n^m = \sum_{k=1}^{j_m-1} b_k^m - \eta, \text{ where } \eta = S_n^m \left(\sum_{k=1}^{j_m-1} b_k^m \right) \Big/ c_n .$$

We first establish that the specified delays can be guaranteed by the EDF policy at each node. Instead of using the extension of Theorem 1 to general shaper envelopes, it will be simpler to argue indirectly as follows: we will show that the specified delays are guaranteed when the RCS discipline uses GPS as the scheduling policy at each node. Since EDF is better than GPS in the single node case, it will follow that the same delay guarantees can at a minimum be provided when the EDF scheduling policy is employed.

Observe that according to (19), we have that $\overline{A}_v(\tau) \leq \delta_v^m + \rho_v \tau$ for any connection v that is multiplexed with connection n on the same output link of node m . It is also true that $c_n \geq \rho_n$. This follows by definition for a connection in class (b). For a connection in class (a), observe that because of (20) and the fact that $\widehat{s}_k$, $k = 1, 2, \ldots$, is nondecreasing we have $c_n = \widehat{s}_{k^*} \geq \widehat{s}_{k_q} \geq \rho_n$. Applying Corollary 1 (where we replace $\widetilde{\delta}_n \leftarrow \delta_n^m$), we conclude that the delay $D_n^m = 0$ can be guaranteed under the GPS policy for any node m for which $s_1^m \geq c_n$. For a node m, where $s_k^m < c_n$, $k = 1, \ldots, j_m - 1$, $s_{j_m}^m \geq c_n$, we apply part 2 of Corollary 1 and, therefore, we first need to show that

$$S_n^m \left(\sum_{k=1}^{j_m-1} b_k^m \right) \leq \frac{c_n \delta_n}{c_n - \rho_n}.$$

This is trivially true for a connection in class (b) since $c_n \delta_n / (c_n - \rho_n) = \infty$. If connection n belongs to class (a), observe that from the definition of $\widehat{s}_{k^*}$, j_m and $\widehat{S}_n(\tau)$, we have (see Figure 5),

$$\begin{aligned} S_n^m \left(\sum_{k=1}^{j_m-1} b_k^m \right) &\leq \widehat{S}_n \left(\sum_{k=1}^{k^*-1} \widehat{b}_k \right) \\ &\leq \delta_n \\ &\leq \frac{\widehat{s}_{k^*} \delta_n}{\widehat{s}_{k^*} - \rho_n}. \end{aligned}$$

Thus, we have established that in both cases (a) and (b), the specified delay bound can be guaranteed at node m. Next, we need to establish that the end-to-end delay guarantee of the RCS discipline as given by (13), does not exceed $\overline{D}_n^*$. The delays $D(\mathcal{A}_m, \mathcal{A}_{m+1})$ are all zero since the traffic shapers are identical. Recall that the input traffic envelope for connection n, $\overline{I}_n(\tau) = \delta_n + \rho_n \tau$, and so from (8), the delay in the first traffic shaper is

$$D\left(\overline{I}_n,\ \mathcal{A}_1\right) = \frac{\delta_n}{c_n}.$$

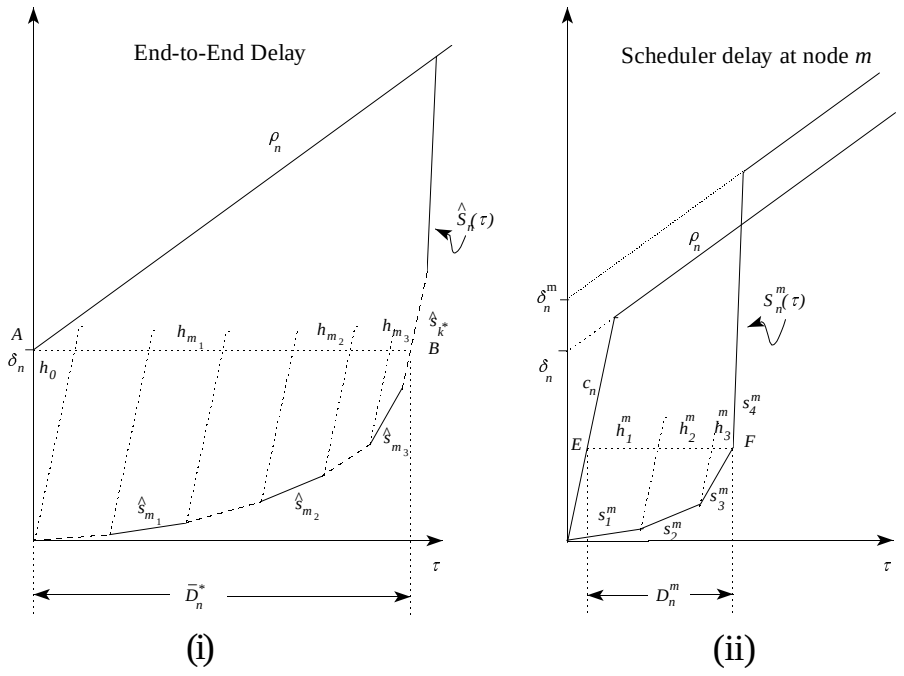

Figure 5: Delay Decomposition of a Class (a) Connection

Therefore, it suffices to show that

$$\frac{\delta_n}{c_n} + \sum_{m=1}^{M} D_n^m = \overline{D}_n^*. \tag{22}$$

Let M_0 be the set of nodes for which $D_n^m > 0$. Obviously then, $\sum_{m=1}^{M} D_n^m = \sum_{m \in M_0} D_n^m$. Assume first that connection n belongs to class (a). Observe that the set of slopes $\widehat{s}_k$, $k = 1, \ldots, k^* - 1$, can be partitioned into subsets F_m, $m \in M_0$, where

$$F_m = \left\{ \widehat{s}_l : \widehat{s}_l = s_k^m, \text{ for some } k = 1, \ldots, j_m - 1 \right\}.$$

We denote by m_k the index l for which $\widehat{s}_l = s_k^m$, i.e., $\widehat{s}_{m_k} = s_k^m$. For the rest of the discussion, it is best to use geometric arguments. Referring to Figure 5.i, draw lines with slope $\widehat{s}_{k^*}$ from all the points in $\widehat{S}_n(\tau)$ where the slope changes and remains less than $\widehat{s}_{k^*}$. These lines intersect segment AB (corresponding to the delay $\overline{D}_n^*$) and divide it into segments of length h_k, $0 \leq k \leq k^* - 1$, where segment h_k corresponds to slope $\widehat{s}_k$, $1 \leq k \leq k^* - 1$. Denote by h_{m_k} the segment that corresponds to $\widehat{s}_{m_k}$. Since by construction $h_0 = \delta_n / c_n$, we then have

$$\overline{D}_n^* = \frac{\delta_n}{c_n} + \sum_{m \in M_0} \sum_{k=1}^{j_m - 1} h_{m_k}. \tag{23}$$

Similarly, in Figure 5.ii, draw lines with slope $\widehat{s}_{k^*}$ from all the points in $S_n^m(\tau)$ where the slope changes and remains less than $\widehat{s}_{k^*}$. These lines intersect segment EF (corresponding to the delay D_n^m) and divide

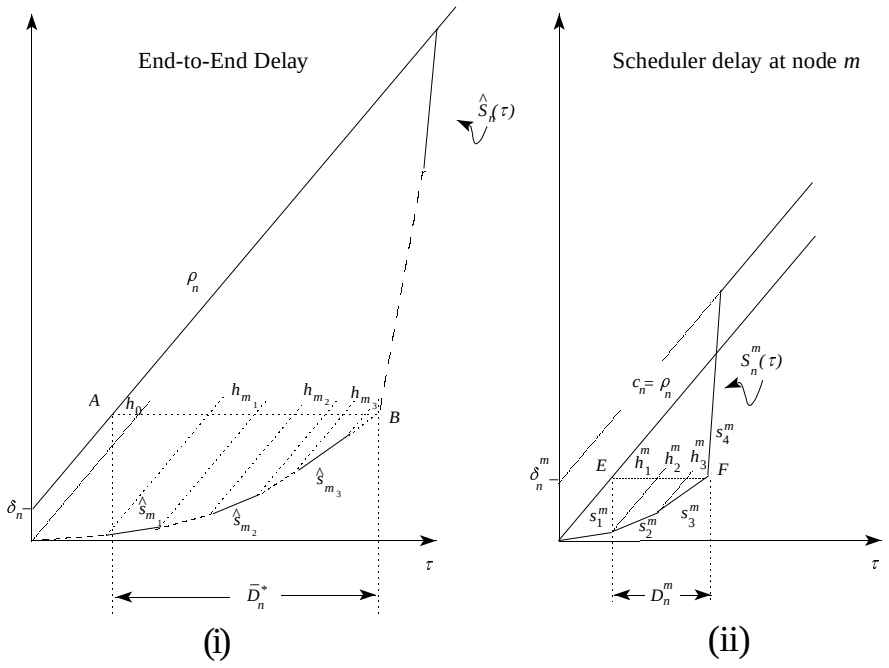

Figure 6: Delay Decomposition of a Class (b) Connection

it into segments h_k^m, $1 \leq k \leq j_m - 1$ (in the figure we have $j_m - 1 = 3$). We can then write,

$$D_n^m = \sum_{k=1}^{j_m - 1} h_k^m. \tag{24}$$

Using the facts that $\widehat{s}_{m_k} = s_k^m$ and that $\widehat{b}_{m_k} = b_k^m$, it can be easily seen that $h_{m_k} = h_k^m$. Taking into account (23) and (24), we conclude the correctness of (22).

Similar arguments can be made for a connection that belongs to class (b). The main difference is that we now draw lines with slope ρ_n. Figure 6 illustrates the construction in this case.

Note: In the course of the previous argument we showed that the delay guarantees provided by a pure GPS policy can also be achieved by an RCS discipline working with worst case delays at each node, where the scheduling policy at each node is GPS. If we replace GPS with the (simpler) EDF scheduling policy at each node, we are not only assured that we can still guarantee the GPS end-to-end delays, but we also create a service discipline that is better than GPS. This is due to the fact that in the single node case, EDF is better that GPS. That is, there are delay vectors that can be guaranteed by EDF but cannot be guaranteed by GPS no matter what weights are chosen. For example, consider a link of capacity r, where two connections are multiplexed and $\overline{I}_n(\tau) = \delta_n + \rho_n \tau$, $n = 1, 2$, with $\rho_1 + \rho_2 \leq r$. Using Theorem 1 with $L = 0$, we can see that the delays that can be guaranteed by EDF policy are

$$D_1 = \frac{\delta_1}{r}, \quad D_2 = \frac{\delta_2}{r - \rho_1} + \frac{\delta_1}{r},$$

For GPS on the other hand, it can be seen from the construction in [11, Section VI.C], that in order to guarantee $D_1^* = \delta_1/r$ we need to specify $\phi_2 = 0$, and then the minimum guaranteed delay for connection 2 is

$$D_2^* = \frac{\delta_2}{r-\rho_1} + \frac{\delta_1}{r-\rho_1}.$$

The difference between the GPS and EDF delay guarantees for connection 2 is

$$D_2^* - D_2 = \frac{\rho_1 \delta_1}{r(r-\rho_1)},$$

which can be quite large. Similar examples can be given for the packetized model when comparing PGPS to NPEDF. In this section, we have shown how "proper" selection of the traffic shapers allows us to construct an RCS discipline that outperforms GPS. In general, it is of interest to determine how the choice of shaper envelopes impacts the performance of rate controlled service disciplines. This is the topic of the next section.

Note that so far we have always used identical shapers at all nodes. Different shapers could however, be used at each node albeit at the cost of greater complexity. The question then is whether the use of different shapers affords sufficient benefits that compensate for the increase in complexity. We address this issue in the next section, together with an investigation of how shaper envelopes impact the performance of RCS disciplines.

5 Traffic Shaper Properties

In this section we discuss some interesting properties of traffic shapers in the context of RCS disciplines. Using these properties, the search for "good" traffic shapers can be significantly simplified.

Consider connection n that traverses nodes $1, \ldots, M$, of a network where an RCS discipline is used. Let $\mathcal{A}_n^m$ be the shaper for connection n at node m. We need the following simple but important observation.

Lemma 3. *If we replace $\mathcal{A}_n^m$ with a shaper $\mathcal{B}_n^m$ such that $\overline{B}_n^m(\tau) \leq \overline{A}_n^m(\tau)$, $\tau \geq 0$, then the scheduler delay D_k^m is still guaranteed for any connection k (including connection n).*

Proof. Observe that since $\overline{B}_n^m(\tau) \leq \overline{A}_n^m(\tau)$, $\tau \geq 0$, $\overline{A}_n^m(\tau)$ is also an envelope for the traffic exiting $\mathcal{B}_n^m$. By definition, D_k^m remains an upper bound on the delay of any connection k traffic as long as connection n still has envelope $\overline{A}_n^m(\tau)$. ■

We will show next (Proposition 2), that it suffices to restrict attention to RCS disciplines that for the same connection, use identical shapers at all nodes. We first need some notation. We write $\mathcal{A}_1 \succeq \mathcal{A}_2$ (or $\mathcal{A}_2 \preceq \mathcal{A}_1$) whenever $\overline{A}_1(\tau) \geq \overline{A}_2(\tau)$, $\tau \geq 0$. We denote by $\mathcal{A}_1 \wedge \mathcal{A}_2$ the arrangement of $\mathcal{A}_1$ and $\mathcal{A}_2$ in series. Since the output of shaper $\mathcal{A}_1$ has envelope $\overline{A}_1(\tau)$, it follows that

$$D\left(\mathcal{A}, \mathcal{A}_1 \wedge \mathcal{A}_2\right) \leq D\left(\mathcal{A}, \mathcal{A}_1\right) + D\left(\mathcal{A}_1, \mathcal{A}_2\right). \tag{25}$$

Also, observe that by (11),

$$\mathcal{A}_i \succeq \mathcal{A}_1 \wedge \mathcal{A}_2, \quad i = 1, 2. \tag{26}$$

Proposition 2. *Consider connection n that traverses nodes $1, 2, \ldots M$ and let $\overline{I}_n(\tau)$ be its envelope at the input to the first shaper. Given any RCS discipline π that uses shapers $\mathcal{A}_n^m$, and guarantees scheduler delays D_n^m, $1 \leq m \leq M$, the RCS discipline π' that uses the same scheduling policy at all nodes as π, and shapers*

$$\mathcal{B}_n^m = \mathcal{B} = \wedge_{m=1}^{M} \mathcal{A}_n^m,$$

can guarantee the same end-to-end delays as π, to all connections.

Proof. By (26) we have that $\mathcal{A}_n^m \succeq \mathcal{B}_n^m$ and, therefore, by Lemma 3, π' can guarantee the same scheduling delays to all connections. Since for any connection $k \neq n$, the shapers remain the same, this implies that policy π' guarantees the same end-to-end delays. Consider next connection n. Taking into account that $D\left(\mathcal{B}_n^m, \mathcal{B}_n^{m+1}\right) = D\left(\mathcal{B}, \mathcal{B}\right) = 0$, (13) implies that π' can provide the following end-to-end delay guarantees.

$$\overline{D}_n^{\pi'} = D\left(\overline{I}_n, \mathcal{B}\right) + \sum_{m=1}^{M} D_n^m + \sum_{m=1}^{M} T_n^{(m,m+1)}.$$

Finally observe that by (25)

$$D\left(\overline{I}_n, \mathcal{B}\right) \leq D\left(\overline{I}_n, \mathcal{A}_n^1\right) + \sum_{m=1}^{M-1} D\left(\mathcal{A}_n^m, \mathcal{A}_n^{m+1}\right).$$

Using (13) again we conclude that $\overline{D}_n^{\pi'} \leq \overline{D}_n^{\pi}$. ■

Note: According to Proposition 2 we can restrict our attention to disciplines that use identical shapers at all nodes. However, formula (13) can still be useful in a heterogeneous environment, where the various nodes are not designed to support identical shapers.

In the rest of this section, we consider disciplines that use identical shapers, i.e. $\mathcal{A}_n^m = \mathcal{A}_n$. Then, the end-to-end delay guarantee for connection n becomes

$$\overline{D}_n = D\left(\overline{I}_n, \mathcal{A}_n\right) + \sum_{m=1}^{M} D_n^m + \sum_{m=1}^{M} T_n^{(m,m+1)}. \tag{27}$$

We consider next, the problem of constructing the "smallest" shaper that causes a specified maximum delay on the input traffic $\overline{I}_n(\tau)$. Specifically, given $d \geq 0$, we want to construct a shaper $\mathcal{A}_n(d)$ such that $D\left(\overline{I}_n, \mathcal{A}_n(d)\right) \leq d$ and $\mathcal{A}_n(d) \preceq \mathcal{A}$, for any shaper $\mathcal{A}$, with $D\left(\overline{I}_n, \mathcal{A}\right) \leq d$. Recall that $\overline{U}_n(\tau)$ denotes the input traffic envelope of connection n before the first packetizer in the network (see Figure 1). We further assume that the input traffic envelope, $\overline{U}_n(\tau)$, is an increasing, concave, piecewise linear function with a finite number of slopes. In Appendix B, we show that these assumptions on the input traffic envelope do not entail any essential loss of generality. We can write $\overline{U}_n(\tau)$ in the form (see Figure 7)

$$\overline{U}_n(\tau) = \min_{k=1,\ldots,K} \left\{\delta_{n,k} + \rho_{n,k}\tau\right\}, \tag{28}$$

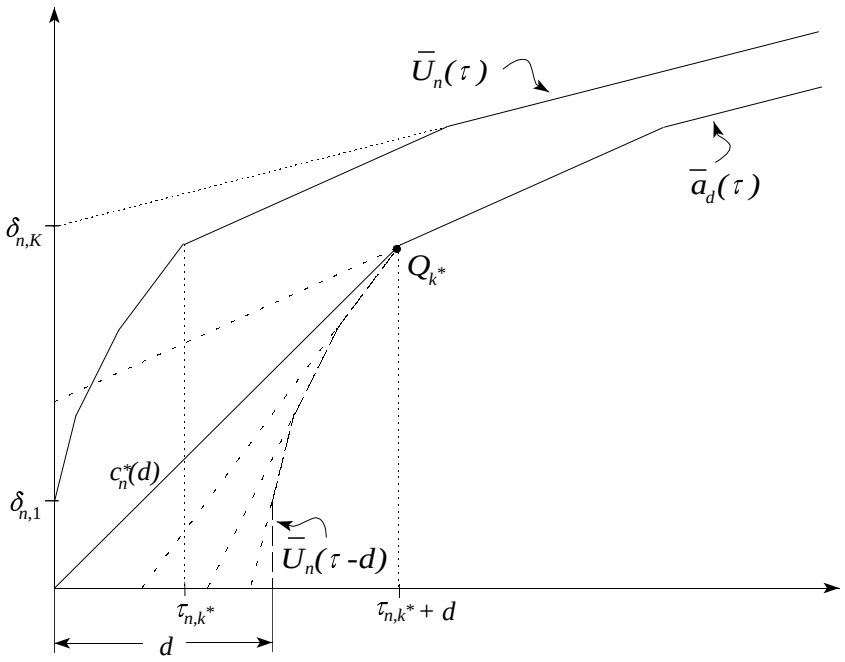

Figure 7: Construction of Smallest Envelope Function

where $\rho_{n,k} > \rho_{n,k+1}$, $\delta_{n,k} < \delta_{n,k+1}$ and when $K \geq 3$,

$$\frac{\delta_{n,k} - \delta_{n,k-1}}{\rho_{n,k-1} - \rho_{n,k}} < \frac{\delta_{n,k+1} - \delta_{n,k}}{\rho_{n,k} - \rho_{n,k+1}}, \quad k = 2, \ldots, K-1.$$

Let $\tau_{n,1} = 0$ and $\tau_{n,k} = (\delta_{n,k} - \delta_{n,k-1}) / (\rho_{n,k-1} - \rho_{n,k})$, $2 \leq k \leq K$. At the point $P_k = (\tau_{n,k}, \delta_{n,k} + \rho_{n,k}\tau_{n,k})$, the slope of the envelope, $\overline{U}_n(\tau)$, changes from $\rho_{n,k-1}$ to $\rho_{n,k}$. According to (1) the envelope of the traffic entering the first shaper is,

$$\overline{I}_n(\tau) = L + \min_{k=1,\ldots,K} \{\delta_{n,k} + \rho_{n,k}\tau\}.$$

Now, let $\overline{A}(\tau) = L + \min_{j=1,\ldots,J} \{\delta_j + \rho_j \tau\}$ be the envelope of $\mathcal{A}$. According to (7), $D\left(\overline{I}_n, \mathcal{A}\right) = \infty$ when $\min_{j=1,\ldots,J} \{\rho_j\} < \rho_{n,K}$, while $D\left(\overline{I}_n, \mathcal{A}\right) \leq \delta_{n,K}/\rho_{n,K}$ whenever $\min_{j=1,\ldots,J} \{\rho_j\} \geq \rho_{n,K}$. Therefore, it is sufficient to restrict our attention to the range $0 \leq d \leq \delta_{n,K}/\rho_{n,K}$. For the next proposition, it will be helpful to refer to Figure 7.

Proposition 3. *Let $0 \leq d \leq \delta_{n,K}/\rho_{n,K}$. Let k^* be the smallest index k such that the line with slope $\rho_{n,k}$ passing through the point $Q_k = \left(\tau_{n,k} + d, \overline{U}_n(\tau_{n,k})\right)$ intersects the y-axis at nonnegative values, i.e.,*

$$k^* = \min_{k=1,\ldots,K} \left\{k : \overline{U}_n(\tau_{n,k}) - \rho_{n,k}(\tau_{n,k} + d) \geq 0\right\}.$$

Then, the envelope of the smallest shaper $\mathcal{A}_n(d)$, such that $D\left(\overline{I}_n, \mathcal{A}_n(d)\right) \leq d$ is

$$\overline{A}_n(d)(\tau) = L + \overline{a}_d(\tau),$$

where, denoting $c_n^*(d) := \overline{U}_n(\tau_{n,k^*})/(\tau_{n,k^*} + d)$,

$$\overline{a}_d(\tau) = \begin{cases} c_n^*(d)\tau & \text{if} \quad 0 \leq \tau < \tau_{n,k^*} + d, \\ \overline{U}_n(\tau - d) & \text{if} \quad \tau \geq \tau_{n,k^*} + d. \end{cases}$$

Proof. Observe first that the index k^* always exists since

$$\begin{aligned} \overline{U}_n(\tau_{n,K}) - \rho_{n,K}(\tau_{n,K} + d) &= \delta_{n,K} + \rho_{n,K}\,\tau_{n,K} - \rho_{n,K}(\tau_{n,K} + d) \\ &= \delta_{n,K} - \rho_{n,K}\,d \;\geq\; 0. \end{aligned}$$

Next, we show that $\overline{A}_n(d)(\tau)$ corresponds to a shaper envelope function. For this, it is sufficient to show that $\overline{A}_n(d)(\tau)$ is a concave function, which will follow by construction, if we show that $c_n^*(d) \geq \rho_{n,k^*}$; but this is a consequence of the definition of k^*. To show that $D\left(\overline{I}_n, \mathcal{A}_n(d)\right) \leq d$, recall that according to (10) we can write

$$D\left(\overline{I}_n, \mathcal{A}_n(d)\right) = \max_{\tau \geq 0} \left\{ \left(\overline{A}_n^{(-1)}(d)\left(\overline{I}_n(\tau)\right) - \tau \right)^+ \right\},$$

and that by construction, $\overline{A}_n^{(-1)}(d)\left(\overline{I}_n(\tau)\right) - \tau \leq d$, for all $\tau \geq 0$. Finally, we need to show that $\overline{A}_n(d)(\tau) \leq \overline{A}(\tau)$ for the envelope of any other shaper $\mathcal{A}$ such that $D\left(\overline{I}_n, \mathcal{A}\right) \leq d$. To see this, observe that if $\overline{A}_n(d)(\tau) > \overline{A}(\tau)$ for some $\tau \geq \tau_{n,k^*} + d$, then $\overline{A}^{(-1)}\left(\overline{A}_n(d)(\tau)\right) > \tau$. Also, by construction, $\overline{A}_n(d)(\tau) = \overline{I}_n(\tau - d)$, $\tau \geq \tau_{n,k^*} + d$. From (10), for all $\tau \geq \tau_{n,k^*} + d$ we have,

$$\begin{aligned} D\left(\overline{I}_n, \mathcal{A}\right) &\geq \overline{A}^{(-1)}\left(\overline{I}_n(\tau - d)\right) - (\tau - d) \\ &> \tau - (\tau - d) = d, \end{aligned}$$

a contradiction. Therefore, $\overline{A}_n(d)(\tau) \leq \overline{A}(\tau)$ for all $\tau \geq \tau_{n,k^*} + d$. Using the inequalities $\overline{A}_n(d)(0) = L \leq \overline{A}(0)$, $\overline{A}_n(d)(\tau_{n,k^*} + d) \leq \overline{A}(\tau_{n,k^*} + d)$ and the concavity of $\overline{A}(\tau)$, we conclude that we also have $\overline{A}_n(d)(\tau) \leq \overline{A}(\tau)$ for $0 \leq \tau \leq \tau_{n,k^*} + d$ as well. ■

Using now Lemma 3 and Proposition 3, we easily conclude that given the input envelope function $\overline{I}_n(\tau)$, and a maximum shaper delay d, it is sufficient to restrict our attention to RCS disciplines that use shapers with envelopes of the form $\overline{A}_n(d)(\tau)$.

Corollary 2. *Given an RCS discipline that for connection n uses shaper $\mathcal{A}$ at all nodes, the RCS discipline that uses the shaper with envelope $\overline{A}_n(d)(\tau)$, where $d = D\left(\overline{I}_n, \mathcal{A}\right)$, can guarantee the same end-to-end delays to all connections.*

From the above discussion we see that given $\overline{I}_n(\tau)$, the search for the appropriate shaper envelope, is reduced to the one-parameter family $\overline{A}_n(d)(\tau)$. We can further constrain the range of the parameter d by taking into account the link speeds, r^m, along the path of connection n. This is done in the next proposition, where it is shown that it is sufficient to restrict attention to envelopes $\overline{A}_n(d)(\tau)$ whose maximum slope $c_n^*(d)$ (peak rate) is not larger than the minimum of r^m.

Proposition 4. *Consider connection n with input traffic envelope $\overline{U}_n(\tau)$ that traverses nodes $1, \ldots M$ with corresponding output link speeds r^m. Then, given an RCS discipline that uses shaper envelope $\overline{A}_n(d)(\tau)$, there is an RCS discipline using shaper envelope $\overline{A}_n(d')(\tau)$ with peak rate $c_n^*(d')$ such that*

$$\rho_{n,K} \leq c_n^*(d') \leq \min\left\{\min_{m=1,\ldots,M}\{r^m\}, c_n^*(d)\right\} \leq \min\left\{\min_{m=1,\ldots,M}\{r^m\}, c_n\right\},$$

where c_n is the peak rate of $\overline{U}_n(\tau)$, i.e., $c_n = \rho_{n,1}$ if $\delta_{n,1} = 0$ and $c_n = \infty$ otherwise, which guarantees the same end-to-end delays to all connections.

Proof. Observe first that by the design of $\overline{A}_n(d)(\tau)$, for all d, $0 \leq d \leq \delta_{n,K}/\rho_{n,K}$, we have $\rho_{n,K} \leq c_n^*(d)$ and

$$c_n^*(d) \leq c_n.$$

Denote by $U_n^{m+1}[t, t+\tau]$ the connection n traffic entering node $m+1$ in the interval $[t, t+\tau]$. Then, since the output link of node m has speed r^m,

$$U_n^{m+1}[t, t+\tau] \leq r^m \tau.$$

Therefore, for the traffic exiting the packetizer at node $m+1$, we have (see (1))

$$I_n^{m+1}[t, t+\tau] \leq B^m(\tau) = L + r^m \tau.$$

Therefore, we can replace $\mathcal{A}_n(d)$ with $\mathcal{B}^m \wedge \mathcal{A}_n(d)$, without altering the shaper delay or the scheduler delay at node $m+1$, $m = 1, \ldots M-1$. Also, by introducing a shaper with envelope $B^M(\tau)$ at the exit point of connection n, we do not affect the end-to-end delay guarantees. Using Proposition 2, we conclude that the delay guarantees are not affected if we employ the RCS discipline that uses shapers

$$\mathcal{B}_n^m = \mathcal{B}_n = \mathcal{A}_n(d) \wedge_{m=1}^{M} \mathcal{B}^m,$$

But then, for the peak rate c_n' of the envelope of shaper $\mathcal{B}_n$, we have

$$c_n' \leq \min\left\{\min_{m=1,\ldots,M}\{r^m\}, c_n^*(d)\right\} \leq \min\left\{\min_{m=1,\ldots,M}\{r^m\}, c_n\right\}.$$

Let $d' = D(\overline{I}_n, \mathcal{B}_n)$. Using Corollary 2, we can replace $\mathcal{B}_n$ with shaper $\mathcal{A}_n(d')$ without altering the delay guarantees for any connection. Since by design $\mathcal{A}_n(d') \preceq \mathcal{B}_n$, we must have $c_n^*(d') \leq c_n'$ and the proposition follows. ■

In the important special case of shapers used in the ATM standards [1] and those proposed for the Internet as well [15], we have

$$\overline{U}_n(\tau) = \min\left\{c_n\tau, \delta_n + \rho_n\tau\right\},\ c_n > \rho_n.$$

In this case, $\tau_{n,2} = \delta_n/\left(c_n - \rho_n\right)$, $k^* = 2$,

$$\overline{a}_d(\tau) = \begin{cases} \tau\left(c_n\delta_n\right)/(\delta_n + d(c_n - \rho_n)) & \text{if} \quad 0 \le \tau < \tau_{n,2} + d, \\ \delta_n + \rho_n(\tau - d) & \text{if} \quad \tau \ge \tau_{n,2} + d. \end{cases}$$

and the range of d is determined by the inequalities

$$\rho_n \le \frac{c_n\delta_n}{\delta_n + d(c_n - \rho_n)} \le \min\left\{\min_{m=1,\ldots,M}\left\{r^m\right\}, c_n\right\}.$$

Therefore, to specify an RCS discipline, one has to determine the single parameter d as well as the scheduler delays, D_n^m for each node m along the path of connection n. The determination of these parameters is an interesting design problem, which is the subject of ongoing research and is not addressed in this paper.

The use of traffic shapers at each hop can introduce extra delays for the traffic of connection n, even if there is no congestion in the network. While this leads to a reduction of jitter and buffer requirements at each node in the network, there may be instances where the resulting increase in the average delay is undesirable. In the next section we describe some simple modifications to the RCS discipline that make it work conserving, without compromising the end-to-end delay guarantees that can be provided.

6 Work Conserving System

Assume that the link scheduler used in the RCS discipline is by itself work conserving. However, if we consider the traffic shaper and the scheduler as a single system, it is evident that this system is not work-conserving since there may be instants in time when there are packets in the system even though the link is idle. In what follows we outline a modification to the system which will make it work conserving, while maintaining the same guaranteed deadlines for the accepted connections. As a result, the output link will no longer be idle when there are packets in the system thus improving the average delay seen by the packets. A similar approach and motivation can be found in [10] for a system where reshaping is performed based on timestamps carried in each packet, and in [6] for the AIRPORT policy proposed in that paper.

To clarify the exposition, we use the model of [10] to represent both the shaper and the scheduler at an output link of node m. Instead of holding up the packets in the shaper, we maintain two queues in the system: Q_e^m is a queue of packets that are eligible for scheduling, i.e., have been reshaped, and Q_γ^m is a queue of not yet eligible packets. Eligibility is determined by the shaper which stamps an eligibility time on the packets and enqueues them in Q_γ^m. The eligibility time is the earliest time the packet could have left the shaper, for the output of the shaper to be in conformance with the pre-specified traffic envelope. The delay of the packet in the scheduler is calculated based on its eligibility time. When a packet in Q_γ^m becomes eligible for scheduling, *viz.* its eligibility time equals the current time, it is promoted to Q_e^m. The

scheduler in the non-work conserving discipline, π_{NW}, only selects packets in Q_e^m for transmission on the output link. Once a packet has completed its transmission it is removed from Q_e^m and the scheduler repeats the above process.

Packets from each of the connections at node m enter Q_e^m in conformance with their respective traffic envelopes. The call admission criteria, ensures that packets in Q_e^m can be scheduled without violating their deadlines. Note that π_{NW} is non-work conserving since packets can be queued in Q_γ^m, but are not considered for transmission by the scheduler, even though the link may be idle.

We now develop a work-conserving discipline π_W, by modifying the scheduler in π_{NW} as follows. Whenever Q_e^m is empty, ineligible packets from Q_γ^m are transmitted (in any order). Next, we specify the operation of scheduler during periods when Q_e^m is non-empty. If Q_e^m is non-empty at time t, define a Q_e^m-*busy period* at t to be the largest closed interval containing t, during which Q_e^m is non-empty. Let t_0 denote the start of one such busy period. Note that at time t_0 it is possible that an ineligible packet is being served, in which case let t_s denote the time that the ineligible packet begins transmission; otherwise define $t_s := t_0$. Let q be the packet that begins transmission at time t_s. Consider the sequence of packet arrivals consisting of packet q, whose arrival time is set to t_s, along with the other packets that arrive to Q_e^m during the corresponding Q_e^m-busy period. Assume that this sequence is fed to the scheduler in π_{NW} with packet q being the first packet to ever arrive at that scheduler. The scheduler in π_W then schedules this sequence of packets in the same manner as the scheduler in π_{NW}. Note that if the scheduler in π_{NW} is NPEDF (FCFS, PGPS, Fixed-Priority scheduler, etc.) the corresponding scheduler in π_W is again NPEDF (FCFS, PGPS, Fixed-Priority scheduler, etc.) during a Q_e^m-busy period. The next proposition shows that the end-to-end delay guarantees are not affected when the service discipline at each node is modified to be work-conserving, as defined above.

Proposition 5. *Let connection n traverse nodes $1, \ldots, M$. The above modification to the service discipline does not increase the guaranteed upper bound on the end-to-end delays. If $\overline{D}_n$ is the end-to-end delay guarantee for connection n, we still have,*

$$\overline{D}_n \leq D\left(\overline{I}_n, \mathcal{A}_n^1\right) + \sum_{m=1}^{M-1} D\left(\mathcal{A}_n^m, \mathcal{A}_n^{m+1}\right) + \sum_{m=1}^{M} D_n^m + \sum_{m=1}^{M} T_n^{(m,m+1)}.$$

Proof. Assume for clarity in the exposition that the propagation delays, $T_n^{(m,m+1)}$, $1 \leq m \leq M$, are zero. We first establish that with the above modification to the service discipline, the scheduler delays at node m, $1 \leq m \leq M$, are still upper bounded by D_n^m .

Lemma 4. *Under discipline π_W, packets of any connection n, are not delayed by more than D_n^m at the scheduler in node m, $1 \leq m \leq M$.*

The proof of the above lemma can be found in Appendix A. We denote by $t_i^{l,m}$, the timestamp with which the ith packet is enqueued in Q_γ^m; $t_i^{l,m}$ is the time that the ith packet would leave shaper $\mathcal{A}_n^m$ in conformance with the traffic envelope $\overline{A}_n^m(\tau)$. The time at which the packet leaves Q_γ^m (to be transmitted

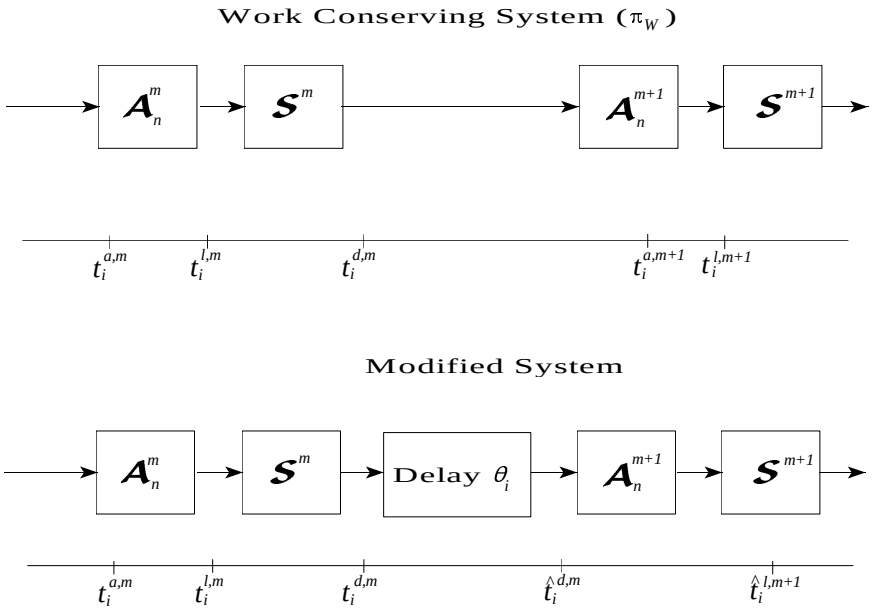

Figure 8: Original (work conserving) system and the modified system

on the link or promoted to Q_e^m) is denoted by $t_i^{a,m}$. If the link is idle, the packet may be transmitted before it becomes eligible, i.e., $t_i^{a,m} \leq t_i^{l,m}$. The departure time of the ith packet from the scheduler is denoted as $t_i^{d,m}$. Similarly, let $\tau_{i,a}$ be the arrival time of the ith packet of connection n to the first traffic shaper, and let $\tau_{i,d}$ be the time it arrives at its destination. Since $t_n^{a,m} \leq t_n^{l,m}$, $1 \leq m \leq M$, we can write

$$\begin{aligned} \tau_{i,d} - \tau_{i,a} &\leq t_n^{l,M} - \tau_{i,a} + \tau_{i,d} - t_n^{a,M} \\ &= \sum_{m=1}^{M-1} \left(t_n^{l,m+1} - t_n^{l,m}\right) + t_n^{l,1} - \tau_{i,a} + \tau_{i,d} - t_n^{a,M}. \end{aligned}$$

Since $\tau_{i,d} - t_n^{a,M} \leq D_n^M$ by Lemma 4, and $t_n^{l,1} - \tau_{i,a} \leq D\left(\overline{I}_n, \mathcal{A}_n^1\right)$ by definition, it suffices to show that for $1 \leq m \leq M-1$,

$$t_i^{l,m+1} - t_i^{l,m} \leq D_n^m + D(\mathcal{A}_n^m, \mathcal{A}_n^{m+1}). \tag{29}$$

Let $\mathcal{S}_m$ be the system consisting of the scheduler at node m. Consider the modified system which is same as the work conserving system operating under π_W except for a delay system inserted between $\mathcal{S}_m$ and shaper $\mathcal{A}_n^{m+1}$ as shown in Figure 8. The delay system delays packet i by $\theta_i = D_n^m + t_i^{l,m} - t_i^{d,m}$; therefore, packet i departs the delay system at time $\widehat{t}_i^{d,m} = D_n^m + t_i^{l,m}$. First we verify that $\theta_i \geq 0$,

$$\theta_i = D_n^m + t_i^{l,m} - t_i^{d,m}$$

$$\geq \quad D_n^m + t_i^{a,m} - t_i^{d,m} \tag{30}$$

$$\geq \quad 0. \tag{31}$$

Inequality (30) follows because packets never depart the shaper later than they are supposed to, i.e., $t_i^{l,m} \geq t_i^{a,m}$, and (31) follows from Lemma 4. Let $\widehat{t}_i^{l,m+1}$ be the timestamp with which packet i is enqueued in Q_γ^{m+1} in the modified system. From Lemma 1, we conclude that

$$t_i^{l,m+1} - t_i^{d,m} \leq \widehat{t}_i^{l,m+1} - t_i^{d,m}. \tag{32}$$

Adding $t_i^{d,m} - t_i^{l,m}$ to both sides of (32) we have

$$\begin{aligned} t_i^{l,m+1} - t_i^{l,m} &\leq \widehat{t}_i^{l,m+1} - t_i^{l,m} \\ &= \widehat{t}_i^{l,m+1} - \widehat{t}_i^{d,m} + \widehat{t}_i^{d,m} - t_i^{l,m}. \end{aligned}$$

Since for all i, $\widehat{t}_i^{d,m} = t_i^{l,m} + D_n^m$, the traffic exiting the delay system has envelope $\overline{A}_n^m(\tau)$. Therefore,

$$\widehat{t}_i^{l,m+1} - \widehat{t}_i^{d,m} \leq D(\mathcal{A}_n^m, \mathcal{A}_n^{m+1}).$$

It follows that

$$t_i^{l,m+1} - t_i^{l,m} \leq D(\mathcal{A}_n^m, \mathcal{A}_n^{m+1}) + D_n^m$$

as desired. When $T_n^{(m,m+1)} > 0$, the same reasoning applies, provided that system $\mathcal{S}_m$ consists of the scheduler at node m, and the link $(m, m+1)$, i.e., the bound on the delay at $\mathcal{S}_m$ is now $D_n^m + T_n^{(m,m+1)}$. ■

7 Conclusions

In this paper, we have established that RCS disciplines offer a powerful solution to provide end-to-end delay and throughput guarantees in high speed networks. We showed that the main disadvantage of these service disciplines, namely that of summing worst case delays at each node to determine end-to-end delay bounds, can be overcome through "proper" reshaping of the source traffic. In particular, we have shown that controlling the peak rate of a connection as a function of its delay requirements is critical to efficient network QoS provisioning. How to perform this reshaping was also investigated in the paper, and illustrated by designing RCS disciplines that outperform GPS. This is significant since the bounds available for these policies take dependencies between nodes into account.

In addition to their efficiency, RCS disciplines are also relatively simple to implement, and offer the flexibility to accommodate a wide range of implementation constraints. For example, it is possible to use different schedulers and shapers at different nodes depending on the capabilities available locally. Furthermore, because we also showed that guarantees are not affected when operating in a work conserving manner, i.e., reshaping traffic only in case of congestion, RCS disciplines also enable us to offer low average delays when the network is not congested. Finally, note that the greater flexibility of RCS disciplines also introduces new and interesting problems, e.g., how to best split a given end-to-end delay budget into local delay bounds, and addressing them is the topic of ongoing work.

Acknowledgements

We are indebted to Raju Rajan for suggesting the proof to Lemma 1 as well as numerous insightful comments on early versions of this paper. We would like to thank Abhay Parekh for many fruitful discussions regarding the comparative performance of GPS and EDF, and Subir Varma for prompting us to look into the issue of the efficiency of RCS disciplines as well as for many comments which helped improve the paper. We would also like to thank Armand Makowski for many helpful discussions and comments on an earlier draft.

Appendix

A Lemma Proofs

Proof of Lemma 1. Let $\overline{A}(\tau)$ denote the envelope of the shaper with $\overline{A}(0) \geq L$. Also, let $s_i^{(1)}$, $s_i^{(2)}$ denote the arrival times, and $f_i^{(1)}$, $f_i^{(2)}$ denote the departure times of the ith packet at the shapers of system $\mathcal{S}_1$ and $\mathcal{S}_2$ respectively (see Figure 2). By definition, $s_i^{(2)} = s_i^{(1)} + \theta_i$, with $\theta_i \geq 0$, and therefore it suffices to show that

$$f_i^{(1)} \leq f_i^{(2)}, \quad i = 1, 2, \ldots$$

Since $\overline{A}(0) \geq L$, the first packet leaves the shaper instantaneously in both the systems, i.e. $f_1^{(1)} = s_1^{(1)}$ and $f_1^{(2)} = s_1^{(2)}$. Therefore, we have $f_1^{(1)} \leq f_1^{(2)}$.

Let l_i denote the length of the ith packet, and assume that

$$f_i^{(1)} \leq f_i^{(2)}, \quad i = 1, 2, \ldots, m. \tag{33}$$

From (2) we can compute the departure time for the $(m+1)$th packet in system $\mathcal{S}_1$ as,

$$\begin{aligned}
f_{m+1}^{(1)} &= \min\{t \geq s_{m+1}^{(1)} : \overline{A}(t - f_i^{(1)}) \geq \sum_{k=i}^{m+1} l_k,\ i = 1, 2, \ldots m\} \\
&\leq \min\{t \geq s_{m+1}^{(2)} : \overline{A}(t - f_i^{(1)}) \geq \sum_{k=i}^{m+1} l_k,\ i = 1, 2, \ldots m\} \qquad (34) \\
&\leq \min\{t \geq s_{m+1}^{(2)} : \overline{A}(t - f_i^{(2)}) \geq \sum_{k=i}^{m+1} l_k,\ i = 1, 2, \ldots m\} \qquad (35) \\
&= f_{m+1}^{(2)},
\end{aligned}$$

where (34) is because $s_{m+1}^{(2)} \geq s_{m+1}^{(1)}$, and (35) follows from the non-decreasing nature of the the shaper envelope, $\overline{A}(\tau)$, and the induction hypothesis, (33). ■

Proof of Lemma 2. The busy period containing t is defined as the largest closed interval containing t, during which the backlog of connection n is positive. Note that only traffic arriving during a busy period can have positive delays and, therefore, we only need to consider the delays of such traffic (since by definition $D_n^* \geq 0$).

Assume that connection n traffic arrives at time t_b which is within a busy period of connection n that starts at time t_0. Then, since the order of service within a connection is FCFS, the delay of the connection n traffic arriving at time t_b is given by,

$$\begin{aligned} d(t_b) &= \min_{t' \geq t_b} \{t' : S_n[0,t'] \geq I_n[0,t_b]\} - t_b \\ &= \min_{t' \geq t_b} \{t' : S_n[0,t_0) + S_n[t_0,t'] \geq I_n[0,t_0) + I_n[t_0,t_b]\} - t_b \\ &= \min_{t' \geq t_b} \{t' : S_n[t_0,t'] \geq I_n[t_0,t_b]\} - t_0 - (t_b - t_0). \end{aligned}$$

The last equality follows from the fact that since t_0 is the start of a connection n busy period, $S_n[0,t_0) = I_n[0,t_0)$. Setting $t = t' - t_0$ and $\tau = t_b - t_0$, and observing that

$$\min_{t' \geq t_b} \{t' : S_n[t_0,t'] \geq I_n[t_0,t_b]\} = \min_{t \geq \tau} \{t : S_n[t_0,t+t_0] \geq I_n[t_0,t_b]\} + t_0,$$

we then get

$$d(t_b) = \min_{t \geq \tau} \{t : S_n[t_0,t+t_0] \geq I_n[t_0,t_b]\} - \tau.$$

Now, since the connection n traffic satisfies $I[t,t+\tau] \leq \delta_n + \rho_n \tau, \;\; t,\tau \geq 0$, from Lemma 10 in [11], we conclude that

$$S_n[t_0,t+t_0] \geq S_n(t).$$

Since by definition we also have

$$I_n[t_0,t_b] \leq \overline{I}_n(t_b - t_0),$$

it follows that

$$\{t : S_n[t_0,t+t_0] \geq I_n[t_0,t_b]\} \supseteq \left\{t : S_n(t) \geq \overline{I}_n(t_b - t_0)\right\},$$

and therefore,

$$\min_{t \geq \tau} \{t : S_n[t_0,t+t_0] \geq I_n[t_0,t_b]\} \leq \min_{t \geq \tau} \left\{t : S_n(t) \geq \overline{I}_n(t_b - t_0)\right\}.$$

Recalling that $\tau = t_b - t_0$, we finally get

$$d(t_b) \leq \min_{t \geq \tau} \left\{t : S_n(t) \geq \overline{I}_n(\tau)\right\} - \tau.$$

■

Proof of Lemma 4. We concentrate on the system operating according to π_W as defined in Section 6 and repeat some notation for the sake of clarity. We denote by $t_i^{l,m}$, the timestamp with which the ith packet is enqueued in Q_γ^m; recall that $t_i^{l,m}$ is the time that the ith packet would leave shaper $\mathcal{A}_n^m$ in conformance with the traffic envelope $\overline{A}_n^m(\tau)$. The time at which the packet leaves Q_γ^m (to be transmitted on the link or promoted to Q_e^m) is denoted as $t_i^{a,m}$ and we say that the packet *arrives at the scheduler* at time $t_i^{a,m}$. If the link is idle, the packet may be transmitted before it becomes eligible, i.e., $t_i^{a,m} \leq t_i^{l,m}$. The departure time of the ith packet from the scheduler is denoted as $t_i^{d,m}$. We need to show that for any packet i,

$$t_i^{d,m} - t_i^{a,m} \leq D_n^m$$

D_n^m has to be larger than the time it takes to transmit a complete packet, and so packets that are scheduled before they become eligible can never miss their deadline. All the eligible packets are scheduled in Q_e^m-busy periods, and so it suffices to show that packets that enter Q_e^m are never delayed by more than D_n^m.

Let $[t_0, t_f]$ be a Q_e^m-busy period. At time t_0 a packet either starts transmission or is in the process of being transmitted and recall that $t_s \leq t_0$ is the time that this packet starts transmission. If $t_s = t_0$ (in this case a packet from Q_e^m starts transmission at t_0), then the traffic of all connections arriving to the scheduler in $[t_s, t_f]$ are conformant to their respective traffic shapers. In addition, in the interval $[t_s, t_f]$ the operation of the scheduler in π_W is identical to the one in π_{NW} if t_s were the start of the first busy period. Thus, for $t_s = t_0$ the result is true by the definition of D_n^m.

Now assume that $t_s < t_0$, i.e., an ineligible packet from connection j starts transmission at t_s. Observe that in a busy period $[t_0, t_f]$, the scheduler in π_W only schedules packets that are in Q_e^m, except for the packet that is being transmitted at the start of the busy period. We will show next that the packets of all connection that have been transmitted in the interval $[t_s, t_f]$ have arrived to the scheduler in conformance with their respective traffic envelopes. The truth of the lemma will then follow as before.

Recall that $A_n^m[t_1, t_2]$ denotes the traffic from connection n that is promoted to Q_e^m in the interval $[t_1, t_2]$. Let $\widehat{A}_n^m[t_1, t_2]$ denote the connection n traffic that arrives at the scheduler in the interval $[t_1, t_2]$. We need to show that

$$\widehat{A}_n^m[t_1, t_2] \leq \overline{A}_n^m(t_2 - t_1), \quad t_s \leq t_1 \leq t_2 \leq t_f.$$

Since we are only concerned with node m here, we drop the superscript m for the sake of clarity. By the definition of t_0, we have that for any connection n,

$$A_n[t_1, t_2] \leq \overline{A}_n(t_2 - t_1), \quad t_0 \leq t_1 \leq t_2 \leq t_f.$$

For a connection $n \neq j$, we have in addition, $A_n[t_s, t_0) = 0$ and therefore,

$$A_n[t_1, t_2] \leq \overline{A}_n(t_2 - t_1), \quad t_s \leq t_1 \leq t_2 \leq t_f, \; n \neq j. \tag{36}$$

Note also that by definition $\widehat{A}_n[t_1, t_2] = A_n[t_1, t_2]$, $t_s \leq t_1 \leq t_2 \leq t_f$, holds for $n \neq j$ since connection j is the only connection which transmits an ineligible packet in $[t_s, t_f]$.

Consider next connection j, and let p_j be the packet that starts transmission at t_s. Let τ_e denote the eligibility time of packet p_j. If $\tau_e \geq t_f$, then clearly $\hat{A}_j[t_1, t_2] \leq L \leq \overline{A}_j(0)$, $t_s \leq t_1 \leq t_2 \leq t_f$, since no more packets from connection j will be transmitted in $[t_s, t_f]$. Now suppose $t_s \leq \tau_e < t_f$. Then, all other packets of connection j will arrive after τ_e. For the case when $t_s \leq t_1 \leq \tau_e$ and $\tau_e \leq t_2 \leq t_f$, we have

$$\hat{A}_j[t_1, t_2] \leq \overline{A}_j(t_2 - \tau_e) \leq \overline{A}_j(t_2 - t_1).$$

The other cases can be similarly checked. ■

B Subadditive Traffic Envelopes

In Proposition 3, we presented a method to obtain the traffic shaper which given an input envelope and a shaper delay has minimal envelope. There, we made the assumption that the input traffic envelope $\overline{I}_n(\tau)$ is a concave, increasing, piecewise linear function with finite number of slopes. Using such functions, we can approximate arbitrarily closely any concave increasing envelope. This means that by using the construction in Proposition 2 we can construct shapers that, for given concave increasing input envelope and shaper delay, are arbitrarily close to optimal. However, a general input envelope satisfies a weaker property than concavity, namely subadditivity [2]. In this appendix we outline how the method of Proposition 2 can be applied to subadditive input envelopes as well.

Let us consider a nondecreasing, piecewise linear function, $\overline{I}_n(\tau)$, with finite number of slopes and such that $\lim_{\tau \to \infty} \overline{I}_n(\tau)/\tau > 0$. Such functions can approximate arbitrarily closely any nondecreasing function in an appropriate sense (using the Skorohod metric [14, Chapter VI]). Let $\widehat{I}_n(\tau)$ be the minimal concave function such that $\overline{I}_n(\tau) \leq \widehat{I}_n(\tau)$, $\tau \geq 0$. It can be seen that $\widehat{I}_n(\tau)$ is increasing, piecewise linear with finite number of slopes and, therefore, can be written in the form of (28), $\widehat{I}_n(\tau) = \min_{k=1,\cdots,K} \left\{ \widehat{\delta}_{n,k} + \widehat{\rho}_{n,k}\tau \right\}$. Given d such that $0 \leq d \leq \widehat{\delta}_{n,K}/\widehat{\rho}_{n,K}$, we construct the minimal envelope $\widehat{A}_n(d)(\tau)$ corresponding to $\widehat{I}_n(\tau)$ using Proposition 3. We claim that $\widehat{A}_n(d)(\tau)$ is also the envelope of the minimal shaper $\mathcal{A}_n(d)$ which provides delay bound d to input traffic with envelope $\overline{I}_n(\tau)$. To see this, consider a shaper $\mathcal{A}$ with envelope $\overline{A}(\tau)$ such that $D(\overline{I}_n, \mathcal{A}) \leq d$, $\overline{I}_n(0) \geq L$, and assume for the moment that $\overline{A}(\tau_0) < \widehat{A}_n(d)(\tau_0)$, for some $\tau_0 \geq 0$. Since $\overline{A}(\tau)$ is concave and $\widehat{I}_n(\tau)$ is the minimal concave function such that $\overline{I}_n(\tau) \leq \widehat{I}_n(\tau)$, $\tau \geq 0$, it can be seen that $D(\widehat{I}_n, \mathcal{A}) \leq d$. Since by design we also have $D(\widehat{I}_n, \mathcal{A}_n(d)) = d$, we conclude from (8) that

$$\begin{aligned} D\left(\widehat{I}_n, \mathcal{A}_n(d) \wedge \mathcal{A}\right) &= \max\left\{ D\left(\widehat{I}_n, \mathcal{A}_n(d)\right),\ D(\widehat{I}_n, \mathcal{A}) \right\} \\ &\leq d. \end{aligned}$$

But the shaper $\mathcal{A}_n(\tau) \wedge \mathcal{A}$ has envelope $\overline{B}(\tau) = \min\left\{ \widehat{A}_n(d)(\tau),\ \overline{A}(\tau) \right\}$. Since $\overline{A}(\tau_0) < \widehat{A}_n(d)(\tau_0)$, the envelope $\overline{B}(\tau)$ is strictly smaller than $\widehat{A}_n(d)(\tau)$, which contradicts the optimality of $\mathcal{A}_n(d)$.

References

[1] *ATM UNI Specification, Version 3.1,* ATM Forum, September 1994.

[2] C.-S. Chang. Stability, queue length and delay of deterministic and stochastic queueing networks. *IEEE Transactions on Automatic Control*, 39(5):913–931, May 1994.

[3] D. Clark, S. Shenker, and L. Zhang. Supporting real-time applications in an integrated packet network: Architecture and mechanisms. In *Proceedings of ACM SIGCOMM'92*, Baltimore, MD, August 1992.

[4] R. L. Cruz. A calculus for network delay, part I: Network elements in isolation. *IEEE Transactions on Information Theory*, 37:114–131, January 1991.

[5] R. L. Cruz. A calculus for network delay, part II: Network analysis. *IEEE Transactions on Information Theory*, 37:132–141, January 1991.

[6] R. L. Cruz. Service burstiness and dynamic burstiness measures: A framework. *Journal of High Speed Networks*, 1(2):105–127, 1992.

[7] A. Demers, S. Keshav, and S. Shenker. Analysis and simulation of a fair queueing algorithm. *Journal of Internetworking: Research and Experience*, 1:3–26, January 1990.

[8] L. Georgiadis, R. Guérin, and A. Parekh. Optimal multiplexing on a single link: Delay and buffer requirements. Research Report RC 19711 (97393), IBM, T. J. Watson Research Center, August 1994. Short version appeared in *Proceedings of INFOCOM'94*.

[9] P. Goyal and H. Vin. Generalized guaranteed rate scheduling algorithms: A framework. Technical Report TR-95-30, Department of Computer Sciences, University of Texas at Austin, 1995.

[10] D. D. Kandlur, K. G. Shin, and D. Ferrari. Real-time communication in multi-hop networks. In *Proceedings of INFOCOM'91*, pages 300–307, 1991.

[11] 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 Transactions on Networking*, 1(3):344–357, June 1993.

[12] A. K. Parekh and R. G. Gallager. A generalized processor sharing approach to flow control in integrated services networks: The multiple node case. *IEEE/ACM Transactions on Networking*, 2(2):137–150, April 1994.

[13] A. K. J. Parekh. *A Generalized Processor Sharing Approach to Flow Control in Integrated Services Networks*. PhD thesis, Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, Cambridge, MA 02139, February 1992. No. LIDS-TH-2089.

[14] D. Pollard. *Convergence of Stochastic Processes*. Springer-Verlag, 1984.

[15] S. Shenker and C. Partridge. Specification of guaranteed quality of service. Internet Draft, `draft-ietf-intserv-guaranteed-svc-04.txt`, Integrated Services WG, IETF, 1995.

[16] H. Zhang and D. Ferrari. Rate-controlled service disciplines. *Journal of High Speed Networks*, 3(4):389–412, 1994.

[17] Q. Zheng and K. Shin. On the ability of establishing real-time channels in point-to-point packet-switching networks. *IEEE Transactions on Communications*, 42(3):1096–1105, March 1994.