Technical University of Budapest
Department of Telecommunications and Telematics

Evaluation of Quality of Service and Network Performance in ATM Networks

Sándor Molnár

Budapest, 1995

Sándor Molnár
Department of Telecommunications & Telematics
Technical University of Budapest,
Sztoczek u. 2., H-1111 Budapest, Hungary
Tel: + 36 1 463 3889
Fax: + 36 1 463 3107
molnar@bme-tel.ttt.bme.hu

Budapesti Műszaki Egyetem
Távközlési és Telematikai Tanszék

ATM hálózatok teljesítőképességének és minőségi jellemzőinek elemzése

Molnár Sándor

Ph.D. disszertáció

Tudományos vezető

Tatai Péter és Søren Blaabjerg

Budapest, 1995

Sándor Molnár
Department of Telecommunications & Telematics
Technical University of Budapest,
Sztoczek u. 2., H-1111 Budapest, Hungary
Tel: + 36 1 463 3889
Fax: + 36 1 463 3107
molnar@bme-tel.ttt.bme.hu

Kivonat

Az elmult évtizedekben az új távközlési szolgáltatások igényei valamint a távközlési világ azon törekvése, hogy az összes távközlő szolgálatot ugyanazon hálózaton valósítsák meg a szélessávú hálózati koncepció (B-ISDN, Broadband Integrated Services Digital Network) kifejlesztését eredményezte. A B-ISDN átviteli technikájának az ATM-et (Asynchronous Transfer Mode) választották, melynek kidolgozása számos kutatási feladat megoldását igényli.

A dolgozat távközlési hálózatok teljesítőképességi és minőségi jellemzőinek vizsgálatával foglalkozik, kiemelten az ATM alapú hálózatok működésének minőségi elemzésével.

Az első rész néhány objektív beszédtorzítási mérték értékelésének kutatási eredményeit ismerteti. A vizsgálatokban a mértékek kiértékelése szabványosított páros összehasonlításos módszerrel történt. A kutatás a beszédkódolók és kommunikációs hálózatok tervezéséhez elterjedten használt spektrális burkoló objektív mértékek alkalmazhatósági korlátainak kimutatását, valamint az objektív mértékek egy továbbfejlesztési irányát eredményezte.

A dolgozat második részének témája az ATM hálózatok hívásszintű vizsgálata, melynek egyik célja annak kimutatása, hogy a B-ISDN forgalom leírására mennyire alkalmas a hagyományos telefonforgalomnál használt Poisson/exponenciális jellemzés. A rész további célja új forgalomleírási módszerek és linkblokkolási mértékek kidolgozása. Ez a rész tartalmazza a linkfoglaltság analízisét különböző érkezési és kiszolgálási folyamatok esetén, és két approximációs módszert is ismertet. Az első módszer a BPP (Bernuolli-Poisson-Pascal) eloszláson alapul, míg a második módszer egy entrópia maximalizációs eljárás. A módszerek a várható értéket és a szórást használják a linkfoglaltság eloszlásának becslésére. A szórás számítására az általánosított csúcsossági mérték kerül bemutatásra egy új analitikus módszerrel. A dolgozat az approximációs eljárások segítségével számos új link blokkolási mértéket, valamint ezek hatékony alkalmazásának illusztrálására egy új ATM hálózatdimenzionálási algoritmust is ismertet.

A harmadik rész ATM hálózatok cellaszintű elemzésével, az egyik legfontosabb minőségi paraméter, a cellakésleltetés ingadozásának leírásával foglalkozik. Néhány eljárást mutat be a cellakésleltetés ingadozásának analízisére, mind egyetlen, mind több ATM multiplexer után. Az egymultiplexeres esetre két új modellezési módszer kerül bemutatásra. Mindkét eljárás figyelembeveszi a háttérforgalom csomósodási jellemzőjét, mely az eddigi modelleknél sokkal pontosabb jellemzést eredményez. A bemutatott Markovi módszer egzakt leírást tesz lehetővé, míg a diffúziós modell egy hatékony approximációs eljárás. A cellafolyam jellemzőinek változása kerül elemzésre a többmultiplexeres esetnél, amint az áthalad a multiplexálási fokozatokon. Ebben a részben az $nTri/D/1$ sorbanállási modell egy új megoldási módszere is megtalálható. A dzsitteres cellafolyamok összemultiplexálásának szimulációs eredményekkel támogatott analízise is bemutatásra kerül. A dolgozat végül néhány ATM forgalomszabályozási eljárás és hálózati elem tervezésére ad irányelveket.

Technical University of Budapest
Department of Telecommunications and Telematics

Evaluation of Quality of Service and Network Performance in ATM Networks

Sándor Molnár

Ph.D. dissertation

Scientific supervisor

Péter Tatai and Søren Blaabjerg

Budapest, 1995

Sándor Molnár
Department of Telecommunications & Telematics
Technical University of Budapest,
Sztoczek u. 2., H-1111 Budapest, Hungary
Tel: + 36 1 463 3889
Fax: + 36 1 463 3107
molnar@bme-tel.ttt.bme.hu

Abstract

The progress of developing a cost-effective telecommunication network which support a wide variety of services yielded the standardization of B-ISDN (Broadband Integrated Services Digital Network). The ATM (Asynchronous Transfer Mode) has been chosen the target transfer mode for B-ISDN which calls for resolving several challenging performance issues.

This thesis is dedicated to the performance evaluation of telecommunication networks, particularly ATM networks, and covers several different fields of the subject including both user-oriented Quality of Service and network provider-oriented Network Performance parameters.

The first part of the thesis presents a performance evaluation study on the spectral objective measures of speech quality. In this study the Equivalent Noise Paired Comparison Test is applied for the evaluation of the candidate objective measures which is a successful standardized method of subjective speech quality assessment. The results of this research shows the applicability limitations of these measures, which are widely used for designing speech coding and communication systems, and indicates a direction of developing more accurate objective speech quality measures.

The second part of the thesis studies several issues of call scale performance evaluation of ATM networks. It introduces a robustness and sensitivity analysis of link occupancy investigating the effect of various arrival and service processes for showing the unreliability of the classical Poisson/exponential description of B-ISDN traffic. This part presents two methods to approximate the link occupancy distributions based on matching the mean and the variance. The first approximation based on the BPP distribution while the second approach using an entropy maximization method. It is also shown that the concept of generalized peakedness provides an efficient tool for finding the variance of the occupancy distribution and a new closed form expression of the generalized peakedness is derived. Several link blocking measures are proposed based on the approximations and their applicability are demonstrated in a new ATM network dimensioning algorithm.

In the part on cell scale performance evaluation one of the most important performance parameter the Cell Delay Variation in both single and cascaded ATM multiplexers are analyzed. Two methods are suggested to evaluate the CDV in a single ATM multiplexer. The first approach is an exact Markovian method taking into account the burstiness of the interfering background traffic. The second one is a diffusion approximation providing a very efficient way for computing point process characteristics of the perturbated cell stream. In the analysis of cascaded ATM multiplexers the characterization of a CBR cell stream going through several queues is investigated. A new solution method for $nTri/D/1$ queue is also derived. Evaluation studies of the superposition of CDV affected CBR cell streams after both single or cascaded multiplexers are presented. Finally, the thesis introduces some guidelines for designing Traffic Control functions and dimensioning network elements in ATM networks.

Acknowledgements

I would like to express my sincerest gratitude and appreciation for my advisor, Péter Tatai, for his support. Péter provided me the background of speech processing and turned my interest to the research of the speech quality assessment. The completion of this thesis in a large part due to his continuous encouragement, critical comments, helpful suggestions and his nice personality.

I am especially grateful to Søren Blaabjerg for his guidance and help during my research. I had the opportunity to work with Søren more times at Ellemtel in Sweden and at the Technical University of Denmark. During these study periods I have learned much from Søren in the field of teletraffic theory and his inspirations and ideas initialized my research activities related to ATM networks. I highly enjoyed these study tours due to his kind help and delighting company. Without the help of Søren this thesis would have never been performed.

I would also like to thank Tamás Henk for the lots of encouragement and help, András Faragó for the fruitful discussions and suggestions, and Géza Gordos for the continuous support.

Special thanks to all members and Ph.D. students of the Department of Telecommunication and Telematics, and particularly, to the members of the High Speed Network Laboratory for their many kinds of help during my research. Also a great thank goes to Henning Christiansen for his kind help in Denmark and for providing me with the simulation program applied in several thesis results.

The research reported herein was supported by Ellemtel. Ellemtel provided opportunities for me to participate study tours and conferences. Moreover, the cooperation between Ellemtel and the Department of Telecommunication and Telematics afforded a possibility for me to taking part in an interesting research project on ATM networks. I have benefited greatly from the fruitful discussions with the colleagues from Ellemtel giving me many inputs for my research. I am especially thankful to Miklós Boda for all his support during my work.

I am very thankful to my family for their various help. Throughout the years of my study, they provided me a stable background and no doubt that their unflagging inspirations and priceless love made it possible for me to carry out my research.

Contents

List of Tables

List of Figures

List of Abbreviations

AAL	ATM Adaptation Layer
ATM	Asynchronous Transfer Mode
B-ISDN	Broadband Integrated Services Digital Network
BPP	Bernoulli-Pascal-Poisson
CAC	Connection Admission Control
CDV	Cell Delay Variation
CBR	Constant Bit Rate
CPN	Customer Premises Network
DMAP	Discrete Markov Arrival Process
GCRA	Generic Cell Rate Algorithm
ITU-T	International Telecommunication Union Telecommunication Standardization Sector
LAN	Local Area Network
MAN	Metropolitan Area Network
MAP	Markov Arrival Process
ME	Maximum Entropy
NNI	Network-to-Network Interface
NP	Network Performance
NPC	Network Parameter Control
NT	Network Termination
NRM	Network Resource Management
QOS	Quality Of Service
TE	Terminal Equipment
UNI	User-to-Network Interface
UPC	Usage Parameter Control
VBR	Variable Bit Rate
VCC	Virtual Channel Connection
VCI	Virtual Channel Identifier
VPC	Virtual Path Connection
VPI	Virtual Path Identifier

Part I

General Introduction

Chapter 1

Performance Issues in Telecommunications Networks

A rapid evolution can be observed in the telecommunication world. The enormous progress to make a service-independent and cost effective telecommunication network which can cope with the needs of the growing new service demands yielded the standardization of B-ISDN (Broadband Integrated Services Digital Network) and its switching and multiplexing technique the ATM (Asynchronous Transfer Mode) [22, 96, 21, 93].

ATM is a fast packet switching technique based on virtual channel connections using fixed-size packets called cells. The size of an ATM cell is 53 bytes, five of which are reserved for the cell header followed by 48 bytes user information. Each cell header has virtual channel and virtual path identifiers (VCI/VPI), denoting the routing address which are used in multiplexing. The cell header includes minimal functionality to reduce the intermediate node processing and contains, among others, payload and priority indicators and one octet for a one-bit header error forward correction and for the self-delineation of cell boundaries.

ATM is a connection oriented technique but connectionless services can also be supported. This technique provides a great flexibility in terms of the bit rate assigned to connections. Moreover, ATM provides inherent statistical multiplexing which may result in significant multiplexing gain. The concept of ATM is a promising technique with several appealing characteristics and as the basic technique of B-ISDN intended to support telecommunication services of both present and future.

However, there are disadvantages with ATM, the most significant two being the *Cell Delay Variation* and cell assembly delay. The Cell Delay Variation (CDV), a phenomena which is not experienced in synchronous networks but one of the most important performance characteristics of ATM networks [3, 50, 14, 100]. A good understanding and characterizing of CDV is needed in order to perform a proper performance engineering in ATM networks. The main part of this thesis addresses this issue.

1.1 Critical Aspects of Performance Evaluation

Performance evaluation is an important part of traffic engineering. Traffic engineering, which is discussed in detail by ITU [47], is about the functional relationship between traffic, resources and performance. Performance evaluation methods is concerned with the same three-way relationship with the main purpose to assess the feasibility of a particular network design under different traffic conditions.

Several challenging performance issues need to be resolved before ATM networks become a reality [48, 106, 98]. The ATM introduced various new and unsolved problems (performance analysis of switching architectures, traffic characterization, evaluation of statistical multiplexing, performance analysis of traffic and congestion control mechanisms, etc).

In spite of the fact that great and growing research activity has been put on this field, resulting in many new models and methods in the last decade, the performance evaluation is still a challenging and unsolved issue of ATM research. This thesis contributes some results to the performance evaluation which could help the performance engineering in ATM networks.

Performance is important to end users when selecting telecommunication services and equipments and also important to network providers when designing and operating telecommunication facilities. Different performance indicators have been developed to measure the quality of services and networks. The *Quality of Service* (QOS) contains user oriented performance measures which are intended to measure how the user is satisfied with the service. In contrast, *Network Performance* (NP) contains network provider oriented performance measures which are expressed measurable characteristics of network elements. The following two sections give a short overview about the QOS and NP measures of B-ISDN [86].

1.2 The Concept of Quality of Service

A typical user is not interested in how a particular service is provided but rather in comparing one service with another in terms of user oriented measures. The objective measures of the B-ISDN systems accurately reflects the user-perceived quality in terms of defined *Quality of Service* (QOS) parameters. The QOS is intended to determine the degree of satisfaction of a user of the service. These user oriented parameters are defined between service access points and take into account all aspects of the service from user's point of view focusing on user-perceivable effects. The QOS parameters can objectively be measured and should be assured by the service provider [48].

The ATM technology introduced many new impairments not experienced in synchronous networks, such as Cell Delay Variation (CDV) and cell loss [93]. The origins and effects of these new types of distortions must be understood to control the network and assure QOS parameters. New distortions can be observed in the transmission of compressed voice and video through ATM networks due to these new impairments. These distortions are mostly not easily quantified by traditional methods. Examples of these

new distortions for voice include gaps in speech, long echo-free delays, bursts of errored bits, speech clipping and phonemic distortions. For example, bursty, short interruptions due to the discarding of cells during congested periods or the misdelivery of cells. The system dependent delays due to cell assembly or queueing, are typical in ATM networks, but do not occur in synchronous networks. Examples of impairments for video include blocking, image persistence and jerky motion. All impairments due to the nature of ATM influence also the codec design, for example CDV complicates the decoder synchronization.

Real QOS measures of users' satisfaction are subjective in nature i.e. depend on user actions and subjective opinions and rather difficult to specify. However, QOS measures restricted to directly observable and measurable characteristics of the service can be defined [48]. Examples of these bearer service QOS parameters are *access delay, incorrect access probability, access denial probability, user information error probability, user information misdelivery probability, service availability, etc.*.

In the research of speech quality assessment, discussed independently from ATM context in the thesis, several measures have been developed [95]. In contrast to the above described bearer service QOS parameters they are intended to cope with the subjective nature of users' quality judgements. The thesis deals with these QOS *speech quality measures*.

1.3 The Concept of Network Performance

A network provider is concerned with the effectiveness of the network, therefore from a network provider's point of view different parameters, *Network Performance* (NP) parameters, are used for purposes of system design, configuration, operation and maintenance. These provider oriented parameters are defined and measurable between network connection elements. NP parameters determine the QOS parameters but they do not necessarily describe the quality in a way that is meaningful to users. For example the *link blocking*, which is an investigated NP parameter of the thesis, is an important performance measure of a link but it does not give much quality information to the user. On the other hand this measure is essential for a network provider to configurate the network.

The ITU-T specified the following NP parameters of the ATM Layer [49]:

- Cell error ratio
- Cell loss ratio
- Cell misinsertion rate
- Severely errored cell block ratio
- Cell transfer delay
  - Mean cell transfer delay
  - Cell delay variation

For CDV measure two parameters have been defined [49], namely the 1-point CDV and a 2-point CDV. The 1-point CDV is defined on the basis of a sequence of consecutive cell arrivals at a single Measuring Point (MP). The 1-point CDV is the difference between the theoretical and the actual cell arrival time at a MP. The 2-point CDV uses the observations of two MPs. The 2-point CDV value is defined as the difference between the absolute cell transfer delay between two MPs and a reference cell transfer delay. The details of these CDV parameters can be found in Chapter 7.2.

This thesis is focusing on the evaluation of *Cell Delay Variation* mainly related to the 1-point CDV measure of the ATM layer.

Chapter 2

Overview of the Thesis

This thesis consists several studies on the performance evaluation of telecommunication networks. It covers various different types of performance measures including both user-oriented QOS and network provider-oriented NP parameters.

The first part of thesis concerns the objective speech quality measures which are QOS parameters. This research has been performed independently from ATM context and has been focused on the performance evaluation of spectral objective speech quality measures. In the second part of the thesis ATM call scale NP evaluation results are presented. Finally, several studies of cell scale NP evaluations of ATM networks are discussed in the third part. The results of the thesis can be categorized as follows:

- QUALITY OF SERVICE:
    1. Speech quality
- NETWORK PERFORMANCE:
    1. Call scale: Link blocking
    2. Cell scale: Cell delay variation

The main goal of the thesis is to get a good insight into the behaviour of the network and its elements and analyze what parameters of the traffic and network are significant and how they influence the QOS and NP parameters under investigation. The final aim of the study, through getting appropriate models, to perform a proper performance engineering in ATM networks.

More specifically, the goals of the thesis are the following:

- To evaluate spectral objective speech quality measures.
- To evaluate the link occupancy distribution with various arrival and service processes.

- To evaluate the occupancy peakedness for arrival and service processes with different variability.
- To derive simple approximations for the link occupancy distribution and link blocking measure.
- To provide further developments of the concept of the generalized peakedness (applications in blocking measures, computation in case of Coxian holding time, computation in case of load sharing).
- To give refined ATM network dimensioning algorithms using two-parameter description of the traffic.
- To evaluate the CDV due to a single ATM multiplexer.
- To evaluate the CDV due to cascaded ATM multiplexers.
- To evaluate the CDV in an ATM multiplexer receiving a superposition of CDV affected CBR cell streams.
- To give designing guidelines for ATM Traffic Control and Network Element dimensioning

Part II

QOS Performance Evaluation

Chapter 3

Performance Evaluation of Objective Speech Quality Assessment Methods

3.1 Introduction

The telecommunication networks have broadened their scope to embrace several new and different types of communication services, but the transmission of voice remains the central issue of the telecommunication world due to the fact that the speech is the most effective way of human communication. The speech communication services represent a wide and steadily growing field of applications for digital communication systems [88, 54]. These applications include: public commercial telephone networks, mobile communications, satellite communications, private communication lines, switched networks, cellular telephones, voice storage services, etc.

From an economical point of view a communication channel should transmit as much information as possible, therefore efficient speech digitization and compression methods are needed. These are particularly important in the case of radio communication systems and voice storage applications where the bandwidth and information capacity are severely limited therefore low bit-rate speech coding methods are needed. In the following low bit-rate means transmission speed below the standard 64 kbit/s.

All speech coding methods, however, have an undesirable side-effect, namely the degradation of speech quality. The fidelity of speech and the reduction of its bit-rate contradict each other. The efficiency of speech digitization and compression can be measured directly by the resulting transmission bit-rate, but the fidelity of speech can not be interpreted easily because of its subjective nature. Some interpretations will be introduced in the dissertation.

In order to evaluate speech coding and transmission systems, speech quality assessment methods are needed [35, 95, 61, 63, 40]. These are very important not only for optimizing coding algorithms but also for planning effective communication systems. There are two categories of such methods: *subjective* and *objective* speech quality assessments. The subjective methods are based on standardized procedures which use humans to judge the quality of speech. In contrast, the objective methods eliminate human judgments from the

assessment procedure and give computable results based on measurable physical quantities. The main problem of finding a good objective speech quality assessment method, however, is that its results should highly correlate with users' opinion, so once again one has to resort to subjective tests in order to "calibrate" objective measures.

The research results presented in this Chapter concerns the evaluation of the most popular and widely used spectral envelope objective speech quality measures. This study has been performed as a part of the research of the Department of Telecommunication and Telematics, Technical University of Budapest in the project of the speech coding and quality evaluation. The goal of this research was to find objective measures which can be used for accurate quality assessment of speech coders. The group of spectral envelope objective measures has been chosen for evaluation because previously published research results have shown that these measures found to be the most successful objective predictors of subjective speech quality ratings [33, 95]. During the research the performance of a wide range of spectral envelope measures has been evaluated based on comparing their results with subjective test results. It will be shown that none of the known spectral objective measures can provide accurate characterization of the speech quality and a two-parameter quality characterization is a more promising way.

The Chapter organized as follows. Section 3.2 reviews on subjective speech quality methods and presents experimental results on CELP coders obtained by the Equivalent Noise Comparison Test. A review of the objective speech quality measures can be found in Section 3.3 and evaluation results of a broad class of spectral envelope objective measures are presented. Finally, Section 3.4 concludes the results of this Chapter.

3.2 Subjective Speech Quality Assessment

The speech quality depends primarily on human perception, so subjective quality assessment methods implies humans as referees. There are two categories of subjective measures: *utilitarian* and *analytic*. Utilitarian methods measure speech quality on a unidimensional scale, therefore results can be summarized by a single number, which is capable of comparing communication systems directly. Analytic methods generate their results on a multidimensional scale reflecting various speech quality components.

A category of utilitarian methods focused on speech intelligibility called *Intelligibility Tests* consists of articulation tests, rhyme tests and speech interference tests [95]. *Articulation Tests* [61] and its modified version the *Equivalent Loss Method* [61] are the widely used methods of this category.

Another category of utilitarian methods is *Quality Tests*. The intelligibility tests are unable to measure the speech quality when speech is highly intelligible, and this is exactly what we expect from most speech services. So methods are needed which can measure other attributes such as pleasantness or naturalness. For these purpose new methods have been developed and the most widely used methods are the *Mean Opinion Score* (MOS) [61] and the *Paired-Comparison Methods* [95].

The analytic methods attempt to obtain different attributes of quality of perceived

speech by exploiting the phenomenon that listeners usually agree on the degree to which speech impairment is present, but vary in their preference of that degradation. Therefore analytic methods generate a multidimensional characterization of the speech quality. Some methods have been developed which produce this kind of parametric description of speech such as *Paired Acceptability Rating Method* (PARM), *Quality Acceptance Rating Test* (QUART) and *Diagnostic Acceptability Measure* (DAM) [104]. Although the analytic methods provide a fairly good description of speech quality, such investigations are difficult and time consuming.

From the wide range of subjective methods the *Equivalent Noise Paired Comparison Test* [63] has been chosen for the performance evaluation of the spectral objective measures. This method is a utilitarian method from the group of the *Paired-Comparison Methods* [95]. I have chosen this method because of the following reasons:

- It is a standardized subjective quality assessment method (ITU Recommendations P.81 [52]).
- It can provide very quick tests (e.g. 20 listeners are sufficient for a quick test as compared to the MOS which needs several hundred or thousand listeners).
- It is sensitive to slight differences between speech samples (e.g. the MOS cannot distinguish fine differences between speech samples assessed by E-grade votes in case of high quality speech).
- It has a good modeling ability to distortions produced by logarithmic quantizers which is important because the purpose of this research was to find objective quality assessment methods intended to characterize the quality of codecs which produce similar distortions.

3.2.1 The Equivalent Noise Paired Comparison Test

In case of this method two signals are presented to listeners, who are asked to choose the better one. It is a forced comparison, "Equal" quality answer is not allowed. The percentage of the signal under test chosen as the preferred one is the preference score. The method uses a reference signal with varied signal-to-noise (SN) ratio. In the test the quality is defined as the SN ratio of reference signal corresponding to 50% preference level. ITU recommends a reference signal generator device called *Modulated Noise Reference Unit* (MNRU) for such tests in ITU Recommendations P.81 [52]. MNRU contains a white noise source modulated by the speech signal and the generated multiplicative noise is added to the speech signal at varied level to produce the reference signal (see Figure 3.1).

Such a reference signal has a speech component and a speech-amplitude correlated noise component with flat frequency spectrum. The signal-to-noise ratio (denoted by Q [dB]) can be set in MNRU and it is constant over a wide dynamic range. Therefore its subjective effect is very similar to that of the distortion produced by logarithmic quantizers

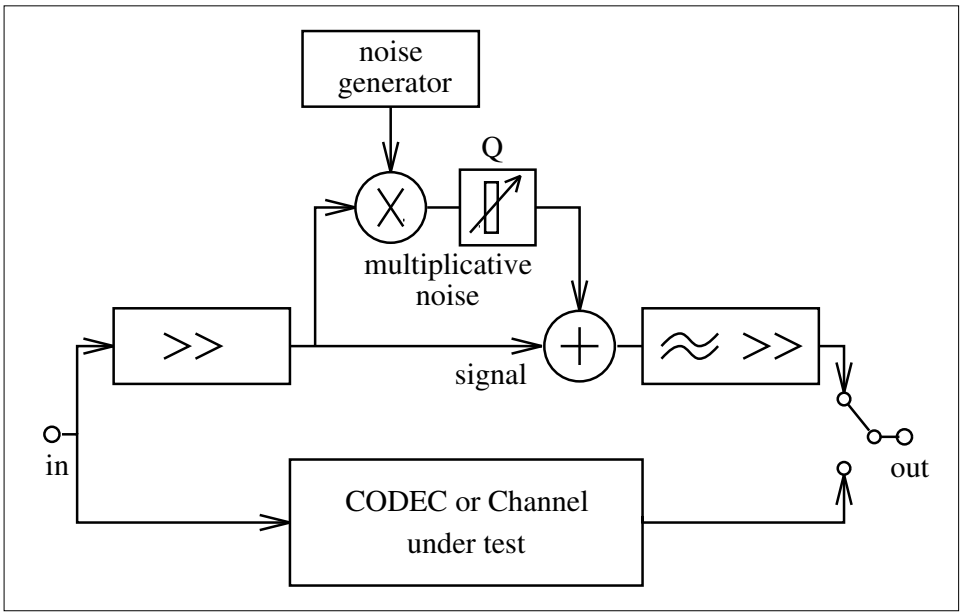

Figure 3.1: Equivalent Noise Comparison Test

(standard PCM systems). Modifying MNRU with a noise-shaping filter, Q becomes almost independent from frequency.

Although it is sometimes difficult to compare signals with different types of impairments, the great advantage of the Paired Comparison Tests is that they provide highly precise assessment on an absolute scale of speech quality even with about 20 listeners. Therefore they are ideal for quick tests. The computer assisted speech quality assessment system Qualiphon developed at DTT, TUB [35] is also based on the Equivalent Noise Paired Comparison Test. In contrary to this, quality tests based on MOS require hundreds or thousands of listeners to provide reliable results and the scale is relative, varies with time, country etc., like marks in schools. Nevertheless, thorough international investigations usually include MOS test because it can take into account the various kinds of deterioration on a single scale.

3.2.2 Experimental Results

One of the promising algorithms at low bit-rate, 8 kbit/s and below, is the CELP (Code Excited Linear Prediction) coding. A real time CELP codec at a rate of 4800 and 3200 bit/s has been chosen for evaluation as presented in the following.

The CELP codec used for subjective testing is realized by a DSP in real time [43]. It conforms to the Proposed US Federal Standard 1016 jointly developed by the US DoD and AT&T Bell Laboratories for 4800 bit/s voice coders.

In the CELP coder a Linear Prediction (LP) analysis is applied to estimate a 10^{th} order LP filter by means of the autocorrelation method, using a 30 ms Hamming window. The results of the LP analysis, the linear prediction coefficients are transformed to Line

Spectrum Pairs (LSP) coefficients. The optimum excitation sequences for the LP synthesis filter are obtained from an adaptive and a stochastic codebook, thus the CELP algorithm can be considered as a two-stage vector quantizer for the speech signal. The algorithm includes a joint optimization process for code vector index and the quantized gain factor.

For subjective speech quality assessment of the CELP coder the Equivalent Noise Paired Comparison Test is applied with MNRU using the Qualiphon system [35]. The speech material for the test consisted of two sets of sentences. Each set contained 5 pairs of sentences. In each sentence-pair there was a distorted (CELP) sentence, and a reference sentence with a predetermined level of multiplied noise. The order of distorted and reference sentence and the changing of distortion level are randomized. In order to eliminate sentence dependency, in the second set of sentence-pairs the role of reference and test sentences was exchanged. The number of listeners was 25. The test procedure was carried out with both 4800 and 3200 bit/s CELP codecs.

The 50% preference point gives the Q value relating to the tested codec. After averaging all tests, the Q value is 11.7 dB at 3200 bit/s and 17.1 dB at 4800 bit/s rate [87].

The acceptable limit for commercial telephone service is about Q=20 dB. According to ITU Recommendations G.113 [51] this is about equivalent to the performance of 14 asynchronously tandemed 8-bit PCM codecs, i.e. 14 qdu (quantizing distortion units) based on a $15log_{10}n$ summation law, where n is the number of qdu's [52].

As can be seen neither the 3200 nor the 4800 bit/s CELP codecs meet the requirements. Based on preliminary tests, however, it is expected that the 7200 bit/s CELP codec (not available in hardware form yet) can provide an acceptable speech quality. It can also be concluded from the results that there is a significant quality difference between the performance of the two codecs (5.4 dB).

3.3 Objective Speech Quality Assessment

However good assessment of speech quality can be provided by subjective tests, they have several disadvantages. Namely, they are expensive, slow, difficult to handle, non-repeatable due to the fact that human listeners' decisions depend on the test conditions and on their personal disposition. Especially the time consuming nature of subjective measures excludes their use in the design and optimization of speech coding systems and communication systems.

Computable objective measures of speech quality based on measured physical parameters are much more desirable [36, 33, 18, 62, 95]. They are cheap, simple, repeatable and fast in comparison with subjective measures, but they can be applied only if they predict subjective speech quality sufficiently well. So the task is to find an objective measure which can be efficiently computed from the original and distorted speech data set, and which highly correlates with subjective tests.

This task is not easy to solve because the human speech perception process is very complex and poorly understood. It involves also the grammar and other diverse factors such as the speakers' attitude and emotional state. People use a lot of redundant information in

speech and, as a result, certain slight distortion effects could cause complete intelligibility loss, while other more extensive distortion products may be almost unperceivable. To perform a quality assessment, the objective measures should take into consideration semantic, prosodic, syntactic, phonetic, etc. information of speech. Of course, no objective measures provide all these, and speech coding systems generally do not produce e.g. semantic distortions but only a fraction of all possible distortions. Accordingly, it is possible to find objective measures showing high correlation with subjective results.

Waveform distortion measures are defined in time domain and based on some kind of discrepancy between the original and the distorted speech waveform. These type of measures are known as variants of *Signal-to-Noise Ratio* (SNR) where noise is usually defined as the difference between the original and distorted signal. Owing to an inevitable coding and transmission delay, precise synchronization is necessary between the two waveforms.

The conventional SNR defined by Eq. 3.1, which has been used for a long time:

$$SNR = 10log_{10}\frac{\sum_{j=1}^{N} x^2(j)}{\sum_{j=1}^{N}[x(j) - y(j)]^2} \tag{3.1}$$

where $x(j)$ and $y(j)$ denote the samples of the original input and the distorted output speech signals, respectively, and N is the number of speech samples considered. The correlation (R) of this measure with subjective measures ranges from 0.24 and somewhat higher values may be obtained using a multiple regression procedure [70].

As the conventional SNR is inadequate to predict subjective quality, the so-called segmental SNR has been proposed:

$$SEGSNR = \frac{1}{\sqrt{M}}\sum_{i=1}^{M} SNR_i \tag{3.2}$$

where SNR_i is defined as in Eq. 3.1 in the signal frame i, and M is the number of frames. Here one frame is a segment of speech, usually 10...30 ms long.

The SEGSNR is based on the experimental fact that the inherently nonstationary speech can be considered approximately stationary for such a short interval as a frame (10$\sim$30ms). Since distortion effects depend on speech statistics, a measure which is the average of objective measures calculated separately for each frame provides a better quality indicator than overall measures.

The heuristic method using arithmetic mean of logarithmic quantities in Eq. 3.2 corresponds to equal weighting of high and low level sounds of an utterance. This is justified by the investigations resulting in high R values which are above 0.77 [95].

Further developments in SNR have resulted in the *Frequency Variant Segmental SNR* ($R = 0.93$) [95] which takes account of the distribution of distortion products in frequency. There are other variants, too, as the *Granular Segmental SNR*, *Articulation Index* and its improved method the *Speech Transmission Index* [95].

These methods, however, can be applied only to waveform coders, which attempt to reproduce the signal shape. More efficient coding algorithms exploit also the insensitivity of human perception to phase information and they reproduce waveforms that show little resemblance to the original speech. In this case the distortion products can no longer be separated by a simple subtraction in the time domain.

Spectral Distortion Measures are defined as discrepancy between the original and distorted speech spectra in frequency domain. The *Spectral Distortion* (SD) defined by Eq. 3.3 provides a logarithmic spectral distortion measure which can be computed by FFT:

$$SD = \left[\frac{1}{\pi} \int_0^{\pi} \left[ln \frac{S_x(\omega)}{S_y(\omega)} \right]^2 d\omega \right]^{\frac{1}{2}} \tag{3.3}$$

$S_x(\omega)$ and $S_y(\omega)$ denote the original and distorted speech spectrum, respectively. Experiments have yielded $R = 0.6$ for SD [95].

A popular and successful type of the spectral measures, which are comparing the differences in the spectral envelope, is outlined in detail in the following section.

Another successful direction of the research is the *Auditory Distortion Measures*. This measures try to model the human perception mechanism [42, 4]. One of the successful measure from this type is the *Bark Distance Measure* [105] based on the Bark spectrum, which reflects the ear's nonlinear scale of frequency and amplitude. $R = 0.85 - 0.98$ is obtained for BSD [105].

Another approach to construct good objective measures is to combine different objective measures using multiple regression analysis to maximize the correlation with subjective results. Although some improvements have been yielded in this way and the correlation coefficients have varied from 0.8 to 0.996 (the upper limit was found only for a restrictive class of waveform coders), the multidimensional scaling of objective measures has shown that the various objective measures are not independent. This means that even though these measures are arithmetically quite dissimilar, they all measure practically the same features of speech quality, hence no significant improvement can be achieved by combining them into one composite measure.

A more promising approach is to design objective measures which predict individual qualities of speech and to combine them into a new measure called *Parametric Objective Measure* [95]. Parameters like those used for DAM seem suitable because they indicate perceptually different aspects of speech quality. Recently such a measure has been developed with good correlation coefficient ($R = 0.82$) over broad classes of distortions [95].

The short overview above illustrates that the design of an appropriate objective measure is not a trivial task. Since the speech production and perception mechanisms are very complex, no simple objective measure can ever be expected. In fact, objective measures can be designed only for a limited and well defined class of distortions. The fast progress in speech coding, processing and transmission systems, however, necessitates a continuous research to cope with the ever increasing variety of degradations. In the next section various spectral measures are presented, which are all evaluated in the dissertation.

3.3.1 Spectral Envelope Distortion Measures

The spectral measures presented in this section are all based on the spectrum envelope distortion computed by Linear Predictive Coding (LPC) [34]. This group of measures found to be one of the best candidate distortion measures of speech quality [33, 95] and this was the reason to chose this group of objective measures for evaluation. They attempt to provide macroscopic (smoothed) spectral information, which is believed to be the most important speech signal characteristic.

The LPC measures compare the all-pole spectral models of the original and distorted speech (Figure 3.2).

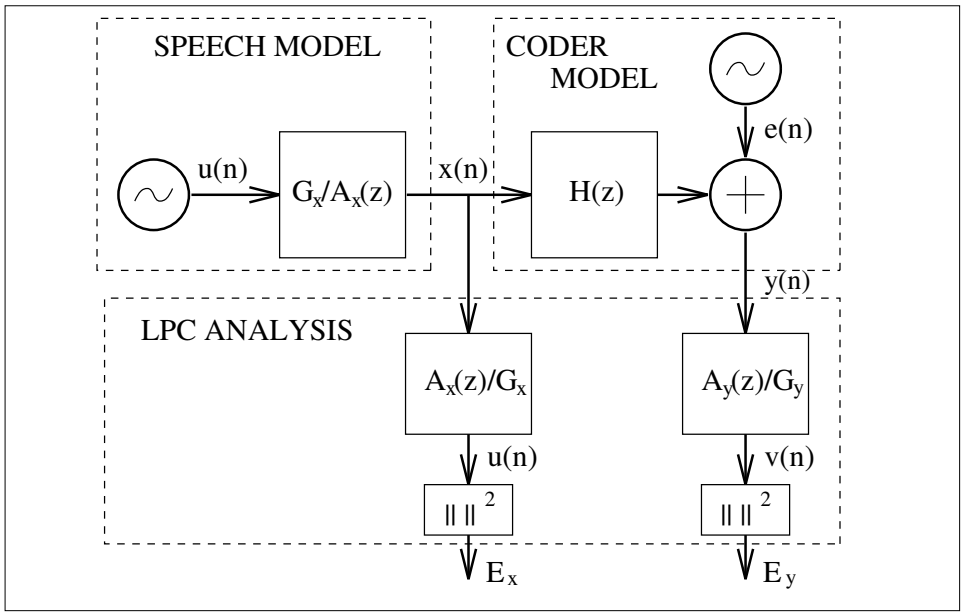

Figure 3.2: Measuring Coder Performance with LPC Measures

The speech model defined in Figure 3.2 consists of the all-pole filter $G_x/A_x(z)$ which is exited by the normalized excitation source $u(n)$. This excitation source is defined to be the normalized residual error in the LPC model of the original speech $x(n)$ and therefore the output of the speech model is exactly $x(n)$. The coder model is composed of a time-varying filter $H(z)$, to model the linear distortions (i.e., attenuation, delay, band limiting), and an additive noise $e(n)$, to account for nonlinear distortions like quantizing noise in the coder. The coder output $y(n)$ is filtered by the inverse filter $A_y(z)/G_y$ to produce the normalized residual $v(n)$. The $A_x(n)$ and $A_y(n)$ filters are defined by the LPC coefficients $a_x(k)$ and $a_y(k)$ for $k = 0, ..., p$ in case of a p-order LPC prediction obtained by LPC analysis [34]:

$$A_x(z) = \sum_{k=0}^{p} a_x(k) z^{-k} \quad \text{and} \quad A_y(z) = \sum_{k=0}^{p} a_y(k) z^{-k} \tag{3.4}$$

with $a_x(0) = a_y(0) = 1$.

The LPC measures are based on the LPC analysis of both the original $x(n)$ and distorted $y(n)$ speech and defined as shown in the following.

The *Log Likelihood Ratio* (LR) measure expresses the dissimilarity between all-pole models of the original and distorted speech waveforms. It is defined by Eq. 3.5.

$$LR = ln\left[\frac{1}{\pi}\int_0^{\pi}\left|\frac{A_y(e^{j\omega})}{A_x(e^{j\omega})}\right|^2 d\omega\right] = ln\left[\frac{\bar{a}_y R_x \bar{a}_y^T}{\bar{a}_x R_x \bar{a}_x^T}\right] \tag{3.5}$$

where $A_x(z)$ and $A_y(z)$ are the analysis filters given by LPC coefficient vectors $\bar{a}_x$ and $\bar{a}_y$ of the original and distorted speech, respectively, and R_x is the autocorrelation matrix of the original speech. The elements of R_x are defined as follows:

$$r_x(|i-j|) = \sum_{n=1}^{N-|i-j|} x(n)x(n+|i-j|) \quad \text{for} \quad |i-j| = 0,1,2,...,p. \tag{3.6}$$

where N is the length of the frame and p is the prediction order used in LPC. A simple interpretation of LR can be given in Figure 3.2: LR is the log ratio of the LPC residual energies of the distorted and original speech: $LR = ln(E_y/E_x)$ [103]. The correlation of LR with subjective measures is around $R = 0.59$ according to [95]. Other popular variants of LR are the *Itakura-Saito Measure*, the *COSH Measure* and the *Weighted Log Itakura-Saito Measure* [95], which provide similar performance but originate from a somewhat different mathematical reasoning.

The *Cepstrum Distance Measure* (CD) is based on cepstral coefficients $c(k)$ [95, 40] which can be obtained from the roots $z_x(i)$ and $z_y(i)$ of the polynomial $A_x(z)$ and $A_y(z)$, respectively.

$$c_x(k) = -\frac{1}{k}\sum_{i=1}^{p} z_x^k(i) \quad \text{and} \quad c_y(k) = -\frac{1}{k}\sum_{i=1}^{p} z_y^k(i) \tag{3.7}$$

Based on these coefficients the CD is defined by

$$CD = \left[[c_x(0) - c_y(0)]^2 + 2\sum_{k=1}^{p}(c_x(k) - c_y(k))^2\right]^{\frac{1}{2}} \tag{3.8}$$

Using Parseval's relation it can be seen that CD is equivalent to a log spectral distance of cepstrally smoothed spectra. For CD the correlation with subjective results has ranged from 0.8 to 0.9 [61, 63].

To overcome the asymmetric property of the LR measure the *COSH measure* (COSH) has been developed [46]. It measures (Eq. 3.9) the spectral distance between the original and distorted speech samples weighted by the *cosh* function:

$$COSH = \left[\frac{2}{\pi}\int_0^{\pi}(cosh[V(e^{j\omega})] - 1)d\omega\right]^{\frac{1}{2}} = min_{\left[\frac{G_x}{G_y}\right]^2} COSH =$$

$$= \left[2 \left[\frac{\sum_{i=-p}^{p} L_y(i) r_x(i) \sum_{i=-p}^{p} L_x(i) r_y(i)}{E_x E_y} \right]^{\frac{1}{2}} - 2 \right]^{\frac{1}{2}} \tag{3.9}$$

where

$$V(e^{j\omega}) = ln \left[\frac{G_x^2}{|A_x(e^{j\omega})|^2} \right] - ln \left[\frac{G_y^2}{|A_y(e^{j\omega})|^2} \right] \tag{3.10}$$

and $L_x(i)$, $L_y(i)$ are the maximum likelihood coefficients defined by

$$L_x(i) = \sum_{j=0}^{p-|i|} a_x(j) a_x(j+|i|) \quad L_y(i) = \sum_{j=0}^{p-|i|} a_y(j) a_y(j+|i|). \tag{3.11}$$

E_x and E_y are the LPC residual energies of the segment of the original and distorted speech, respectively. G_x and G_y are the gain factors of the original and distorted all-pole models, respectively.

The *Energy Ratio* (ER) [95] is also a frequently used measure, which has close relationship with the LR measure. ER is defined by Eq. 3.12:

$$ER = \left[\frac{\bar{a}_y R_x \bar{a}_y^T}{\bar{a}_x R_x \bar{a}_x^T} \right]^{0.125} \tag{3.12}$$

The *Linear LPC Measure* (LILPM) and the *Log LPC Measure* (LOLPM) compute the linear and logarithmic distance between the LPC coefficients of the original and distorted speech, as defined by Eq. 3.13 and Eq. 3.14 [95], respectively.

$$LILPM = \left[\frac{1}{p} \sum_{i=1}^{p} [a_x(i) - a_y(i)]^2 \right]^{\frac{1}{2}} \tag{3.13}$$

$$LOLPM = \frac{1}{p} \sum_{i=1}^{p} 20 log_{10} \left| \frac{a_y(i)}{a_x(i)} \right| \tag{3.14}$$

Similarly to LILPM and LOLPM, measures can be defined by computing the linear or logarithmic distance between the Partial Correlation (PARCOR) coefficients of the original and distorted speech $(k_x(i), k_y(i))$ [34, 95]. These PARCOR measures are called *Linear PARCOR Measure* (Eq. 3.15) and *Log PARCOR Measure* (Eq. 3.16), respectively [95].

$$LIRCM = \frac{1}{p} \sum_{i=1}^{p} |k_x(i) - k_y(i)| \tag{3.15}$$

$$LORCM = \frac{1}{p} \sum_{i=1}^{p} 20 log_{10} \left| \frac{k_y(i)}{k_x(i)} \right| \tag{3.16}$$

Another type of very successful measures are the *Linear Area Ratio* (LIAR) and the *Log Area Ratio* (LOAR) measures [95]. They use the same definition as LIRCM and LORCM (Eq. 3.15 and Eq. 3.16) but with the so called Area Ratio coefficients defined by

$$ar_x = \frac{1 + k_x(i)}{1 - k_x(i)} \quad ar_y = \frac{1 + k_y(i)}{1 - k_y(i)} \tag{3.17}$$

All measures above give results for one segment of the speech. For the full speech sample a global measure can be computed by averaging these results.

3.3.2 Evaluation of Spectral Envelope Distortion Measures

The spectral envelope distortion measures introduced above have been evaluated on the same real time CELP codec [43] at a rate of 4800 and 3200 bit/s that were applied for the subjective test described in Section 3.2.2 because the purpose of the research was to evaluate the performance of these measures by comparing their results with subjective test results. (Details about the CELP codec can be found in that Section.) The characteristics of the LPC measures and investigated speech samples can be found in Table 3.1.

Number of Recordings	10
Length of Recordings	4 sec
Length of Segments	32 ms
Sampling Frequency	8 kHz
Windowing Technique	Hamming
LPC Technique	Autocorrelation
Degree of Prediction	30

Table 3.1: Characteristics of Investigated Speech Material and LPC Measures

The digital recordings of the same 10 sentences were processed by each rate of the coders that were used for the subjective quality evaluation of the coders in Section 3.2.2. The results of the applied LPC measures and the relative difference of their results between the 3200 bit/s and 4800 bit/s CELP coders can be seen in Table 3.2 [85].

Codec	**LR**	**CD**	**COSH**	**ER**	**LILPM**
CELP 3k2	0.707	1.741	1.445	1.097	0.336
CELP 4k8	0.727	1.688	1.463	1.097	0.349
Relative Difference	2.8%	3.1%	1.2%	0.0%	3.8%
Codec	**LOLPM**	**LIRCM**	**LORCM**	**LIAR**	**LOAR**
CELP 3k2	9.114	0.102	8.325	0.342	1.967
CELP 4k8	9.485	0.105	8.596	0.346	2.020
Relative Difference	4.0%	2.9%	3.2%	1.2%	2.7%

Table 3.2: Results of Spectral Objective Speech Quality Measures

It can be seen from the results that all investigated spectral measures have indicated almost the same quality degradation. All the results are the same within a range of 4% relative error, which also includes measuring and numerical computation errors.

In contrast, the results of the subjective test have clearly indicated a significant quality difference of 5.4 dB between the 3200 bit/s and 4800 bit/s CELP coders in the MNRU signal-to-noise ratio.

According to our opinion the explanation of this is the following: Both CELP coders introduce approximately the same amount of linear distortions but a different value of quantizing noise. These objective measures are mostly sensitive to the linear distortion of the speech and they are not able to capture the effect of the nonlinear distortion i.e. the quantizing noise.

However, some of the spectral measures can also indicate the presence of nonlinear distortion. For example, the Log Likelihood Ratio weights linear and nonlinear distortion equally [19], and many investigations have shown that this measure has very high correlation with subjective quality [95], but the ability of indicating the noise difference in the described investigations has been found not adequate.

It can also surprisingly be noted that even the widely accepted Cepstral Distortion measure, which is believed to be the best [40, 62, 61, 46, 63], is not able to perform proper speech quality characterization.

This broad range of spectral envelope objective measures have been developed and used for the evaluation of speech coding systems all over the world [95]. Many of them have found to be successful in predicting the quality of a wide range of speech coders. Thereby they are strongly suggested for general usage. The above presented investigation clearly shows the limitations of these objective measures. Although they measure different characteristics of the deviation of spectral envelopes and to some extent they can also take into account the nonlinear distortions of coders, they are not appropriate to characterize the very popular and widely used CELP and similar type of coders. Roughly speaking we can conclude that all these LPC measures are properly indicating the linear distortions only and they are not very sensitive to the quantizing noise.

The final conclusion is that these known spectral envelope objective measures are not always appropriate alone to characterize the speech quality and one has to be careful when applying them for quality assessment of speech coders and should be aware of their limitations.

3.4 Summary and Further Research

Various known and widely applied spectral objective quality measures have been investigated (Log Likelihood Ratio, LPC Cepstrum Distance Measure, COSH Measure, Energy Ratio, Linear LPC Measure, Log LPC Measure, Linear PARCOR Measure, Log PARCOR Measure, Linear Area Ratio, Log Area Ratio) in the case of a 4800 bit/s and 3200 bit/s CELP (Code Exited Linear Prediction) coder [43] and the results are compared to subjective quality tests [87, 85].

The subjective test has been performed by Paired Comparison [63] based on the MNRU (Modulated Noise Reference Unit) standardized by ITU [52]. These tests showed a significant quality difference of 5.4 dB between the 3200 bit/s and 4800 bit/s CELP coders in the MNRU signal-to-noise ratio.

In contrast, the results of all investigated spectral envelope objective measures are almost the same (within a range of relative difference smaller than 4%) for both transmission speeds. The inadequate performance of these measures can be explained by identifying that they are not sufficiently sensitive to the nonlinear distortions of coders like the quantizing noise.

Based on the above results it can be concluded that these known spectral measures are not appropriate alone to characterize the speech quality.

There is no promising way to get a spectral objective speech quality measure which is able to capture correctly both the linear and nonlinear distortions of speech. A proposal for a two-parameter objective speech quality measure, which could provide a better characterization of speech quality, is the following: one parameter should be a correct measure of the linear distortion in terms of the smoothed (envelope) spectral difference. It could be e.g. the Cepstral Distortion measure. The other parameter should focus on the nonlinear distortion. This measure could be a comparing of the residual energies of the original and distorted speech filtered by their own analysis filters. The comparison in the residual signals could be done e.g. on the basis of segmental signal-to-noise ratio.

Part III

Network Performance Evaluation on Call Scale

Chapter 4

Performance Evaluation of Link Occupancy

4.1 Introduction

In the beginning of this century Erlang published his famous formula [12] for the loss probability of a telephone system and now we have a historical view on the great success of his formula and the good applicability of the Poisson process in traditional telephone networks.

With the introduction of B-ISDN now we have the questions: can we use the Poisson process for characterizing the call arrival process or not? Can we use the exponential distribution for modeling call holding time or not? So far we have only a few measurements from real B-ISDN environments to give a definite answer. Furthermore, it is difficult to predict good traffic models of future services. On the other hand, we can expect that in many cases of B-ISDN supporting a large variety of services, the arrival process of new connection request will differ from service to service and will often be quite different from a Poisson process [99, 15, 60]. It is also expected that the holding time in most cases will differ significantly from the exponential distribution. However, an analysis is needed to investigate the robustness of the classical Poisson/exponential assumption which is the main purpose of this Chapter.

More general processes and queueing models are needed to achieve appropriate call scale models in ATM. It also means that in many cases the mathematical models with several nice properties (e.g. insensitivity properties) thanks to the Poisson process, which we successfully use in telephone networks, cannot be applied. These expectations actualize the investigation of networks with $G/G/c$ queues both as loss and delay systems. For mathematical reason the $G/G/\infty$ queue is also of interest.

The objective of this Chapter is to provide an insight into the behaviour of link occupancy. The analysis is carried out through the analysis of the $G/G/c$ queue and the focus is on the occupancy distribution and the generalized peakedness which is a quite successful variability measure and probably the best candidate for a B-ISDN traffic characterization.

This study provides a tool for robustness and sensitivity analysis of link occupancy investigating its deviation effects from the classical Poisson/exponential description of B-ISDN traffic.

The Chapter is organized as follows. In the next two sections the concepts of the performed analysis of the occupancy distribution and occupancy peakedness are given, respectively. In Section 4.3.2 a new result for the generalized peakedness in case of Coxian holding time is shown. Finally the analysis results and concluding remarks can be found in the last two sections of this Chapter.

4.2 The Occupancy Distribution

An appropriate model for the investigation of the call scale traffic in B-ISDN is the $G/G/c$ loss system. The link occupancy analysis based on this model of a single link. So far one of the few approaches to obtain the exact steady state distribution of a $G/G/c$ system is by restricting the arrival process to be of the type *Markovian Arrival Process*, [90] and to assume the holding time distribution to be of *phase type* [39, 91, 84]. Then the system can be analyzed through the usual Markovian approach [89, 39, 84]. However, the size of the state space very quickly exceeds any in practice manageable size even when the sparseness of the transition matrix is taken into account, thereby limiting the feasible models to those with a rather limited number of phases in the arrival process, a very limited number of phases of the holding time distribution and only a small number of servers. Furthermore any exact solution approach is far to complicated to be used for dimensioning purposes.

The analysis is significantly simplified when the holding time distribution is exponential, and the solution for the $GI/M/\infty$ and $GI/M/c$ has also been known for about fifty years [94]. However, deriving the occupancy distribution is numerically challenging and in practice impossible when the offered traffic and number of servers grow. It has been common practice to approximate the occupancy distribution of the $GI/M/\infty$ and $GI/M/c$ systems by using the state dependent Poisson process of the type BPP [23, 53]. When the offered traffic increases the BPP distribution converges towards a normal distribution [30], and for high capacity systems it has been suggested to use the normal distribution [23].

In the occupancy distribution study homogeneous traffic situation is considered where each call requires a unit of capacity but in the investigations for the generalized peakedness (Section 4.3), approximate occupancy distributions, blocking measures (Chapter 5) and ATM network dimensioning algorithm (Chapter 6) can be extended for heterogeneous traffic as shown in Section 6.2. In the model under study an infinite server group is offered traffic from a stationary process S with arrival intensity m. All servers act independently of each other and have the same holding time distribution H which is assumed general but with finite mean $1/\mu$.

In order to obtain the stationary occupancy distribution the arrival process is restricted to be renewal and the interarrival time to be of phase-type [39, 91, 84]. Two approaches have been considered for finding the distribution. In the first approach the results and

procedure of Ramaswami and Neuts [97] has been applied which allows the holding time to be general. In the second approach the holding time is exponential which ensures a Markovian system [102].

4.2.1 The Occupancy Distribution in Case of General Holding Time

A phase type (PH) distribution is defined as a distribution function for the time until absorption in an $(n+1)$ state Markov chain with n transient state and one absorbing state. The infinitesimal generator is of the form

$$Q = \begin{bmatrix} T & \boldsymbol{T}^\circ \\ \mathbf{0} & 0 \end{bmatrix} \tag{4.1}$$

where $T = (T_{ij})$ is a non-singular $n \times n$ matrix such that $T_{ii} < 0$ and $T_{ij} \geq 0$ for $i \neq j$, and $\boldsymbol{T}^\circ \geq \mathbf{0}$ is an n-vector satisfying $T\boldsymbol{e} + \boldsymbol{T}^\circ = \mathbf{0}$, where $\boldsymbol{e}' = (1, ..., 1)$. A phase type renewal process can be obtained by resetting the Markov chain Q according to an initial probability vector $(\boldsymbol{\alpha}, \alpha_n + 1)$ after each transition to the absorbing state. We consider as a model of an arrival process this PH renewal process, which has the infinitesimal generator $Q^* = T + T^\circ A^\circ$, where $A^\circ = diag(\alpha_1, ..., \alpha_n)$ and $T^\circ = (\boldsymbol{T^\circ, ..., T^\circ})$.

The investigated model is the $PH/G/\infty$ queueing system, where the holding time distribution is arbitrary. Ramaswami and Neuts [97] derives the basic system of differential equations and obtains

$$\frac{\partial}{\partial t} G(z,t) = [(T + T^\circ A^\circ) + (z-1)\{1 - H(t)\}T^\circ A^\circ]G(z,t) \tag{4.2}$$

with the initial condition $G(z,0) = I$, where $G(z,t)$ is the generating function of the $n \times n$ matrix $G_k(t)$ with elements

$$G_k^{ij}(t) = P[X(t) = k, J(t) = j | X(0) = 0, J(0) = i], \quad t \geq 0 \tag{4.3}$$

for $k \geq 0$, $i, j = 1, ..., n$, where the number of occupied servers and the phase of the arrival process at time $t+$ are denoted by $X(t)$ and $J(t)$, respectively.

Based on the basic system of differential equations we get an infinite system of differential equations for the time dependent occupancy distribution. Truncated at a sufficient large value of the index k they can be solved numerically. Defining $\boldsymbol{g}_k(t) = G_k(t)\boldsymbol{e}$, $k \geq 0$ where the i^{th} entry of $\boldsymbol{g}_k(t)$ is given by

$$g_{ki}(t) = P[X(t) = k | X(0) = 0, J(0) = i], \quad t \geq 0$$

that is,

$$\frac{d}{dt}\boldsymbol{g}_0(t) = [T + H(t)T^\circ A^\circ]\boldsymbol{g}_0(t), \quad \boldsymbol{g}_0(0) = \boldsymbol{e} \tag{4.4}$$

and for $k \geq 1$

$$\frac{d}{dt}\boldsymbol{g}_k(t) = [T + H(t)T^{\circ}A^{\circ}]\boldsymbol{g}_k(t) + \{1 - H(t)\}T^{\circ}A^{\circ}\boldsymbol{g}_{k-1}(t), \quad \boldsymbol{g}_k(0) = \mathbf{0} \tag{4.5}$$

The main advantages of this approach that it produces both the transient and the steady state occupancy distributions, moreover, the complexity of the solution is independent on the holding time distribution. However, for steady state analysis, this solution approach is rather complex, and we have experienced difficulties in the numerical calculations for cases in which the steady state probability of an empty system is very small [76].

4.2.2 The Occupancy Distribution in Case of Exponential Holding Time

For the special case when the holding time is exponential Takács [102] provides a simple method to get the occupancy distribution just before an arrival. However, it should be noted, that even though this analytical solution exists, obtaining numerical values from this solution is challenging and in practice limited to small offered traffic and small number of servers.

The occupancy distribution just before an arrival of the $GI/M/c$ queue can be obtained by

$$p_k^a = \sum_{r=k}^{c}(-1)^{r-k}\binom{r}{k}B_r \tag{4.6}$$

where B_r is the r^{th} binomial moment of $\{p_k^a\}$ and is given by

$$B_r = C_r\frac{\sum_{j=r}^{c}\binom{m}{j}\frac{1}{C_j}}{\sum_{j=0}^{c}\binom{m}{j}\frac{1}{C_j}} \tag{4.7}$$

where $C_0 = 1$ and

$$C_r = \prod_{i=1}^{r}\left(\frac{\Phi(i\mu)}{1-\Phi(i\mu)}\right), \quad r = 1, 2, ... \tag{4.8}$$

with c number of servers and the Laplace-Stieltjes transform of the distribution function of the interarrival time is denoted by Φ.

There is a relation between the occupancy distribution just prior to an arrival p_k^a and the occupancy distribution at an arbitrary time p_k (see Theorem 4 of Chapter 4 in [102]):

$$p_k = \frac{mp_{k-1}^a}{k\mu}, \quad k \in \{1, ..., c\} \quad \text{and} \quad p_0 = 1 - \frac{m}{\mu}\sum_{k=1}^{c}\frac{p_{k-1}}{k} \tag{4.9}$$

Thereby the stationary occupancy distribution of the $GI/M/c$ queue at an arbitrary time

can be computed by Eq. 4.9.

4.3 The Generalized Peakedness

In the traditional telephone traffic theory, the issue of variability in the arrival process of connection requests has been investigated mostly for overflow traffic in systems with alternative routing [26].

Two measures has been intensively used. The most straightforward is the squared coefficient of variation of the interarrival time between two consecutive connection requests [16]. In the case where the arrival process is well described by a renewal process this gives a complete second order characterization [16]. However in the general case, it only gives one component in the characterization [17]. Another important disadvantage comes from the fact that the blocking probability and the occupancy distribution also depends on the distribution and not only the mean of the holding time in the case without Poisson arrivals. A more accurate variability measure is therefore needed. The *generalized peakedness* measure as defined by Eckberg [27] has the advantage that it is a complete second order characterization of the arrival process and furthermore also takes the holding time process into account.

The definition is as follows: Assume that the arrival of connection requests are offered to a link with infinite capacity. Let $L(t)$ be the amount of bandwidth occupied at time t. Then the generalized peakedness $Z(t)$ is defined as:

$$Z(t) = \frac{Var\{L(t)\}}{E\{L(t)\}} \tag{4.10}$$

On a route in a real network it is possible only to measure the actual occupied bandwidth and here only accepted connections contributes. However, since also the amount of blocked connections needs to be monitored, it is possible by combining the occupancy distribution of carried call and the process of connection requests which are blocked to obtain an estimate of the occupancy distribution in the infinite capacity case and thereby get a measured estimate of the peakedness of the connection request on the route.

4.3.1 Computation of Generalized Peakedness

Eckberg has provided formulas for the generalized peakedness [27] assuming only that the arrival process is stationary. Let $U(x)$ denote the renewal function of the process S that is: $U(x) = E[N(a,b)]$ with $N(a,b)$ denoting the number of arrivals in the interval $]a,b]$, when an arrival occurred at time a. Furthermore, define

$$H_2^c(x) = \int_{-\infty}^{\infty} (1 - H(u))(1 - H(u - x))du \tag{4.11}$$

According to formula (3) in [27] then the peakedness Z of the complementary holding time distribution $1 - H$ is

$$Z(H) = 1 + 2\mu \int_{0-}^{\infty} H_2^c(x) dU(x) - \frac{m}{\mu} \tag{4.12}$$

By Eq. 4.12 the arrival stream is characterized in terms of a *peakedness functional* which takes *complementary holding time distributions* as arguments and maps them into *peakedness values*. The intuitive concept of the peakedness functional is that if a given complementary distribution characterizes the "reaction time" of the arrival stream with a system, then the resulting peakedness value is a potentially useful measure of stream variability with respect to that system.

If the holding time is exponentially distributed with mean $1/\mu$, the peakedness formula reduces to:

$$Z_{exp}(\mu) = 1 + U^*(\mu) - \frac{m}{\mu} \tag{4.13}$$

in which U^* denotes the Laplace-Stieltjes transform of U. Restricting the holding time distributions to be of the class which,

$$1 - H(x) = \int_{-\infty}^{\infty} e^{-xt} a(t) dt, \quad x > 0 \tag{4.14}$$

and a is a generalized function [67], Eckberg [27] obtains the following relation between the peakedness of H and the peakedness of an exponential holding time distribution with mean $1/\mu$:

$$Z(H) = 1 + 2\mu \int_{0-}^{\infty} \alpha(y)(Z_{exp}(y) - 1) dy \tag{4.15}$$

where

$$\alpha(y) = a(y) \int_{-\infty}^{\infty} \frac{a(x)}{x + y} dx \tag{4.16}$$

4.3.2 Generalized Peakedness in Case of Coxian Holding Time Distributions

In this section a new closed form expression for the generalized peakedness in case of Coxian holding time is derived [80, 83, 77].

Consider a Coxian distribution represented by a weighted sum of generalized Erlang distributions as shown in Figure 4.1. For simplicity reasons it is assumed that $\lambda_i \neq \lambda_j$ for $i \neq j$ but the general case can be included with only minor changes in the expressions. For the Coxian distributions $a(t)$ in Eq. 4.14 can be written as a series of delta functions[1]:

$$a(t) = \sum_{i=1}^{n} \left(\prod_{j=1}^{i-1} \frac{\lambda_j}{\lambda_j - \lambda_i} \right) \left(\sum_{j=i}^{n} p_j \prod_{k=i+1}^{j} \frac{\lambda_k}{\lambda_k - \lambda_i} \right) \delta_{\lambda_i} \tag{4.17}$$

[1] When $\lambda_i = \lambda_j$ for some i and j, derivatives of delta functions appear in the expression for $a(t)$.

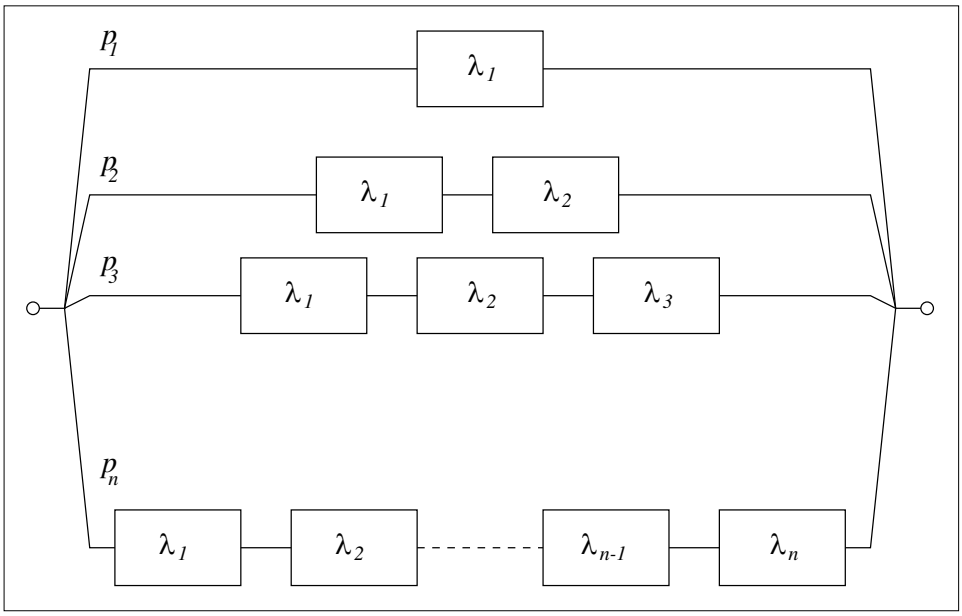

Figure 4.1: Coxian Distribution Represented as a Weighted Sum of Generalized Erlang Distributions

where λ_i ($\lambda_i \neq \lambda_j$ for $i \neq j$) and p_i, $1 \leq i \leq n$, are the intensities and branch probabilities of an n-branch Coxian distribution, respectively (see Figure 4.1). By applying Eq. 4.16 the result of the integral is

$$b(y) = \sum_{i=1}^{n} \left(\prod_{j=1}^{i-1} \frac{\lambda_j}{\lambda_j - \lambda_i} \right) \left(\frac{1}{y + \lambda_i} \right) \left(\sum_{j=i}^{n} p_j \prod_{k=i+1}^{j} \frac{\lambda_k}{\lambda_k - \lambda_i} \right) \tag{4.18}$$

Multiplying by $a(y)$ and substituting the obtained $\alpha(y)$ into Eq. 4.15 we get the peakedness:

$$Z = 1 + 2\mu \sum_{l=1}^{n} p_l \sum_{i=1}^{l} \left(\prod_{j=1}^{i-1} \frac{\lambda_j}{\lambda_j - \lambda_i} \right) \left(\prod_{k=i+1}^{l} \frac{\lambda_k}{\lambda_k - \lambda_i} \right) b(\lambda_l) \left(Z_{exp}(\lambda_l) - 1 \right) \tag{4.19}$$

where Z_{exp} is given in Eq. 4.13.

The importance of this formula is that it provides a closed form expression of the peakedness for any stationary arrival process for which the Laplace transform of the renewal function is available. It includes e.g. renewal processes, Markov renewal processes and doubly stochastic processes. Moreover, the holding time can be any Coxian distribution which covers a very large subset of phase type distributions (e.g. generalized Erlang or hyperexponential distributions) and has the advantage that any distribution can be approximated by Coxian distributions with arbitrary accuracy [20].

In heterogeneous traffic environment the peakedness of the aggregated multirate traffic can be computed by Eq. 6.3 as shown in Section 6.2 using Eq. 4.19 for computing the peakedness of the individual calls.

4.4 Analysis Results

4.4.1 Occupancy Distribution

In this section analysis results based on the method of section 4.2.1 can be found. These results give an overview on how the arrival processes and service processes influence the occupancy distribution [76].

The numerical study have concentrated on the occupancy distribution of a $PH/PH/\infty$ system with a mean occupancy of 10 servers by setting the arrival intensity $m = 10$ and mean holding time $1/\mu = 1$. In the method of section 4.2.1 a mean of 10 servers was close to the upper limit of where the numerical approach was stable. In the Figures c_a^2 and c_h^2 denote the squared coefficient of variation of the interarrival time and holding time, respectively.

In the first scenario the arrival process is fixed and the occupancy distribution is investigated as the holding time distribution changes. Fig. 4.2 and Fig. 4.3 show the occupancy distributions when the arrival process is smooth (Erlang-4 arrival process) and bursty (Hyperexponential arrival process with $c_a^2 = 20$), respectively. For this smooth arrival process the distribution is bell shaped in the considered range while for the peaky hyperexponential arrival process the shape of the occupancy distribution changes dramatically from a bell shaped form when the holding time is very variable towards a curve with very large probabilities of an almost empty systems and a slowly decreasing tail at high occupancy levels when the holding time distribution is Erlang-4.

In the second scenario the holding time distribution is fixed and the occupancy distribution is investigated as the arrival process changes. Fig. 4.4 and Fig. 4.5 give plots corresponding to Erlang-4 and Hyperexponential ($c_h^2 = 20$) holding time distributions, respectively. In cases when the holding time has small variability (Erlang-4 case) the probability mass is moving away from the mean to smaller occupancy levels with increasing burstiness of the arrival process and finally it loses the bell shape type. For high holding time variability (Hyperexponential with $c_h^2 = 20$) the distributions are bell shaped for all investigated arrival processes.

From the results it can be seen that the behaviour of the system is rather complex and the sensitivity of occupancy distribution from the holding time distribution and arrival process depend highly on each other. In such investigations the inputs are two distributions (arrival process and holding time) and the output is also a distribution (occupancy distribution) which makes finding simple behaving rules quite difficult. In the following section the output of the analysis is the generalized peakedness which allows us to take a deeper look into the system behaviour and conclude important practically useful system properties.

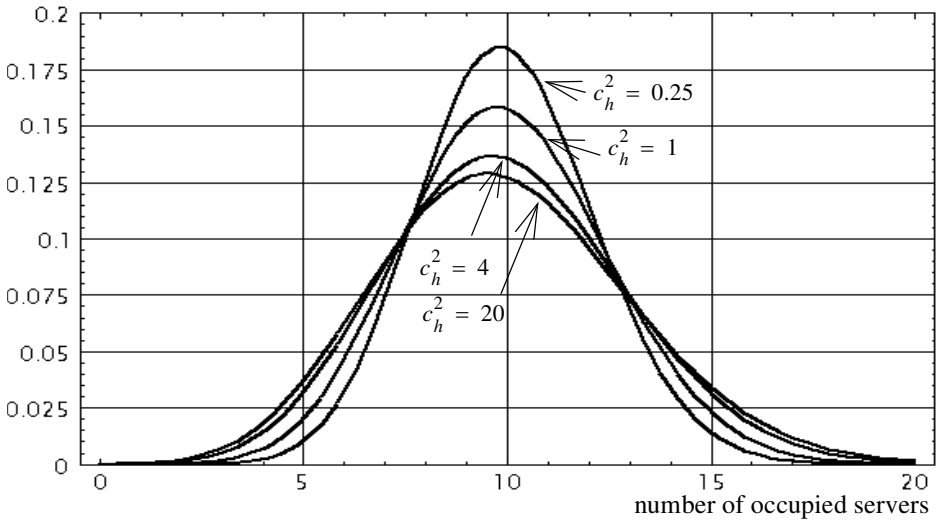

Figure 4.2: Occupancy Distributions with Smooth Arrival Process (Erlang-4)

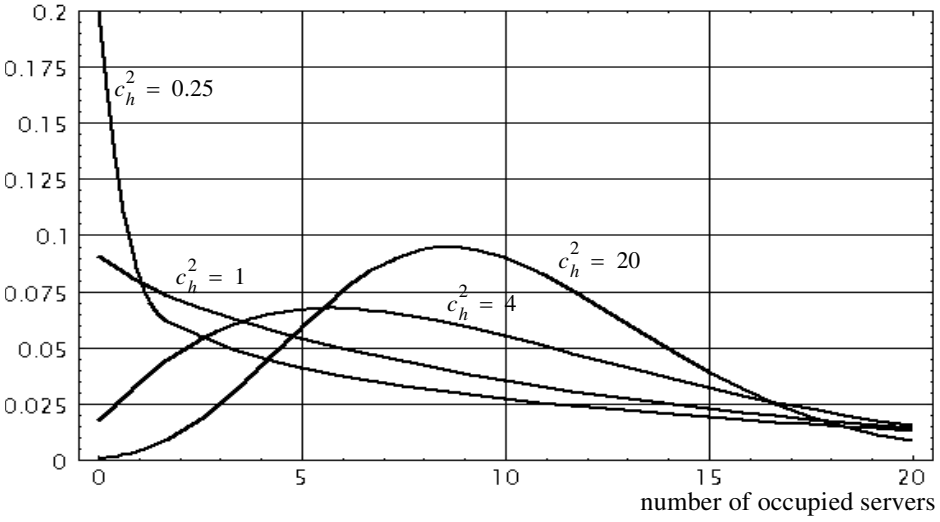

Figure 4.3: Occupancy Distributions with Bursty Arrival Process (Hyperexponential with $c_a^2 = 20$)

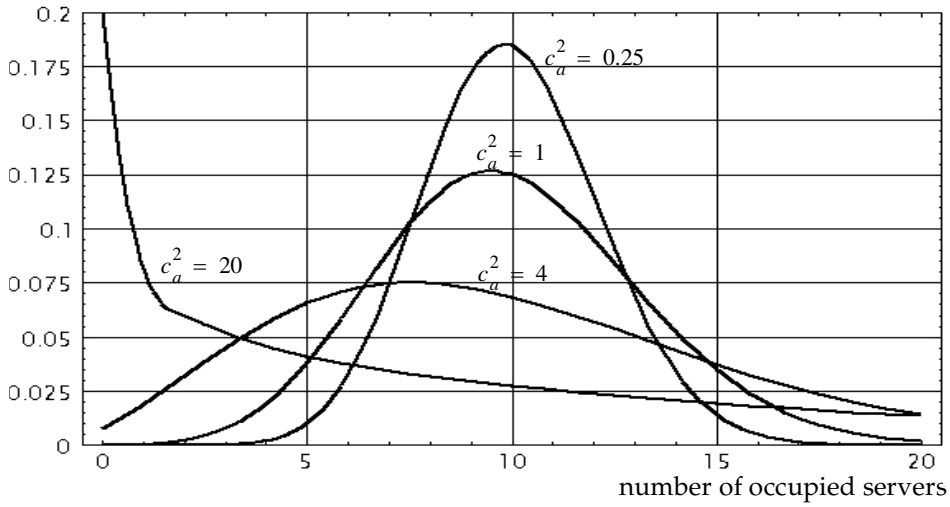

Figure 4.4: Occupancy Distributions with Small Holding Time Variability (Erlang-4)

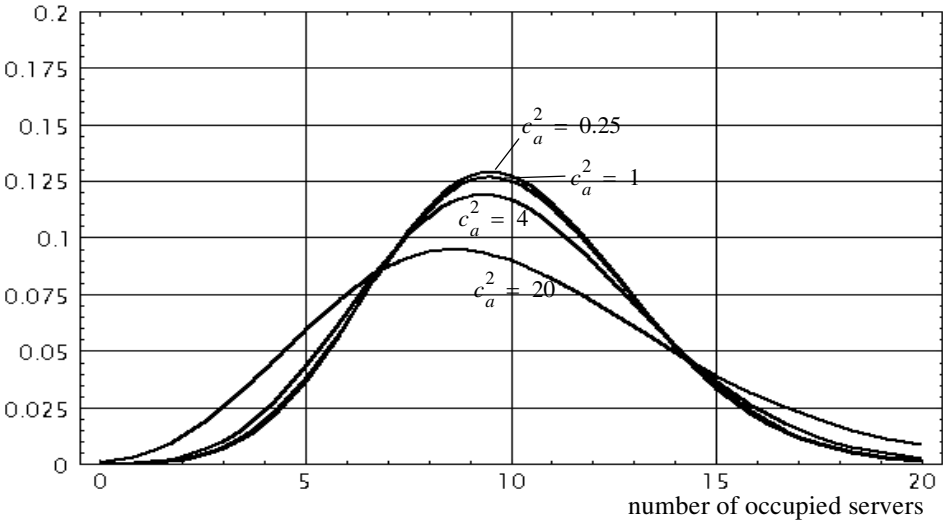

Figure 4.5: Occupancy Distributions with High Holding Time Variability (Hyperexponential with $c_a^2 = 20$)

4.4.2 Occupancy Peakedness

The results in this section are based on the method of Section 4.3 and the parameters of the numerical examples (mean holding time, etc.) are the same as in the previous section [76].

In Fig. 4.6 the peakedness as a function of the squared coefficient of variation of the holding time distribution is shown for eight different phase type renewal arrival processes ranging from an Erlang-4 to a hyperexponential with squared coefficient of variation equal to 20. For smooth arrival processes the peakedness increases with increasing holding time variability, which is expected. For peaky arrival processes the peakedness decreases with increasing holding time variability. The reason is that when the holding time is regular the number of occupied servers follows the peaky nature of the arrival process, and when the variability of the holding time increases it randomizes the number of occupied servers and a decrease in peakedness follows.

It can also be seen that the peakedness is very holding time dependent for arrival processes with very high variability, and the sensitivity is largest for holding time distributions with small variability.

The peakedness as a function of the squared coefficient of variation of the interarrival time distribution is shown for eight different holding time distributions in Fig. 4.7. Fig. 4.7 shows that the peakedness increases almost linear with the squared coefficient of variation of the interarrival time. The relative difference from a first order Taylor approximations is always below 6% in the investigated range. The slope of the lines increases with decreasing holding time variability. In cases with a highly variable holding time distributions ($c_h^2 > 15$) the peakedness is almost independent of the variability of the arrival process.

The results show a quite unexpected and interesting property: even for very bursty arrival processes ($c_a^2 = 20$) the occupancy distribution has small variability (the peakedness is $Z < 2$) in case of bursty service processes ($c_h^2 = 20$). This smoothing effect of the holding time also indicates that the variability of the holding time is a significant characteristics of the system.

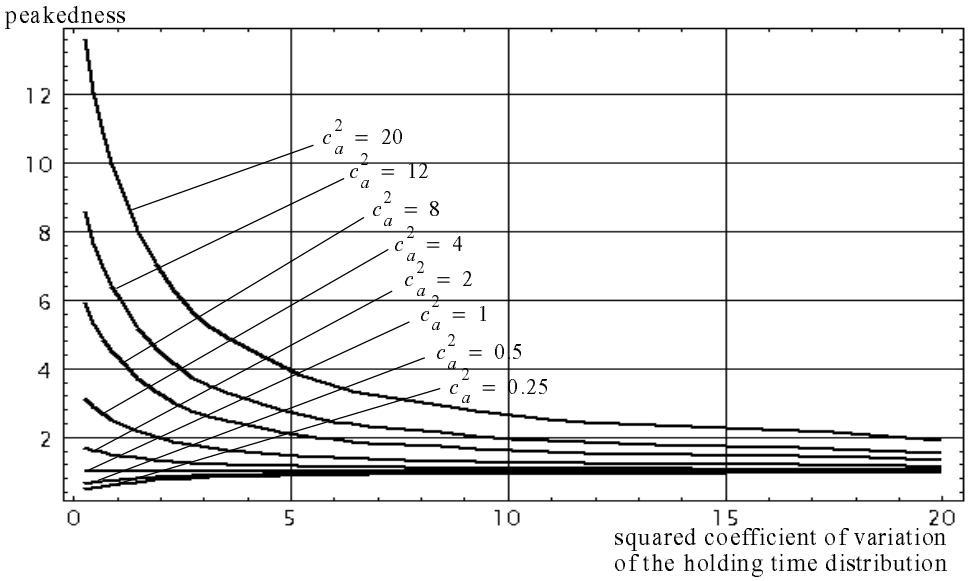

Figure 4.6: Peakedness of the Occupancy Distribution as a Function of the Squared Coefficient of Variation of the Holding Time

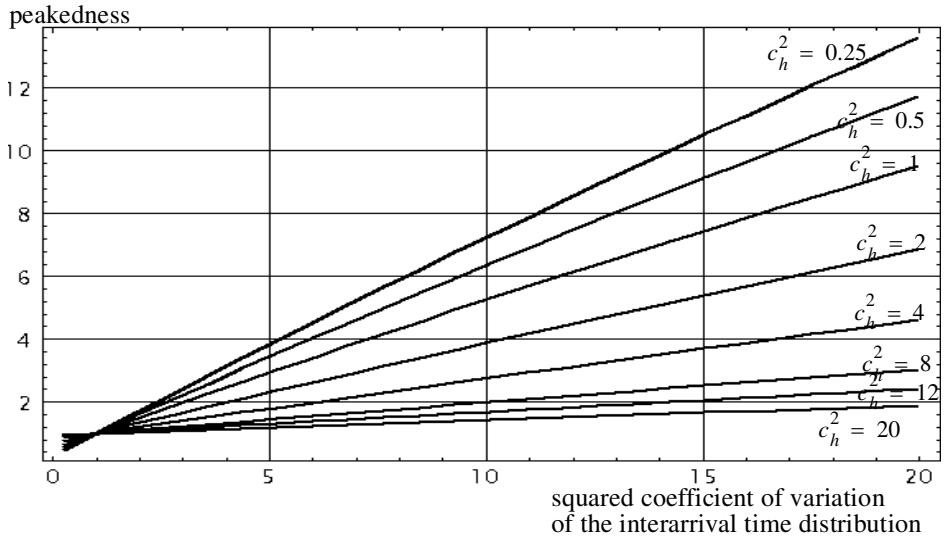

Figure 4.7: Peakedness of the Occupancy Distribution as a Function of the Squared Coefficient of Variation of the Interarrival Time

4.5 Summary of Results

In this Chapter a robustness and sensitivity analysis of link occupancy with respect to the arrival and the service processes is presented. Based on the results I have concluded some basic and interesting properties of both link occupancy distribution and occupancy peakedness.

An important practical conclusion can be established from these results: *the traditional Poisson/Exponential description of B-ISDN is quite vulnerable to deviations from these classical assumptions*, thereby if these assumptions are not fulfilled the Poisson process and the exponential distribution, which are widely applied in traditional telephone networks, cannot be accepted in B-ISDN environment for modeling the arrival process and the holding time, respectively.

It means that an accurate characterization of both processes are necessary. It can also be noted that the holding time distribution should be correctly taken into account for developing blocking probability measures etc. because if the arrival process differ from Poisson the nice property that the occupancy distribution depends only on the mean of the holding time does not hold.

A traffic variability measure based on the concept of the generalized peakedness with a new formula is also presented.

Chapter 5

Approximations for Link Occupancy Distributions and Link Blocking Measures

5.1 Introduction

One of the most important NP characteristics of the call level is the end-to-end blocking [15]. It is defined as the probability of a call being blocked from all the routes that it can use to travel from its origination node to its destination node. In general the calculation of blocking probabilities in such networks is very difficult and in practice it is often based on decomposition into series of fixed routing problems with link-by-link decompositions [25]. Therefore the key element of any end-to-end blocking computation procedure is the proper link blocking computation.

Even considering only a single link the calculation of blocking probability is not trivial and differs significantly from the traditional circuit-switching networks [15]. In the traditional telephone networks the estimation of blocking probabilities has a long tradition initiated with the pioneering work of Erlang [12]. Many years of experience has shown that the Poisson process is a reasonable model of the process of telephone call requests, and its nice mathematical properties have to a great extend simplified the analysis of such networks resulting in properties such as product form solutions and insensitivity of holding time distributions [56].

In order to analyze overflow traffic in e.g alternative routing, it has been necessary to include less tractable arrival processes like *renewal*, or *Markov Modulated Poisson Processes*. For these cases it has normally been assumed that the holding time distribution is exponential [101, 23]. Alternatively, the class of state dependent Poisson processes has been used for the arrival process. In B-ISDN the characterization of arrival process and holding time will likely be different from the classical description (see the previous Chapter) but in this case the computation of the link occupancy distribution and thereby the link blocking measures are rather complicated.

The fact that it is relatively easy to obtain the mean and variance of the distribution,

but difficult to obtain the distribution itself, calls for approximations. In this Chapter two approximations for the occupancy distributions are presented based on matching the mean and variance. The issue of how the third and higher moments affects the occupancy distribution is not investigated in this Chapter but I refer to the work by Holtzman [45] in which some bounds have been derived based on the Laplace transform of the occupancy distribution.

The first approximation is a BPP approximation while the second one is the distribution which is obtained by maximizing the entropy subject to matching of the mean and variance. It is shown how the maximum entropy method enforces a distribution which is a discrete version of a normal density function restricted to the positive line. Finally, approximations for the blocking probability of the finite capacity system is suggested based on the derived occupancy distribution of the infinite system.

5.2 The BPP Approximation

The BPP (Bernoulli-Poisson-Pascal) arrival process [23] is a state dependent Poisson process characterized by two parameters α and β such that the Poissonian arrival intensity when k servers are occupied is $\alpha + k\beta$ (see Figure 5.1). In the case $\beta = 0$ it reduces to

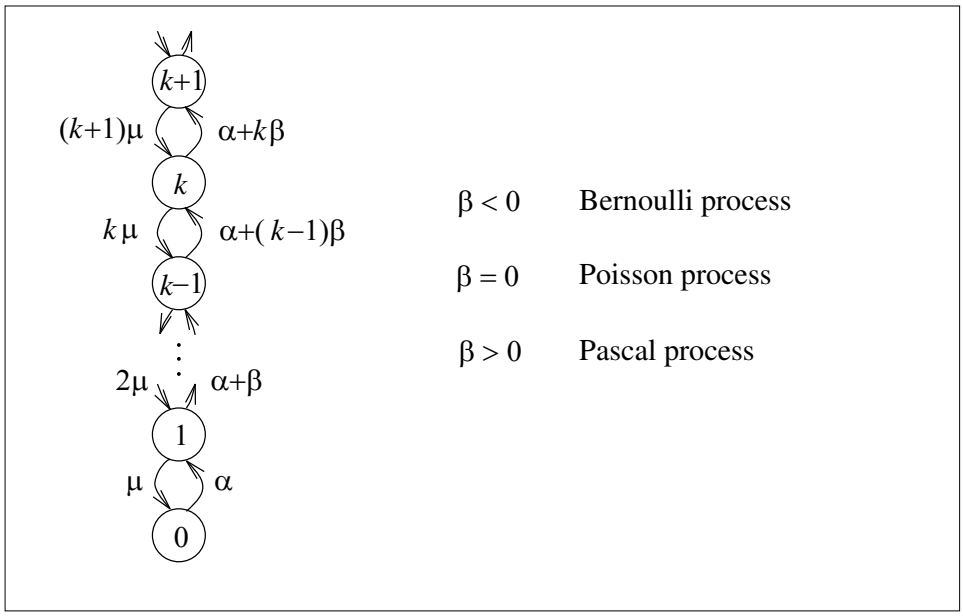

Figure 5.1: The BPP Process

the plain Poisson process, for $\beta < 0$ it represents a process of less variability than Poisson (finite source model), and for $\beta > 0$ it represents a process of higher variability than Poisson [23].

The mean and variance of the occupancy distribution turns out to be:

$$M = \frac{\alpha}{\mu - \beta} \quad V = \frac{\alpha\mu}{(\mu - \beta)^2} \tag{5.1}$$

where $1/\mu$ is the mean holding time. *I suggest to approximate the occupancy distribution with the BPP distribution with the mean M and variance V obtained by the concept of the generalized peakedness* presented in Chapter 4.3 [76, 80, 79, 83, 77]. This approximation has the advantage that it takes into account both the arrival and service processes and in the approximation the exact mean and variance of occupancy distribution is used. Thereby this approximation produces more accurate results (see numerical examples) compared to other usually applied methods which uses only the characterization of the arrival process (e.g. the BPP approximation in [23]). With the mean and the variance the parameters α and β should be chosen as:

$$\beta = 1 - \frac{1}{Z} \quad \alpha = M(1 - \beta) \tag{5.2}$$

assuming a mean holding time of 1. The BPP distribution can now be obtained by

$$p_i = \begin{cases} p_{i-1}\dfrac{\alpha + (i-1)\beta}{i} & \text{if } i > 0 \\ 0 & \text{if } \beta < 0 \text{ and } i > N \end{cases} \tag{5.3}$$

and

$$p_0 = \begin{cases} (1-\beta)^{\frac{\alpha}{\beta}} & \text{if } \beta \neq 0 \\ e^{-\alpha} & \text{if } \beta = 0 \end{cases} \tag{5.4}$$

where in the case of $\beta < 0$ the β is modified to the closest value such that $N = -\frac{\alpha}{\beta}$ must be an integer.

Finally I suggest to truncate and renormalize the infinite BPP distribution in order to obtain the link occupancy distribution.

$$p_i^{tr} = \begin{cases} p_i & \text{if } \beta < 0 \text{ and } c > N \\ \dfrac{p_i}{\sum_{j=0}^{c} p_j} & \text{otherwise} \end{cases} \tag{5.5}$$

where c is the capacity of the link.

5.3 The Maximum Entropy Approximation

The concept of entropy appears in the mathematical theory of interconnecting networks and in the queueing theory [6, 65]. In queueing theory its existence is known only in the 1-server case and usually it has been applied in a way where only average quantities like mean queue length and utilization has been matched. In this Section an application of this

technique on a many server loss system is shown where also the variance is matched.

The basic idea of the method is based on Bernoulli's principle of insufficient reason [41], which states that all events over a sample space should have the same probability unless there is evidence to the contrary. The entropy plays as a measure of the certainness of the outcome of an event. The more uncertain the value of a random variable is, the bigger the entropy is. In order to fulfill the Bernoulli's principle the entropy has to be maximized under the constrains of the mean and the variance which we would like to be matched.

Consider a stationary stochastic process $X(t)$ in a discrete state space and let p_i be the probability of being in state i. The entropy of $X(t)$ with stationary distribution $\{pi\}$ is defined as:

$$H(\boldsymbol{p}) = -\sum_i p_i ln p_i \tag{5.6}$$

The idea here is, as an approximation for the occupancy distribution, to take the one which maximizes the entropy under the constrains that

- *it should be a proper probability distribution i.e.* $\sum_{i=0}^{\infty} p_i = 1$
- *the mean should be correct i.e.* $\sum_{i=0}^{\infty} i p_i = E(X)$
- *the second moment should be correct i.e.* $\sum_{i=0}^{\infty} i^2 p_i = E(X^2)$

and I suggest to use the concept of generalized peakedness to get the variance of the occupancy distribution [76, 80, 79, 83, 77]. Similarly to the proposed approximation in Section 5.2 this method also uses both the arrival and service process characteristics. Moreover, it utilizes the information of the two-parameter matching but also ensures that the approximate distribution will be maximum uniform. In Chapter 8.4.1 of [41] the following theorem is presented and proved:

THEOREM: The probability mass function $\{pi\}$ which maximizes

$$H(\boldsymbol{p}) = -\sum_i p_i ln p_i \tag{5.7}$$

subject to

$$\sum_i^{\infty} p_i = 1 \quad \text{and} \quad \sum_i f_j(i) p_i = \bar{f}_j \quad \text{for } 1 \le j \le k \tag{5.8}$$

(where $\{\bar{f}_j | (1 \le j \le k)\}$ are prescribed mean values of functions $\{f_j\}$) is:

$$p_i = g \prod_{j=1}^{m} x_j^{f_j(i)} \tag{5.9}$$

where g is the normalization constant.

If this result is applied to a single server queue and the mean queue length is matched, the queue length distribution which maximizes entropy is geometrical thus yielding the exact distribution for the queue length in the $M/M/1$ case.

When matching the first and second moment in the infinite server case, a straightforward application of the theorem shows that the occupancy distribution which maximizes entropy is:

$$p_i = P\{X = i\} = g x_1^i x_2^{i^2} = g e^{i l n x_1} e^{i^2 l n x_2} \tag{5.10}$$

Thus the maximum entropy approach enforces a distribution which comes from sampling the normal density function at non-negative integer values. The three equations needed to match the mean, variance and obtain a proper distribution is:

$$\sum_{i=0}^{\infty} g x_1^i x_2^{i^2} = 1 \quad \sum_{i=0}^{\infty} g i x_1^i x_2^{i^2} = E\{X\} \quad \sum_{i=0}^{\infty} g i^2 x_1^i x_2^{i^2} = E\{X^2\} \tag{5.11}$$

The equations has been solved by heuristic methods. However, applying the corresponding continuous approach fitting a normal density function restricted to the positive line yields a system of equation which can be solved in an exact way as demonstrated in [83]. The numerical results indicates that the difference between the continuous and the discrete results are very small. Here I also suggest to truncate and renormalize the infinite distribution in order to get the link occupancy distribution.

5.4 Link Blocking Measures

In this Section some practically applicable and accurate link blocking measures are presented based on different approximate link occupancy distributions [76, 80, 79, 83, 77]. It is possible to construct different types of blocking measures like time congestion, call congestion or traffic congestion and the usage of them can be chosen from the application point of view. In this Section the emphasis is on the traffic congestion measures, which are applied in ATM network dimensioning in Chapter 6, but call and time congestion measures are also possible to define. I use the classical definition of traffic congestion in the proposed link blocking probabilities:

$$TC = \frac{OT - CT}{OT} \tag{5.12}$$

where OT and CT denote the offered traffic and carried traffic, respectively. I suggest to find the carried traffic from the approximate occupancy distributions obtained by truncating a renormalizing the infinite capacity distribution. It should be noted that the truncation and renormalization procedure yields the correct distribution in cases when the system under consideration is a time reversible Markov process (see e.g. corollary 1.10 in [55]), which is the case for e.g. the BPP arrival process.

5.4.1 Traffic Congestion Based on the Exact Infinite Capacity Occupancy Distribution

The exact infinite capacity occupancy distribution can be obtained in case of phase type arrivals by the method presented in Chapter 4. Based on this distribution a traffic congestion measure can be defined after truncation and renormalization of the distribution.

I suggest the following traffic congestion measure:

- Compute the exact occupancy distribution by the method of Section 4.2.1.
- Compute the finite capacity distribution by truncating and renormalizing the infinite capacity exact distribution.
- Compute a Traffic Congestion (TC) by Eq 5.12 where the carried traffic is computed by finding the mean of the approximate finite capacity distribution.

5.4.2 Traffic Congestion Based on the BPP Approximation

By applying the BPP approximation procedure a simple but accurate (see numerical examples) traffic congestion measure can be defined.

I suggest to obtain the BPP traffic congestion by the following procedure:

- Compute the approximate occupancy distribution by the BPP method presented in Section 5.2.
- Compute the finite capacity distribution by truncating and renormalizing the infinite capacity distribution.
- Compute a Traffic Congestion (TC) by Eq 5.12 where the carried traffic is computed by finding the mean of the approximate BPP distribution.

5.4.3 Traffic Congestion Based on the ME Approximation

Based on the ME method I define the following traffic congestion measure:

- Compute the approximate occupancy distribution by the ME method presented in Section 5.3.
- Compute the finite capacity distribution by truncating and renormalizing the infinite capacity distribution.
- Compute a Traffic Congestion (TC) by Eq 5.12 where the carried traffic is computed by finding the mean of the approximate ME distribution.

5.4.4 Traffic Congestion in Case of Renewal Input and Exponential Holding Time

The solution to the $GI/M/c$ and $GI/M/\infty$ systems has been known [94]. Both the occupancy distribution at an arbitrary point in time and the occupancy distribution at an arbitrary arrival can be found (see Section 4.2.2). Even though an analytical solution exists, obtaining numerical values from the solution is challenging and in practice limited to small offered traffic and small value of the capacity.

In the following we show that *the call congestion equals to the traffic congestion in the $GI/M/c$ system therefore through the method of Section 4.2.2 we can get the exact traffic congestion measure.* For the $GI/M/c$ system the following relation between the occupancy distribution p_i at an arbitrary time and the occupancy distribution p_i^a just prior to an arrival exists (see Section 4.2.2):

$$p_i = E\{L\}\frac{p_{i-1}^a}{i}, \quad \text{for } i \in \{1, ..., c\} \quad \text{and} \quad p_0 = 1 - E\{L\}\sum_{i=1}^{c}\frac{p_{i-1}^a}{i} \tag{5.13}$$

$E\{L\}$ is the mean of the number of occupied servers in the corresponding infinite server system. Using Eq. 5.13 the call congestion can be written as

$$\begin{aligned} CC &= 1 - \sum_{k=0}^{c-1} p_k^a = 1 - \sum_{k=0}^{c-1}\frac{p_{k-1}(k+1)}{E\{L\}} = 1 - \frac{\sum_{k=1}^{c} kp_k}{E\{L\}} = \\ &= \frac{E\{L\} - E\{L_{tr}\}}{E\{L\}} = \frac{OT - CT}{OT} = TC \end{aligned} \tag{5.14}$$

Here $E\{L_{tr}\}$ is the mean of the number of occupied servers in the finite server system. This calculation shows that the call congestion equals to the traffic congestion thereby the method of Section 4.2.2 is appropriate to compute TC.

5.4.5 Call Congestion Based on the Delbrouck Method

Delbrouck presented an approximate call congestion formula in [23]:

$$CC = TIC(1 + \frac{c}{OT}(Z - 1)) \tag{5.15}$$

where CC and TIC are the call and time congestions, respectively. c denotes the capacity of the link and Z is the peakedness.

I suggest to use the concept of the generalized peakedness to find Z in Eq. 5.15 for computing call congestion.

5.5 Numerical Results

In this section the evaluation results of both link occupancy approximations and link blocking measures are shown. In the evaluation study the methods are investigated different arrival and service processes. For evaluating the results of the BPP approximation the results of the classical BPP method is also given for comparison. In the evaluation of the blocking measures arrival processes with not only different burstiness but with the same burstiness and different distributions are also investigated.

5.5.1 The Occupancy Distributions

In the numerical study the same examples have been chosen as in Section 4.4 where the mean occupancy is 10 by setting the arrival intensity $m = 10$ and mean holding time $1/\mu = 1$. Since the maximum entropy essentially gives a normal distribution, the approximations was of interest to evaluate at systems with a higher occupancy. Therefore we have considered the case with exponential holding time and a mean occupancy of 50 servers also, and solved the traditional steady state equations of this system. The occupancy distributions are evaluated for both smooth and bursty arrival and service processes to identify the most important characteristics.

In Figure 5.2-Figure 5.4 the BPP and ME approximations are evaluated for the case with an Erlang-4 arrival process and three different holding time distributions. In order to investigate the effect of taking into account only the arrival process the classical BPP approximation is also plotted (denoted by BPPI) where the holding time distribution is assumed to be exponential. The BPP approximation presented in Section 5.2 denoted by BPPII. As expected then the BPPI approximation which does not take into account the holding time distribution is unacceptable. However, the BPPII which matches the mean and variance of the occupancy distribution is very accurate while the maximum entropy distribution is a little inaccurate mainly because the maximum of the distribution is forced to be very close to the mean since the restriction to the positive line changes almost nothing compared to an ordinary normal distribution.

In Figure 5.5-Figure 5.7 the approximations are evaluated for the case with a peaky non balanced hyperexponential arrival process. In Figure 5.5 with an Erlang-4 holding time distribution none of the approximations are able to capture the "knee" of the exact distribution. When the holding time distribution is exponential the maximum entropy comes out as the most accurate while for a hyperexponential holding time distribution again the BPPII is very accurate.

Figure 5.8-Figure 5.9 show how the accuracy of the ME approximation increases when the mean occupancy of the system increases.

Finally, in Figure 5.10-Figure 5.12 the BPP and ME approximations are evaluated on the system with a mean occupancy of 50 exponential servers. For the case with Erlang-4 arrivals the accuracy is very good while in the hyperexponential case the fact that the maximum of occupancy distribution is shifted away from the mean is not captured by the truncated normal distribution dictated from the maximum entropy method.

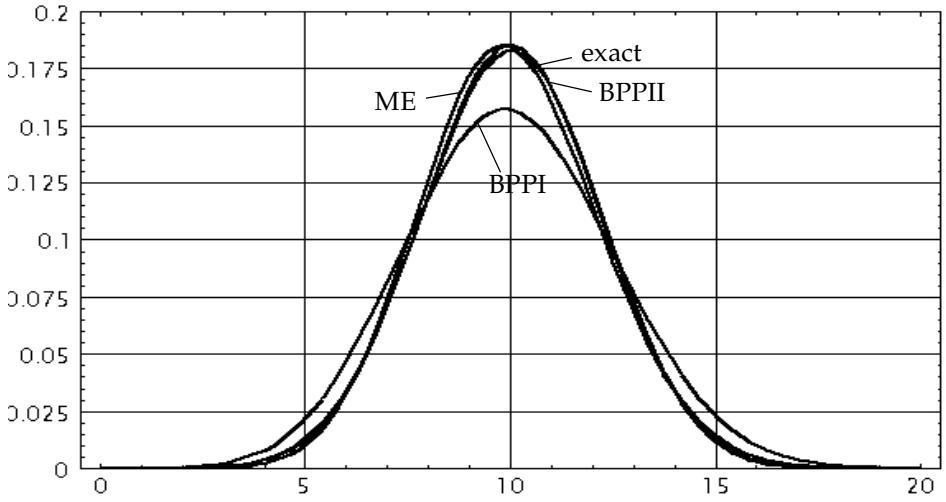

Figure 5.2: Occupancy Distributions with Erlang-4 Interarrival and Holding Time Distributions

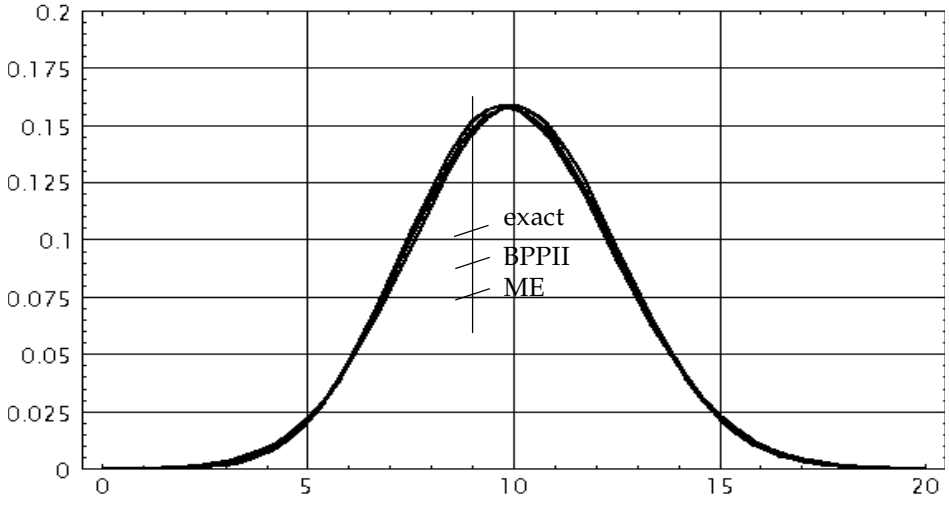

Figure 5.3: Occupancy Distributions with Erlang-4 Interarrival and Exponential Holding Time Distributions

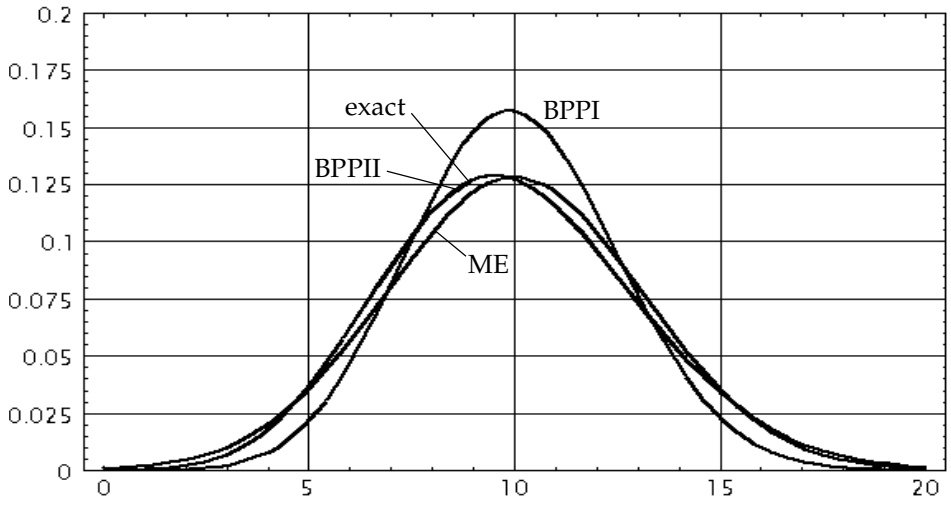

Figure 5.4: Occupancy Distributions with Erlang-4 Interarrival and Hyperexponential ($c_h^2 = 20$) Holding Time Distributions

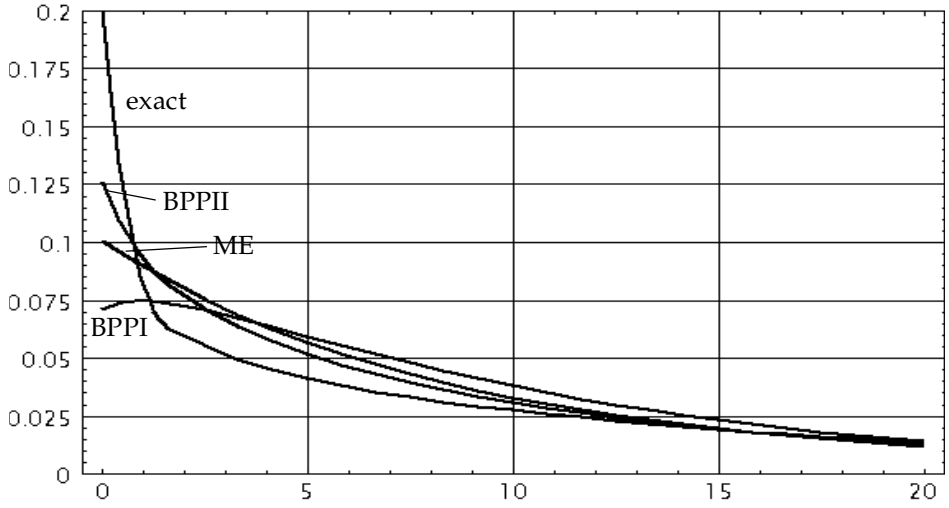

Figure 5.5: Occupancy Distributions with Hyperexponential ($c_h^2 = 20$) Interarrival and Erlang-4 Holding Time Distributions

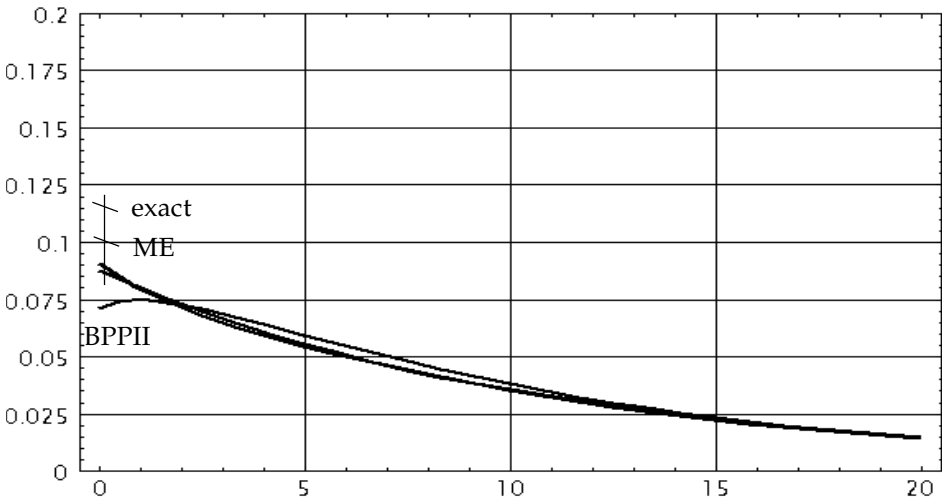

Figure 5.6: Occupancy Distributions with Hyperexponential ($c_h^2 = 20$) Interarrival and Exponential Holding Time Distributions

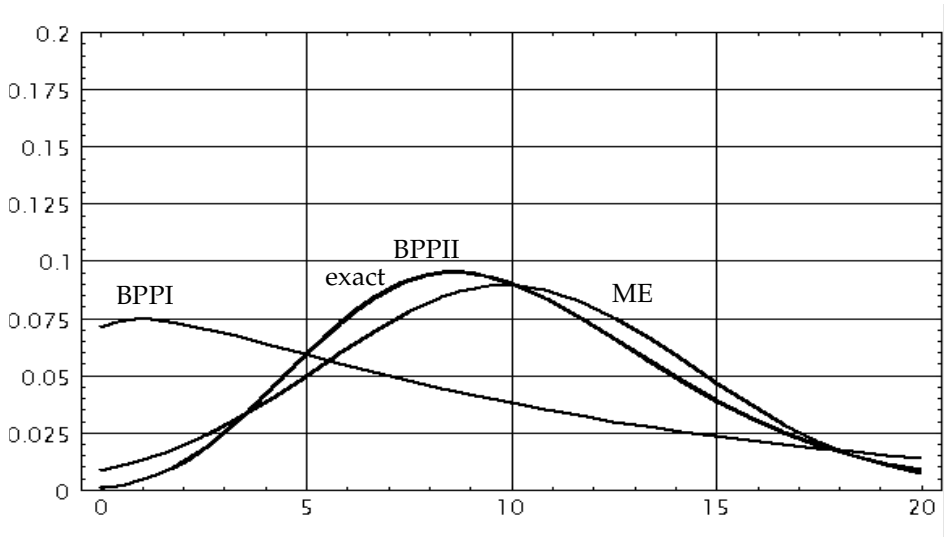

Figure 5.7: Occupancy Distributions with Hyperexponential ($c_h^2 = 20$) Interarrival and Hyperexponential ($c_h^2 = 20$) Holding Time Distributions

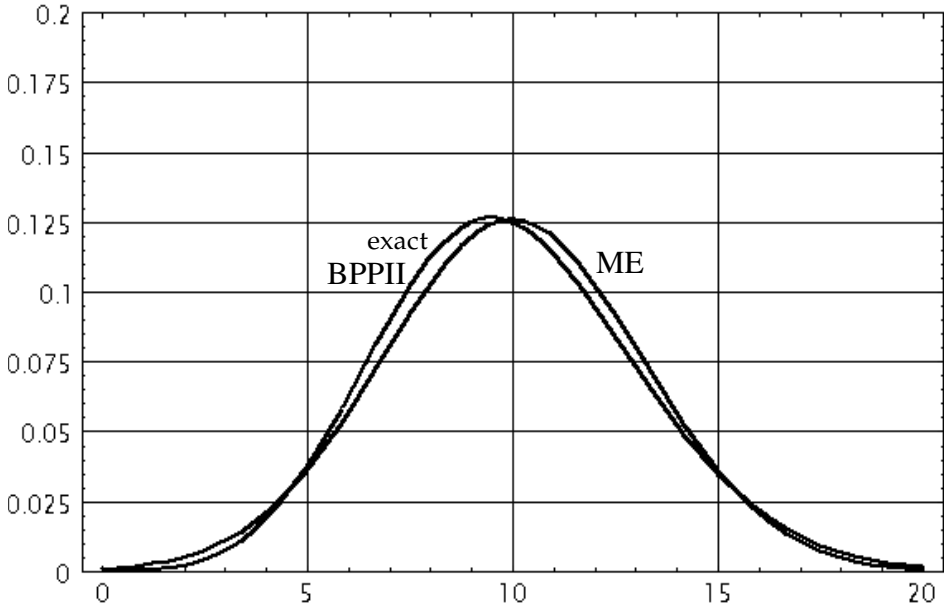

Figure 5.8: Occupancy Distributions with Poisson Arrivals (mean occupancy=10)

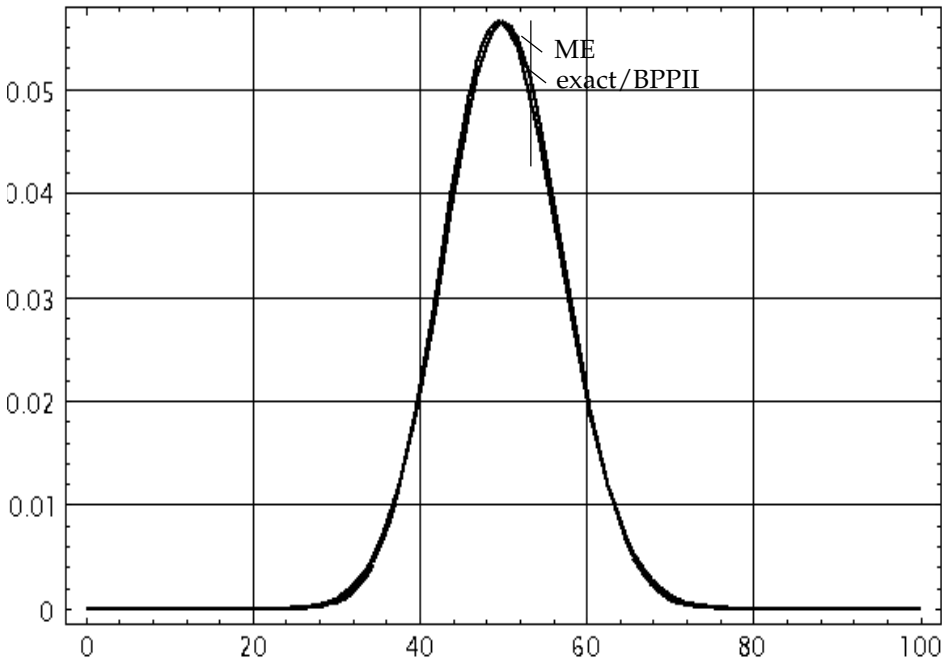

Figure 5.9: Occupancy Distributions with Poisson Arrivals (mean occupancy=50)

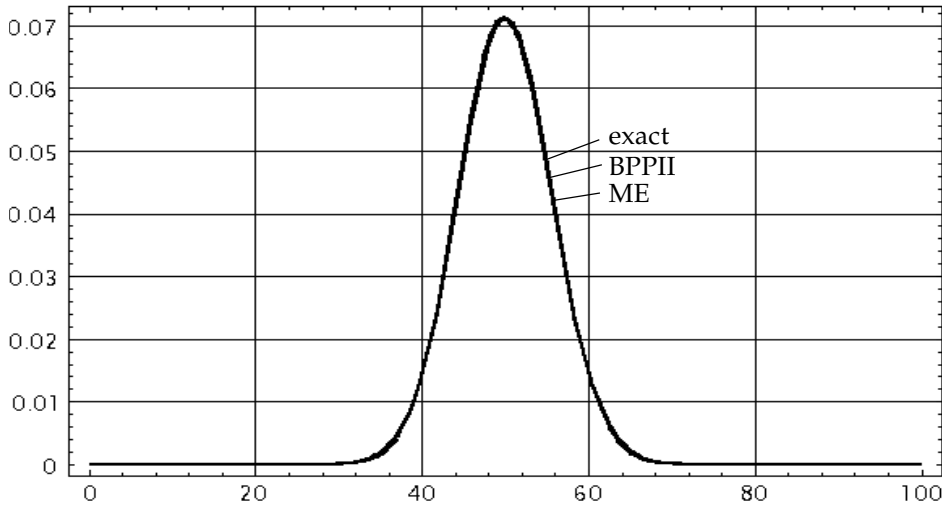

Figure 5.10: Occupancy Distributions with Erlang-4 Interarrival and Exponential Holding Time Distributions

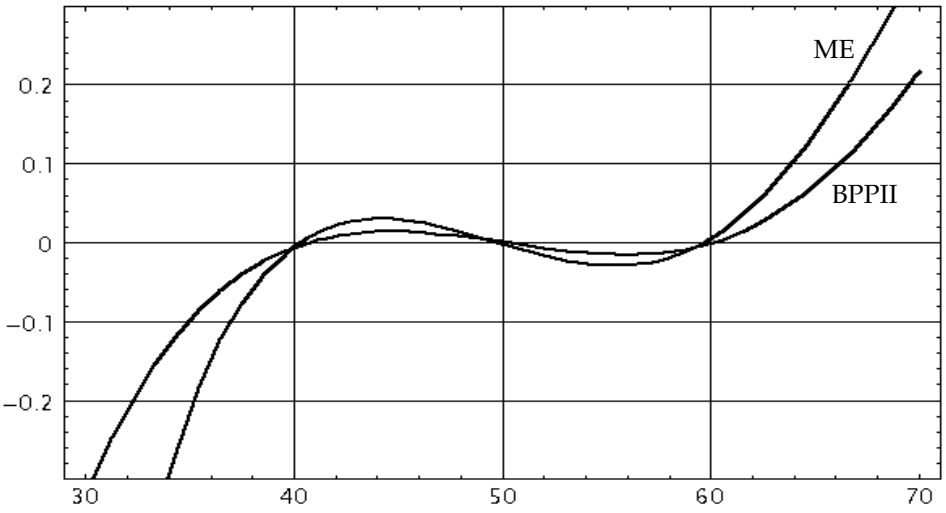

Figure 5.11: Relative Error of the Approximations of Case Figure 5.10

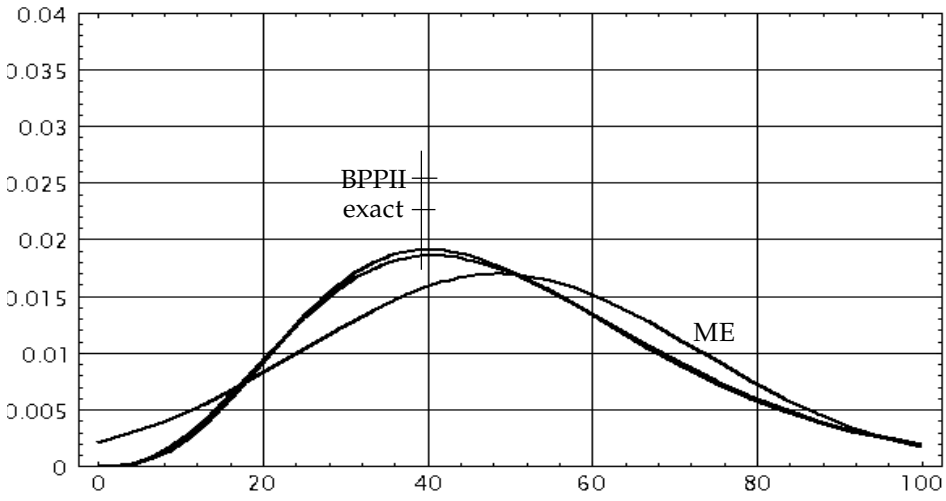

Figure 5.12: Occupancy Distributions with Hyperexponential ($c_h^2 = 20$) Interarrival and Exponential Holding Time Distributions

5.5.2 The Link Blocking Measures

We have found the exact blocking probabilities of the $GI/M/c$ case for 8 different interarrival time distributions. Varying the squared coefficient of variation of the interarrival time (c_a^2) from 1/3 to 9 and also varying the type of the distributions we have investigated a broad range of the arrival distributions to evaluate the approximate blocking probabilities. For smooth arrival processes the c_a^2 of 1/3 with Erlang-3 distribution and the convolution of the exponential and the deterministic distributions (M and D) are considered. In the $c_a^2 = 1$ case not only the Poisson arrival process (M), but also a Coxian distribution (Cox) of two branches with Erlang-2 and Erlang-3 distributions are investigated. For peaky arrival processes the c_a^2 of 3 and 9 cases with Hyperexponential balanced (Hyp-bal) and unbalanced (Hyp-unbal) distributions are evaluated. In all cases the offered traffic is kept equal to 10 and the number of servers is 15. Then we have evaluated four different approximative blocking probabilities.

Column 2, 3, 4, 5 and 6 are the blocking probabilities described in the Section 5.4. To get exact results the method of Section 5.4.4 has been applied. "Truncated" denotes the measure of Section 5.4.1. In the column "BPP" the traffic congestions of the BPP method (Section 5.4.2) is shown which in these cases equals to the call congestions. The traffic congestions based on the truncated normal distribution (Section 5.4.3) are shown in the column "Tr. normal".

c_a^2	**Type**	**Exact**	**Truncated**	**BPP**	**Tr. normal**
1/3	Erlang-3	1.54%	1.49%	1.30%	1.67%
1/3	M and D	1.45%	1.39%	1.29%	1.65%
1	M	3.65%	3.65%	3.65%	3.88 %
1	Cox	3.72%	3.73%	3.65%	3.88%
3	Hyp-bal	7.99%	9.27%	9.86%	10.36%
3	Hyp-unbal	10.53%	10.90%	10.86%	11.51%
9	Hyp-bal	15.38%	21.86%	20.71%	23.48%
9	Hyp-unbal	25.52%	26.64%	26.18%	29.91%

Table 5.1: Approximate and Exact Blocking Probabilities (offered traffic=10, capacity=15)

The BPP approximation underestimates in the cases with regular arrival processes while the truncated normal distribution overestimates for peaky arrival processes. We can conclude from the Table 5.1 that all the approximations performs reasonable well and provide satisfactory measures of link blocking taking into account the simplicity of the methods and no better similarly simple approximation is known.

5.6 Summary of Results

This Chapter has introduced two new approximate methods for computing the link occupancy in case of general arrival and service processes with their performance evaluation. The first method is a BPP approximation while the second one is an entropy maximization procedure. The methods are based on matching the mean and the variance of the exact distribution and for the variance computation the concept of generalized peakedness is suggested. From the evaluation results it can be seen that these approximations show a better performance than e.g. the classical BPP approximation which does not take into account the holding time. Thereby we can conclude that the exponential assumption of holding time can lead to misleading results and the usage of generalized peakedness is justified.

Based on the approximations some link blocking measures have been developed and evaluated. These measures are simple to compute, therefore good candidates for practical applications, but on the other hand they can cope with other types of traffic than Poisson. An application of the proposed measures in network dimensioning procedures is shown in the following Chapter.

Chapter 6

ATM Network Dimensioning

6.1 Introduction

One of the important issues of the configuration and management of large ATM networks is how to partition the network into a number of logical (virtual) subnetworks that share the capacity of the same physical network [29]. In the the network design and dimensioning of traditional telephone networks several methods have been developed [32, 57] but there are more reasons (e.g. the different nature of B-ISDN traffic, more complex routing functions, more flexible reconfiguration facilities, etc.) that make the applicability of traditional methods very restricted in ATM networks.

Such partitioning is required because of various service classes in a B-ISDN environment have very different characteristics and they also demand very different control mechanisms (e.g. delay sensitive traffic like voice communication and loss sensitive traffic like data communication) and their management is much easier if traffic classes having similar nature are grouped into logical subnetworks. Moreover, large business users may require virtual leased networks with a guaranteed grade of service which can be safely realized by logical resource separation. Furthermore, the logical subnetworks makes different call level management possible for different traffic classes which is combined with cell level management like priority queueing provide us with an effective framework of B-ISDN configuration and management.

In ATM network dimensioning we have some given logical subnetworks and traffic demand for each route, where we want to find partition of the physical capacities related to subnetworks that maximizes the total carried traffic or revenue. The characterization of traffic at call level should be accurate enough to provide the designer with reliable tools for dimensioning the transmission and switching capacities.

The experience with telephone traffic is that the Poisson process which is described by a single parameter constitutes a natural and accurate model for the arrival of call attempts, and its memoryless property ensures that the so-called insensitivity property i.e. most quantities of interest depends on the distribution of the holding time only through the mean. It is highly unlikely that the Poisson property carries over to many other services in a B-ISDN context. The connection request process for some services may

have a rather regular pattern while for other services it may come out very bursty. The first way to think of to solve this problem is to include in the traffic demand matrix a two parameter description, one parameter for the usual demand and a parameter characterizing the variability of the arrival of connection requests.

This Chapter describes a two-parameter description of traffic and illustrates its applicability for both ATM link and network partitioning. These results clearly indicate the powerful applicability and relevance of the theoretical results of Chapter 4-5. Also a new ATM network dimensioning algorithm is suggested using the two-parameter description of traffic. Finally, a new formula for peakedness calculation in case of load sharing is presented.

6.2 Two-parameter Description of Traffic

In the suggested two-parameter description of traffic one parameter describes the usual mean of the traffic demand and the other parameter characterizes the variability of the traffic demand.

As described in Section 4.3 two measures has been widely used for characterizing the variability of traffic. One of them is the squared coefficient of variation of the interarrival time between two consecutive connection requests [16], which provides a full second order characterization in case of arrival processes well modeled by renewal processes [16], but in the general case this description is not adequate [17].

The generalized peakedness measure (see Section 4.3 for details) provides a complete second order characterization of the arrival process and furthermore also takes the service process into account. *I suggest the generalized peakedness for the variability measure of B-ISDN traffic* [73, 83].

Based on the above two-parameter description of traffic streams the following model of the aggregated multirate traffic is considered: we have N number of independent traffic classes. Let ν_i and z_i denote the offered arrival rate and the generalized peakedness of calls from traffic class i, respectively. So each traffic stream is characterized by (ν_i, z_i). Let A_i denote the bandwidth demand of calls from class i. The variance of offered traffic (ω_i^2) can be computed by $\omega_i^2 = z_i \nu_i$. The mean bandwidth occupancy ν, variance of occupancy ω^2 and the peakedness of the aggregated traffic z in case of infinite capacity link will be

$$\nu = \sum_{i=1}^{N} \nu_i A_i \tag{6.1}$$

$$\omega^2 = \sum_{i=1}^{N} \omega_i^2 A_i^2 \tag{6.2}$$

$$z = \frac{\omega^2}{\nu} \tag{6.3}$$

6.3 Link Partitioning

In this section we consider a link partitioning task using two approximations based on matching the mean (ν) and the variance (ω^2) of the occupancy distribution in the infinite capacity case and computing blocking probabilities derived from the truncated occupancy distributions. The theoretical details of the approximations (BPP and ME) and applied blocking measures are described in Chapter 5 [83].

6.3.1 The Model and the Solution of Link Partitioning

We consider a single broadband transmission link with C capacity which is used by N number of traffic classes. Each traffic stream is characterized by the average arrival rate and peakedness of calls (ν_i, z_i) requiring A_i bandwidth of capacity. As link blocking measure BM I suggest the BPP or ME type measures instead of Erlang formula described in Section 5.4. The task is to find the optimal partition of the link related to traffic classes $(C_1, C_2, ..., C_N)$ that maximizes the approximated total carried traffic [83]:

$$\text{Maximize} \sum_{i=1}^{N} \nu_i A_i (1 - BM_i)^{A_i} \tag{6.4}$$

with $BM_i = BM(\nu_i A_i, \nu_i z_i A_i^2, C_i)$ subject to $\sum_{i=1}^{N} C_i = C$ and $C_i \geq 0$ which is a simple linear programming problem.

6.3.2 Numerical Example

Here we consider a simple numerical example which demonstrates the effect of using the variability measure and illustrates the link partitioning task.

Consider a 150 Mbit/s link which is loaded to 140 Mbit/s and carries two different traffic types: Traffic type 1 is 2 Mbit/s circuit emulation with peakedness of 0.25, and the Traffic type 2 is 2 Mbit/s frame relay with peakedness of 15. Consider a situation where 60 Mbit/s traffic offered to the link from both traffic types and we would like to share the capacity of the link to the two traffic streams such that the total carried traffic will be maximum.

If we do not use any variability measure we are restricted to share the capacity only based on the offered traffic and using e.g. Erlang formula to compute the blocking probabilities. This way we get the equally partition solution: 70-70 Mbit/s. By using the peakedness as a variability measure and computing the blocking probabilities based on the above described methods by the BPP and the Maximum Entropy approximations and partition the capacity such that the total carried traffic is maximized we get the following results (Figure 6.1):

From the result we can see that the bursty traffic (Traffic type 2) requires a bigger capacity and the smooth traffic (Traffic type 1) requires smaller capacity compared to the

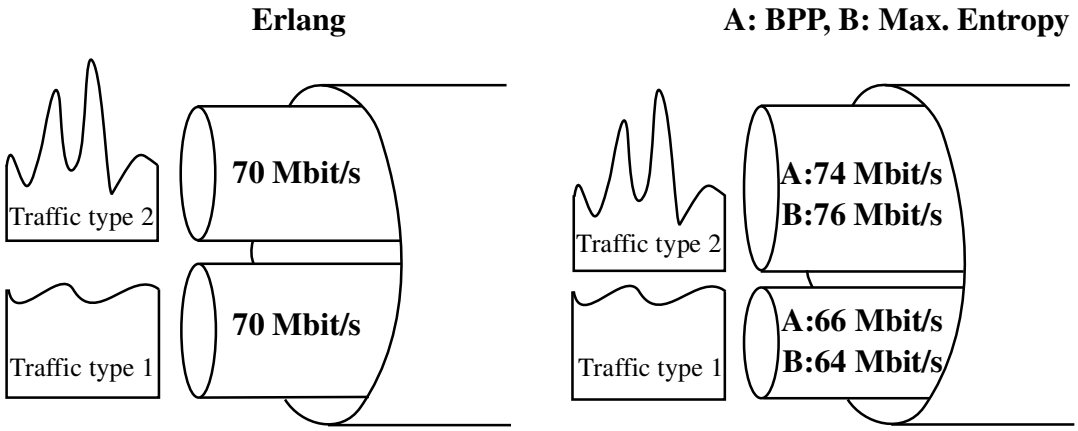

Figure 6.1: Link Partitioning

equally partitioning case. Also we can conclude that the BPP and Maximum Entropy methods give practically the same results.

The results clearly illustrates that in order to achieve the optimum capacity sharing related to the maximum total carried traffic we need to take into account the variability measure.

6.4 Network Partitioning

In this section the effect of the variability measure in an ATM network configuration problem is demonstrated and a new ATM dimensioning algorithm is proposed [73, 83].

6.4.1 The Model

We consider a loss network with J logical links, labeled $j = 1, 2, ..., J$ operating under fixed routing. Let R be the set of routes in the network and let $r \in R$ be a specific route. Let C_{phys} be the vector of given physical capacities and let $C = (C_1, C_2, ..., C_J)$ be the vector of logical link capacities. Let S be a matrix in which the j^{th} entry in the i^{th} row is 1 if logical link j needs capacity on the i^{th} physical link, otherwise 0. Then the condition that the sum of logical link capacities on the same physical link cannot exceed the physical capacity can be expressed by $SC \leq C_{phys}$. A call on route r has ν_r arrival rate and ω_r^2 variance with A_{jr} units of capacity requirements on logical link j. The variance can be obtained by using the peakedness z_r of the traffic on route r: $\omega_r^2 = z_r \nu_r$.

6.4.2 The Algorithm

Consider an ATM network, as described by the model in the previous section, with some given logical subnetworks and traffic demand for each route, where we want to find the partition of the physical capacities related to the subnetworks that maximizes the total carried traffic. For this problem a solution can be found in [29] which is based on the Erlang fixpoint method [57].

Now we consider an extension of this network dimensioning algorithm with using the variability measure. The main purpose of this extension is to improve the network dimensioning algorithm to fulfill the expected nature of the future ATM, where the call arrival process will differ significantly from the Poisson process and the holding time distribution will be deviate from the exponential distribution.

For dimensioning purposes the traffic offered to route r is characterized by the mean ν_r and the peakedness z_r. Based on these two parameters we are using the proposed blocking measures (BPP and ME measures, see Section 5.4) instead of the Erlang formula to compute link blocking probabilities on link j in the dimensioning algorithm : $B_j = BM(\rho_j, \sigma_j^2, C_j)$, where BM denotes a BPP or ME type blocking measure.

For the computation of the aggregated offered traffic to logical link j we use the same reduced load and link independence assumption as in [29] and so:

$$\rho_j := (1 - B_j)^{-1} \sum_r A_{jr} \nu_r \prod_i (1 - B_i)^{A_{ir}}$$

For the calculation of the variance of the aggregated offered traffic to logical link j we simple assume that the variance of the offered traffic is thinned by the same factor as the mean. It means that we keep the peakedness at the same value. However, it should be noted that the changing of the peakedness of the traffic stream going through the network is affected by the congestions on each link and very dependent on the burstiness of the offered traffic. This effect is rather complex and we use this simple approach as a first approximation for the changing of the peakedness. Therefore the variance of the aggregated offered traffic to logical link j can be computed by

$$\sigma_j^2 := (1 - B_j)^{-1} \sum_r A_{jr}^2 \omega_r^2 \prod_i (1 - B_i)^{A_{ir}}$$

Now based on the Erlang fixed point method [29, 57] and the above described considerations I propose the following new network dimensioning algorithm:

1. Set $B_j := 0, \;\; a_j := 1$ for each j.

2. Solve the linear programming problem

$$\text{Maximize} \quad \sum_j a_j C_j$$

$$\text{Subject to} \quad SC \leq C_{phys} \;\text{ and }\; C \geq 0.$$

3. Compute new values for the mean and the variance of the aggregated offered traffic to logical link j by

$$\rho_j := (1 - B_j)^{-1} \sum_r A_{jr} \nu_r \prod_i (1 - B_i)^{A_{ir}}$$

and

$$\sigma_j^2 := (1 - B_j)^{-1} \sum_r A_{jr}^2 \omega_r^2 \prod_i (1 - B_i)^{A_{ir}}$$

4. Compute new values for the blocking probabilities by

$$B_j := BM\left(\rho_j, \sigma_j^2, \tilde{C}_j\right)$$

where $\tilde{C}_1, \ldots, \tilde{C}_J$ come from Step 2 as a solution of the linear programming problem.

5. Set

$$a_j := -\log(1 - B_j), \qquad j = 1, \ldots, J.$$

6. If all variables differ from their previous value by less then a given error threshold, then stop, else repeat from Step 2.

The algorithm has the following main characteristics:

- The input of the algorithm are the physical capacities of the network and the traffic demand characterized by the mean and the peakedness on each route.
- The output of the algorithm are the logical link capacities on each physical link which defines the logical subnetworks.
- The applied link blocking probabilities are the BPP or ME type measures defined in Section 5.4.
- The basis of the algorithm is the Erlang fixpoint method using the reduced load and link independence assumptions.
- The objective function of the algorithm is the total carried traffic which is maximized by the algorithm.

So far there is no proof for the unique solution of this heuristic algorithm but numerical experiences show that the algorithm finds the solution with quite fast convergence. (It should be noted that even for the original Erlang fixpoint method the unique solution is proved only for the special case if the traffic is homogeneous [56].)

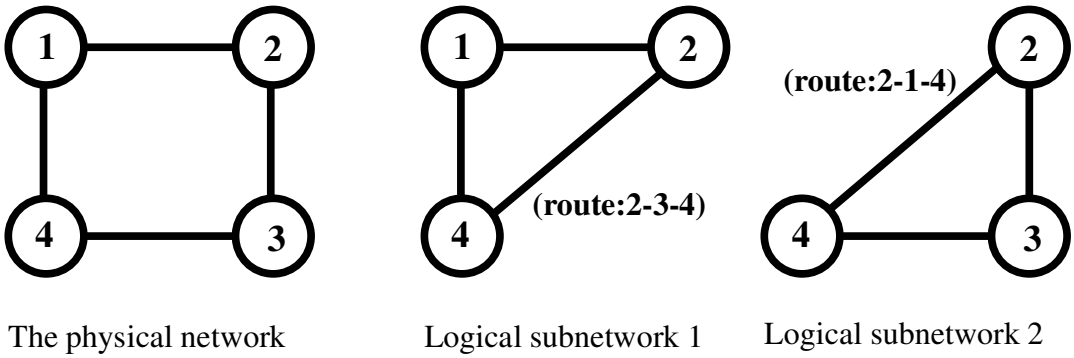

Figure 6.2: A Network Example with Two Logical Networks

6.4.3 Numerical Example

The algorithm is demonstrated in a small network example. In this 4 nodes network arranged in a ring carries two fully connected logical subnetworks with 3 nodes as shown in Figure 6.2.

The matrix S corresponding to the network with routes

$$R = (\{1,2\},\{1,4\},\{2,3,4\},\{2,3\},\{3,4\},\{2,1,4\})$$

is

$$S = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 \end{bmatrix}$$

Each physical link of the network has 45 Mbit/s capacity i.e. $C_{phys} = (45, 45, 45, 45)$ and each logical subnetwork carries two types of traffics:

- Frame relay with effective bandwidth: 0.75 Mbit/s, mean holding time: 60 s
- DS-1 circuit emulation with effective bandwidth: 1.5 Mbit/s, mean holding time: 480 s

The arriving rate of calls from different traffic types are set such that the load be equally shared among the traffic types on a link. The matrix of arriving rate (1/sec) of calls on routes R corresponding to the two traffic types is

$$\nu = \begin{bmatrix} 0.256 & 0.256 & 0.256 & 0.256 & 0.256 & 0.256 \\ 0.016 & 0.016 & 0.016 & 0.016 & 0.016 & 0.016 \end{bmatrix}$$

The partitioning results from the original fixpoint method (which does not take into account any variability measure) and from the extended method (which uses the peakedness of the traffic as described in the previous section, and we used peakedness of 1 to

the logical subnetwork 1 and peakedness of 15 to the logical subnetwork 2) are shown in Table 6.1.

Link	Fixpoint	Fixpoint-BPP
1-2	22.5-22.5	15.8-29.2
2-3	22.5-22.5	13.8-31.2
3-4	22.5-22.5	13.8-31.2
4-1	22.5-22.5	15.8-29.2

Table 6.1: Capacity Partitioning (capacity to logical subnetwork 1 (Mbit/s) - capacity to logical subnetwork 2 (Mbit/s))

The results show that the logical subnetwork 2, which carries rather bursty traffic, requires more capacities (and the logical subnetwork 1, which carries smooth traffic, requires smaller capacities) on the links compared to results of the original fixpoint method which equally partitioned the links between the two logical subnetworks. The optimal partitioning corresponding to the maximum of the carried traffic can be obtained by the above link partitioning and indicates that the traffic variability has influence on the optimal dimensioning which shows the importance of the variability measure.

6.5 Peakedness Calculation in Case of Load Sharing

In network dimensioning it is frequently requested to share the load among several routes between the origin-destination (O-D) pair. The optimization task of the network dimensioning in this case is given jointly with load sharing. The term *load sharing* refers to the fact that the traffic load offered to an O-D pair is shared among a set of allowed routes.

In order to provide the proposed dimensioning algorithm or other algorithms which are using the peakedness for characterizing traffic variability with the load sharing facility the peakedness of the shared traffic on each route must be computed. In this Section a simple formula of the generalized peakedness can be derived for this purpose.

Consider a network where the call process at the origin node is characterized by the average rate ν and the peakedness z. The call holding time is assumed exponential with mean $1/\mu$. Furthermore the call process is described by a renewal process with probability density function (pdf) f and let U denote the renewal function defined by $U(x) = E[N(a, a+x)]$ where $N(a, b)$ denoting the number of arrivals in the interval $]a, b]$.

The traffic is allowed to be shared among K number of routes between an O-D pair and the rule of the sharing is random splitting with parameter p_k on route k. The shared traffic can be characterized by the average rate νp_k and the peakedness z_k (Figure 6.3). Of course the load sharing parameter should satisfy the following constraints for each O-D pair:

$$\sum_{k=1}^{K} p_k = 1 \tag{6.5}$$

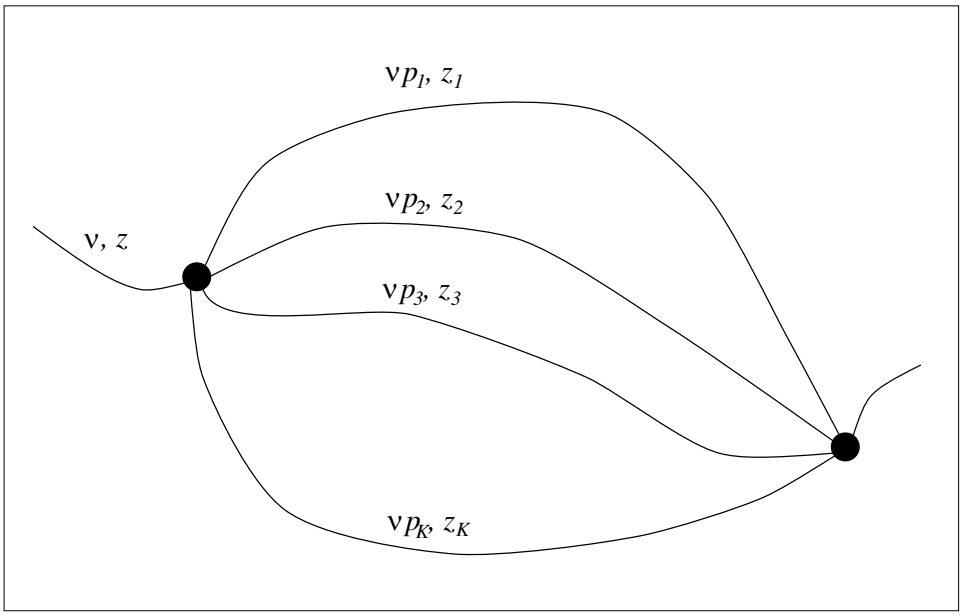

Figure 6.3: The Load Sharing Problem

For an ordinary renewal process the Laplace transform of f can be written as (see e.g. Chapter 4 in [16])

$$\hat{f} = \frac{\hat{U}}{1 + \hat{U}} \tag{6.6}$$

where $\hat{U}$ denote the Laplace-Stieltjes transform of the renewal function U. The Laplace-Stieltjes transform of the renewal function of the call process can be expressed (see Section 4.3) by

$$\hat{U} = z - 1 + \frac{\nu}{\mu} \tag{6.7}$$

The pdf of the interarrival time of the call process shared on route k can be written as

$$f_k = \sum_{n=1}^{\infty} p_k (1 - p_k)^{n-1} f^{*n} \tag{6.8}$$

where f^{*n} denotes the n-fold convolution of f. From Eq. 6.8 the Laplace transform of f_k will be

$$\hat{f}_k = \frac{p_k \hat{f}}{1 - (1 - p_k)\hat{f}} \tag{6.9}$$

Substituting Eq. 6.6 into Eq. 6.9 we will get:

$$\hat{f}_k = \frac{p_k \hat{U}}{1 + p_k \hat{U}} \tag{6.10}$$

Now substituting Eq. 6.10 into the same equation as Eq. 6.6 on the shared traffic:

$$\hat{U}_k = \frac{\hat{f}_k}{1 - \hat{f}_k} \tag{6.11}$$

we get

$$\hat{U}_k = p_k \hat{U} \tag{6.12}$$

The peakedness of the traffic on route k can now be expressed by

$$z_k = 1 + p_k \hat{U} - \frac{p_k \nu}{\mu} \tag{6.13}$$

Finally substituting Eq. 6.7 into Eq. 6.13 we get

$$z_k = 1 + p_k z - p_k \tag{6.14}$$

Based on Eq. 6.14 we can compute the peakedness of the traffic on route k and we have the needed characterization of the traffic on each route k by the average rate $p_k \nu$ and the peakedness z_k.

6.6 Summary of Results

In this Chapter a new approach that characterizes the traffic demand at the call level in a refined way, namely, by using a two-parameter description with the generalized peakedness instead of the traditional one-parameter characterization is presented. This approach contributes to the more accurate description of traffic demands at call level in order to provide the network designer and manager with precision tools to handle traffic demands and their consequences in dimensioning and related issues, while retaining simplicity, algorithmic feasibility and practical applicability.

A link and a network dimensioning problem with a new algorithm are also presented demonstrating the applicability of the two-parameter description of traffic. Results compared to the results based on one-parameter description are given illustrating the importance of the variability measure. Moreover, a new simple formula for the generalized peakedness in case of load sharing is derived which can be applied for upgrading network dimensioning algorithms with load sharing facility.

Part IV

Network Performance Evaluation on Cell Scale

Chapter 7

Performance Evaluation of a Single ATM Multiplexer

7.1 Introduction

Cells may encounter different traffic conditions in buffers of ATM multiplexers, thereby the transfer delay of any cell of a given connection is a random variable. This phenomenon is covered by the term *Cell Delay Variation* (CDV) (for details of CDV parameter definitions see Section 7.2), which is an important NP parameter of ATM networks [14, 100] investigated by the standardization bodies [49, 50] and ATM Forum [2, 3].

A thorough understanding of how CDV arises, how it is affected by traffic from other sources, by its own bitrate, by the number switch buffers and multiplexing stages passed etc., is of great importance concerning Traffic Control (e.g. setting of the Usage Parameter Control (UPC) parameters, designing Call Admission Control (CAC) mechanisms etc.) and Network Element Dimensioning (e.g. dimensioning of buffers of ATM multiplexers and playout buffers at the receiving end of the connection, etc.), see Chapter 9.

In the past few years many investigations have helped towards an understanding of CDV. In [13] a thorough investigation of the $D + GI^{[X]}/D/1$ queue, based on generating functions and the theory of holomorfic function, was performed with the main objective of finding the steady state waiting time at any time including the waiting time seen by the CBR arrivals. In [38] another more direct Markovian approach was applied and the emphasis was on the study of the point process properties of the departure process of the CBR stream. The results made an accurate setting of the UPC parameters possible. By modeling the queueing behaviour between CBR arrivals by a Brownian motion a computationally much simpler model was derived in [10].

The purpose of this Chapter to present new models and results of the performance evaluation of CDV due to a single ATM multiplexer. The application of the results for dimensioning appropriate traffic control functions and network elements are also demonstrated in Chapter 9.

The Chapter is organized as follows. First the definition of CDV parameters are outlined in Section 7.2. An exact Markovian and a diffusion method for modeling the CDV in

a single ATM multiplexer are shown in Section 7.3 and 7.4, respectively. The performance evaluation of both methods is given in Section 7.5. The issue of the superposition of CDV affected CBR cell streams are investigated in Section 7.6 and Section 7.7 summarizes the results of this Chapter.

7.2 Definition of CDV Parameters

Cells experience different delays mainly due to the statistical fluctuations in the queue lengths in ATM networks. This variation in cell delays referred to as *Cell Delay Variation*. In the Recommendation I.356 [49] the cell transfer performance parameters are specified by ITU-T and these CDV parameter definitions are outlined in this section. (It should be noted that beside these ITU-T definitions of CDV several other definitions are also used, e.g. in the RACE Atmospheric project CDV is the variance of the transfer delay of a particular connection.) The results of the thesis mostly related to the 1-point CDV parameter.

The *1-point CDV* y_k for a given cell is defined as the difference between the theoretical time c_k and the actual cell arrival time a_k at the single measuring point, i.e. $y_k = c_k - a_k$. If T denotes the negotiated Peak Emission Interval, the set of theoretical times is recursively defined by

$$\begin{aligned} c_{k+1} &= c_k + T \qquad \text{if } y_k \geq 0 \\ &= a_k + T \qquad \text{if } y_k < 0 \end{aligned} \tag{7.1}$$

with $c_0 = a_0 = 0$. The 1-point CDV parameter (y_k) is positive when the cell arrives "early" and this cell is considered to belong to a "clump". On the other hand, the 1-point CDV parameter is negative when the cell arrives "late" and in this case a "gap" is observed between the preceding cell and the present cell.

The *modified 1-point CDV* y'_k is defined in the same way as the previous, taking into consideration the CDV tolerance τ. This CDV definition mirrors the Peak Cell Rate Reference Algorithm, that is given in Annex 1 to Recommendation I.371 [50], and can be used for the conformance test of cells. The value of the modified 1-point CDV parameter is $y'_k := c'_k - a_k$. The set of theoretical times is recursively obtained as follows:

$$\begin{aligned} c'_{k+1} &= c'_k + T \qquad \text{if } \tau \geq y'_k \geq 0 \\ &= a_k + T \qquad \text{if } y'_k < 0 \\ &= c'_k \qquad\quad \text{if } y'_k > \tau \end{aligned} \tag{7.2}$$

A cell is "conforming" to the pair (T, τ) if and only if its modified 1-point CDV value (y'_k) is smaller than the CDV tolerance τ.

The *2-point CDV* v_k is defined on the basis of observation of corresponding cell arrivals at two measuring points. For cell no. k it is defined as the difference between the absolute cell transfer delay x_k between the two measurement points and a reference cell transfer delay $d_{1,2}$, i.e. $v_k = x_k - d_{1,2}$. The 2-point CDV parameters can be used for the proper

allocation of the equipments or network sections that are responsible for CDV on an end-to-end connection.

7.3 A Markovian Solution Method

In this section an exact Markovian solution method for the evaluation of CDV in a single ATM multiplexing stage and the characterization of the CDV affected CBR cell stream are presented [81, 74, 75]. This solution method is the generalization of the approach of [38]. In contrast to [38] where the background traffic is modeled by Poisson process in this method the background traffic modeled by a batch Bernoulli process with general batch size distribution thereby allowing to model the burstiness of the background traffic. This provides us with a more powerful tool for ATM multiplexer modeling and yields to a realistic model.

7.3.1 The Model

Consider a single server infinite capacity discrete time FIFO queue receiving the superposition of a CBR stream with interarrival time T and an interfering background stream which arrives to the queue in the form of independent and identically distributed batches with a general batch size distribution denoted by $b(k)$ (Figure 7.1).

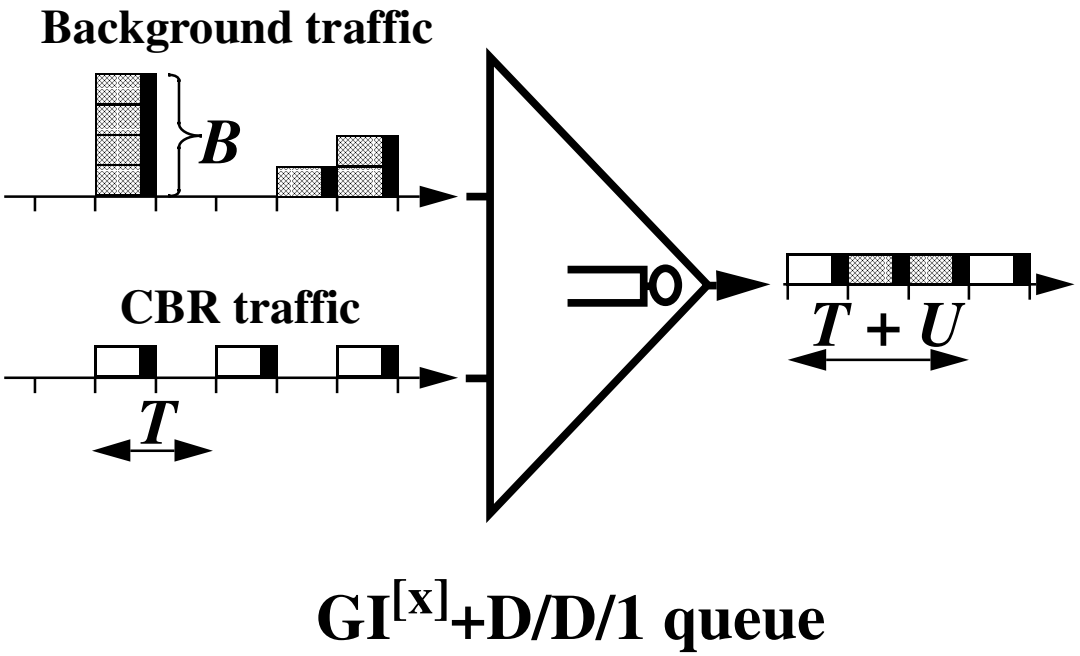

Figure 7.1: The FIFO Model

The cells arriving from the background cell stream in the same timeslot are served in random order, but when there is also a cell arriving from the CBR stream it is served

first. This serving rule will give a tight lower bound on the amount of CDV which the background stream puts on the CBR cell stream [68]. With the assumption that all actions in the queue occur at the timeslot boundaries, the queue length seen by an arriving CBR cell is a Markov process and it equals the waiting time of the CBR cell.

7.3.2 The Computation of the Transition Matrix

In this section a new solution method for the computation of the transition matrix of the model is derived.

Let W_n denote the waiting time of cell no. n. The element (j,k) of the transition matrix $\mathbf{Q}$ is defined as follows:

$$q_{j,k} = P\{W_i = k \mid W_{i-1} = j\} \quad j,k \geq 0 \tag{7.3}$$

The transition matrix $Q = \{q_{j,k}\}$ can be computed by

$$q_{j,k} = Q(j,k-1) - Q(j,k) \tag{7.4}$$

where

$$Q(j,k) = P\{W_i > k \mid W_{i-1} = j\} = \sum_{n \geq 0} P_n(j,k) \tag{7.5}$$

with $P_n(j,k) = P\{W_i > k \mid W_{i-1} = j \text{ and } n \text{ arrivals in }]Ti - T, Ti] \}b^{T*}(n)$, where $b^{T*}(n)$ is the T-fold convolution of the batch size distribution, and evaluated in n it represents the probability of n arrivals in $]Ti - T, Ti]$. It can be shown (see Appendix A) that

$$P_n(j,k) = \begin{cases} 0 & j+1 \geq T \text{ and } n \leq T+k-j-1 \text{ or} \\ & \quad j+1 < T \text{ and } k \geq n \\ b^{T*}(n) & n > T+k-j-1 \\ \sum_{s=1}^{n-k} b^{s*}(k+s) \times & \\ \quad \times b^{(T-s)*}(n-k-s)\frac{T-n+k}{T-s} & j+1 < T \text{ and } k < n \leq T+k-j-1 \end{cases} \tag{7.6}$$

The stationary waiting time distribution can be found from solving the equilibrium system $\mathbf{w} = \mathbf{wQ}$ where $\mathbf{w} = (w_0, w_1, ..., w_i, ...)$ and $\mathbf{Q} = \{q_{j,k}\}$. In numerical applications, a truncation of the state space is in general necessary.

For the steady state queue length distribution, [8] provide a closed form expression for the generating function. [8] also provide the generating function of the waiting time for cell no. $n+1$ conditioning on the event that waiting time for cell n is i. From this the entries in the transition matrix can in principles be found. However, this requires not only an inversion but also the determination of $T-1$ boundary probabilities, and we have found the direct approach more appealing.

7.3.3 Characterization of CDV Affected CBR Cell Stream

In this section the characterization of the CDV affected CBR cell stream is outlined [38]. This approach can be used for characterizing the departure process of the CBR cell stream from any networks or network elements (see Figure 7.2) where the dependence between the waiting times of successive cells is Markovian. In our case this characterization is used

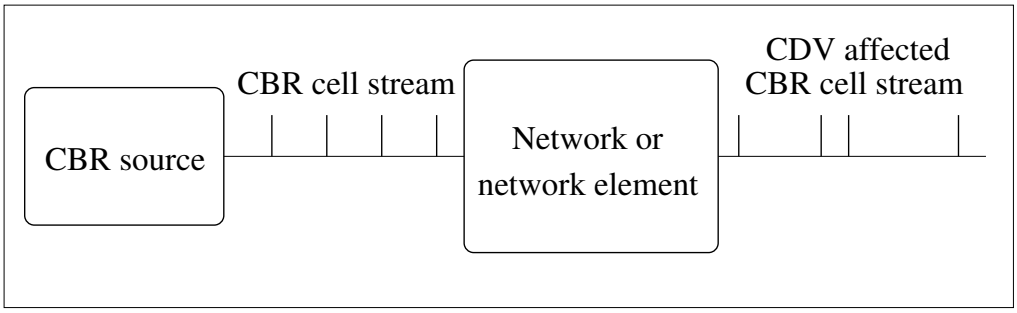

Figure 7.2: The CDV Affected CBR Cell Stream

for the specific case when the CBR cell stream is multiplexed with a background stream in a FIFO queue as described in Section 7.3.1. In the previous section the computation of the transition matrix and the stationary waiting time distribution have been derived which are the input data for the CDV characterization presented below.

Consider a CBR cell stream with interarrival time T. Choose the time such that cell no. n is transmitted at time nT from the source. Let W_n denote the waiting time of cell no. n.

Two assumption are made in the model:

- The sequence W_n is assumed stationary with distribution $w_k = P\{W_i = k\}$.
- The dependence between waiting times of successive cells is assumed first order Markovian, and characterized by the transition matrix $\mathbf{Q}$ in which entry (j,k) is:

$$q_{j,k} = P\{W_i = k \mid W_{i-1} = j\} \quad j,k \geq 0 \tag{7.7}$$

Define $\tau_n = nT + W_n + 1$. Thus τ_n denotes the departure time of cell no n. (Strictly speaking, τ_n is the time at which cell n starts the service i.e. one time unit before it leaves the queue.) Define the shifted interdeparture time of cell no. n as: $U_n = \tau_n - nT - \tau_0$ (see Figure 7.3). It is a fundamental random variable from which all CDV characteristics of interest can be derived. Let $f_n(k) = P\{U_n = k\}$ be the probability distribution of U_n for cell no. n. It can be seen [38] that

$$f_n(k) = \begin{cases} \sum_{i=0}^{\infty} w_i q_{i,i+k}^{(n)} & \text{if } k \geq -nT+1 \\ 0 & \text{otherwise} \end{cases} \tag{7.8}$$

where $q_{i,i+k}^{(n)}$ denotes entry (j,k) in the n'th power of the transition matrix $\mathbf{Q} = \{q_{j,k}\}$.

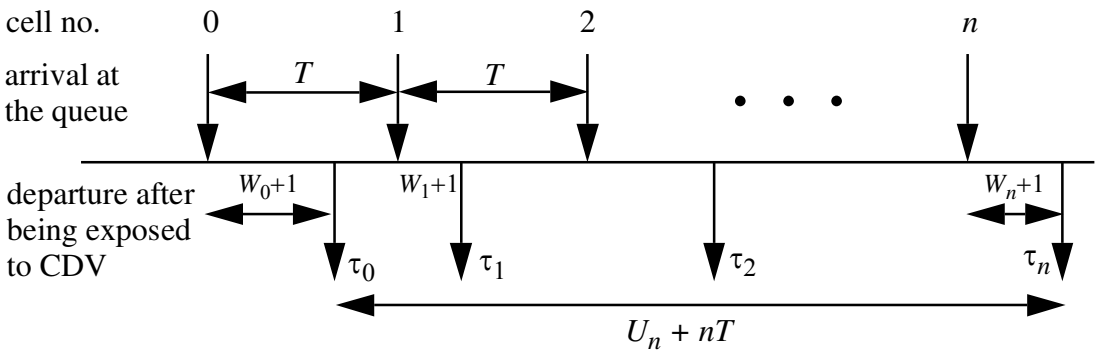

Figure 7.3: The shifted interdeparture time is to be seen as the difference between the actual departure of cell n (τ_n) and the expected departure time ($\tau_0 + nT$).

Next it is shown that how the usual point process characteristics can be derived from U_n. We first investigate the *interval representation.*

7.3.4 Interdeparture Time Distributions

The interdeparture time distribution between cell no. 0 and n is simply obtained by the translation nT of the shifted interdeparture time distribution U_n i.e.

$$P\{\tau_n - \tau_0 = k\} = P\{U_n = k - nT\} = f_n(k - nT) \tag{7.9}$$

7.3.5 Index of Dispersions for Intervals

The index of dispersion can be obtained from the definition as:

$$c_n^2 = \frac{nVar(\tau_n - \tau_0)}{E(\tau_n - \tau_0)^2} = \frac{Var(U_n)}{nT^2} = \frac{\sum_{k=-nT+1}^{\infty} k^2 f_n(k)}{nT^2} \tag{7.10}$$

since $E(U_n) = 0$.

For the *counting representation* the number of counts is fundamental and is measured in two ways:

7.3.6 Number of Departures in a Window Starting Just After a Departure

Let $N(\tau_j, \tau_j + t)$ denote the number of departures in a window of length t starting just after an arbitrary CBR cell departure τ_j. The probability distribution is

$$\begin{aligned} P\{N(\tau_j, \tau_j + t) = n\} &= P\{U_n + nT \leq t\} - P\{U_{n+1} + (n+1)T \leq t\} \\ &= \sum_{j \leq -nT+t} f_n(j) - \sum_{j \leq -(n+1)T+t} f_{n+1}(j). \end{aligned} \tag{7.11}$$

7.3.7 Number of Departures in an Arbitrary Window

Based on a basic result in point process theory (see e.g. section 4.2 in [17]) a simpler formula for the distribution of the number of departures in an arbitrary window than the one presented in [38] can be obtained. Let t_0 denote an arbitrary point in time and $N(t_0, t_0 + t)$ denote the number of departures in $]t_0, t_0 + t]$. Then

$$\begin{aligned} P\{N(t_0, t_0 + t) \geq n\} = P\{Y + X_2 + \ldots + X_n \leq t\} = \\ \frac{1}{T} \sum_{u=1}^{t} \sum_{k=1}^{u} (f_{n-1}(k - [n-1]T) - f_n(k - nT)) \end{aligned} \tag{7.12}$$

in which Y denote the forward recurrence time, X_i denote the interdeparture time between cell no. $i-1$ and i. From (7.12) we obtain

$$P\{N(t_0, t_0 + t) = n\} = \frac{1}{T} \sum_{u=1}^{t} \sum_{k=1}^{u} f_{n-1}(k - (n-1)T) - 2f_n(k - nT) + f_{n+1}(k - (n+1)T) \tag{7.13}$$

7.3.8 Limit Distributions

The Markovian assumption implies that in the limit $n \to \infty$, $q_{j,k} \to w_k$ independent of j yielding the limit distribution

$$f_\infty(k) = \sum_{i=0}^{\infty} w_i w_{i+k} \tag{7.14}$$

and

$$P\{N(\tau_j, \tau_j + t) = n\} = \sum_{j=-(n+1)T+t+1}^{-nT+t} f_\infty(j) \tag{7.15}$$

7.4 A Diffusion Method

The approach presented in the previous sections is attractive because it is exact and the fundamental solution of the transition matrix is given in closed form. However, the computational burden is high and the numerical complexity increases without bounds

when the load of the multiplex approaches one. Therefore an accurate approximation is needed which maintain the important qualities of the model but for which the fundamental point process quantities can be easily computed. As a candidate a diffusion approximation which fulfills these requirements is suggested in this section [81, 74, 75]. This method is a generalization of the diffusion method of [10] where the background traffic is Poisson. In this new method, similarly as in the Markovian exact method, the background traffic modeled by a batch Bernoulli process with general batch size distribution. Therefore this method also has the ability to model the burstiness of the background traffic.

7.4.1 The Model

The diffusion model is based on the idea that the evolution of the queue length (or virtual waiting time) between CBR arrivals is approximated by a reflected Brownian motion. Markovian dependencies are assumed between successive CBR cell waiting times in the queue as in the former section.

Let $\tilde{W}_t$ denote the waiting time a fictitious observer would experience if he joined the diffusion queue at time t (the virtual waiting time at time t). The probability of $\tilde{W}_t \leq x$ conditioned on $\tilde{W}_0 = y$ is, for a Brownian motion with drift m (m assumed smaller than zero in order to ensure a stable queue), variance σ^2 and a reflection in zero, in Section 2.8 of [64] derived to be:

$$P\{\tilde{W}_t \leq x \mid \tilde{W}_0 = y\} = \begin{cases} \Phi(\frac{x-y-mt}{\sigma\sqrt{t}}) - e^{2\frac{mx}{\sigma^2}}\Phi(\frac{-x-y-mt}{\sigma\sqrt{t}}), & \text{for } x \geq 0 \\ 0, & \text{for } x < 0 \end{cases} \quad (7.16)$$

where Φ denotes the standard Gaussian probability distribution. The right hand side of formula (7.16) is a distribution function in x for all $y \geq 0$ and all $t > 0$, and it converges towards the exponential distribution with mean $\frac{-\sigma^2}{2m}$, independent of initial condition y, when t tends to infinity.

The diffusion approximation is used for modeling the arrivals of cells belonging to both the CBR stream and the background stream, i.e. the queue length is not increased by one at each CBR cell arrival. These arrivals as well as the batch arrivals and the departures are taken into account by a proper choice of drift and variance in the diffusion process as shown in the next subsection.

The waiting time of CBR cell no. n, is approximated by:

$$W_n = \tilde{W}_{nT} \quad (7.17)$$

that is *the waiting time that CBR cell no. n experience in the queue is approximated by the virtual waiting time in the diffusion queue at time nT.*

A disadvantage with formula (7.16) is that there is a positive probability that the waiting time between time t and time $t+u$ decreases more than u which of course in the original $GI^{[x]} + D/D/1$ queue is impossible. If t is the time of CBR arrival no. n and $u = T$ then this would imply that CBR cell no. $n+1$ departs the queue before cell no. n. This weakness is inherent to the model and the model should only be used for values of

T for which this probability is low, implying that T cannot be chosen too small.

7.4.2 Drift and Variance Computation

As in the diffusion model for the M/G/1 queue (see section 2.8 in [64]) the drift is $m = \rho - 1$. The key idea of choosing the most appropriate variance is that *we match the decay rates of the stationary waiting time distribution in the diffusion approximation with the asymptotic decay rate of the $D + GI^{[x]}/D/1$ queue.*

It is known from [13] that the queue length seen by an arriving CBR cell is asymptotically geometric i.e. $P(Q > r) \approx (1/z_\infty)^r$, where z_∞ is the dominant root of the corresponding z-transform which is found by solving the equation

$$B(z)^T - z^{T-1} = 0. \tag{7.18}$$

The root z_∞ is the root with smallest module outside the unit disk, and it is unique and real (see [13] In Appendix A3).

Since the diffusion decay rate is $\frac{-2m}{\sigma^2}$ we get:

$$\sigma^2 = -\frac{2m}{\ln(z_\infty)}. \tag{7.19}$$

7.4.3 The Distribution of the Interdeparture Time

The shifted interdeparture time in the diffusion context is: $\tilde{U}_t = \tilde{W}_t - \tilde{W}_0$ (think of $t = nT$), and

$$\begin{aligned} P\{\tilde{U}_t \le x\} &= \int_0^\infty P\{\tau_t - \tau_0 - t \le x \mid \tau_0 = y\} dP\{\tau_0 \le y\} \\ &= \int_0^\infty P\{\tilde{W}_t \le x + y \mid \tilde{W}_0 = y\} dP\{\tilde{W}_0 \le y\} \end{aligned} \tag{7.20}$$

The computation carried out in [10] shows that the time dependent shifted interdeparture time distribution function is given as:

$$\tilde{F}_t(x) = P\{\tilde{U}_t \le x\} = \begin{cases} \frac{1}{2} + \frac{1}{2}\Phi(\frac{x-mt}{\sigma\sqrt{t}}) - \frac{1}{2}e^{\frac{2mx}{\sigma^2}}\Phi(-\frac{x+mt}{\sigma\sqrt{t}}), & \text{for } x \ge 0 \\ \frac{1}{2}e^{-\frac{2mx}{\sigma^2}}\Phi(\frac{x-mt}{\sigma\sqrt{t}}) + \frac{1}{2}\Phi(\frac{x+mt}{\sigma\sqrt{t}}), & \text{for } x < 0 \end{cases} \tag{7.21}$$

From this expression it is seen that the probability distribution function $\tilde{F}_t$ for $\tilde{U}_t$ fulfills the relation: $\tilde{F}_t(x) + \tilde{F}_t(-x) = 1$, thus implying that the density function for $\tilde{U}_t$ is symmetric.

As in the exact case a limit distribution emerges as $t \to \infty$ From (7.21) it is easily seen that

$$\tilde{F}_\infty(x) = P\{\tilde{U}_\infty \le x\} = \begin{cases} 1 - \frac{1}{2}e^{\frac{2mx}{\sigma^2}}, & \text{for } x \ge 0 \\ \frac{1}{2}e^{-\frac{2mx}{\sigma^2}}, & \text{for } x < 0 \end{cases} \tag{7.22}$$

that is in the limit $t \to \infty$ the shifted interdeparture time is a two sided exponential distribution with rate $-\frac{2m}{\sigma^2}$ and $\frac{2m}{\sigma^2}$ respectively.

The interdeparture time distribution between cell no. k and $k+n$ is obtained from the shifted interdeparture time distribution as

$$F_n(x) = P\{\tau_{k+n} - \tau_k \leq x\} = P\{\tau_n - \tau_0 \leq x\} = P\{\tilde{U}_{nT} \leq x - nT\} = \tilde{F}_{nT}(x - nT) \quad (7.23)$$

7.4.4 Index of Dispersions for Intervals

For IDI computation the variance of $\tilde{U}_t$ is needed, and in [10] it is found to be:

$$\begin{aligned} Var(\tilde{U}_t) &= \frac{\sigma^4}{2m^2}(2\Phi(-\frac{m}{\sigma}\sqrt{t}) - 1) + (2\sigma^2 t + m^2 t^2)\Phi(\frac{m}{\sigma}\sqrt{t}) \\ &+ (\frac{\sigma^3}{m}t^{1/2} + m\sigma t^{3/2})\varphi(\frac{m}{\sigma}\sqrt{t}) \end{aligned} \quad (7.24)$$

where φ is the standard normal density function. By applying the expansion $\Phi(-x) = \frac{\varphi(x)}{x}(1 - \frac{1}{x^2} + O(x^{-4}))$, for $x \to \infty$, see e.g. problem 7.7.1 in [30] and evaluating $Var(\tilde{U}_t)$ at $t = nT$ we arrive at:

$$c_n^2 = \frac{\sigma^4}{2m^2T^2}\frac{1}{n} + O(n^{-3/2}e^{-\frac{m^2}{\sigma^2}nT}), \quad n \text{ large} \quad (7.25)$$

which clearly shows that the relative variance of the departure process decreases with increasing time scales, i.e. the process is less bursty than a renewal process.

7.4.5 Number of Departures in a Window Starting Just After a Departure

Consider an interval of the form $]\tau_j, \tau_j + t]$ starting just after the departure of cell no. j. Then

$P\{N(\tau_j, \tau_j + t) \geq n\} = P\{\tilde{U}_{nT} + nT \leq t\} = \tilde{F}_{nT}(t - nT) =$

$$\begin{cases} \frac{1}{2} + \frac{1}{2}\Phi(\frac{t}{\sigma\sqrt{nT}} - \frac{(1+m)}{\sigma}\sqrt{nT}) - \frac{1}{2}e^{\frac{2m}{\sigma^2}(t-nT)}\Phi(-\frac{t}{\sigma\sqrt{nT}} + \frac{(1-m)}{\sigma}\sqrt{nT}), & \text{for } t \geq nT \\ \frac{1}{2}e^{-\frac{2m}{\sigma^2}(t-nT)}\Phi(\frac{t}{\sigma\sqrt{nT}} - \frac{(1+m)}{\sigma}\sqrt{nT}) + \frac{1}{2}\Phi((\frac{t}{\sigma\sqrt{nT}} - \frac{(1-m)}{\sigma}\sqrt{nT}), & \text{for } t < nT \end{cases} \quad (7.26)$$

7.4.6 The Number of Departures in an Arbitrary Window

Let t_0 denote an arbitrary point in time and consider the interval $]t_0, t_0 + t]$. Similarly as in (7.12) we can write

$$P\{N(t_0, t_0 + t) \geq n\} = P\{Y + X_2 + .. + X_n \leq t\} = \frac{1}{T}\int_0^t (F_{n-1}(u) - F_n(u))du \quad (7.27)$$

in which Y denote the forward recurrence time, X_i denote the interdeparture time between cell no. $i-1$ and i, and $F_i(u)$ denotes the distribution function of $\sum_{j=1}^{i} X_j$. Since $F_n(u) = P\{\tilde{U}_{nT} \leq u - nT\} = \tilde{F}_{nT}(u-nT)$ is given in closed form in formula (7.26) we may obtain the probability distribution of the number of departures in an arbitrary window by integrating (7.26) with respect to t. A complication arises due to the non-zero probability of cell no. $n+k$ departing the queue before cell no. n. If cell no. n is the first departure in the window it may happen that cell no. $n+k$ is left of the window. However, if cell no. $n+p$ is the first departure after the end of the window it may happen that cell no. $n+p+k$ is inside the window. The easiest way to approximately capture this is by extending the integration in (7.27) from 0 down to $-\infty$ i.e. $P\{N(t_0, t_0+t) \geq n\} \approx \frac{1}{T}\int_{-\infty}^{t}(F_{n-1}(u) - F_n(u))du$. The result of the integration then is:

$$P\{N(t_0, t_0 + t \geq n\} = G_{n-1}(t) - G_n(t) \tag{7.28}$$

in which $G_n(t)$ for $t < nT$ is given as:

$$\begin{aligned} G_n(t) = \quad & \frac{\sigma^2}{(-4m)}\frac{1}{T}\{e^{-\frac{2m}{\sigma^2}(t-nT)}\Phi(\frac{t}{\sigma\sqrt{nT}} - \frac{(1+m)}{\sigma}\sqrt{nT}) \\ - \quad & \Phi(\frac{t}{\sigma\sqrt{nT}} - \frac{(1-m)}{\sigma}\sqrt{nT})\} \\ + \quad & \frac{1}{2T}\{(t-(1-m)nT)\Phi(\frac{t}{\sigma\sqrt{nT}} - \frac{(1-m)}{\sigma}\sqrt{nT}) \\ + \quad & \sigma\sqrt{nT}\varphi(\frac{t}{\sigma\sqrt{nT}} - \frac{(1-m)}{\sigma}\sqrt{nT})\} \end{aligned} \tag{7.29}$$

and for $t > nT$, $G_n(t)$ is given as:

$$\begin{aligned} G_n(t) = \quad & \frac{\sigma^2}{(-4m)}\frac{1}{T}\{e^{\frac{2m}{\sigma^2}(t-nT)}\Phi(\frac{-t}{\sigma\sqrt{nT}} + \frac{(1-m)}{\sigma}\sqrt{nT}) \\ - \quad & \Phi(\frac{-t}{\sigma\sqrt{nT}} + \frac{(1+m)}{\sigma}\sqrt{nT})\} \\ + \quad & \frac{1}{2T}\{((1+m)nT - t)\Phi(\frac{-t}{\sigma\sqrt{nT}} + \frac{(1+m)}{\sigma}\sqrt{nT}) \\ + \quad & \sigma\sqrt{nT}\varphi(\frac{-t}{\sigma\sqrt{nT}} + \frac{(1+m)}{\sigma}\sqrt{nT})\} + \frac{t-nT}{T} \end{aligned} \tag{7.30}$$

where we have utilized the symmetry relation $\tilde{F}_l(x) + \tilde{F}_l(-x) = 1$ to obtain (7.30).

7.5 Numerical Evaluation of the Markovian and Diffusion Models

In this Section a series of numerical results are presented to illustrate the behavior of both the exact model as well as the diffusion approximation.

In order to keep the Section short we have fixed the interarrival time of the CBR traffic to $T = 20$, as well as the load on the multiplex, which is fixed to $\rho = 0.8$.

The influence of the burstiness of the background traffic is investigated by using the following 3 batch size distributions:

- Smooth background traffic: *Binomial* distribution with parameters n and p where

$$b(k) = \binom{n}{k} p^k (1-p)^{n-k} \tag{7.31}$$

  In the examples $n = 2$ and $p = 0.375$. For this case the peakedness (the ratio of the variance and mean of the number of arriving cells) is $Z = 0.625$. A binomial batch size distribution is appropriate when the background traffic consists of only a few sources or it may be used to model a background traffic with negative correlations.

- Bursty background traffic: *Generalized negative binomial* (Pascal) distribution with parameters n and p where

$$b(k) = \binom{n+k-1}{k} p^k (1-p)^n \tag{7.32}$$

  In the examples $n = 0.75$ and $p = 0.5$ with peakedness $Z = 2$. Since a batch of arrivals in a slot in practice always consist of arrivals from different sources a negative binomial distribution is, seen in isolation, not an appropriate batch size distribution. However, applying it anyway provides a simple way to model positive correlations in the background traffic with a model that does not include correlations. This technique cannot be expected to work when the burstiness of the background traffic is very large. However in this case the CBR traffic would possible need some kind of delay priority and a completely different model would be needed.

- *Poisson* background traffic is also used, since it is the classical case. Furthermore, it is an appropriate model for the case where the background traffic consists of a large number of sources.

Figure 7.4 and Figure 7.5 depict the effect of the burstiness of the background traffic on the shifted interdeparture time distribution. In Figure 7.4 it is the distribution between cell no. 0 and 1 ($n = 1$) while in Figure 7.5 it is the distribution between cell no. 0 and 5 ($n = 5$).

In Figure 7.6 the effect of the load on the shifted interdeparture time distribution can be seen. The background traffic is Poisson ($Z = 1$) and $n = 1$.

By comparing Figure 7.4 and Figure 7.6 it can be observed that the burstiness of the background traffic has a stronger influence on the shifted interdeparture time distribution than the load. Since it is the load which can be controlled or estimated, while the burstiness is highly unknown, this gives rise to practical difficulties in the dimensioning of the UPC parameters.

The diffusion approximation performs reasonable well, but for the smooth background traffic case the inaccuracy is slightly larger than for the other cases. The somewhat disappointing result for the Binomial case when $n = 5$ is due to the fact that the steady state waiting time distribution of the $D+GI[Binomial]/D/1$ deviates more from an exponential distribution than the others.

Figure 7.7 and Figure 7.8 depict the distributions of the number of cell departures in a window for the three different burstiness of the background traffic and the load, respectively. Also here we observe the strong effect a variation of the burstiness of the background traffic has on the distribution. The performance of the diffusion approximation is very good, in the figures the distribution curves almost coincide.

Finally, Figure 7.9 depicts the index of dispersions for intervals for different burstiness of the background traffic. While some slight inaccuracies are present in the Binomial and Pascal case, it is clearly demonstrated that the actual shape of the IDI curve is very well approximated over the entire range of n.

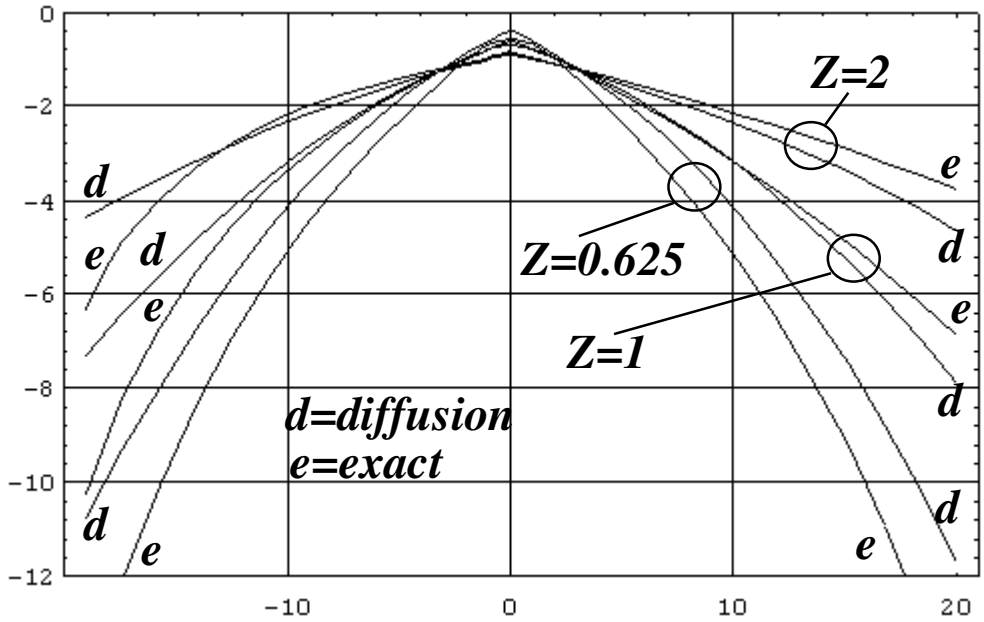

Figure 7.4: Probability mass function of the shifted interdeparture time on a loglinear plot for background traffic with different burstiness ($n = 1$, $T = 20$, $\rho = 0.8$).

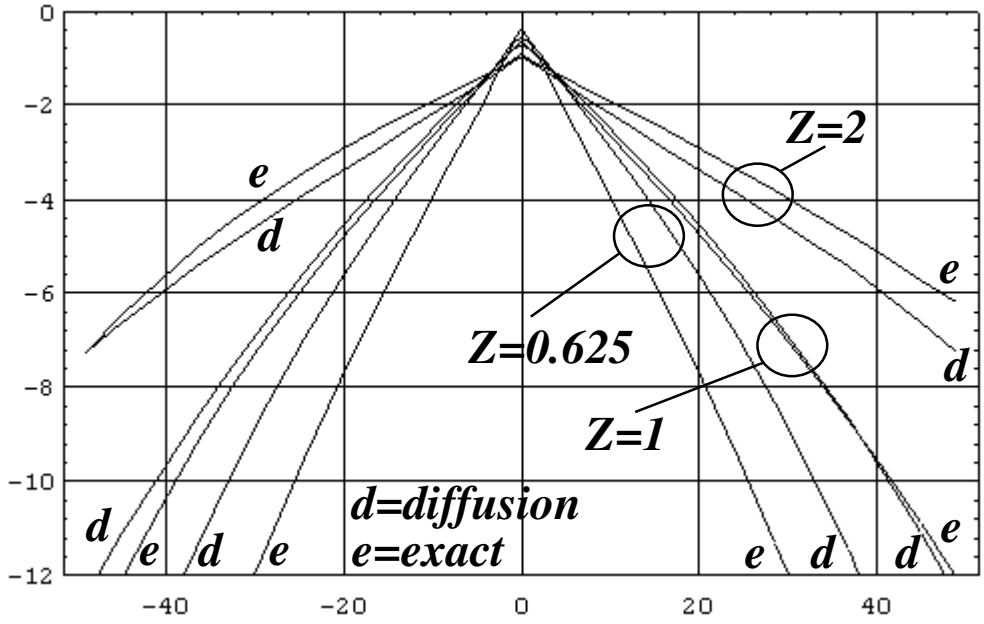

Figure 7.5: Probability mass function of the shifted interdeparture time on a loglinear plot for background traffic with different burstiness ($n = 5$, $T = 20$, $\rho = 0.8$).

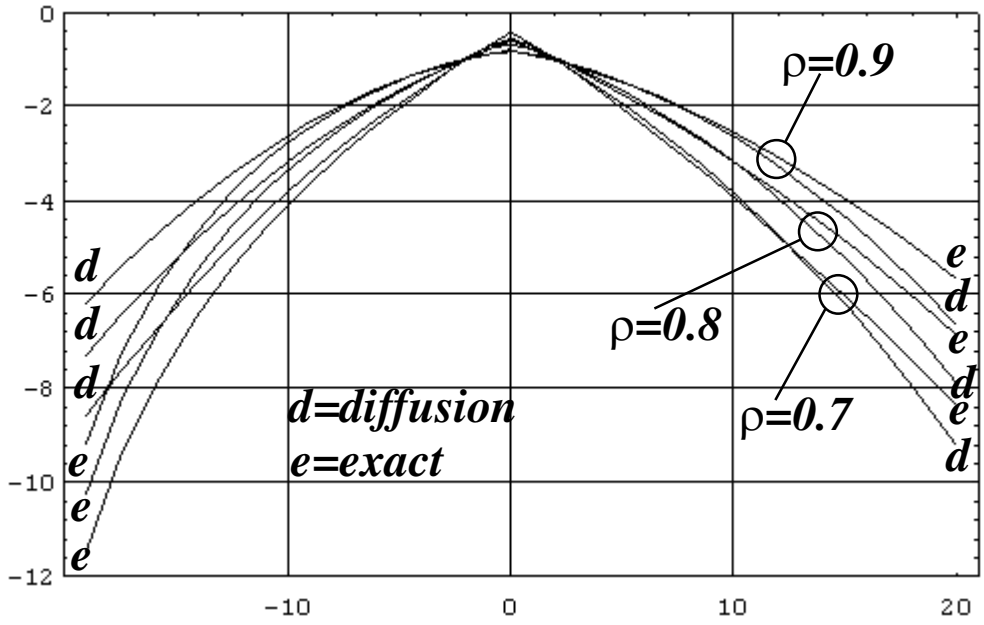

Figure 7.6: Probability mass function of the shifted interdeparture time on a loglinear plot for different loads ($n = 1$, $T = 20$, $Z = 1$).

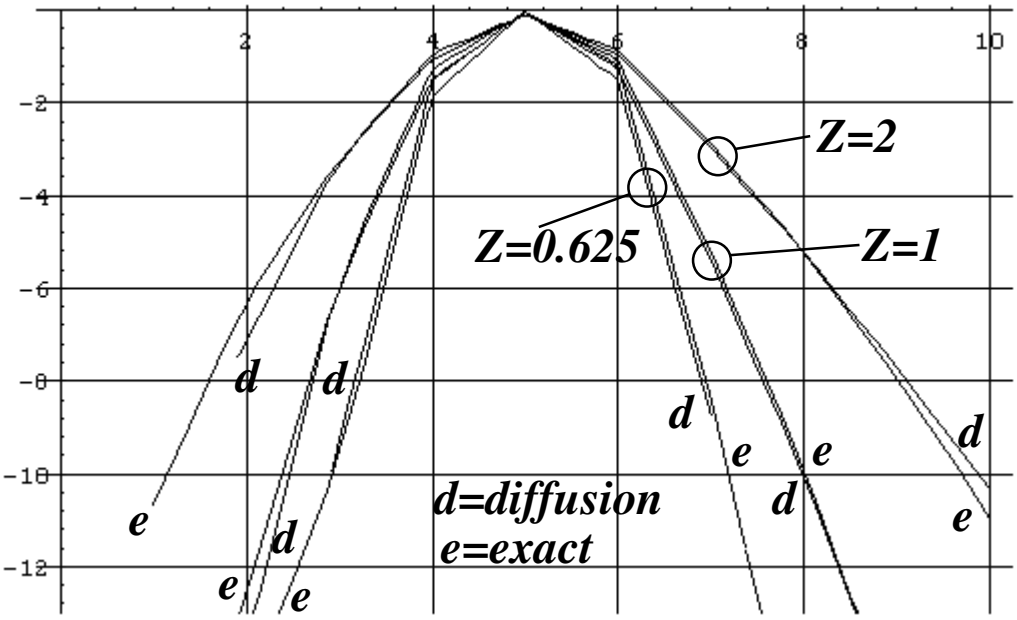

Figure 7.7: Probability mass function of the number of cell departures in a window of length $5T$ on a loglinear plot for background traffic with different burstiness ($T = 20$, $\rho = 0.8$).

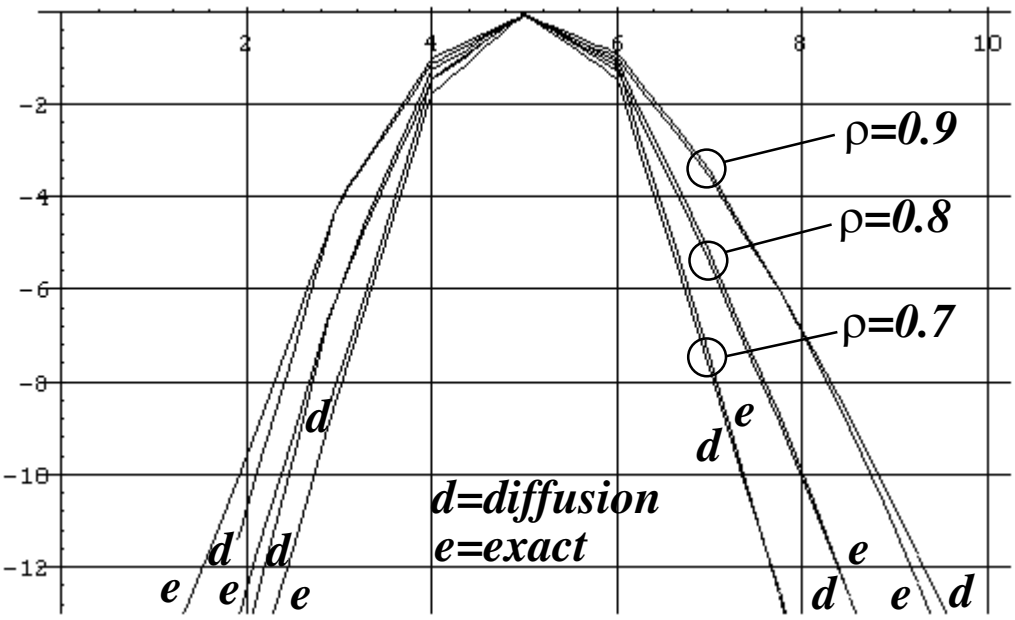

Figure 7.8: Probability mass function of the number of cell departures in a window of length $5T$ on a loglinear plot for different loads ($T = 20$, $Z = 1$).

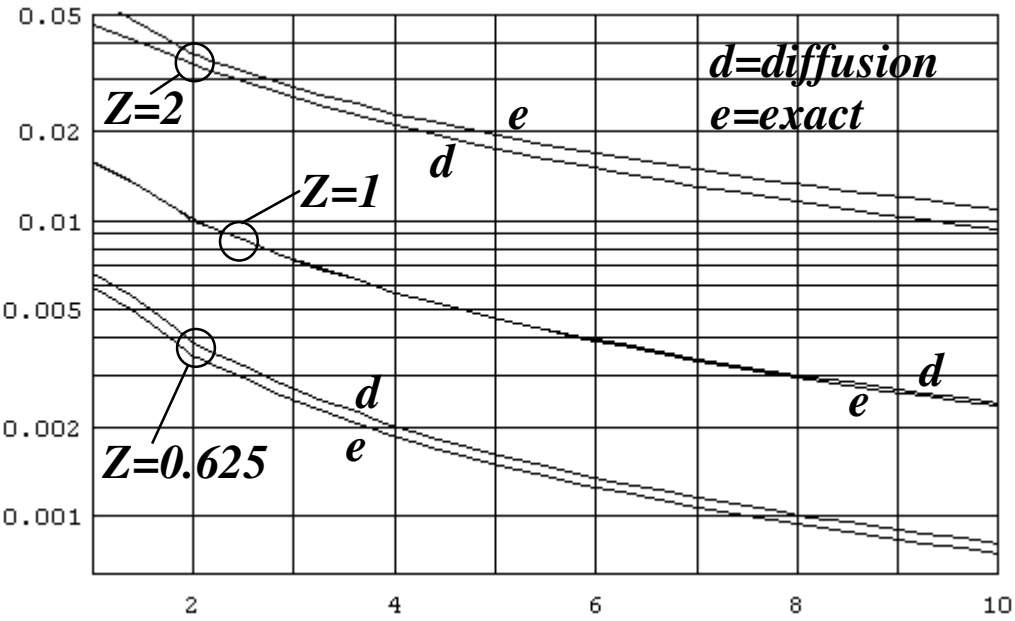

Figure 7.9: Index of dispersions for intervals for background traffic with different burstiness ($T = 20$, $\rho = 0.8$).

7.6 The Superposition of CDV Affected CBR Cell Streams

In this section the performance evaluation of an ATM multiplexer receiving a number of CDV affected CBR cell streams is investigated [78, 82]

7.6.1 The Model

We consider a multiplexer offered traffic from a number of CBR cell streams which prior to the arrival to the multiplexer have been exposed to CDV in a single multiplexing stage (see Figure 7.10).

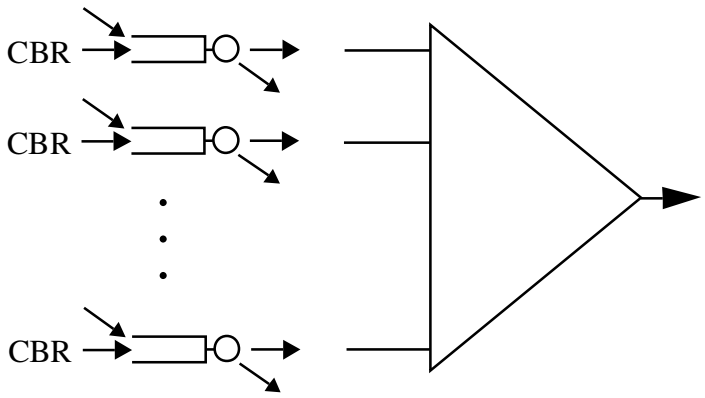

Figure 7.10: Superposition of CDV Affected CBR Cell Streams

7.6.2 The Analysis Method

For queues with constant service times it has turned out that the so-called Beneš Result, see [99] Section 5.3, is a very powerful tool to obtain either the exact queue length distribution or at least a tight approximation. The Beneš Result valid for stable queues ($\rho < 1$) offered a general stationary arrival process is:

$$P\{W_t > r\} = \sum_{n=1}^{\infty} P\{N(t-n,t) = n+r\} P\{W_{t-n} = 0 | N(t-n,t) = n+r\} \qquad (7.33)$$

where W_t is the virtual waiting time at time t. The difficult term in the above expression is $P\{W_{t-n} = 0 | N(t-n,t) = n+r\}$. Applying the so-called "local load approximation", see section 5.3.2 in [99] for details, we end up with:

$$P\{W_t > r\} = \sum_{n=1}^{\infty} P\{N(t-n,t) = n+r\} - \rho \sum_{n=1}^{\infty} P\{N(t-n,t) = n+r | \text{one arrival at (t-n)}\} \qquad (7.34)$$

i.e. an expression containing only terms related to the arrival process.

In Section 7.4, expressions for the number of departures in a window from a single CDV affected CBR cell stream were derived. Now a finite number of independent CDV affected CBR cell streams are multiplexed in the buffer. Therefore the distribution of the number of departures from the superposition before it enters the multiplexing queue can be found from a convolution of the individual distributions. The number of departures from a single stream is given in Eq. 7.26 and Eq. 7.28. An approximation for the virtual waiting time distribution can now be obtained from Eq. 7.34.

7.6.3 Numerical Results

To illustrate the approach two examples are presented. In the first example 8 CBR sources are exposed to CDV in a single multiplexing stage (see Figure 7.10) and after that the 8 streams are offered to a FIFO with a load of 0.8 implying that the peak rate of the CBR sources are 0.1 corresponding to $T = 10$. The amount of CDV is controlled by the load of the interfering traffic on the CDV creating queues as described in Section 7.4.

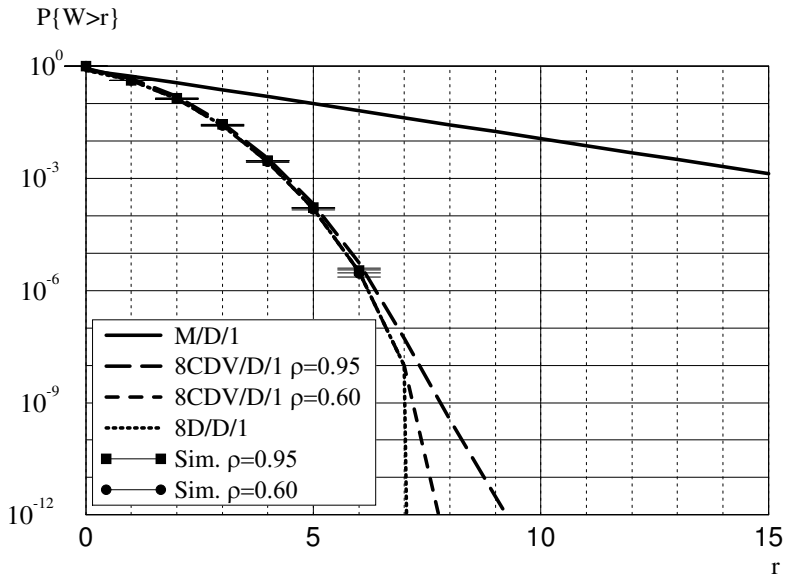

Figure 7.11: Comparison of the Virtual Waiting Time Distributions of 8CDV/D/1 Queues (load=0.8)

For a load of 0.6 in the CDV creating queues the queueing result differs very little from the result of the $8D/D/1$ queue while for a CDV creating load of 0.95 the difference become slightly larger. However, the delay performance is still much better than for the $M/D/1$ case.

A simulation of the virtual waiting time has also been carried out and the analytically computed values fall within the 95% confidence interval over the entire range for which simulation is feasible.

In the second example 16 CDV affected CBR sources are multiplexed in a FIFO with a load of 0.8 implying $T = 20$.

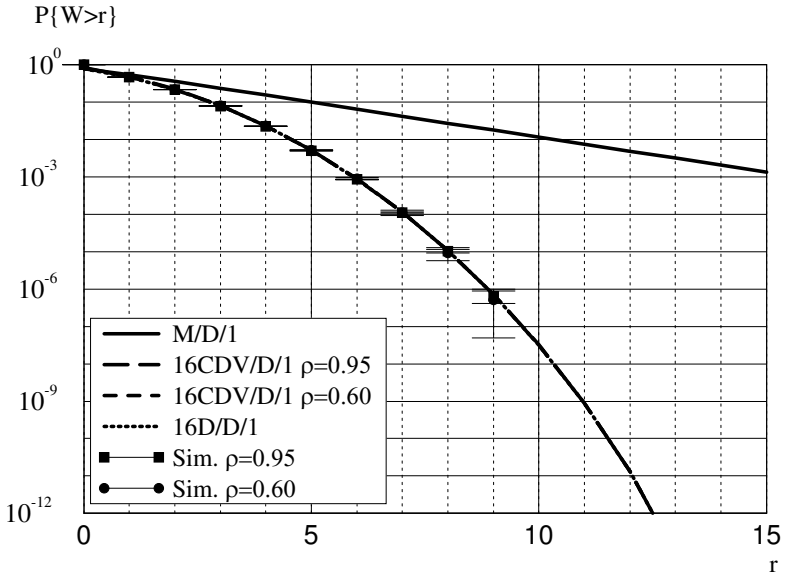

Figure 7.12: Comparison of the Virtual Waiting Time Distributions of $16CDV/D/1$ Queues (load=0.8). (Notice that the $16D/D/1$ curve and both $16CDV/D/1$ curves visually coincide.)

For this case the results are practically identical to the results for the $16D/D/1$ queue even in the case with a CDV creating load of 0.95. For more numerical results see [78]. Again simulations have been carried out and they support the analytical results.

An overall conclusion is that, except for the case with multiplexing a very few high speed CBR cell streams, the effect of CDV in a single multiplexing queue is neglectable.

7.7 Summary of Results

In the first part of this Chapter the CDV in an ATM multiplexer that is fed by the superposition of a CBR stream and a background stream is investigated.

I have presented a generalization of the approach of [38] to cope with other types of background traffic than Poisson. In the new exact method the background traffic is modeled by a batch arrival traffic with general distribution of batch size providing us with a flexible tool to investigate the impact of various types of interfering background traffic with arbitrary burstiness. The numerical investigations show that the variability of the background traffic is a key parameter in the determination of the CDV incurred on the CBR stream at the multiplexer.

An efficient approximation method by the generalization of a diffusion model developed in [10] is presented. This model has also been extended and evaluated for other types of background traffic than Poisson. From the numerical results we can conclude that the diffusion approach works reasonable well in general, that the accuracy is highest when the CBR stream is small compared to the background traffic, and when the load is not

unreasonable small. Also the peakedness of the background traffic should be kept not too far from 1. It can also be shown that the diffusion method is second moment exact in the important heavy traffic limit case for all values of T [81, 74]. The largest advantage of the diffusion model is that it provides a very efficient way to compute the point process characteristics of a CBR stream which have been perturbated by an interfering background stream in an ATM multiplexer.

In the second part of this Chapter an ATM multiplexer receiving CBR cell streams which prior to the arrival to the multiplexer have been exposed to CDV in a single multiplexing stage is investigated. The analysis results, which is based on the diffusion method and the Beneš result, show that except for the case with multiplexing a very few high speed CBR cell streams, the effect of CDV in a single multiplexing queue is negligible.

Chapter 8

Performance Evaluation of Cascaded ATM Multiplexers

8.1 Introduction

All results and models of the previous Chapter concern CDV due to a single ATM multiplexer. However, cells progressing along a virtual connection experience random delays in several multiplexing stages. It may happen not only within the public ATM networks but also in the Customer Premises Networks. Therefore the study of CDV due to tandem queues is also of great importance concerning the dimensioning of traffic control and network elements. It should be noted that it is a rather complicated and more difficult issue than the single queue case.

First results concerning the behaviour of CDV in a networking environment can be found in [7] and [69], where the results in [69] are based on the same authors results [68] for the single multiplexer case and now the issue is in the focus of CDV research [14].

The aim of this Chapter to model and analyze the CDV originated from the CBR cell stream have been passed through several multiplexing stages. The conclusions and results of this Chapter are used in Chapter 9 in the design of traffic control functions and network elements.

In Section 8.2 a characterization of CBR cell streams going through cascaded ATM multiplexers is given. The performance evaluation of the simple and widely applied renewal model of CDV affected CBR cell streams is shown in Section 8.3. The analysis of the superposition of CBR cell streams exposed to CDV after several multiplexing stages and a new solution of the $nTri/D/1$ queue are presented in Section 8.4-8.5. Finally, the Chapter is concluded in Section 8.6.

8.2 Characterization of CBR Cell Streams Going Through Cascaded ATM Multiplexers

8.2.1 Model Overview

Consider a discrete time queueing model consisting of M ATM queues in series. There are two types of arrivals to the system. First we have a reference connection entering the first queue and passing successively through all M queues. Secondly, for each queue k we have an interfering traffic entering queue k and immediately after service completion in queue k leaves the system (see Figure 8.1).

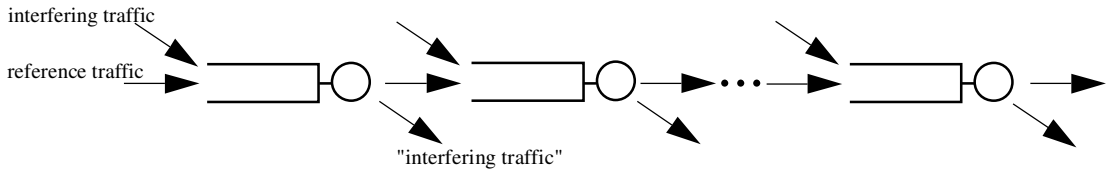

Figure 8.1: Queues in Series with Interfering Traffic

Let $\{X_n^{k-1}\}_n$ for $k = 1, ..., M$ be the sequence of interarrival times of the reference traffic to queue k. X_n^{k-1} is the interarrival time between reference cell no. n and $n+1$ to queue k.

The interfering traffic is modeled as a batch Bernoulli process, that is the number of arrivals in successive times slots are independent identically distributed with common distribution f and generating function F. For simplicity it is assumed that the interfering traffic statistically is the same at each queue. Y_i^k denotes the number of interfering arrivals in timeslot i at queue k. The output process of the reference connection from queue k is a function of the input processes $\{X_n^{k-1}\}_n$ and $\{Y_n^k\}_n$.

8.2.2 Analysis

The basis for the analysis is the approach developed by Berg and Resing [7]. Observing the empty slots between successive departures from queue k of cells of the reference cell stream, it is due to either departures of cells from the interfering stream or idleness of server in queue k. The last case is difficult to handle but if all queues are in overload, empty slots between two successive departures of the reference cell stream are all caused by departures from interfering cells. So, in the analysis a *heavy traffic assumption* is used which implies that at each point in time the k^{th} queue is non-empty.

Under this assumption all slots between two adjacent reference cell arrivals consists of interfering cells, implying that

$$X_n^k = 1 + \sum_{i=1}^{X_n^{k-1}} Y_i^k \tag{8.1}$$

where for $i = 1, ..., X_n^{k-1}$ the Y_i^k is the number of interfering arrivals in slot i. To arrive at Eq. 8.1 we have also assumed that the reference connection has priority over the interfering cells. A similar expression would be valid if the interfering cells had priority over the reference connection [7].

Eq. 8.1 shows that the process is a branching process with immigration. Furthermore we get

$$P_k(s) = sP_{k-1}(F(s)) \tag{8.2}$$

where $P_k(s)$ denotes the generating function of X_n^k. From the general theory of branching processes it is possible to arrive at the following results [7]:

1. Limit interdeparture time distribution

Assume that $E(Y_i^k) < 1$. When $k \to \infty$, then $X_n^k \to X_n^\infty$ in distribution where the generating function of X_n^∞ satisfy $P(s) = \prod_{j=1}^{\infty} F_j(s)$ in which $F_j(s)$ is recursively defined by $F_1(s) = s$ and $F_j(s) = F(F_{j-1}(s))$.

2. Asymptotically a renewal process

Let X_n^0 and X_{n+p}^0 be two possibly dependent interarrival times. Define: $(X_n^{k+1}, X_{n+p}^{k+1}) = \left(1 + \sum_{i=1}^{X_n^k} Y_i^k, 1 + \sum_{i=1}^{X_{n+p}^k} Y_i^{k,p}\right)$ where Y_i^k and $Y_i^{k,p}$ are independent of each other and satisfying the assumptions of the previous result. Then $(X_n^k, X_{n+p}^k) \to (X_n^\infty, X_{n+p}^\infty)$ in distribution, where X_n^∞, X_{n+p}^∞ are independent.

From these two results two important conclusions can be drawn:

1. When the reference cell stream has passed a sufficient number of queues the interdeparture time distribution has converged towards a limit distribution which only depends on the interfering traffic and not on the original characteristics of the reference stream except the rate.

2. When the reference stream has passed a sufficient amount of queues it becomes a renewal stream.

Section 6 in [7] argues by an interpolation argument between the light traffic case (no interfering traffic) and the heavy traffic case that the two results should also hold in moderate traffic with only the speed of the convergence decreased.

In the analysis we focus on how the squared coefficient of variation of the interdeparture time changes as the cell stream passes through the network and we will derive formulas for the important case when the reference cell stream is a CBR cell stream. From Eq. 8.2

we get

$$E(X_n^{k+1}) = E(Y_i)(1 + E(X_n^k)) \quad (8.3)$$

$$Var(X_n^{k+1}) = (1 + E(X_n^k))Var(Y_i) + E^2(Y_i)Var(X_n^k) \quad (8.4)$$

Consider the case where the reference connection is a CBR connection with rate $1/T = 1-p$ and $Var(X_n^0) = 0$. Assume that the load on each of the queues are 1, and assume that the variance of the interfering stream $Var(Y_i) = p$ corresponding to Poisson traffic. From Eq. 8.4 it can be seen that

$$Var(X_n^k) = \frac{p}{1-p}\sum_{j=0}^{k-1} p^{2j} = \frac{p(1-p^{2k})}{(1-p)(1-p^2)} = \frac{T^2(T-1)}{2T-1}\left(1 - \left(\frac{T-1}{T}\right)^{2k}\right) \quad (8.5)$$

and from Eq. 8.5 we obtained the squared coefficients of variation of the CBR stream after queue k:

$$c_k^2 = \frac{T-1}{2T-1}\left(1 - \left(\frac{T-1}{T}\right)^{2k}\right) \quad (8.6)$$

For the limit case in which the number of queues to be passed is infinite and in the case with Poissonian interference, the CDV affected CBR cell stream will have a squared coefficient of variation

$$c_\infty^2 = \frac{T-1}{2T-1} \quad (8.7)$$

An important implication of Eq. 8.7 is that *the squared coefficient of variation in the Poisson interfering case never exceeds 1/2* [78, 82].

8.2.3 Numerical Examples

Consider the case in which the reference stream has interarrival time $T = 10$, corresponding to $p = 0.9$. Then Table 8.1 shows how the squared coefficient of variation of the reference stream varies as it passes through the queues of the network.

c_k^2 $(T = 10)$	M = 0	M = 1	M = 2	M = 4	M = 6	M = 10	M = 25
Computed	0	0.090	0.163	0.270	0.340	0.416	0.471
Simulation	0	0.088	0.159	0.264	0.333	0.412	0.470

Table 8.1: The Squared Coefficient of Variation as a Function of the Number of Queues

For the case $T = 20$, corresponding to $p = 0.95$ the values of the squared coefficient of variation of the reference stream have also been computed (Table 8.2) and the results have been verified by simulation as shown in both Tables. The confidence intervals of the simulation ranges from ± 0.001 to ± 0.005 for all M.

c_k^2 $(T = 20)$	M = 0	M = 1	M = 2	M = 4	M = 6	M = 10	M = 25
Computed	0	0.048	0.090	0.164	0.224	0.313	0.450
Simulation	0	0.047	0.089	0.162	0.220	0.308	0.444

Table 8.2: The Squared Coefficient of Variation as a Function of the Number of Queues

As the Tables show then the analytical approach is in accordance with the simulation results for the investigated cases with load of 1. When the load is less than one, the increase in squared coefficient of variation as a function of the number of queues is slower.

8.3 Evaluation of the Renewal Approximation

Taking into account the dependencies between successive waiting times of the CBR cells leads rather complicated solution methods. This is one of the the reason that the CDV affected CBR cell stream are often be modeled as a renewal process [9, 37, 44, 58]. The renewal approximation has the nice property that in the special case when the CBR cell stream is going through a single queue, and the interfering traffic is Poisson, the renewal approximation becomes exact in the limit where the overall load on the multiplexer approaches 1 [82]. Moreover, one may argue that a cell stream becomes a renewal stream after passing through an infinite number of queues in the heavy limit case (see the results of the previous Section), therefore it could be a good approximation for characterizing a cell stream which has been passed through several queues.

In order to investigate the relevance of the renewal approximation a simulation was carried out [82]. In Figure 8.2 the index of dispersions for intervals (IDI) are shown for the case of $T = 10$. The IDI was measured after passing through $M = 1$, 2, 4, 6, 10 and 25 queues.

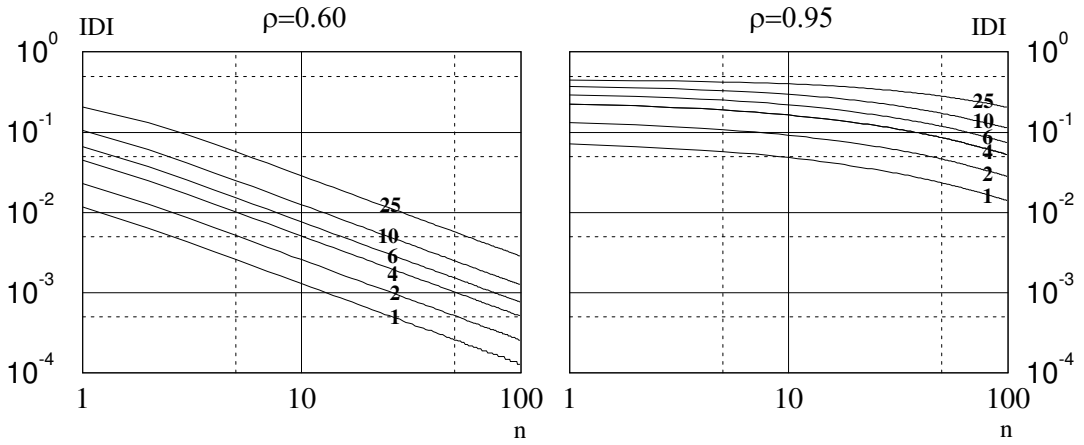

Figure 8.2: Index of Dispersions for Intervals (IDI) for Two Different Loads

For the case with a load of 0.6 the IDI have the same shape for all values of M, the

curves are simply shifted upwards when M increases. Therefore we may conclude that the CDV affected cell stream has not approached a renewal process even after 25 queues. For the case with a load of 0.95 the picture is the same. However, here the curves are much closer to a renewal even for $M = 1$. *For modeling purposes it is therefore only justified to use a renewal approximation when the load is close to 1.*

An important practical conclusion of the result is that for UPC dimensioning with small overallocation factor and receiving a single CDV affected CBR cell stream the use of a renewal approximation it is justified only when the load is very close to 1.

8.4 The Superposition of CDV Affected CBR Cell Streams

Within a network if we consider an ATM multiplexer than we often have a superposition of cell streams where even the input streams have CDV because they have been passed through several queues before entering the multiplexer. What will be the effect of the multiplexing in these cases? How much will be the CDV in the output cell stream? These problems are addressed and investigated in this section [78, 82].

8.4.1 The Model

Consider the superposition of N number of CBR cell streams which prior to the arrival to the multiplexer have been exposed to CDV in M number of multiplexing stages as shown in Figure 8.3. This queueing system is denoted by NCDVM/D/1.

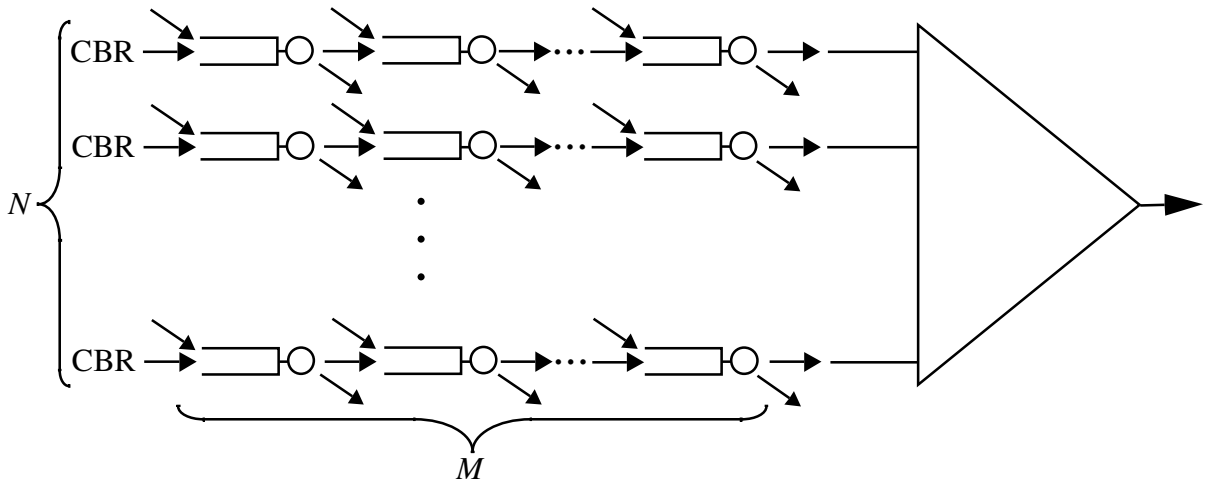

Figure 8.3: Superposition of CDV Affected (after many multiplexing stages) CBR Cell Streams

8.4.2 The Analysis Method

For queueing analysis the Beneš Result is used as in Section 7.6, but now the CDV affected CBR cell streams are modeled as renewal processes with Erlang interdeparture distributions chosen with the appropriate squared coefficient of variation which is obtained by the method presented in Section 8.2. The distributions of number of departures from a single stream are computed by Eq. 8.8 and Eq. 8.9–Eq. 8.10 (see e.g. Chapter 3 in [16] for details).

The distribution of number of departures in a window of size t starting just after the departure of cell no. j at τ_j in case of Erlang interdeparture distribution with n stages is obtained by

$$P\{N(\tau_j, \tau_j + t) = r\} = \sum_{m=rn}^{rn+n-1} \frac{\left(\frac{nt}{T}\right)^m e^{-\frac{nt}{T}}}{m!} \tag{8.8}$$

where T denotes the period of the CBR cell stream. In case of a window of size t starting at an arbitrary time t_0 the distribution of number of departures for $r \geq 1$

$$P\{N(t_0, t_0+t) = r\} = \frac{1}{n} \sum_{l=rn}^{rn+n-1} (rn+n-l) \frac{\left(\frac{nt}{T}\right)^l e^{-\frac{nt}{T}}}{l!} + \frac{1}{n} \sum_{l=rn-n}^{rn-1} (l-rn+n) \frac{\left(\frac{nt}{T}\right)^l e^{-\frac{nt}{T}}}{l!} \tag{8.9}$$

and for $r = 0$

$$P\{N(t_0, t_0 + t) = 0\} = \frac{1}{n} \sum_{l=0}^{n-1} (n-l) \frac{\left(\frac{nt}{T}\right)^l e^{-\frac{nt}{T}}}{l!} \tag{8.10}$$

By the convolution of the individual distributions we can get the distribution of the number of departures from N multiplexed cell streams. Finally, an approximation for the virtual waiting time distribution can be obtained by Eq. 7.34.

8.4.3 Numerical Examples

As in Section 7.6.3 we consider a scenario where 8 or 16 CBR cell streams, which have been exposed to CDV due to many multiplexing stages (all of them with load of 1), are multiplexed in a queue with a load of 0.8, see Figure 8.3.

From the values given in the Table 8.1 and 8.2 each CDV affected stream is approximated by a renewal stream with an Erlang-n interarrival time distribution.

The values for n have been taken in case of $T = 10$ for $n = 11$ (corresponding to passage of a single queue), $n = 4$ (corresponding to passage of 4 queues), and $n = 2$ (corresponding to an infinite number of queues).

In case of $T = 20$ the values of $n = 21$ (corresponding to passage of a single queue), $n = 6$ (corresponding to passage of 4 queues), and $n = 2$ (corresponding to an infinite number of queues) have been chosen.

As in Section 7.6.3, the local load approximation of the Beneš Result have been used for finding the virtual waiting time distribution.

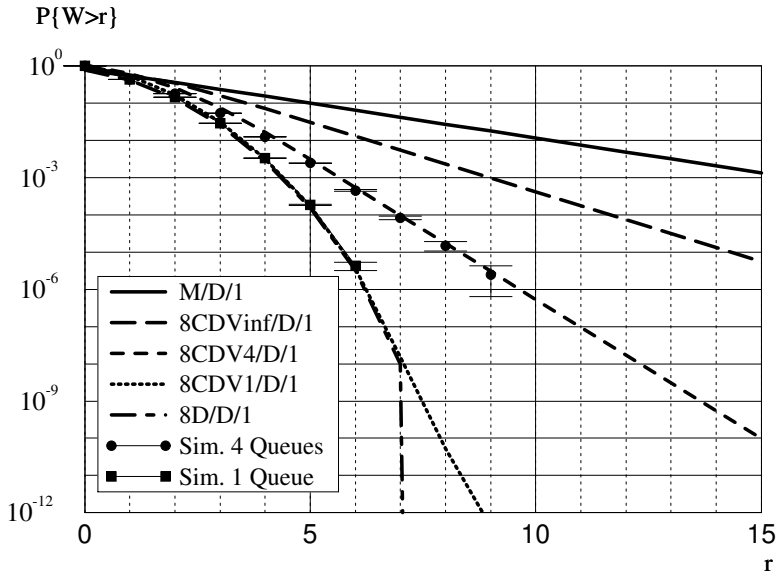

Figure 8.4: Comparison of the Virtual Waiting Time Distributions of 8CDVM/D/1 Queues (load=0.8)

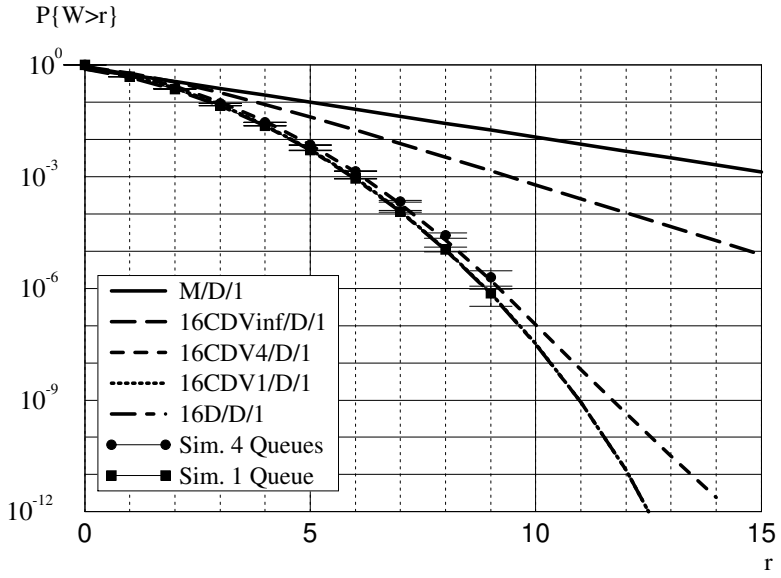

Figure 8.5: Comparison of the Virtual Waiting Time Distributions of 16CDVM/D/1 Queues (load=0.8)

For the superposition of 8 and 16 cell streams the cases where each CBR cell stream has passed through 1 queue and 4 queues respectively with interfering traffic have been simulated. The simulation results of the virtual waiting time are given with 95% confidence intervals. The analytical results are in agreement with the simulations.

As the result shows then the effect of CDV becomes more and more apparent as the number of queues which the CBR streams has passed increases. It should also be noted that when T is large the effect of CDV becomes apparent only after passing through a considerable number of queues. However, an important result of the model described in Section 8.2 is that *as long as we have Poissonian interference then the delay performance of the multiplexing of a number of CDV affected CBR cell streams will not be worse than the superposition of the same number of renewal processes with squared coefficients of variation 1/2.*

8.5 A Method Based on the $nTri/D/1$ Queue

To evaluate an ATM multiplexer a $G/D/1$ queueing model is considered (General cell interarrival distribution, Deterministic service time, single server system), where the basic problem is to find the best approximation of the arrival processes. Due to the cascaded queueing in ATM networks to find an accurate model for the cell arrival process is a quite difficult and challenging task. Various models have been developed to describe the cell scale queueing problem [99].

A frequently used simple model is the $Geo(N)/D/1$ queueing model. The $Geo(N)/D/1$ queue has Bernoulli arrival process (the short-hand notation "Geo" stands for the geometrically distributed interarrival time). The time is slotted and it is assumed that an arrival at any input port in any timeslot has the same constant probability. The arrivals are independent of each other. The probability of a specific number of arrivals to a given fixed output is given by the binomial distribution. This can be an appropriate model of an ATM switch because the number of cells arriving to the switch during cell transmission time is bounded by the number of inlets to the switch and this model takes it into account. Increasing the number of inlets (N) this queue becomes an $M/D/1$ queue as N goes to infinity. The calculation of the $Geo(N)/D/1$ queueing model is simple [99].

The $M/D/1$ queueing model assumes a Poisson arrival process. This approximation could be acceptable when the arrival process is a superposition of a large number of sparse renewal processes. Also we can choose this model in the case when the arrival process is unknown, only the arrival rate is known. In contrast to the previous model, in this model the number of arrivals during a service time could exceed any finite bound, which is not a realistic assumption, but many investigations have shown [99], that this model is acceptable if the system load is low. The calculation of the queue length distribution in this model is simple and well known [99].

To model the superposition of CBR traffic the $nD/D/1$ queue is considered as basic model. The input process is a superposition of n independent periodic sources with period D. The phases of different sources are random. Compared to the $M/D/1$ queueing model this is an improvement, because it takes into account the periodic nature of the cell emissions, the exact periodicity, however, is an idealistic assumption. The regularly behavior of this arrival process compared to the previous models yields shorter queues with the same mean arrival rate. The calculation of the queue length distribution in this model is not trivial, but applying the Beneš formula a tractable solution can be obtained [99].

A generalization of the $nD/D/1$ queue is the $\sum D_i/D/1$, where the arrival process is a heterogeneous mix of N periodic sources with different arrival rates. The model is general, but there is no exact solution for it. Exact solution is known only for a restricted case [99]. However, some accurate approximations have been developed for this general model [99].

To overcome the inconveniences of the $M/D/1$ and $nD/D/1$ queues but also retain their advantages the $nTri/D/1$ queue has been developed [28], which is the object of the next Section.

8.5.1 The $nTri/D/1$ Queueing Model

Observing the real ATM traffic it can be concluded that the $M/D/1$ queueing model yields results, which are too pessimistic, because this model does not take into account the negative correlations between cells. In contrast, the results obtained by the $nD/D/1$ queueing model is too optimistic because of the idealistic periodicity assumption. In a real ATM network Cell Delay Variation (CDV) can be observed, which is a result of a number of queues. Therefore a CBR traffic becomes a jittered traffic after going through the network (the periodic cell stream becomes a spread cell stream). In order to overcome such drawbacks of these models the $nTri/D/1$ queueing model has been developed [28] for the finite buffer case. In this model the arrival process is a superposition of n independent streams, in which each stream has exactly one arrival in each frame. The arrival times have uniform distribution within a frame and are independent from the next frame (the "Tri" denotes the triangular interarrival time distribution). This model takes into account both the negative correlation between cells and the CDV phenomenon. The assumption that the CDV has uniform distribution is certainly not true, but acceptable as an approximation for cases when we know nothing about the CDV. This model is the object of the next Section with the same restrictions as in [28], that is, all streams are synchronized.

8.5.2 The Solution of the $nTri/D/1$ Queue

Different solutions have been developed [99, 92, 1] to solve the call scale queueing models, most of them are based on Markov-chains. Recently a new solution method have been investigated, which is originally invented by Beneš [5]. This approach provides a very powerful tool to solve both cell and burst scale queueing problems [99]. In this Section a new solution is presented for the $nTri/D/1$ queue based on the Beneš approach [71, 72]. The main motivation to perform this task was to get a simpler solution than the solution based on the Markov-chain approach [28], which is rather complex and the calculations are very time-consuming.

The Beneš formula gives the overflow probability for $G/D/1$ queueing systems in the following form:

$$P\{Q_t > r\} = \sum_{s=1}^{\infty} P\{N_s = s + r\} P\{Q_{t-s} = 0 | N_s = s + r\} \tag{8.11}$$

where

- Q_t: queue length at time t
- N_s: number of cell arrivals in $(t - s, t)$
- s: the investigated time-window (cell service time is chosen as time unit)

If this formula is interpreted it can be seen that the probability of exceeding a certain buffer level depends on two probabilities in the summation. The first one is the probability that

exactly $s+r$ cells arrive in a window of size s. The second one is a conditional probability of that the queue is empty at time t, given that exactly $s+r$ cell arrivals occurred in the window. The derivation of this formula can be found in [99, 92, 72]

First a new exact solution is presented for the first probability and three different approximations will be used for the second one. Before presenting the results some more model-description notations are introduced: (see Figure 8.6.)

- D: frame size (the period of CBR cell streams)
- A: distance from the start of the window to the first boundary of the next frame(a specific value of A denoted by a)
- k: number of cell arrivals in a frame = number of sources (n)
- ρ: load (the number of arrivals divided by the frame size)

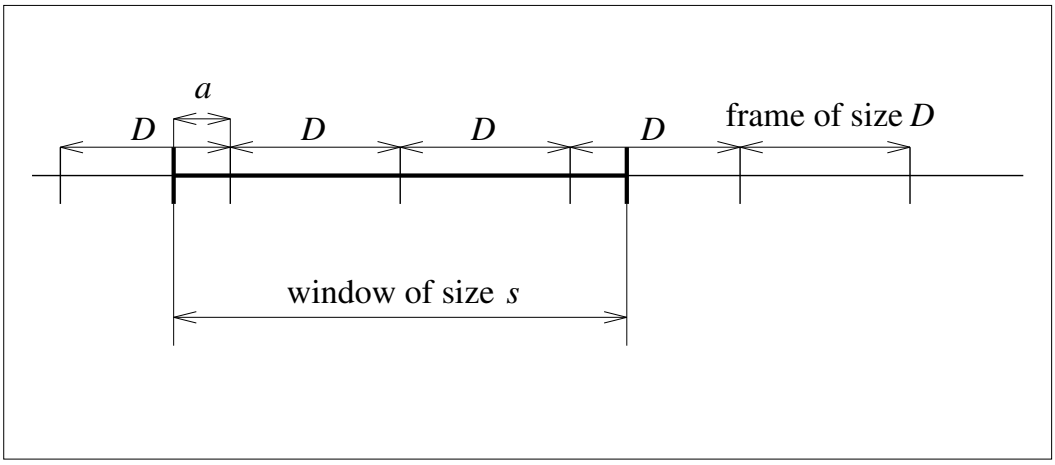

Figure 8.6: The Window on the Frame Flow

The probability of a specific number of arrivals in the window is the following:

$$P\{N_s = i\} = \begin{cases} \int_0^s P_\alpha\{N_s = i | A = a\} P\{A \in (a, a+da)\} + \\ \int_s^D P_\beta\{N_s = i | A = a\} P\{A \in (a, a+da)\} & \text{if } s \leq D \\ \int_0^D P_\gamma\{N_s = i | A = a\} P\{A \in (a, a+da)\} & \text{if } s > D \end{cases} \tag{8.12}$$

where $P\{A \in (a, a+da)\} = \frac{da}{D}$ and

$$P_\alpha\{N_s = i|A = a\} = \begin{cases} \sum_{l=0}^{i} \binom{k}{l} P_1^{k-l} P_2^l \binom{k}{i-l} P_3^{i-l} P_4^{k-i+l} & \text{if } i \le k \\ \sum_{l=i-k}^{k} \binom{k}{l} P_1^{k-l} P_2^l \binom{k}{i-l} P_3^{i-l} P_4^{k-i+l} & \text{if } k < i \le 2k \\ 0 & \text{if } i > 2k \end{cases} \tag{8.13}$$

with $P_1 = \frac{D-a}{D}$, $P_2 = 1 - P_1$, $P_3 = \frac{s-a}{D}$ and $P_4 = 1 - P_3$,

$$P_\beta\{N_s = i|A = a\} = \begin{cases} \binom{k}{i} P_1^i P_2^{k-i} & \text{if } i \le k \\ 0 & \text{if } i > k \end{cases} \tag{8.14}$$

with $P_1 = \frac{s}{D}$ and $P_2 = 1 - P_1$,

$$P_\gamma\{N_s = \left\lfloor \frac{s-a}{D} \right\rfloor k + i|A = a\} = \begin{cases} \sum_{l=0}^{i} \binom{k}{l} P_1^{k-l} P_2^l \binom{k}{i-l} P_3^{i-l} P_4^{k-i+l} & \text{if } i \le k \\ \sum_{l=i-k}^{k} \binom{k}{l} P_1^{k-l} P_2^l \binom{k}{i-l} P_3^{i-l} P_4^{k-i+l} & \text{if } k < i \le 2k \\ 0 & \text{if } i > 2k \end{cases} \tag{8.15}$$

with $P_1 = \frac{D-a}{D}$, $P_2 = 1 - P_1$, $P_3 = \frac{s-\left(a+\lfloor \frac{s-a}{D} \rfloor D\right)}{D}$ and $P_4 = 1 - P_3$. The derivation of results can be found in Appendix B.

For the second probability of Eq. 8.11 three approximations will be shown. Applying these approximations an upper bound, a lower bound and an acceptable accurate formula can be obtained for Eq. 8.11.

An upper bound

A very simple upper bound of Eq. 8.11 is obtained if the following obvious approximation is used

$$P\{Q_{t-s} = 0|N_s = s + r\} \approx 1 \tag{8.16}$$

yielding the upper bound:

$$P\{Q_t > r\} < \sum_{s=1}^{\infty} P\{N_s = s + r\} \tag{8.17}$$

A lower bound

It is known that in any stationary single server queueing system the probability of finding the system empty equals $1-\rho$ where ρ is the load. Using this an approximation the following can be obtained for the second probability of Eq. 8.11:

$$P\{Q_{t-s}=0|N_s=s+r\}\approx 1-\rho \tag{8.18}$$

I conjecture that this approximation yields a lower bound of Eq. 8.11 because the system is overloaded in the window and the number of arrivals is fixed in the frame. The lower bound:

$$P\{Q_t>r\}>(1-\rho)\sum_{s=1}^{\infty}P\{N_s=s+r\} \tag{8.19}$$

An approximation based on local load

Using local load instead of load a better approximation can be obtained:

$$P\{Q_t>r\}\cong\sum_{s=1}^{\infty}[P\{N_s=s+r\}-\rho P\{N_s=s+r|\text{one arrival at (t-s)}\}] \tag{8.20}$$

where $P\{N_s=s+r\}$ is given by Eq 8.12 and $P\{N_s=s+r|\text{one arrival at (t-s)}\}=$

$$=\begin{cases}\int_0^s P_\delta\{N_s=i|A=a\}P\{A\in(a,a+da)\}+ & \\ \int_s^D P_\xi\{N_s=i|A=a\}P\{A\in(a,a+da)\} & \text{if } s\le D\\ \int_0^D P_\eta\{N_s=i|A=a\}P\{A\in(a,a+da)\} & \text{if } s>D\end{cases} \tag{8.21}$$

where $P\{A\in(a,a+da)\}=\frac{da}{D}$ and

$$P_\delta\{N_s=i|A=a\}=$$
$$\begin{cases}\sum_{l=0}^{i}\binom{k-1}{l}P_1^{k-l-1}P_2^l\binom{k}{i-l}P_3^{i-l}P_4^{k-i+l} & \text{if } i\le k-1\\ \sum_{l=i-k}^{k-1}\binom{k-1}{l}P_1^{k-l-1}P_2^l\binom{k}{i-l}P_3^{i-l}P_4^{k-i+l} & \text{if } k-1<i\le 2k-1\\ 0 & \text{if } i>2k-1\end{cases} \tag{8.22}$$

with $P_1=\frac{D-a}{D}$, $P_2=1-P_1$, $P_3=\frac{s-a}{D}$ and $P_4=1-P_3$,

$$P_\xi\{N_s=i|A=a\}=\begin{cases}\binom{k-1}{i}P_1^iP_2^{k-i-1} & \text{if } i\le k-1\\ 0 & \text{if } i>k-1\end{cases} \tag{8.23}$$

with $P_1 = \frac{s}{D}$ and $P_2 = 1 - P_1$,

$$P_\eta\{N_s = \left\lfloor \frac{s-a}{D} \right\rfloor k + i | A = a\} =$$

$$\begin{cases} \sum_{l=0}^{i} \binom{k-1}{l} P_1^{k-l-1} P_2^l \binom{k}{i-l} P_3^{i-l} P_4^{k-i+l} & \text{if } i \leq k-1 \\ \sum_{l=i-k}^{k-1} \binom{k-1}{l} P_1^{k-l-1} P_2^l \binom{k}{i-l} P_3^{i-l} P_4^{k-i+l} & \text{if } k-1 < i \leq 2k-1 \\ 0 & \text{if } i > 2k-1 \end{cases} \quad (8.24)$$

with $P_1 = \frac{D-a}{D}$, $P_2 = 1 - P_1$, $P_3 = \frac{s-\left(a+\lfloor \frac{s-a}{D} \rfloor D\right)}{D}$ and $P_4 = 1 - P_3$. The derivation of these results can be found in Appendix C.

The main advantage of the new solution is that the computation time is efficiently decreased. Moreover the approach can be generalized to the model, where the relative phases of input streams are random. Furthermore the approach can be used for the case where there are more than one arrivals in an input stream, leading to a realistic extension of the model.

8.5.3 Numerical Examples

The numerical results given below have been calculated for the $nTri/D/1$ queue by the solution methods of the previous Section. Numerical tests have also justified the correctness of the solution by comparing the results to the outcome of the solution based on Markov-chains [28]. Results for the lower bound, upper bound, the local load approximation and related results of the $nD/D/1$ and $M/D/1$ queues have been obtained and shown in this Section.

The number of sources is 100, the frame size is 120 (consequently the load is 0.833) in the considered example. In Figure 8.7, the probability of exceeding a certain buffer level is shown as a function of the buffer occupancy level for the upper bound, the lower bound and the approximation.

For comparison, the results of the $M/D/1$ and the $nD/D/1$ queueing models are also shown in Figure 8.8. For the $nTri/D/1$ queue the local load approximation is plotted.

It can be observed that the $M/D/1$ queueing model differs significantly from the $nD/D/1$ queueing model. It is because the arrival process of the $nD/D/1$ queue is more regular. The performance of the $nTri/D/1$ queue is somewhere between that of the $M/D/1$ and the $nD/D/1$ queues. This shows that the $nTri/D/1$ queue is not as regular as the $nD/D/1$ queue (because the arrival time of the cell is uniformly distributed within the frame and independent from the next frame), but it also takes into account the periodic nature of the cell stream (because there is only one arrival in a frame from a specific source) and that is why it is not as random as the $M/D/1$ queue.

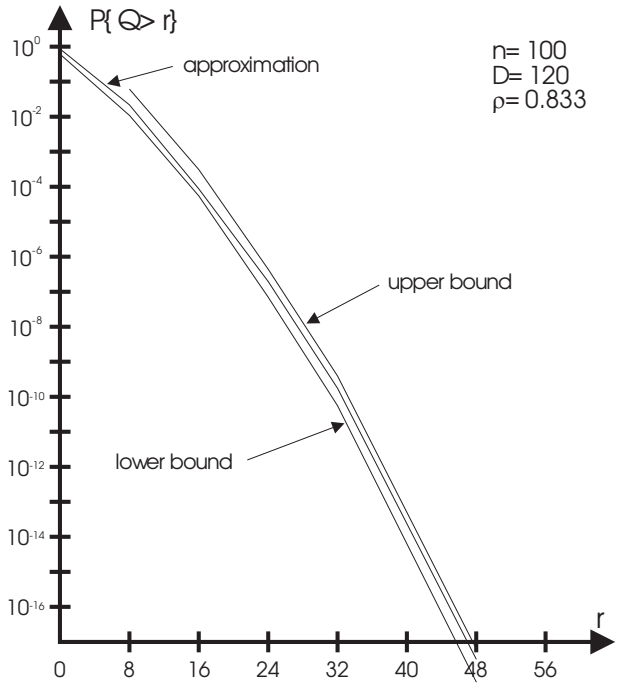

Figure 8.7: The Probability of Exceeding a Certain Buffer Level ($P\{Q > r\}$) as a Function of the Buffer Occupancy Level (r) for the $nTri/D/1$ Queue

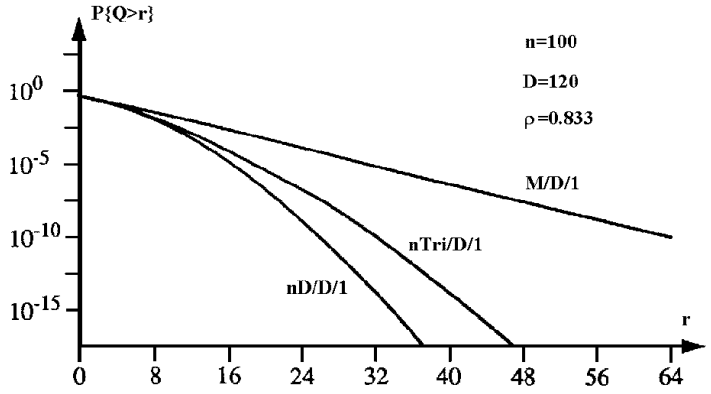

Figure 8.8: The Probability of Exceeding a Certain Buffer Level ($P\{Q > r\}$) as a Function of the Buffer Occupancy Level (r) for Different Queueing Models

8.6 Summary of Results

In this Chapter the CDV phenomenon due to cascaded multiplexing stages is examined. First, we have investigated how the characteristics of the CBR cell stream changes as it has been passed through several multiplexing stages. A formula and an upper bound for the squared coefficient of variation of the interdeparture time are derived. Numerical examples showing the main properties of the alteration of the cell stream profile are given.

Second, the suitability of the widely applied renewal model is investigated. It has been shown that the renewal approximation is not acceptable characterization of CDV affected CBR cell stream except for the case when the load approaches 1 in the multiplexing stages.

Third, an analysis of an ATM multiplexer receiving CBR cell streams which prior to the arrival to the multiplexer have been exposed to CDV in several multiplexing stage is presented. Numerical results illustrate how the delay performance changing as the cell stream has been passed through different number of multiplexing stages.

Finally, a new solution of the $nTri/D/1$ queue is derived which is a a possible candidate model for describing the multiplexed CDV affected CBR cell streams if we have no knowledge about the CDV. Lower bound, upper bound and an approximation for the buffer occupancy are derived and an evaluation of the model is also given.

Chapter 9

Designing Guidelines for ATM Traffic Control and Network Element Dimensioning

9.1 Introduction

The CDV phenomenon has several undesired effects on the successful actions of Traffic Control and appropriate design of Network Elements. In order to design them properly a CDV analysis is required to make a decision whether the CDV has to be taken into account or not. Moreover, if the CDV cannot be neglected the question is how to use the CDV information in the design of traffic control functions or dimensioning of network elements. This Chapter addresses this issue providing some proposals. The influence of the CDV used in the proposals are derived from the results of the analysis in Chapter 7- 8.

Section 9.2- 9.3 and Section 9.4- 9.5 describe the Traffic Control functions under investigation with the CDV impact on these functions and some network element dimensioning problems with their CDV sensitivity, respectively. Finally, the proposals can be found in Section 9.6 and Section 9.7.

9.2 Traffic Control

The primary role of Traffic Control and Congestion Control is to protect the network and the user in order to achieve network performance objectives. An additional role is to optimize the use of network resources [50, 2]. Traffic Control refers to the set of actions taken by the network to avoid congested conditions. ITU-T and ATM Forum have defined Traffic Control functions, see [50, 2]. In this section the functions which are the objectives of this Chapter are briefly outlined:

- Usage Parameter Control (UPC)
- Call Admission Control (CAC)

- Cell Spacing.

The locations of these Traffic Control functions in the B-ISDN access configuration can be seen in Figure 9.1.

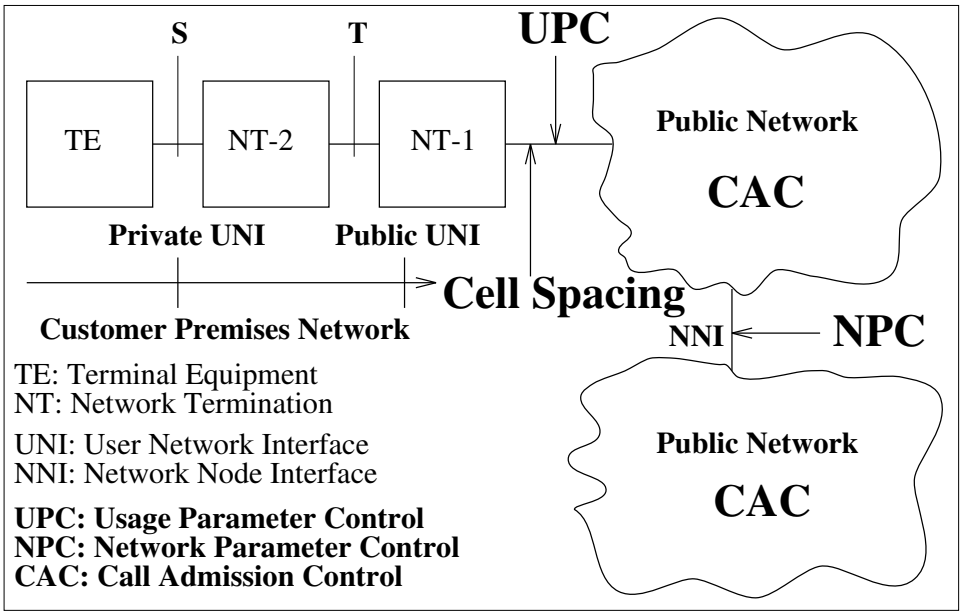

Figure 9.1: Location of Traffic Control Functions

The *Usage/Network Parameter Control* (UPC/NPC) is defined as the set of actions taken by the network to monitor and control traffic, in terms of traffic offered and validity of the ATM connection. The main purpose of UPC/NPC is to protect network resources from malicious as well as unintentional misbehaviour by detecting violations of negotiated parameters and taking appropriate actions [50, 2].

The *Call Admission Control* is defined as the set of actions taken by the network during the call set up phase in order to establish whether a Virtual Channel (VC) or Path (VP) request can be accepted or rejected [50, 2].

The *Cell Spacing* is a function which allows to reduce CDV by buffering the incoming cells of a connection and re-scheduling them on the basis of the connection peak cell rate.

9.3 The Impact of CDV on Traffic Control

The setting of UPC parameters and the CAC decision are complicated due to CDV occuring between a terminal and a UPC point (see Figure 9.1). This CDV prevents the network from correctly estimating the actual traffic characteristics offered to the network even though the characteristics of the cell stream at the terminal is honestly declared. The main reason for this CDV is that the cell stream is disturbed in NT-2 (e.g. PBX or LAN) as going through from the terminal to the network.

Moreover, the CDV within the network also complicates the CAC decision and the setting of NPC parameters because the internetwork traffic is altered in the network mainly due to the buffering of cells in ATM nodes.

CDV introduces fluctuations on the cell interarrival times resulting in much higher instantaneous cell rates then the negotiated peak cell rate. In the case if the CDV is not bounded at the User Network Interface (UNI) [2] and at the Network Node Interface (NNI) [2] the design of a suitable UPC/NPC mechanism and a proper resource allocation is impossible. Furthermore, even though the CDV is bounded its effect shall be taken into account by the UPC/NPC procedure. This is the primary purpose of introducing the CDV Tolerance parameter in the Traffic Contract and the Generic Cell Rate Algorithm is defined for testing cell conformance [50, 2]. This algorithm uses the *peak emission interval* and the *CDV tolerance* parameters thereby CDV analysis is required to set these parameters. Examples of applications of thesis results presented in Chapter 7 for setting these UPC parameters can be found in [11].

The difference between the expected traffic and the actual one also influences the utilization of transmission resources so the proper design of the CAC function also depends on the CDV. A CDV analysis is needed to estimate the actual traffic characteristics to be taken into account.

Concerning the dimension of Cell Spacer the amount of buffer space required to perform this function without cell losses depends on the input CDV. Therefore the maximum admissible CDV shall be specified which also requires a CDV analysis.

9.4 Network Element Dimensioning

In this section the following Network Element Dimensioning problems are addressed:

- Buffer dimensioning
- Play out buffer dimensioning

In the issue of *buffer dimensioning* we consider an ATM multiplexer with outgoing link buffers using FIFO queueing disciplines. The generic task is to determine the buffer requirements for a given input stream of the buffer [31, 66]. The main problem of the buffer dimensioning is to accurately characterize the cell stream offered to the buffer. Moreover, depending on the philosophy of handling burst scale congestion (burst scale delay or loss) [99] and also buffer sizes (e.g. short buffers for delay sensitive applications or large buffers for data communications) different dimensioning approaches should be used [31].

The problem of *play out buffer dimensioning* arises in the design of circuit emulation facilities [14, 59]. The B-ISDN uses the ATM network as a backbone network to provide several broadband services with different ATM Adaptation Layers (AALs) that enhance the basic facilities offered by the backbone network. Some real time services may use circuit emulation facilities which requires a buffer at the receiving end of the connection to play the variable end-to-end cell delays out. This play out buffer sends cells to the receiver

at a constant rate that matches the emission rate by compensating the experienced variable cell transfer delays. The AAL1 has been designed to provide this facility [14].

9.5 The Impact of CDV on Network Element Dimensioning

The CDV has a significant impact on the dimensioning of both buffers in ATM multiplexers and play out buffers at the receiving end of the connections in case of circuit emulation. These buffers are dimensioned based on some kind of input traffic models. These traffic profile is altered in the network due to CDV. In order to make a proper dimensioning of these buffers an accurate characterization of cell stream as going through the network is necessary.

9.6 Designing Proposals for ATM Traffic Control Functions

The proposals are grouped according to the proper model of the Customer Premises Network (CPN):

9.6.1 Customer Premises Network Modeled by a Single FIFO Multiplexer

Based on the results of Chapter 7 I have the following proposals and conclusions concerning the design of traffic control functions:

A. When the Customer Premises Network (CPN) can be modeled as a single FIFO multiplexing stage and *peak rate allocation* is applied:

1. The CAC function does not have to take into account the CDV effect.
As I have shown in Chapter 7 the CDV effect after a single FIFO queue is negligible. Therefore no need to make any effort to design a CAC function using any CDV information.

2. No cell spacing is needed.
Because of the small amount of CDV no reason to perform any spacing function. The CAC dimensioning without CDV can be performed even without any spacer.

3. UPC dimensioning based on modeling the CPN with the $M + D/D/1$ queue yields overestimation of CDV Tolerance.

The results of Chapter 7 clearly indicates that if the CPN is modeled by the $M+D/D/1$ queueing model it will result in overestimation of the CDV Tolerance.

4. An accurate model of CPN can be used for UPC dimensioning with modeling the background traffic by batch Bernoulli process ($GI^{[x]} + D/D/1$ queue) thereby taking into account the burstiness of the background traffic.

(a) The Poisson background traffic is not sufficient for proper modeling if we have a small number of multiplexed connections.

(b) For choosing a proper model for background traffic an attempt to match the peakedness should be made.

(c) In case of a small number of sources the background traffic is smooth and should be modeled by batch size distributions (e.g. Binomial) where the peakedness of the background traffic is smaller than 1.

This proposal is based on the results of the Chapter 7 where it has been shown that the burstiness of the background traffic has a significant impact on the CDV.

B. When the Customer Premises Network (CPN) can be modeled as a single FIFO multiplexing stage and *statistical multiplexing* is applied:

1. An accurate model of CPN can be used for UPC dimensioning with modeling the background traffic by batch Bernoulli process ($GI^{[x]} + D/D/1$ queue) thereby taking into account the burstiness of the background traffic.

(a) For choosing a proper model for background traffic an attempt to match the peakedness should be made.

(b) In cases when the background traffic has positive correlations the background traffic is bursty and could be modeled by distributions (e.g. Pascal) where the peakedness is bigger than 1.

These proposals are also based on the results of the Chapter 7.

9.6.2 Customer Premises Network Modeled by Cascaded FIFO Multiplexers

When the CPN can be modeled as cascaded FIFO multiplexing stages and peak rate allocation is applied I give the following proposals:

1. If no cell spacing applied the CAC function has to take into account the CDV effect.
Because of the tandemed FIFO queues may result in significant CDV (see Chapter 8) its effect should be taken into account in the design of the CAC function. However, if we

do so the CAC design may become rather complicated thereby in some cases where the extra delay introduced by the cell spacer can be allowed it is better to apply a cell spacer which reduces the CDV drastically and in this case the CAC function can be designed without any CDV information.

In cases where no spacing function applied the CAC design shall be based on the so called Worst Case Traffic (WCT), i.e., the most demanding traffic pattern among the ones compliant with the Traffic Contract. Therefore the identification of the WCT is needed. Some recent studies show that in some cases different traffic patterns can be worst than the full-rate on-off pattern which is believed the WCT so far [24].

2. The application of cell spacing is recommended.

As described above if the extra delay due to the cell spacer is allowed the application of the spacer is recommended at the network ingress resulting in more simpler CAC procedures.

3. UPC dimensioning based on modeling the CPN with a number of $M + D/D/1$ queues yields overestimation of CDV Tolerance.

Similarly as in the previous section we can conclude the overestimation property of the $M + D/D/1$ queueing model (see Chapter 8).

4. An appropriate approximation (upper bound) for UPC dimensioning can be obtained by modeling the output cell stream of the CPN with a renewal process with squared coefficient of variation

$$c_k^2 = \frac{T-1}{2T-1}\left(1 - \left(\frac{T-1}{T}\right)^{2k}\right)$$

where 1/T is the rate of the CBR cell stream and k is the number of queues that have been passed. The formula is valid in cases where the cell stream has been altered by background traffic which can be modeled by Poisson processes at each stage through the network.

Here also the results of Chapter 8 are applied for characterizing the cell stream after several multiplexing stages.

9.7 Designing Proposals for Network Element Dimensioning

From the CDV analysis results I have the following conclusion and proposal concerning buffer and play out buffer dimensioning:

1. Buffer dimensioning in ATM networks: When the background traffic can be well modeled by Poisson processes at each stage through the network the buffer dimensioning based on modeling the single or tandemed ATM multiplexers with the $M/D/1$ queue provides overestimation of needed buffers.

This conclusion based on the results of Chapter 8 and point out that the dimensioning of buffers based on the widely applied and suggested $M/D/1$ queueing model yields overestimation of buffer sizes.

2. Play out buffer dimensioning in case of circuit emulation: An appropriate model of the cell process at the receiving end of an ATM connection for dimensioning a play out buffer is a renewal process with squared coefficient of variations of 0.5 in cases where the background processes can be modeled by Poisson processes through the multiplexing stages.

From the results of Chapter 8 we can see that in the heavy load case and with Poissonian interference the squared coefficient of variations of the CBR cell stream going through ATM multiplexers will never exceed 0.5. Therefore using a renewal stream with squared coefficient of variations of 0.5 for modeling the cell stream at the receiving end of the connection is a worst case approximation for the traffic but it is more close to the reality than the Poisson assumption which is too conservative.

9.8 Summary

The alteration of the cell stream characteristics due to CDV complicates both the design of traffic control procedures and network element dimensioning. This Chapter has introduced some practically applicable proposals concerning in which cases the CDV has to be taken into account in the issues of dimensioning of traffic control functions and network elements and provided some suggestions how it can be done.

Part V

Concluding Remarks

Chapter 10

Summary of the Dissertation

This dissertation covers various fields of performance evaluation of telecommunication networks. It is particularly devoted to several unsolved and challenging performance problems that arise in ATM networks varying from the call level description of B-ISDN traffic to the characterization of Cell Delay Variation. There is a growing interest in the telecommunication world in solving these performance issues and the dissertation is a contribution which could help the development of a proper performance engineering in ATM networks.

Part I introduces the different concepts of quality of telecommunication networks and reviews the main performance evaluation issues of the dissertation.

It is impossible to design speech coding systems and communication networks based on only subjective speech quality assessment and it is desirable that speech quality be assessed by objective methods based on measured physical parameters. Several candidates of speech quality objective measures have been developed to fulfill this requirement. However, the problem of finding a good objective measure is that it should correspond well with the subjective assessment values which depend on several phenomenons of the poorly understood human speech perception. Therefore careful attention should be taken to use these measures and identification of the applicability limitations of any objective speech quality measure is necessary.

This issue is addressed in Part II and a widely accepted and successful group of objective speech quality measures is chosen for evaluation. The results clearly indicates that none of the investigated spectral envelope objective speech quality measures are appropriate alone to characterize the speech quality. The limitations of these measures are also pointed out, namely, their sensitivity to nonlinear distortions is not satisfactory.

Part III presents different results of ATM call scale performance evaluation. The problem of a proper call scale traffic characterization is addressed and a robustness and sensitivity analysis of link occupancy investigating the impact of deviations of arrival process and holding time from the classical Poisson/exponential description is analyzed. It has been shown that the traditional Poisson/exponential description of B-ISDN is quite

vulnerable to deviations from these classical assumptions resulting in the conclusion that in cases when these assumptions are not fulfilled this classical description cannot be accepted and a more accurate characterization of call scale traffic is necessary.

The need for developing general traffic models related to several issues like developing link blocking formulas which can take into account also the variability of the traffic but still applicable in practice. Two approximation methods for computing the link occupancy distribution are presented based on matching the mean and the variance. A proposal for a variability measure based on the concept of generalized peakedness with a new closed form expression is given and this concept is suggested for the variance computation of the approximations. Based on the approximations several new link blocking measures with their evaluation are shown. These simple measures can be seen as candidate measures of B-ISDN link blocking which are intended to cope with more general call scale traffic than Poisson but also easily applicable in practice.

The applicability of the developed link blocking measures and the new concept of B-ISDN traffic characterization using a two-parameter description of traffic (the mean and the generalized peakedness) is demonstrated in an ATM network dimensioning problem. The dimensioning of large ATM networks into a number of logical subnetworks is also a hot topic of ATM research. So far only a few solutions have been published to solve this problem and all of them assume Poisson input. A new algorithm based on the generalization of the fixpoint method presented where the Poisson assumption is relaxed and the traffic can be general described by the mean and generalized peakedness. Evaluation results with the comparison of the original fixpoint method showing the effect of the variability measure are also given.

Part IV studying various ATM cell level performance evaluation issues with focusing on the modeling and evaluation of Cell Delay Variation. Two new methods are described to evaluate the CDV in a single ATM multiplexer. Both methods take into account the burstiness of the background traffic which provide a more realistic model. The first approach is an exact Markovian solution while the second one a diffusion approximation which is a candidate in several fields of ATM design considering the good practical applicability. The evaluation results show that there is a significant effect of background burstiness on CDV which is therefore cannot be neglected.

The issue of how the characteristics of a CBR cell stream alters as it is going through several ATM multiplexing stages is also addressed and a formula for computing the squared coefficient of variations of the interdeparture time is derived. It is shown that an upper bound can be given for the burstiness of the cell stream.

The relevance of the renewal description of the CDV affected CBR cell stream is investigated and simulation results show that it is only acceptable in cases if the load is very close to 1 in the multiplexing stages.

Studies on the impact of multiplexing CDV affected CBR cell streams which prior to the arrival to the multiplexer have been passes through single or cascaded multiplexing stages are presented. It can be concluded that the CDV due to a single multiplexing stage can be neglected, but after several stages it may become significant as shown by numerical

examples with simulation verification.

A new solution method derived for the $nTri/D/1$ queue, which can be used for modeling the superposition of CDV affected CBR sources in cases when we know nothing about the CDV. Upper bound, lower bound and an accurate approximation is demonstrated for the buffer occupancy distribution.

Finally the dissertation provides some guidelines for designing traffic control functions and network elements, which also illustrates the practical applicability of the results of Part IV. Namely, the issue of CAC, UPC/NPC, cell spacer, buffer and play out buffer dimensioning are addressed.

Chapter 11

Areas for Further Research

The inadequate performance of all investigated spectral envelope objective speech quality measures and their identified disadvantage calls for more research of objective measures. An idea of a two-parameter objective measure with sensitivity to the linear and nonlinear distortions of speech is presented and the evaluation of this new measure one of the possible research topic of the future.

Concerning call level traffic characterization of B-ISDN an important area of the future to analyze the traffic based on collected measurements from real B-ISDN environments. However, it should be noted, that since, for the time being, there is no traffic produced by real B-ISDN and we are restricted only to measurements from a few experimental ATM networks providing only a limited number of services. As soon as we have some data from real B-ISDN the performance of the proposed traffic description and link blocking measures should be evaluated.

A related and important future research topic is the problem of estimating the variability measure. This problem and a possible solution is briefly outlined in the dissertation but to perform it in practice requires a significant research.

The further developing of the proposed ATM dimensioning algorithm for considering e.g. revenue maximization and load sharing or using the concept of two-parameter description and the developed link blocking measures in other algorithms are possible directions of future work.

The practical applications of the new CDV models (e.g. for setting the parameters of UPC) is one of the most important research area of future work related to the results of Part IV. This area is also considered in the dissertation (see Chapter 9) and in my related publications (e.g. [11]) but still there are several unsolved issues of applying these models in practical designing problems of ATM networks.

Bibliography

[1] H. Akimaru and K. Kawashima. *Teletraffic: Theory and Applications*. Springler-Verlag, 1993.

[2] ATM Forum. *ATM User-Network Interface Specification Version 3.0*, September 1993. Englewood Cliffs, NJ: Prentice Hall.

[3] ATM Forum. *Traffic and Congestion Control*, April 1993. Draft Besaline Document.

[4] J. G. Beerends and J. A. Stemerdink. A perceptual audio quality measure based on a psychoacoustic sound representation. *Journal of the Audio Engineering Society*, 40(12):963–978, 1992.

[5] V. E. Beneš. *General Stochastic Processes in the Theory of Queues*. Addison-Wesley, 1963.

[6] V. E. Beneš. *Mathematical Theory of Interconnecting Networks*. Academic Press, 1965.

[7] J.L.v. Berg and J.A.C. Resing. The change of traffic characteristics in ATM networks. *COST 242 TD(92)040*, 1992.

[8] C. Bisdikian, W. Matragi, and K. Sohraby. A study of the jitter in ATM multiplexers. In *IFIP TC6 High Speed Networks and their Performance*, pages 219–235, Raleigh, NC, USA, October 1993.

[9] S. Blaabjerg. Estimating the effect of cell delay variation by an application of the heavy limit theorem. *COST 242 TD(92)017*, 1992.

[10] S. Blaabjerg. Cell delay variation in a FIFO queue: A diffusion approach. In *IFIP TC6 High Speed Networks and their Performance*, pages 237–256, Raleigh, NC, USA, October 1993.

[11] S. Blaabjerg and S. Molnár. Methods for UPC dimensioning of a CDV perturbated cell stream. *RACE BRAVE Workshop*, June 14-15 1995. Milano, Italy.

[12] E. Brockmeyer, H. L. Hallstrom, and A. Jensen. *The Life and Works of A.K. Erlang*. Academic Press, Copenhagen, 1948.

[13] P. Brown and A. Simonian. Perturbation of a periodic flow in a synchronous server. In P.J. Courtois and G. Latouche, editors, *PERFORMANCE'87*. Elsevier Science Publishers, 1988.

[14] COST 242. *Cell Delay Variation in ATM Networks*, October 1994. Interim report.

[15] COST 242. *Multi-Rate Models for Dimensioning and Performance Evaluation of ATM Networks*, June 1994. Interim report.

[16] D.R. Cox. *Renewal Theory*. Methuen, 1962.

[17] D.R. Cox and P.A.W. Lewis. *The Statistical Analysis of Series of Events*. Methuen, 1966.

[18] R. E. Crochiere, L. R. Rabiner, N. S. Jayant, and J. M. Tribolet. A study of objective measures of speech waveform coders. In *Proceedings of the 1978 Zurich Seminar on Digital Communications*, pages H1.1–H1.7, 1979.

[19] R. E. Crochiere, J. M. Tribolet, and L. R. Rabiner. An interpretation of the log likelihood ratio as a measure of waveform coder performance. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28(3):318–323, 1980.

[20] A. Cumani. On the canonical representation of homogeneous markov processes modelling failure-time distributions. *Microelectronics Reability*, (22):583–602, 1982.

[21] L. G. Cuthbert and J-C Sapanel. *ATM the broadband telecommunications solution*. The Institute of Electrical Engineers, London, 1993.

[22] M. de Prycker. *ATM Solutions for B-ISDN*. Prentice Hall, New Jersey, 1990.

[23] L. E. N. Delbrouck. A unified approximate evaluation of congestion functions for smooth and peaky traffic. *IEEE Trans. on Comm.*, 29:85–91, February 1981.

[24] B. Doshi. Deterministic rule based traffic descriptors for broadband ISDN: worst case behavior and connection admission control. In *GLOBECOM'93*, Orlando, 1993.

[25] Z. Dzinog and J. W. Roberts. Congestion probabilities in a circuit switched integrated services network. *Performance Evaluation*, 7:267–284, 1987.

[26] Dziong, Piero, and Koerner. An adaptive call routing strategy for circuit switched networks. In *Seventh Nordic Teletraffic Seminar*, 1987.

[27] A. E. Eckberg. Generalized peakedness of teletraffic processes. In *ITC-10*, Montreal, 1983. Session 4.4b.

[28] B. Eklundh, K. Söllberg, and B. Stavenow. Asychronous transfer modes - options and characteristics. In *ITC-88*, Torino, 1988.

[29] A. Faragó et al. Resource separation - an efficient tool for optimizing ATM network configuration. In *NETWORKS'94*, Budapest, September 1994.

[30] W. Feller. *An Introduction to Probability Theory and Its Applications*, volume I. Wiley, 1968. Third edition.

[31] Steve Fuhrmann and Jean-Yves Le Boudec. Burst and cell level models for ATM buffers. In *Proceedings of the 13th ITC*, Copenhagen, 1991.

[32] A. Girard. *Routing and Dimensioning in Circuit-Switched Networks.* Addison-Wesley, 1990.

[33] D. J. Goodman, C. Scagliola, R. E. Crochiere, L. R. Rabiner, and J. Goodman. Objective and subjective performance of tandem connections of waveform coders with an LPC vocoder. *The Bell System Technical Journal*, 58(3):601–629, 1978.

[34] G. Gordos and Gy. Takács. *Digital Speech Processing.* Műszaki Kiadó, Budapest, 1983. (In Hungarian).

[35] G. Gordos, P. Tatai, et al. On digitized speech quality assessment. Technical report, Institute of Communication Electronics, Technical University of Budapest, 1987–1990. (In Hungarian).

[36] R. M. Gray, A. Buzo, A. H. Gray, and Y. Matsuyama. Distortion measures for speech processing. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28(4):367–376, 1980.

[37] F. Guillemin, P. Boyer, A. Dupuis, and L. Romoef. Peak rate enforcement in ATM networks. In *IEEE INFOCOM'92*, Florence, 1992.

[38] F. Guillemin and J. W. Roberts. Jitter and bandwidth enforcement. In *IEEE GLOBECOM'91*, Phoenix, AZ, USA, December 1991. paper no. 9.

[39] S. Hansen. *Phase Type Distributions in Queueing Theory.* PhD thesis, IMSOR, Lungby, Denmark, 1983. Ph.D Thesis no. 41.

[40] L. Hanzó and L. Hinsenkamp. On the subjective and objective evaluation of speech coders. *Budavox Review*, (2):6–8, 1987.

[41] P. G. Harrison. *Performance Modelling of Communications Networks and Computer Architectures.* Addison Wesley, 1992.

[42] H. Hermansky. Perceptual linear predictive (PLP) analysis of speech. *Journal of the Acoustical Society of America*, 87(4):1738–1752, 1990.

[43] L. A. Hernández-Gómez et al. Real-time implementation and evaluation of variable rate CELP coders. In *IEEE Proc. Int. Conf. Acoust., Speech, Signal Processing*, pages 585–588, 1991.

[44] F. Hűbner. Dimensioning of a peak cell rate monitor algorithm using discrete-time analysis. In *ITC-14*, Antibes Juan-les-Pins, France, June 1994.

[45] J. M. Holtzman. The accuracy of the equivalent random method with renewal input. *Bell System Technical Journal*, 52:1673–1679, 1973.

[46] K. Itoh, N. Kitawaki, and K. Kakehi. Objective quality measures for speech waveform coding systems. *Review of the Electrical Communications Laboratories*, 32(2), 1984.

[47] ITU-T Recommendation E.700. *Series on Traffic Engineering.*

[48] ITU-T Recommendation I.350. *General Aspects of Quality of Service and Network Performance in Digital Networks, Including ISDNs*, 1993. Helsinki.

[49] ITU-T Recommendation I.356. *B-ISDN ATM Layer Cell Transfer Performance*, 1993. Helsinki.

[50] ITU-T Recommendation I.371. *Traffic Control and Congestion Control in B-ISDN*, 1992. Geneva, Switzerland.

[51] ITU-T Recommendations. *ITU-T Recommendations G.101–G.181.*

[52] ITU-T Recommendations. *ITU-T Recommendations of the P Series.*

[53] V. B. Iversen. The exact evaluation of multi service loss systems with access control. *Teleteknik*, 31, 1987. (English Edition).

[54] N. S. Jayant. High-quality coding of telephone speech and wideband audio. *IEEE Communications Magazine*, pages 10–20, January 1990.

[55] F. P. Kelly. *Reversibility and Stochastic Networks.* Wiley, 1979.

[56] F. P. Kelly. Blocking probabilities in large circuit switched networks. *Adv. Appl. Prob.*, 18, 1986.

[57] F. P. Kelly. Loss networks. *The Annals of Applied Probability*, 1(3), 1991.

[58] F.P. Kelly. Mathematical models of multi service networks. *Complex Stochastic Systems and Engineering*, September 1993.

[59] F.P. Kelly and P.B. Key. Dimensioning playout buffers for an ATM network. In *11th UK Teletraffic Symposium*, IEE Savoy Place, London, 1994.

[60] Z. Kenesi, A. Vidács, and S. Molnár. Fractals in telecommunications. *Magyar Távközlés*, July 1995. (In Hungarian).

[61] N. Kitawaki, M. Honda, and K. Itoh. Speech quality assessment methods for speech-coding systems. *IEEE Communications Magazine*, 22(10):26–33, October 1984.

[62] N. Kitawaki, K. Itoh, M. Honda, and K. Kakehi. Comparison of objective speech quality measures for voiceband codecs. In *Proceedings of IEEE ICASSP'82*, pages 1000–1003, 1982.

[63] N. Kitawaki and H. Nagabuchi. Quality assessment of speech coding and speech synthesis systems. *IEEE Communications Magazine*, 26(10):36–44, October 1988.

[64] L. Kleinrock. *Queueing Systems II: Computer Applications*. John Wiley, Chichester, 1976.

[65] D.D. Kourvatsos. Maximum entropy and the G/G/1/N queue. *Acta Informatica*, 23, 1986.

[66] H. Kröner. Statistical multiplexing of sporadic sources - exact and approximate performance analysis. In *ITC-13*, Copenhagen, 1991.

[67] M. J. Lighthill. *Fourier Analysis and Generalized Functions*. Cambridge University Press, 1958.

[68] W. Matragi, C. Bisdikian, and K. Sohraby. Jitter calculus in ATM networks: Single node case. In *INFOCOM'94*, June 1994.

[69] W. Matragi, K. Sohraby, and C. Bisdikian. Jitter calculus in ATM networks: Multiple node case. In *INFOCOM'94*, June 1994.

[70] B. J. McDermott, C. Scagliola, and D. J. Goodman. Perceptual and objective evaluation of speech processed by adaptive differential PCM. *Bell Syst. Tech. J.*, 57(5):1597–1618, June 1978.

[71] S. Molnár. An application of the Beneš approach to the queueing performance of a superposition of cell delay variated CBR sources. *EUA document*, October 15 1992. Lund, Sweden.

[72] S. Molnár. Congestion control and queueing models in ATM networks. *Technical Report*, February 17 1993. Ser. Electrical Engineering, Department of Telecommunications and Telematics, Technical University of Budapest, No. TUB-TR-93-EE02, Budapest, Hungary.

[73] S. Molnár. Network dimensioning based on refined blocking measures. *Internal Report*, May 1994. Department of Telecommunications and Telematics, Technical University of Budapest, Budapest, Hungary.

[74] S. Molnár and S. Blaabjerg. Cell delay variation in an ATM multiplex. *submitted to the Performance Evaluation.*

[75] S. Molnár and S. Blaabjerg. Correlations in ATM cell streams exposed to cell delay variation. *submitted to the Journal on Communications, Special Issue on ATM Networks.*

[76] S. Molnár and S. Blaabjerg. On two simple approximations for the G/G/∞ and G/G/c systems. *EUA document*, June 25 1993. Lund, Sweden.

[77] S. Molnár and S. Blaabjerg. Blocking probabilities in B-ISDN. *Internal Report*, Oct 1994. Department of Telecommunications and Telematics, Technical University of Budapest, Budapest, Hungary.

[78] S. Molnár and S. Blaabjerg. The effect of multiplexing CDV affected CBR cell streams. *COST 242 TD(94)014*, May 25–26 1994. Budapest, Hungary.

[79] S. Molnár and S. Blaabjerg. Methods for computing B-ISDN link blocking probabilities. In *Technical Conference on Computer Aided Methods and Technical Management in Electrical Engineering Education*, Budapest, Hungary, June 9–10 1994.

[80] S. Molnár and S. Blaabjerg. On some simple blocking approximations for the G/G/c queue. *COST 242 TD(94)001*, February 10–11 1994. Barcelona, Spain.

[81] S. Molnár and S. Blaabjerg. Generalized CDV models. *COST 242 TD(95)016*, January 18-19 1995. Cambridge, UK.

[82] S. Molnár, S. Blaabjerg, and H. Christiansen. On the superposition of a number of CDV affected cell streams. In *International Conference on Local and Metropolitan Communication Systems LAN&MAN*, Kyoto, Japan, December 7–9 1994.

[83] S. Molnár, A. Faragó, T. Henk, and S. Blaabjerg. Towards precision tools for ATM network design, dimensioning and management. *accepted to the Periodica Polytechnica.*

[84] S. Molnár, P. Pozsgai, and Á. Rétfalvi. Phase type queueing models. *Magyar Távközlés*, July 1995. (In Hungarian).

[85] S. Molnár and P. Tatai. Subjective and objective speech quality assessment methods. In *Proceedings of the Speech Research '92 Conference*, Budapest, Hungary, September 24–25 1992. (In Hungarian).

[86] S. Molnár and P. Tatai. Quality parameters of ATM networks. *Magyar Távközlés*, September 1994. (In Hungarian).

[87] S. Molnár, P. Tatai, and Z. Jánosy. Speech quality assessment for low-bit rate coding. *Journal on Communications*, 43, July–September 1992.

[88] C. Mossotto. Speech technology and telecommunications. *CSELT Technical Reports*, 20(1):5–13, 1992.

[89] M. F. Neuts. *Matrix Geometric Solutions in Stochastic Models.* John Hopkins University Press, 1981.

[90] M.F. Neuts. A versatile markovian point process. *J. Appl. Prob.*, 16, 1979.

[91] B. F. Nielsen. *Modelling of Multiple Access Systems with Phase-Type Distributions.* PhD thesis, IMSOR, Lungby, Denmark, 1988. Ph.D Thesis no. 49.

[92] I. Norros, J. W. Roberts, A. Simonian, and J. T. Wirtamo. The superposition of variable bit rate sources in an ATM multiplexer. *IEEE Journal on Selected Areas in Communications*, 9(3):187–197, April 1991.

[93] R. O. Onvural. *Asynchronous Transfer Mode Networks: Performance Issues.* Artech House, Boston-London, 1993.

[94] C. Palm. Intensitätsschwankungen im fernsprechverkehr. *Ericsson Technics*, 44, 1943.

[95] S. R. Quackenbush, T. P. Barnwell III, and M. A. Clements. *Objective Measures of Speech Quality.* Prentice Hall, 1988.

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

[97] V. Ramaswami and M. F. Neuts. Some explicit formulas and computational methods for the infinite-server queues with phase-type arrivals. *Journal of Applied Probability*, 17, 1980.

[98] J. S. Richters and C. A. Dvorak. A framework for defining the quality of communications services. *IEEE Communications Magazine*, pages 17–23, October 1988.

[99] J. Roberts, editor. *COST 242: Performance evaluation and design of multiservice networks.* Commission of the European Communities, October 1991.

[100] H. Saito. *Teletraffic Technologies in ATM Networks.* Artech House, 1994.

[101] R. Syski. *Introduction to Congestion Theory in Telephone System.* Oliver and Boyd, 1960.

[102] L. Takács. *Introduction to the Theory of Queues.* Oxford University Press, 1962.

[103] P. Tatai. Comments on objective quality measures in speech encoding. *Budavox Review*, (4):20–24, 1989.

[104] W. D. Voiers. Diagnostic acceptability measure for speech communication systems. In *IEEE Proc. Int. Conf. Acoust., Speech, Signal Processing*, pages 204–207, May 1977.

[105] S. Wang, A. Sekey, and A. Gersho. Auditory distortion measure for speech coding. In *IEEE Proc. Int. Conf. Acoust., Speech, Signal Processing*, pages 493–496, 1991.

[106] S. Wolf, C. A. Dvorak, R. F. Kubichek, C. R. South, R. A. Schaphorst, and S.D. Voran. How will we rate telecommunications system performance ? *IEEE Communications Magazine*, pages 23–29, October 1991.

Appendix A

The Derivation of the Transition Matrix

In this Appendix the derivation of the transition matrix computation of the exact Markovian model is given.

The elements of the transition matrix can be obtained by

$$q_{j,k} = Q(j, k-1) - Q(j, k) \tag{A.1}$$

where

$$\begin{aligned} Q(j,k) = P\{W_i > k \mid W_{i-1} = j\} = \\ \sum_{n=0}^{\infty} (P\{W_i > k \mid W_{i-1} = j, N(iT-T, iT) = n\} \\ P\{N(iT-T, iT) = n \mid W_{i-1} = j\}) \end{aligned} \tag{A.2}$$

where $N(t_1, t_2) = n$ denotes the event of n arrivals in $]t_1, t_2]$.

The second term of the sum in (A.2) can be obtained by the T-fold convolution of the batch size distribution, i. e.

$$P\{N(iT-T, iT) = n \mid W_{i-1} = j\} = P\{N(iT-T, iT) = n\} = b^{T*}(n) \tag{A.3}$$

The first factor of each term in the sum of (A.2) can be derived as follows:

If $j + 1 \geq T$ and $n \leq T + k - j - 1$ or $j + 1 < T$ and $k \geq n$ then

$$P\{W_i > k \mid W_{i-1} = j, N(iT-T, iT) = n\} = 0 \tag{A.4}$$

If $n > T + k - j - 1$ then

$$P\{W_i > k \mid W_{i-1} = j, N(iT-T, iT) = n\} = 1 \tag{A.5}$$

These formulas are seen by investigating the queue length behaviour, and it is easy to see that the queue length cannot be or must be exceed k, respectively.

For the remaining case when $j+1 < T$ and $k < n \leq T+k-j-1$ the Beneš formula, see section 5.3.2. in [99], can be applied:

$$\begin{aligned} &P\{W_i > k \mid W_{i-1} = j, N(iT-T, iT) = n\} = \\ &= \sum_{s=1}^{n-k} P\{N(iT-s, iT) = k+s \mid W_{i-1} = j, N(iT-T, iT) = n\} \times \\ &\times P\{W(iT-s) = 0 \mid W_{i-1} = j, N(iT-s, iT) = k+s, N(iT-T, iT) = n\} \end{aligned} \tag{A.6}$$

where $W(t)$ denotes the queue length at time t.

The first probability in (A.6) can be derived as follows:

$$\begin{aligned} &P\{N(iT-s, iT) = k+s \mid W_{i-1} = j, N(iT-T, iT) = n\} = \\ &= \frac{P\{N(iT-s, iT) = k+s, N(iT-T, iT) = n\}}{P\{N(iT-T, iT) = n\}} = \\ &= \frac{P\{N(iT-s, iT) = k+s, N(iT-T, iT-s) = n-k-s\}}{P\{N(iT-T, iT) = n\}} = \end{aligned}$$

In the numerator the number of arrivals in the two adjacent windows are independent of each other, therefore

$$\begin{aligned} &= \frac{P\{N(iT-s, iT) = k+s\} P\{N(iT-T, iT-s) = n-k-s\}}{P\{N(iT-T, iT) = n\}} = \\ &= \frac{b^{s*}(k+s) b^{(T-s)*}(n-k-s)}{b^{T*}(n)} \end{aligned} \tag{A.7}$$

The second probability in (A.6) can be obtained by using the equivalence with the Beneš analysis of the $nD/D/1$ queue, see Chapter 6.2.1 in [99] for details, so we have

$$P\{W(iT-s) = 0 \mid W_{i-1} = j, N(iT-s, iT) = k+s, N(iT-T, iT) = n\} = \frac{T-n+k}{T-s} \tag{A.8}$$

Putting (A.7) and (A.8) into (A.6) we get the first term of (A.2):

$$\begin{cases} 0 & j+1 \geq T \text{ and } n \leq T+k-j-1 \text{ or} \\ & \quad j+1 < T \text{ and } k \geq n \\ 1 & n > T+k-j-1 \\ \sum_{s=1}^{n-k} \frac{T-n+k}{T-s} \times & \\ \quad \times \frac{b^{s*}(k+s) b^{(T-s)*}(n-k-s)}{b^{T*}(n)} & j+1 < T \text{ and } k < n \leq T+k-j-1 \end{cases} \tag{A.9}$$

Finally, introducing $P_n(j,k)$ as the product of the two terms in (A.2) we have

$$Q(j,k) = \sum_{n \geq 0} P_n(j,k) \tag{A.10}$$

with

$$P_n(j,k) = \begin{cases} 0 & j+1 \geq T \text{ and } n \leq T+k-j-1 \text{ or} \\ & \quad j+1 < T \text{ and } k \geq n \\ \underset{n-k}{b^{T*}}(n) & n > T+k-j-1 \\ \sum_{s=1} b^{s*}(k+s) \times & \\ \quad \times b^{(T-s)*}(n-k-s)\frac{T-n+k}{T-s} & j+1 < T \text{ and } k < n \leq T+k-j-1 \end{cases} \tag{A.11}$$

Appendix B

The Distribution of Number of Arrivals in a Window

In this Appendix the derivation of the distribution of the number of arrivals in a window for the $nTri/D/1$ queue is given. Depending on the position of the window in relation to the frame flow (see Figure 8.6), different cases can be distinguished.

B.1 The Case When the Window Size is Smaller Then the Frame Size

B.1.1 The Subcase of $A < s$

This case is illustrated in Figure B.1.

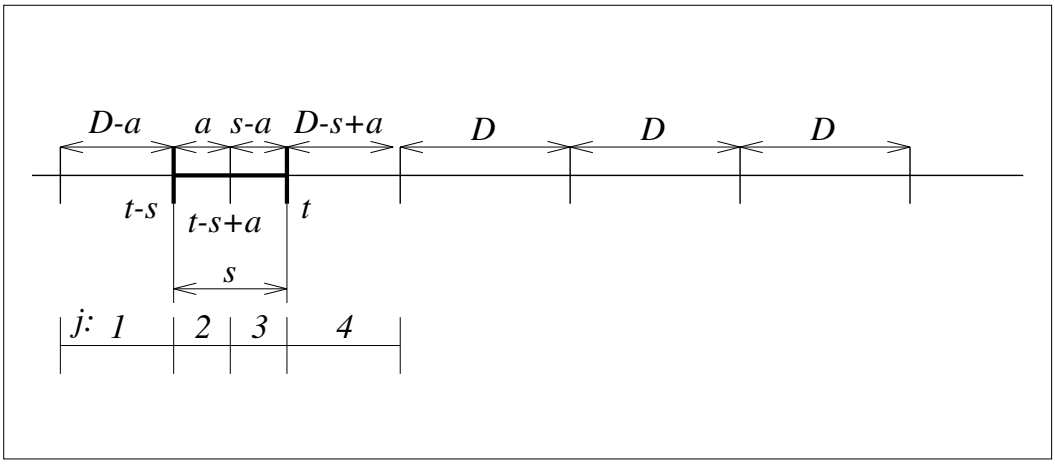

Figure B.1: The Case of $a < s \leq D$

The k arrivals are uniformly distributed in each frame, so we can easily conclude the probability of any one arrival in a specific interval from Figure B.1. These are

$$P_1 = \frac{D-a}{D}, \quad P_2 = 1 - P_1, \quad P_3 = \frac{s-a}{D}, \quad \text{and} \quad P_4 = 1 - P_3 \tag{B.1}$$

where $P_j = P\{any\ one\ arrival\ in\ interval\ j\ from\ a\ specific\ source\}$. In order to calculate the conditional probability of a specific number of arrivals in the window given $A = a$ we can collect all possible events which are different from each other in the sharing of arrivals in interval 2 and 3. Collecting these events we obtain

$$P_\alpha\{N_s = i | A = a\} = \begin{cases} \sum_{l=0}^{i} \binom{k}{l} P_1^{k-l} P_2^l \binom{k}{i-l} P_3^{i-l} P_4^{k-i+l} & \text{if } i \le k \\ \sum_{l=i-k}^{k} \binom{k}{l} P_1^{k-l} P_2^l \binom{k}{i-l} P_3^{i-l} P_4^{k-i+l} & \text{if } k < i \le 2k \\ 0 & \text{if } i > 2k \end{cases} \tag{B.2}$$

where l and $i-l$ denote the numbers of arrivals in interval 2 and 3, respectively. The notation α is used for denoting the case of $a < s \le D$.

B.1.2 The Subcase of $A \ge s$

This is the case when the window is within a frame. The probability of anyone arrival in the intervals denoted by 1 and 2 are obtained from Figure B.2:

$$P_1 = \frac{s}{D}, \quad \text{and} \quad P_2 = 1 - P_1 \tag{B.3}$$

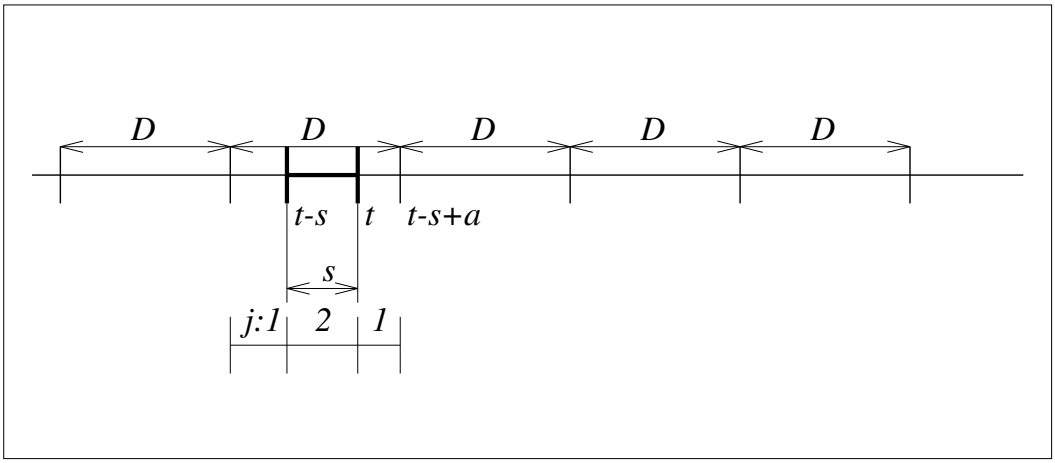

Figure B.2: The Case of $s \le a \le D$

The conditional probability:

$$P_\beta\{N_s = i | A = a\} = \begin{cases} \binom{k}{i} P_1^i P_2^{k-i} & \text{if } i \le k \\ 0 & \text{if } i > k \end{cases} \tag{B.4}$$

where β denotes the case of $s \le a \le D$. Now we have the conditional probabilities for both subcases of $s \le D$. From these conditional probabilities we can get the probability of a specific number of arrivals in the window for this case:

$$P\{N_s = i\} = \int_0^s P_\alpha\{N_s = i | A = a\} P\{A \in (a, a + da)\} + \tag{B.5}$$

$$\int_s^D P_\beta\{N_s = i | A = a\} P\{A \in (a, a + da)\} \tag{B.6}$$

where $P\{A \in (a, a + da)\} = \frac{da}{D}$, because A has uniformly distribution over the frame.

B.2 The Case When the Window Size is Larger Then the Frame Size

In the present case (see Figure B.3) it is known that the frame contains minimum $\lfloor \frac{s-a}{D} \rfloor k$ number of arrivals (e.g. in Figure B.3 it is $2k$), where $\lfloor x \rfloor$ denotes the integer part of x. The number of arrivals can be different only in interval 2 and 3.

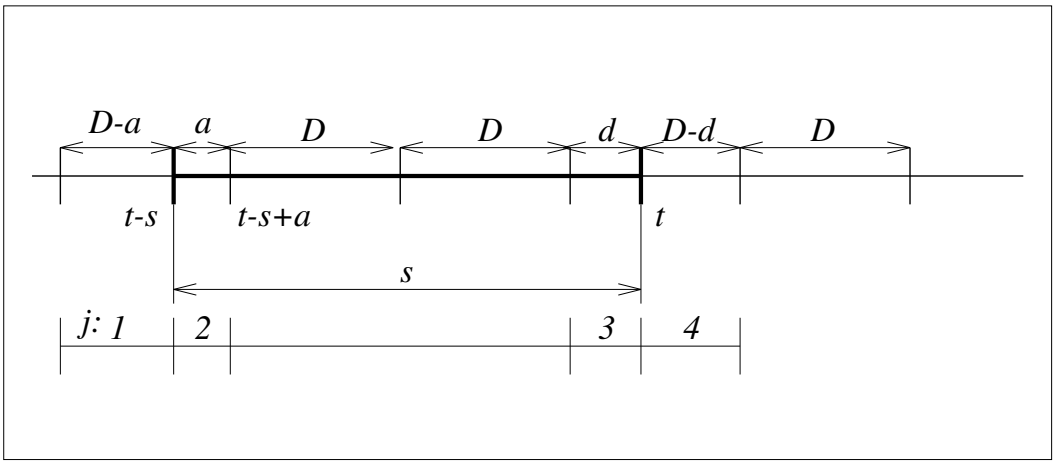

Figure B.3: The Case of $s > D$

From Figure B.3 the probabilities of any one arrival in the intervals is obtained:

$$P_1 = \frac{D - a}{D}, \quad P_2 = 1 - P_1, \quad P_3 = \frac{s - \left(a + \lfloor \frac{s-a}{D} \rfloor D\right)}{D}, \quad \text{and} \quad P_4 = 1 - P_3 \tag{B.7}$$

Similarly as above the conditional probabilities can be obtained:

$$P_\gamma\{N_s = \left\lfloor \frac{s-a}{D} \right\rfloor k + i | A = a\} =$$
$$\begin{cases} \sum_{l=0}^{i} \binom{k}{l} P_1^{k-l} P_2^l \binom{k}{i-l} P_3^{i-l} P_4^{k-i+l} & \text{if } i \le k \\ \sum_{l=i-k}^{k} \binom{k}{l} P_1^{k-l} P_2^l \binom{k}{i-l} P_3^{i-l} P_4^{k-i+l} & \text{if } k < i \le 2k \\ 0 & \text{if } i > 2k \end{cases} \tag{B.8}$$

From the conditional probability we obtain

$$P\{N_s = i\} = \int_0^D P_\gamma\{N_s = i | A = a\} P\{A \in (a, a+da)\} \tag{B.9}$$

where $P\{A \in (a, a+da)\} = \frac{da}{D}$, as above.

Appendix C

The Local Load Approximation

In this Appendix the derivation of the overflow probability of the $nTri/D/1$ queue using the local load approximation is given.

Assuming local stationarity around $t-s$, it can be written that

$$P\{Q_{t-s}=0|N_s=s+r\}\approx 1-\rho' \tag{C.1}$$

where ρ' is the local load at $t-s$ given $s+r$ arrivals in the window. The local load defined by

$$\rho'=\frac{E\{N(t-s-\Delta t,t-s)|N_s=s+r\}}{\Delta t} \tag{C.2}$$

Because of the probability of more then one arrival is $O(\Delta t^2)$ Eq. C.2 can be rewritten as

$$\rho'=\frac{P\{N(t-s-\Delta t,t-s)=1|N_s=s+r\}}{\Delta t} \tag{C.3}$$

Applying $P(B|A)=\frac{P(A|B)P(B)}{P(A)}$ for Eq. C.3 we obtain

$$\rho'=\frac{P\{N(t-s-\Delta t,t-s)=1\}}{\Delta t}\times\frac{P\{N_s=s+r|N(t-s-\Delta t,t-s)=1\}}{P\{N_s=s+r\}} \tag{C.4}$$

where for the first term we can conclude that it is the load. Insertion of Eq. C.1 and Eq. C.4 into Eq. 8.11 finally we get

$$P\{Q_t>r\}\cong\sum_{s=1}^{\infty}[P\{N_s=s+r\}-\rho P\{N_s=s+r|\text{one arrival at (t-s)}\}] \tag{C.5}$$

The main advantage of Eq. C.5 is that it contains only quantities related to the arrival process.

The derivation of $P\{N_s=s+r\}$ is shown in Appendix B. The calculation of $P\{N_s=s+r|\text{one arrival at (t-s)}\}$ can be performed similarly in the following way:

C.1 The Case When the Window Size is Smaller Then the Frame Size

C.1.1 The Subcase of $A < s$

This case is illustrated in Figure C.1.

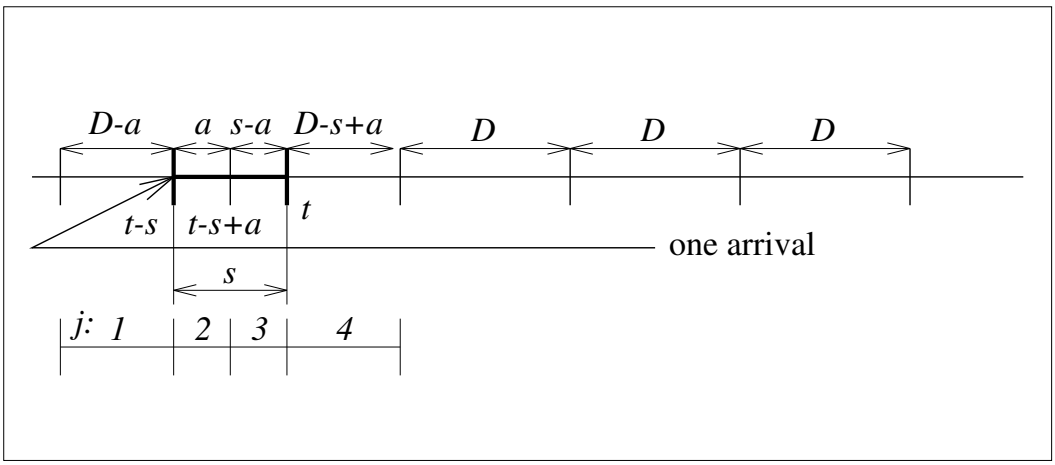

Figure C.1: The Case of $a < s \leq D$

The probability of any one arrival in a specific interval from Figure C.1:

$$P_1 = \frac{D-a}{D}, \quad P_2 = 1 - P_1, \quad P_3 = \frac{s-a}{D}, \quad \text{and} \quad P_4 = 1 - P_3 \tag{C.6}$$

The conditional probability:

$$P_\delta\{N_s = i | A = a\} = \begin{cases} \sum_{l=0}^{i} \binom{k-1}{l} P_1^{k-l-1} P_2^l \binom{k}{i-l} P_3^{i-l} P_4^{k-i+l} & \text{if } i \leq k-1 \\ \sum_{l=i-k}^{k-1} \binom{k-1}{l} P_1^{k-l-1} P_2^l \binom{k}{i-l} P_3^{i-l} P_4^{k-i+l} & \text{if } k-1 < i \leq 2k-1 \\ 0 & \text{if } i > 2k-1 \end{cases} \tag{C.7}$$

where l and $i-l$ denote the numbers of arrivals in interval 2 and 3, respectively. The notation δ is used for denoting the case of $a < s \leq D$.

C.1.2 The Subcase of $A \geq s$

The probability of any one arrival in the intervals denoted by 1 and 2 are obtained from Figure C.2:

$$P_1 = \frac{s}{D} \quad \text{and} \quad P_2 = 1 - P_1 \tag{C.8}$$

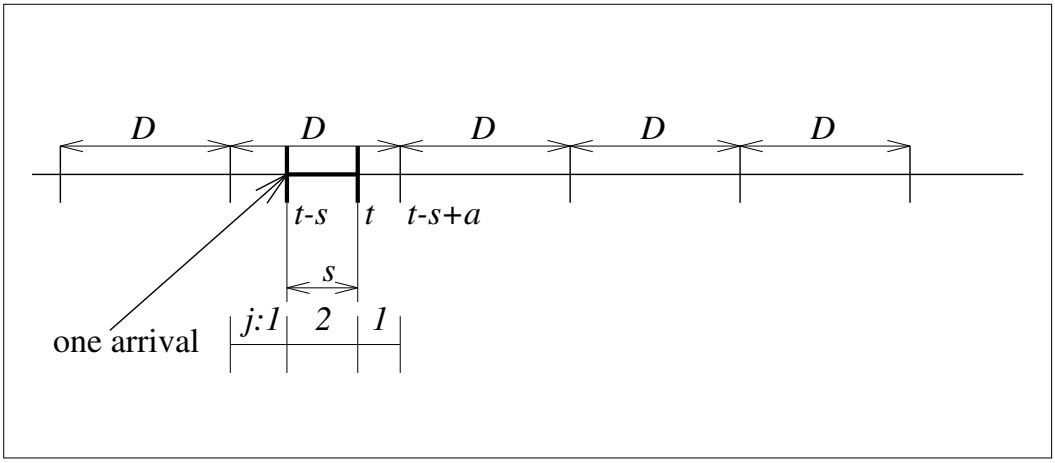

Figure C.2: The Case of $s \leq a \leq D$

The conditional probability:

$$P_\xi\{N_s = i | A = a\} = \begin{cases} \binom{k-1}{i} P_1^i P_2^{k-i-1} & \text{if } i \leq k-1 \\ 0 & \text{if } i > k-1 \end{cases} \tag{C.9}$$

where ξ denotes the present case. From these conditional probabilities the probability of a specific number of arrivals in the window can be obtained as:

$P\{N_s = s + r | \text{one arrival at (t-s)}\} =$

$$= \begin{cases} \int_0^s P_\delta\{N_s = i | A = a\} P\{A \in (a, a + da)\} + \\ \int_s^D P_\xi\{N_s = i | A = a\} P\{A \in (a, a + da)\} \end{cases} \tag{C.10}$$

where $P\{A \in (a, a + da)\} = \frac{da}{D}$.

C.2 The Case When the Window Size is Larger Then the Frame Size

The probabilities of one arrival in the intervals from Figure C.3:

$$P_1 = \frac{D - a}{D}, \quad P_2 = 1 - P_1, \quad P_3 = \frac{s - \left(a + \lfloor \frac{s-a}{D} \rfloor D\right)}{D}, \quad \text{and} \quad P_4 = 1 - P_3 \tag{C.11}$$

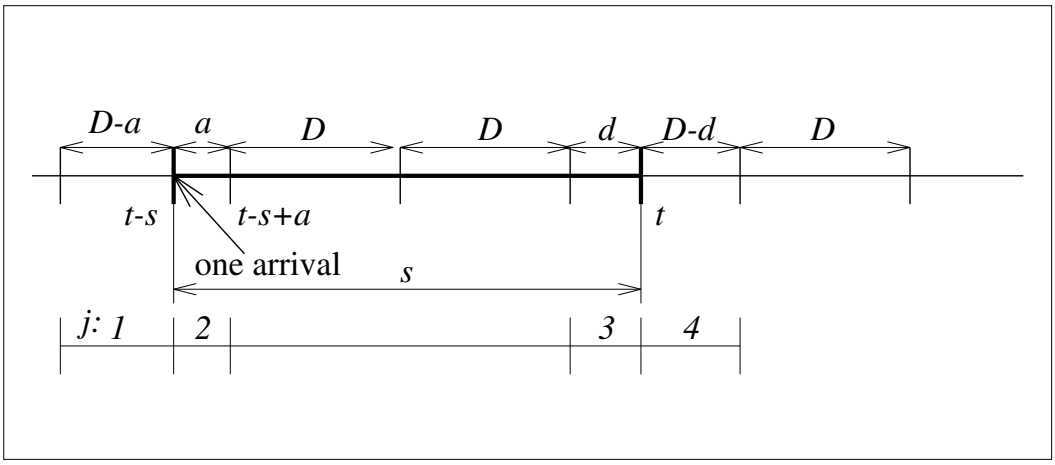

Figure C.3: The Case of $s > D$

The conditional probabilities from Figure C.3:

$$P_\eta\{N_s = \left\lfloor \frac{s-a}{D} \right\rfloor k + i | A = a\} =$$

$$\begin{cases} \sum_{l=0}^{i} \begin{pmatrix} k-1 \\ l \end{pmatrix} P_1^{k-l-1} P_2^l \begin{pmatrix} k \\ i-l \end{pmatrix} P_3^{i-l} P_4^{k-i+l} & \text{if } i \le k-1 \\ \sum_{l=i-k}^{k-1} \begin{pmatrix} k-1 \\ l \end{pmatrix} P_1^{k-l-1} P_2^l \begin{pmatrix} k \\ i-l \end{pmatrix} P_3^{i-l} P_4^{k-i+l} & \text{if } k-1 < i \le 2k-1 \\ 0 & \text{if } i > 2k-1 \end{cases} \quad \text{(C.12)}$$

From the conditional probability we obtain

$$P\{N_s = s + r | \text{one arrival at (t-s)}\} =$$

$$= \int_0^D P_\eta\{N_s = i | A = a\} P\{A \in (a, a+da)\} \quad \text{(C.13)}$$

where $P\{A \in (a, a+da)\} = \frac{da}{D}$.