Journal of Mathematical Sciences, Vol. 92, No. 3, 1998

STABILITY AND PERFORMANCE ANALYSIS OF RATE-BASED FEEDBACK FLOW CONTROLLED ATM NETWORKS

V. Sharma (Bangalore, India) and J. Kuri (Montreal, Canada)

UDC 519.2

Motivated by the ABR class of service in ATM networks, we study a continuous-time queueing system with a feedback control of the arrival rate of some of the sources. The feedback regarding the queue length or the total workload is provided at regular intervals (variations on it, especially the EPRCA algorithm, are also considered). The propagation delays can be nonnegligible. For a general class of feedback algorithms, we obtain the stability of the system in the presence of one or more bottleneck nodes in the virtual circuit. We also obtain rates of convergence to the stationary distributions and finiteness of moments. For the single bottleneck case, we provide algorithms to compute the stationary distributions and the moments of the sojourn times in different sets of states. We also show analytically (by showing the continuity of stationary distributions and moments) that for small propagation delays, we can provide feedback algorithms which have higher mean throughput, lower probability of overflow, and lower delay jitter than any open-loop policy.

1. Introduction

Recently, at an ATM forum, a new service category, available bit rate (ABR), was announced. Earlier, ATM networks were standardized for constant-bit-rate (CBR) and variable-bit-rate (VBR) services, where the network provides an explicit guarantee of service. These services require an a priori declaration of certain traffic descriptors. But for data traffic it may be difficult to announce these descriptors at the time of connection. Therefore, for such traffic ABR service has been developed, where the network provides the source the best-effort service. However, to provide a certain upper bound on the probability of packet loss, the network may require such a source to flow-control its traffic. The two main proposals discussed at the forum for flow-control for this service were rate-based and credit-based flow control [6]. The credit-based scheme, which is similar to window flow control, has been analyzed in [24]. However, finally the rate-based mechanism has been approved [6]. This has resulted in a flurry of research activity on this problem. Some earlier studies on this problem, motivated by the analysis of congestion control in datagram networks, may be found in [10, 17, 18, 25]. More recent studies, motivated by high-speed networks, include [2, 3, 7, 13, 14, 16, 19, 21, 28]. Siu and Tzeng [21] provide a general survey of the different proposals discussed at ATM forums for the flow control of ABR traffic. Most of these studies use a deterministic or a fluid model of the network. Kawahara et al. [16], Altman et al. [3], and Wang and Sengupta [27] model the system as a discrete-time queue. Analysis of such systems in the standard continuous queueing framework, although of obvious practical interest, is still not available. This paper provides such a study. Our methods and results differ from the previous studies.

Our main model is shown in Fig. 1. A single queue with infinite buffers is provided with two input streams. This queue may represent a bottleneck node on the virtual circuit of an ABR source. One of the input streams, the uncontrolled stream (called stream 2), is a Poisson arrival process with rate λ and represents all the other traffic of the network passing through that node. The ABR source (stream 1) is represented by an infinite data source, and the packets from this source are extracted as a Poisson process (we also consider packets extracted at regular intervals) with instantaneous intensity at time t as $\lambda(t)$. The intensity $\lambda(t)$ is controlled based on the feedback information available at regular intervals of length T (this is not really required by our analysis; these intervals could be time varying but upper bounded). Below, we will consider one more variation. We denote the queue length at time t by $q(t)$ and the total workload by $V(t)$. For exponential service times the feedback information is $q(nT) = q_n$, while for general service times it is $V(nT) = V_n$. In high-speed networks, there can be considerable propagation delay of traffic in both directions. With a delay of time D_b the feedback information reaches the ABR source. At that time, based on q_n and the previous rate, the controller changes the rate of the ABR source. In between feedback instants, the rate $\lambda(t)$ remains the same. The propagation delay of packets from the source to the bottleneck node is denoted by D_f. We assume D_f, D_b, and T to be constants. The service times of the two traffic streams will be independent identically distributed (i.i.d.) with

Proceedings of the Seminar on Stability Problems for Stochastic Models, Hajdúszoboszló, Hungary, 1997, Part I.

 1072-3374/98/9203-3930$20.00

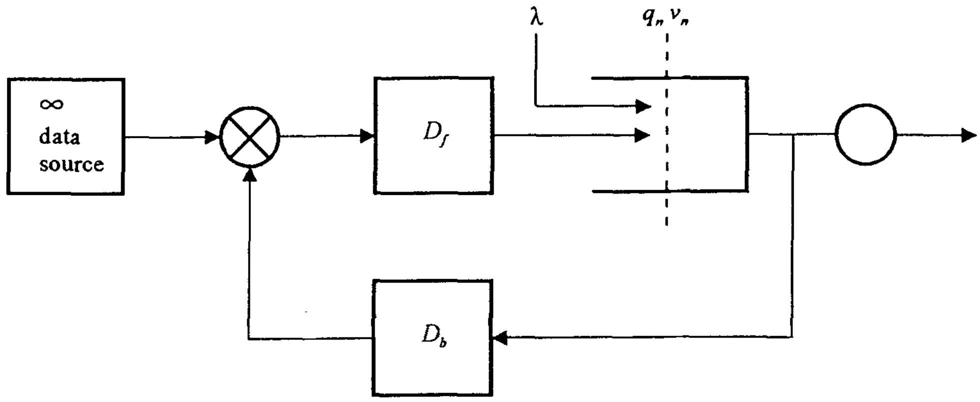

Fig. 1

general service time distributions $s(1)$ and $s(2)$, respectively. Our model is flexible enough to include the 1-bit feedback as well as the explicit-rate feedback schemes proposed more recently.

We obtain the stability of this system under a natural rate condition. We also obtain algorithms and analytical expressions of the stationary distributions of the queue lengths, waiting times, and moments of various passage times.

To obtain a good feedback control, one needs to consider some appropriate performance indices. Commonly used criteria for high-speed networks are maximizing mean throughput, minimizing delay jitter and mean delay, and (in the case of finite buffers) minimizing stationary probability of overflow. In order to optimize these contradictory criteria, recently, minimizing a new criterion, the distance of stationary queue length from a fixed constant N_0, $B > N_0 > 0$ (B is the buffer length), has been attempted (see, e.g., [27]). Intuitively one sees that choosing N_0 appropriately will provide a good balance between the three criteria. We show analytically that for our model, optimizing this criterion (distance from N_0) will indeed provide better system performance (with respect to all three criteria mentioned above) as compared to any open-loop policy (i.e., fixed arrival rate for the ABR source) if D_b, D_f, and T are sufficiently small. In order to show this, we prove the continuity of stationary distributions and moments of $\{q(t)\}$ and $\{V(t)\}$, which is of independent interest. These results will be further corroborated by the computational result we provide at the end.

We will also consider a variation of this system which will make our model close to the Enhanced Proportional Rate Control Algorithm (EPRCA) [6]. For this system, instead of sending the packets as a Poisson process with rate λ_n, the ABR source sends packets at regular intervals of time $1/\lambda_n$. Also, instead of sending feedback at intervals of length T, the feedback is sent after every N (a constant) packets of the ABR source have been sent. In addition, λ_n can be a function of $\lambda_{n-1}, \ldots, \lambda_{n-m}, V_n, V_{n-1}, \ldots, V_{n-m+1}$.

The basic model described above will be generalized in various ways. We prove the stability if the stream 2 is a Markov-modulated Poisson arrival stream. The infinity data source can be replaced by a local queue at the ABR source where the packets joining the local queue are being generated by another Poisson stream. There can be more than one ABR controlled source feeding the bottleneck node. The buffer length at the queue can be finite. Instead of one bottleneck queue, there can be several (bottleneck or otherwise) queues on the virtual circuit of the ABR source (see Fig. 2).

To save space, we have omitted the main proofs in the paper. They will appear elsewhere. We have also obtained extensive numerical and simulation results on our basic model.

The paper is organized as follows. In Sec. 2, all the stability results are provided. Section 3 provides the algorithms for stationary distributions, comparison of open-loop and closed-loop policies, a continuity result, and several results on passage times.

2. Stability

Let us first assume that the propagation delays $D_f = D_b = 0$. Then at time nT, the workload V_n (for exponential service, it could be q_n) becomes available to the rate controller. Now the controller updates its rate to λ_n:

$$\lambda_n = f(V_n, \lambda_{n-1}), \tag{1}$$

where f is some measurable, deterministic function (generalization to a random function will be considered later on). In the basic model of Fig. 1, because of infinite data source assumptions the ABR source will input data to the queue

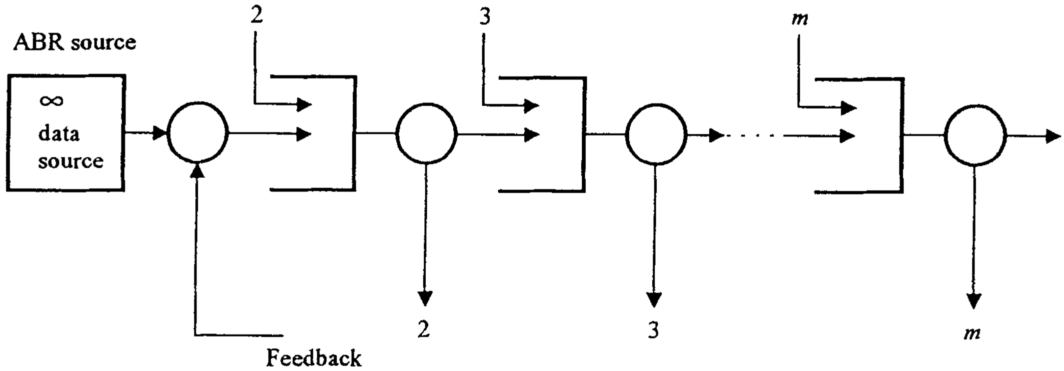

Fig. 2

during time $[nT, (n+1)T]$ as a Poisson process with rate λ_n. Now because of i.i.d. service times, $\{(V_n, \lambda_{n-1})\}$ is a Markov chain.

If the propagation delays D_f and/or D_b are not zero, then the queue length (workload) information will reach the controller D_b times later and the traffic arriving at the queue will arrive D_f time after being sent from the controller. Therefore, (1) will have to be appropriately modified. But the stability argument provided in the next theorem will continue to hold.

For the stability of the system $\{(V_n, \lambda_{n-1})\}$, we further make

ASSUMPTION A. (i) *There are* N_1 *and* $\overline{\lambda}_1$ *such that if* $V_n = 0, V_{n+1} = 0, \ldots, V_{n+N_1} = 0$, *then* $\lambda_{n+N_1} = \overline{\lambda}_1$ *for any* λ_{n-1};
(ii) *there is a* $\overline{\lambda} \geq 0$ *such that* $\lambda_n \leq \overline{\lambda}$ *for all* n;
(iii) *there are* $\overline{N} \geq 0$ *and* $N_2 \geq 0$ *such that if* $V_n > \overline{N}, V_{n+1} > \overline{N}, \ldots, V_{n+m} > \overline{N}$ *for* $m \geq N_2$, *then* $\lambda_{n+m} \leq \overline{\lambda}_2$, *where* $\overline{\lambda}_2$ *is a constant with*

$$\overline{\lambda}_2 \mathbf{E}[s(1)] + \lambda \mathbf{E}[s(2)] < 1. \tag{2}$$

Assumption A is quite general and most of the algorithms considered in the literature for (1) satisfy it. Furthermore, assumptions (ii) and (iii) are obviously necessary for the stability to hold. We will consider one specific case in the next section.

THEOREM 1. *Let the system in Fig. 1 satisfy Assumption A. Then* $\{(V_n, \lambda_{n-1})\}$ *has a unique stationary distribution and starting from any initial conditions, the system will converge in total variation to the stationary distribution.*

If we assume that $\mathbf{E}[s(i)^m] < \infty$, $i = 1, 2$, for some $m > 1$, then we obtain (e.g., see [15]) $n^{m-1}\|(V_n, \lambda_{n-1}) - \pi\| \to 0$ as $n \to \infty$, where $\|\cdot\|$ denotes the total variation norm. Also, by the regenerative theory ($\hat{\tau}$ denotes the regeneration length),

$$\mathbf{E}_\pi[V^m] = \frac{1}{\mathbf{E}\hat{\tau}} \mathbf{E}\left[\sum_{k=1}^{\hat{\tau}} V_k^m\right] \leq \frac{1}{\mathbf{E}\hat{\tau}} \mathbf{E}[\hat{\tau}(\hat{\tau}T)^m]$$

and hence $\mathbf{E}_\pi[V^m] < \infty$ for $m \geq 1$ if $\mathbf{E}[\hat{\tau}^{m+1}] < \infty$. We can further obtain the central limit theorem for $\{V_n\}$ (see [23]) by ensuring that

$$\mathbf{E}_{(0,\overline{\lambda}_1)}\left[\left(\sum_{k=1}^{\hat{\tau}} V_k\right)^2\right] \leq T^2 \mathbf{E}_{(0,\overline{\lambda}_1)}[(\hat{\tau})^4] < \infty.$$

If in (1) f is not deterministic but random, i.e., the distribution of λ_n depends upon (V_n, λ_{n-1}) (independent of anything else), then Theorem 1 still holds even if Assumption A(i) is generalized to have λ_{n+N_1} having a specific distribution.

If the queue in Fig. 1 has a finite buffer, then Theorem 1 holds without the rate condition in Assumption A(iii).

As remarked earlier, if the service times are exponential, instead of total workload information V_n, the queue length information would be sufficient. For the case of general service times, the above proof would hold for the process $\{(q_n, \lambda_{n-1})\}$ under some extra conditions (to take care of residual service times).

If the propagation delays D_f and D_b are nonzero, then by keeping track of the arrival rates to the queue and the workload information available (see next section for an example) during any interval $[nT, (n+1)T]$, the proof of Theorem 1 will continue to hold (just replace N_2T by $N_2T + D_b + D_f$ in (2)).

Now we consider a variation on the modeling assumptions of the system in Fig. 1 which will make this system closer to the EPRCA algorithm proposal for the ABR service (see, e.g., [6]). For simplicity, we will call this class of systems EPRCA algorithms. We assume that when the ABR source is assigned the rate λ_n, instead of sending its packets (cells) as a Poisson process with rate λ_n, it sends packets at regular intervals of time $1/\lambda_n$. Also, after it has sent a fixed number (say N) of cells, it receives feedback on the total workload or the queue length. Thus, now the feedback is not received at regular intervals of time T but rather, the interval between nth and $(n+1)$st feedback is N/λ_n. In addition to the assumptions of Theorem 1, now we further assume that there is an upper bound $\overline{T}$ on the feedback interval, i.e., if λ_n becomes very small, then, instead of sending the next feedback after time N/λ_n, it will be sent after time $\overline{T}$ and $\mathbf{P}\{\overline{\lambda}s(1) < 1\} > 0$. If we denote the workload at time of nth feedback by V_n and the rate of the ABR source immediately after it by λ_n, then the stability of the Markov chain (V_n, λ_n) can be obtained as in Theorem 1. In the next section, we will also indicate how to obtain the transition probability matrix and the stationary distributions for this system.

We consider one more variation which is needed for the EPRCA algorithm. Instead of (1), consider $\lambda_n = g(V_n, V_{n-1}, \lambda_{n-1}, \lambda_{n-2})$, where g satisfies Assumption A. Then proceeding as in Theorem 1, we can show that the mean intervisit time to the state sequence $(V_n, V_{n-1}, \lambda_{n-1}, \lambda_{n-2}) = (0, 0, \overline{\lambda}_1, \overline{\lambda}_1)$ is finite. Since such epochs make the process $\{(V_n, \lambda_{n-1})\}$ regenerate, the conclusions of Theorem 1 hold. The generalization to the case $\lambda_n = g(V_n, V_{n-1}, \dots V_{n-m+1}, \lambda_{n-1}, \lambda_{n-2}, \dots, \lambda_{n-m})$ for some finite m is obvious.

Now we consider the case where the uncontrolled stream (stream 2) is a Markov-modulated Poisson process modulated by a finite-state, irreducible Markov chain $\{X(t)\}$ (i.e., at time t if $X(t) = i$, then the instantaneous arrival rate will be $\lambda_2(i)$). Let its intensity at time t be $\lambda_2(t)$. Then $X_k \equiv X(kT)$ is also a finite-state, irreducible, aperiodic chain with a stationary distribution π. Let Assumption A be satisfied with λ replaced by $\mathbf{E}_\pi[\lambda_2(t)]$. Start Markov chain $\{(V_n, \lambda_{n-1}, X_n)\}$ in state $(0, \overline{\lambda}_1, i)$ at time 0 with i being a fixed state of $\{X_n\}$. Then, as in Theorem 1, we can show that $(V_n, \lambda_{n-1}) = (0, \overline{\lambda}_1)$ at a finite mean time (say k) (in (2), λ will now be replaced by $\max_i \lambda_2(i)$). Since $\{X_k\}$ is finite and irreducible, there is a $0 \le k_1 < \infty$, such that from any state j, X_k can reach i in k_1 steps with a positive probability. Therefore, when $(V_n, \lambda_{n-1}) = (0, \overline{\lambda}_1)$, within k_1 steps $(V_n, \lambda_{n-1}, X_n) = (0, \overline{\lambda}_1, i)$ with a positive probability. This shows that, by starting the chain $\{(V_n, \lambda_{n-1}, X_n)\}$ in state $(0, \overline{\lambda}_1, i)$, in a finite mean time it will revisit the state. Also, the intervisit time has an aperiodic distribution. Therefore, Theorem 1 holds for this case. The nonzero propagation delays can be taken care of as above.

Next we consider the case of more than one (say m) ABR controlled users sharing the queue. Let the controlled rates be $\lambda_n = (\lambda_n^1, \lambda_n^2, \dots, \lambda_n^m)$ and let

$$\lambda_n^k = f^k(V_n, \lambda_{n-1}), \quad k = 1, \dots, m. \tag{3}$$

The service-time distributions of different users can be different. We assume that Assumption A is satisfied with the following modifications (in the terminology of Assumption A): there is an N_1 (common for all ABR sources) such that if $V_n = 0, \dots, V_{n+N_1} = 0$, then $\lambda_{n+N_1}^k = \overline{\lambda}_1^k$, $k = 1, \dots, m$. There is a $\overline{\lambda}$ upper bounding all the rates. There is an $\overline{N} \ge 0$ and N_2 such that if $V_n > \overline{N}, \dots, V_{n+m} > \overline{N}$ for $m \ge N_2$, then $\lambda_{n+N_1}^k \le \overline{\lambda}_2^k$, $k = 1, \dots, m$, where

$$\sum_{k=1}^{m} \overline{\lambda}_2^k \mathbf{E}[s^k(1)] + \lambda[s(2)] < 1, \tag{4}$$

$s^k(1)$ being a generic service time of the kth ABR source. Then Theorem 1 continues to hold. The propagation delays can be included as before.

Next we consider the system in Fig. 2. We assume that there is one ABR source and the interfering traffic entering at each queue leaves the system after service at that queue. Let $\lambda(i)$, $i = 1, 2$, be the arrival rate of the interfering source $i+1$ and let $s(i)$ be its generic service time. Also, let $V_n(i)$ be the workload at time nT at queue i and let $V_n = \max_i(V_n(i))$. We use this V_n in (1) for the update of the arrival rates. Also, let $\overline{\lambda}$ upper bound all rates and let $\overline{\lambda}_2$ in Assumption A(iii) be such that the total intensity at each queue is less than one. This is in conformance with the proposals discussed at the ATM forum [21].

THEOREM 2. *For the system in Fig. 2, let Assumption* A *(with the above-mentioned modifications) hold for each queue. Also, let* $\mathbf{E}[s(1, i)] > 0$ *for* $i = 1, \dots, m-1$. *Then the Markov chain* $\{(V_n(1), \dots, V_n(m), \lambda_{n-1})\}$ *has a unique stationary distribution and starting in any initial conditions, the chain converges to the stationary distribution in total variation.*

As before, we can include the propagation delays in the above analysis to obtain the stability result under the same conditions.

3. Performance Analysis

Once we have the stability of the system, it is important to obtain some quantitative and/or qualitative estimates of the various performance indices: transient and stationary moments and distributions of the queue lengths, waiting times, probability of overflow (for the finite buffer case), and waiting-time jitter for the various classes of traffic. In this section, we obtain some algorithms to compute the above quantities and provide comparison and continuity results for the system in Fig. 1. For example, we will identify an optimal closed-loop policy in a class of policies when the propagation delays and T are zero. We will show that this policy is better than an optimal open-loop policy when the propagation delays and T are small. We will also indicate a method for computing an optimal feedback policy when $T > 0$ while $D_b = D_f = 0$. In the next section, we will consider some specific examples and provide numerical results.

For the system in Fig. 1, the transition probability function of the Markov chain $\{(V_n, \lambda_{n-1})\}$ is given by (when $D_f = D_b = 0$)

$$\mathbf{P}\{V_{n+1} \varepsilon A, \lambda_n = f(v, \lambda') \mid V_n = v, \lambda_{n-1} = \lambda'\}$$

$$= \begin{cases} \text{The probability that the workload at time } T \text{ in an } M \mid GI \mid 1 \text{ queue is in set } A \text{ given} \\ \text{that at } t = 0\text{, the workload is } v \text{ and the arrival process is Poisson with rate } f(v,\lambda') + \lambda \\ \text{and service times a mixture of } s(1) \text{ and } s(2) \text{ (in the ratio of } f(v,\lambda')/(\lambda + f(v,\lambda'))\text{;} \\ 0, \quad \text{otherwise.} \end{cases} \tag{5}$$

If the service times are exponential, then, as mentioned in the last section, V_n can be replaced by the queue length q_n at time nT. The Markov chain $\{(q_n, \lambda_{n-1})\}$ has a countable state space (if λ_n takes values in a countable set). Then the transition matrix for $\{q_n, \lambda_{n-1})\}$ involves knowing the transient probabilities of an $M \mid M \mid 1$ queue. Unlike the $M \mid GI \mid 1$ case considered above, the transient probabilities of an $M \mid M \mid 1$ can be more easily calculated. For a specific f (to be described below), we have calculated these probabilities (using efficient formulas in [11, 22]) and then obtained the stationary distribution of the Markov chain $\{(q_n, \lambda_{n-1})\}$.

Once we have the stationary probability π of $\{(q_n, \lambda_{n-1})\}$, $\mathbf{E}_\pi[\lambda_{n-1}]$ provides the mean throughput of the ABR source. Now we obtain the stationary distribution of the continuous-time process $\{(q(t), \lambda(t))\}$ (or $\{(V(t), \lambda(t))\}$ as the case may be). Since the regeneration epochs of (q_n, λ_{n-1}) are also the regeneration epochs of $\{(q(t), \lambda(t))\}$, Theorem 1 provides the existence and uniqueness of the stationary distribution and convergence to the stationary distribution starting from any initial distribution. Also one can show, using the arguments in [8, Chap. 4], that the stationary distribution of $\{(q(t), \lambda(t))\}$ is given by

$$\hat{\pi}(A) = \frac{1}{T} \int_0^1 \mathbf{P}[(q(u), \lambda(u)) \in A] \, du$$

if (q_0, λ_{-1}) has the stationary distribution π. Thus, by PASTA, the arrivals of the uncontrolled source 2 see the number of customers in the queue with distribution $\hat{\pi}$. The waiting-time distribution of these customers can be calculated using the fact that any customer in the queue is of the ABR source with probability $\mathbf{E}_\pi[\lambda_n]/(\mathbf{E}_\pi[\lambda_n] + \lambda)$ (this way we can calculate the distribution of all the customers in the queue at the time of arrival). Also, using Little's law, from $\hat{\pi}$ we can calculate the mean sojourn time of all the customers in the system. This provides the mean delay $\mathbf{E}W$ of the customers. Since we know the mean delay $\mathbf{E}\{W(2)\}$ of the uncontrolled customers, the mean delay $\mathbf{E}[W(1)]$ of the ABR source is given by

$$\mathbf{E}W = \mathbf{E}[W(1)] \frac{\mathbf{E}_\pi[\lambda_n]}{\mathbf{E}_\pi[\lambda_n] + \lambda} + \mathbf{E}[W(2)] \frac{\lambda}{\mathbf{E}(\lambda_n) + \lambda}.$$

If the queue has a finite buffer length, then in (5) we need the transient probability of a finite queue. Again, for exponential service time, we can explicitly calculate the transition probabilities and the stationary distributions. Then $\mathbf{P}_{\hat{\pi}}(q > 0)$ provides the mean throughput of the system while $\mathbf{E}_\pi[\lambda_n]\mathbf{E}[s(1)] + \lambda \mathbf{E}[s(2)]$ is the mean input intensity. Thus, the stationary fraction of information loss for the system is $\mathbf{P}_{\hat{\pi}}(q > 0)/(\mathbf{E}_\pi[\lambda_n]\mathbf{E}[s(1)] + \lambda \mathbf{E}[s(2)])$. Also, using PASTA, the stationary probability of packet loss of type 2 is $\mathbf{P}_{\hat{\pi}}(q = B)$, where B is the buffer length. Now one can also calculate the stationary probability of packet loss for the ABR source.

If the sampling interval T is also zero and at any time t, $\lambda(t) = f(q(t))$, then for the exponential service time μ, $\{q(t)\}$ becomes a pure birth–death process whose stationary distribution is explicitly available. By continuity results to be provided later, this can approximate the distributions for small D_b, D_f, and T.

Now we consider the case where the propagation delays D_b and D_f are nonzero. First consider the case where $D_b + D_f < T$. Assume that the feedback is being sent back at time instants nT, $n \geq 1$. During the time interval $[nT, (n+1)T]$, the traffic of the uncontrolled source (assuming there is no propagation delay — although it can also be taken care of) arriving at the queue is Poisson with rate λ generated in the time interval $[nT, (n+1)T]$. During time $[(n-1)T + D_f + D_b, nT + D_b + D_f]$, the ABR traffic arriving at the queue has rate λ_{n-1} and during time $[nT + D_b + D_f, (n+1)T + D_f + D_b]$ it has rate λ_n. This information can easily be incorporated in calculating the transition probability matrix in (5). Now it is easy to see how to take care of the situation when $D_b + D_f \geq T$.

Using transient probabilities of the $M \mid M \mid 1$ queue, one can also construct a procedure for calculating the transition probability matrix and the stationary distributions for (q_n, λ_n) for the EPRCA systems, described in Sec. 2. Then one can also calculate the time stationary distributions, mean throughput, and the waiting-time distribution of the two types of customers.

Next we consider a particular feedback policy in more detail. The motivation for this policy is the desire to keep the queue length (workload for the general service times) within a certain interval $[N_0 - \delta, N_0 + \delta]$ for some specified N_0 and δ. As discussed in the introduction, this criterion satisfies the following conflicting performance indices: it provides good mean throughput (when $N_0 - \delta > 0$) by making the probability $\mathbf{P}_{\bar{\pi}}(q = 0)$ low, provides low probability of overflow by making $\mathbf{P}_{\bar{\pi}}(q = B)$ low (if $N_0 + \delta$ is small compared to B), and reduces delay jitter by increasing the probability $\mathbf{P}_{\bar{\pi}}(q \in [N_0 - \delta, N_0 + \delta])$. To increase $\mathbf{P}_{\bar{\pi}}(q \in [N_0 - \delta, N_0 + \delta])$, we consider the following policy:

$$\lambda_n = \begin{cases} \lambda_{n-1} + \Delta_1 & \text{if } q_n < N_0 - \delta, \\ \lambda_{n-1} & \text{if } q \in [N_0 - \delta, N_0 + \delta], \\ \lambda_{n-1} - \Delta_2 & \text{if } q_n > N_0 + \delta, \end{cases} \tag{6}$$

where $\Delta_1 > 0$ and $\Delta_2 > 0$ are appropriately specified. This feedback policy is close to the various proposals discussed at the ATM forum. We can modify (6) by ensuring that $\lambda_n \geq \lambda_{\min}$ and $\lambda_n \leq \lambda_{\max}$. If we make sure that λ_n in (6) is upper bounded, then this policy satisfies the requirements of Theorem 1. We will also consider a simpler form of (6):

$$\lambda_n = \begin{cases} \hat{\lambda} + \Delta_1 & \text{if } q_n < N_0 - \delta, \\ \hat{\lambda} & \text{if } q_n \in [N_0 - \delta, N_0 + \delta], \\ \hat{\lambda} - \Delta_2 & \text{if } q_n > N_0 + \delta, \end{cases} \tag{7}$$

where $\hat{\lambda}$ is appropriately specified (we will discuss it further later on). If Δ_2 is chosen to satisfy the rate condition in Assumption A(iii), this algorithm also satisfies the conditions in Theorem 1. As mentioned above, for general service times, in (6) and (7) we will replace q_n by V_n.

Now we will show analytically that the feedback policy (7) is strictly better than the open-loop policy $\lambda_n \equiv \hat{\lambda}$ for any $\Delta_1 > 0$ and $\Delta_2 > 0$ (with the mentioned restriction on Δ_2) whenever $D_b = D_f = T = 0$, simultaneously with respect to the following criteria: mean throughput, stationary probability of overflow (when $N_0 + \delta < \beta$) and that $\mathbf{P}_{\bar{\pi}}(q \in [N_0 - \delta, N_0 + \delta])$ is higher in the case of (7). Then with the continuity result to be provided later, it will also indicate that for small positive D_b, D_f, and T, the policy (7) will still be strictly better than the open-loop policy. Some of these results have been shown by others (see the references in the introduction) using simulations and numerical examples for the fluid flow and discrete queue models, but not analytically. Since we will consider general i.i.d. service times, instead of queue length, we will consider V_n in (7) and now the buffer length B will indicate the maximum amount of work that can be stored in the buffer. Also, since $T = 0$, we actually consider $V(t)$ and $\lambda(t)$ with the obvious modifications in (7). The process $\{(V(t), \lambda(t))\}$ corresponds to the algorithm (7), while $\{V(t)\}$ will correspond to the open-loop policy $\lambda(t) = \hat{\lambda}$.

To obtain the above-cited result, we form the following construction. First, construct Poisson arrival streams with rates λ and $\hat{\lambda} + \Delta_1$. The service-time sequences for the two kinds of traffic are $\{s_k(1)\}$ and $\{s_k(2)\}$. Now, by thinning the arrival process with rate $\hat{\lambda} + \Delta_1$ we first form a Poisson process with rate $\hat{\lambda}$ and then by thinning this thinned process we obtain a Poisson process with rate $\hat{\lambda} - \Delta_2$. Form the process $\{V'(t)\}$ by using the above Poisson streams with rates λ and $\hat{\lambda}$ and the service streams given above. From the process $\{V'(t)\}$ we construct another process $\{V''(t)\}$ as follows. First assume that $V'(0) \in [N_0 - \delta, N_0 + \delta]$. Let τ_1 be the first time $V'(t) \notin [N_0 - \delta, N_0 + \delta]$. If $V'(\tau_1^+) > N_0 + \delta$, then form $\{V''(t)\}$ as follows: $V''(t) = V'(t)$ for $t \leq \tau_1$ and at time τ_1 use the Poisson stream with rate $\hat{\lambda} - \Delta_2$ constructed above. Let τ_2'' be the time after τ_1 when $V''(t)$ comes back to the set $[N_0 - \delta, N_0 + \delta]$ (then $V''(\tau_2'') = N_0 + \delta$). Also, let $\tau_2 > \tau_1$ be the first time $V'(t)$ returns to the set $[N_0 - \delta, N_0 + \delta]$. By construction, $\tau_2 - \tau_1 \geq \tau_2'' - \tau_1$ for each sample path. Again, let $\tau_3 > \tau_2$ be the time when $V'(t)$ next exits the set $[N_0 - \delta, N_0 + \delta]$. Now we define $V''(t + \tau_2'') = V'(\tau_2 + t)$ for $0 \leq t \leq \tau_3(\omega) - \tau_2(\omega)$. In this way, by patching and shifting we construct the process $V''(t)$. If $V'(\tau_1^+) < N_0 - \delta$, then of course instead of using the Poisson stream with rate $\hat{\lambda} - \Delta_2$ we would have used the stream with rate $\hat{\lambda} + \Delta_1$. One can easily show that the distribution of $\{V''(t), t \geq 0\}$ is same as that of $\{V(t)\}$. Also, we have the following inequalities, sample pathwise:

$$\int_0^t \mathbf{1}\{V'(s) \in [N_0 - \delta, N_0 + \delta]\}\, ds \leq \int_0^t \mathbf{1}\{V''(s) \in [N_0 - \delta, N_0 + \delta]\}\, ds,$$

$$\int_0^t \mathbf{1}\{V'(s) = 0\}\, ds \geq \int_0^t \mathbf{1}\{V''(s) = 0\}\, ds, \tag{8}$$

and

$$\int_0^t \mathbf{1}\{V'(s) = B\}\, ds \geq \int_0^t \mathbf{1}\{V''(s) = B\}\, ds$$

for all $t \geq 0$. When $\Delta_1 > 0$, $\Delta_2 > 0$, if we take the expected value in the above inequalities, these become strict (we have a strict inequality even in stochastic ordering with respect to $\leq st$). Also, if we divide by t and take $t \to \infty$, in the limit we get strict inequalities. This proves all the claims we made above and in fact also shows that these inequalities are valid for any time t.

The proof of the above result also shows the following fact. If in (7) we replace Δ_1 and Δ_2 by $\Delta_1' \geq \Delta_1$ and $\Delta_2' \geq \Delta_2$ and denote the corresponding workload process by $\{\tilde{V}(t)\}$, then the process $\{\tilde{V}(t)\}$ will be better than the process $\{V(t)\}$ with respect to all the above criteria and the corresponding inequalities in (8) continue to hold. Thus, in the class of algorithms of the type (7), with $\hat{\lambda}$ fixed, the unique optimal algorithm will be to make Δ_2 and Δ_1 as large as is allowed. In fact, the optimality holds for all algorithms (1) with $\lambda(t) = f(q(t))$ if we impose the restriction that $\lambda(t) = \hat{\lambda}$ for $q(t) \in [N_0 - \delta, N_0 + \delta]$. There is the question of choosing the optimal $\hat{\lambda}$ (with respect to maximizing $\mathbf{P}_{\hat{\pi}}(V \in [N_0 - \delta, N_0 + \delta])$). If the service times of the two types of traffic are exponential (can have different means), then $\{q(t)\}$ corresponding to (7) is a birth–death process. Choose Δ_1 and Δ_2 to be the largest possible and then calculate the stationary probability $\mathbf{P}_{\hat{\pi}}(q(t) \in [N_0 - \delta, N_0 + \delta])$ (explicitly available, as mentioned before) as a function of $\hat{\lambda}$ and maximize it. One additional advantage of this optimal algorithm is that it minimizes the frequency of change of rate of the ABR source, which is useful in practice. If we do not impose the restriction that $f(q(t)) =$ constant for $q(t) \in [N_0 - \delta, N_0 + \delta]$, the following algorithm will increase $\mathbf{P}_{\pi}[q(t) \in [N_0 - \delta, N_0 + \delta]]$: $f(q(t)) = \min(\lambda)$ for $q(t) > N_0 + \delta$, $f(q(t)) = \overline{\lambda}$ for $q(t) < N_0 - \delta$, and $f(N_0 - \delta) = \tilde{\lambda}_1$, $f(N_0 + \delta) = \tilde{\lambda}_2$, where $\tilde{\lambda}_1$ (and similarly $\tilde{\lambda}_2$) is chosen to optimize the mean sojourn time of $q(t)$ in the set $[N_0 - \delta, N_0 + \delta]$ when $q(0) = N_0 - \delta$ (we will calculate these quantities in what follows). The frequency of change of rate of the ABR source will also be less than the above-mentioned optimal algorithm. At the cost of increasing the frequency of change of rate, we can further improve the algorithm: $f(q(t))$ is same as above for $q(t) \notin [N_0 - \delta, N_0 + \delta]$, but $f(i) = \tilde{\lambda}_i$, where $\tilde{\lambda}_i$ maximizes the mean sojourn time of $q(t)$ in the set $[N_0 - \delta, N_0 + \delta]$ when starting in the state $i \in [N_0 - \delta, N_0 + \delta]$.

The above discussion studies the optimal value λ^* of $\hat{\lambda}$ given N_0 and δ. But even N_0 and δ need to be properly chosen. Start with $N_0 = B/2$ (for finite B; otherwise with a value close to the number of customers served in time $D_b + D_f$, as proposed in [27]). Then, using some iterative optimization (e.g., steepest descent), calculate optimal λ^*, N_0, and δ with respect to maximizing $\mathbf{P}_{\hat{\pi}}[q \in [N_0 - \delta, N_0 + \delta]]$ or some other function of mean throughput, probability of overflow, mean waiting time, and variance of waiting time. We will report elsewhere the stochastic approximation algorithms we have used to obtain such optimal policies.

Now we calculate the mean sojourn time of $q(t)$ in various sets of states for a given algorithm in (1) with $\lambda(t) = f(q(t))$, using the theory of birth-death processes (when again, $D_b = D_f = T = 0$). These can be used to obtain better algorithms, as mentioned in the previous section, but they are also of interest in themselves for any given algorithm. The mean sojourn times of a process in any set of states can of course be calculated from the stationary distributions, but the following analysis provides more detailed results. It will also be relevant for nonzero D_b, D_f, and T because of the continuity result to be provided.

To obtain the sojourn times in different subsets of states we use the results in [26] for the passage times in birth-death process. Let T_{mn} be the first passage time from state m to n for our queueing process and let ${}^rT_{m,n}$ be the conditional first passage time from m to n given that state r is not visited in the meanwhile. For our purposes, the quantities of most interest are the distribution and moments of the sojourn time in the set $[N_0 - \delta, N_0 + \delta]$ starting in states $N_0 - \delta$ or $N_0 + \delta$. Of course, for the finite buffer case with buffer length B, $T_{i,B+1}$, the first time to overflow, starting in state $i \leq B$ will also be very useful.

From [26], we have the Laplace–Stieltjes transform (LST) of the distribution and moments of $T_{i,i+1}$. This provides (by taking $i = N_0 + \delta$) the mean intervisit time to the congested set of states of queue length $>N_0 + \delta$, a quantity of interest to us. From this we can also obtain the moments of the sojourn times in the set $[0, N_0 - \delta - 1]$ starting in the state $N_0 - \delta - 1$. Now we calculate the moments of sojourn time in the set $[N_0 - \delta, N_0 + \delta]$, starting in the state $N_0 + \delta$. We have again from [26], the LST, and the first two moments of ${}^{N_0-\delta-1}T_{N_0+\delta,N_0+\delta+1}$ and the probability that starting in state $N_0 + \delta$, we reach $N_0 + \delta + 1$ before reaching $N_0 - \delta - 1$. We further need the moments of ${}^{N_0+\delta+1}T_{N_0+\delta,N_0-\delta}$. But by [26],

$$ {}^{N_0+\delta+1}T_{N_0+\delta,N_0-\delta} \overset{\text{dist}}{=} {}^{N_0-\delta}T_{N_0-\delta+1,N_0+\delta+1} $$

and also,

$$ {}^{N_0-\delta}T_{N_0-\delta+1,N_0+\delta+1} = {}^{N_0-\delta}T_{N_0-\delta+1,N_0-\delta+2} + {}^{N_0-\delta}T_{N_0-\delta+2,N_0-\delta+3} + \cdots + {}^{N_0-\delta}T_{N_0+\delta,N_0+\delta+1}, $$

where each term on the RHS can be taken independent of others. Also, the LST and moments of each term on the RHS is available in [24]. The above results provide the LST and moments of the sojourn time in the set $[N_0 - \delta, N_0 + \delta]$, starting in the state $N_0 + \delta$.

Next we calculate the LST and moments of the sojourn time in the set $[N_0 - \delta, N_0 + \delta]$, starting in the state $N_0 - \delta$. We have already obtained the information for ${}^{N_0-\delta-1}T_{N_0-\delta,N_0+\delta+1}$. We also have available the probability that starting in the state $N_0 - \delta$, the process reaches $N_0 - \delta - 1$ before reaching $N_0 + \delta + 1$ [24]. Again from [24], as above we can calculate ${}^{N_0+\delta+1}T_{N_0-\delta,N_0-\delta-1}$, and now we obtain the desired result.

It is also of interest to calculate the sojourn times in the states $[N_0 + \delta + 1, \infty)$. If from state i upward the arrival rate stays same, then the distribution of the sojourn time in the set $[i, \infty)$ is same as that of the busy period in the corresponding $M \mid M \mid 1$ queue. In fact, using the busy period of an $M \mid GI \mid 1$ queue, we can also obtain the sojourn times in congested states for the $\{V(t)\}$ process with general service times. Similarly, we can calculate the distribution of the first time to reach the state 0 from the state $N_0 - \delta - 1$. It is possible to obtain other sojourn times for the general service time distributions, using the results in the literature, but they can be substantially more complicated to compute.

The above results can be used as approximations when D_b, D_f, and T are small, using the continuity results to be provided in Theorem 3. But we can also use these techniques directly for nonzero (not necessarily small) D_b, D_f, and T to obtain useful information and also to select a good set of N_0, δ, Δ_1, Δ_2, $\hat{\lambda}$ in algorithm (7) (or other similar algorithms). For example, at time nT (taking $D_b = D_f = 0$ for now), we know (q_n, λ_n). This allows us to calculate moments of sojourn times in various states in time $[nT, (n+1)T]$. We can also calculate various probabilities: for example, if $q_n \in [N_0 - \delta, N_0 + \delta]$, the probability that $q(t)$ will cross over the boundary $N_0 + \delta$ during this time.

Now we provide the promised continuity of the process $\{q(t)\}$. We consider the case of only nonzero T; the propagation delays can be included with minimal changes. We consider a sequence of feedback intervals $T_m \to 0$ (taken to be $1/m$ in the following proof), and the corresponding sequence of workload processes is denoted by $\{V^m(t), t \geq 0\}$. Also, we restrict ourselves to the case of algorithms of the type $\lambda_n = f(V_n)$, where f satisfies the conditions in Theorem 1 because only for these algorithms did we identify the optimal policy. We denote the stationary distribution of $\{V^m(t)\}$ by π^m and that of the limiting process $\{V(t)\}$ by π. All these processes are assumed to have sample paths in the space $D[0, \infty)$ which is considered with the Skorokhod topology (see [5, Chap. 4] and [12, Chap. 3]). The following continuity result is more complicated than usual because $\{V^m(t), t \geq 0\}$ are not Markov processes ($\{V^m(t), \lambda^m(t)\}$ is a nonhomogeneous Markov process) and the limiting process is not a Wiener process.

THEOREM 3. *Under the notation and the assumptions given above, if* $V^m(0) \Rightarrow V(0)$, *f is uniformly continuous,* $\lambda^m \to \lambda$, $s^m(i) \Rightarrow s(i)$, $i = 1, 2$, *and* $\mathbf{E}[s^m(1) + s^m(2)] \leq M_1 < \infty$ *for all m large enough, then* $\{V^{(m)}(t), t \geq 0\} \Rightarrow \{V(t), t \geq 0\}$ *($\Rightarrow$ denotes weak convergence). If, in addition,* $(\overline{\lambda}\mathbf{E}[s^m(1)] + \lambda^m \mathbf{E}[s^m(2)])/(\overline{\lambda} + \lambda^m) < 1$ *and* $\mathbf{E}[s^m(i)] \to \mathbf{E}[s(i)]$, $i = 1, 2$, *then* $\pi^m \Rightarrow \pi$. *Furthermore, if* $\sup_m \mathbf{E}[(s^m(i))^{1+\alpha'}] < \infty$, $i = 1, 2$, $\alpha' > 0$, *then for* $\alpha < \alpha'$ *the αth moment of the stationary moments also converges.*

In the above theorem, we could easily have also included continuity with respect to the function f, i.e., if the function f corresponding to system m is f^m, then we would have the continuity result if $\sup_t |f^m(t) - f(t)| \to 0$. Similarly, instead of $T_m \downarrow 0$ we could have take $T_m \to T > 0$. Of course, as mentioned above, for system m we could include the $D_b^m > 0$ and $D_f^m > 0$ with $D_b^m \to 0$, $D_f^m \to 0$.

The above continuity result shows that, under the conditions provided, the optimal stationary probability of being in $[N_0 - \delta, N_0 + \delta]$ is a continuous function of (D_b, D_f, T) (see [4, Theorem 4.2.2]) and the set of optimal policies is upper semicontinuous. Furthermore, by Theorem 4.2.4 of [4], the set of $\in$-optimal policies is lower semicontinuous. Since at $(D_b, D_f, T) = (0, 0, 0)$ we have a unique optimal policy (we do not know about the uniqueness for nonzero D_b, D_f, T), we obtain the continuity of the optimal policy at $(0, 0, 0)$.

Now we show that $\{V^m(t), t \geq 0\} \Rightarrow \{V(t), t \geq 0\}$ also implies the convergence of distributions and moment of the various passage times we have considered earlier in this section. We show the convergence for the sojourn time in the set $[N_0 - \delta, N_0 + \delta]$; the others will follow in the same way. Let $V^m(0) = y = V(0)$, $y \in [N_0 - \delta, N_0 + \delta]$. Then, denoting the respective sojourn times by τ^m and τ, $\mathbf{1}\{\tau^m > x\} = \mathbf{1}\{\sup_{0 \leq t \leq x} V^m(t) \leq N_0 + \delta\}\mathbf{1}\{\inf_{0 \leq t \leq x} V^m(t) \geq N_0 - \delta\}$. From [12, Problem 26, p. 153], the functions $f_i : D[0,\infty) \to D[0,\infty)$, $i = 1, 2$, defined by $f_1(x)(t) = \sup_{s \leq t} x(s)$ and $f_2(x)(t) = \inf_{s \leq t} x(s)$ are continuous in the Skorokhod topology. Since $\{V(t), t \geq 0\}$ is piecewise continuous and at any t, $V(t)$ is continuous for almost all samples paths, $\mathbf{1}\{\tau > x\}$ is a bounded continuous function and hence $\mathbf{P}[\tau^{(n)} > x] \to \mathbf{P}[\tau > x]$ for each x. Also, since for $V(t) \leq N_0 + \delta$, the arrival intensity is lower bounded (away from zero), the first time to cross $N_0 + \delta$ from i is stochastically upper bounded (with the upper bound having all moments) for all $\{V^m(t)\}$ and $\{V(t)\}$ and hence so are $\{\tau^m, m \geq 0\}$. Therefore, all moments of $\tau^{(m)}$ also converge to that of τ.

The above continuity results are for $\{V^m(t), t \geq 0\}$. If the service times are exponential, we can obtain all these continuity results for the processes $\{q^m(t), t \geq 0\}$ in the same way (in fact, the proofs will be some what easier).

In Theorem 3, for $\pi^m \Rightarrow \pi$ we have assumed $\overline{\lambda}\mathbf{E}[s^m(1)] + \lambda^m\mathbf{E}[s^m(2)] < \lambda^m + \lambda$. This is harmless in practice because even when the workload is low at a time, because of the nonzero D_b, D_f, T, it may be desirable not to overload the system. However, this condition can be removed by imposing extra conditions on $s^m(i)$.

The continuity of the optimal probability in $[N_0 - \delta, N_0 + \delta]$ shows that for small D_f, D_b, and T the optimal closed-loop policy will be strictly better than an optimal open-loop policy. Now let us consider the case where $D_f = D_b = 0$ but T is large. Then, as $T \to \infty$, irrespective of the value of q_n, q_{n+1} will have a distribution close to the stationary distribution of the queue with arrival rate $\lambda + \lambda_n$. If in the interval $[nT, (n+1)T]$ the steady-state performance dominates, which will happen for large T (for exponential service times, it will happen even at moderate T), then, obviously, the optimal open-loop policy will be the best. The effect of nonnegative D_f and D_b can only enhance the arguments.

In the framework of algorithm (7), instead of sending feedback at regular intervals nT, one may just send feedback at the times the queue length crosses the boundary of $[N_0 - \delta, N_0 + \delta]$, informing only in which region the queue has entered. This is the system studied in [27] for discrete time queues and in fact is closer in spirit to the several actual proposals considered at the ATM forum [21]. The analysis of this system is actually simpler than the system we have studied till now, and when $D_b = D_f = 0$, it is just a birth–death process. Thus, the stability (in Theorem 1), comparison and optimality results of this section, and the passage-time calculations hold exactly. Even with D_b and D_f nonzero, the stability of this system follows just as in Theorem 1. Also the continuity of Theorem 3 (as $D_b \to 0$ and $D_f \to 0$) holds without the extra condition for convergence of the stationary distributions.

REFERENCES

1. S. Asmussen, *Applied Probability and Queues*, Wiley, New York (1987).
2. E. Altman and T. Basar, *Optimal Rate Control for High-Speed Telecommunication Networks*, Preprint (1995).
3. E. Altman, F. Baccelli, and J.-C. Bolot, *Discrete Time Analysis of Adaptive Rate Control Mechanisms*, Preprint (1995).
4. B. Bank, J. Guddat, D. Klatte, B. Kummer, and K. Tammer, *Nonlinear Parametric Optimization*, Birkhauser Verlag, Basel (1983).
5. P. Billingsley, *Convergence of Probability Measures*, Wiley, New York (1968).
6. F. Bonomi and K. W. Fendick, "The rate based flow control framework for the available bit rate ATM service," in: *IEEE Network*, March/April (1995), pp. 25–39.
7. F. Bonomi, D. Mitra, and J. B. Seery, "Adaptive algorithms for feedback-based flow control in high speed wide-area ATM networks," *IEEE JSAC*, **13**, 1267–1282 (1995).
8. A. A. Borovkov, *Stochastic Processes in Queueing Theory*, Springer, New York (1976).
9. A. Brandt, P. Franken, and B. Lisek, *Stochastic Stationary Models*, Wiley, New York (1990).
10. D. M. Chiu and R. Jain, "Analysis of the increase and decrease algorithms for congestion avoidance in computer networks," in: *Computer Networks and ISDN Systems*, Vol. 17 (1989), pp. 1–84.
11. B. W. Conolly and C. Langaris, "On a new formula for the transient state probabilities for $M \mid M \mid 1$ queues and computational implications," *J. Appl. Probab.*, **30**, 237–246 (1993).
12. S. N. Ethier and T. G. Kurtz, *Markov Processes, Characterization and Convergence*, Wiley, New York (1986).
13. K. W. Fendick, M. A. Rodrigues, and A. Weiss, "Analysis of a rate based feedback control strategy for long haul data transport," *Performance Evaluation*, **16**, 67–84 (1992).
14. K. W. Fendick and M. A. Rodrigues, "Asymptotic analysis of adaptive rate control for diverse sources with delayed feedback," *IEEE Trans. Inf. Theor.*, **40**, 2008–2025 (1994).
15. V. V. Kalashnikov, *Mathematical Methods in Queueing Theory*, Kluwer, Dordrecht (1994).

16. K. Kawahara, Y. Oie, and H. Miyahara, "Performance analysis of reactive congestion control for ATM networks," *IEEE JSAC*, **13**, 651-660 (1995).
17. S. Keshav, "A control theoretic approach to flow control," in: *Proceedings ACM, SIGCOMM'91* (1991), pp. 3-15.
18. A. Mukherjee and J. C. Strikwerda, "Analysis of dynamic congestion control protocols. A Fokker-Planck approximation," in: *Proceedings ACM, SIGCOMM'91* (1991), pp. 159-169.
19. R. S. Pazhyannur and R. Agrawal, "Feedback based flow control of B-ISDN/ATM networks," *IEEE JSAC*, **13**, 1252-1266 (1995).
20. W. A. Rosenkrantz, "Some martingales associated with queueing and storage processes," *Z. Wahrscheinlichkeits-theor. verw. Geb.*, **58**, 205-222 (1981).
21. K.-Y. Siu and H.-Y. Tzeng, "Intelligent congestion control for ABR service in ATM networks," in: *Proceedings ACM, SIGCOMM'95* (1995), pp. 81-106.
22. O. P. Sharma, *Markovian Queues*, Harwood, New York (1990).
23. V. Sharma, "Reliable estimation via simulation," *Queueing Systems*, **19**, 169-192 (1995).
24. V. Sharma and R. Mazumdar, "Stability and performance of some window and credit based flow controlled in the presence of background traffic" (submitted).
25. S. Shenker, "A theoretical analysis of feedback flow control," in: *Proceedings ACM, SIGCOMM'90* (1990), pp. 156-165.
26. V. Sumita, "On conditional passage time of birth-death processes," *J. Appl. Probab.*, **21**, 10-21 (1984).
27. Y. T. Wang and B. Sengupta, "Performance analysis of a feedback congestion control policy," in: *Proceedings ACM, SIGCOMM'91* (1991), pp. 149-157.
28. N. Yin and M. G. Hluchyj, "On closed-loop rate control for ATM cell relay networks," in: *INFOCOM'94* (1994), pp. 99-108.

Department of Electrical Engineering,
Indian Institute of Science,
560012 Bangalore,
India

Ecole Polytechnique,
Department de Génie Electrique et de Genie Informatique,
Case Postale 6079, Succ A, H3C 3A7 Montreal,
Canada