Invited contribution to the 8-th Australian Teletraffic Research Seminar, Melbourne, Dec. 1993.

# Discrete-time analysis technique and application to usage parameter control modelling in ATM systems

**P. Tran-Gia**
*Institute of Computer Science, University of Würzburg*
*Am Hubland, D-97074 Würzburg, Germany*
*Tel: +49-931-8885509, Fax: +49-931-8884601,*
*email: trangia@informatik.uni-wuerzburg.de*

**Abstract** *Most of modern communication systems are based on operation of fixed-size data packets like cells or slots. Due to this property discrete-time modelling techniques are increasingly appropriate for performance evaluation purposes. In the first part of this paper basic discrete-time analysis techniques are outlined, where the GI/GI/1 queue with bounded delay is taken as an example to illustrate the methods. The second part deals with discrete-time algorithms to analyse usage parameter control functions in ATM systems, where two submodels are discussed. The first submodel is used to investigate the traffic shaping function using a cell process spacer, where the characterization of cell traffic passing the spacer is taken into account. The second submodel deals with the dimensioning of cell delay variation tolerance, which appears in the generic cell rate algorithm used in the user-network interface of ATM networks.*

## 1. Discrete-time modelling and analysis technique

### 1.1 General

An increasing number of modern computer networks and communication systems operate with fixed-size packets like cells in asynchronous transfer mode (ATM) systems or slots in distributed queue dual bus (DQDB) subnetworks. Due to the cell-based nature of these system concepts, discrete-time queueing models are the natural choice and offer various advantages, compared to conventional methods. For the analysis of discrete-time models methods operating in discrete-time and discrete probability environments are needed. These methods can be subdivided into two main subclasses: i) analysis methods dealing directly with probability distributions in time domain and ii) analyses in transform domain.

The main purpose of this paper is to illustrate the use of discrete-time analysis techniques, considering models dealing with usage parameter control (UPC) functions in ATM systems at the user-network interface (UNI). In these analyses, performance measures of interest

are usually the cell blocking probability (e.g. in the range of $10^{-5}$ to $10^{-9}$) or the waiting time distribution function with the percentiles. Thus, simulation methods normally reach their limits very fast due to excessive computing time and the need of appropriate analysis techniques is obvious.

This paper is organized as follows. In this section basic discrete-time analysis techniques are outlined, where the GI/GI/1 queue with bounded delay is taken as example. This model is the underlying model for the performance evaluation of the submodels in the next section. Section 2 presents performance analyses of UPC functions in ATM systems, where two submodels are discussed. The first submodel is used to investigate the traffic shaping function using a generic cell process spacer, where the characterization of cell traffic passing the spacer is taken into account. The second submodel deals with the dimensioning of cell delay variation tolerance, which appears in the generic cell rate algorithm used in the user-network interface of ATM networks.

## 1.2 Discrete-time GI/GI/1 queue

Discrete-time models appear more frequently in performance evaluation of modern communication systems and play an increasingly important role. On the one hand, new system structures and principles often employ discrete or discretized basic time and data units. As discussed above, examples are the concept of cells in ATM networks or time slots in high-speed local and metropolitan area networks (e.g. DQDB). On the other hand, system parameters and input values are often based on measured data, which are given in the form of histograms. They are discrete-time by nature. These facts lead to the development of discrete-time models in performance analyses in the recent literature.

For the analysis of this class of models, conventional methods operating in continuous time are obviously inappropriate. Due to the lack of discrete-time methods they are used in some cases in an approximate sense. In these studies equivalent continuous-time model components are employed, e.g. the discrete-time stochastic arrival and service processes are approximately described by means of random variables with well-known time-continuous types of distribution functions. A small number of studies [1, 2, 21, 23, 24, 29, 32, 33, 34, 37] deals with direct analysis approaches for discrete-time models. Most of these studies take into account the discrete-time analysis of basic queueing models like single server systems [1, 2, 23, 24, 32] or queueing networks [29]. Some other studies present discrete-time analysis of general polling systems, overload control models in communication switching systems [33], routing mechanisms [37] or multiplexing schemes in modern communication system architectures [34]. A comprehensive survey can be found in [21].

### 1.2.1 Algorithm for the GI/GI/1 queue in time domain

We consider here the basic discrete-time GI/GI/1 queueing system. The term discrete-time indicates here that the time axis is slotted in equidistant time units of length $\Delta t$. The service time $B$ is generally distributed and the arrival process is a general stochastic process characterized by a generally distributed interarrival time $A$. Arriving customers

finding a busy server have to join an infinite capacity queue. Waiting customers in the queue will be treated according to a first-in, first-out (FIFO) service discipline.

The main performance measures of interests are here the distribution functions of the waiting time and the inter-departure interval. We should refer to studies appearing in the literature dealing with the calculation of the waiting time distribution function of the GI/GI/1 queue [1, 22, 23, 24, 25, 31, 32]. Most of these methods are related to solutions of the Lindley integral equation [25], which is a special form of Wiener-Hopf equations. Most of studies consider methods operating in Laplace domain. They are based on techniques like spectral factorization, numerical estimation of poles and roots of the system function, determination of quadratic factor of polynomials, as well as separation of functions having convolutions in frequency domain. Ackroyd [1] presented an efficient algorithm for the calculation of the waiting time distribution of the discrete-time GI/GI/1 queue, where discrete transform techniques (e.g. the Cepstrum concept [27], phase unwrapping technique etc.) and fast convolution algorithms are used. Using the same approach, a discrete-time analysis of the idle time and the inter-departure process is given by Tran-Gia in [32, 35].

As mentioned above, the random variables (RVs) are of discrete-time nature, i.e. the time axis is divided into intervals of unit length $\Delta t$. As a consequence, samples of those random variables are integer multiples of $\Delta t$; the time discretization is equidistant. We further assume that the discrete distributions have finite lengths. For a discrete-time random variable $X$, we use the notation $x(k) = \Pr\{X = k\}, -\infty < k < +\infty$, for the distribution (probability mass function) of $X$, $X(k) = \sum_{i=-\infty}^{k} x(i), -\infty < k < +\infty$, for the distribution function of $X$, $EX$ for the mean and $c_X$ for the coefficient of variation of $X$.

The basic relationship for the analysis of the discrete-time GI/GI/1 queue is equivalent to the well-known Lindley integral equation for the continuous-time system [22, 25, 31]. It will be outlined in the following.

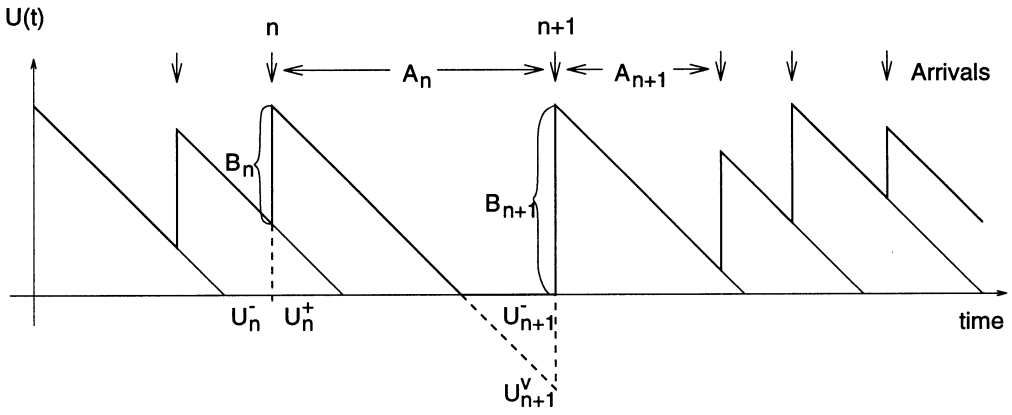


Figure 1: *Unfinished work process in GI/GI/1 model*

We observe a test customer (number $n$) which joins the system and sees upon its arrival an amount $U_n^-$ of unfinished work in the system (cf. Fig. 1). The service time of the test customer is denoted by the RV $B_n$ and the interarrival time, i.e. the interval until

the $(n+1)$-st arrival, by $A_n$. The corresponding distributions $a_n(k)$ and $b_n(k)$ exist for $k = 0,...,N_A - 1$ and $k = 0,..., N_B - 1$, respectively. The unfinished work $U_n^-$ indicates the amount of time units the server has to work out before being able to serve the test customer. Assuming the FIFO service discipline, its waiting time $W_n$ with the corresponding distribution

$$w_n(k) = \Pr\{W_n = k \quad \text{time units of length } \Delta t\} \tag{1}$$

are identical with $U_n^-$ and its distribution $u_n^-(k)$, respectively. The following recursive relationship between the waiting time distributions of two successive customers can be found [1, 24, 32]:

$$u_{n+1}^-(k) = \pi_0(u_n^-(k) \star a_n(-k) \star b_n(k)) = \pi_0(u_n^-(k) \star c_n(k)), \tag{2}$$

or

$$w_{n+1}(k) = \pi_0(w_n(k) \star c_n(k)) \tag{3}$$

where the term $c_n(k) = a_n(-k) \star b_n(k)$ is often called the *system function* of the GI/GI/1 system. The symbol "$\star$" denotes the discrete convolution operation:

$$a_3(k) = a_1(k) \star a_2(k) = \sum_{j=-\infty}^{+\infty} a_1(k-j) \cdot a_2(j) \tag{4}$$

and $\pi_0(.)$ the following operator:

$$\pi_0(x(k)) \quad = \quad \begin{cases} x(k) & k > 0 \\ \sum_{i=-\infty}^{0} x(i) & k = 0. \\ 0 & k < 0 \end{cases} \tag{5}$$

Considering the service and arrival processes to be recurrent and the observed state process of the discrete-time GI/GI/1 model to be in statistical equilibrium, the index $n$ of the test customer can be suppressed. We arrive at the stationary state equation for the analysis of the waiting time distribution of the GI/GI/1 system:

$$w(k) = \pi_0(w(k) \star c(k)). \tag{6}$$

Eqns. (3) and (6) represent discrete forms of the Lindley integral equation [25], which is well-known in the context of GI/GI/1 analysis. According to this equation the waiting time distribution of the $(n+1)$-st customer can be expressed as a function of the waiting time distribution of the $n$-th customer and the system function. Using this relationship (cf. [1, 32]) the equilibrium waiting time distribution can be iterativly calculated as schematically

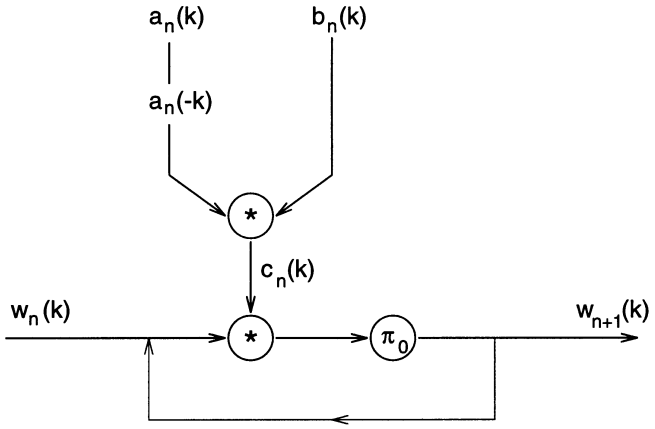


Figure 2: *Computational diagram of the algorithm in time domain*

depicted in Fig. 2. The iteration procedure may start e.g. by assuming the first customer finding an empty system.

For large vector sizes of the arrival and service time distributions, the discrete convolution operation can efficiently be implemented using discrete transforms and convolution algorithms, e.g. the fast Fourier transform (FFT) (based on the discrete Fourier transform (DFT)) [16, 27]. For the analysis of a GI/GI/1 system under stationary conditions, the number of iteration cycles needed and in accordance with it, the computing efforts depend strongly on the parameters of the system to be investigated. In comparison with algorithms in transform domain, e.g. the spectral factorization in Laplace- or Z-domain or the separation of maximum and minimum phase systems using the cepstrum concept (see next subsection), the algorithm in probability domain (or time domain) is very robust with respect to the type of interarrival and service processes. Furthermore, the algorithm in time domain as illustrated in Fig. 2 is also applicable to more general G/G/1 systems with time- or state-dependent interarrival and service time distributions, e.g. systems with workload-oriented overload control [33] or with alternating input processes [37]. The algorithm can also be modified to deal with GI/GI/1 queueing systems with bounded delay.

### 1.2.2 Algorithm for the GI/GI/1 queue in transform domain

From eqn. (6), which was given for the *waiting time distribution*, an analogous form of the waiting time *distribution function* defined by $W(k) = \sum_{i=-\infty}^{k} w(i)$ can be obtained:

$$W^{-}(k) + W(k) = c(k) \star W(k), \tag{7}$$

where $W^{-}(k)$ consists of components of the convolution $c(k) \star W(k)$ lying on the negative time axis. In Z-transform domain, we obtain the following fundamental equation:

$$W^{-}_{ZT}(z) \quad \cdot \quad \frac{1}{w_{ZT}(z)} \quad = \quad \frac{c_{ZT}(z) - 1}{1 - z^{-1}}, \tag{8}$$

where the transfer function of the GI/GI/1 system in Z-domain is contained:

$$S_{ZT}(z) \quad = \quad \frac{c_{ZT}(z) - 1}{1 - z^{-1}}. \tag{9}$$

It can be shown (cf. [1, 35]) that for distributions $a(k)$ and $b(k)$ with finite length, the function $W_{ZT}^{-}(z)$ stands for the Z-transform of a maximal phase system (cf. [27]). Furthermore, the term $\frac{1}{w_{ZT}(z)}$ corresponds to the Z-transform of a minimal phase system. This knowledge leads to solutions of eqn. (8) using pole and root allocation schemes [24] or in conjunction with the use of the Cepstrum concept [1]. The application of the Cepstrum to the waiting time analysis of GI/GI/1 queueing model will be discussed below.

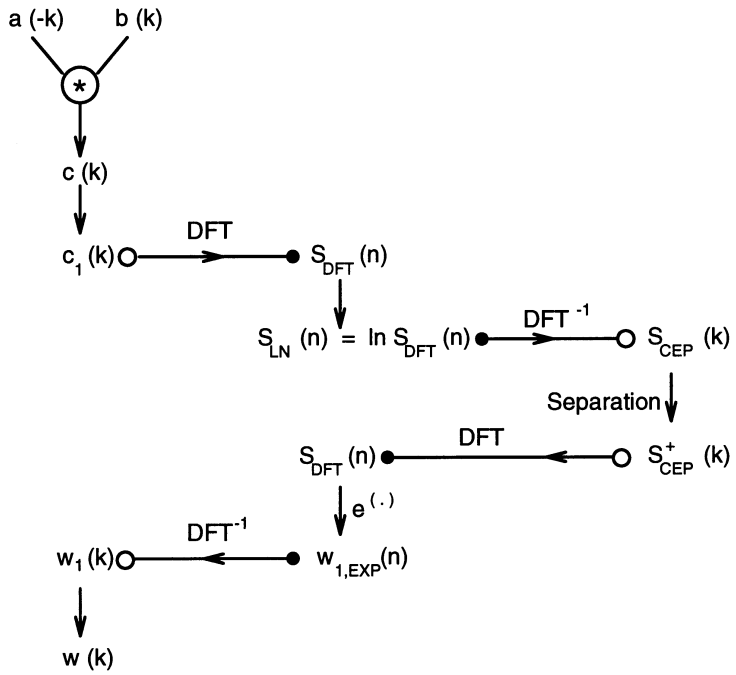


Figure 3: *The Cepstrum algorithm for GI/GI/1 analysis*

In the same way as in signal processing techniques, the Cepstrum concept is employed here to separate maximal and minimal phase systems [27]. Thus, the term $\frac{1}{w_{ZT}(z)}$ or consecutively, the waiting time distribution $w(k)$ in time domain, can be filtered out of the transfer function $S_{ZT}(z)$ after transforming it into the Cepstrum domain. The algorithm is illustrated in Fig. 3 containing the following major steps (cf. Ackroyd [1]):

1. Calculation of the transfer function $S_{ZT}(z)$ out of the system function $c(k) = a(-k) \star b(k)$. Since $c(k)$ is of finite length, $S_{ZT}(z)$ can be equivalently represented by the DFT $S_{DFT}(n)$.

2. Calculation of the complex Cepstrum

$$S_{CEP}(k) = DFT^{-1}\{\ln[S_{DFT}(n)]\}. \tag{10}$$

3. Separation of $S^{+}_{CEP}(k)$, which consists of non-negative components of $S_{CEP}(k)$ . The function $S^{+}_{CEP}(k)$ is the Cepstrum of the unnormalized waiting time distribution $w_1(k)$.

4. Inverse transformation of $S^{+}_{CEP}(k)$ to get $w_1(k)$ and normalization of $w_1(k)$ to obtain finally the waiting time distribution w(k) of the GI/GI/1 system.

Out of the waiting time distribution, further performance measures of interest, like the idle time of the server and the interdeparture distribution can be derived [32].

## 1.3 The GI/GI/1 queue with bounded delay

We consider in this subsection the case of the GI/GI/1 system with bounded delay, i.e. the waiting time of customers is limited to a maximum value of $L$. Customers who arrive and would have to wait longer than a threshold value, say $L-1$, are rejected.

This modification of the basic GI/GI/1 system has been used in modelling of overload control strategies in switching systems [34] and in backpressure mechanisms in reservation-based access mechanisms [36]. In this paper, this model will be used for performance evaluation of UPC functions in ATM systems like the spacing device and the peak cell-rate monitoring algorithm.

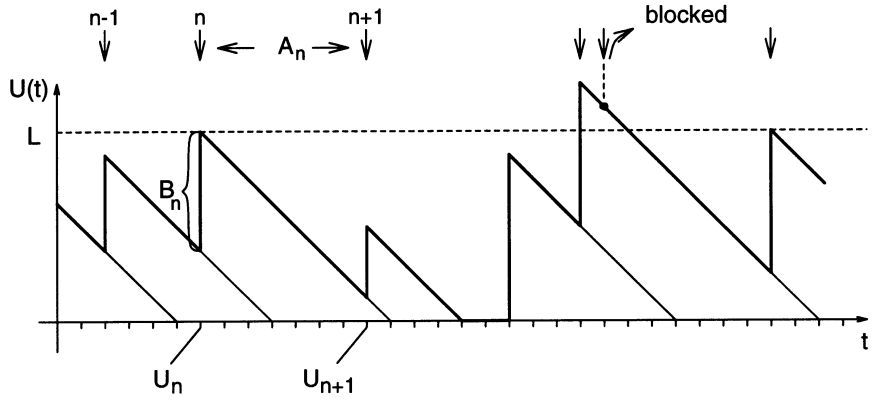


Figure 4: *Sample path of GI/GI/1 queue with bounded delay*

Following the derivation in [34] the analysis of the GI/GI/1 queue will be outlined. We denote by $A_n$ the RV for the interarrival time between the $n$-th and the $(n+1)$-st customer, $B_n$ the RV for the service time of the $n$-th customer and again, $U_n$ the RV for the unfinished work in the system immediately prior to the $n$-th customer arrival.

A snapshot of the state process development in the system is shown in Fig. 4. Observing the $n$-th customer in the system and the condition for customer acceptance upon arrival instant, the following *conditional random variables* for the workload seen by an arriving customer are introduced:

$$U_{n,0} = U_n | U_n < L \ , \qquad U_{n,1} = U_n | U_n \geq L, \tag{11a}$$

$$U_{n+1,0} = U_{n+1}|U_n < L, \quad U_{n+1,1} = U_{n+1}|U_n \geq L. \tag{11b}$$

The distributions of these random variables, adjusted by normalization, are:

$$u_{n,0}(k) = \frac{\sigma^{L-1}[u_n(k)]}{\Pr\{U_n < L\}} = \frac{\sigma^{L-1}[u_n(k)]}{\sum_{i=0}^{L-1} u_n(i)} \tag{12a}$$

$$u_{n,1}(k) = \frac{\sigma_L[u_n(k)]}{\Pr\{U_n \geq L\}} = \frac{\sigma_L[u_n(k)]}{\sum_{i=L}^{\infty} u_n(i)} \tag{12b}$$

where $\sigma^m(.)$ and $\sigma_m(.)$ are operators which truncate parts of a probability distribution function. The results of these operations are unnormalized distributions defined by:

$$\sigma^m[x(k)] = \begin{cases} x(k) & k \leq m \\ 0 & k > m \end{cases} \tag{13a}$$

$$\sigma_m[x(k)] = \begin{cases} 0 & k < m \\ x(k) & k \geq m \end{cases} \tag{13b}$$

Observing the development of the process (cf. Fig. 4) together with the maximum delay $L-1$, the following relationships between the RVs and their distributions are obtained:

1. $U_n < L$: customer acceptance

$$U_{n+1,0} = U_{n,0} + B_n - A_n \tag{14a}$$
$$u_{n+1,0}(k) = \pi_0[u_{n,0}(k) \star b_n(k) \star a_n(-k)]. \tag{14b}$$

2. $U_n \geq L$: customer rejection

$$U_{n+1,1} = U_{n,1} - A_n \tag{15a}$$
$$u_{n+1,1}(k) = \pi_0[u_{n,1}(k) \star a_n(-k)]. \tag{15b}$$

The distribution of the workload seen by the $(n+1)$-st customer is:

$$u_{n+1}(k) = \Pr\{U_n < L\} \cdot u_{n+1,0}(k) + \Pr\{U_n \geq L\} \cdot u_{n+1,1}(k). \tag{16}$$

From eqns. (14b), (15b) and (16), we finally arrive at a recursive relation to calculate the workload at arrival epochs of customers:

$$\begin{aligned} u_{n+1}(k) &= \pi_0[\sigma^{L-1}[u_n(k)] \star b_n(k) \star a_n(-k)] + \pi_0[\sigma_L[u_n(k)] \star a_n(-k)] \\ &= \pi_0[(\sigma^{L-1}[u_n(k)] \star b_n(k) + \sigma_L[u_n(k)]) \star a_n(-k)] \end{aligned} \tag{17}$$

Using this equation an algorithm to calculate the the workload prior to customer arrivals can be found. The algorithm can be used for both stationary and nonstationary traffic conditions. Under stationary conditions the index $n$ and $(n+1)$ in this equation can be suppressed. The computational diagram is depicted in Fig. 5.

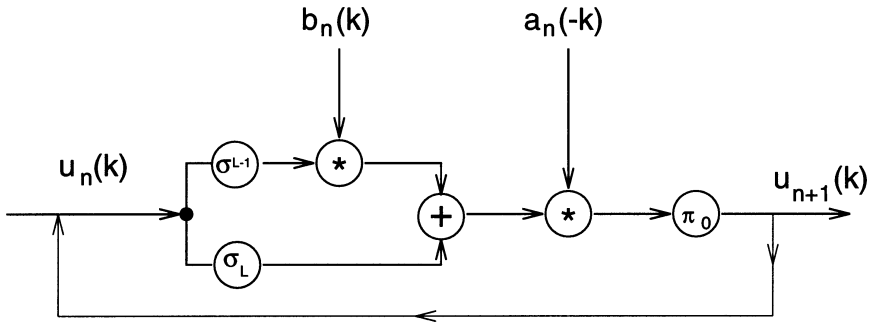


Figure 5: *Computational diagram for GI/GI/1 queue with bounded delay*

Finally, the customer rejection probability in statistical equilibrium is:

$$B = \sum_{i=L}^{\infty} u(i). \tag{18}$$

This performance measure will be used e.g. to compute the probability for a cell to be compliant due to the generic cell rate algorithm as discussed in the next section.

# 2. Discrete-time analysis of usage parameter control functions

## 2.1 User-network interface and usage parameter control

In the current standardization process of ATM networks the design issues concerning the user-network interface (UNI) in accordance with the usage parameter control (UPC) are crucial points. The main functions of the UPC/UNI illustrated in Fig. 6 are under discussion in standardization bodies (cf. CCITT recommendation [9]). We will briefly describe major UNI/UPC functions in the following with focus to modelling aspects.

Fig. 6 shows a number of virtual channel connections (VCC) multiplexed at the network edge. The multiplexed cell stream has to pass physical layer functions and is multiplexed further with operation and maintenance (OAM) cell traffic before entering the ATM network. Due to the connection admission control (CAC) a VCC is accepted or rejected. In the case of being accepted by the network, the VCC is thought of to have a contract with the network: the VCC user agrees to keep the negotiated traffic characteristics during the connection holding time, the network guarantees a predefined quality of service (QoS). Once a connection is established, policing is provided to guarantee the desired QoS for all connections according to their traffic contracts.

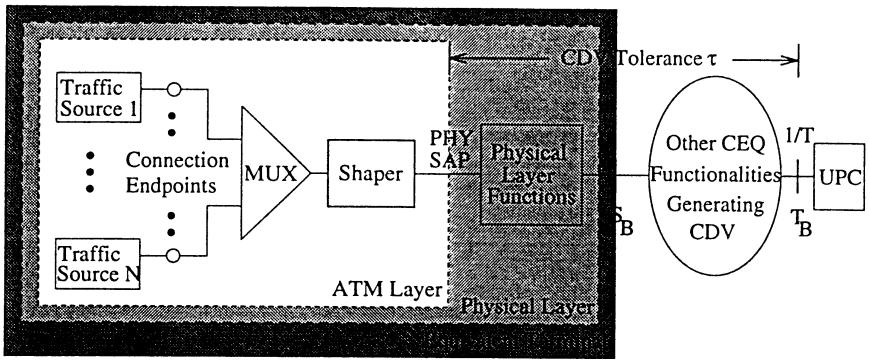


Figure 6: *Reference configuration from CCITT Draft Rec. I.371*

In the first phase of ATM system implementation and testing the most important connection parameter is the *peak cell rate*; this parameter is already standardized by various standardization bodies. The peak cell rate $R_i$ of the VCC $i$ is defined as the inverse of the minimum time $T_i$ between the emissions of two cells from this connection. Currently there are discussions about the introduction of a *sustainable cell rate* in the standards. A preliminary definition of this measure can be found e.g. in Draft Rec. I.371 of CCITT ([8, 9]).

We observe now the cell stream of a connection crossing the multiplexer to enter the ATM network, as shown in Fig. 7. The traffic process I generated in higher protocol layers has to pass a spacing device, which ensures that the peak cell rate $R_i$ of the VCC $i$ is not exceeded. This results in the cell process marked by II in Fig. 7. In [7, 32, 36] a *cell spacer* was suggested as traffic shaper. Another approach using a *spacing policer* was presented in [9]. Different virtual channel connections with different peak cell rates may be multiplexed. It should be noted that the arising spacer delay can affect the end-to-end performance drastically. This phenomenon will be the major subject of the performance model in section 2.2.

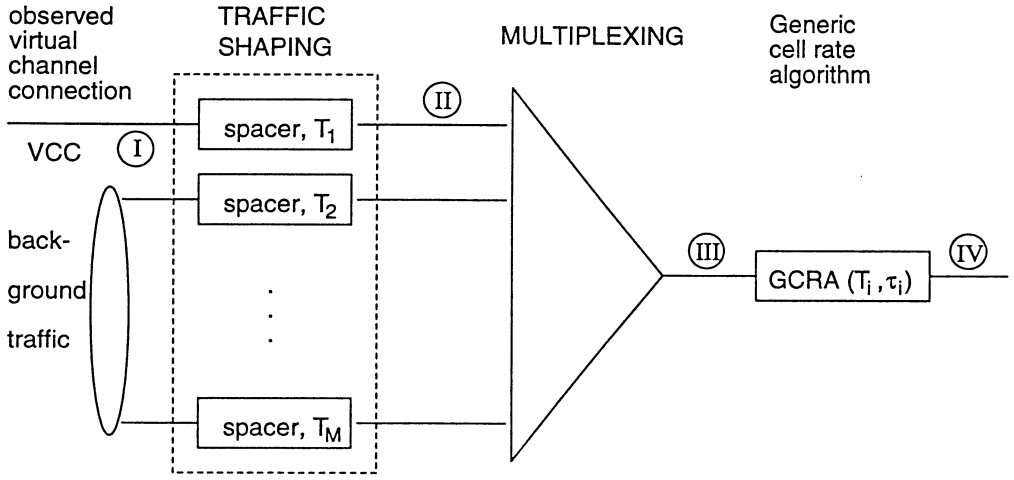


Figure 7: *Cell traffic spacing and cell delay variation*

According to the multiplexing function and to additional superposition of OAM cell traffic, a cell stream, which has been already spaced, can be again perturbed. At point III, the minimum distance $T_i^\star$ between cells which belong to a connection $i$ can be smaller than $T_i$ and thus the peak cell rate $R_i^\star$ at point III can momentarily be higher than $R_i$. This effect is often refered to as the *cell delay variation* (CDV), which can be observed in cell process modifying phenomena like cell clumping or cell dispersion, which strongly influence the cell process characteristics and the short-term rate of a VCC through the ATM network and thus, its quality of service provided by the ATM system [4]. Cell clumping can lead to short-term network congestions. Cell dispersion causes the necessity for larger buffer at the receiving site of the ATM network if the customer equipment has tight delay constraints (e.g. play-out buffer for voice/video). In a more general context, CDV can be introduced by different factors, e.g. i) multiplexing cells from different ATM connections, ii) UPC and *network parameter control* (NPC), iii) segmentation and reassembly functions in the ATM adaptation layer (AAL) and iv) other network and protocol functionalities.

To control the cell delay variation, the generic cell rate algorithm (GCRA) is introduced. By applying the GCRA to control the peak cell rate, it is also called the *virtual scheduling algorithm* (VSA). It can be shown that this algorithm is equivalent to a *continuous-state leaky bucket algorithm* (cf. [8]). The main function of the GCRA is to limit the CDV by controlling the cell stream and to declare accordingly cells during a heavy-traffic phase to be *compliant* or *non-compliant*. This is done connection-wise using two parameters: the minimal inter-cell interval $T_i$ and the CDV tolerance $\tau_i$. Thus, the algorithm is denoted by GCRA($T_i$,$\tau_i$). Details of the algorithm will be described in Section 2.3.1, where an analysis to calcutate the probability for a cell to be non-compliant is presented. The analysis also delivers the distribution function of the inter-cell process at point IV (cf. Fig. 7).

It should be noted that the location and the order of the three basic function blocks i) the spacer, ii) the multiplexer and iii) the GCRA as shown in Fig. 7 represent only one of various architectural possibilities, which are organized in accordance with the ATM switching system design. The GCRA can be performed e.g. at the same time with the multiplexing function.

Details about the operation of the cell spacer and the generic cell rate algorithm will be discussed later in this section.

## 2.2 Analysis of a cell spacer

The aim of the performance study in this section is an analytical treatment of a spacing device used for cell shaping functions. The cell process spacer is often discussed in recent literature [6, 7, 13, 15, 38]. The analysis is based on a discrete-time GI/D/1 queueing system with bounded delay. The cell arrival process which is subject to spacing can arbitrarily be chosen. The algorithm aims at the calculation of the spacer output process in term of the cell inter-departure time distribution, which gives insights to understand the traffic stream forming properties of the spacing mechanism. It should be noted that the spacer output process is in general non-renewal; its description using distributions will be done in a cumulative sense.

### 2.2.1 Traffic shaping using a spacer

As discussed above the peak cell rate $R_i$ of a virtual channel connection in ATM is defined as the inverse of the minimum cell inter-arrival time $T_i$. This parameter is part of the traffic contract between the VCC and the network. The basic function of a cell spacer is to keep the cells generated by a VCC to be at least $T_i$ apart. This is illustrated in Fig. 8.

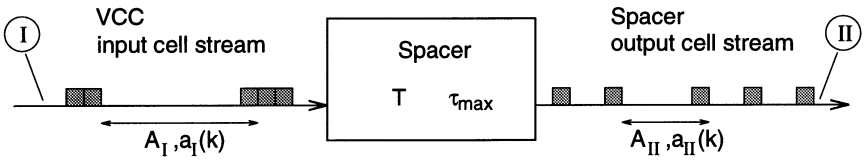


Figure 8: *Basic spacer model*

The time axis is discretized by the cell duration $\Delta t$. The cell process of the VCC is described by the arbitrary distributed random variable $A_I$ (cf. the mark I in Fig. 7 and Fig. 8) for the cell interarrival time. The spacer output process is represented by the discrete-time RV $A_{II}$ for the inter-departure time. The minimum inter-cell interval for the observed VCC is $T$. We consider further a maximum spacer delay $\tau_{max}$; cells which would have to wait in the spacer for longer than $\tau_{max}$ are rejected.

### 2.2.2 Spacer modelling with bounded delay GI/D/1

We introduce the random variable $U(t)$ for the time-dependent spacer state, which stands for the amount of unfinished work in the spacer at time $t$. According to the spacing scheme a cell which arrives at time $t_0$ and sees the spacer in the state $U(t_0) = k$ has to wait $k$ time units in the spacer before being issued.

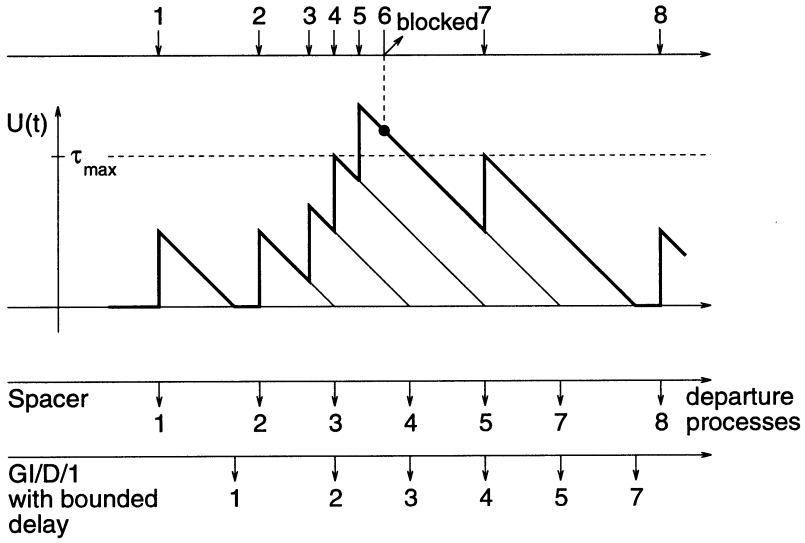


Figure 9: *Spacer state process and output stream*

A sample path of $U(t)$ is depicted in Fig. 9, where the spacer input process, the spacer output process and — for comparison purposes — the output process of a GI/D/1 queue with bounded delay are shown. Cells 1 and 2 find the spacer empty and pass it without delay. Cells 3, 4 and 5 arrive at the spacer and see a positive spacer state $U(t)$; they have to be spaced according to the value of $U(t)$ upon their arrivals. Cell number 6 would have to wait longer than $\tau_{max}$ and is thus rejected, since it is a non-compliant cell. It can be seen that an accepted cell increases $U(t)$ by an amount of $T$, which is thought of as the *virtual service time* of the spacing process.

As can be observed in Fig. 9, the output process of the spacer is *similar* to the output of an equivalent GI/D/1 queue, where a delay of $T$ is the only difference. This fact is also used in [38]. A spacer analysis using a $GI^{[X]}/D/1$ queue, which is similar to the discrete-time analysis described here, can be found in Hübner et al. [18].

The analysis used to deliver the results in this section is summarized as follows. For the case of a spacer with unlimited delay ($\tau_{max} \rightarrow \infty$) the equivalent model is a GI/D/1 queue, for which the analysis in both time and transform domains as discussed in the first section, can be used. Given the case of a spacer with bounded delay, the state analysis as presented in the previous section can be employed. It should be noted that for the calculation of the output process the use of the equivalent queue of type $GI^{[X]}/D/1$ is more advantagous.

### 2.2.3 Spacer dimensioning issues

To illustrate the influence of the spacer on the cell process we consider a spacer, which enforces the peak cell rate of a VCC with a minimum inter-cell interval of $T = 15$. The input process $A_I$ is assumed to be negative-binomially distributed.

Fig. 10 depicts how the spacer affects and forms the cell process for the case of $E[A_I] = 20$ and $c_{A_I} = 0.5$, where the distribution functions of the input and the output process of the spacer are shown. This is done for different values of maximum spacer delay $\tau_{max}$. The shape of the inter-departure distribution function indicates the truncation function performed by the spacer.

Fig. 11 illustrates the main function of the spacer, i.e. to smooth a bursty, high-variance input cell stream. To illustrate the smoothing effect quantitatively, we take the case of $E[A_I] = 6$ and draw the coefficient of variation $c_{A_I}$ of the inter-departure process as a function of the coefficient of variation of the input cell stream. As expected, $c_{A_{II}}$ is smaller than $c_{A_I}$ due to the cell spacing function. This effect is observed for all values of the maximum spacer delay.

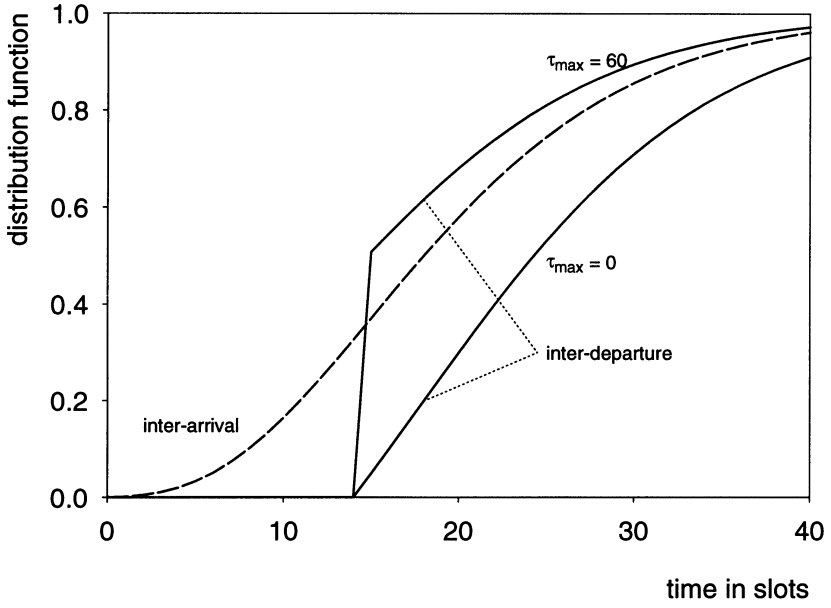


Figure 10: *Illustration of cell spacing function*

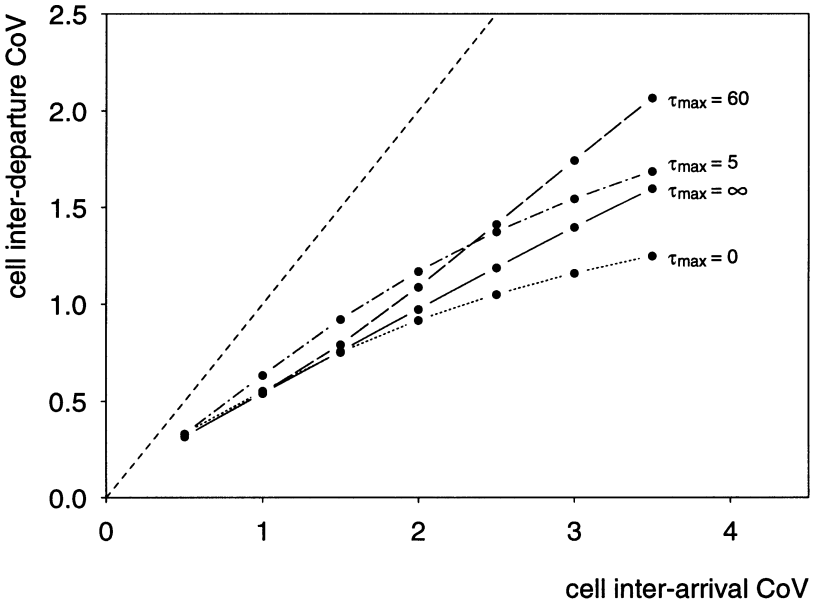


Figure 11: *Process smoothing by cell spacing*

## 2.3 Cell Delay Variation modelling

The aim of the performance study in this section is a discrete-time analysis of the generic cell rate algorithm, which is used to monitor the peak cell rate of a VCC in ATM environments. The analysis is again based on a discrete-time GI/D/1 queueing system with bounded delay. The cell arrival process to be controlled by the GCRA can be arbitrarily chosen. The purpose is the calculation of the probability $B$ for a cell to be non-compliant. Subsequently, dimensioning aspects of the cell delay variation tolerance in the UPC are studied.

### 2.3.1 The Generic Cell Rate Algorithm

As discussed above, CDV can be introduced by various factors like the multiplexing function, the UPC, the NPC or the segmentation and reassembly functions. To prevent the cell clumping and dispersion to cause network congestion, the GCRA is recommended [8], which has been first applied to monitor the peak cell rate.

The GCRA, as depicted in Fig. 7, monitors each VCC individually. We take into account the virtual channel connection $i$ and suppress the index $i$ in the notation GCRA($T_i$,$\tau_i$). Thus the parameters of the GCRA($T$,$\tau$) for the VCC $i$ are the target minimum inter-cell interval $T$ and the CDV tolerance $\tau$. The main problem is to scale $\tau$ for a VCC in such a way that the probability for a cell to be non-compliant (and rejected or tagged) is low while the CDV is reduced. At the same time, $\tau$ must be dimensioned to prevent overload caused by non-conforming VCC's.

The working mode of the GCRA($T$,$\tau$) is as follows. The algorithm determines whether a cell is generated too close to the last cell, indicating that the connection generates cells with a higher rate than the negotiated rate, or not. Distinction is made between the *theoretical arrival time* $TAT_j$ and the *actual arrival time* $t_j$ of a cell $j$. According to the characteristic of the cell process of the VCC $i$ (marked by III in Fig. 7), cells can be marked as *compliant* or *non-compliant*. The way to treat the next cell (number $(j+1)$) is as follows:

1. Estimate the theoretical arrival time $TAT_{j+1}$ of cell $j+1$

   a) $TAT_{j+1} = t_j + T$ if $t_j \geq TAT_j$:
   cell $j$ is generated after its expected arrival. It should be noted that the late generation of this cell does not allow for an earlier generation of the next cell.

   b) $TAT_{j+1} = TAT_j + T$ if $t_j < TAT_j$ and cell $j$ is compliant:
   cell $j$ is generated prior to its expected arrival but is still within the CDV tolerance. The TAT of the next cell is set as if cell $j$ had been generated at its TAT and not earlier.

   c) $TAT_{j+1} = TAT_j$ if $t_j < TAT_j$ and cell $j$ is non-compliant:
   cell $j$ is generated prior to its expected arrival and lies outside the tolerance. Cells which are identified as non-compliant can optionally be tagged or rejected (cf. [8]). It is assumed here that non-compliant cells are discarded. The TAT for the next cell is not modified in this case.

2. The next cell $j+1$ arrives at $t_{j+1}$. It will be considered as

   a) compliant if $t_{j+1} \geq TAT_{j+1} - \tau$

   b) non-compliant if $t_{j+1} < TAT_{j+1} - \tau$.

The algorithm guarantees that cells from a VCC enter the ATM network at the $T_B$ reference point (cf. Fig. 6) with a long term rate of at most $1/T$. It should be noted that the tolerance $\tau$ could be chosen to be larger than the target minimum inter-cell time $T$. Furthermore, due to the CDV tolerance $\tau$, the smallest inter-cell interval of the cell process at point IV in Fig. 7 is tolerated by the GCRA to be shorter than the (theoretical) target minimum inter-cell interval $T$. The peak cell rate $1/T$ could momentarily be exceeded and deliberately violated.

### 2.3.2 Discrete-time model for the Generic Cell Rate Algorithm

The discrete-time random variable $U(t)$ for the time-dependent GCRA state is introduced, which represents the remaining time to the next theoretical arrival time. The value of $U(t)$ can be thought of as the *virtual unfinished work* according to the GCRA function. A cell which arrives at time $t_0$ and sees the GCRA in the state $U(t_0) = k$ will be considered as non-compliant for $k \geq \tau$ and compliant for $k < \tau$.

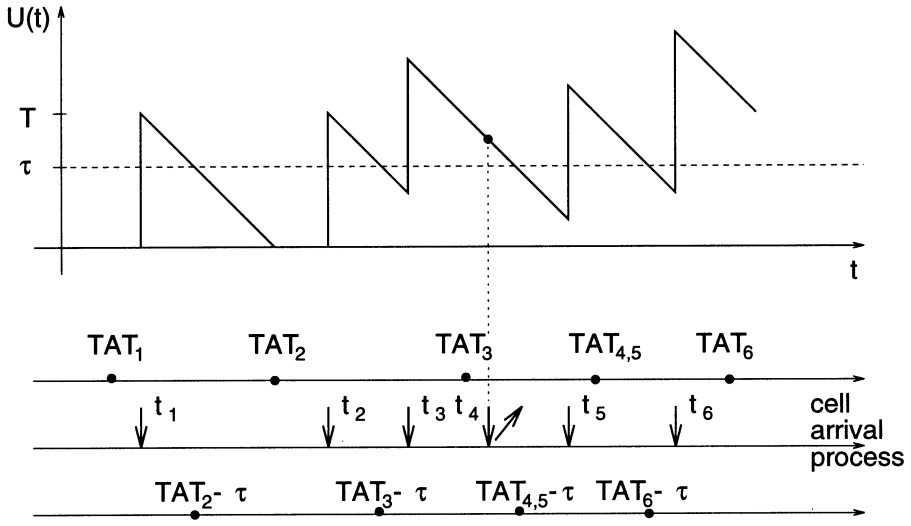


Figure 12: *Generic Cell Rate Algorithm state process*

A sample path of $U(t)$ is depicted in Fig. 12, where the theoretical arrival time $TAT_j$, the actual arrival time $t_j$ and the tolerated early arrival time (control measure) $TAT_j - \tau$ are shown. Cells 1 and 2 arrive after their TAT and are compliant. The actual arrival time of cell 3 lies before its TAT but is still within the tolerance (i.e. after $TAT_3 - \tau$); cell 3 is therefore compliant. Cell 4 arrives at the GCRA before the tolerance interval ($t_4 < TAT_4 - \tau$) and is non-compliant.

As can be observed in Fig. 12, the state process of the virtual unfinished work of the GCRA is identical to the unfinished work process of a GI/D/1 queue with bounded delay. A compliant (and accepted) cell increases $U(t)$ by an amount of $T$.

The modelling approach here was first presented by Hübner [17]. The assumptions made are quite general. The cell process of the VCC entering the GCRA function is described by the arbitrarily distributed random variable $A_{III}$ (cf. the mark III in Fig. 7) for the cell interarrival time. The GCRA output process is represented by the discrete-time RV $A_{IV}$ for the interdeparture time. In general, this process is non-renewal. Its distribution function given in this analysis is a cumulative one. To compute the probability for a cell to be non-compliant the blocking probability given in eqn. (18) can be taken.

### 2.3.3 Dimensioning issues of Generic Cell Rate Algorithm

In the following we take results from [17] to illustrate the use of the analysis. The traffic scenario depicted in Fig. 7 is considered, where the GCRA is used to monitor the cell process generated by a constant bit rate (CBR) source. The background traffic is characterized by a negative-binomially distributed inter-cell interval $X$. A hybride approach using both simulation and analysis is employed in [17], where the inter-cell distribution of the process $A_{III}$ for the CBR source after being multiplexed with the background traffic is obtained using simulation.

Unless noticed otherwise, CBR source traffic is assumed to have higher priority by the multiplexer as background traffic. In the numerical results the inter-cell time of the observed CBR source is chosen at $T^\star = 24$. If the CBR source fulfills the traffic contract, its cell rejection probability $B$ should be below a defined quality of service, say $10^{-9}$, when the monitor parameters $T$ and $\tau$ are chosen appropriately. For this range of blocking probabilities, analytical methods are advantagous. Normally, without an elastic tolerance, the parameter $T$ in the $GCRA(T, \tau)$ is the same as $T^\star$. In [6] is was stated that in some cases $T$ must be chosen significantly smaller than $T^\star$ to be able to meet the performance requirements. In Table 1, taken from [17], minimum values of $\tau$ are listed for different choices of $T$ to keep $B$ below $10^{-9}$.

| $T =$ | 23 | 22 | 21 | 20 |
|---|---|---|---|---|
| $E[X] = 4,\ c_X = 1,\ \tau \geq$ | 8 | 5 | 3 | 1 |
| $E[X] = 4,\ c_X = 2,\ \tau \geq$ | 37 | 29 | 26 | 23 |
| $E[X] = 2,\ c_X = 1,\ \tau \geq$ | 18 | 14 | 11 | 9 |
| $E[X] = 2,\ c_X = 2,\ \tau \geq$ | 107 | 65 | 50 | 42 |

Table 1: *Dependency of minimum $\tau$ on $T$ (for $B < 10^{-9}$)*

It can be clearly seen that $\tau$ depends not only on the background traffic intensity, as already observed in [15], but is also strongly influenced by the coefficient of variation $c_X$. This has to be incorporated in the dimensioning of the CDV tolerance as illustrated in

Fig. 13, where the influence of varying $\tau$ and $c_X$ on the probability $B$ for CBR cells to be non-compliant, with $T^\star = 24$, $E[X] = 2$, and $T = 20$, is shown.

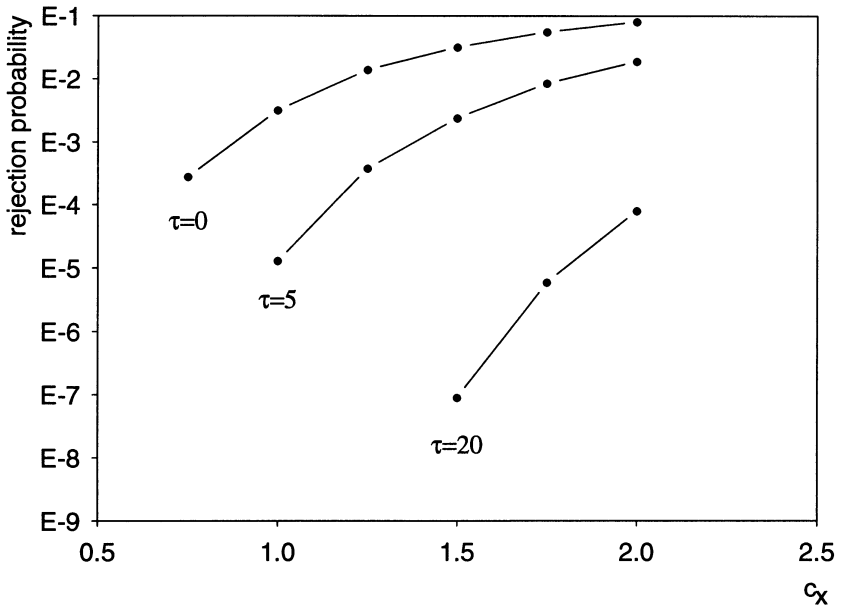


Figure 13: *Dimensioning of CDV tolerance*

It can be observed in Fig. 13 that the cell blocking probability can be remarkably reduced by increasing the tolerance value $\tau$. However, a choice of larger $\tau$ would also enlarge the burst length with higher cell rate than the target peak cell rate in the VCC contract.

If the amount of non-compliant cells is too large although the traffic contract is fulfilled by the VCC user (e.g. due to high CDV caused by multiplexing stages), there are different possibilities to scale the parameters of the GCRA function to monitor the agreed cell blocking probability $B$. The first possibility is to dimension $T < T^\star$, which means the introduction of an additional elastic tolerance. This does not seem to be able to be incorporated in the standards. The second possibility is to increase the CDV tolerance $\tau$. This must be done in a careful way. If $\tau$ is too large, the number of back-to-back cells which are still recognized as compliant cells increases. As a consequence, the cell clumping effect arises, which can lead to network congestion. Furthermore, for a large $\tau$, cells from a traffic source which violates the traffic contract can not be appropriately controlled.

## 3. Conclusion and outlook

The purpose of this paper is to present methods of discrete-time modelling and analysis, which gain importance due to the development of modern computer networks and communication systems operating with fixed-size packets like cells in ATM systems or slots in DQDB subnetworks.

The analysis technique and related algorithms have been presented by means of the treatment of the discrete-time GI/GI/1 queue and GI/GI/1 queue with bounded delay, where algorithms dealing directly with probability distributions in time domain and analysis techniques in transform domain have been outlined.

Subsequently, discrete-time performance analyses of usage parameter control functions in ATM systems have been presented. Submodels for the traffic shaping function using cell process spacer and for the dimensioning of cell delay variation tolerance in conjunction with the generic cell rate algorithm used in the user-network interface of ATM networks have been discussed.

It has been shown that discrete-time models are well-suited for this class of models. In particular, the discrete-time analysis technique delivers not only mean values but also the entire distribution functions of the cell traffic processes. Furthermore, in modelling cases where very small blocking probabilities are subjects of the performance investigation, discrete-time analysis is advantagous compared to simulations.

**Acknowledgement** The author would like to thank Frank Hübner for stimulating discussions and for the numerical results. Helps and careful reviews provided by Sandra Heilmann, Michael Ritter, Alexander Schömig and Kurt Tutschku are appreciated.

## References

[1] Ackroyd M.H., "Computing the waiting time distribution for the G/G/1 queue by signal processing methods", IEEE COM-28(1980) 52-58.

[2] Ackroyd M.H., "Iterative computation of the M/G/1 queue length distribution via the Discrete Fourier Transform", IEEE COM-28(1980) 1929-1932.

[3] ATM Forum, "Traffic Control and Congestion Control", Draft Baseline Document, April 1993.

[4] Bernabei F., Gratta L., Listanti M., Sarghini A., "Analysis of ON-OFF Source Shaping for ATM Multiplexing", INFOCOM 1993, pp. 11a.3.

[5] Bogert B.P., Healy M.J.R., Tukey J.W., "The Quefrency Alanysis of Time Series for Echoes: Cepstrum, Pseudo-Autocovariance, Cross-Cepstrum, and Saphe Cracking", Proc. Symp. Time Series Analysis, Ed.: M. Rosenblatt, Wiley 1963, 209-243.

[6] Boyer P., Guillemin F.M., Servel M.J., Coudreuse J.-P., "Spacing Cells Protects and Enhances Utilization of ATM Network Links", IEEE Network Vol. 6(1992) No.5.

[7] Brochin F.M., "A Cell Spacing Device for Congestion Control in ATM Networks", Performance Evaluation, Vol.16, No.1-3, November 1992, 107-127.

[8] CCITT Draft Recommendation I.371, "Traffic Control and Congestion Control in B-ISDN", June 1992.

[9] CCITT Study Group XVIII Contribution D.2373, "A Proposal for a Definition of a Sustainable Cell Rate Traffic Descriptor", January 1993.

[10] COST 224 Final Report, "Performance evaluation and design of multiservice networks", J.W. Roberts (editor), Commission of the European Communities (EUR 14152), October 1991.

[11] Daley D.J., "Notes on Queueing Output Processes", Math. Methods in Queueing Theory, Springer, 1974, 351-358.

[12] Gravey A., Boyer P., "Cell Delay Variation Specification in ATM Networks", Proc. IFIP Workshop TC6, Modelling and Performance Evaluation of ATM Technology, La Martinique, January 1993.

[13] Guillemin F.M., Boyer P., Romoeuf L., "The Spacer-Controller: Architecture and First Assessments", Workshop on Broadband Communications, Estoril, Portugal, January 1992, 313-323.

[14] Guillemin F.M., Monin W., "Limitation of Cell Delay Variation in ATM Networks", ICCT, Beijing, China, September 1992.

[15] Guillemin F.M., Monin W., "Management of Cell Delay Variation in ATM Networks", GLOBECOM 1992, 128-132.

[16] Henrici P., "Fast Fourier Methods in Computational Complex Analysis", Siam Review, 21(1979) 481-527.

[17] Hübner F., "Dimensioning of a Peak Cell Rate Monitor Algorithm Using Discrete-Time Analysis", University of Würzburg, Institute of Computer Science Research Report Series, Report No.59, March 1993, to appear in Proc. 14-th International Teletraffic Congress, Antibes Juan-les-pins, France, June 1994.

[18] Hübner F., Tran-Gia P., "A Discrete-time Analysis of Cell Spacing in ATM Systems", University of Würzburg, Institute of Computer Science, Research Report Series, Report No. 66, June 1994.

[19] Hübner F., "Discrete-time Performance Analysis of finite-capacity Queueing Models for ATM Multiplexers", University of Würzburg, Institute of Computer Science, Ph.D. Dissertation, 1993.

[20] Hluchyi, M. G., Yin, N., "On the Queueing Behavior of Multiplexed Leaky Bucket Regulated Sources", INFOCOM 1993, paper 6a.3.

[21] Hunter J.J., "Mathematical Techniques of Applied Probability", Vol.1: Discrete Time Models: Basic Theory, Academic Press, 1983.

[22] Kingman J.F.C., "Inequalities in the Theory of Queues", J. Roy. Stat. Soc. B32(1970) 102-110.

[23] Kobayashi H., "Stochastic Modelling: Queueing Models; Discrete-Time Queueing Systems", in : Part II, Louchard G., Latouche G. (eds.), "Probability Theory and Computer Science", Academic Press 1983.

[24] Konheim A.G., "An Elementary Solution of the Queueing System GI/G/1", SIAM J. Comp., 4(1975) 540-545.

[25] Lindley D.V., "The Theory of Queues with a Single Server", Proc. of the Cambridge Philosophical Society, 48(1952) 277-289.

[26] Meisling T., "Discrete-Time Queueing Theory", Operations Research 6(1958) 96-105.

[27] Oppenheim A.V., Schafer R.W., "Digital Signal Processing", Prentice Hall 1975.

[28] Pack C.D., "The Output of an M/D/1 Queue", Operations Research 23(1975)4, 750-760.

[29] Pujolle G., Claude J.P., Seret D., "A Discrete Queueing System with a Product Form Solution", Proc. Int. Seminar on Comp. Networking and Perf. Evaluation, pp. 3.4, Tokyo 1985.

[30] Reiss L.K., Merakos L.F., "Shaping of Virtual Path Traffic for ATM B-ISDN", INFOCOM 1993, paper 2a.4.

[31] Smith W.L., "On the Distribution of Queueing Times", Proc. of the Cambridge Philosophical Society, 49(1953) 449-461.

[32] Tran-Gia P., "Discrete-time analysis for the interdeparture distribution of GI/G/1 queues", Proc. Semin. Teletraffic Analysis and Comp. Perform. Eval., Amsterdam, The Netherlands, 1986.

[33] Tran-Gia P., "Analysis of a load-driven overload control mechanism in discrete-time domain", Proc. 12th International Teletraffic Congress, Torino 1988.

[34] Tran-Gia P., Ahmadi H., "Analysis of a Discrete-Time $G^{[x]}/D/1 - S$ Queueing System with Applications in Packet-Switching Systems", INFOCOM 1988, 861-870.

[35] Tran-Gia P., "Discrete-Time Analysis of Performance Models in Computer and Communication Systems", 46th Report on Studies in Congestion Theory, University of Stuttgart, 1988.

[36] Tran-Gia P., Dittmann R., "A discrete-time analysis of the cyclic reservation multiple access protocol", Performance Evaluation, Vol. 16, 1992, 185-200.

[37] Tran-Gia P., Rathgeb E., "Performance Analysis of Semidynamic Scheduling Strategies in Discrete-Time Domain", Proc. INFOCOM '87, San Francisco, March/April 1987, IEEE Computer Society Press 1987, 962-970.

[38] Wallmeier E., Worster T., "The Spacing Policer, an Algorithm for Efficient Peak Bit Rate Control in ATM Networks", Int. Switching Sympos. 14, October 1992, paper A5.5.