University of Würzburg
Institute of Computer Science
Research Report Series

Analysis of a Queueing Model with Delayed Feedback and its Application to the ABR Flow Control

Michael Ritter

Report No. 164 January 1997

Institute of Computer Science, University of Würzburg
Am Hubland, D-97074 Würzburg, Germany
Tel.: +49 931 888 5505, Fax: +49 931 888 4601
e-mail: ritter@informatik.uni-wuerzburg.de

Abstract

A common method for regulating the input flow in today's telecommunication networks is the implementation of reactive traffic control mechanisms. By limiting the user's access according to the current network load situation, the throughput is optimized while congestion is avoided. In ATM networks, for example, a rate-based feedback control is applied to dynamically adjust the transmission rates of connections belonging to the ABR service category. This paper presents a discrete-time analysis of a queueing model with non-deterministic arrivals and delayed feedback. The binary feedback information is used to regulate the input flow by changing the inter-arrival distribution accordingly. Applying this model, the performance of the ABR flow control can be studied with respect to the variability of the cell generation process.

Keywords

ATM, binary feedback, discrete-time analysis, rate-based flow control, signaling delay.

1 Motivation

In integrated telecommunication networks based on the *Asynchronous Transfer Mode* (ATM), the control and management of services with different traffic characteristics and *Quality of Service* (QoS) demands is a main issue. Depending on the predictability of the traffic volume and the real-time requirements of an application, the most appropriate ATM service category is chosen for transmission. For each of the service categories, a number of distinctive traffic control mechanisms are invoked at connection setup in order to guarantee the requested QoS for the connection itself and to avoid QoS degradations for other active connections.

In the ATM context, traffic control mechanisms can be grouped into two major classes. The first class is formed by preventive mechanisms, such as *Connection Admission Control* (CAC) and *Usage Parameter Control* (UPC). They are applied for each of the service categories and protect the network from overload situations due to application misbehavior and system malfunction. Reactive control mechanisms, such as *Feedback Flow Control* (FFC), belong to the second class. They regulate the traffic flow of a connection in order to avoid network overload and to obtain a high network throughput. Applications which are typically controlled by this kind of mechanism are loss-sensitive services without real-time requirements. The corresponding service category is called *Available Bit Rate* (ABR) service category, which utilizes the transmission capacity left over by real-time connections of higher priority.

Due to the large product of bandwidth and delay in today's telecommunication networks, the design of feedback controls is a difficult task. Since the information about transmission rate adjustments at least requires the propagation delay to arrive from the bottleneck to the sending end system, data has to be buffered during overload periods and the throughput temporary decreases when the available capacity increases. To define an appropriate control mechanism and to find suitable control parameter sets, extensive analyses and simulation studies have to be carried out. Objectives are a small buffer space required to avoid losses and a high throughput.

This paper presents a discrete-time analysis of a queueing model with delayed feedback and non-deterministic customer arrivals. Binary feedback information is used to regulate the input flow by changing the inter-arrival distribution according to the buffer space available in the bottleneck queue. The model allows to investigate different control strategies by adjusting a small number of system parameters. Furthermore, the performance of the binary mode of the ABR flow control defined by the ATM Forum [2] can be studied with respect to the variability of the cell generation process.

The organization of the paper is as follows. In Section 2 we introduce the feedback-controlled queueing model and outline its analysis. We look at specific time instants within a control cycle to obtain information about the dynamics of the queue length and the arrival process. The accuracy of our analysis, which is of an approximate nature, is discussed in Section 3. Section 4 shows how our queueing model can be applied to investigate the influence of the variability of the cell generation process on the performance of the ABR flow control mechanism. We conclude the paper with final comments in Section 5.

2 Queueing model and its analysis

Our queueing model operates in the discrete-time domain and consists of a single server queue fed by a feedback controlled source, cf. Figure 1. The service time T is deterministic and the queue length is infinite. After every fixed number N of customers served, binary feedback information about the current congestion status of the queue is periodically signaled back to the source. To monitor congestion, two thresholds for the queue length are used. Congestion is detected if the queue length exceeds the upper threshold Q_H and regarded as terminated if the queue length becomes shorter than the lower threshold Q_L.

When the signaling information, which is delayed by a fixed amount of time τ, arrives at the source, the customer generation process is adjusted accordingly. Customer generation is throttled down each time congestion is indicated and accelerated during periods without congestion. We assume that the customers are generated according to a general inter-arrival distribution between two consecutive arrivals of feedback information.

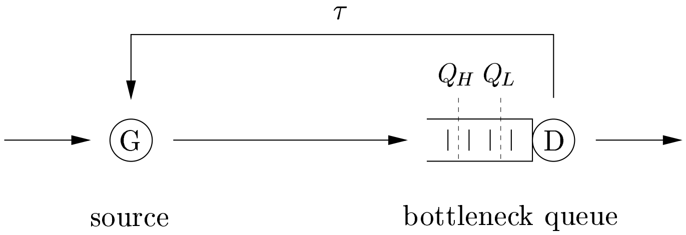


Figure 1: *Queueing model with binary feedback and signaling delay*

Depending on the feedback information signaled, the inter-arrival distribution is changed. In the following we consider a number of distributions $c_k(\cdot)$ with $k = 1, \ldots, m$. If $c_k(\cdot)$ represents the inter-arrival distribution used currently for customer generation, then $c_{k+1}(\cdot)$ and $c_{k-1}(\cdot)$ are used after the arrival of the next negative and positive feedback information, respectively. I.e., the source index k is increased during congested periods and decreased otherwise. For $k = 1$ no further decrease and for $k = m$ no further increase of the source index occurs. To guarantee stability, we assume that

- $\mathrm{E}[c_k(\cdot)] < \mathrm{E}[c_{k+1}(\cdot)] \quad \text{for} \quad k = 1, \ldots, m-1$,
- $0 < \mathrm{E}[c_1(\cdot)] < T$,
- $T < \mathrm{E}[c_m(\cdot)] < \infty$,

where $\mathrm{E}[x(\cdot)]$ denotes the mean of the distribution $x(\cdot)$. These assumptions are not necessarily required, even for the analysis we present. Other configurations may also lead to a stable system operation. Furthermore, the source index k can be changed in a non-linear manner or by more than one at each arrival of feedback information.

In the literature, a large variety of queueing systems is investigated under non-stationary conditions. Besides theoretical studies, which concentrate on access regulation schemes [7] and transient system dynamics for time-varying arrival processes [14, 19], a large number of papers has been published on feedback controls for communication networks. Most of them consider saturated, i.e. greedy, sources and use a differential equation approach to analyze the stability of the control law [8, 9, 10, 13], the system dynamics [4, 6, 15], the buffer size required to avoid losses [17, 20] and fairness issues [5]. Others study the performance of control laws for non-deterministic service times with respect to the design and stability of the law [1, 12], the throughput [3] and the buffer requirements [18].

In contrast, our queueing model considers a time-varying stochastic arrival process instead of saturated sources. A similar assumption is also made in [11, 16], where the authors use Markov-modulated arrival processes to investigate the performance of feedback controls. A drawback of their analytical approaches is the computation of performance measures which are observed for a time interval of infinite length. These measures do not give insight in the dynamical system behavior, which occurs due to the delayed signaling of feedback information and stands for the periodic oscillation of the queue length and the source index in case of our model. The analysis we present in the following allows to compute system characteristics at specific time instants within a control cycle. By control cycle we denote a period consisting of one decrease and one increase phase of the source index k.

Figure 2 shows the typical evolution of the source index and the queue length over a control cycle observed for our model. As long as congestion is signaled, the source index is increased, which throttles down the arrival process. When the queue length falls below the lower threshold Q_L at t_1, congestion is released and after the signaling delay, i.e. at t_2, the source index is decremented each time positive feedback is received. In our model, the total signaling delay consists of the fixed delay τ and the time until the next feedback information is generated.

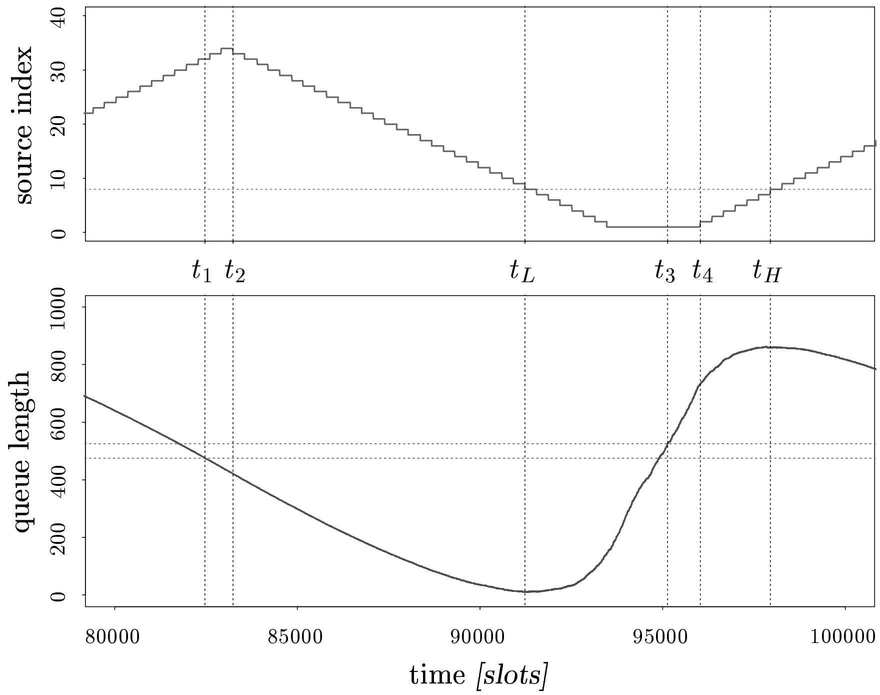


Figure 2: *Example evolution of the source index and the queue length*

By decrementing the source index k the arrival process is accelerated and the queue length has its minimum value within this cycle when the mean arrival rate exceeds the service rate $1/T$. In Figure 2, the minimum queue length is reached at t_L. From now on, the queue length starts to increase while the source index further decreases. Congestion is detected when the queue length exceeds the upper threshold Q_H at t_3. Again, a feedback delay of τ plus the time to generate the next feedback information is required until the source reacts to the detection of congestion at t_4 by increasing its index k. Within a control cycle, the maximum queue length is reached at t_H, i.e. when the mean arrival rate falls below the service rate $1/T$.

Each control cycle contains all of the six time instants mentioned above. In general, the queue length and the source index observed at these instants differ for each control cycle due to the non-deterministic nature of the arrival process. To analyze this feedback controlled queueing model we first compute the system state in equilibrium at $t_1, \ldots, t_4$ expressed by the distributions $q_{t_j}(\cdot)$ and $s_{t_j}(\cdot)$ for $j = 1, \ldots, 4$. The vector elements $q_{t_j}(i)$ and $s_{t_j}(k)$ denote the probability that the queue length is equal to i and the source index is equal to k at the time t_j, respectively. Based on these distributions, we derive the queue length distributions $q_L(\cdot)$ and $q_H(\cdot)$ at t_L and t_H. When knowing the system state distributions, the system dynamics, i.e. the variation of the source index and the queue length over a control cycle, can be investigated.

We start with the system state distributions at t_1 when congestion is released. At this time, the queue length drops to the lower threshold Q_L and we obtain

$$q_{t_1}(i) \quad = \quad \delta(i - Q_L) \quad , \tag{1}$$

where $\delta(\cdot)$ denotes the *Kronecker Delta* with $\delta(0) = 1$ and 0 otherwise. For the source index distribution $s_{t_1}(\cdot)$ we have to assume a hypothetical distribution. The real distribution can be obtained by performing the following steps iteratively.

To compute the source index distribution $s_{t_2}(\cdot)$ we need to know how often the index k is increased during the signaling delay due to receiving outdated feedback information, cf. Figure 2. As mentioned above, the signaling delay consists of the fixed delay τ and a variable delay until the next feedback information is generated. This additional delay varies between 1 slot, i.e. the next customer served causes the generation of feedback information, and NT slots, i.e. feedback information was generated right before congestion release. Assuming the length of this period to be uniformly distributed, which is motivated by the independence of the service and arrival process, we obtain for the source index distribution at t_2

$$\begin{aligned} s_{t_2}(k) \quad = \quad & (1 - (\tau/NT - \lfloor \tau/NT \rfloor)) \cdot s_{t_1}(k - \lfloor \tau/NT \rfloor) + \\ & (\tau/NT - \lfloor \tau/NT \rfloor) \cdot s_{t_1}(k - \lfloor \tau/NT \rfloor - 1) \quad , \end{aligned} \tag{2}$$

where $\lfloor x \rfloor$ denotes the largest integer smaller than x. In other words, during the signaling delay the index k is either increased by $\lfloor \tau/NT \rfloor$ or by $\lfloor \tau/NT \rfloor + 1$, depending on the length of the variable part of the delay. For a control law where the index increases in a non-linear manner or by more than one at each arrival of feedback information, the two expressions above have to be adapted accordingly. Of course, the index is upper bounded by $k = m$, which is not considered in Equation (2) and in the following in order to increase the readability of the equations. The same holds for the lower bound $k = 1$.

To derive the queue length distribution $q_{t_2}(\cdot)$ from the system state distributions at t_1, we consider each source index with $s_{t_1}(k) \neq 0$ separately. For a particular k we obtain the queue length distribution $q_{t_2}^k(\cdot)$ by adding the number of arrivals to and subtracting the number of departures from the queue length at t_1 for each possible signaling delay, i.e.

$$q_{t_2}^k(i) \quad = \quad \frac{1}{NT} \sum_{t=1}^{NT} \pi_0 \left[q_{t_1}(i) \star a_k^{\tau+t}(i) \star d^{\tau+t}(-i) \right] \quad . \tag{3}$$

Again, we assume a uniform distribution for the variable part of the signaling delay. The operator $\star$ denotes the discrete convolution operation and the function $\pi_0[\cdot]$, which is defined by

$$\pi_0[x(i)] \quad = \quad \begin{cases} 0 & : \quad i < 0 \\ \sum\limits_{j=-\infty}^{0} x(j) & : \quad i = 0 \\ x(i) & : \quad i > 0 \end{cases} \quad , \tag{4}$$

suppresses the calculation of negative queue lengths. Denoting the distribution for the number of customers generated in an interval of length t by $\alpha_k^t(\cdot)$ for a source index equal to k, we obtain the distribution $a_k^t(\cdot)$ by

$$a_k^t(i) \quad = \quad \alpha_k^{t-NT\cdot\lfloor t/NT \rfloor}(i) \star \alpha_{k+1}^{NT}(i) \star \cdots \star \alpha_{k+\lfloor t/NT \rfloor}^{NT}(i) \quad . \tag{5}$$

Thus, each interval between two consecutive arrivals of feedback information is considered separately. The length of the first interval is determined by t, since the signaling delay ends at t_2 with the arrival of positive feedback. For the computation of $\alpha_k^t(\cdot)$ we assume that the time until the first customer generation is distributed according to the recurrence time distribution of $c_k(\cdot)$. This is, of course, not valid in general and the segmentation into separate intervals also involves an inaccuracy. However, as we will see later, the error is rather small and neglectable for practical purposes.

Finally, the distribution $d^t(\cdot)$ expressing the number of departures during t time slots is determined by

$$d^t(i) \quad = \quad \delta(i - \lfloor t/T \rfloor) \quad . \tag{6}$$

Knowing the distributions $q_{t_2}^k(\cdot)$, we arrive at the queue length distribution $q_{t_2}(\cdot)$ by summing up the distributions weighted by the corresponding probability $s_{t_1}(k)$:

$$q_{t_2}(i) \quad = \quad \sum_{k=1}^{m} s_{t_1}(k) \cdot q_{t_2}^k(i) \quad . \tag{7}$$

Next, we outline the derivation of the system state distributions at t_3. At this time, the upper threshold Q_H is exceeded, congestion is detected and feedback information for increasing the source index will be signaled. Clearly, the queue length distribution $q_{t3}(\cdot)$ is given by

$$q_{t_3}(i) \quad = \quad \delta(i - Q_H) \quad . \tag{8}$$

For the computation of the source index distribution $s_{t_3}(\cdot)$ we proceed as follows. Again, each index with $s_{t_2}(k) \neq 0$ is considered separately. The queue length distribution $\hat{q}^{k-j}(\cdot)$

just before decreasing the source index the j-th time after receiving the congestion release feedback at t_2 can be approximated by

$$\hat{q}^{k-j}(i) \;=\; \eta^{Q_H}\left[\hat{q}^{k-j+1}(i)\right] \star \alpha_{k-j}^{NT}(i) \star d^{NT}(-i) \tag{9}$$

for $j > 0$, if we set $\hat{q}^k(i) = q_{t_2}(i)$ (cf. Equation (3)). The function $\eta^{Q_H}[x(\cdot)]$ extracts the part lower than Q_H from the distribution $x(\cdot)$ and normalizes it. This operation must be performed since congestion is detected and t_3 is already reached if the upper threshold Q_H is exceeded.

Consequently, we obtain the source index distribution $s_{t_3}^k(\cdot)$ at t_2 for a particular index k by

$$s_{t_3}^k(l) \;=\; \sum_{j=Q_H}^{\infty} \hat{q}^l(j) \cdot \left(1 - \sum_{l'=l+1}^{k} s_{t_3}^k(l')\right) \quad . \tag{10}$$

The probabilities for $l \geq k$ are equal to zero. The source index distribution $s_{t_3}(\cdot)$ is finally determined by the weighted sum of the individual distributions, i.e.

$$s_{t_3}(l) \;=\; \sum_{k=1}^{m} s_{t_2}(k) \cdot s_{t_3}^k(l) \quad . \tag{11}$$

The computation of $s_{t_3}(\cdot)$ does not take into consideration periods where the queue is empty. During such periods, the feedback information arrives at the source with a lower frequency than $1/NT$. The error is, however, neglectable if the signaling delay is more than one order of magnitude shorter than the length of a control cycle, which is generally the case.

To compute the system state distributions observed at t_4 out of the distributions observed at t_3, we proceed similar as for the computation of the measures at t_2. In principal, we only have to consider a decrease of the source index instead of an increase. The behavior regarding the signaling delay and the length of the periods between two consecutive arrivals of feedback information are the same.

For the computation of the source index distribution $s_{t_4}(\cdot)$, Equation (2) changes to

$$\begin{aligned} s_{t_4}(k) \;=\; & (1 - (\tau/NT - \lfloor \tau/NT \rfloor)) \cdot s_{t_3}(k + \lfloor \tau/NT \rfloor) \; + \\ & (\tau/NT - \lfloor \tau/NT \rfloor) \cdot s_{t_3}(k + \lfloor \tau/NT \rfloor + 1) \quad . \end{aligned} \tag{12}$$

In the Equations (3) and (7), the indices t_1 and t_2 are substituted by t_3 and t_4, respectively, to derive the queue length distribution $q_{t_4}(\cdot)$. Equation (5) has to be changed to

$$a_k^t(i) \;=\; \alpha_k^{t-NT\cdot\lfloor t/NT \rfloor}(i) \star \alpha_{k-1}^{NT}(i) \star \cdots \star \alpha_{k-\lfloor t/NT \rfloor}^{NT}(i) \quad , \tag{13}$$

since the source index is now decreased during the signaling delay.

In order to enable an iterative computation of the system state distributions, we finally have to compute the source index distribution $s_{t_1}(\cdot)$ from the distributions obtained for the instant t_4. Since an increase phase of the source index can be treated similarly to a decrease phase, we can make use of the findings for the step from t_2 to t_3. The queue length distributions $\hat{q}^{k+j}(\cdot)$ just before increasing the source index the j-th time after receiving the congestion detection feedback at t_4 are approximated by

$$\hat{q}^{k+j}(i) \quad = \quad \eta_{Q_L}\left[\hat{q}^{k+j-1}(i)\right] \star \alpha^{NT}_{k+j}(i) \star d^{NT}(-i) \quad . \tag{14}$$

The function $\eta_{Q_L}[x(\cdot)]$ extracts the part higher than Q_L from the distribution $x(\cdot)$ and normalizes it. To obtain the source index distribution for a particular index k, Equation (10) has to be exchanged by

$$s^k_{t_1}(l) \quad = \quad \sum_{j=-\infty}^{Q_L} \hat{q}^l(j) \cdot \left(1 - \sum_{l'=k}^{l-1} s^k_{t_1}(l')\right) \quad . \tag{15}$$

In Equation (11), the indices t_2 and t_4 are substituted by t_4 and t_1, respectively.

Applying the analysis described by the Equations (1) to (15) iteratively, we can compute the system state distributions in equilibrium at the time instants $t_1, \ldots, t_4$. Besides these performance measures, we are also interested in the minimum and maximum queue length over a control cycle, which are attained at t_L and t_H, cf. Figure 2. Since these values also vary for each control cycle, we derive the distributions $q_L(\cdot)$ for the minimum and $q_H(\cdot)$ for the maximum queue length observed in equilibrium in the following.

A good approximation for the distribution of the minimum queue length is obtained by computing the queue length distribution at the time just before the mean customer generation rate exceeds the service rate due to the decrease of the source index k. From this instant on, more customers arrive at the queue than depart from it and thus, the queue length begins to increase. Starting with $s_{t_2}(\cdot)$ and $q_{t_2}(\cdot)$, Equation (9) allows to compute the queue length distribution $\hat{q}^{k-*}_{k'}(\cdot)$ just before the source index falls to a k' where the customer generation rate exceeds the service rate. By adding the distributions $\hat{q}^{k-*}_{k'}(\cdot)$ for all initial values of the source index k at t_2 and weighting them accordingly, we arrive at an estimation for the distribution of the minimum queue length:

$$q_L(i) \quad = \quad \sum_{k=1}^{m} s_{t_2}(k) \cdot \hat{q}^{k-*}_{k'}(i) \quad . \tag{16}$$

Similar considerations can be made to obtain the maximum queue length distribution $q_H(\cdot)$, starting with $s_{t_4}(\cdot)$ and $q_{t_4}(\cdot)$.

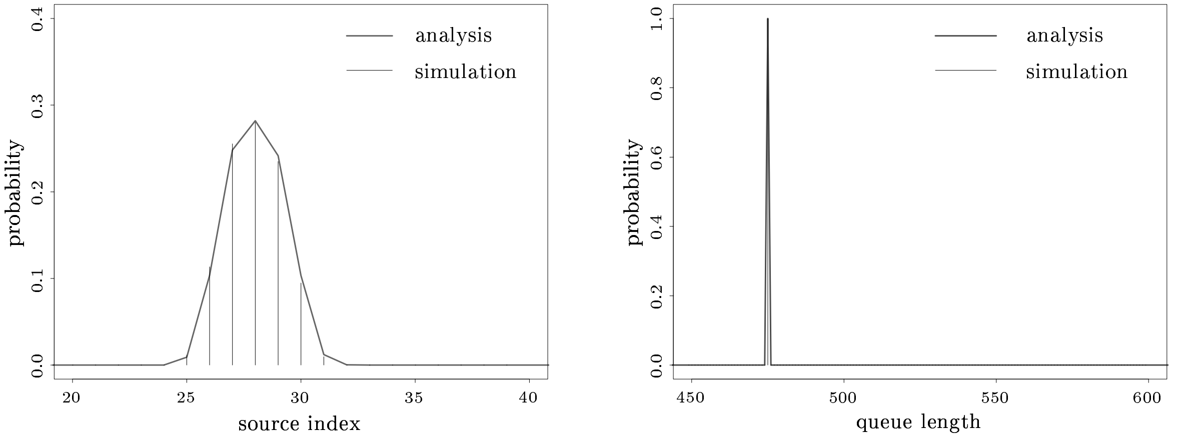
3 Approximation accuracy and system performance

Due to neglecting periods where the queue is empty, the decomposition of the control cycle into four time intervals and the assumption made in Equation (5), several small approximation steps are involved in the analysis outlined in the last section. In the following we present a number of numerical results to discuss the accuracy of our approach.

We start with the system state distributions observed at $t_1, \ldots, t_4$. To compare results computed using our analysis with simulation results, we consider a queue with a deterministic service time of $T = 10$ slots and set the thresholds for congestion monitoring to

$Q_L = 475$ and $Q_H = 525$. The signaling information is sent every $N = 32$ customers served, delayed by a fixed time of $\tau = 320$ slots. For customer generation we use a geometric distribution with mean $\mu = 2$ shifted according to the source index. I.e., the whole distribution is shifted by k units for the source index k. The maximum index is set to $m = 64$.

Figure 3 shows the source index and the queue length distribution at t_1, i.e. the time congestion is released. The histogram represents the simulation results and the solid lines join the results computed with our analysis.



(a) *Source index distribution at t_1*

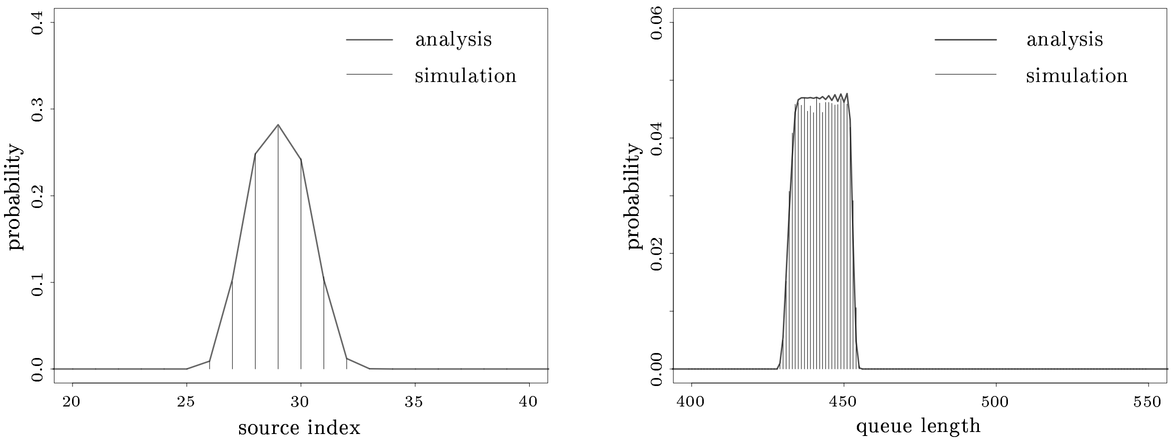
(b) *Queue length distribution at t_1*

Figure 3: *System state distributions at congestion release* (t_1)

Looking at the source index distribution depicted in Figure 3(a), we observe a high accuracy of our analytical approach. For the system configuration we used, the source index varies around $k = 28$ when congestion is released. The analytical results for the queue length distribution are exact, since the queue length equals to the lower threshold Q_L at t_1, cf. Figure 3(b).

The next instant we look at is the time t_2 just before the source index starts to decrease due to receiving the first congestion release information. At this time, the index has its largest value within a control cycle. Figure 4 shows the corresponding distributions for the source index and the queue length. As before, we observe a good approximation of the simulation results in both cases. The source index now varies around $k = 29$, which is explained by the fixed signaling delay of $\tau = 320$ slots. This is the time required to serve $N = 32$ customers. The mass of the queue length distribution at t_2 is located below the lower threshold Q_L, since the service rate $1/T$ is higher than the mean customer arrival rate during the feedback delay.

If we take a closer look at the analytical results in Figure 4(b), we observe a number of small peaks a the right hand side of the distribution. These peaks occur due to the separate calculation of the number of departures from the queue for each interval bounded by two consecutive arrivals of feedback information, cf. Equation (6).

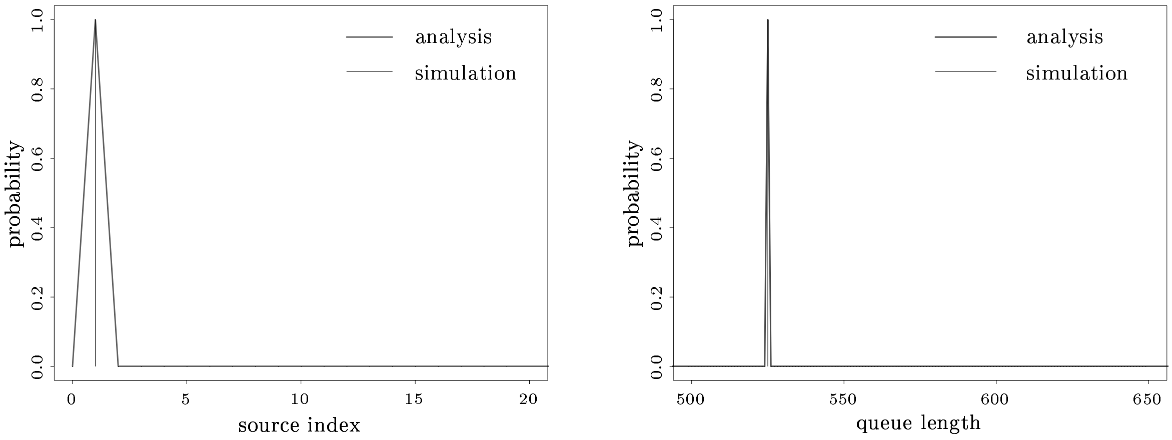


(a) *Source index distribution at t_2*

(b) *Queue length distribution at t_2*

Figure 4: *System state distributions before the decrease of the source index* (t_2)

The distributions for the time when congestion is detected, i.e. t_3, are shown in Figure 5. The source index attains the minimum value $k = 1$ within each control cycle in case of the parameter set we used for this example, cf. Figure 5(a). For the particular control law we look at, it is in general difficult to find parameter sets where both the queue does not become empty and the source index starts to increase before $k = 1$ is reached. The distribution for the queue length in Figure 5(b) is also deterministic, since the number of customers queued equals to the upper threshold Q_H at t_3.

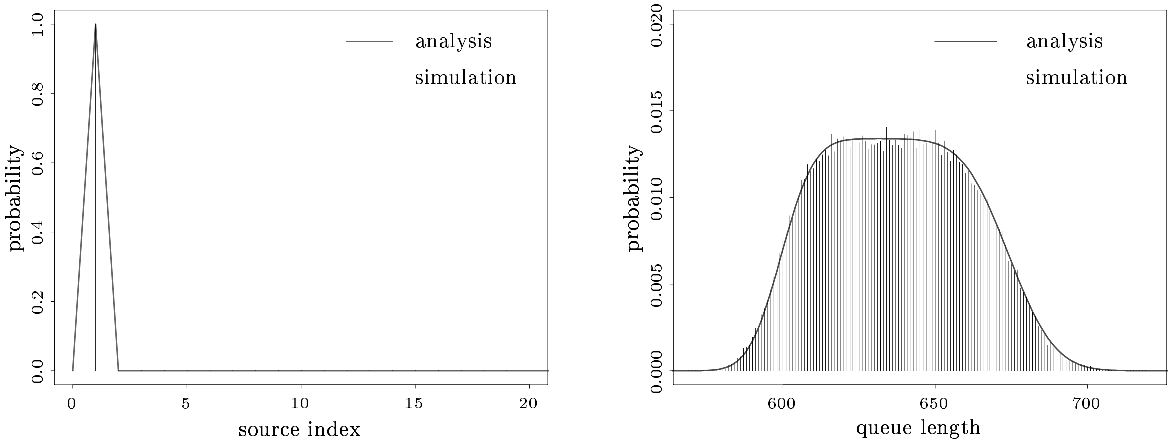


(a) *Source index distribution at t_3*

(b) *Queue length distribution at t_3*

Figure 5: *System state distributions at congestion detection* (t_3)

The next instant we look at is the time t_4 just before the first congestion detection information arrives at the source. Since the source index has already attained its minimum at t_3, the index can not be decreased further during the signaling delay and the corresponding distribution is the same as that measured at t_3, cf. Figure 6(a). The mass of the queue length distribution depicted in Figure 6(b) is now located beyond the upper threshold Q_H, which is explained by a mean arrival rate higher than the service rate $1/T$. Again, the approximation accuracy of our approach is high.

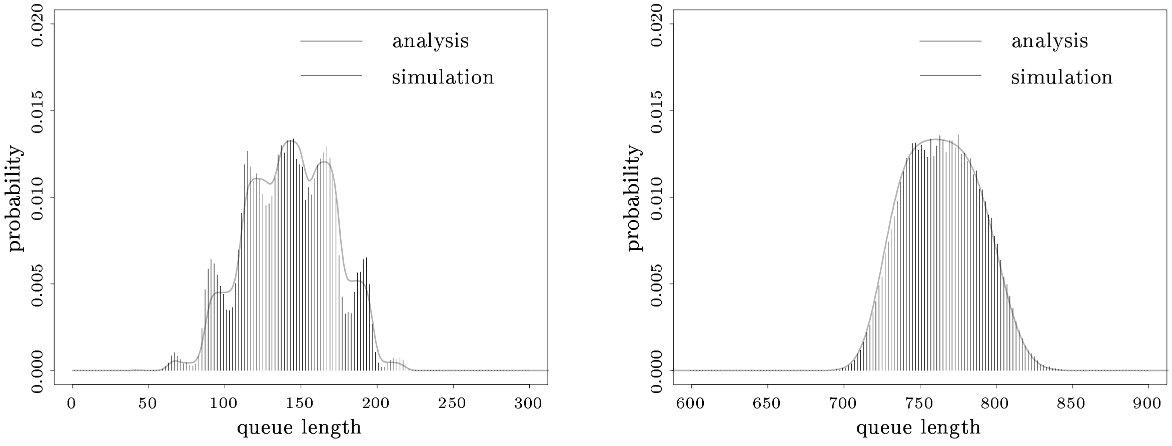


(a) *Source index distribution at t_4*

(b) *Queue length distribution at t_4*

Figure 6: *System state distributions before the increase of the source index (t_4)*

To investigate the dynamics of the queue length over a control cycle we make use of the queue length distributions at t_L and t_H, where the queue length has its local minimum and maximum, respectively. Figure 7 shows these distributions for our example parameter set.



(a) *Queue length distribution at t_L*

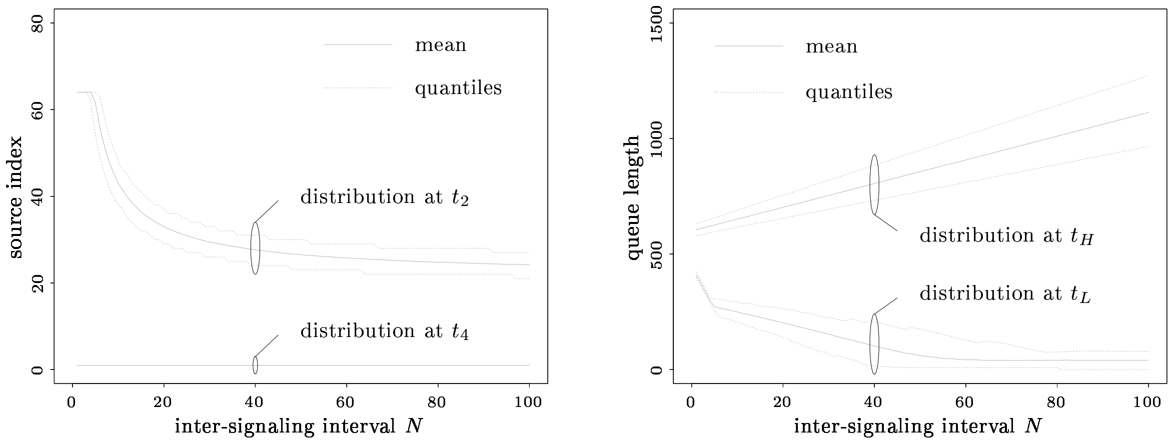
(b) *Queue length distribution at t_H*

Figure 7: *Queue length distributions at the local minimum (t_L) and maximum (t_H)*

Looking at the distribution at t_L in Figure 7(a), we observe a good agreement between simulation and analysis. The peaks result from the source index distribution at t_2, where each peak corresponds to a particular index k. In case of the simulation results the peaks are more clear than in case of our analysis. This inaccuracy is explained by the separate evaluation of the customer arrivals and departures for each interval bounded by two consecutive arrivals of feedback information. In such an interval, the first arrival is assumed to occur after a time period which follows the recurrence time distribution of the current arrival process, cf. Equations (5) and (13). The consequence is a more spread queue length distribution for each initial index k.

This behavior is not observed for the queue length distribution at t_H depicted in Figure 7(b), where the single peak results from the deterministic source index at t_4. Due to the low source index between t_4 and t_H, the customer arrivals occur more frequently and the assumption of the recurrence time distribution only involves a minor approximation error. If the lower bound for the source index is set to a higher value, we observe the same spreading effect of the queue length distribution as described above.

After discussing the accuracy of our analysis we present a number of numerical results to show the influence of the most important system parameters on the performance of the queueing model. In Figure 8, the mean and both 10^{-3}-quantiles of the source index and the queue length distribution are plotted as a function of the length of the inter-signaling interval N. The other system parameters are set as described above.



(a) *Source index distribution*

(b) *Queue length distribution*

Figure 8: *System performance for different signaling frequencies*

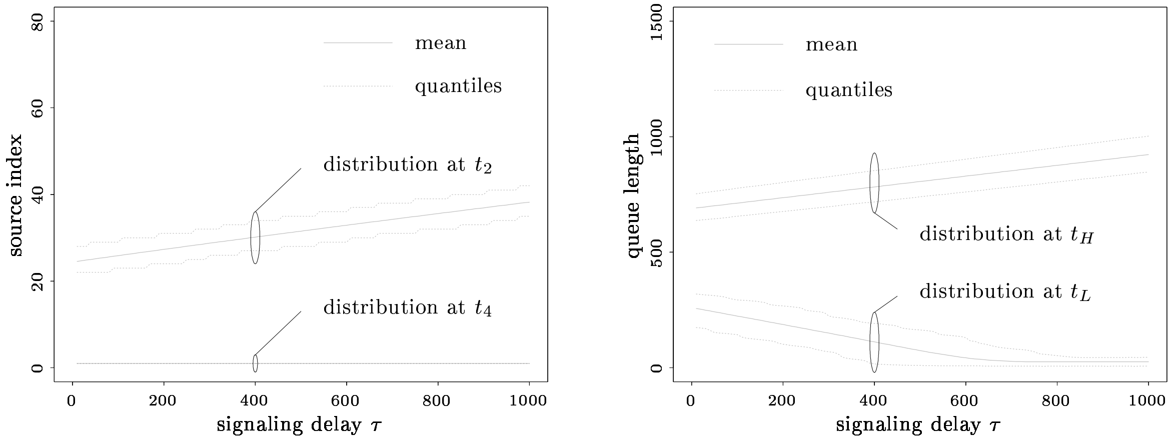
If the length of the inter-signaling interval increases, we observe an exponential decrease of the mean and the quantiles of the source index distribution measured at t_2, i.e. the time the index has its maximum value over a control cycle, cf. Figure 8(a). The variability, which is expressed by the difference of the quantiles, remains almost the same. The stepwise decrease results from the discrete nature of the source index. Since the minimum source index is attained during each control cycle, the mean and the quantiles measured at t_4 are constant and equal to one. For system parameter sets where the minimum index is not attained, the mean and the quantiles measured at t_4 increase exponentially.

In contrast, the mean and the quantiles of the queue length distribution at t_L and t_H

decrease and increase linearly with the length of the inter-signaling interval, respectively (cf. Figure 8(b)). I.e., the variability of the queue length over a control cycle increases linearly with the number of customers which have to be served until the next signaling information is generated. The bend in case of the distribution measured at t_L results from the upper limit for the source index, which is attained for $N \leq 5$. Looking at the quantiles, we observe a slight and linear increase of the variability of the local minimum and maximum within a control cycle with the length of the inter-signaling interval N.

An optimal choice for the signaling frequency is the generation of feedback information after every 20 to 40 customers served. For such a choice, the variability of the source index does only decrease slightly with the length of the inter-signaling interval and the queue length required to avoid customer loss is reasonable short.

With the plots depicted in Figure 9 we investigate the influence of the fixed signaling delay τ on the system behavior. The mean and the quantiles of both the source index distribution and the queue length distribution show a linear dependency on the fixed delay. For an increasing delay τ, the difference between the minimum and maximum values within a control cycle increases while the variability for each measure itself is almost not affected. The source index distribution at t_4 shows the same behavior if a parameter set is chosen where the minimum $k = 1$ is not attained.



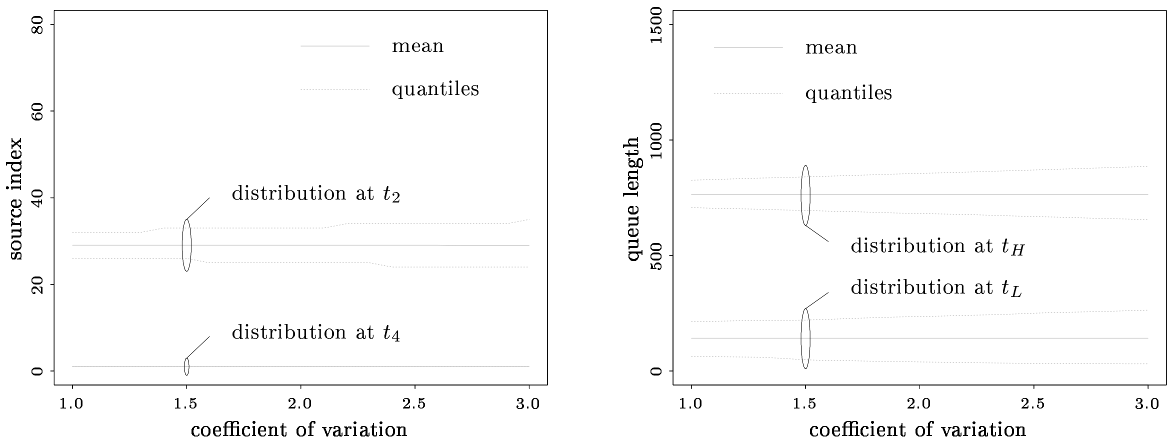
(a) *Source index distribution*

(b) *Queue length distribution*

Figure 9: *Influence of the signaling delay τ on the system performance*

Our last example in this section considers the variability of the arrival process. Instead of the geometric distribution used for the examples before we now use a negative-binomial distribution to generate the customer arrivals. The mean is again $\mu = 2$ and the coefficient of variation is varied between 1.0 and 3.0. The inter-arrival distributions $c_k(\cdot)$ are generated in the same way and all other system parameters are set as described.

In Figure 10(a) we observe an independency of the mean source index measured at t_2 and t_4. The same holds for the mean queue length measured at the local minimum and maximum within a control cycle, cf. Figure 10(b).



(a) *Source index distribution*

(b) *Queue length distribution*

Figure 10: *Robustness against the variation of the arrival process*

Regarding the variability of the local minimum and maximum of both system state distributions, a slight linear increase with the coefficient of variation can be recognized in Figure 10. However, the control law we investigated in this section is quite robust against the variation of the basic process used to generate the inter-arrival distributions.

4 Application to the ABR flow control mechanism

In ATM networks, a feedback flow control is applied for connections belonging to the *Available Bit Rate* (ABR) service category. The control mechanism developed by the ATM Forum [2] supports two signaling modes: a simple binary mode and an enhanced explicit rate mode. In the following we show how our queueing model can be used to evaluate the performance of the binary signaling mode with respect to the variability of the cell generation process.

A distinguishing feature of ABR sources is their ability to submit cells into the network at a variable but controlled rate, the *Allowed Cell Rate* (ACR). The ACR is controlled by the return of *Resource Management* (RM) cells, which are sent periodically by the source end systems after each submission of N_{rm} data cells and are looped back by the destination end systems. As the RM cells travel through the network, the switches provide information about their current congestion status by modifying the content of these cells. When an RM cell arrives back at the source, the ACR is reset based on the information carried by the RM cell. If congestion is indicated, the source must decrease its ACR, otherwise it may increase its ACR by a fixed amount.

In the binary mode of the flow control mechanism, the *Explicit Forward Congestion Indication* (EFCI) bit located in the header of data cells is set by the switches to indicate congestion for a particular connection. According to this information, the destination modifies the *Congestion Indication* (CI) bit of backward RM cells. At the return of an

RM cell carrying positive feedback, i.e. no congestion was detected by the switches, the source may increase its ACR by a fixed increment negotiated at connection setup. If the CI bit is set, i.e. congestion was detected, the ACR is decreased by an amount relative to its current value.

Figure 11 shows the typical evolution of the ACR and the queue length over a control cycle for an ABR connection running through a single bottleneck switch. In contrast to the control law studied in the last section, the rate now decreases in an exponential manner during congested periods.

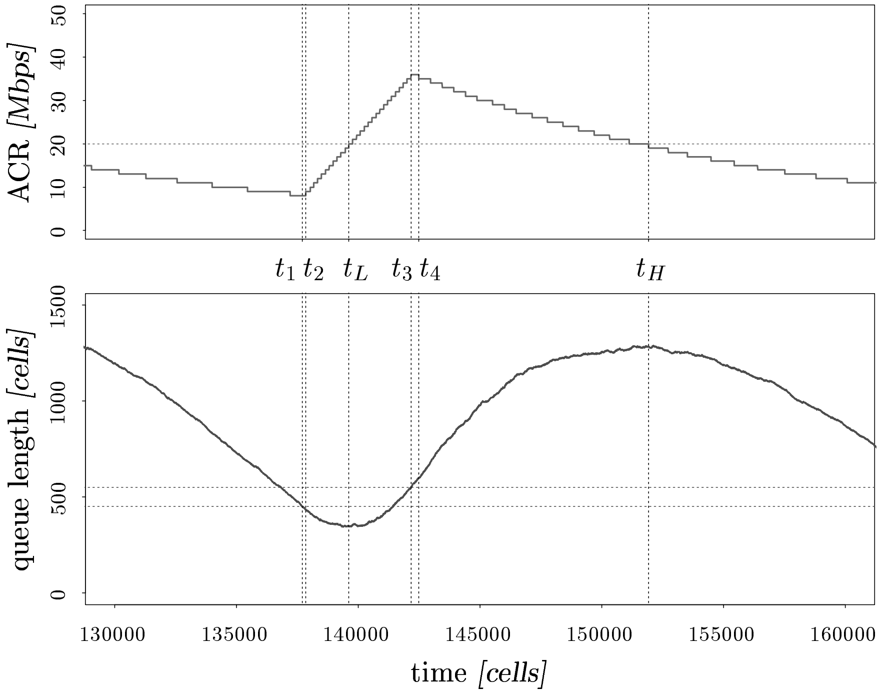


Figure 11: *Example evolution for the binary mode of the ABR flow control*

To apply our queueing model, we look at the cell arrival rate which is valid directly at the switch instead of the ACR of the source. The choice of this observation point has no influence on the results for the performance measures we compute and was also used in [17, 18]. Consequently, the fixed signaling delay τ consists of the round trip propagation delay of the ABR connection we look at.

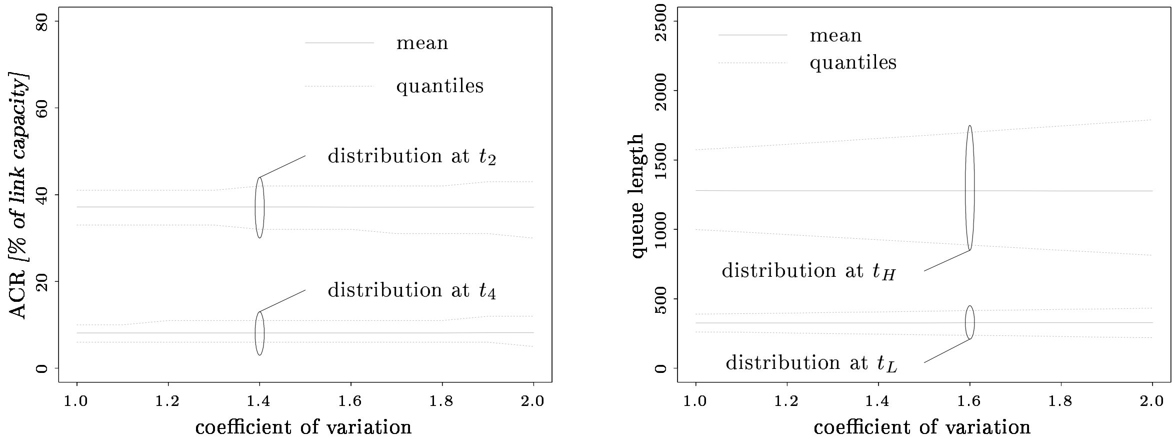
For the performance analysis we use the following parameter set. The fixed delay is equal to $\tau = 80$ slots or cell times in the ATM context, which corresponds to a distance of approximately 100 *km* for a 34 *Mbps* ATM link. Since the RM cells are typically sent every $N_{rm} = 32$ data cells transmitted, we use a inter-signaling interval of $N = 32$ slots. This agrees well with the findings derived from Figure 8. The thresholds for congestion monitoring are set to $Q_L = 450$ and $Q_H = 550$ and one fifth of the link capacity is available to the ABR connection, i.e. $T = 5$.

To map the ACR to the source index defined for our model we make the following assumption. Each source index $k = 1, \ldots, 100$ represents an ACR equal to k percent of the link capacity. Due to the much higher granularity of the ACR, the analysis always considers the k which fits best to the current value. When receiving positive feedback information, i.e. no congestion was detected, the ACR is increased by one percent of the link capacity. In case of negative feedback we reduce the ACR by one percent of its current value.

Here we concentrate on the performance of the binary signaling mode with respect to the variability of the cell generation process. Dependencies on other system parameters, such as the feedback delay and the control parameters, have already been studied in [17] and other papers. Of course, the authors assumed saturated sources, but similar results can be expected.

Therefore, we use a negative-binomial distribution with a mean equal to $100/k$ for each inter-arrival distribution $c_k(\cdot)$ and vary the coefficient of variation. To avoid batch arrivals, the distributions $c_k(\cdot)$ are shifted by one. This type of arrival process is not directly compatible with the ABR concept, where the cell transmission rate is limited according to the ACR. It allows, however, a clear representation of the dependencies. For dimensioning purposes, other configurations can be used.

Figure 12 shows the mean and the quantiles of the ACR and the queue length distribution measured at the local minimum and maximum within a control cycle.



(a) *Cell rate distribution*

(b) *Queue length distribution*

Figure 12: *Robustness of the binary signaling mode for ABR connections*

We observe the same dependencies as for the control law studied in the last section, cf. Figure 10. The mean values are almost insensitive to the coefficient of variation and the variability of the local measures linearly increases with the variation. Thus, the exponential decrease of the transmission rate during congested periods has only a minor influence on the robustness of the feedback control.

5 Conclusions

In this paper we presented an approximate analysis of a discrete-time queueing model with non-deterministic arrivals and binary feedback, which is applicable to investigate the binary mode of the ABR flow control. We derived distributions for the local minimum and maximum of the state of the arrival process, expressed by a source index, and the queue length within a control cycle. Numerical results have shown a high accuracy of our approach although a number of small approximation steps are involved.

From our parameter study we conclude that the variability of the arrival process has only a minor influence on the variation of the local minimum and maximum of the source index and the queue length. Furthermore, the long-term mean of these extreme values is almost not affected. This holds in general for each control law, irrespective of the way the source index is increased and decreased.

Regarding the binary mode of the ABR flow control, we observe a robustness against the variation of the arrival process if the frequency of variations is high compared to the frequency of the system state dynamics.

Acknowledgments

The author would like to thank Prof. P. Tran-Gia for the stimulating discussions during the course of work. The financial support of the Deutsche Telekom AG (Technologiezentrum Darmstadt) is appreciated.

References

[1] E. Altman. F. Baccelli and J.-C. Bolot. *Discrete-Time Analysis of Adaptive Rate Control Mechanisms.* High Speed Networks and Their Performance (C-21), H. G. Perros and Y. Viniotis (ed.), North-Holland 1994, pp. 121-140.

[2] The ATM Forum Technical Committee. *Traffic Management Specification.* Version 4.0, April 1996.

[3] C. Blondia and O. Casals. *Throughput Analysis of the Explicit Rate Congestion Control Mechanism in ATM Networks.* 10th ITC Specialists Seminar, Lund, 1996, pp. 89-101.

[4] J.-C. Bolot and A. U. Shankar. *Dynamical Behavior of Rate-Based Flow Control Mechanisms.* ACM SIGCOMM Computer Communication Review 20(2), 1990, pp. 35-49.

[5] F. Bonomi, D. Mitra and J. B. Seery. *Adaptive Algorithms for Feedback-Based Flow Control in High-Speed, Wide-Area ATM Networks.* IEEE Journal on Selected Areas in Communications 13(7), 1995, pp. 1267-1283.

[6] D. Cavendish, Y. Oie, M. Murata and H. Miyahara. *Proportional Rate-Based Congestion Control Under Long Propagation Delay.* International Journal of Communication Systems 8, 1995, pp. 79-89.

[7] A. I. Elwalid and D. Mitra. *Analysis and Design of Rate-Based Congestion Control of High Speed Networks, I: Stochastic Fluid Models, Access Regulation.* Queueing Systems 9, 1991, pp. 29-64.

[8] A. I. Elwalid. *Analysis of Rate-Based Congestion Control for High-Speed Wide-Area Networks.* IEEE ICC '95, Seattle, 1995, pp. 1948-1943.

[9] K. W. Fendick, M. A. Rodrigues and A. Weiss. *Analysis of a Rate-Based Control Strategy with Delayed Feedback.* IEEE Communications Magazine, October 1991, pp. 74-82.

[10] R. Izmailov. *Adaptive Feedback Control Algorithms for Large Data Transfers in High-Speed Networks.* IEEE Transactions on Automatic Control 40(8), 1995, pp. 1469-1471.

[11] K. Kawahara, Y. Oie, M. Murata and H. Miyahara. *Performance Analysis of Reactive Congestion Control for ATM Networks.* IEEE Journal on Selected Areas in Communications 13(4), 1995, pp. 651-661.

[12] S. Keshav. *A Control-Theoretic Approach to Flow Control.* ACM SIGCOMM Computer Communication Review 21(4), 1991, pp. 3-15.

[13] A. Mukherjee and J. C. Strikwerda. *Analysis of Dynamic Congestion Control Protocols – A Fokker-Plank Approximation.* ACM SIGCOMM Computer Communication Review 21(4), 1991, pp. 159-169.

[14] G. F. Newell. *Queues with Time-Dependent Arrival Rates: I — The Transition Through Saturation.* Journal of Applied Probability 5, 1968, pp. 436-451.

[15] H. Ohsaki, M. Murata, H. Suzuki, C. Ikeda and H. Miyahara. *Rate-Based Congestion Control for ATM Networks.* ACM SIGCOMM Computer Communication Review 25(2), 1995, pp. 60-72.

[16] R. S. Pazhyannur and R. Agrawal. *Feedback-Based Flow Control of B-ISDN/ATM Networks.* IEEE Journal on Selected Areas in Communications 13(7), 1995, pp. 1252-1266.

[17] M. Ritter. *Network Buffer Requirements of the Rate-Based Control Mechanism for ABR Services.* IEEE INFOCOM '96, San Francisco, 1996, pp. 1090-1097.

[18] M. Ritter. *The Effect of Bottleneck Service Rate Variations on the Performance of the ABR Flow Control.* To be presented at the IEEE INFOCOM '97, Kobe, 1997.

[19] D. Tipper and M. K. Sundareshan. *Numerical Methods for Modeling Computer Networks Under Nonstationary Conditions.* IEEE Journal on Selected Areas in Communications 8(9), 1990, pp. 1682-1695.

[20] N. Yin and M. G. Hluchyj. *On Closed-Loop Rate Control for ATM Cell Relay Networks.* IEEE INFOCOM '94, Toronto, 1994, pp. 99-108.