Some Optimal Traffic Regulation Schemes for ATM Networks: A Markov Decision Approach

Mohamed Abdelaziz and Ioannis Stavrakakis, *Senior Member, IEEE*

***Abstract*—In this paper, some new traffic regulation schemes are defined in terms of a relief-spacing (or spacing of the allowance for cell delivery to the network) function. The class of open-loop traffic regulators (TR's) is defined in terms of relief-spacing functions which depend on some user-state; this class may be viewed as an extension of the Spacer-Controller defined in terms of some constant (user-state independent) relief-function. The optimal open-loop TR's are derived by formulating proper optimization problems and applying a Markov decision approach. Numerical results illustrate the improved performance of the optimal open-loop TR over that of the (constant relief-spacing) Spacer-Controller. Finally, the class of closed-loop TR's is defined in terms of relief-spacing functions which depend on both some user- and some network- state information and its optimal element is derived. The improved performance under the optimal closed-loop TR over that of the optimal open-loop TR is illustrated and their difference determines the performance gain if feedback information can become available on time.**

I. Introduction

PREVENTIVE CONTROL is considered to be a promising approach for traffic congestion management in the emerging high-speed asynchronous transfer mode (ATM) networks. An extended survey of mechanisms—called bandwidth enforcement mechanisms or traffic regulators (TR)-developed for the implementation of such control may be found in [2]. The object function of a TR is to control the flow of the user traffic to the network in a way that unacceptable network congestion be avoided; no network state information is assumed to be available to the users.

The prevailing approach for the design of efficient TR's has been the following. At first, certain key measures of basic traffic characteristics—which have a significant impact on the network performance—are identified. For instance, such measures can be the cell rate and burst length of the user-traffic. Then, the user and the network agree on some limits on the values of these measures (contract), so that the network be able to plan based on the maximum expected amount of stress (potential for congestion) coming from the particular user. Network congestion may then be controlled through an effective TR, whose basic function is to provide for the smoothing of the user traffic and prevent network congestion. A TR may also be seen as a mechanism which enforces compliance with the contract agreements. Smoothing of the traffic occurs by not allowing the values of key measures associated with the actual traffic delivered to the network to exceed some predetermined limits. It also leads to a higher utilization of network resources [3]. There is a trade-off relationship between the regulation level of the user traffic and the improvement of network performance. While the network performance can be improved through increased regulation of the user traffic streams, excessive regulation may result in unacceptable performance at the user premises as illustrated in [1].

The insightful description of the basic functions that need to be implemented by a TR, presented in [1], as well as numerical results, have pointed to the ineffectiveness of the leaky bucket TR and the improved performance delivered by Spacer-Controller type TR's [1], [10], [14], [17]. By defining the relief rate (or intensity, or function) to be the rate of *allowance* for cell delivery to the network, the class of $\sigma - Relief$ TR's was introduced in [1], as a practical implementation of a fixed (and equal to σ) relief function in the slotted ATM environment.

In this paper, the class of constant $\sigma - Relief$ TR's introduced in [1], is extended to include TR's implementing nonfixed, state-dependent relief functions. The class of user-state dependent traffic regulation schemes is defined and studied. An open-loop relief-spacing regulation scheme is defined to be one whose relief function depends only on user-state information; the user-state is defined in terms of the user-queue and source states. The optimal elements in the resulting class of the TR's are identified by defining some meaningful optimality criteria and following a Markov decision approach. Finally, the class of closed-loop relief-spacing TR's is also defined and studied. The associated relief function depends on both user and network state information; the latter is defined in terms of the state of the network-queue receiving the regulated user-traffic. It should be emphasised that the closed-loop traffic regulation approach is not proposed as an alternative to the open-loop one unless the user and the network buffers are not far apart and the propagation delay is assumed to be negligible. Rather, the closed-loop approach is considered to establish the magnitude of performance improvement that can be achieved through the availability of network-state information and to also gain insight towards the development of more efficient open-loop schemes or hybrid schemes.

In [12], an open-loop regulation concept was introduced by proposing a two-level shaper for the regulation of the traffic generated by an on/off Markov source. Two different

Manuscript received November 1993; revised April 1994; approved by IEEE/ACM Transactions on Networking Editor H. Miyahara. This work was supported in part by the Advanced Research Project Agency (ARPA) under Grant F49620-93-1-0564 monitored by the Air Force Office of Scientific Research (AFOSR).

The authors are with the Department of Computer Science and Electrical Engineering of the University of Vermont, Burlington, VT 05405 USA.

IEEE Log Number 9404854.

1063-6692/94$04.00 © 1994 IEEE

cell-delivery (service) rates were considered in a continuous-time setup: a service rate equal to the source peak cell rate when the user-queue is full, to guarantee loss-free performance at the user premises; a lower service rate (referred to as intermediate), otherwise. The work presented in [12] was extended in [13] where an M-level shaper was introduced and analyzed for M =2,3. Although the problems considered in [12], [13] may be viewed as special cases of the general open-loop formulation presented in this paper, no attempts were made to identify optimal elements in the limited class of shapers considered there.

II. The Open-loop Traffic Regulation Scheme

Almost the entirety of the TR's which have been proposed in the past can be classified as open-loop since their regulating parameters do not depend on the network-state. With the exception of [12], [13], as explained in the introduction, the open-loop TR's proposed in the past are also user-state-independent. In this section, the class of open-loop, user-state-dependent, relief-spacing TR's is introduced and its performance at both the user and network premises is evaluated.

In this paper, an allowance for cell delivery to the network provided by the TR is defined to be a unit of relief and, thus, the TR's are characterized as relief-spacing; a relief unit is also equivalent to a user-service unit. A relief-spacing TR is defined to be one which is completely determined in terms of its relief-spacing function, defined on some appropriate space. The $\sigma - Relief$ TR in [1], is completely determined by the constant value σ of the relief intensity function. To avoid practical problems associated with the implementation of relief-rates in a time-slotted environment, the relief-spacing TR's considered here, are defined in terms of the relief-spacing function (in slots), $T(.)$, rather than the not always feasible relief rate $\sigma(.)$; $T(.)$ is in general, a function of some selected system state, and takes positive integer values.

Let S_s and S_u denote the state-space of the Markov source to be regulated and the state space of the user-queue-occupancy process, respectively. An open-loop, user-state-dependent relief-spacing TR is completely defined in terms of the associated function $T(s,q_u),(s,q_u)\epsilon S_s \times S_u$ which determines the next slot at which a relief unit becomes available, given that the system's state is (s,q_u) in the slot at which the latest relief-unit was available. The class of such TR's is described in terms of the class of all possible mappings:

$$T : S_s \times S_u \rightarrow Z_B^+ \equiv \{1,2,...,B\}, B < \infty \qquad (1)$$

and is denoted by $\langle R - T(s,q_u)\rangle$, that is,

$$\langle R - T(s,q_u)\rangle \equiv \{T(s,q_u),(s,q_u)\epsilon S_s \times S_u : T(s,q_u)\epsilon Z_B^+\}.$$

An example of a relief-spacing function $T(s,q_u)$ is shown in Fig. 1. A Spacer-Controller type of TR [1], [10], [14], [17] can easily be viewed as a relief-spacing TR with constant relief-spacing function T. Such a TR does not utilize locally available user-state information and is expected to deliver suboptimal performance. The performance of the proposed class of open-loop TR's defined in terms of the general relief-spacing function $T(s,q_u)$ is considered in the next two subsections.

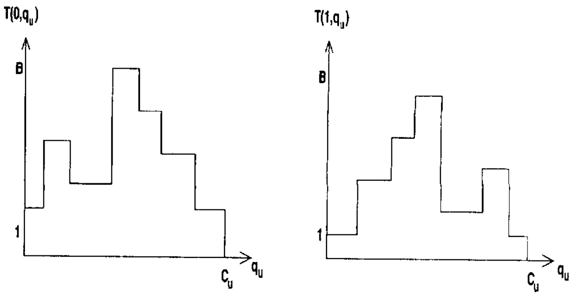


Fig. 1. An example of the open-loop regulation function.

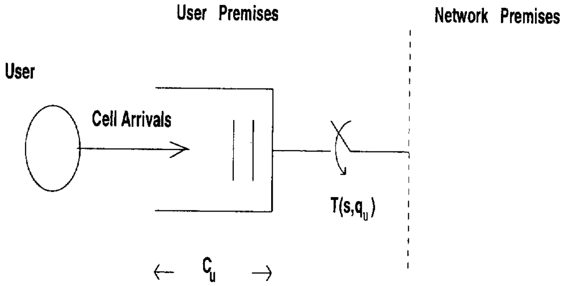


Fig. 2. A queueing model for the relief-spacing TR.

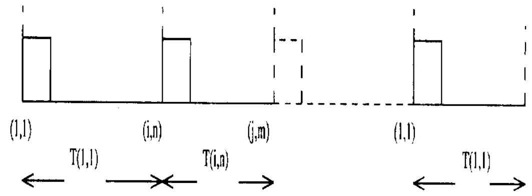


Fig. 3. A service pattern realization for the open-loop, relief-spacing TR.

A. Performance Analysis at the User Premises for the Open-Loop Relief-Spacing TR

A queueing model for a relief-spacing TR is shown in Fig. 2. The server provides service (relief) to the finite-capacity user-queue of size C_u according to a pseudo-periodic pattern determined by the relief-spacing function $T(s,q_u)$, as follows. Let t denote the beginning of a slot at which the server visits the queue and let s and q_u be the state of the user-source and user-queue occupancy process, respectively, at that time. The server remains at the user queue for one time slot and provides one unit of service if a cell is available. Then the server takes a vacation of length $T(s,q_u)-1$ and returns to the user-queue at time $t+T(s,q_u)$; if $T(s,q_u)-1=0$, the server is assumed to return to the user-queue instantaneously. If the user queue is empty at the time of server visit, the server will only switch away after providing service to the first arriving cell. A service pattern realization in time is shown in Fig. 3.

In the sequel, the cell loss probabilities induced at the user premises are calculated for the queueing system shown in

Fig. 2 under an arbitrary service (relief-spacing) policy defined by the mapping (1).

Let $S_s = \{0, 1\}$ denote the state space of the two-state Markov source generating one cell when in state 1 (on) and no cells when in state 0 (off); the transition probabilities from state i to state j are denoted by $p(i, j)$, the stationary state probabilities by $\pi(i)$, the source cell rate by λ and the source burstiness coefficient is defined by $\gamma = p(1, 1) - p(0, 1)$. Let $S_u = \{1, .., C_u\}$ denote the state space of the user-queue occupancy process, assuming some finite user-queue capacity C_u.

Let $\{I_l, Q_l^u\}_{l \geq 0}$ denote the two-dimensional process describing the state of processes $\{I_l\}_{l \geq 0}$ and $\{Q_l^u\}_{l \geq 0}$ at the beginning of the slot at which the lth cell departure from the user-queue occurs (departure slot), where $\{I_l\}_{l \geq 0}$ ($\{Q_l^u\}_{l \geq 0}$) denote the state of the source (user-queue-occupancy) at the beginning of the lth departure slot; a relief unit is available to the user-queue at that time slot. It is assumed that cell departures from the user-queue occur at the end of a departure slot. The state of the source is declared at the beginning of a slot, but cell generation due to the visit to that state are declared at the end of the slot. It may be easily established that $\{I_l, Q_l^u\}_{l \geq 0}$ is a Markov chain embedded at the beginning of the cell departure slots. Let $A = [a(i_1, n_1, i_2, n_2)]$ denote the transition probability matrix of $\{I_l, Q_l^u\}_{l \geq 0}, i_1, i_2 \epsilon S_s, n_1, n_2 \epsilon S_u$. Calculation of these probabilities (leading to the stationary probabilities of $\{I_l, Q_l^u\}_{l \geq 0}$) will be facilitated by the description of the length of the sequence of interdeparture intervals at the boundaries of which $\{I_l, Q_l^u\}_{l \geq 0}$ is defined. Furthermore, the average length of interdeparture intervals will be utilized directly for the derivation of the cell loss probability, as it will be shown later.

Let a_m be a random variable describing the number of cells generated by the source in m consecutive time slots; let $(i \overset{m}{\rightarrow} j, a_m = k)$ denote the joint event that the source state moves from i to j in m steps and $a_m = k$, with probability $f^m(i, j, k)$, $i, j \epsilon S_s, 0 \leq k \leq m$, derived in Appendix A. Let $ID(i, n), (i, n) \epsilon S_s \times S_u$, denote the time in slots between two consecutive cell departures, given that $\{I_l, Q_l^u\}_{l \geq 0}$ was in state (i,n) at the beginning of the first one. It is easy to express ID(i, n) in terms of the relief-spacing function $T(i, n)$ governing the service of the user-queue, as shown below.
For $i \epsilon S_s$:

$$ID(i,n) = \begin{cases} T(i,n) & \text{if } n \geq 2 \\ T(i,1) & \text{if } n = 1, \\ & (i \overset{T(i,1)}{\rightarrow} j, a_{T(i,1)} = k), 1 \leq k \\ T(i,1)+1 & \text{if } n = 1, (i \overset{T(i,1)}{\rightarrow} 1, a_{T(i,1)} = 0) \\ T(i,1)+k & \text{if } = 1, (i \overset{T(i,1)}{\rightarrow} j, a_{T(i,1)} = 0), \\ & (0 \overset{k-2}{\rightarrow} 0, a_{k-2} = 0), \\ & (0 \overset{1}{\rightarrow} 1, a_1 = 0), 2 \leq k. \end{cases}$$

Starting from some departure slot in state (i, n) with $n \geq 2$, the next departure slot will appear exactly after $T(i, n)$ slots. The same holds if $n = 1$ and $a_{T(i,1)} \geq 1$. If $n = 1$ and $a_{T(i,1)} = 0$, the next departure slot will appear after $T(i, 1)$ slots plus the time it will take for the source to switch to state 1 afterwards. The expected value of ID(i, n), $\overline{ID}(i, n)$, (to be utilized in the cell loss probability calculation) is obtained by applying the expectation operator to the above expression.
For $i \epsilon S_s$:

$$\begin{aligned} \overline{ID}(i,n) &= T(i,n), n \geq 2 \\ \overline{ID}(i,1) &= T(i,1) \textstyle\sum_{k=1}^{T(i,1)} \sum_{j \epsilon S_s} f^{T(i,1)}(i,j,k) \\ &+ T(i,1) f^{T(i,1)}(i,1,0) \\ &+ \textstyle\sum_{j \epsilon S_s} \sum_{k=2}^{\infty} (T(i,1)+k) f^{T(i,1)}(i,j,0) \\ &\quad p^{k-2}(0,0) p(0,1) \\ &= T(i,1) + f^{T(i,1)}(i,1,0) + \tfrac{f^{T(i,1)}(i,0,0)}{p(0,1)}. \end{aligned} \tag{2}$$

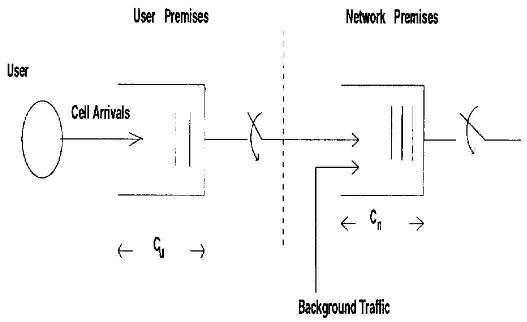


Fig. 4. The integrated queueing system.

Following a similar reasoning to that in the description of the interdeparture process, the following expressions for the transition probabilities $a(i_1, n_1, i_2, n_2), i_1, i_2 \epsilon S_s, n_1, n_2 \epsilon S_u$, of $\{I_l, Q_l^u\}_{l \geq 0}$ can be easily obtained.

For $i_1 \epsilon S_s, i_2 \epsilon S_s, n_1 \geq 2$:

$$\begin{aligned} a(i_1, n_1, i_2, n_2) &= f^{T(i_1,n_1)}(i_1, i_2, n_2 - n_1 + 1), \\ &\quad n_1 - 1 \leq n_2 \leq C_u - 1 \\ a(i_1, n_1, i_2, C_u) &= \textstyle\sum_{k=C_u-n_1+1}^{T(i_1,n_1)} f^{T(i_1,n_1)}(i_1, i_2, k). \end{aligned}$$

For $i_1 \epsilon S_s, i_2 \epsilon S_s$:

$$\begin{aligned} a(i_1, 1, i_2, 1) &= f^{T(i_1,1)}(i_1, i_2, 1) + p(1, i_2) \times \\ &\quad (f^{T(i_1,1)}(i_1, 1, 0) + f^{T(i_1,1)}(i_1, 0, 0)) \\ a(i_1, 1, i_2, n_2) &= f^{T(i_1,1)}(i_1, i_2, n_2), 2 \leq n_2 \leq C_u - 1 \\ a(i_1, 1, i_2, C_u) &= \textstyle\sum_{k=C_u}^{T(i_1,1)} f^{T(i_1,1)}(i_1, i_2, k), \\ &\quad C_u \leq k \leq T(i_1, 1). \end{aligned}$$

The steady-state probabilities $\pi(i, n), i \epsilon S, n \epsilon S_u$ are obtained from the matrix equation $\overline{\Pi} = A\overline{\Pi}$, $\overline{\Pi} = [\pi(i, n)]$ with the normalizing condition $\overline{\Pi}\overline{e}' = 1$ where $\overline{e}$ is a unity vector. The cell loss probability L_u at the user premises can then be obtained from the expression:

$$L_u = \frac{\sum_{i \epsilon S_s} \sum_{n \epsilon S_u} L_u(i,n) \pi(i,n)}{\lambda \sum_{i \epsilon S_s} \sum_{n \epsilon S_u} \overline{ID}(in) \pi(i,n)} \tag{3}$$

where $L_u(i, n)$ denotes the expected number of cells lost over the interval $ID(i, n)$ and is given by

$$L_u(i,n) = \sum_{j \epsilon S_u} \sum_{k=C_u-n+1}^{T(i,n)} (k - C_u + n - 1) f^{T(i,n)}(i,j,k). \tag{4}$$

B. Performance Analysis at the Network Premises for the Open-Loop Relief-Spacing TR

To establish the performance of the relief-spacing TR at the network premises, the network-queue receiving the regulated user traffic is studied. The integrated system of user-queue and network-queue shown in Fig. 4, is considered for this purpose. The network-queue is assumed to be served according to the First-In–First-Out (FIFO) service policy and to be of finite capacity C_n. In addition to the regulated user traffic, the network-queue is also fed by a background Bernoulli cell traffic, with rate λ_b, representing other network traffic. Primarily to introduce a new class of TR's and develop a procedure for the determination of its optimal elements, the simple Bernoulli model is considered for the background traffic; other models can also be considered as explained later.

In [12], the performance of the proposed traffic shapers is analyzed using an approximate model. The parameters of the output shaped traffic are matched to a 3-state Markov chain model and a network queue fed by a number of such Markov chains is studied to yield an approximate expression for the cell loss probability. Similar approximate analyses for a multiplexer loaded by a set of leaky bucket regulated traffic streams are presented in [13], [16], where a similar traffic matching technique is employed. Contrary to the above mentioned approaches, the analysis adopted in this paper is exact, since no approximations are involved in the description of the regulated traffic. The transition probabilities of the integrated model and the cell loss probability at the network premises are obtained using a similar analysis to that presented in Section II-A and can be found in Appendix B.

The number of states in the state space of the three dimensional process governing the behavior of the system—see Appendix B—is $2 \times C_u \times C_n$, thus requiring large storage space for the resulting transition probability matrix A. In order to solve for systems with larger than 2000 states, the renewal based recursive equations technique used in [1], was used. In that technique, the storage space requirement is eliminated by computing the transition probabilities—instead of storage in some matrix—in each iteration of the recursive algorithm, thus trading computing time for storage space. Systems containing as many as 5000 states were solved in conveniently short time.

III. Optimal Relief-spacing Traffic Regulation Schemes

The class of open-loop, user-state-dependent relief-spacing TR's, $\langle R - T(s, q_u)\rangle$, contains the sub-class of fixed (user-state-independent) relief-spacing TR's (Spacer-Controllers). It will be interesting to compare the performance of the "best" performing $R - T(s, q_u)$ TR to that of Spacer-Controllers, to establish the magnitude of performance improvement that can be achieved by using $R - T(s, q_u)$ TR's. In the next subsection the optimal element in $\langle R - T(s, q_u)\rangle$ is identified by formulating a Markov decision based optimization problem. Some numerical results are presented in Section III-B.

A. Optimal TR's Based on a Markov Decision Approach

The objective in regulating the user-traffic in an ATM (multiplexing) environment is to increase the network utilization while guaranteeing the Quality of Service (QoS) of all supported users. An effective TR delivers to the network a user-traffic that induces low level of network stress. The network then will potentially be in a position to not only provide the necessary QoS but also accommodate a larger number of services (users).

Let the Quality of Service Margin (QoSM) be defined as the maximum amount of "disturbance" of the user traffic which can be tolerated before the necessary QoS is not provided. The QoSM is basically described in terms of probabilistic measures on cell delay and/or loss (as for the QoS). Let QoSMR_u and QoSMR_n denote the value of the QoSM Reduction at the user and network premises, respectively. A TR may be adopted only if

$$\mathrm{QoSMR}_u + \mathrm{QoSMR}_n \leq \mathrm{QoSM}.$$

Note that by increasing (decreasing) the level of traffic regulation, QoSMR_u increases (decreases). Thus, an optimal TR could be defined to be the one balancing these trends optimally, minimizing $\mathrm{QoSMR}_u + \mathrm{QoSMR}_n$. Such an optimality criterion may not lead necessarily to a well performing TR in a real networking environment, due to network traffic unpredictability and the lack of direct consideration of QoS requirements associated with other supported services. In addition, this optimality criterion does not seem to lead to a tractable optimization formulation which would identify the optimal TR in a given class. Since QoSMR_u basically measures the performance of the TR at the user premises, it can be calculated in terms of the TR and the source characteristics. Thus, a feasible optimization problem, leading to the derivation of some optimal TR, could be set up in terms of the following objective:

$$\min: \ \mathrm{QoSMR}_n \quad \text{subject to: } \mathrm{QoSMR}_u \leq w. \qquad (5)$$

In this paper, the cell loss probability is considered to be the measure of the provided QoS. $w = 0$ corresponds to an optimization formulation identifying the loss-free at the user premises TR, which minimizes the network induced losses. It should be noted that the optimization formulation in (5), allows for the consideration of QoS measures (QoSMR units) which are different at the user and network premises. This is particularly appealing for the derivation of optimal open-loop TR's, as it will be seen below.

Let L_u and L_n denote the cell loss probabilities at the user and network premises, respectively. The optimal element in the class of open-loop, user-state-dependent relief-spacing TR's, $\langle R - T(s, q_u)\rangle$, according to the criterion in (5) will be:

$$R_5 - T(s, q_u) = \arg\{\min_{T: S_s \times S_u \to Z_B^+} \{L_n\} : L_u \leq w\}. \qquad (6)$$

Since $B^{2 \times C_u}$ elements are contained in $\langle R - T(s, q_u)\rangle$, it is apparent that exhaustive search for $R_5 - T(s, q_u)$ as described above, is unrealistic for most practical cases. A different and more tractable formulation of the optimization problem for the open-loop case is developed below.

In view of the definition of the class $\langle R - T(s, q_u)\rangle$ and the presentation in Section II, it is easy to establish that given a function $T(s, q_u)$, the evolution of the semi-Markov process $\{I_l, Q_l^u\}_{l\geq 0}$ is probabilistically completely determined. Similarly, given the state of $\{I_l, Q_l^u\}_{l\geq 0}$, the evolution of the relief-spacing process $\{T_l\}_{l\geq 0}$ is completely determined; T_l denotes the relief-spacing applied at the $\underline{lth}$ departure slot, where $T_l = T(I_l, Q_l^u)$. By interpreting the relief-spacing process $\{T_l\}$ as a set of decisions, each taken upon a visit to some state (s, q_u) of the Markov process $\{I_l, Q_l\}$, and by associating a certain cost with each decision, a total cost associated with a given relief-spacing process $\{T_l\}$ can be computed. Thus the optimal relief-spacing function (or $R - T(s, q_u)$ TR), in the sense that it minimizes some performance measure (cost), can be identified by following a semi Markov decision approach, provided that the cost during decision instances can be determined completely in terms of the action $T(I_l, Q_l^u)$ taken and the state of $\{I_l, Q_l^u\}_{l\geq 0}$. The advantage of such a formulation of the optimization problem is that non-exhaustive, computationally appealing approaches exist for the derivation of the optimal policy, such as the policy and value iteration algorithms and the linear programming approach [15].

Unfortunately, the cost associated with the optimization problem described in (5), that is cell loss at the network premises, cannot be determined by the state of $\{I_l, Q_l^u\}_{l\geq 0}$ and $\{T_l\}_{l\geq 0}$ at the departure (decision) instants, since the state of the network-queue is not available in the open-loop case. In view of the Bernoulli model for the background traffic (see Section II–B), knowledge of the state of $\{I_l, Q_l^u, Q_l^n\}_{l\geq 0}$ (see Appendix B) would be sufficient. The approach can be easily extended to incorporate other types of background processes such as a batch arrival process. This will require the substitution of the proper values for the set of probabilities $b^T(k)$ used in Appendix B to describe the probability that k background arrivals occur in T time slots. Correlated models for the background traffic (e.g. a Markov source) can also be incorporated at the expense of increased system size.

To develop a semi Markov decision formulation of the problem of deriving the optimal open-loop TR defined here, a network performance measure (cost)—to be minimized—which can be determined by the state of the semi-Markov process $\{I_l, Q_l^u\}_{l\geq 0}$, and the action taken, is considered. In this section, the variance of the cell interdeparture process is considered as a measure of the smoothness of the regulated traffic and the potentially induced network congestion. Thus, the optimal $R - T(s, q_u)$ is defined by

$$R_v - T(s, q_u) = \arg\{\min_{T:S_s\times S_u\to Z_B^+}\{v(T)\} : L_u \leq w\} \quad (7)$$

where $v(T)$ denotes the variance of the cell interdeparture process given by

$$v(T) = \frac{\sum_{(i,n)\epsilon S_s\times S_u} c(i,n,T(i,n))\pi(i,n)}{\sum_{(i,n)\epsilon S_s\times S_u} \tau(i,n,T(i,n))\pi(i,n)}$$

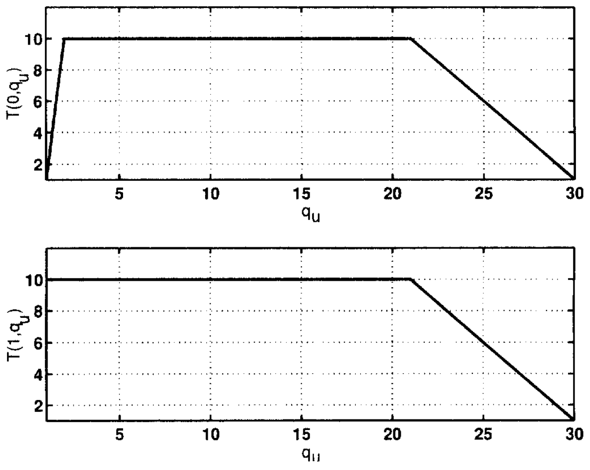


Fig. 5. Relief-spacing function for $R_v - T(s, q_u)$.

where $\tau(i, n, T(i, n)) = \overline{ID}(i, n)$ is the expected time until the next decision (departure) instant given that $T(i, n)$ is the action (relief spacing) chosen in the present decision instant at which the semi-Markov process is in state (i, n); $c(i, n, T(i, n)) = (\overline{ID}(i, n) - \frac{1}{\lambda})^2$ is the expected cost until the next decision instant given that $T(i, n)$ is the action chosen in the present decision instant at which the semi-Markov process is in state (i, n). The linear programming algorithm outlined in Appendix C is used to identify the optimal element $R_v - T(s, q_u)$.

B. Numerical Results—Fixed Versus Optimal Relief-Spacing

In this section, some numerical results are presented on the performance of some optimal $\langle R - T(s, q_u)\rangle$ TR's. To illustrate the performance improvement obtained by utilizing the user-state-dependent relief-spacing functions, the subclasses

$$\langle R - T\rangle \equiv \{T(s, q_u) = c, (s, q_u)\epsilon S_s \times S_u : c\epsilon Z_B^+\}$$

and

$$\langle R - T(q_u)\rangle \equiv$$
$$\{T(s, q_u) = T(q_u), (s, q_u)\epsilon S_s \times S_u : T(q_u)\epsilon Z_B^+\}$$

are also considered. Note that the relief-spacing function is user-state-independent in $\langle R - T\rangle$ and is source-state-independent in $\langle R - T(q_u)\rangle$. Let $R_v - T$ and $R_v - T(q_u)$ denote the optimal elements in these classes in the sense that they minimize the variance of the cell interdeparture process; $R_v - T$ can be easily derived through an exhaustive search procedure; the linear programming algorithm has been used for the derivation of $R_v - T(q_u)$ and $R_v - T(s, q_u)$.

Fig. 5 shows the relief-spacing function for $R_v - T(s, q_u)$ for loss-free performance at the user premises [$w = 0$ in (5)], for $\lambda = 0.09$, $\gamma = 0.8$ and $C_u = 30$; a maximum relief-spacing $B = 10 \approx \frac{1}{\lambda}$ has been considered. The improvement in performance obtained through the adoption of the user-state-dependent relief-spacing TR's is illustrated in Fig. 6, where the cell loss probability at the network premises, L_n, is plotted for the loss-free optimal policies $R_v - T, R_v - T(q_u)$

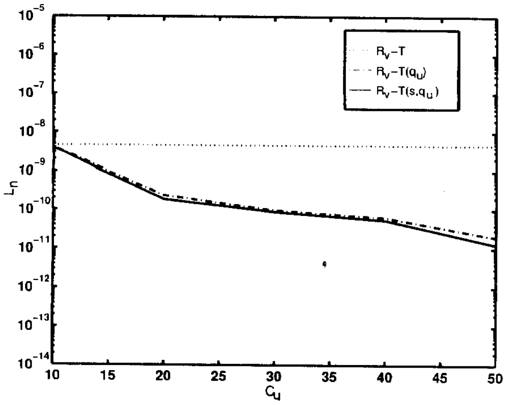


Fig. 6. Cell loss probability L_n versus C_u, induced by R_v-T, $R_v-T(q_u)$, and $R_v-T(s,q_u)$.

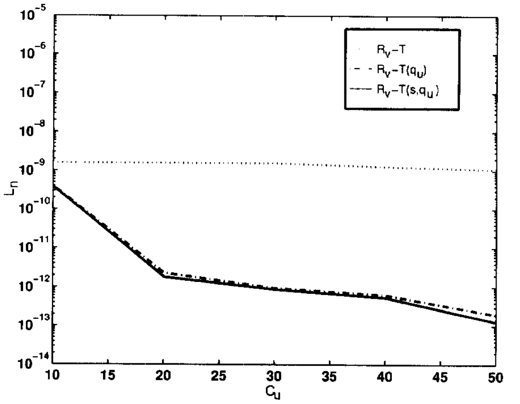


Fig. 7. Cell loss probability L_n versus C_u, induced by R_v-T, $R_v-T(q_u)$, and $R_v-T(s,q_u)$ for $L_u \leq 10^{-8}$.

and $R_v - T(s,q_u)$; $C_n = 50, 10 \leq C_u \leq 50$, $\lambda = 0.09$, $\gamma = 0.8$ and $\lambda_B = 0.8$. It can be observed that lower cell loss probabilities L_n can be achieved by $R_v - T(q_u)$ or $R_v - T(s,q_u)$ as C_u increases.

The performance of $R_v - T$, $R_v - T(q_u)$ and $R_v - T(s,q_u)$ under the constraint $L_u \leq 10^{-8}$ is presented in Fig. 7; the system characteristics are otherwise identical to those of the system associated with Fig. 6, where $L_u = 0$. Notice that a lower value of L_n can be achieved if losses are allowed at the user premises compared to that under loss-free performance at the user premises. The tradeoff between L_n and L_u is shown in Fig. 8; $C_n = 50$, $C_u = 40$, $\lambda = 0.09$, $\gamma = 0.6$ and $\lambda_b = 0.8$. It can be observed that L_n increases as L_u decreases, implying that a very low constraint on L_u may result in substantially increased value of L_n, resulting in a degradation of the overall performance.

IV. The Closed-loop Traffic Regulation Scheme

Motivated by the potential for improved performance when network state information can be available at the user premises, the class of closed-loop relief-spacing traffic regulation schemes is introduced in this section. Unlike the class of open-loop TR's presented in Section II, the relief-spacing function associated with the closed-loop TR's is assumed to depend on the network-state as well as the user-state.

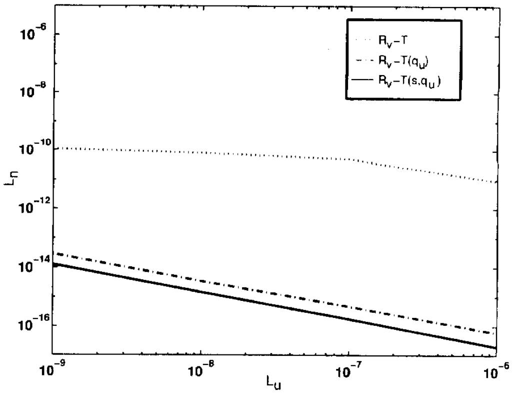


Fig. 8. L_n versus L_u tradeoff.

In this paper, the network premises has been defined to consist of the network-queue receiving directly the regulated user-traffic together with some Bernoulli background traffic. With this definition, the network-state is sufficiently described by the network-queue occupancy process $\{Q_l^n\}_{l \geq 0}$. Accordingly, if the user and the network queues are not far apart, it may not be unrealistic to assume that such information can be timely available, and therefore the effects of propagation delay are assumed to be negligible. In fact, such network-feedback-based traffic regulation schemes have been recently proposed, in an attempt to improve on the relatively poor performance of totally preventive congestion control approaches [18]. More complex network models incorporating both propagation delay effects and procedures for network state determination are beyond the scope of this work and will be considered in future studies.

The queueing model of Fig. 4, considered for the study of the closed-loop regulation scheme, is interesting in its own, with potential applications elsewhere. The optimal closed-loop, user-state-dependent relief-spacing TR minimizing, for instance, total cell losses, could provide for an optimal policy for cell load distribution in a cascade of two queues (of capacities C_u and C_n) with external inputs and/or for an optimal distribution of some available total buffering capacity $C = C_u + C_n$ to the two queues.

A. Formulation of the Optimization Problem

A closed-loop, user-state-dependent relief-spacing TR is defined in terms of the associated relief-spacing function $T(s,q_u,q_n), (s,q_u,q_n)\epsilon S_s \times S_u \times S_n$ and is denoted by $R - T(s,q_u,q_n)$. The class of such TR's is defined in terms of all possible mappings

$$T : S_s \times S_u \times S_n \rightarrow Z_B^+ = \{1,2,...,B\}, B < \infty$$

and is denoted by

$$\langle R - T(s,q_u,q_n)\rangle \equiv$$
$$\{T(s,q_u,q_n),(s,q_u,q_n)\epsilon S_s \times S_u \times S_n : T(s,q_u,q_n)\epsilon Z_B^+\}.$$

In view of the discussions in Sections II and III and the presentation in Appendix B, it is easy to establish that a semi-Markov decision approach can be followed for the derivation of optimal elements in $\langle R - T(s, q_u, q_n)\rangle$. The semi-Markov process $\{(I_l, Q_l^u, Q_l^n)\}_{l\geq 0}$ defined at departure slots is used to completely describe the evolution of the system under a given relief-spacing function $T(s, q_u, q_n)$ and determine the induced performance measure (cost) involved in the optimization problem formulation. The optimal TR's in $\langle R - T(s, q_u, q_n)\rangle$—($< R_t - T(s, q_u, q_n)\rangle$ and $\langle R_w - T(s, q_u, q_n)\rangle$—are derived in the following two senses:

$$R_t - T(s, q_u, q_n) = \arg\{\min_{T: S_s \times S_u \times S_n \rightarrow Z_B^+} \{L_u + L_n\}\} \quad (8)$$

$$R_w - T(s, q_u, q_n) = \arg\{\min_{T: S_s \times S_u \times S_n \rightarrow Z_B^+} \{L_n\} : L_u \leq w\}. \quad (9)$$

Notice that the calculation of L_n in terms of the semi-Markov process and the action taken, is now possible and, thus, it can be part of the quantity to be minimized. This is contrary to the open-loop case where only some traffic smoothness measure (determined by the user-state) was possible to consider.

The optimization formulation in (9), implies that the corresponding closed-loop relief-spacing TR's do not treat the user-traffic as a best effort service since it guarantees a certain cell loss probability L_u. This formulation is a constrained optimization one, which may be more suitably tackled by employing the linear programming approach leading to the derivation of the optimal element $R_w - T(s, q_u, q_n)$. However, since the state-space of the semi-Markov process $\{I_l, Q_l^u, Q_l^n\}_{l\geq 0}$ can be very large ($2 \times C_u \times C_n$ states) and, consequently, the number of variables ($2 \times B \times C_u \times C_n$) and constraints which need to be stored under the linear programming algorithm, this approach can be easily made intractable. This is not the case with the unconstrained optimization problem (8), where the more efficient value iteration algorithm presented in Appendix D is employed for the derivation of $R_t - T(s, q_u, q_n)$.

By utilizing the fact that no losses can occur at the user-queue when the relief spacing $T(s, q_u, q_n)$ is less than the remaining queue capacity ($C_u - q_u + 1$), then, $R_0 - T(s, q_u, q_n)$ which is the optimal loss-free element at the user premises (($w = 0$) in $R_w - T(s, q_u, q_n)$) can be simply identified as a member of the sub-class of loss-free at the user premises closed-loop, user-state-dependent relief-spacing TR's, $\langle R_f - T(s, q_u, q_n)\rangle$, determined by,

$$\begin{aligned}&\langle R_f - T(s, q_u, q_n)\rangle \\ &\quad \equiv \{T(s, q_u, q_n), (s, q_u, q_n) \epsilon S_s \times S_u \times S_n \\ &\quad : T(s, q_u, q_n) \epsilon Z_B^+, T(s, q_u, q_n) \leq C_u - q_u + 1\} \quad (10)\end{aligned}$$

and therefore, $R_0 - T(s, q_u, q_n)$ can be obtained by solving the equivalent unconstrained optimization problem,

$$R_0 - T(s, q_u, q_n) = \arg\{\min_{T: R - T(s, q_u, q_n) \epsilon \langle R_f - T(s, q_u, q_n)\rangle} \{L_n\}\} \quad (11)$$

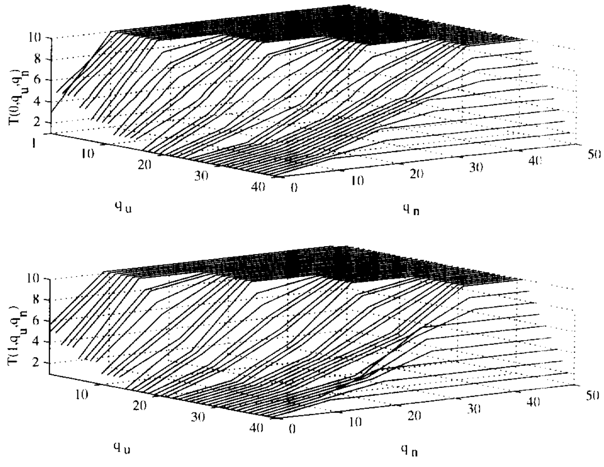


Fig. 9. Relief-spacing function for $R_0 - T(s, q_u, q_n)$.

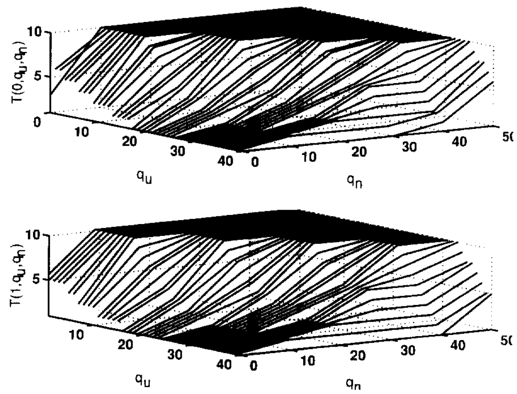


Fig. 10. Relief-spacing function for $R_t - T(s, q_u, q_n)$.

by applying the more efficient value iteration algorithm with a cost defined in terms of number of cells lost at the network premises over a decision interval.

B. Numerical Results

Fig. 9 presents the relief-spacing function for $R_0 - T(s, q_u, q_n)$ for $C_u = 40, C_n = 50, \lambda = 0.09, \gamma = 0.8, \lambda_B = 0.6$ and $B = 10$. Notice that as q_n increases, the optimal relief-spacing T approaches its limit $B = 10$ while guaranteeing no losses at the user premises; for instance, when $q_u = C_u$ then $T = 1$ to guarantee no losses.

Fig. 10 presents the relief-spacing function for $R_t - T(s, q_u, q_n)$for $C_u = 40, C_n = 50, \lambda = 0.09, \gamma = 0.8, \lambda_B = 0.6$ and $B = 10$. The optimal tradeoff between L_u and L_n that $R_t - T(s, q_u, q_n)$ achieves, leads to a maximum relief-spacing when q_u is small and q_n moderate or large, and minimum relief-spacing when q_u is large and q_n is small or moderate. It is also observed that for large values of q_u or q_n, the relief-spacing is larger when the source is off, compared to that when the source is on, since more losses are more likely to occur in the immediate future when the source is on.

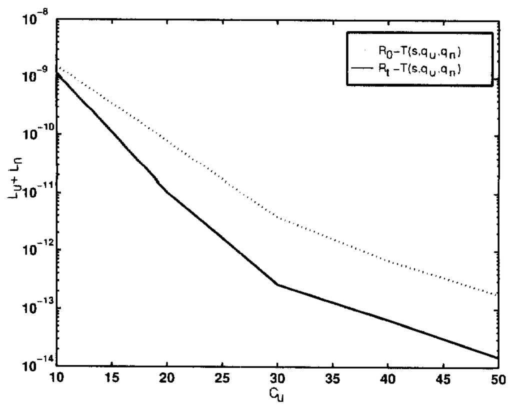


Fig. 11. $L_u + L_n$ versus C_u under $R_0 - T(s, q_u, q_n)$ and $R_t - T(s, q_u, q_n)$.

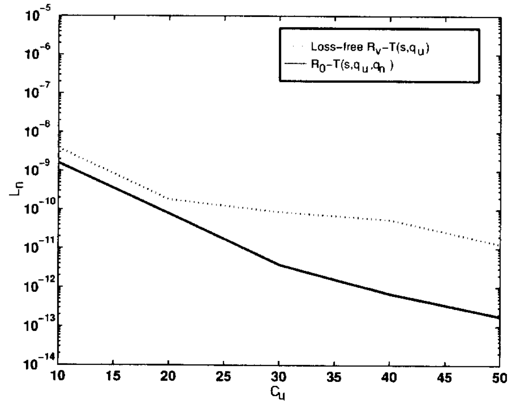


Fig. 12. L_n versus C_u for the optimal open- and closed-loop TR's.

In both Figs. 9 and 10, it may be observed that the optimal policies tend to operate under low value of relief-spacing for small values of both q_u and q_n, keeping the steady-state probability that the system operates in the region of larger values of q_u and q_n small. That is, as expected, the optimal policies tend to operate less in the critical (loss inducing) region, by providing faster relief when in noncritical regions (small buffer occupancies).

In Fig. 11, the relative performance of $R_0 - T(s, q_u, q_n)$ and $R_t - T(s, q_u, q_n)$ is presented, as a function of C_u. As expected, $R_t - T(s, q_u, q_n)$ outperforms $R_0 - T(s, q_u, q_n)$ since it is the optimal element in a class containing the subclass whose optimal element is $R_0 - T(s, q_u, q_n)$. This becomes clear by replacing L_n by $L_n + L_u(L_u = 0)$ in (8) and comparing to (9). In addition to improving performance, the global minimization formulation leading to $R_t - T(s, q_u, q_n)$ may also be more fair, in the sense that multiplexed traffic at the network premises is not over-penalized to guarantee $L_u = 0$.

In Fig. 12, the performance of the optimal open- and closed-loop, user-state-dependent relief-spacing TR's inducing zero cell loss at the user premises is presented in terms of L_n versus C_u plots. As expected, the employment of additional (that is, network state) information in the closed-loop scheme, improves the performance achieved by the optimal open-loop scheme, for any value of C_u. It should be noted that the difference in performance increases as C_u increases since a more effective utilization of the network-queue information can be achieved increasing the advantage of the closed-loop scheme. In Figs. 13-15, some relief-spacing functions for $R_t - T(s, q_u, q_n)$ obtained under various background traffic rates λ_B and source burstiness coefficients γ, are presented. These figures illustrate the adaptation of the optimal TR, $R_t - T(s, q_u, q_n)$, to the change in service requirements of both the regulated traffic and the background traffic. From Figs. 13 and 14, it can be seen that as the arrival rate of the background process (λ_B) decreases while the regulated traffic's parameters are held constant, the optimal relief- spacing TR provides lower relief-spacing for smaller values of q_n thus taking advantage of the reduction in service demand by the background process and offering this available capacity to the regulated process. Similar results can be observed with regard to the burstiness of the Markovian source γ in Figs. 14 and 15; for a given system state, the optimal TR under a higher value of γ provides smaller (or equal at most) relief-spacing compared to that under a lower value of γ (Fig. 15); a faster service to the queue is expected to alleviate the increased queueing problems under a higher value of γ. Similar results were also observed with respect to the arrival rate of the regulated traffic.

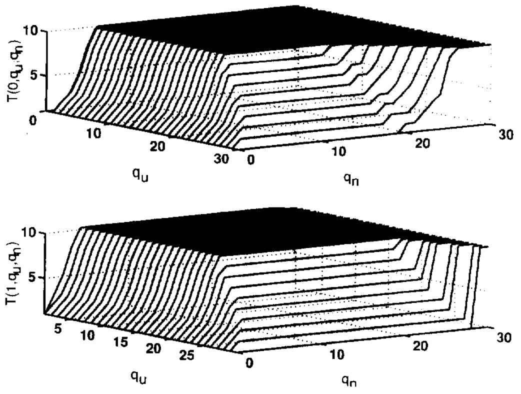


Fig. 13. Relief-spacing function for $R_t - T(s, q_u, q_n)$ $\lambda_B = 0.9$, $\gamma = 0.8$, and $\lambda = 0.09$.

V. Conclusion

In this paper, some new traffic regulation schemes are defined in terms of a relief-spacing (or spacing of the allowance for cell delivery to the network) function. This function may depend only on some user-state (open-loop relief-spacing traffic regulator (TR)) or on both some user and network-state (closed-loop relief-spacing traffic regulator).

Meaningful criteria for optimizing the performance of the proposed classes are defined and a queueing model is presented to analyze their performance. The developed analysis is based on the representation of a relief-spacing traffic regulator in terms of decisions taken upon visiting the states of the

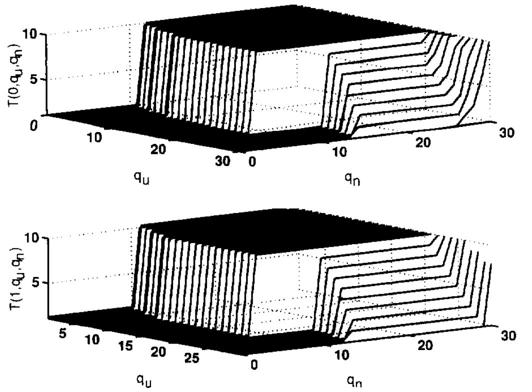


Fig. 14. Relief-spacing function for $R_t - T(s, q_u, q_n)$ $\lambda_B = 0.8$, $\gamma = 0.8$, and $\lambda = 0.09$.

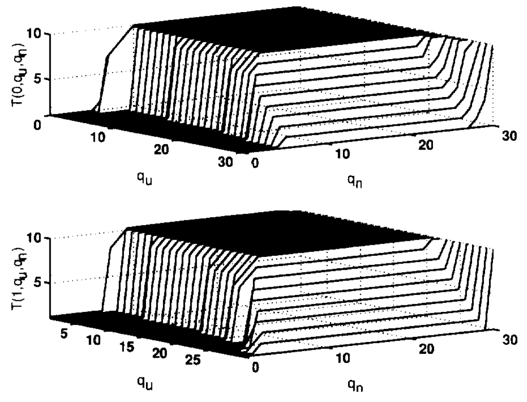


Fig. 15. Relief-spacing function for $R_t - T(s, q_u, q_n)$ $\lambda_B = 0.8$, $\gamma = 0.6$, and $\lambda = 0.09$.

Markovian process describing the evolution of the system. By associating an appropriate cost with each decision, different cost functions, such as cell loss probability, can be defined. Then Markov decision theory approaches can be followed to determine the optimal TR in a given class according to the predetermined criteria. Thus the computationally intractable exhaustive search for the optimal TR is avoided.

When the distance between the user and the network premises is large, purely preventive (open-loop) congestion control should be considered. The optimal open-loop TR derived in this work could be adopted in this case; numerical results have illustrated its improved performance compared to that of the (fixed) Spacer-Controller. When the user and the relevant network premises (that is, the first multiplexing stage within the network) are co-located the optimal closed-loop TR derived in this work could be adopted. Numerical results have shown its improved performance compared to that of the optimal open-loop TR. In addition to its potential applicability for the improved traffic regulation when relevant propagation delays are negligible, the optimal closed-loop traffic regulator can serve as a scheduler for the optimal traffic load distribution in a cascade of two queues with external inputs.

Appendix A

The probabilities $f^m(i,j,k)$ are computed recursively as follows.

- For $m = 1$, clearly:

$$f^1(i,j,k) = \Pr(i \xrightarrow{1} j, a_1 = k) = p(i,j)g(i,k), \\ i\epsilon S, j\epsilon S, k\epsilon[0,1]$$

where:

$$g(i,k) = \Pr\{k \text{ cells are generated from state } i\}.$$

- For $m > 1$:

$$f^m(i,j,0) = \sum_{l=0}^{1} f^{m-1}(i,l,0) f^1(l,j,0).$$

- For $m > 1, 1 \le k \le m$:

$$f^m(i,j,k) = \sum_{l=0}^{1} [f^{m-1}(i,l,k) f^1(l,j,0) \\ + f^{m-1}(i,l,k-1) f^1(l,j,1)].$$

Appendix B

In this Appendix, the cell loss probabilities at the network queue shown in Fig. 4, are computed. The presented analysis is similar to the one presented in Section II–A for the cell loss probability at the user premises and the notation used in that section is adopted. In the following, let $\{I_l\}_{l\ge 0}, \{Q_l^u\}_{l\ge 0}, \{Q_l^n\}_{l\ge 0}$ denote the state of the source, user-queue-occupancy process and the network-queue-occupancy process at the beginning of the lth departure slot, respectively. It may be easily established that $\{I_l, Q_l^u, Q_l^n\}_{l\ge 0}$ is a Markov chain embedded at the beginning of the cell departure slots.

Let $A = [a(i_1, n_1, m_1, i_2, n_2, m_2)]$ denote the transition probability matrix of $\{I_l, Q_l^u, Q_l^n\}_l \ge 0, i_1, i_2 \epsilon S_s$, $n_1, n_2 \epsilon S_u, m_1, m_2 \epsilon S_n$ where $S_n = \{0, 1, .., C_n\}$ denotes the state of the network-queue-occupancy process; let $b^T(k)$ be the probability that k background arrivals occur in T time slots. The following expressions for the transition probabilities $a(i_1, n_1, m_1, i_2, n_2, m_2), i_1, i_2 \epsilon S_s, n_1, n_2 \epsilon S_u, m_1, m_2 \epsilon S_n$, of $\{I_l, Q_l^u, Q_l^n\}_{l\ge 0}$ can be easily obtained.

For $i_1 \epsilon S_s, i_2 \epsilon S_s, T = T(i,n) > m_1 \ge 0$:

$$\begin{aligned} a(i_1, n_1, m_1, i_2, n_2, m_2) = \\ & f^T(i_1, i_2, n_2 - n_1 + 1) b^T(m_2 - m_1 + T - 1), \\ & \{n_1 \ge 2, n_1 - 1 \le n_2 \le C_u - 1\}, \\ & \{n_1 = 1, 2 \le n_2 \le C_u - 1\}, \\ & \{1 \le m_2 \le C_n - 1\} \\ a(i_1, n_1, m_1, i_2, C_u, m_2) = \\ & \textstyle\sum_{k=C_u-n_1+1}^{T} f^T(i_1, i_2, k) \\ & \cdot b^T(m_2 - m_1 + T - 1), n_1 \ge 1, \\ & 1 \le m2 \le C_n - 1. \end{aligned}$$

For $i_1 \epsilon S_s, i_2 \epsilon S_s, m_1 \geq T, T = T(i, n)$:

$$\begin{aligned}
&a(i_1, n_1, m_1, i_2, n_2, m_2) = \\
&\qquad f^T(i_1, i_2, n_2 - n_1 + 1) b^T(m_2 - m_1 + T - 1), \\
&\qquad \{n_1 \geq 2, n_1 - 1 \leq n_2 \leq C_u - 1\}, \\
&\qquad \{n_1 = 1, 2 \leq n_2 \leq C_u - 1\}, \\
&\qquad m_1 - T + 1 \leq m2 \leq C_n - 1, \\
&a(i_1, n_1, m_1, i_2, C_u, m_2) = \\
&\qquad \textstyle\sum_{k=C_u-n_1+1}^{T} f^T(i_1, i_2, k) b^T(m_2 - m_1 + T - 1), \\
&\qquad n_1 \geq 1, m_1 - T + 1 \leq m_2 \leq C_n - 1.
\end{aligned}$$

For $i_1 \epsilon S_s, i_2 \epsilon S_s, C_n \geq m_1 \geq 0, T = T(i_1, n_1)$:

$$\begin{aligned}
&a(i_1, n_1, m_1, i_2, n_2, C_n) = \\
&\quad \textstyle\sum_{k=C_n-m_1+T-1}^{T} f^T(i_1, i_2, n_2 - n_1 + 1) b^T(k), \\
&\quad \{n_1 \geq 2, n_1 - 1 \leq n_2 \leq C_u - 1\}, \\
&\quad \{n_1 = 1, 2 \leq n_2 \leq C_u - 1\} \\
&\quad a(i_1, n_1, m_1, i_2, C_u, C_n) = \\
&\quad \textstyle\sum_{k_1=C_u-n_1+1}^{T} \sum_{k_2=C_n-m_1+T+1}^{T} \\
&\quad \cdot f^T(i_1, i_2, k_1) b^T(k_2), \\
&\quad n1 \geq 1.
\end{aligned}$$

For $i_1 \epsilon S_s, i_2 \epsilon S_s, T(i_1, n_1) - 1 \geq m_1 \geq 0$:

$$\begin{aligned}
&a(i_1, n_1, m_1, i_2, n_2, 0) = \\
&\qquad \textstyle\sum_{k=0}^{T-m_1-1} f^T(i_1, i_2, n_2 - n_1 + 1) b^T(k), \\
&\qquad \{n_1 \geq 2, n_1 - 1 \leq n_2 \leq C_u - 1\}, \\
&\qquad \{n_1 = 1, 2 \leq n_2 \leq C_u - 1\}, \\
&\qquad T = T(i_1, n_1) \\
&a(i_1, n_1, m_1, i_2, C_u, 0) = \\
&\qquad \textstyle\sum_{k_1=C_u-n_1+1}^{T} \sum_{k_2=0}^{T-m_1-1} \\
&\qquad \cdot f^T(i_1, i_2, k_1) b^T(k_2), \\
&\qquad n_1 \geq 1, \\
&\qquad T = T(i_1, n_1).
\end{aligned}$$

For $i_1 \epsilon S_s, i_2 \epsilon S_s, C_n \geq m_1 \geq 0$:

$$\begin{aligned}
&a(i_1, 1, m_1, i_2, 1, m_2) = \\
&\qquad \cdot f^T(i_1, i_2, 1) b^T(m_2 - m_1 + T - 1) \\
&\qquad + f^T(i_1, 1, 0) p(1, i_2) b^{T+1}(m_2 - m_1 + T) \\
&\qquad + \textstyle\sum_{k=2}^{\infty} f^T(i_1, 0, 0) p^{k-2}(0, 0) p(0, 1) p(1, i_2) \\
&\qquad b^{T+k}(m_2 - m_1 + T + k - 1), \\
&\qquad 0 \leq m_2 \leq C_n, \\
&\qquad T = T(i_1, 1) \\
&a(i_1, 1, m_1, i_2, 1, C_n) = \\
&\qquad \textstyle\sum_{k_b=C_n-m_1+T-1}^{T} f^T(i_1, i_2, 1) b^T(k_b) \\
&\qquad + \textstyle\sum_{k_b=C_n-m_1+T}^{T+1} f^T(i_1, 1, 0) p(1, i_2) b^{T+1}(k_b) \\
&\qquad + \textstyle\sum_{k=2}^{\infty} \sum_{k_b=C_n-m_1+T-1+k}^{T+k} f^T(i_1, 0, 0) \\
&\qquad p^{k-2}(0, 0) p(0, 1) p(1, i_2) b^{T(i_1,1)+k}(k_b), \\
&\qquad T = T(i_1, 1) \\
&a(i_1, 1, m_1, i_2, 1, 0) = \\
&\qquad \textstyle\sum_{k_b=0}^{T-m_1-1} f^T(i_1, i_2, 1) b^T(k_b) \\
&\qquad + \textstyle\sum_{k_b=0}^{T-m_1} f^T(i_1, 1, 0) p(1, i_2) b^{T+1}(k_b) \\
&\qquad + \textstyle\sum_{k=2}^{\infty} \sum_{k_b=0}^{T+k-m_1-1} f^T(i_1, 0, 0) p^{k-2}(0, 0) p(0, 1) \\
&\qquad p(1, i_2) b^{T+k}(k_b), \\
&\qquad T = T(i_1, 1).
\end{aligned}$$

The steady-state probabilities $\pi(i, n, m), i \epsilon S_s, n \epsilon S_u, m \epsilon S_n$ are obtained from the matrix equation $\overline{\Pi} = A\overline{\Pi}$, $\overline{\Pi} = [\pi(i, n, m)]$ with the normalizing condition $\overline{\Pi}'e = 1$ where e is a unity vector. The cell loss probability L_n at the network premises can then be obtained from the expression:

$$L_n = \frac{\sum_{i \epsilon S_s} \sum_{n \epsilon S_u} \sum_{m \epsilon S_n} L_n(i, n, m) \pi(i, n, m)}{\lambda' \sum_{i \epsilon S_s} \sum_{n \epsilon S_u} \sum_{m \epsilon S_n} \overline{ID}(i, n) \pi(i, n)} \quad (12)$$

where λ' denotes the effective total arrival rate at the network queue and where $L_n(i, n)$ denotes the expected number of cells lost at the network-queue over the interval $ID(i, n)$ and is given by

$$\begin{aligned}
&L_n(i, n, m) = \\
&\quad \textstyle\sum_{k=C_n-m+T-1}^{T} (k - C_n + m + T - 1) b^T(k), \\
&\quad n > 1, T = T(i, n) \\
&L_n(i, 1, m) = \\
&\quad \textstyle\sum_{i2} \sum_{k_u=1}^{T} \sum_{k_b=C_n-m+T-1}^{T} (k_b - C_n + m + T - 1) \\
&\quad f^T(i, j, k_u) b^T(k_b) \\
&\quad + \textstyle\sum_j \sum_{k_b=C_n-m+T}^{T+1} (k_b - C_n + m - T) f^T(i, 1, 0) \\
&\quad p(1, j) b^{T+1}(k_b) \\
&\quad + \textstyle\sum_j \sum_{k=2}^{\infty} \sum_{k_b=C_n-m+T+k-1}^{T+k} \\
&\quad \cdot (k_b - C_n + m - T) f^T(i, 0, 0) p^{k-2}(0, 0) \\
&\quad \cdot p(0, 1) p(1, j) b^{T+1}(k_b), \\
&\quad T = T(i, 1).
\end{aligned} \qquad (13)$$

APPENDIX C

In this Appendix, the linear programming approach is applied for the derivation of the optimal element in the class $R_v - T(s, q_u)$ defined in (7). First let $A(i_1, n_1) \equiv \{T_{\min}(i_1, n_1), ..., T_{\max}(i_1, n_1)\}$, $T_{\min}(i_1, n_1)$ $(T_{\max}(i_1, n_1))$ is the min (max) allowable relief-spacing from state (i_1, n_1). let $\pi(i_1, n_1, T)$ be the steady-state probability of being in state (i_1, n_1) and taking the action $T \epsilon A(i_1, n_1)$. Let $u(i_1, n_1, T) = \frac{\pi(i_1, n_1, T)}{\tau(i_1, n_1, T)}$, where $\tau(i_1, n_1, T) = \overline{ID}(i_1, n_1)$ is the expected time till the next decision epoch if action (relief-spacing) T is taken from state (i_1, n_1). Based on the above, the linear programming formulation of the optimization problem is as follows:

$$\text{minimize} \sum_{i_1 \epsilon S_s} \sum_{n_1 \epsilon S_u} \sum_{T \epsilon A(i,n)} c(i_1, n_1, T) u(i_1, n_1, T).$$

Subject to:

$$u(i_1, n_1, T) \geq 0$$

$$\begin{aligned}
&\textstyle\sum_{T \epsilon A(i,n)} u(i_2, n_2, T) - \sum_{i_1 \epsilon S_s} \sum_{n_1 \epsilon S_u} \sum_{T \epsilon A(i,n)} \\
&\quad \cdot a(i_1, n_1, i_2, n_2, T) u(i_1, n_1, T) = 0, i_2 \epsilon S, n_2 \epsilon S_u
\end{aligned}$$

$$\textstyle\sum_{i_1 \epsilon S_s} \sum_{n_1 \epsilon S_u} \sum_{T \epsilon A(i_1,n_1)} \tau(i_1, n_1, T) u(i_1, n_1, T) = 1$$

$$\begin{aligned}
&\textstyle\sum_{i_1 \epsilon S_s} \sum_{n_1 \epsilon SQ} \sum_{T \epsilon A(i_1,n_1)} [\tau(i_1, n_1, T) L_u(i_1, n_1, T) \\
&\quad - w\lambda\tau^2(i_1, n_1, T)] u(i_1, n_1, T) \leq 0.
\end{aligned}$$

If $u^*(i_1, n_1, T)$ is an optimal basic solution for the above linear program, then for each state (i_1, n_1) there will be at most one action T such that $u^*(i_1, n_1, T) > 0$, implying that $T^*(i, n) = T$. If for some state (i_1, n_1), $u^*(i_1, n_1, T) = 0$ for all T, $T^*(i_1, n_1)$ is chosen arbitrarily as some T such that $a(i_1, n_1, i_2, n_2, T) > 0$ and $u^*(i_2, n_2, T_2) > 0$.

In the above linear program, $\tau(i_1, n_1, T)u(i_1, n_1, T)$,represents the steady-state probability of being in state (i_1, n_1) and taking the action T. Note that only stationary policies are considered, i.e, policies where the same action $T(i_1, n_1)$ is taken at every visit to the state (i_1, n_1). Therefore, $\tau(i_1, n_1, T)u(i_1, n_1, T)$ represents the steady-state probability $\pi(i_1, n_1)$.

The first set of constraints represent the balance equations requiring that for any state $(i_2, n_2)\epsilon S_s \times S_u$ the long-run average number of transitions from state (i_2, n_2) per unit time be equal to the long-run average number of transitions into state (i_2, n_2) per unit time. The second constraint requires that the sum of the steady-state probabilities be equal to 1 and finally the last constraint represents the cell loss QoS; as in (5), that must be met at the user premises and which is obtained through a simple manipulation of the expression for the cell loss probability (3) presented in Section II–A.

IX. Appendix D

In the following, the value iteration algorithm used in the derivation of the optimal closed-loop, relief-spacing TR's (8) and (11), is outlined. This algorithm computes recursively a sequence of value functions $V_l(i_1, n_1, m_1)$ associated with each state $(i_1, n_1, m_1)\epsilon S_s \times S_u \times S_n$, approximating the minimal average cost per unit time (slot), (starting with an arbitrarily chosen function $V_o(i_1, n_1, m_1), (i_1, n_1, m_1)\epsilon S_s \times S_u \times S_n)$, as follows:

$$\begin{aligned} V_l(i_1, n_1, m_1) = {} & \min\nolimits_{T\epsilon A(i_1,n_1,m_1)} \\ & \{\tfrac{1}{\tau(i_1,n_1,m_1,T)} \times [c(i_1, n_1, m_1, T) \\ & +\tau \textstyle\sum_{i_1} \sum_{n_1} \sum_{m_1} a(i_1, n_1, m_1, i_2, n_2, m_2, T) \\ & \cdot V_{l-1}(i_2, n_2, m_2)] + (1 - \tfrac{\tau}{\tau(i_1,n_1,m_1,T)}) \\ & \cdot V_{l-1}(i_1, n_1, m_1)\} \end{aligned}$$

where $a(i_1, n_1, m_1, i_2, n_2, m_2, T)$ denote the transition probabilities of $\{I_l, Q_l^u, Q_l^n\}_{l\geq 0}$, from state (i_1, n_1, m_1) at the beginning of a departure slot (decision epoch) to state (i_2, n_2, m_2) at the beginning of the following departure slot, if the relief-spacing function is such that $T(i_1, n_1, m_1) = T$; these probabilities are equal to those found in appendix B when $T(i_1, n_1, m_1) = T$. $\tau(i_1, n_1, m_1, T)$ is the expected time till the next decision epoch if action T is taken in the state (i_1, n_1, m_1) and is equal to $\overline{ID}(i_1, n_1)$ given by (2) in Section II–A, when $T(i_1, n_1) = T$. $A(i_1, n_1, m_1)$ is the set of possible actions from state (i_1, n_1, m_1); defined in Appendix C. τ is a number chosen for aperiodicity purposes such that $0 \leq \tau \leq \min_{i_1,n_1,m_1,T\epsilon S_s\times S_u\times S_n\times S_B} \tau(i_1, n_1, m_1, T)$. $c(i_1, n_1, m_1, T)$ represents the cost associated with state (i_1, n_1, m_1) when the relief-spacing function is such that $T(i, n, m) = T$. For the class $\langle R_w - T(s, q_u, q_n)\rangle$, $w = 0$, $c(i_1, n_1, m_1, T)$ is set equal to the expected number of cells lost at the network buffer over a decision interval, if $T(i_1, n_1, m_1) = T$, and is obtained from the expression (13) for $L_n(i_1, n_1, m_1)$ with $T(i_1, n_1) = T$ (appendix B). For the class $\langle R_t - T(s, q_u, q_n)\rangle$, $c(i_1, n_1, m_1, T)$ is set equal to the sum of the expected losses over a decision interval at both the user premises (4) and the network premises (11).

Notice that $V_l(i_1, n_1, m_1)$ represents a long-term minimal cost rate per unit time (slot) induced by observing the system following a visit to state (i_1, n_1, m_1). For large l, the difference $V_l(i_1, n_1, m_1) - V_{l-1}(i_1, n_1, m_1)$ approaches the average cost rate and the policy T minimizing the above value function induces a cost arbitrarily close to that of the optimal policy [15].

References

[1] I. Stavrakakis, M. Abdelaziz, and D. Hoag, "A user relief approach to congestion control in ATM networks," in *Asynchronous Transfer Mode Networks*. New York: Plenum, 1993, pp. 135-155.

[2] J. Bae and T. Suda, "Survey of traffic control schemes and protocols in ATM networks," *Proc. IEEE*, vol. 79, pp. 170–189, Feb. 1991.

[3] E. Rathgeb, "Modeling and performance comparisons of policing mechanisms for ATM networks," *IEEE J. Select. Areas Commun.*, vol. 9, pp. 325–224, Apr. 1991.

[4] Special Issue on Congestion Control in High Speed Networks, *IEEE Commun. Mag.*, Oct. 1991.

[5] W. Leland, "Window based congestion management in broadband ATM networks: The performance of three access-control policies," in *IEEE GLOBECOM'89 Conf.*, pp. 1794–1800.

[6] G. Gallasi, G. Rigolio, and L. Fratta, "ATM: Bandwidth assignment and bandwidth enforcement policies," in *IEEE INFOCOM'91*, pp. 1788-1793.

[7] K. Bala, I. Cidon, and K. Sohraby, "Congestion control for high speed packet switching networks," in *ICC'90*, 1990, pp. 520–526.

[8] P. Boyer, "A Congestion Control for ATM," in *Int. Teletraf. Cong. Sem.*, NJ, Oct. 1990, p. 4.3.

[9] B. Lague, C. Rosenberg, and F. Guillemin, "A generalization of some policing mechanisms," in *IEEE INFOCOM'92*, pp. 767-775.

[10] F. Guillemin, P. Boyer, A. Dupuis, and L. Romoeuf, "Peak rate enforcement in ATM networks," in *IEEE INFOCOM'92*, pp. 753-758.

[11] G. Rigolio and L. Fratta, "Input rate regulation and bandwidth assignment in ATM networks: An Integrated approach," vol. ITC-13, pp. 141-146, 1991.

[12] F. Bernabei, L. Gratta, M. Listanti, and M. Testa, "Analysis of Two Level Shaping for Multiplexing of ON-OFF ATM Sources," in *ICC'93 Conf.*, pp. 1380-1385.

[13] ——, "Analysis of ON-OFF source shaping for ATM multiplexing," in *IEEE INFOCOM'93*, pp. 1330-1336.

[14] C. Ohta, H. Tode, M. Yamamoto, H. Okada, and Y. Tezuke, "Peak rate regulation scheme for ATM networks and its performance," in *IEEE INFOCOM'93*, pp. 680-689.

[15] H. C. Tijms, *Stochastic Modeling and Analysis: A Computational Approach*. New York: Wiley, 1986.

[16] H. W. Lee and J. W. Mark, "ATM network traffic characterization using two types of on-off sources," IEEE *INFOCOM'93*, pp. 152-159.

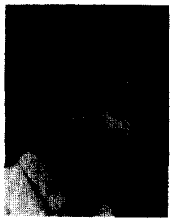
[17] F. M. Brochin, "A cell spacing device for congestion control in ATM networks," *Perform. Eval.*, pp. 107-127, 1992.

[18] I. Rubin and K. D. Lin, "A burst level adaptive input-rate flow control scheme for ATM networks," in *IEEE INFOCOM'93*, pp. 386-394.

Mohamed Abdelaziz (S'94) received the B.S. and M.S. degrees in electrical engineering from Cairo University, Egypt in 1988 and 1991, respectively. He is currently a candidate for the Ph.D. degree and a graduate research assistant in the Department of Electrical Engineering and Computer Science at the University of Vermont.

His research interests include performance analysis of communication networks, discrete-time queueing theory and resource allocation and scheduling in high-speed networks.

Mr. Abdelaziz has been a member of the ACM since 1992.



Ioannis Stavrakakis (S'85-M'89-SM'93) received the Diploma in electrical engineering from the Aristotelian University of Thessaloniki, Thessaloniki, Greece, 1983, and the Ph.D. degree in electrical engineering from the University of Virginia in 1988.

Since 1988, he has been with the Department of Electrical Engineering and Computer Science, University of Vermont, where he is currently an Associate Professor. His research interests are in stochastic system modeling, teletraffic analysis and discrete-time queueing theory, with primary focus on the design and performance evaluation of Broadband Integrated Services Digital Networks (B-ISDN).

Dr. Stavrakakis is a member of the IEEE Communications Society, Technical Committee on Computer Communications. He has organized and chaired sessions, and has been a technical committee member, for conferences such as GLOBECOM, ICC, and INFOCOM.