Optimal Flow Control Schemes that Regulate the Burstiness of Traffic

Takis Konstantopoulos, *Member, IEEE,* and Venkat Anantharam, *Member, IEEE*

***Abstract*—The problem of designing burst reducing flow controllers for traffic in an ATM network is studied. By requiring that the output flow obey certain burstiness constraints, it is shown that an optimal design exists and that it can be easily implemented in real time. Two versions of the problem are considered. The first one places constraints on the buffer size and the second one on the maximum delay that a cell can experience. Both problems are solved for arbitrary traffic processes. To treat the problems in this generality we introduce reflection mappings and use them, in a rather novel way, to establish optimality results. As a by-product of our analysis and methods, the optimality of the popular leaky bucket flow control scheme is also established.**

I. INTRODUCTION

THERE is a need to rethink basic flow control and resource allocation issues to deal with the novel kinds of traffic encountered in a Broadband Integrated Services Digital Network (B-ISDN). This includes traffic with widely differing characteristics, such as audio, data and video. Some of this traffic is expected to be of a highly bursty nature, making the traditionally employed models, such as Poisson or renewal processes, of doubtful relevance in deriving design insights. Several analyses of samples of traffic have pointed to the deficiency of traditional traffic models in the new applications context; see, for instance, Leland *et al.* [12], and Jagerman and Melamed [10]. There is a serious need to incorporate explicit burstiness modeling in the traffic models used for performance analysis.

A simple and versatile class of models was introduced by Cruz [5], and has been further developed by Low and Varaiya, [13], [14]. Let A be a traffic process. Following Cruz [5], we say that A obeys (σ, ρ) constraints if there are positive constants σ, ρ, such that

$$A[0,t] \leq \sigma + \rho t, \text{ for all } t \geq 0$$
$$A(s,t] \leq \sigma + \rho(t-s), \text{ for all } t \geq s \geq 0$$

where $A(s,t]$ denotes the amount of traffic over the time interval $(s,t]$. Intuitively, ρ poses a constraint on the long-term average rate and σ on the instantaneous bursts. A more precise characterization is that A obeys (σ, ρ) constraints iff when A is fed to a server with service rate ρ that is started empty at time s and has an infinite buffer, the backlog in the buffer never exceeds σ at any time. There are a number of reasons why this class of flows is useful to consider. First of all, the above is an intuitively appealing characterization of burstiness, measured by the backlog induced in a simple queueing system, and as such is amenable to analysis and so to intelligent design choices. Second, it is a very simple characterization. Users may have to negotiate the parameters of the flows that they offer the network; the parameters σ and ρ are a simple pair with which to work, representing burstiness and average offered load.

Our focus in this paper is on the issue of flow control to ensure that a flow obeys burstiness constraints in the above sense. We examine the question of how a user, constrained to offer a flow obeying (σ, ρ) constraints can optimally control his source to ensure this requirement. Our key result is that there is a very simple answer to this question, involving recursively updating a quantity, which, for precision, we call the *virtual backlog*. This quantity can also be shown to work as a sufficient statistic to study various natural allocation problems within a B-ISDN network, as shown in [1].

One should mention that the leading contender for a standard for implementing the B-ISDN concept is asynchronous transfer mode (ATM). In ATM traffic is segmented into fixed length cells (the standard calls for cells of length 53 bytes) which are routed through the network along virtual circuits. One of the most popular flow control schemes for traffic consisting of ATM cells is the "leaky bucket" flow control scheme; see, for example, Eckberg and Lucantoni [8]. This scheme has been shown in practice to be very effective in regulating burstiness and is also simple to implement. In [2] we analyzed a model of the leaky bucket, driven by a stationary and ergodic arrival point process, and showed that it is burst-reducing, in the sense that the delay induced by its output is stochastically reduced. A similar result was also obtained independently by Low and Varaiya [13] and by Kuang [11]. These results suggest that the leaky bucket is a good flow

Manuscript received January 12, 1994; revised January 9, 1995; approved by IEEE/ACM TRANSACTIONS ON NETWORKING Editor D. Towsley. The work of T. Konstantopoulos was supported in part by the NSF under Grant NCR 92-11343. The work of V. Anantharam was supported in part by the NSF under Grant NCR 88-57731.

T. Konstantopoulos is with the Department of Electrical and Computer Engineering, University of Texas at Austin, Austin, TX 78712 USA (e-mail: takis@alea.ece.utexas.edu).

V. Anantharam is with the Department of Electrical Engineering and Computer Science, University of California at Berkeley, Berkeley, CA 94720 USA (e-mail: ananth@eecs.berkeley.edu).

IEEE Log Number 9413393.

control scheme. It is interesting that an examination of the protocol shows that the output flow from the leaky bucket obeys (σ, ρ) constraints.

Previous work on flow control for high-speed networks has been similar in that it has focused mainly on the performance evaluation and comparison between different proposed schemes. In this paper we take a harder look at the problem of designing an optimal flow control scheme. We pose the design problem as follows: we are given a target burstiness parameter pair (σ, ρ). The flow controller is allowed to delay incoming cells by at most d time units, $d \geq 0$. At each point of time the flow controller can either reject or transmit part of the delayed cells or the incoming traffic. The decision of the controller (whether and how much to transmit or reject) at each point of time t is allowed, in principle, to depend on the *entire past* of the input process up to time t. The only requirement is that the flow control scheme should create an output traffic stream that must satisfy the (σ, ρ) constraints. Every policy that obeys the above requirements is called a delay-constrained feasible policy. The goal is to find the best delay-constrained feasible policy in the sense that the rejected traffic is as small as possible. It turns out that there is an optimal policy that is rather simple in nature and can be implemented by recursively updating the quantity we call the virtual backlog. It appears to be very easy to implement in real time.

A second design problem can be posed as follows. Besides the target burstiness parameter pair (σ, ρ), we are also given a bound K on the maximum number of cells that can be stored in the flow controller's memory. The goal is as above, namely, the design of a scheme that obeys the above requirements (buffer-constrained flow controller) and which rejects the smallest possible amount of traffic. Once again, it turns out that there is an optimal policy that can be easily implemented by recursively updating a virtual backlog. Further, this policy is essentially the leaky bucket scheme.

The paper is organized as follows. In Section II we consider flow control for arbitrary traffic process models. In particular, fluid models as well as point processes are covered by our analysis in a unified way. This unified approach is based on the concept and properties of the reflection mapping with (possibly) time-varying boundaries. These are summarized in Appendix A. We first solve the problem of optimal instantaneous flow control (Section II-A). Subsequently we consider buffer-constrained flow controllers (Section II-B). Finally we deal with the case of delay-constrained flow controllers (Section II-C). The topics are organized in increasing order of complexity. In Section III we show how, from the results of Section II, to obtain the optimal flow controllers in discrete time. In particular, we re-discover a simplified algorithm for the results of [3] concerning the discrete time optimal flow controller with delay constraints (Section III-B). The algorithm for the optimal discrete time flow controller with buffer constraints is also given (Section III-A). Appendix B summarizes the formulas for discrete time reflection mappings used to obtain the results of Section III. In Section IV we establish the optimality of the Leaky Bucket flow control scheme. Briefly, not only does the leaky bucket produce a (σ, ρ) constrained output but it also does so in an optimal way. This result is, to the best of our knowledge, the first one concerning the optimality of the Leaky Bucket. Finally, some concluding remarks are presented in Section V.

II. Flow Control for General Traffic Processes

The system operates in continuous or in slotted time. We consider optimal flow control problems for general traffic processes. We shall allow processes that are fluid-like, that is, messages arrive at an arbitrary time-varying rate, as well as processes that have instantaneous jumps (bulk arrivals). More generally, a traffic process can be a combination of the above types. A suitable object for representing a traffic process is a nonnegative, sigma-finite measure on the real line. The notion of sigma-finiteness excludes the possibility of an infinite volume of traffic on a finite interval of time. Let $\mathcal{M}$ be the collection of all sigma-finite measures. Henceforth all such measures are restricted to $[0, \infty)$. Our purpose is to provide a formulation and design methodology for various types of flow control problems associated with traffic processes in $\mathcal{M}$.

In [3] we formulated and solved optimal flow control problems in discrete time. The proof techniques in discrete time are conceptually easier, because induction techniques can be used. The general optimal flow control design problem will be considered in the sequel. Because of a) the generality of traffic processes considered and b) the fact that we are working in continuous time, it is not possible to use induction type arguments to prove optimality. We thus need a more global approach to the matter. It turns out that the structure of the optimal flow controller as well as the proof of its optimality relies on the construction and properties of the so-called *reflection mappings*, which are reviewed in Appendix A. Their use not only achieves easier proof schemes but also allows for a wider interpretation of the burstiness constraints.

To precisely state what is meant by burstiness constraints, we recall that a traffic process $A \in \mathcal{M}$ is (σ, ρ) *constrained with initial virtual backlog* $\sigma_0 \leq \sigma$ iff

$$A[0, t] \leq \sigma_0 + \rho t, \text{ for all } t \geq 0,$$
$$A(s, t] \leq \sigma + \rho(t - s), \text{ for all } t \geq s \geq 0.$$

The terminology *virtual backlog* will be made more apparent through the course of development of the flow controller dynamics. More generally, we may replace $\{\rho t\}$ by another measure $R \in \mathcal{M}$. We then say that

Definition 1: $A \in \mathcal{M}$ is (σ_0, σ, R) constrained iff

$$A[0, t] \leq \sigma_0 + R[0, t], \text{ for all } t \geq 0,$$
$$A(s, t] \leq \sigma + R(s, t], \text{ for all } t \geq s \geq 0. \tag{1}$$

The measure R is a parameter of the design. The reader should notice that intervals $(s, t]$ are conveniently taken to be open on the left and closed on the right. The initial interval $[0, t]$ is taken to be closed on both sides. Occasionally we shall use the notation

$$A_t := A[0, t].$$

Note that such a process is right continuous with left limits (càdlàg ; see Appendix A) for each t.

A flow controller is regarded as a map, say φ, mapping $\mathcal{M}$ into itself. Naturally, such a map should be *causal*, in the sense that, for any $A_1, A_2 \in \mathcal{M}$,

$$\varphi(A_1^{[0,t)} + A_2^{[t,\infty)})^{[0,t)} = \varphi(A_1)^{[0,t)}, \text{ for all } t \geq 0, \quad (2)$$

where A^S denotes the restriction of the measure A on the subset S of the set $\mathbf{R}$ of real numbers. Besides causality, a *feasible flow controller* should map arbitrary traffic processes into traffic processes with a predescribed degree of burstiness as defined by the parameters σ_0, σ and the measure R. In other words, a feasible φ should be such that $\varphi(A)$ is (σ_0, σ, R) constrained, for any $A \in \mathcal{M}$.

We can consider two optimal flow control problems. One that poses a constraint on the buffer size, say K, that can be used (i.e., on the maximum number of messages that can ever be stored in the flow controller's memory), and one that poses a constraint on the maximum delay, say d, that an incoming message can experience before it is transmitted or rejected. (The discrete time version of the latter problem is precisely the one that was considered in [3].) The constraint on the buffer size can be expressed by the requirement that

$$\begin{aligned}&\varphi(A)[0,t] \leq A[0,t], t \geq 0,\\ &\varphi(A)(s,t] \leq K + A(s,t], t \geq s \geq 0,\end{aligned} \quad (3)$$

while the delay constraint can be expressed by

$$\begin{aligned}&\varphi(A)[0,t] \leq A[0,t], t \geq 0,\\ &\varphi(A)(s,t] \leq A(s-d,t], t \geq s \geq 0.\end{aligned} \quad (4)$$

with $A(s-d,t]$ defined to be $A[0,t]$ for $s < d$. Thus, given $\sigma \geq 0, 0 \leq \sigma_0 \leq \sigma, R \in \mathcal{M}, 0 \leq K \leq \infty, 0 \leq d \leq \infty$, we have

Definition 2: A *buffer-constrained* feasible flow controller φ is a causal flow controller that produces a (σ_0, σ, R) constrained output traffic stream that also satisfies (3).

Definition 3: A *delay-constrained* feasible flow controller φ is a causal flow controller that produces a (σ_0, σ, R) constrained output traffic stream that also satisfies (4).

A flow controller φ^* is *buffer-optimal* (respectively, *delay-optimal*) iff

$$\varphi^*(A)[0,t] \geq \varphi(A)[0,t] \quad (5)$$

for all $A \in \mathcal{M}$, $t \geq 0$, and all buffer-constrained (respectively, delay-constrained) feasible flow controllers φ. The two problems will be treated separately in the sequel.

Let us first start by expressing the constraints (1) in a more convenient way. Consider the càdlàg process (function)

$$x_t = \sigma - \sigma_0 + A_t - R_t. \quad (6)$$

(Recall that we use the shorthand $A_t = A[0,t]$.) Then the inequalities (1) are easily seen to be equivalent to[1]

$$x_t - \inf_{0 \leq s \leq t} x_s \wedge 0 \leq \sigma, \text{ for all } t \geq 0. \quad (7)$$

But $x_t - \inf_{0 \leq s \leq t} x_s \wedge 0$ is precisely the reflection of x_t upwards at zero (see Appendix A). The advantage of expressing (1) in the equivalent form (7) is that the reflected quantity can be "updated dynamically". So, to verify that a traffic process $A \in \mathcal{M}$ satisfies (σ_0, σ, R) constraints, it suffices to verify that the reflection of its associated "free process" x_t, as defined in (6), is below σ for all t. Equivalently, one could express the constraints (1) by defining

$$y_t = \sigma_0 + R_t - A_t = \sigma - x_t$$

and by requiring that its reflection downwards at σ, i.e., the process $\zeta_t = \inf_{0 \leq s \leq t}(y_t - y_s + \sigma) \wedge y_t$, stay above zero for all $t \geq 0$. This follows from the fact that the reflection operator commutes with affine transformations (see Appendix A). Note that, for $t > 0$, it is precisely ζ_{t-} that one would refer to as the *virtual backlog* corresponding to the traffic process A. Our observations so far lead to

Fact 1: The following are equivalent.

1) $A \in \mathcal{M}$ is (σ_0, σ, R) constrained.
2) The reflection of $\sigma - \sigma_0 + A_t - R_t$ upwards at zero stays below σ for all t.
3) The reflection of $\sigma_0 + R_t - A_t$ downwards at σ stays above zero for all t.

It is convenient to exemplify the method of the reflection mappings in the special case of an instantaneous flow controller, i.e., a φ with $K = 0$ in Definition 2 (or, what is the same, $d = 0$ in Definition 3). This is done in the sequel.

A. Instantaneous Optimal Flow Controllers

An instantaneous flow controller φ should satisfy (3) with $K = 0$, i.e.,

$$\begin{aligned}&\varphi(A)[0,t] \leq A[0,t], t \geq 0,\\ &\varphi(A)(s,t] \leq A(s,t], t \geq s \geq 0.\end{aligned}$$

We are looking for the best φ, in the sense of (5), which is also causal and produces a (σ_0, σ, R) constrained $\varphi(A)$. Consider the map φ^* defined via the following procedure. Given $A \in \mathcal{M}$ define its associated free process

$$x_t = \sigma - \sigma_0 + A_t - R_t$$

and reflect it at zero and σ

$$\begin{aligned}q_t^* &= x_t + \ell_t^* - u_t^*\\ &= [\sigma - \sigma_0 + A_t - R_t] + \ell_t^* - u_t^*,\end{aligned} \quad (8)$$

where $q^* = \mathcal{R}_{0,\sigma}(x)$, $\ell^* = \mathcal{L}_{0,\sigma}(x)$, $u^* = \mathcal{U}_{0,\sigma}(x)$ are, respectively, the reflected process, the lower boundary process, and the upper boundary process (see Appendix A for the notation used). The map φ^* maps, by definition, A into B^* defined by

$$B_t^* := A_t - u_t^*. \quad (9)$$

We note that B^* is indeed an element of $\mathcal{M}$. To see this, first note that $B_0^* \geq 0$ because if $u_0^* = 0$ then $B_0^* = A_0 \geq 0$, while if $u_0^* > 0$, we necessarily have $q_0^* = \sigma$ and $l_0^* = 0$, so that (8) gives $B_0^* = \sigma_0 + R_0 \geq 0$. Second, for $t > 0$, if $u_t = u_{t-}$ we have $B_t = A_t - u_t \geq A_{t-} - u_{t-} = B_{t-}$, while if $u_t > u_{t-}$ we necessarily have $q_t = \sigma$ and $\ell_t = \ell_{t-}$, so that (8) yields

$$B_t = \sigma_0 + R_t - \ell_t \geq \sigma_0 + R_{t-} - \ell_{t-} \geq A_{t-} - u_{t-} = B_{t-}.$$

[1] The operations min and max are supposed to have priority over plus and minus; thus $-a \wedge b$ means $-(a \wedge b)$.

Lemma 1: (Feasibility) i) The greedy flow controller φ^* is causal; ii) $\varphi^*(A)$ is (σ_0, σ, R) constrained; iii) $\varphi^*(A)[0,t] \leq A[0,t], \varphi^*(A)(s,t] \leq A(s,t], t \geq s \geq 0$.

Proof: The causality of φ^* follows immediately from the causality of the reflection mapping (Property 1 of Appendix A). Let $B^* = \varphi^*(A)$, as defined by (9). To show ii) we use Fact 1. According to it, one has to ensure that the reflection of $\sigma - \sigma_0 + B_t^* - R_t$ upwards at zero stays below σ. To this end, rewrite (8) as

$$q_t^* = [\sigma - \sigma_0 + B_t^* - R_t] + \ell_t^*.$$

Since ℓ_t^* increases only if q_t^* is zero, it follows, by the uniqueness of the reflection mapping, that q^* is also the reflection of $\sigma - \sigma_0 + B_t^* - R_t$ upwards at zero. Of course, $q_t^* \leq \sigma$, because it was originally defined as the reflection of x at zero and σ. Finally, iii) is immediate from (9) since u_t^* is nonnegative and nondecreasing. □

The main result is the optimality of the greedy flow controller.

Theorem 1: (Optimality) For any $A \in \mathcal{M}$ and any feasible φ,

$$\varphi^*(A)_t \geq \varphi(A)_t, t \geq 0.$$

Proof: Let $B^* = \varphi^*(A)$, as in (9), and $B = \varphi(A)$. The assumption that φ is feasible means that B is (σ_0, σ, R) constrained and that $B[0,t] \leq A[0,t], B(s,t] \leq A(s,t]$, $t \geq s \geq 0$. Reflect $\sigma - \sigma_0 + B_t - R_t$ at zero and σ, and let q, ℓ, u be, respectively, the reflected, lower boundary, upper boundary processes:

$$q_t = [\sigma - \sigma_0 + B_t - R_t] + \ell_t - u_t.$$

Since B is (σ_0, σ, R) constrained, the upper boundary process u is identically equal to zero. Rewrite the above as

$$q_t = [\sigma - \sigma_0 + A_t - R_t] + \ell_t - (A_t - B_t). \quad (10)$$

Since $B[0,t] \leq A[0,t], B(s,t] \leq A(s,t], t \geq s \geq 0$, the latter term $A_t - B_t$ is nonnegative and nondecreasing. Compare (10) with (8) and use the minimality property of the reflection mapping (Property 2 of Appendix A) to obtain

$$\ell_t^* \leq \ell_t, u_t^* \leq A_t - B_t.$$

To finish the proof, read the latter inequality as $B_t \leq A_t - u_t^* = B_t^*$. □

We thus established the optimality of the greedy flow controller. The interpretation of the construction should be clear, especially in the case where the arrival traffic process A and the process R are absolutely continuous with derivatives, say, a_t, r_t, respectively. Then, from (9), we see that B^* is also absolutely continuous with derivative equal to a_t whenever $q_t^* < \sigma$ and r_t whenever $q_t^* = \sigma$. The expression (9) for the mapping $\varphi^* : A \mapsto B^*$ precisely takes care of the cases where A is allowed to have jumps as well.

B. Flow Control with Buffer Constraints

Consider the optimization problem among the class of buffer-constrained feasible flow controllers φ (Definition 2). We follow the method introduced above. Namely, we first define a special flow controller φ^* by performing a suitable reflection operation on the free process $x_t = \sigma - \sigma_0 + A_t - R_t$ associated with an arrival process A. Then we show that φ^* is indeed feasible. Finally, we show that it is optimal in the sense of (5).

The suitable reflection operation that defines φ^* is a reflection of x at zero and $\sigma + K$: letting $q^* = \mathcal{R}_{0,\sigma+K}(x)$, $\ell^* = \mathcal{L}_{0,\sigma+K}(x)$, $u^* = \mathcal{U}_{0,\sigma+K}(x)$ be the reflected, upper boundary, and lower boundary processes, respectively, we have

$$q_t^* = [\sigma - \sigma_0 + A_t - R_t] + \ell_t^* - u_t^*. \quad (11)$$

Define $\varphi^* : A \mapsto B^*$ by

$$B_t^* := \begin{cases} A_t - u_t^* & \text{if } q_t^* < \sigma \\ \sigma_0 + R_t - \ell_t^* & \text{otherwise.} \end{cases} \quad (12)$$

We argue that $B^* \in \mathcal{M}$. For this, first note that $B_0^* \geq 0$. Indeed, if $u_0^* = \ell_0^* = 0$, then if $q_0^* < \sigma$ we have $B_0^* = A_0 \geq 0$, while if $q_0^* \geq \sigma$, we have $B_0^* = \sigma_0 + R_0 \geq 0$. If $u_0^* > 0$, we necessarily have $q_0^* = \sigma + K$ and $\ell_0^* = 0$, so that $B_0^* = \sigma_0 + R_0 \geq 0$. If $\ell_0^* > 0$, we necessarily have $q_0^* = 0$ and $u_0^* = 0$, so that $B_0^* = A_0 \geq 0$. Second, if $t > 0$ is such that $q_{t-}^* < \sigma$ and $q_t^* < \sigma$, we necessarily have $u_t^* = u_{t-}^*$, so that $B_t^* = A_t - u_t^* \geq A_{t-} - u_{t-}^* = B_{t-}^*$. If $q_{t-}^* < \sigma$ and $q_t^* \geq \sigma$, we necessarily have $\ell_t^* = \ell_{t-}^*$, so that

$$B_t^* = \sigma_0 + R_t - \ell_t^* \geq \sigma_0 + R_{t-} - \ell_{t-}^* \geq A_{t-} - u_{t-}^* = B_{t-}^*.$$

If $q_{t-}^* \geq \sigma$ and $q_t^* < \sigma$ we necessarily have $u_t^* = u_{t-}^*$, so that

$$B_t^* = A_t - u_t^* \geq A_{t-} - u_{t-}^* \geq \sigma_0 + R_{t-} - \ell_{t-}^* = B_{t-}^*.$$

Finally, if $q_{t-}^* \geq \sigma$ and $q_t^* \geq \sigma$, we necessarily have $\ell_t^* = \ell_{t-}^*$, so that

$$B_t^* = \sigma_0 + R_t - \ell_t^* \geq \sigma_0 + R_{t-} - \ell_{t-}^* = B_{t-}^*.$$

Note that we can use (11) to write (12) in either of the following ways:

$$B_t^* = A_t - u_t^* - (q_t^* - \sigma)^+ \quad (13)$$
$$= \sigma_0 + R_t - \ell_t^* - (\sigma - q_t^*)^+. \quad (14)$$

Lemma 2: (Feasibility) i) The greedy flow controller φ^* is causal; ii) $B^* = \varphi^*(A)$ is (σ_0, σ, R) constrained; iii) for all $t \geq s \geq 0$, $B_t^* \leq A_t$, and $B_t^* - B_s^* \leq K + A_t - A_s$.

Proof: Causality is immediate from the causality of the reflection mapping (Property 1 of Appendix A). Assertion ii) will follow if we ensure that the reflection of $\sigma - \sigma_0 + B_t^* - R_t$ upwards at zero does not exceed σ. Use the identity

$$\sigma - (\sigma - q)^+ = \sigma \wedge q$$

to rewrite (14) in the form

$$\sigma \wedge q_t^* = [\sigma - \sigma_0 + B_t^* - R_t] + \ell_t^*.$$

By definition, ℓ_t^* increases only when $q_t^* = 0$, or equivalently, only when $q_t^* \wedge \sigma = 0$. Thus, $q_t^* \wedge \sigma$ is the sought reflection of

$\sigma - \sigma_0 + B_t^* - R_t$ upwards at zero; and certainly $q_t^* \wedge \sigma$ does not exceed σ. To show iii) we use the following fact.

Fact 2: Let $A, B \in \mathcal{M}$. Then the following are equivalent.

1) $B_t \leq A_t$, $B_t - B_s \leq K + A_t - A_s$, for all $t \geq s \geq 0$.
2) The reflection of $\{A_t - B_t\}$ downwards at K stays above zero for all t.

The proof of this is essentially the same as that of Fact 1.

From (13) we have

$$(q_t^* - \sigma)^+ = [A_t - B_t^*] - u_t^*.$$

By definition, u_t^* increases only when $q_t^* = K + \sigma$, or, equivalently, only when $(q_t^* - \sigma)^+ = K$. Thus the reflection of $A_t - B_t^*$ downwards at K is precisely equal to the nonnegative process $(q_t^* - \sigma)^+$. Owing to Fact 2, this establishes iii) and so the lemma has been proved. □

Theorem 2: (Optimality) For any $A \in \mathcal{M}$ and any buffer-constrained feasible flow controller φ,

$$\varphi^*(A)_t \geq \varphi(A)_t, t \geq 0.$$

Proof: Consider an arbitrary buffer-constrained feasible flow controller φ, let $B = \varphi(A)$, and let $B^* = \varphi^*(A)$. We re-express the two basic properties of the map $\varphi : A \mapsto B$ via Facts 1 and 2. First, the inequalities $B_t \leq A_t$, $B_t - B_s \leq K + A_t - A_s$ for all $t \geq s \geq 0$ are equivalent to the fact that the reflection of $A_t - B_t$ downwards at K stays above zero, that is if we let

$$q_t^1 = [A_t - B_t] + \ell_t^1 - u_t^1, \tag{15}$$

with q^1, ℓ^1, u^1 being the reflected, upper boundary (at K), and lower boundary (at zero) processes, respectively, then $\ell^1 \equiv 0$. Second, the inequalities $B_t \leq \sigma_0 + R_t$, $B_t - B_s \leq \sigma + R_t - R_s$ for all $t \geq s \geq 0$ are equivalent to the fact that the reflection of $\sigma - \sigma_0 + B_t - R_t$ upwards at zero stays below σ. So, if we let

$$q_t^2 = [\sigma - \sigma_0 + B_t - R_t] + \ell_t^2 - u_t^2, \tag{16}$$

with q^2, ℓ^2, u^2 being the reflected, upper boundary (at σ), and lower boundary (at zero) processes, respectively, then $u^2 \equiv 0$. Add up the previous two displays (15) and (16), to obtain

$$q_t^1 + q_t^2 = [\sigma - \sigma_0 + A_t - R_t] + \ell_t^2 - u_t^1.$$

Compare this with the expression (11) for q_t^*. Since ℓ_t^2, u_t^1 are nondecreasing and since $q_t^1 + q_t^2 \in [0, \sigma + K]$, it follows, by minimality (Property 2 of Appendix A), that

$$u_t^* \leq u_t^1, \tag{17}$$

$$\ell_t^* \leq \ell_t^2. \tag{18}$$

Use inequality (17) in the expression (15) for q_t^1 to obtain

$$0 \leq q_t^1 \leq A_t - B_t - u_t^* = B_t^* - B_t + (q_t - \sigma)^+,$$

where we also used (13). Likewise, use inequality (18) in the expression (16) for q_t^2 to obtain

$$\sigma \geq q_t^2 \geq \sigma - \sigma_0 + B_t - R_t + \ell_t^* = B_t - B_t^* + \sigma - (\sigma - q_t)^+,$$

where we also used (14). From the last two displays we readily obtain

$$B_t - B_t^* \leq \min\{(q_t - \sigma)^+, (\sigma - q_t)^+\} = 0,$$

where we used the obvious identity $\min\{x^+, (-x)^+\} \equiv 0$. □

The interpretation of the optimal flow controller is especially interesting in this case, since, it turns out that it is the *Leaky Bucket* flow control scheme, an accepted standard for flow control in ATM networks. For more details, see Section IV below.

C. Flow Control with Delay Constraints

A delay-constrained feasible flow controller $\varphi : A \mapsto B$ should satisfy inequalities (4), i.e.,

$$B[0,t] \leq A[0,t], B(s,t] \leq A(s-d,t], t \geq s \geq 0, \tag{19}$$

where for $s < d$, $A(s-d,t]$ is interpreted as $A[0,t]$. The problem is to find the best such φ. The idea behind the optimal construction hinges on the fact that the above inequalities can also be interpreted via reflection mappings. To see this, write

$$A(s-d,t] = A(s-d,s] + A(s,t]$$

and let

$$\psi_s = A(s-d,s].$$

Then inequality (19) can be written as

$$\inf_{0 \leq s \leq t}\{A(s,t] - B(s,t] + \psi_s\} \wedge (A_t - B_t) \geq 0, \text{ for all } t \geq 0. \tag{20}$$

We interpret the nonnegative càdlàg process $\{\psi_s, s \geq 0\}$ as an upper (time-varying) boundary. The quantity in the left hand side of (20) is then the reflection of the process $\{A_t - B_t, t \geq 0\}$ downwards at the upper boundary ψ (see Appendix A). We now need to use

Fact 3: Let $A, B \in \mathcal{M}$, and ψ a nonnegative càdlàg function. Then the following are equivalent.

1) $B_t \leq A_t, B_t - B_s \leq \psi_s + A_t - A_s$, for all $t \geq s \geq 0$.
2) The reflection of $\{A_t - B_t\}$ downwards at $\{\psi_t\}$ stays above zero for all t.

The proof of this fact is also essentially identical to that of Fact 1.

Fact 3 suggests to define a greedy flow controller φ^* in a manner analogous to that of the buffer-constrained problem. Consider the reflection of $\sigma - \sigma_0 + A_t - R_t$ upwards at zero and downwards at $\{\sigma + \psi_t\}$. This yields

$$q_t^* = [\sigma - \sigma_0 + A_t - R_t] + \ell_t^* - u_t^*, \tag{21}$$

with ℓ^*, u^* being the lower and upper boundary processes, respectively; ℓ_t^* increases only when $q_t^* = 0$, i.e., $\int 1\{q_t^* > 0\}d\ell_t^* = 0$, while u_t^* increases only when $q_t^* = \sigma + \psi_t$, i.e., $\int 1\{q_t^* < \sigma + \psi_t\}du_t^* = 0$. The greedy flow controller is then defined precisely by the same expression as before, namely (12), or, equivalently, (13) or (14). We argue that $B^* \in \mathcal{M}$. The proof for this is identical to that used in the buffer constrained problem, with the only difference being that to show $B_0^* \geq 0$ in the case that $u_0^* > 0$, we need to observe that $q_0^* = \sigma + \psi_0^* \geq \sigma$ and $l_0^* = 0$, so that $B_0^* = \sigma_0 + R_0 \geq 0$.

The structure of this φ^* may at this point seem rather abstract. But in discrete time (i.e., when all measures considered are supported on the integers) φ^* is exactly the optimal flow controller studied in [3]. In Section 3 below we shall derive, starting from the definition (12), an algorithm for the implementation of φ^* that is simpler than the one presented in [3].

Lemma 2: (Feasibility) i) The greedy flow controller φ^* is causal; ii) $B^* = \varphi^*(A)$ is (σ_0, σ, R) constrained; iii) for all $t \geq s \geq 0$, $B^*_t \leq A_t$, and $B^*(s,t] \leq A(s-d,t]$.

Proof: Precisely as in the proof of Lemma 2 one observes that $\sigma \wedge q^*_t$ is the reflection of $\sigma - \sigma_0 + B^*_t - R_t$ upwards at zero, which establishes ii). Also, $(q^*_t - \sigma)^+$ is the reflection of $A_t - B^*_t$ downwards at $\{\psi_t\}$, and this establishes iii). The causality of φ^* is a direct consequence of the causality of the reflection mapping with time-varying boundaries. □

Theorem 3: (Optimality) For any $A \in \mathcal{M}$ and any delay-constrained feasible flow controller φ,

$$\varphi^*(A)_t \geq \varphi(A)_t, t \geq 0.$$

Proof: Exactly as the proof of Theorem 2, with Kreplaced by the time-varying boundary $\{\psi_t\}$. □

III. The Structure of Discrete Time Optimal Flow Controllers and Simple Implementation Algorithms

The design of optimal flow controllers was considered in the previous section. In this section we look into the structure of the optimal flow controllers in discrete time. In particular we will derive explicit recursions for their implementation. Both buffer-constrained and delay-constrained flow controllers will be examined. In Section III–A we derive recursions for optimal buffer-constrained flow controllers. In Section III–B we derive recursions for optimal delay-constrained flow controllers. Recursions similar to the latter ones were derived in [3] using entirely different techniques.

The discrete time arrival process is described by a sequence of nonnegative numbers $(a_n, n \geq 0)$. We write $A_n = \sum_{k=0}^{n} a_k$, $n \geq 0$ and set $A_{-1} = 0$ for notational convenience. To this sequence we may associate the càdlàg function $(A_t, t \geq 0)$ defined by

$$A_t = A[0,t] = \sum_{k=0}^{\infty} 1\{k \leq t\} a_k$$

thought of as an arrival process in continuous time, i.e. as an element of $\mathcal{M}$. Note that this continuous time traffic process is piecewise constant with jumps only at the nonnegative integers. Similarly, we assume given a sequence $(r_n, n \geq 0)$ of nonnegative numbers; write $R_n = \sum_{k=0}^{n} r_k$, $n \geq 0$, setting $R_{-1} = 0$ for notational convenience; and associate to this sequence the càdlàg function $(R_t, t \geq 0)$ defined by

$$R_t = R[0,t] = \sum_{k=0}^{\infty} 1\{k \leq t\} r_k$$

which we interpret as the element of $\mathcal{M}$ prescribing the desired (σ_0, σ, R) burstiness constraint that the regulated output must satisfy (in addition to the buffer or delay constraint on the flow controller). From the already derived structure of the optimal flow controllers we then see that all the derived processes, i.e. the free process, the regulated process, and the upper and lower boundary processes, will be càdlàg piecewise constant with jumps only at the nonnegative integers. This is the situation dealt with in Appendix B.

A. Structure of the Optimal Flow Controller with Buffer Constraints

Recall the way that the greedy flow controller φ^* was defined in Section II, (11) and (12). The construction is valid for all values of the allowable memory K, including $K = 0$ and $K = \infty$. Using the formulas of Appendix B we can write explicit recursions. See also Cruz [7] for similar recursions in a queueing framework. With $A_n = \sum_{k=0}^{n} a_k$, $R_n = \sum_{k=0}^{n} r_k$, consider the reflection of the free process

$$x_n = \sigma - \sigma_0 + A_n - R_n, n \geq 0. \tag{22}$$

at 0 and $\sigma + K$. The reflected process is

$$q^*_n = x_n + \ell^*_n - u^*_n, \tag{23}$$

where ℓ^*_n, u^*_n are the lower boundary, upper boundary processes, respectively. For notational convenience, we let $x_{-1} = \sigma - \sigma_0 = q^*_{-1}$, which is consistent with the above equations, and with the interpretation of $(\sigma - q_n)^+$ as the amount of token left after the decision at time n. From (41), (42), and (43) of Appendix B we get

$$q^*_n = \left[q^*_{n-1} + \Delta x_n\right]_0^{\sigma+K} \tag{24}$$

$$\Delta\ell^*_n = \ell^*_n - \ell^*_{n-1} = -(q^*_{n-1} + \Delta x_n) \wedge 0 \tag{25}$$

$$\Delta u^*_n = u^*_n - u^*_{n-1} = (q^*_{n-1} + \Delta x_n - \sigma - K) \vee 0. \tag{26}$$

Denote by $B^*_n = \sum_{k=0}^{n} b^*_k$ the flow controller's output process. The defining relations for B^*_n, (13), (14), now yield

$$B^*_n = A_n - u^*_n - (q^*_n - \sigma)^+ \tag{27}$$

$$= \sigma_0 + R_n - \ell^*_n - (\sigma - q^*_n)^+. \tag{28}$$

We are looking for a recursion for $b^*_n = \Delta B^*_n = B^*_n - B^*_{n-1}$. It is convenient to define

$$\lambda^*_n = (q^*_{n-1} - \sigma)^+ \tag{29}$$

$$\mu^*_n = \sigma \wedge q^*_{n-1} \tag{30}$$

$$\sigma^*_n = \sigma - \mu^*_n = (\sigma - q^*_{n-1})^+. \tag{31}$$

Note that $\lambda^*_0 = (q^*_{-1} - \sigma)^+ = 0$, $\mu^*_0 = \sigma - \sigma_0$, and $\sigma^*_0 = \sigma_0$.

We claim that b^*_n satisfies

$$b^*_n = (a_n + \lambda^*_n) \wedge (r_n + \sigma^*_n). \tag{32}$$

The proof of this is a series of straightforwardly tedious computations, which we undertake in the sequel. From the first of the defining relations for B^*_n (27) and (29) we have

$$b^*_n = a_n - \Delta u^*_n - \lambda^*_{n+1} + \lambda^*_n. \tag{33}$$

From (23) and (25) we obtain

$$\begin{aligned} q^*_n &= q^*_{n-1} + \Delta x_n + \Delta\ell^*_n - \Delta u^*_n \\ &= q^*_{n-1} + \Delta x_n - \Delta u^*_n - (q^*_{n-1} + \Delta x_n) \wedge 0 \\ &= (q^*_{n-1} + \Delta x_n - \Delta u^*_n) \vee (-\Delta u^*_n). \end{aligned}$$

The definition (29) of λ_n^* together with the above display gives

$$\begin{aligned}\lambda_{n+1}^* &= (q_n^* - \sigma) \vee 0 \\ &= (q_{n-1}^* + \Delta x_n - \Delta u_n^* - \sigma) \vee (-\Delta u_n^* - \sigma) \vee 0 \\ &= (q_{n-1}^* + \Delta x_n - \Delta u_n^* - \sigma) \vee 0.\end{aligned}$$

Substitute the latter display into (33) and use $\Delta x_n = a_n - r_n$ (from (22)) to obtain

$$\begin{aligned}b_n^* &= (a_n - \Delta u_n^*) \wedge (\sigma + r_n - q_{n-1}^*) + \lambda_n^* \\ &= a_n \wedge (\sigma + K + r_n - q_{n-1}^*) \wedge (\sigma + r_n - q_{n-1}^*) + \lambda_n^*, \\ &\quad \text{by (26)} \\ &= a_n \wedge (\sigma + r_n - q_{n-1}^*) + \lambda_n^* \\ &= (a_n + \lambda_n^*) \wedge (\sigma + r_n - q_{n-1}^* + \lambda_n^*) \\ &= (a_n + \lambda_n^*) \wedge (\sigma + r_n - q_{n-1}^* + (q_{n-1}^* - \sigma)^+), \\ &\quad \text{by (29)} \\ &= (a_n + \lambda_n^*) \wedge (r_n + (\sigma - q_{n-1}^*)^+) \\ &= (a_n + \lambda_n^*) \wedge (r_n + \sigma_n^*), \text{ by (31)}.\end{aligned}$$

So (32) has been proved. We now obtain recursions for the quantities λ_n^* and σ_n^* appearing in the derived formula for b_n^*. From (33) we obtain

$$\lambda_{n+1}^* = \lambda_n^* + a_n - b_n^* - \Delta u_n^*$$

when, by summing from zero up to n and using $\lambda_0^* = 0$,

$$\lambda_{n+1}^* = [A_n - B_n^*] - u_n^*.$$

Notice that $\lambda_{n+1}^* \leq K$. Also, $\Delta u_n^* > 0$ implies $\lambda_{n+1}^* = K$. It follows that $\{\lambda_{n+1}^*\}$ is the reflection of $\{A_n - B_n^*\}$ downwards at K, implying that (see formula (43) of Appendix B)

$$\Delta u_n^* = (\lambda_n^* + a_n - b_n^* - K) \vee 0.$$

Substitute this into (33) to obtain

$$\lambda_{n+1}^* = (\lambda_n^* + a_n - b_n^*) \wedge K. \tag{34}$$

From the second expression (28) for B_n^* and definition (30) we have

$$\mu_{n+1}^* = [\sigma - \sigma_0 + B_n^* - R_n] + \ell_n^*. \tag{35}$$

Notice that $\mu_{n+1}^* \geq 0$. Also, $\Delta \ell_n^* > 0$ implies $\mu_{n+1}^* = 0$. It follows that $\{\mu_{n+1}^*\}$ is the reflection of $\sigma - \sigma_0 + B_n^* - R_n$ upwards at zero, and so (see formula (42) of Appendix B)

$$\Delta \ell_n^* = -(\mu_n^* + b_n^* - r_n) \wedge 0. \tag{36}$$

Expressions (35) and (36) yield

$$\mu_{n+1}^* = (\mu_n^* + b_n^* - r_n) \wedge 0.$$

Finally, substituting the latter into (31), we obtain

$$\sigma_{n+1}^* = \sigma \wedge (\sigma_n^* + r_n - b_n^*). \tag{37}$$

To recapitulate [cf., the recursions (32), (34), and (37)] the algorithm that defines the optimal flow controller with buffer constraint K is as follows:

$$\begin{aligned}b_n^* &= (a_n + \lambda_n^*) \wedge (r_n + \sigma_n^*) \\ \lambda_{n+1}^* &= (\lambda_n^* + a_n - b_n^*) \wedge K \\ \sigma_{n+1}^* &= \sigma \wedge (\sigma_n^* + r_n - b_n^*).\end{aligned}$$

The algorithm is initialized with $\lambda_0 = 0$ and $\sigma_0^* = \sigma_0$. Note that the quantity σ_n^* is the virtual backlog, also encountered in [3]. The physical meaning of this algorithm is clear: send forward as much flow as possible subject to the burstiness constraint given by the virtual backlog. Retain as much of the remaining flow as possible in the buffer (up to K units of flow).

One observation that is probably worth mentioning at this point is that all computations above remain valid if the traffic processes do not necessarily have jumps on the integers $k = 1, 2, \ldots$, but on an increasing sequence of times $T_1 < T_2 < \ldots$

B. Structure of the Optimal Flow Controller with Delay Constraints

The definition of the greedy flow controller in presence of delay constraint d calls for reflection of the free process

$$x_n = \sigma - \sigma_0 + A_n - R_n$$

at 0 and the time-varying upper boundary $\{\sigma + \psi_n\}$, where

$$\psi_n = A(n - d, n] = a_{n-d+1} + \cdots + a_n.$$

The reflected process is

$$q_n^* = x_n + \ell_n^* - u_n^*$$

and satisfies the recursion (set $\alpha_n = 0$ and $\beta_n = \sigma + \psi_n$ in formula (41) of Appendix B)

$$q_n^* = [q_{n-1}^* + \Delta x_n]_0^{\sigma + \psi_n}.$$

The algebra done earlier (Section III-A) goes through provided that we replace K with $\psi_n = a_{n-d+1} + \cdots + a_n$ everywhere in the computations. We omit the computations to save space. The algorithm for the optimal flow controller with delay constraints reads

$$b_n^* = (a_n + \lambda_n^*) \wedge (r_n + \sigma_n^*) \tag{38}$$

$$\lambda_{n+1}^* = (\lambda_n^* + a_n - b_n^*) \wedge (a_{n-d+1} + \cdots + a_n) \tag{39}$$

$$\sigma_{n+1}^* = \sigma \wedge (\sigma_n^* + r_n - b_n^*) \tag{40}$$

with the initialization $\lambda_0^* = 0$ and $\sigma_0^* = \sigma_0$. Again, the quantity σ_n^* is interpreted as the virtual backlog.

IV. Optimality of the Leaky Bucket Flow Controller

It turns out that the optimal flow controller with buffer constraints (Sections II-B, III-A) has the structure of the Leaky Bucket flow controller, a recently accepted standard for open loop flow control in ATM networks. The leaky bucket operates as follows. Cells arrive according to some (bursty) traffic process A. Their transmission is controlled by objects called tokens. Tokens are generated according to a process R (in practice R is a periodic process with rate ρ, i.e., one token is generated every ρ^{-1} time units) and are stored in the token buffer that can hold up to σ tokens. An arriving cell that finds a token in the token buffer is transmitted instantaneously; otherwise, it waits till the generation of the next token. For more details see [2].

Now consider the flow controller φ^* of Section II-B. Its evolution equations in discrete time are given in Section III-A. Let q_t^* be the reflected process (11). Interpret $(\sigma - q_t^*)^+$ as the number of tokens in the system at time t. Similarly, interpret $(q_t^* - \sigma)^+$ as the number of cells in the system at time t. (Note that, at all times, there are either cells or tokens in the system.) A moment of reflection shows that the output process B_t^* given by (12) behaves exactly as the output process of a leaky bucket. We thus have the following result.

Fact 4: The optimal flow controller that produces a (σ_0, σ, R) constrained output traffic process and uses a buffer of size K is a leaky bucket with token buffer size σ, cell buffer size K, token arrival process R and initial number of tokens σ_0.

A leaky bucket with $K = \infty$ is referred to as a (σ, ρ) regulator (cf., [5]) In [2], we showed that this leaky bucket is burst-reducing in the following sense: let A be a stationary and ergodic point process with rate λ and let R be deterministic with rate $\rho > \lambda$. Consider the output process as an input to a queue with deterministic service rate $\mu > \rho$. Then the steady state queue length in the latter queue increases stochastically as σ increases.

V. Concluding Remarks

The problem of designing optimal flow controllers that regulate the burstiness of traffic, while guaranteeing a bound on the buffer or on the delay, was solved in a very general framework. In a sense, our result is "model free." No specific assumptions on the input model have been used. There are two important methodological points that should be emphasized. First, the burstiness constraints of a traffic stream can be summarized by the virtual backlog, which acts as a sufficient statistic, summarizing the past information. This simple realization allows one to pose and study a new class of control and optimization problems of substantial relevance to the design of networks handling burstiness constrained flows [1]. Second, in this work, we realized that constraints on the buffer size and constraints on the delay can be thought of as "generalized (σ, ρ) constraints" that can be expressed via a reflection mapping on a suitable "free process" (Section II). The analysis then hinges on some structural properties of this reflection mapping, the most important of which is its minimality (Appendix A). It is likely that this simple realization made in this paper will be useful in other optimization problems as well, for instance in a multi-user environment.

Since the problems we considered were formulated in a model free context, one could very well pose them in a stochastic context as well, by requiring, for instance, that the arrival process be an arbitrary stochastic process. The greedy flow controller φ^* is then seen to minimize the average loss rate when such averages exist. Indeed, our results hold pathwise.

The Leaky Bucket flow control scheme was shown to be optimal among the class of causal flow controllers. This is a new result, to the best of our knowledge, that should be comforting to network engineers, who plan to use it in any case. In fact, it should be stressed that nowhere in the analysis did we make explicit use of the causality of the decision rule of a feasible flow controller (except, of course, that we cannot borrow flow from the future). Thus, the leaky bucket (a causal controller by itself) is optimal even when one allows the flow control decision to be made noncausally.

Appendix A
The Reflection Mapping and its Properties

Reflection mappings are used throughout the paper in order to define the flow controllers as well as in establishing their optimality. In this appendix we review the definition of the reflection mapping of a "free process" x_t in a–possibly time-varying interval $I_t = [\alpha_t, \beta_t]$. We denote by $D_A[0, \infty)$ the collection of all càdlàg (i.e., right-continuous with left limits) functions mapping $[0, \infty)$ into a space A, a subset of $\mathbf{R}$ for our purposes. One of the earliest references to reflection in one dimension, in a stochastic context, is due to Skorokhod [16]. See also Harrison [9], for a nice discussion of the reflection with two boundaries and for continuous x, Chung and Williams [4], and the references therein.

A. The Reflection Problem

We are given càdlàg functions x_t (the free process), α_t (the lower boundary), and β_t (the upper boundary). Denote by I_t the interval $[\alpha_t, \beta_t]$. (Usually, I_t is constant.) It is assumed that $\alpha_t < \beta_t$ for all $t \geq 0$ (note the strict inequality, which is important for some of the statements below). The case where one of the boundaries is at infinity is allowed. The–so called–Skorokhod's reflection problem looks for functions ℓ_t (*lower boundary process*), u_t (*upper boundary process*) such that

1) The *reflected process*

$$q_t := x_t + \ell_t - u_t$$

is in I_t for all $t \geq 0$.

2) Both ℓ_t, u_t are nondecreasing with $\ell_{0-} = u_{0-} = 0$ and ℓ_t (respectively, u_t) increases only when $q_t = \alpha_t$ (respectively, $q_t = \beta_t$), i.e.,

$$\int_{\mathbf{R}_+} 1(q_t > \alpha_t) d\ell_t = 0, \int_{\mathbf{R}_+} 1(q_t < \beta_t) du_t = 0.$$

It turns out that the solution to this problem is unique and can be constructed by successive approximations (cf., Harrison [9]). The mapping $x \mapsto (\ell, u, q)$ is referred to as "the reflection mapping." Our notation is

$$q = \mathcal{R}_{\alpha,\beta}(x), \ell = \mathcal{L}_{\alpha,\beta}(x), u = \mathcal{U}_{\alpha,\beta}(x).$$

Occasionally, we may also write

$$q = \mathcal{R}_I(x), \ell = \mathcal{L}_I(x), u = \mathcal{U}_I(x).$$

When one boundary is present (i.e., when either $\alpha_t = -\infty$ or $\beta_t = +\infty$) then explicit formulas are available.

B. Formulas for Single Boundary Reflection

There are no explicit formulas when both boundaries are present. However, when only one boundary is present (i.e., when either $\alpha_t = -\infty$ or $\beta_t = +\infty$) we do have useful explicit formulas.

1) Lower Boundary Only:

$$\ell_t = -\inf_{0\le s\le t}(x_s - \alpha_s) \wedge 0,$$
$$q_t = x_t + \ell_t = \sup_{0\le s\le t}(x_t - x_s + \alpha_s) \vee x_t.$$

2) Upper Boundary Only:

$$u_t = \sup_{0\le s\le t}(x_s - \beta_s) \vee 0,$$
$$q_t = \inf_{0\le s\le t}(x_t - x_s + \beta_s) \wedge x_t.$$

C. Fundamental Properties

The following properties of the reflection mapping are used in the paper.

1) Causality: Fix time $s > 0$ and consider the free process after s, defined by

$$x_t^s := q_{s-} + x_{s+t} - x_{s-}, t \ge 0.$$

To obtain the reflected process at t time units after s, we may forget entirely the past of x and q before s and consider the reflection of the new free process $\{x_t^s, t \ge 0\}$, i.e.,

$$q_{s+t} = x_t^s + \mathcal{L}_{\alpha,\beta}(x^s)_t - \mathcal{U}_{\alpha,\beta}(x^s)_t, t \ge 0.$$

2) Minimality: The boundary process $\ell_t - u_t$ that maintains q_t within the constraint set is the smallest process (in the sense of smallest variation) that manages to do so. To put it in terms of the ℓ and u, suppose that ℓ_t', u_t' are two increasing processes such that

$$q_t' := x_t + \ell_t' - u_t' \in [\alpha_t, \beta_t], \text{ for all } t \ge 0.$$

Then

$$\ell_t \le \ell_t', u_t \le u_t', \text{ for all } t \ge 0.$$

In addition, the following properties are useful, though we do not use them.

3) Monotonicity: When two free processes x, x' compare in the sense that

$$x_t - x_s \le x_t' - x_s', 0 \le s \le t, \text{ and } x_0 \le x_0'$$

then the reflected process compare in the sense that

$$q_t \le q_t', t \ge 0.$$

4) Affine Scaling: Let T be a nonsingular affine transformation of $\mathbf{R}$, i.e.,

$$T(x) = \lambda x + \mu, \lambda \ne 0.$$

Then, for any interval $I \equiv I_t = [\alpha_t, \beta_t]$, and any free process $x \equiv x_t$

$$T\mathcal{R}_I(x) = \mathcal{R}_{TI}(Tx).$$

In other words,

$$\mathcal{R}_{\alpha,\beta}(x) = \frac{1}{\lambda}\mathcal{R}_{\lambda\alpha+\mu,\lambda\beta+\mu}(\lambda x + \mu) - \frac{\mu}{\lambda}.$$

Appendix B
Formulas for Discrete-Time Reflection

Let $\{x_t, t \ge 0\}$ be a càdlàg process, which is piecewise constant and jumps only at the nonnegative integer points $n = 0, 1, 2, \cdots$ Let x_{-1} be arbitrary, and set $q_{-1} = x_{-1}$. Let $\Delta x_n = x_n - x_{n-1}, n \ge 0$. Consider two boundaries α and β, possibly time varying, in which case we denote them by α_t, β_t, and assume that they are also piecewise constant, càdlàg , with possible jumps on the nonnegative integers, and $\alpha_t \le \beta_t$, for all $t \ge 0$. We denote by q_t the process that is obtained by reflecting x_t in the interval $[\alpha_t, \beta_t]$ and let ℓ_t, u_t be the lower and upper boundary processes, respectively. Clearly $(q_t, t \ge 0)$, $(\ell_t, t \ge 0)$, $(u_t, t \ge 0)$ are piecewise constant as well, with jumps at the nonnegative integers. The purpose of this appendix is to find explicit recursions for the values q_n at the nonnegative integer points $n = 0, 1, 2, \cdots$. We have

$$q_n = x_n + \ell_n - u_n.$$

As in Appendix A, ℓ_t, u_t are the unique nondecreasing processes, with $\ell_{0-} = u_{0-} = 0$ that maintain q_t between α_t and β_t, for all $t \ge 0$, and satisfy

$$\int_{\mathbf{R}_+} 1(q_t > \alpha_t)d\ell_t = \int_{\mathbf{R}_+} 1(q_t < \beta_t)du_t = 0.$$

Taking into account the fact that all processes are càdlàg and piecewise constant, the latter two relations are written as

$$\sum_{n=0}^{\infty} 1\{q_n > \alpha_n\}\Delta\ell_n = 0 = \sum_{n=0}^{\infty} 1\{q_n < \beta_n\}\Delta u_n.$$

For $n \ge 0$, let $\Delta\ell_n = \ell_n - \ell_{n-1}$, $\Delta u_n = u_n - u_{n-1}$, with $l_{-1} = u_{-1} = 0$. For $\alpha \le \beta$ we introduce the "truncation" symbol

$$[x]_\alpha^\beta = \alpha \vee (x \wedge \beta) = (\alpha \vee x) \wedge \beta = \begin{cases} \alpha, & x < \alpha \\ x, & \alpha \le x \le \beta \\ \beta, & x > \beta, \end{cases}$$

and claim that the following recursions yield the q_n, ℓ_n, and u_n that satisfy the aforementioned requirements.

$$q_n = [q_{n-1} + \Delta x_n]_{\alpha_n}^{\beta_n} \quad (41)$$
$$\Delta\ell_n = -(q_{n-1} + \Delta x_n - \alpha_n) \wedge 0 \quad (42)$$
$$\Delta u_n = (q_{n-1} + \Delta x_n - \beta_n) \vee 0. \quad (43)$$

Indeed, observe that, for all $n \ge 0$, $\alpha_n \le q_n \le \beta_n$, and $\Delta\ell_n \ge 0$, $\Delta u_n \ge 0$. Furthermore, if $\Delta\ell_n > 0$ then $q_{n-1} +$

$\Delta x_n - \alpha_n < 0$, implying that $q_n = \alpha_n$. So $\sum_{n=0}^{\infty} 1\{q_n > \alpha_n\}\Delta \ell_n = 0$. Similar conclusion holds for u_n. This proves that (41), (42), and (43) solve the reflection problem for $\{x_n\}$ at the boundaries $\{\alpha_n\}$ and $\{\beta_n\}$.

References

[1] V. Anantharam, "An approach to the control of ATM networks," presented at IEEE Int. Symp. Inform. Theory, Bahia, Brazil, July 1992. See also in *Proc. Int. Teletraffic Congr.*, Indian Inst. of Service, Bangalore, Nov. 1993, pp. 163-167; and in *Proc. 32nd IEEE CDC*, San Antonio, TX, Dec. 1993, pp. 1678–1679.

[2] V. Anantharam and T. Konstantopoulos, "Burst reduction properties of the leaky bucket flow control scheme in ATM networks," in *Proc. 29th Allerton Conf.*, 1991, pp. 302-309. Also submitted to *IEEE Trans. Commun.*.

[3] ——, "An optimal flow control scheme that regulates the burstiness of traffic subject to delay constraints," in *Proc. 32nd IEEE CDC*, San Antonio, TX, 1993, pp. 3606-3610.

[4] K. L. Chung and R. J. Williams, *Introduction to Stochastic Integration.* Boston, MA: Birkhäuser, 1990, 2nd ed.

[5] R. L. Cruz, "A calculus for network delay, Part I: Network elements in isolation" *IEEE Trans. Inform. Theory*, vol. 37, pp. 114-131, 1991.

[6] R. L. Cruz and M. C. Chuah, "A minimax approach to a simple routing problem," *IEEE Trans. Automat. Contr.*, vol. 36, pp. 1424-1435, 1991.

[7] R. L. Cruz and H.-N. Liu, "Single server queues with loss: A formulation," in *Proc. 1993 CISS*, The Johns Hopkins Univ., Baltimore, MD, 1993.

[8] A. E. Eckberg and D. M. Lucantoni, "A traffic/performance analysis of the bandwidth management throughput-burstiness filter," in *Proc. 29th IEEE Conf. Dec. Contr.*, 1988, pp. 2118-2123.

[9] J. M. Harrison, *Brownian Motion and Stochastic Flow Systems.* New York: Wiley, 1985.

[10] D. Jagerman and B. Melamed, "The transition and autocorrelation structure of TES processes: Part I: General theory; Part II: Special cases," *Commun. Stat. Stoch. Models*, vol. 8, no. 1, pp. 194-219, 1992; and Part II in vol. 8, no. 3, pp. 499-527, 1992.

[11] L. Kuang, "On the variance reduction property of buffered leaky bucket," Preprint, 1992.

[12] W. E. Leland, M. S. Taqqu, W. Willinger, and D. V. Wilson, "On the self-similar nature of Ethernet traffic," *IEEE/ACM Trans. Networking*, vol. 2, pp. 1–15, Feb. 1994.

[13] S. Low and P. Varaiya, "A simple theory of traffic and resource allocation in ATM," in *Proc. GLOBECOM*, 1991, pp. 1633-1637.

[14] S. Low, "Traffic management of ATM networks: Service provisioning, routing, and traffic shaping.," Ph.D. dissertation, Univ. of California, Berkeley, 1992.

[15] G. Sasaki, "Input buffer requirements for round robin polling systems with nonzero switchover times," in *Proc. 29th Allerton Conf.*, 1991, pp. 603-612.

[16] A. V. Skorokhod, "Stochastic equations for diffusion processes with boundaries," *Theoretic. Prob. Appl.*, vol. 6, pp. 287-298, 1961.

Takis Konstantopoulos (M'90) received the Diploma in electrical engineering from the National Technical University of Athens, Greece, and the M.Sc. and Ph.D. degrees from the University of California, Berkeley, in 1983, 1985 and 1989, respectively.

He held the position of Chargé de Recherce in INRIA, Sophia Antipolis, France, between 1989 and 1990. He was visiting faculty at the University of California, Berkeley in 1991. In spring 1992 he was a visiting researcher at the Mathematical Sciences Institute of Cornell University. He is currently Assistant Professor with the Department of Electrical and Computer Engineering, at University of Texas at Austin. His main research interests are in performance analysis, modeling and control of communication networks, applied probability and stochastic processes and systems.

Dr. Konstantopoulos is a recipient of an NSF Research Initiation Award and an NSF Faculty Career Development Award. He is currently Associate Editor of *Queueing Systems: Theory and Applications*. He is a member of the Mathematical Association of America, the Institute of Mathematical Statistics, and the Technical Chamber of Greece.

Venkat Anantharam (M'86) received the B.Tech degee in electrical engineering (electronics) in 1980 from the Indian Institute of Technology (IIT), Madras, India, and the M.A. and C.Phil degrees in mathematics and the M.S. and Ph.D. degrees in electrical engineering in 1983, 1984, 1982, and 1986 respectively, all from the University of California, Berkeley.

He is currently an Associate Professor with the Department of Electrical Engineering and Computer Science at the University of California at Berkeley. He was with the School of Electrical Engineering at Cornell University, Ithaca, NY, from 1986 to 1994.

Dr. Anantharam is the recipient of the Phillips India Medal and the President of India Gold Medal from the IIT, an NSF Presidential Young Investigator Award, and an IBM Faculty Development Award. He was a plenary speaker at the 18th Bernoulli Society Conference on Stochastic Processes and Their Applications in 1989, and delivered an invited short course on Stochastic Networks at the Fifth Workshop on Stochastic Analysis of Oslo-Silivri held in Silivri, Turkey, in July 1994.