Telecommunication Systems 5(1996)1–24

Reminder on queueing theory for ATM networks

Paul J. Kuehn
Institute of Communications Switching and Data Technics (IND), University of Stuttgart, Seidenstraße 36, D-70174 Stuttgart, Germany
E-mail: kuehn@ind.uni-stuttgart.de

Queueing theory is a very useful means for performance prediction during the system design phase, for resource dimensioning and for planning of networks according to load and *quality of service* figures. In this paper, an overview is given about traffic models for ATM traffic sources, generic ATM traffic control models and performance evaluation methods.

1. Introduction

Queueing theory forms the classical framework for modelling stochastic service processes which have strong applications in telecommunication systems, computers and computer communication networks. The typical approach is as follows:

(1) modelling the statistical arrivals by a stochastic process,
(2) modelling the system behaviour by a service system consisting of generic components such as servers and queues,
(3) modelling the system operation by generic operational strategies such as queueing disciplines, resource allocation strategies, etc.,
(4) Performance analysis of the system state process, queueing processes and derived quality of service characteristics.

The results of this analysis can be applied to the dimensioning of system resources, such as number of server devices or buffer capacities, for comparison of alternative systems or operational strategies or for studies on the parametric dependence of quality of service quantities, robustness, fairness, and so forth.

Queueing theory, also called teletraffic theory in telecommunication applications, has provided an impressive set of solutions and methods during its 80 years of history. Many basic results are applied as part of resource dimensioning and network optimization processes. Traffic measurements and continuous monitoring have been used to validate these methods.

With the event of packet switching networks for computer communications and, in particular, ATM networks for integrated broadband services, many classical results did either not apply or had to be refined. The most striking reason behind this is the effect that traffic sources for packetized data flows reveal bursty and correlated statistical behaviour affecting system performance in a crucial way. Together with advanced traffic control procedures for ATM networks a set of new traffic problems has been identified which cannot be covered by conventional methods or solutions.

Analytic modelling of queueing systems quickly results in mathematical complexity which does not allow a rigorous solution due to more complex stochastic processes or system structures. For those cases, many approximate analytical methods have been developed. Alternatively, computer simulation methods are commonly used to approach complex models, to get results much quicker or to validate approximate analytic results. Versatile simulation program tools are available to support the application.

Simulation typically has one drawback of large compute time consumption. For ATM networks, in particular, another drawback has been identified which had much less importance before: accuracy for simulating rare events.

In traditional packet networks sophisticated protocol mechanisms were applied to protect against packet losses. Modern high speed networks use lightweight protocols either for reasons of buffer economy or for reasons of real time applications (such as voice or video) where repeated transmissions are not applicable. Lightweight protocols avoid retransmissions, the quality of service has to be achieved by proper traffic control and resource dimensioning. Cell losses in ATM networks must therefore be kept at a very low rate, typically less than 10^{-8}, which is quite difficult to study by simulations. Therefore, analytical queueing theory is the only feasable method of performance prediction and resource dimensioning.

The paper is organized as follows: In section 2 a survey is given on various traffic source models for ATM applications. Section 3 introduces the basic (generic) queueing models for ATM networks. In section 4 an overview is given on various analytical performance evaluation methods. The scope of this paper, however, is too small for an exhaustive presentation. For details it will be referred to the literature. The Appendix gives a collection of ATM traffic related terms which are based on a new edition of teletraffic terms [1].

2. ATM source traffic models

2.1. HIERARCHICAL ATM SOURCE TRAFFIC MODEL

In the asynchronous transfer mode (ATM) the information is packetized into small cells of 53 octets, where 5 octets are used as the header. Conventional fixed bandwidth (bitrate) services result into constant bitrate (CBR) traffic, i.e. a traffic stream where cells are generated by the source equidistantly.

The nature of ATM allows, however, for bitrate reduction by advanced source coding techniques which results into variable bitrate (VBR) traffic, where cells arrive sporadically but at a much lower rate. For VBR voice traffic, e.g., cells are only generated during the talkspurt phase. Video coding makes use of differential coding and prediction which allows to reduce the average cell rate by a factor of 50–100. The superposition of many VBR cell streams allows use of the effect of statistical multiplexing, which promises another gain on top of VBR coding compared to peak bitrate allocation.

To make use of these effects, the nature of these VBR traffic sources has to be captured. A standard approach, which has been adopted by the traffic community and by standardization bodies, is the *hierarchical ATM source traffic model*, see fig. 1, [2].

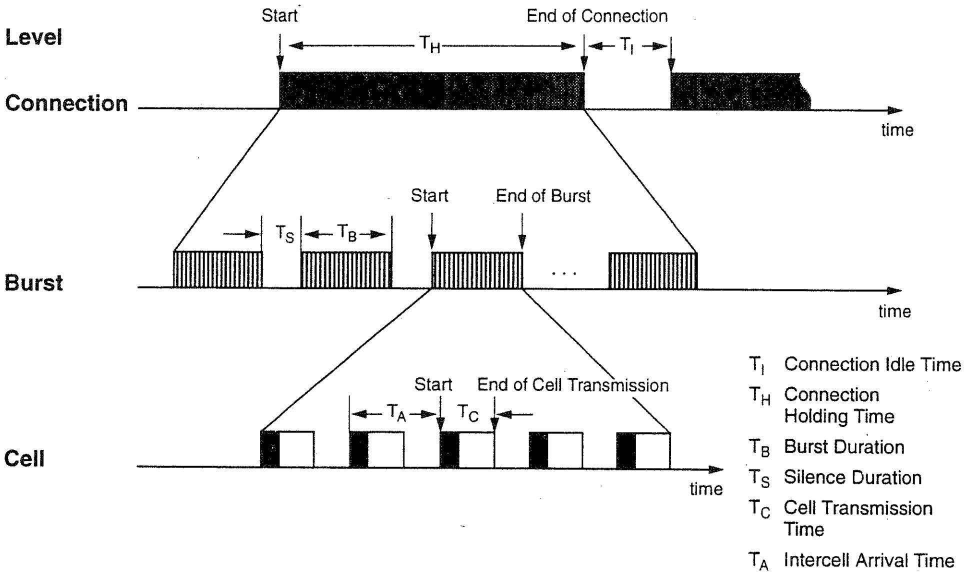


Fig. 1. Hierarchical ATM source traffic model.

In the basic model, three levels are distinguished: connection, burst and cell level. These levels may differ temporally by many orders of magnitude; in case of a voice communication a typical connection may last hundreds of seconds, bursts are in the order of hundreds of milliseconds, where cells may be transmitted over a 155 Mbit/s link within less than 3 microseconds. In general, even more levels could be identified: a dialogue level between burst and connection level and a call level above the connection level for a multimedia communication.

The hierarchical source model may be characterized by the statistics of the various durations, as indicated in fig. 1. For analytical treatment, however, it is not useful to cover all levels of a source; instead, one or two levels are captured by the source model, e.g., the burst and cell level, Analogously, the system behaviour is often considered at different levels (decomposition approach).

2.2. CONTINUOUS TIME SOURCE MODELS

In continuous time models, the slotted nature of constant cell transfer events is neglected, i.e., source models are described by point processes where no restriction exists about the location of the arrival events on the time axis.

(a) *Poisson process (PP)*

In the PP, interarrival times are independent of each other and negative exponentially distributed:

$$A(t) = P\{T_A \leq t\} = 1 - e^{-\lambda t}, \quad t \geq 0,$$

where T_A denotes the random interarrival time and λ the arrival rate. The PP is memoryless. It is adequate for the description of connection arrivals in a pool of many sources. Individual source behaviour on the cell level with bursty characteristic cannot be modelled by the PP; it may be used as the resulting process of the superposition of a large number of cell sources where none of the sources dominates.

(b) *Generally modulated Poisson process (GMPP)*

The GMPP is a doubly stochastic PP where the intensity function $\lambda(t)$ is controlled by another stochastic process with a finite state space $\{1, 2, \ldots, m\}$. The controlling process is also called generator or modulating process. Figure 2 illustrates the generator model for the GMPP and special cases derived from it. (*Note*: Sometimes the GMPP is also called switched Poisson process.)

(c) *Markov modulated Poisson process (MMPP)*

The MMPP is a special case of the GMPP where the modulating process is a finite Markov chain with infinite generator

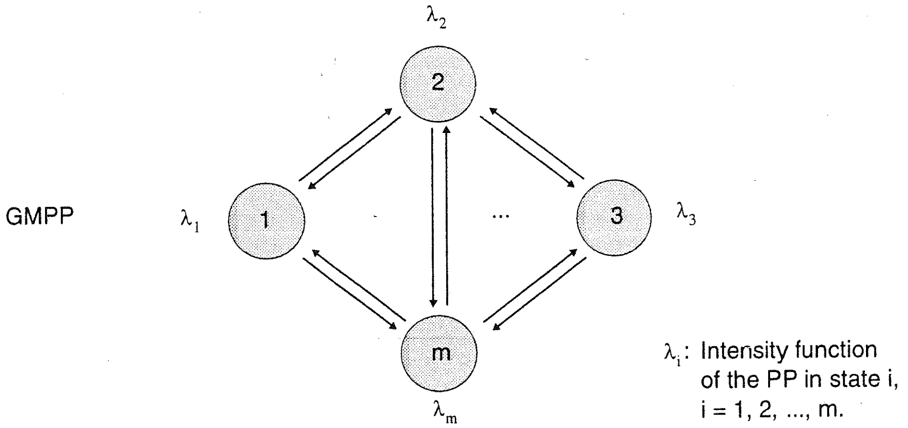
$$Q = (q_{ij}),$$

where q_{ij} is the transition rate for the transition of state i to state j, $j \neq i$ with $-q_{ii} = \Sigma_{j \neq i} q_{ij}$.

The intensity functions of the state-dependent Poisson processes are expressed by the diagonal matrix

$$\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_m).$$

The MMPP is no longer a renewal process; therefore it may be used to characterize traffic sources with some degree of memory [3].



Special Cases:
MMPP: GMPP with finite Markov Chain as modulating process
SPP: 2-State MMPP
IPP: SPP with $\lambda_1 > 0, \lambda_2 = 0$.

Fig. 2. Generally modulated Poisson process (GMPP).

(d) *Switched Poisson process (SPP)*

The SPP is a 2-state MMPP. The SPP is characterized by 4 parameters q_{12}, q_{21}, λ_1 and λ_2 from which the average arrival rate λ and higher order moments can easily be derived.

(e) *Interrupted Poisson process (IPP)*

The IPP is a special case of the SPP where $\lambda_1 \neq 0$ and $\lambda_2 = 0$ (or vice versa). The IPP imitates the bursty nature of a source where state 1 represents the ON-phase and state 2 the OFF-phase. Contrary to the SPP and the MMPP, the IPP is a renewal process; it can equivalently be represented by a hyperexponential inter-arrival distribution function of the order 2 (H_2-process).

(f) *Markovian arrival process (MAP)*

The MAP is a stochastic process which is based on a finite state Markov chain with m states. The transitions are governed by a transition matrix $Q = (q_{ij})$; at each transition an arrival occurs with a certain probability [4].

A generalization of the MAP is the MAP with batch arrivals (BMAP); this process adequately describes the arrival of multiple cells which may, e.g., arrive simultaneously at a buffer from different origins.

(g) *Autoregressive moving average model (ARMA)*

In the ARMA model, the number of cells in the nth interval is modelled by a discrete-state autoregressive moving average process X_i generated by the recursive relation

$$X_i = g(\alpha Z_{i-m} + Y_i + V_i),$$

where $\alpha < 1$ is a constant, Y_i and Z_i are sequences of correlated Gaussian random variables with zero means representing frame and scene correlations of video data streams, respectively, and where V_i is a sequence of uncorrelated Gaussian random variables with zero mean modelling white noise. The function g is a zero memory non-linear operator which transforms the Gaussian distributions to the required distribution $f(X)$. The ARMA model has been successfully used to represent video sources but it restricts itself from queueing analysis [5].

(h) *Fluid-flow approximation model (FFA)*

The Fluid-flow approximation replaces the discrete cell stream by a continuous flow of information where the flow intensity is modulated according to an. e.g., Markov chain or an ON/OFF-mechanism. This model focuses primarily on the capture of the burst nature of traffic sources. The neglection of the discrete cell level yields some advantages in the analysis [6].

(i) *Self-similarity source model (SSS)*

Studies on LAN and on video traffic sources have revealed a so-called self-similar (or fractal) nature, where the bursty behaviour is stretched out on an extremely wide temporal range [7]. In particular, the typical effect of randomization does not apply when sufficiently many bursty sources are superimposed and leads to a worse queueing behaviour.

To characterize the SS behaviour of a source, the aggregated process

$$X^{(m)} = \left(\frac{1}{m}[X_1 + X_2 + \ldots + X_m], \frac{1}{m}[X_{m+1} + X_{m+2} + \ldots + X_{2m}], \ldots\right),$$
$$m = 1, 2, 3, \ldots,$$

of the stochastic process $X = (X_1, X_2, \ldots)$ is considered. If the autocorrelation function $r^{(m)}(k) \to r(k)$ as $m \to \infty$, the stochastic process X is called asymptotically second-order self-similar; the aggregated process has a non-degenerate correlation structure as $m \to \infty$, whereas in the usual source models $r^{(m)}(k) \to 0$ as $m \to \infty$ for all $k \geq 1$. This long-term correlation structure is responsible for an increased cell loss probability in finite buffer multiplexers.

(j) *Renewal process (RP, GI)*

In the RP, also called the *general independent* point process in queueing theory, arrivals occur with interarrival times which are independent of each other and identically distributed (iid). Any of the typical repertoire of interarrival distribution functions may be applied which have to be matched to measured characteristics.

2.3. DISCRETE TIME SOURCE MODELS

Discrete time source models reflect the slotted nature of cell arrivals and departures.

(a) *Deterministic process (DP)*

In the DP, cells arrive equidistantly at the beginning (or at the end) of slot intervals $id\Delta t$, where $i = 0, 1, \ldots$ indicates the slot number, Δt the slot duration, and $d = 1, 2, \ldots$ the interarrival time between successive cell arrivals (as multiples of Δt). The slot time unit Δt can be chosen arbitrarily; in real applications Δt is integer-related with the cell transmission time.

(b) *Bernoulli process (BP)*

In the BP, in each slot interval a cell is generated with a constant probability q independently of previous slots. The BP is therefore memoryless. The interarrival time between two successive cells follows a geometric distribution with minimum of 1 slot Δt:

$$P\{T_A = i\Delta t\} = (1-q)^{i-1}q, \quad i = 1, 2, \ldots .$$

The BP describes the superposition of many individual cell arrival processes sufficiently well as long as none of the individual processes dominates and self-similarity effects are not present. It is often applied for ATM switching network analysis.

The BP can easily be extended to the batch Bernoulli process (BBP); in this case, in each slot interval k simultaneous cell arrivals occur with probability p_k, $k = 0, 1, \ldots$. The BBP results from the superposition of n individual and identical BPs; the number of arrivals at any slot has a binomial distribution $B(n, q)$ with mean nq.

(c) *Generally modulated deterministic process (GMDP)*

The GMDP has been proposed to describe VBR sources with piecewise DPs of different intensities. The GMDP is a doubly stochastic discrete time point process where the intensity is controlled by another stochastic process with a finite state space $\{1, 2, \ldots, m\}$. The controlling process is called generator or modulating process. The duration of the states of the modulating process follow a general discrete distribution

$$f_i(k) = P\{X_i = k\Delta t\}, \quad i = 1, 2, \ldots, m; \; k = 1, 2, \ldots \,.$$

The transition behaviour of the modulating process may be expressed by an $m \times m$ matrix $P = (p_{ij})$, where p_{ij} is the probability for a transition form state i to state j at the end of the sojourn time of state i. During state i cells are generated according to a DP with interarrival time $d_i\Delta t = (1/\lambda_i)$, $i = 1, 2, \ldots, m$ [8, 9].

Figure 3 gives an illustration of the GMDP with a sample realization.

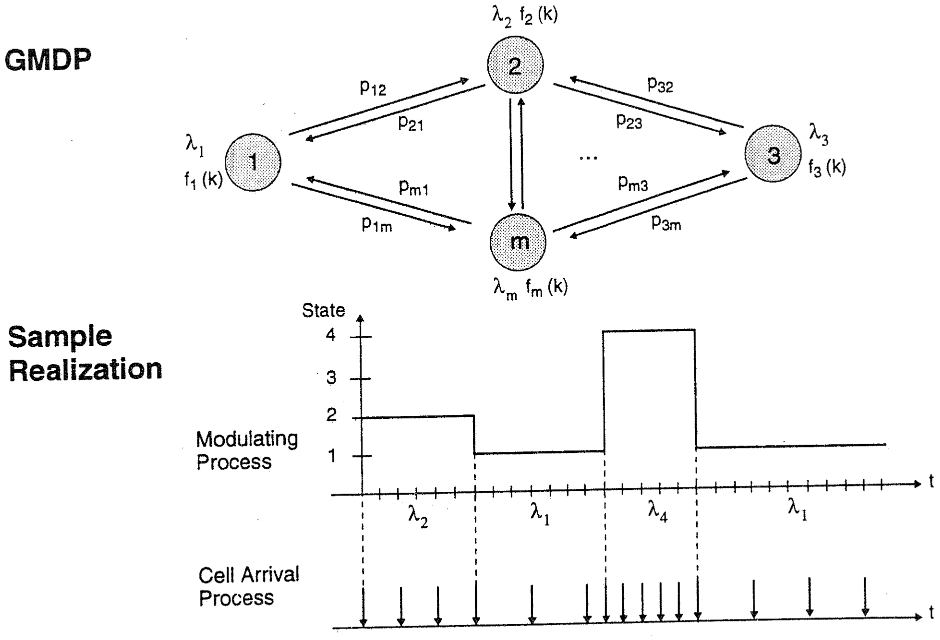


Fig. 3. Generally modulated deterministic process (GMDP).

(d) *Markov modulated deterministic process (MMDP)*

The MMDP is a special case of the GMDP where the modulating process is a finite discrete time Markov chain, i.e., the sojourn times of the states of the modulating process follow a geometric distribution.

(e) *Switched deterministic process (SDP)*

The SDP is a 2-state MMDP with cell arrival rates $\lambda_1 \neq 0$ and $\lambda_2 \neq 0$.

(f) *Interrupted deterministic process (IDP, On/Off-source)*

The IDP is a special case of the SDP where $\lambda_1 \neq 0$ and $\lambda_2 = 0$. In the literature, this source model is also called the On/Off-source, burst/silence model or talkspurt/silence model which is derived from its basic application to model packetized voice [10].

In the special cases of SDP and IDP the modulating process can be described by two parameters α and β, where α and β are the parameters of the geometric distributions of the On-state and Off-state, respectively:

$$\text{On-state:} \quad f_1(k) = (1-\alpha)^{k-1}\alpha,$$

$$\text{Off-state:} \quad f_2(k) = (1-\beta)^{k-1}\beta.$$

The average On-state duration is $\Delta t/\alpha$, the average Off-state duration is $\Delta t/\beta$. The modulating process can be interpreted such that the On-state continues with probability $(1-\alpha)$ and it changes to the Off-state with probability α after each slot. Similarly, the Off-state continues with probability $(1-\beta)$ and it changes to the On-state with probability β after each slot.

(g) *Discrete time Markovian arrival process (DMAP)*

The DMAP is a discrete time stochastic process which is based on a discrete time finite state Markov chain with m states. The transitions are governed by a transition matrix $P = (p_{ij})$, where the probabilities p_{ij} are composed by probabilities c_{ij} (no cell arrival) and d_{ij} (one cell arrival) [11].

Similarly, the DMAP can be generalized to the batch DMAP (D-BMAP) where at a transition batches of cells may be generated.

(h) *Discrete renewal process (DRP)*

In the DRP arrivals occur with interarrival times which are independent of each other and identically distributed. The interarrival times follow a general discrete distribution $f_A(i)$.

3. Generic queueing models for ATM

In this section, the most important basic queueing models for ATM networks are introduced. These models are "generic" in the sense, that many variants exist which, however, can be traced back to these basic models.

3.1. MULTIPLEXER MODEL

Multiplexing of various cell streams onto one outgoing link is the most common ATM queueing model where buffering is necessary to resolve the inherent congestion caused by simultaneously arriving cells. Figure 4 shows the basic *multiplexer model* where *n* cell streams (virtual channel connections) are multiplexed on one outgoing link which forms the single server. The buffer capacity is finite with s cell places. The arrival processes are general (i.e., anyone of the processes introduced in section 2 may be applied). The service times typically are constant as they represent the transmission time for a constantly sized cell on the link.

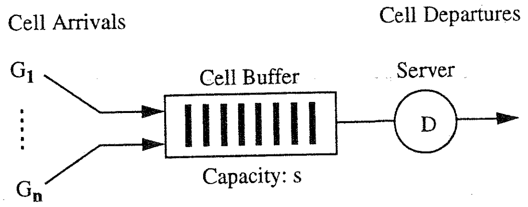


Fig. 4. Multiplexer model.

The engineering issue is to dimension the buffer capacity *s* (or, equivalently, to limit the number of cell streams) such, that a prescribed cell loss is not exceeded. The results of the queueing analysis aim at

(1) the individual cell stream loss probabilities,

(2) the cell delay variations,

incurred by the finite buffer.

As not all cells are equally important, variants of this model have been suggested by introduction of (spacial) priorities: cells may be distinguished with respect to their importance such that high priority cells experience a lower cell loss probability. Various methods can be applied:

- reordering of cells according to their spacial priorities,
- partial buffer sharing.

3.2. USAGE PARAMETER CONTROL (UPC)

To effectively control the quality of service the network operator has to make sure that all virtual path and virtual channel connections which have been set up behave according to the conditions under which they had been accepted. These conditions are defined at the instant of connection admission by providing a set of traffic parameters (the "traffic descriptor"). The UPC function is located at the user-network interface (UNI) and aims at monitoring individual VC or VP cell streams, checking the conformity between the monitored parameters and the negotiated parameters and taking appropriate actions in case of violation. The UPC function is also known as "source policing".

A similar function may be applied at the network node interface (NNI), which is called the network parameter control (NPC).

Actions which may be taken in the case of a "traffic contract"-violation range from tagging (setting the cell loss priority (CLP) bit in the cell header) to cell discarding in case of congestion. The problem is, however, to find an appropriate monitoring function which is robust against short term changes and which allows a safe detection of peak cell rate and mean cell rate violation. The peak cell rate is defined as the inverse of the minimum interarrival time between two successive cells. For the monitoring of the average cell rate the ATM Forum has introduced the notion of the "sustainable cell rate" which defines an upper value of the mean cell rate within a given burst tolerance.

Figure 5 shows the principal mechanism of the source policing function which is based on the "leaky bucket" principle: The LB is based on an up counting (by arriving cells) and a down counting (by a constant clock rate) mechanism. As long as the counter is within the counting range (the bucket size), cells are passed without any action. As soon as the counter crosses the bucket threshold, a proper action is taken (tagging or discarding). In case of discarding the cell traffic stream changes its statistical characteristic.

To achieve more sophisticated UPC mechanisms, such as the sustainable cell rate, variants of the LB mechanism, combinations of various LB-mechanisms or window based mechanisms can be applied.

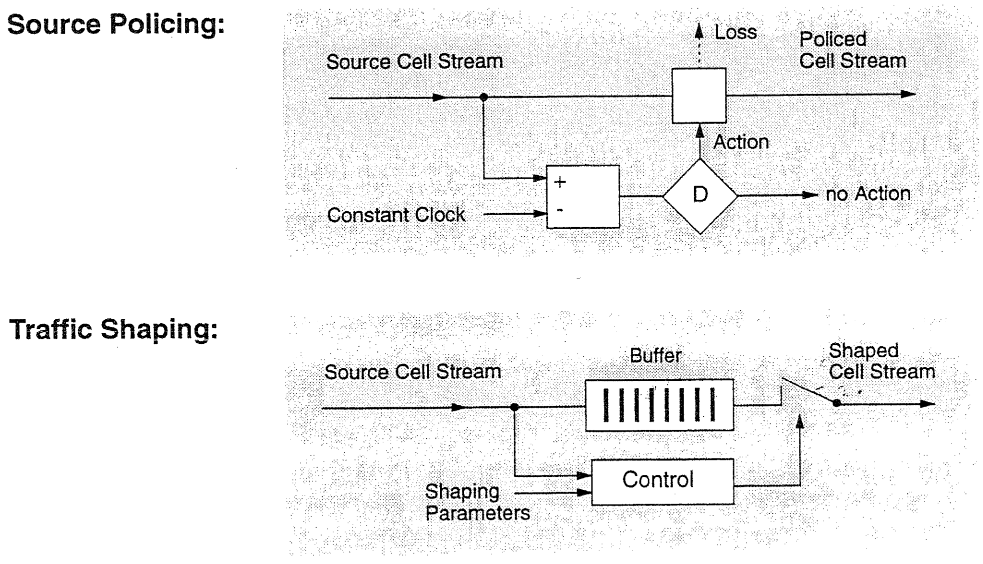


Fig. 5. Basic usage parameter control (UPC) and traffic shaping.

Traffic shaping is another method of traffic control acting on the cell stream of a VC, a VP or on aggregate cell streams. The basic traffic shaping mechanism is also illustrated in fig. 5. Contrary to UPC, traffic shaping delays cells within an intermediate buffer; the traffic shaping control releases the buffered cells in a more regular manner, by which the cell delay variation (CDV) is reduced which has a rather positive effect on cell losses and on the end-to-end cell delay variation.

3.3. CONNECTION ADMISSION CONTROL (CAC)

The CAC function decides whether a new VC or VP connection can be accepted under the given traffic descriptor, the QoS requirements and the already existing load. Typically, the CAC function has to be applied link-by-link from source to destination before a new VC or VP connection is admitted.

Figure 6 illustrates the basic functionality for an ATM link. The CAC algorithm decides on the traffic descriptor of the existing VC/VP connections, the new VC connection and the underlying QoS parameters. If the new loading complies with the QoS requirements of that link and all other ATM links within the path from source to destination the new VC connection may be admitted and entered into the control table; otherwise, the new VC connection will be rejected or a renegotiation process will be started.

The CAC algorithm is quite complex due to the difference between the

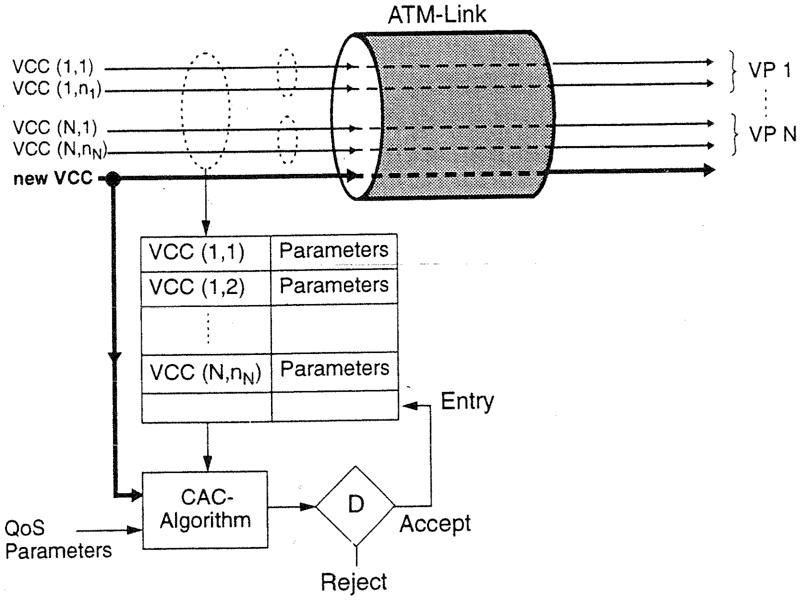


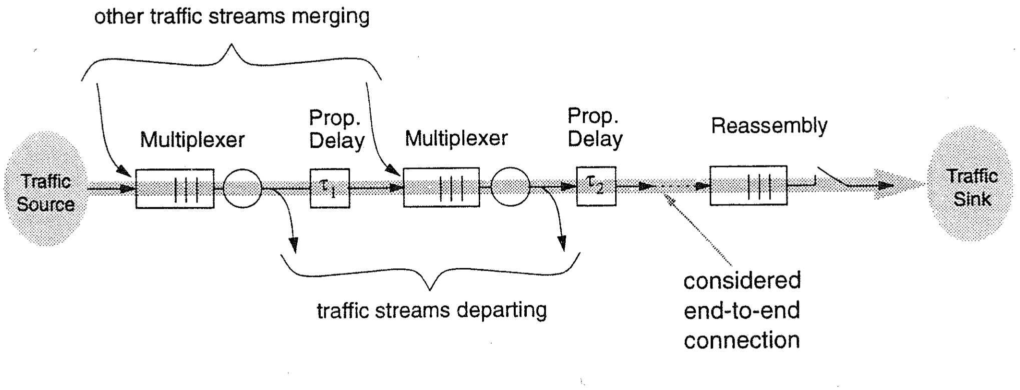
Fig. 6. Basic ATM models: Connection admission control (CAC).

indicated traffic descriptors and the real network load. Therefore, in reality also actual short term measurements should also be applied to the CAC algorithm.

Currently, various CAC algorithms are discussed. In the convolution method, the joint state distribution is calculated under the assumption that each VCC is independent of the others. From the convolved state distribution the blocking probability is derived on which the decision is based. This time consuming algorithm may be simplified by distributional assumptions, e.g. normal distributions, or by bounding considerations. Other CAC algorithms are based on the definition of "equivalent bandwidths" which are used as real constant bandwidths in an STM CAC environment. For the quick estimation of the current loading, neural networks have been suggested, too.

3.4. END-TO-END TRANSFER DELAY MODEL

For the estimation of end-to-end transfer delays for cells of a particular (reference) connection all delay components have to be considered along the connection path between the traffic source and the traffic sink. Figure 7 illustrates the principal delay components.



End-to-End Transfer Delay = Packet Assembly Delay (Traffic Source)
+ Transmission Delays (Links)
+ Queueing Delays (Buffers)
+ Propagation Delay
+ Reassembly Delay

Fig. 7. End-to-end cell transfer delay model.

The end-to-end transfer delay is composed of

(1) the packet assembly delay at the traffic source,

(2) the constant serial transmission delays on the various ATM links,

(3) the serial queueing delays incurred by buffering within the multiplexers,

(4) the (constant) propagation delays which are incurred by the physical travelling of waves along the transmission medium,

(5) the reassembly delay at the traffic sink.

The latter component of the reassembly delay may have different interpretations: In packet transfer applications, the reassembly delay stands for the time to collect and to reassemble the payload of all those cells which have originated from a packet by the segmentation function within the AAL of the sending side. For real time audio/video applications the reassembly buffer is used to compensate the cell delay jitter such that a continuous or synchronized delivery of user information can be provided at the destination.

For the analysis of the queueing delay components it is essential to model a realistic load case in each of the intermediate multiplexer models which results of the merging and departing traffic streams sharing the same link resources with cells of the considered reference connection.

The end-to-end transfer delay is also an essential part for the estimation of the scew which is defined as the difference between presentation times of two related object streams, e.g., a video and an audio stream within a multimedia application.

3.5. ATM SWITCHING NETWORKS (SNs)

There is a great variety of ATM switch architectures which have been designed and implemented. This variety is far too complex to be described in this context and will be referred to the literature [12–14]. Figure 8 illustrates three principal functional architectures which have been applied: shared media, shared buffer and space division. Real switching networks are built from such units, from combinations of them and by using of further principles such as, e.g., sorting networks or internal traffic distribution and resequencing at the output.

In shared medium SNs cell switching is achieved by time division multiplexing on a fast intermediate bus. The central medium (server) is allocated by the control. This principle has an inherent bottleneck, but it can economically be implemented for small switches. The model simplifies in the input part when the transmission capacity of the shared medium is at least as large as the aggregated capacity of all incoming links (nonblocking switch).

In shared memory SNs all arriving cells share one large buffer. This principle is most favourable but it causes severe implementation problems due to the limited memory bandwidth which requires a high degree of parallelism for writing and reading.

In space division SNs elementary switch devices, such as 2 × 2 switches, are used to temporally interconnect input and output buffers. Space division switches experience a special blocking effect (head-of-the-line blocking), when two cells of different input buffers are heading to the same output buffer. In multistage space

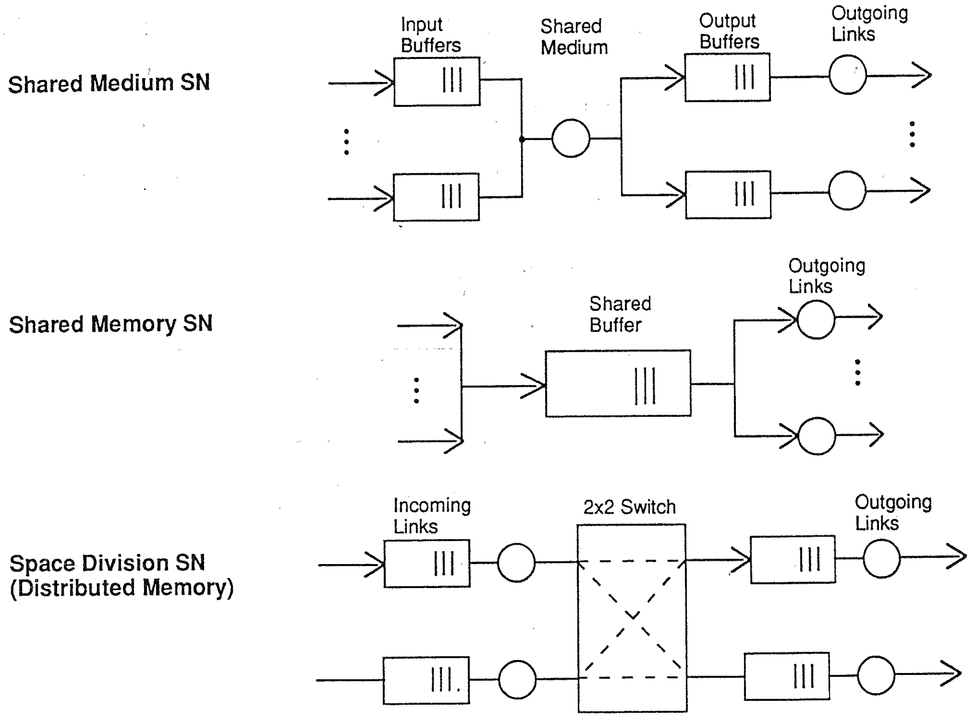


Fig. 8. Principal functional principles of ATM switching networks.

division SNs also back-pressure mechanisms may be used when the output buffers are fully occupied. The advantage of space division SNs is the possibility to build large multistage SNs from basic switch elements exploiting parallel operation and allowing scaling to arbitrary sizes.

3.6 OTHER MODELS

Besides the basic models of sections 3.1–3.6, there exist further models for particular aspects such as, e.g.,

- ATM adaptation layer (AAL) functions
  (packetization, segmentation, reassembly),
- Signalling functions
  (common virtual channel signalling),
- Resource management for multiservice applications
  (allocation/reservation strategies, routing, etc.),

which will not be addressed in this paper and where it is referred to the literature [15–19].

4. Performance analysis and evaluation methods

The performance evaluation requires the analysis of the stochastic processes within ATM networks or within particular subsystems. These stochastic processes depend on arrival processes for cell streams (workload), on the system structure, and on the system organization. Due to the complexity of the mathematical problem, only small system complexes can be solved rigorously; for larger system complexes, approximation methods are necessary.

In the following, a collection of typical performance analysis problems will be referred to. This collection is neither complete nor is it detailed.

4.1. PROCESS PARAMETER MATCHING

The parameters by which the various point processes of section 2 are defined are usually not known per se; instead, other parameters may be known either from measurements or as planning quantities. As an example, from a process the mean cell rate, the IDC (index of dispersion of counts) and the autocovariance function are known; the task is to select a proper point process and to match its parameters, e.g., for a MMDP. The usual way is to derive these quantities from the mathematical definition of the process and solving for the parameters. The mathematical difficulty lies in problems as overdetermination, nonlinear relationships or intractability.

4.2. TRAFFIC STREAM OPERATIONS

As ATM is a fully connectionoriented transfer mode, traffic streams of different connections are multiplexed and demultiplexed on common resources. Real multiplexing is always related to buffering, i.e. the traffic streams change their characteristics more or less through the random queueing delays. This can be taken care by considering output streams of multiplexer devices.

As long as the queueing delays are much smaller than interarrival times, an idealistic multiplexing can be considered as a superposition of various point processes. This problem is difficult in general, but for particular processes exact results are known. Examples are:

- superposition of PPs result in a PP,
- superposition of MMPPs result in a MMPP,
- superposition of BPs result in a BBP.

For more details, see [20, 10].

4.3. BASIC QUEUEING SYSTEM ANALYSIS

The principal queueing model of fig. 5 has many applications as cell multiplexing, source policing or as a component model within a multistage switching network. Therefore, this basic model has been studied under various assumptions. Following the COST group 224, a classification into pure cell scale, pure burst scale and combined burst and cell scale is useful [21]:

- Cell scale multiplexer models

$$\begin{array}{ccc} M/D/1 & BP/D/1 & (GEO/D/1) \\ GI/D/1 & DRP/D/1 & \\ \sum M_i/D/1 & \sum D_i/D/1 & \end{array}$$

- Burst Scale multiplexer models

  Fluid flow approximation multiplexer model

- Combined burst and cell scale multiplexer models

$$\begin{array}{rr} MMPP/D/1 & MMDP/D/1 \\ SPP/D/1 & SDP/D/1 \\ IPP/D/1 & IDP/D/1 \\ MAP/D/1 & DMAP/D/1 \\ BMAP/D/1 & D\text{-}BMAP/D/1 \end{array}$$

The single/multiple dimensional discrete system state processes are analysed by embedded Markov chain methods, semi-Markov processes or matrix geometry [19, 21]. In the pure burst scale analysis a continuous state X is introduced where the arrival intensity of the flow is modulated by On/Off-processes. For the continuous state variable a system of partial difference-differential equations can be derived [22, 23].

More sophisticated variants of the multiplexer model have been suggested for the analysis of spatial priority systems with partial buffer sharing or cell reordering [24].

4.4. QUEUEING NETWORK ANALYSIS

Network models are more complicated to solve. There is no hope of closed form solutions, such as product-form solutions in the case of BCMP queueing networks. Therefore, decomposition methods have to be applied.

Approved decomposition techniques as, e.g., suggested by the author [20] under the consistent approximation hypothesis of renewal processes, do not apply in the case of ATM networks due to the highly correlated traffic stream.

Individual output processes of multiple-connection or multiple service multiplexer models, such as $\Sigma M_i/D/1$, $\Sigma BP_i/D/1$ or $MMPP/D/1$ allow the estimation of the resource sharing influence on individual traffic streams [19].

End-to-end transfer delay analysis is another important issue to estimate the cell delay variations and skew of multi-service connections. Due to the highly correlated nature of the ATM cell streams, independence assumptions and delay convolutions cannot be applied. A successive aggregation method allows, however, for capturing of the basic serial delay effects [25].

5. Conclusion

In this paper we have tried to give an overview on models and approaches of queueing theory for ATM networks. There is a large spectrum of models, methods and standard solutions which could, in this context, only shortly be referenced.

For the future we feel that there are extensions necessary in three directions:

(1) traffic and model validations by experiments and measurements,
(2) tool support for standard problems,
(3) engineering practices,

to fully exploit teletraffic theory for resource dimensioning and QoS guarantees in future broadband networks.

Appendix. ATM traffic related terms

Traffic, teletraffic
Phenomenon of communication activities. In an ATM system, traffic is characterized by cells being transfered across a particular reference point.

Traffic parameter
Value describing the characteristic of the traffic. In an ATM system, typical traffic parameters are cell rates, burstiness, peak duration etc.

Traffic descriptor
Set of traffic parameters specifying the traffic characteristics of a connection at the time of connection set up.

Traffic source
Origination of traffic. Various traffic source types may be distinguished such as for voice, text, data, image or video.

Traffic source model
Mathematical model which describes the traffic source behaviour. In ATM systems, the traffic source model describes the arrival process of cells.

Hierarchical traffic source model
Traffic source model which explicitly takes into account various levels of activity. In

a typical ATM system, at least three levels of activity can be distinguished: connection, burst and cell level.

Cell rate
Ratio of cells transfered and unit of time.

Bit rate
Ratio of bits transfered and unit of time.

Constant bit rate traffic (CBR traffic)
Traffic, where the bit rate is constant during the whole connection.

Variable bit rate traffic (VBR traffic)
Traffic, where the bit rate changes during the connection. Changes may occur continuously or at discrete instants of time.

Peak cell rate
Inverse of the minimum interarrival time of two successive cells.

Average cell rate
Ratio of the number of cells transfered during a specified time t_0 and t_0.

Burst traffic
Traffic, where arrivals (cells in ATM systems) are clustered.

Burstiness
Ratio of peak cell rate and average cell rate.

Arrival process
Stochastic process which describes the random arrivals of events (connection set up requests, packets, cells).

Point process
Finite or infinite sequence of random events on the real (time) axis. The number $X(t)$ of events within the interval $[0, t]$ forms a stochastic process.

Leading function of a point process
The leading function $\Lambda(t)$ of a point process defines the average number of events (points) within the interval $[0, t]$.

Intensity function, rate
The intensity function $\lambda(t)$ is the derivative of $\Lambda(t)$ with respect to t.

Index of dispertion of counts (IDC)
Ratio of the variance and the mean of the number $X(t)$ of events within the interval $[0, t]$.

Renewal process
Point process whose interarrival times between successive events are independent of each other and identically distributed (iid).

Poisson process (PP)
Point process where the number $X(t)$ of events within the interval $[0, t]$ follows the Poisson distribution. The numbers of events within disjunct intervals are independent of each other.

Homogeneous Poisson process
Poisson process with a constant intensity function λ.

Generally modulated Poisson process (GMPP)
Doubly stochastic Poisson process where the intensity function is controlled by another stochastic process with a finite state space. The controlling process is also called generator process or modulating process.

Markov-modulated Poisson process (MMPP)
Special case of the GMPP where the modulating process is a continuous time Markov chain.

Switched Poisson process (SPP)
Special case of the MMPP with two states.

Interrupted Poisson process (IPP)
Special case of the SPP; within one of the states the intensity function is zero.

Markovian arrival process (MAP)
Arrival process whose arrivals are controlled by a continuous time Markov chain. Arrivals are only generated at the instances of state transitions where the number of arrivals depend on the particular state transition.

Discrete MAP (D-MAP)
Arrival process whose arrivals are controlled by a discrete time Markov chain. Arrivals are only generated at the instances of state transitions where the number of arrivals depends on the particular state transition.

Deterministic process (DP)
Point process where the interarrival times are constant.

Bernoulli process (BP)
Discrete time point process, where in each slot a cell arrives with a constant probability independent of the preceding slots.

Generally modulated deterministic process (GMDP)
Doubly stochastic deterministic process where the intensity function is controlled by another discrete time stochastic process with finite state space.

Markov modulated deterministic process (MMDP)
Special case of the GMDP where the modulating process is a discrete time Markov chain.

Switched deterministic process (SDP)
Special case of the MMDP with two states.

Interrupted deterministic process (IDP)
Special case of the SDP; within one of the states the intensity function is zero. This process is also known as the On/Off-source process.

Service process
A stochastic process which describes the characteristics of the service mechanism of the server. In most cases, the service process is described by the distribution function of service times.

Service system
Model which describes the occupations of resources (servers, queues). The service system includes the arrival proces(ses) of service requests (cells, connection set ups), the service proces(ses) and the system organization (such as service strategies, queueing disciplines, priority schedules, etc.).

Delay system, queueing system
Service system where all requests which cannot immediately be served may wait.

Loss system
Service system where all those requests which cannot immediately be served are rejected, i.e. they are lost.

Delay-loss system
Service system with a finite queue capacity. Requests are only lost when all available queue positions are occupied.

Queueing network
Network of service systems, in particular delay systems.

Multiplexer model
Basic ATM service system where several cell arrival streams are multiplexed on (i.e. interleaved in time) and transmitted by one outgoing channel. The standard multiplexer model for ATM cell streams is a delay-loss system with one server and a deterministic service process.

End-to-end connection model
Serial queueing network which describes the connection path from source to destination including all interfering traffic streams.

Traffic control
Control mechanism for a specific control level. Typical control mechanisms are: usage parameter control (UPC), connection admission control (CAC), routing control, end-to-end delay control, network parameter control (NPC), priority/congestion control, spacing, cell loss control, etc.

Traffic control model
Service system with specific functions of the particular control mechanism.

Quality of service (QoS)
Set of all characteristic values of a communication network with respect to requests of a service user. QoS values are observed by the service user.

Network performance (NP)
Set of quantitative values of a communication network or part of a communication network with respect to the traffic handling capabilities. NP values may be defined between two identical reference points (measurement points) of the physical, ATM or AAL layer.

Grade of service (GoS)
Part of the NP which depends on the dimensioning of resources.

ATM cell loss probability
Probability for the loss of a particular cell.

Cell loss ratio (CLR)
Fraction of cells of a particular connection which are lost within a specified interval of time. The CLR value may be referred to an ATM multiplexer, to an ATM node or to an end-to-end virtual channel connection.

Cell transfer delay
Elapsed time of a cell between two reference points in the network, i.e., the time between the instant where the first bit of a cell leaves at the first observation point

and the instant where the last bit of that cell is received at the second observation point. The cell transfer delay is composed of the packetization, propagation, transmission, queueing and reassembly delays.

Cell delay variation (CDV, jitter)
Measure for the random delays of cells within a connection which are caused by buffering effects.
Note: There are several definitions in use for the CDV measure, such as:

- Variance of the transfer delay end-to-end.
- Complementary distribution functions of the transfer delay difference of successive cells.
- Complementary distribution function of the difference between the transfer delay of the cells and the average transfer delay.

Packetization delay
Time which is necessary to accumulate the required number of bits to form an ATM cell.
Note: This time depends on the AAL-type.

Propagation delay
Time which is necessary for a signal (e.g., a bit or a cell) to travel along the transmission medium. The propagation delay depends on the distance between source and destination and on the type of medium.

Transmission delay
Time which is necessary to transmit a cell on a particular medium. The transmission delay is the ratio between cell length (424 bits) and the transmission speed of the link.

Queueing delay
Aggregated time of a cell for intermediate buffering.

Reassembly delay
Time by which a cell is delayed after the arrival at the destination reassembly buffer before it is transfered to the application level. In a packet transfer application, the reassembly delay is the time to collect and to reassemble the payload of all cells which have been originated from a packet by the segmentation function. In a real time application, such as voice or video, the reassembly delay is used to compensate the cell delay jitter for a continuous or synchronized delivery of the user information.

Skew
Difference between the presentation times of two related object streams, e.g., the video and the audio stream in a multimedia application.

References

[1] *ITG-Recommendation 1.2-03 on Teletraffic Terms*, Draft (1995).

[2] F. Schoute, Simple decision rules for acceptance of mixed traffic streams, *Proc. ITC 12* (1988) paper 4.2 A.5.

[3] H. Heffes and D.M. Lucantoni, A Markov modulated characterization of packetized voice and data traffic and related statistical multiplexer performance, IEEE J. SAC 4(6) (1986) 856–867.

[4] D.M. Lucantoni, K.S. Meier-Hellstern and M.F. Neuts, A single server queue with server vacations and a class of non-renewal arrival processes, Univ. of Arizona, Systems and Ind. Engin. Dept., Working Paper 88-008 (1988).

[5] R. Grünenfelder, J. Cosmas, S. Manthorpe and A. Odinma-Okafor, Characterization of video codecs as autoregressive moving average processes and related queueing system performance, IEEE J. SAC 9(3) (1991) 284–293.

[6] P. Castelli, E. Cavallero and A. Tonietti, Policing and cell admission problems in ATM networks, *Proc. ITC 13* (1991) 847–852.

[7] W.E. Leland, M.S. Taqqu, W. Willinger and D.V. Wilson, On the self-similar nature of Ethernet traffic, IEEE/ACM Trans. Networking 2(1) (1994) 1–15.

[8] P.J. Kuehn, Input process model GMDP, RACE 1022 Document UST_123_003_CD_CC (1988).

[9] J.P. Cosmas et al., A review of voice, data and video traffic models for ATM, Euro. Trans. Telecom. 5(2) (1944) 11–26.

[10] K. Sriram and W. Whitt, Characterizing superposition arrival processes in packet multiplexers for voice and data, IEEE J. SAC 4(6) (1986) 833–846.

[11] C. Blondia, A discrete time Markovian arrival process, RACE 1022 Document PRLB_123_0015_CD_CC (1989).

[12] J.Y. Hui, *Switching and Traffic Theory for Integrated Broadband Networks* (Kluwer, London, 1991).

[13] R.O. Onvural, *Asynchronous transfer mode networks: Performance issues* (Artech House, Boston/London, 1994).

[14] R. Händel, M.N. Huber and S. Schröder, *ATM Networks, Concepts, Protocols, Applications* (Addison-Wesley, Wokingham, 1994).

[15] *Proc. Int. Teletraffic Congress (ITC)*.

[16] V.A. Bolotin, J.G. Kappel and P.J. Kuehn (eds.), *Teletraffic Analysis of ATM Systems*, Special Issue: IEEE J. SAC 9(3) (1991).

[17] P.J. Kühn, R. Lehnert and G. Gallassi (eds.), *Teletraffic Research for Broadband-ISDN in the RACE Programme*, Special Issue: Euro. Trans. Telecomm. 5(2) (1994).

[18] V.A. Bolotin, P.J. Kuehn, C.d. Pack and R.A. Skoog (eds.), *Common Channel Signalling Networks: Performance, Engineering, Protocols and Capacity Management*, Special Issue: IEEE J. SAC 12(3) (1994).

[19] H. Saito, *Teletraffic Technologies in ATM Networks* (Artech House, Boston/London, 1994).

[20] P.J. Kuehn, Approximate analysis of general queueing networks by decomposition, IEEE Trans. Commun. 27(1) (1979) 113–126.

[21] J. Roberts (ed.), COST 224, Performance evaluation and design of multiservice networks, Final Report, Commission of the Euro. Communities, EUR 14152 EN (1992).

[22] D. Anick, D. Mitra and M. Sondhi, Stochastic theory of a data handling system with multiple sources, BSTJ 61(8) (1982) 1871–1894.

[23] J. Roberts, Queueing models for variable bitrate traffic streams, AEÜ 47(5/6) (1993) 435–439.

[24] H. Kröner, G. Hébuterne, P. Boyer and A. Gravey, Priority management in ATM switching nodes, IEEE J. SAC 9(3) (1991) 418–427.

[25] H. Kröner, M. Eberspächer, T.H. Theimer, P.J. Kühn and U. Briem, Approximate analysis of the end-to-end delay in ATM networks, *Proc. IEEE INFOCOM '92*, Florence (1992) pp. 978–986.