BUDAPEST UNIVERSITY OF TECHNOLOGY AND ECONOMICS
DEPT. OF TELECOMMUNICATIONS AND MEDIA INFORMATICS

PARSIMONIOUS ESTIMATES OF SOME QUALITY OF SERVICE MEASURES IN TELECOMMUNICATION NETWORKS

Zalán Heszberger

Ph.D Dissertation

Supervised by

Dr. József Bíró and Dr. Tamás Henk
High Speed Networks Laboratory
Dept. of Telecommunications and Media Informatics
Budapest University of Technology and Economics

Budapest, Hungary
2006

Zalán Heszberger

High Speed Networks Laboratory

Dept. of Telecommunications and Media Informatics

Budapest University of Technology and Economics[1]

[1]The reviews and the minutes of the Ph.D. Defense are available from the Dean's Office.

BUDAPESTI MŰSZAKI ÉS GAZDASÁGTUDOMÁNYI EGYETEM
TÁVKÖZLÉSI ÉS MÉDIAINFORMATIKAI TANSZÉK

KEVÉS PARAMÉTERES MINŐSÉGBIZTOSíTÁSI TECHNIKÁK KOMMUNIKÁCIÓS HÁLÓZATOKBAN

Heszberger Zalán

Doktori disszertáció

Tudományos vezető

Dr. Bíró József és Dr. Henk Tamás
Nagysebességű Hálózatok Laboratóriuma
Távközlési és Médiainformatikai Tanszék
Budapesti Műszaki és Gazdaságtudományi Egyetem

Budapest
2006

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

As the Internet becomes popular and widely used in the whole world, demands for guaranteed end-to-end service quality that supports the fast-growing number of real-time applications is increasing. The most popular such applications include VoIP (Voice over Internet Protocol) and Internet radio/television as well as videoconferencing applications. New developments are also urged by the fact that – besides the (wired) public telephone services – new generations of mobile networks are also tend to increasingly integrate into the global networking infrastructure operated at networking level mostly by the Internet protocol suite.

Although traditional best-effort packet delivery does not prevent people from using the network for applications with advanced real time requirements, from financial side the long-term management of such systems is not advisable. From one side it provides no intention for network operators to invest for further resources, from the other it gives no intention for customers to use resources economically. Besides structural redevelopment, the latest optical technologies seem to provide the possibility of resource over-provisioning, however, most of the time this is not a satisfactory solution. Installing e.g. a faster transmission link at one point on the path of a flow generally does not lead to a better end-to-end quality by itself. Also there are naturally scarce resources that cannot be extended arbitrarily like free air frequency channels in the case of wireless transmission.

To provide guarantied transmission quality in packet-switched multiservice networks appropriate planning, implementation and management techniques have to be developed. Also appropriate pricing mechanisms are just as important as advanced network monitoring and resource assessment techniques.

The focus of the dissertation is to provide contribution to the resource assessment part of this wide research area in the form of traffic characterization techniques coupled with service quality guaranties. The main application of these techniques are quite widespread including network/resource/traffic control and management. As the most important common features of the results, they are based on simple traffic measurements and the ultimate target is to offer simple ways of estimating transmission link saturation and packet loss probabilities with corresponding link capacity assessments under given traffic conditions.

In the following introductory chapter, first a short overview is given for QoS (Quality of Service) aware network architectures and algorithms that will serve mainly as a reference for the discussion of later chapters. Next the important properties of the presented traffic characterization techniques are introduced. High-level modeling assumptions and real application requirements are also discussed in detail. Section 1.4 aims to present the specific scope of the dissertation and brings together the most important recent research results in the area. The final section of the introductory chapter provides a detailed guide for later chapters in the book defining the main objectives in the form of an overview of the particular contributing results and methodology of the research.

1.1 Multiservice Network Achitectures and Algorithms

Considerable efforts have been made in the past decades to enhance the current best-effort capable global Internet network architecture to provide means of service differentiation, but asides from some isolated implementations and simple testbeds the problem remained unsolved. Although – as a first attempt – the original description of IP (Internet Protocol) in RFC (Request for Comments) 791, at the birth of the standardized Internet protocol suit back in the year 1981, defined four "Type of Service" bits and three "Precedence" bits provided in each IP packet, service differentiation turned out to be a considerably more complex task.

1.1.1 Important Properties of QoS Achitectures

The first real and most influential attempt to define QoS in telecommunication was carried out by an industry consortium called the ATM Forum. The main purpose of the non-profit organization was to create implementation agreements around the ATM (Asynchronous Transfer Mode) as the enabling technology of the B-ISDN (Broadband Integrated Services Digital Network). Although the technology did not live up to every expectations, it still remains an important core network technology. But most important of all they have created the first consistent concept of supporting multiple levels of QoS for packet traffic based telecommunication. Two of the defined QoS classes CBR (Constant Bit Rate) and rt-VBR (Real-Time Variable Bit Rate) are designed for real-time traffic (ie. where there is an inherent reliance on time synchronisation between end-nodes) as well as non-real time types where accurate timing between source and destination has little importance (nrt-VBR – Non-Real-Time Variable Bit Rate and ABR – Available Bit Rate) or not important at all (UBR - Unspecified Bit Rate). The class representing the most strict requirements is CBR where services that are to be provided should be similar to that of circuit switched connections with dedicated channel capacity. The ATM Forum also defines the concept of a traffic contract, a set of information sent by a traffic source to the network containing the type of traffic to be transported, and the performance requirements of it. The three main parts of the traffic description are:

- type of service description
- traffic parameters of the provided data flow
- the required QoS parameters for the traffic.

The traffic parameters refer to the traffic characteristics of the source like *peak* and *mean rate* or the maximum burst size. The QoS parameters on the other hand represent the requested quality of the connection (like *transfer delay* or *loss ratio*) and quantify end-to-end network performance.

As the Internet protocol suite become widely accepted among network operators the IETF (Internet Engineering Task Force) established an IP layer based QoS architecture: the IntServ or Integrated Services. IntServ is a fine-grained QoS system where every router on a path of a traffic flow keeps resource

reservation information on a per-flow basis. The architecture even defines its own resource reservation protocol the RSVP (Resource ReSerVation Protocol) in RFC 2205. In IntServ terminology Flow Specs consists of the TSPEC as the traffic specifications and RSPEC for the service request specifications. The most typical traffic description parameters in the architecture belongs to the token bucket algorithm, where tokens are defined to represent "licenses" for sending units of data (ie. IP packets). As tokens are generated the corresponding number of packets are allowed to send. If a token is generated but there is no data to send it is put into a "token bucket" to store. If the token bucket fills up, no further tokens can be stored. In this model, the size of the token bucket represents the allowed number of packets that can be sent back to back (ie. at maximum rate) also called the allowed maximum size of burst on a transmission link, while the generation rate of the tokens refers to the allowed long term mean rate of the traffic. The typical service types in IntServ are:

- *Best effort* service for flows without any special transmission requirements
- *Controlled load* for the flows with "mild" QoS requirements with respect to eg. delay or loss rate
- *Guaranteed* service provides absolute and deterministic guaranties on delay and no losses as long as there is no excess traffic as above previously specified.

As described above IntServ is designed to keep per flow guaranties, which obviously cannot scale well. This was the ideology behind the modified concept of aggregated IntServ, where at different levels of the network, the micro flows are to be aggregated and reservation is made accordingly.

Another IP-layer QoS concept called DiffServ or Differentiated Services, learning from the failures of previous architectures, attempts to find a lightweight solution lying between unreliable transmission and fully specified QoS provision. The main idea is to classify flows into a small number of QoS classes and defines priorities among classes only. In the pure version of this concept only the routers on the edge of the network deal with traffic policing (ie. dealing with non-conform or excess traffic), there is not any center entity that manage their process, later extensions however define the Bandwidth Broker entity to

help cooperative and cross-domain operation to provide end-to-end QoS provision. In the DiffServ concept there is no need of keeping (ever-changing) per-flow information in the routers, only a single negotiation is needed between parties (eg. customer and ISP or between ISP's) at the time of service subscription. These contracts or SLA's (Service Level Agreements) contains the requested QoS classes and the amount of traffic belonging to the different classes. Besides the policing work of the edge routers and simple priority based forwarding function of core routers, the bandwidth broker entities can have the task to maintain enough aggregated channel capacity (according to current organizational policies) in the network to satisfy the needs stated in all the SLA's.

Among the most latest researches and developments in the field, significant advancements have been made by the Internet2 non-profit consortium. The mission of the organization is to develop and deploy advanced network applications and technologies. Many of their achievement are organized around the development of IPv6 (IP version 6), which is planed to be the follower of the still widespread IPv4 at the Internet network layer. Although the main urge for migration is the lack of adequate addressing field in IPv4, IPv6 have been armoured by advanced capabilities like QoS-awarness and extended IP multicast and mobility mechanisms.

According to the Internet2 – QoS Working Group investigation, the installation of QoS in the public Internet meets several difficulties in general. The most important reason of which is that different operators with cross-connected networks – which the Internet is made of – depends upon each other, i.e. if some of them does not make the necessary improvements, end-to-end quality will not be able to be provided. Furthermore, when some services are offered for money, it is the provider's best interest to prevent users from using the free alternative by degrading it, which is not considered to be a fair behavior. Instead the group suggests a new architecture introducing *QBone Scavenger Service (QBSS)*, based on the philosophy of "lower than best-effort" service brought up recently. The idea is that traditional best effort traffic should be considered as the premium service which users may voluntarily give up by indicating their traffic as low priority. Similar ideas include *Alternative Best-Effort (ABE)* and *Best Effort Differentiated Services (BEDS)*. The different handling of the traffic flows can be realized by e.g. placing the packets in a separate queue at routers having lower priority in some given sense.

1.1.2 Quality Management Tasks and Algorithms

The first step toward thinking in quality management is correct *network planning* and *dimensioning*. Throughout the process global network user demand should be carefully considered, taking into account a broad class of popular traffic types. After choosing transmission technology, equipment brands and available protocol sets, the most important parameters to plan in this state are the physical capacity of connections between nodes and the size of the total buffer at the nodes. To make these choices one should estimate potential user numbers and the typical bandwidth requirement of the popular traffic types used. Besides taking into account peak user times, the effect of *statistical multiplexing* should also be carefully considered

The main focus of existing networks when providing service quality is to construct suitable *network control and management* functions that can fit into the planed business model. Global network policy is realized in longer terms through negotiations like peering, transit and customer contracts and in shorter terms typically through *traffic management* rules and corresponding algorithms applied on different time scales. The largest time scale in this context (i.e. that traffic management deals with) is the call-level one. At every newcomer traffic flow several decisions have to be made, the most important among them is about admission achieved by *call/connection admission control*. To provide quality to a connection through the calculated path to the target, adequate amount of resources like bandwidth, buffers at switching nodes/routers, should be reserved throughout connection lifetime. The *resource reservation* process can be realized according to the proprietary polices by carefully developed resource reservation protocols.

Besides the above mentioned traffic control functions, several management relates task should also be carried out. Many of them building around the so called *AAA (Authentication, Authorization, and Accounting)*. Suitable pricing and tariffing algorithms at different time scales can be crucial to a network to operate successfully. *Resource management* related tasks are also closely connected to this area.

However, there are functions that are extensively used in almost all of the above mentioned areas. Such tasks are the *network/traffic monitoring*. Monitoring functions are essential at traffic control functions as many of the control decisions should be made according to current state of network, and so

should be performed at wide range of timescales. Management functions like resource management or pricing should also be done based on this process. Network monitoring functions are typically dealing with the state of network devices, resource failures, sending notifications or alarms to network administrators, and also monitors network uplinks using *network traffic measurement procedures*. Such measurements, alternatively, can also be the basis of traffic control decisions like connection admission, or even some type of flow control functions. An important property of monitoring/measurement procedures is about its activeness. Active type monitoring intervene into network operation e.g. by testing resources (testing mail/file/web servers or network links) by sending test requests, or sending test traffic on links to discover link failures or congestion, while passive monitoring only observe operation without involving into it.

One of the major tools for developing efficient traffic management algorithms is *traffic engineering*. The central concept in *traffic engineering* is to use statistical techniques to discover important properties of current and future network states or traffic conditions taking into account user requests. The main results in this area are based on the fact that aggregate properties of long run systems become increasingly more characterizable by statistical descriptors than individual parts e.g. a single traffic flow. Classical results of queuing theory (originally developed for conventional telephone networks) are of great help also in engineering the packet-switched Internet.

1.2 Traffic Modeling and Characterisation

Traffic engineering techniques in traditional telephone systems are mostly based on the formulae developed by A. K. Erlang. The central assumption there is that traffic flow requests can be modeled by Poisson processes due to that in large systems there can be many independent users requesting services with small probabilities. It was proved to be a realistic assumption to estimate the service time distribution with exponentially distributed random variable. These basic assumptions however turned out to be hardly useable in packet switched multi-service networks. In store and forward packet switched networks queuing theory—which in circuit switched networks is used to study queues of users at the edges of the network—for the most part, deals with

buffers at the end-points of communication links. In this case the communicating parties at the two ends of the links are computers, inside the network, not humans, the behavior of which are often highly different. To efficiently characterize, engineer, and manage the Internet, it is of utmost importance to find new ways of modeling aggregate packet flow traffic stemming from a large number of service types.

In traffic engineering, the most important traffic characteristics are statistical parameters like mean rate, peak rate or "burstiness" (ie. carefully chosen function of higher order moments). At the very first step, such characterisation provides means of developing traffic models, like suitable stochastic processes that can describe traffic flows with adequate accuracy. At the second step, the established models are to be refined through the free parameters. If the characterisation and modeling steps are correctly accomplished network performance analysis is to be followed. Such typical performance measures are derived from the queuing behavior in buffers at network nodes or when buffering is negligible, traffic flow rate distribution is considered. Finally, taking also into account of various business related policies, it is ready to provide QoS guaranties to network users. *This dissertation concentrates from the above mentioned steps to the characterisation part, although the investigated measures also contains reference to the performance of the network. The results presented considers traffic flows taking (rate) values from a continuous set (or more specifically positive real values), which assumption is referred to as the fluid flow modeling approach.*

In the last decade, Internet related researches proved that packet flows can be modeled by long-range dependent processes more successfully as opposed to the traditional Markovian modeling approach suggested by the Erlang formulae. This behaviour is presumed partly to be the cause of the frequent appearance of network state properties (like size of transferred files/ bursts or packet-inter-arrival times) having heavy-tailed distribution. Heavy-(or long-)tailed distributions have the characteristic feature that the probability of rare events (towards the tail of a distribution) sums up to be a non-negligible part of the whole. *In the dissertation, the results are built on the bufferless modeling assumption, that is, buffers in the path of a data flow effecting network performance (with respect to that data flow) is considered to be negligible. This assumption can be a good estimation for stream type traffic flows or under certain heavy traffic situations. Under this assumption the properties of the*

autocovariance functions (e.g. short or long range dependence) of traffic flow processes have no influence on the presented results [29]. The distribution of the random variables referring to the rate of the traffic flows at any given time is considered to be arbitrary.

A highly important issue in recent traffic engineering research is based on the fact that in modern packet-switched networks high quality service provisioning necessitates the analysis of distributions of the sum of a large number of random variables (e.g. rate of traffic flows) far from its mean or standard deviation values. This area (developed originally in insurance and risk analysis [11]) is the main scope of large deviation theory. *In the dissertation, the aggregate properties of independent traffic flows are investigated, using large deviation type results. The presented contributions, due to the bufferless fluid flow multiplexing assumption, uses the instantaneous transmission rates of traffic sources as random variables, and provides theorems under this basis.*

1.3 Traffic Characterisation in Telecommunication Networks

A popular classification in packet-switched communication networks distinguish traffic types into two main groups. Stream type flows have the property of high sensitivity of transmission delays or jitter. It is also typical that in such cases data corruption or packet losses have only a second priority. Elastic type flows, belonging to the second group, on the other hand, are more robust against timing properties but claim more errorfree communication lines. As elastic flows, if necessary, can easily be transmitted with minimum resource availability, the engineering problems most often should consider the amount of stream traffic types in the first place.

Real-time (or quasi-real-time) flows are typical examples of stream types. To fulfill timing constraints to such flows buffers must be minimized (or at least limited) on the whole end-to-end transmission path. When a buffer becomes filled up at the end of a link, due to heavy traffic, the link is said to be saturated and subsequently loss occurs. To characterize the performance of the network in such cases the most direct descriptors used are

- the saturation probability, that is the proportion of time that the link is saturated,

- the loss probability, that is the proportion of packets that are lost.

The above characterisation has a simplified form when the system is modeled as bufferless. In this case buffers are considered negligible, and so saturation occurs when transmission rate overcome the bandwidth of the link. This concept is also called the rate envelope multiplexing. When characterising with saturation probability, the performance of the link can be derived easily, which is a useful point for the owner of the link, like the network operator, but tells much less about the experienced service degradation. It basically shows how the link itself performs on a time-line. Loss probability, on the other hand, concentrates on the effectiveness of the link as a data transmission equipment. This type of descriptor can be more informative for the owner of an individual flow, that is the network user. The former type of characterisation is also termed as *resource based*, while the latter is called *stream based*.

Traffic or network performance characterisation in real applications has a series of conditions to perform well. The most important ones are:

- it should not require extreme time or computing power to be determined
- it should be easy enough to be determined in real environments (by making simple measurements possibly aided with prescribed parameters)
- it should be informative, or should be easy to convert to be informative.

Regarding the simplicity part, it is essential for easy applicability that the necessary parameters measured (or queried *a priori*) be easily obtainable, and can be used in a simple way for further processing. The minimum amount of information for such purpose can be the peak rate of the aggregate traffic or of each flow individually on the link under study. Using such information QoS guaranties can be provided, however, effectiveness is hardly fulfilled. To further specify traffic situation, individual mean rates can be obtained, which may provide in most cases enough information to reach good suboptimal decisions regarding the management of link traffic. In a real environment, on the other hand, often even the determination of individual mean rates is difficult, or many times even unaccomplishable. In these cases, aggregate mean rate values provide a good compromise. The measurement or *a priori* acquisition of further description on the traffic like e.g. the variation of the rate distribution or higher moments of it, most often is practically hardly attainable.

With respect to being informative, the descriptor should not only be easily interpretable to concrete network resource characterization metrics like bandwidth, buffer size or money, but should also be easily verifiable for a network operator as well as the service subscriber. Being informative is often conflicting with computational effectiveness. In more crowded parts of a network node, there could be thousands of service claims per seconds that must be decided upon acceptance, rejection or other policy to handle user packets. It is so of utmost importance in these situations that characterisation algorithms work fast enough to keep time with incoming service demands. To fulfill this requirement it is often necessary to look after suboptimal solutions instead of finding the most economical decision, which can be carried out e.g. by suboptimal formulation and/or fast numerical computation.

1.4 Scope of the Dissertation & Related Work

QoS guaranteed packet networks is planned to be capable of guaranteeing predefined service quality according to diverse application needs, hence, based on appropriate modeling and analysis framework, the development of efficient traffic characterization and resource utilization techniques is of the utmost importance. The appropriate and fast quantification of resource usage is a core building element of numerous network/traffic management functions like monitoring, call/connection admission control algorithms, or even (resource based) pricing. During connection admission control a sub-process should quantify the resource required by the ongoing traffic flows together with the newcomers in order to decide whether the QoS requirement is still fulfilled. The decision is based on a rule which maps the QoS requirement onto resource requirement. The resource quantification sub-process needs traffic parameters which are partly measurable and partly known *a priori*. The signaling messages of resource reservation protocols like RSVP have the task to propagate these parameters to the place of admission decision if necessary.

Among these complex and relating tasks the dissertation focuses on direct estimates of bandwidth requirements based on parsimonious traffic characterization under the widely accepted and well-understood bufferless fluid flow multiplexing (*bffm*) modeling framework adopted for the traffic aggregation. Parsimonious in this context means that the estimators only needs the

average aggregate traffic rate and the peak rates of flows as inputs for computation. Using *bffm* to dimension transmission link capacity, characterize network performance or design decision rules for connection admission control algorithms typically is realized through estimating the overflow probability or the expected workload loss ratio under given conditions or alternatively the bandwidth requirements related to these performance meters.

Throughout the dissertation the emission rates of the individual traffic sources are modeled by stationary rate processes and the sources (with one exception) are always assumed to be fully independent. As stated above, stationary transmission is assumed with no buffers considered in its path, hence time dependence of traffic is eliminated, only their rate distribution is considered.

Some of the given estimations in the dissertation are proved to be optimal (i.e. best achievable) under given circumstances, some of them are suboptimal versions of the previous ones yielding, however, more computationally effective algorithms, shown to be necessary in these specific cases to attain practicability.

1.4.1 Background

Hoeffding in his famous paper [16] establishes conservative bounds on the tails of sums of (complex-valued) random variables. Many of his achievements are based on Chernoff's results [8] that assumes only Bernoulli distributions. The variables in Hoeffding's works is generalized to be arbitrary and uses the aggregate mean and variances as input data. Some of the bounds are proved to be optimal, generally the ones concerning independent and identically distributed variables. Most of the results remained unimproved until Talagrand's work [33], where the results are yielded by combining the Hoeffding formulae with large deviation results [36]. The so called "missing factor" refers to a factor depending also on the variation of the random variables. Later on remaining on the same track, quite a few new results are achieved (see e.g. [28], [15]), mostly by mathematicians examining tail distributions under different conditions. Researchers of communication networks started explicitly studying the field in the 1990's with especially the popularization and standardization of the ATM technology.

1.4.2 Related Work

The most important features of the practical research works in the field of communication networks are the extremely strict assumptions on the available information on network states or traffic flows traversing the network. Many works [32], [4] even discuss the effect of non-stationarity of the ever growing traffic. When studying traffic characterization, stochastic modeling assumptions rarely consider more than the mean and the variance of the applied distributions of certain traffic parameters. The majority of the contributions only assumes mean and peak rates of traffic flows to be known. Several papers e.g. on measurement-based traffic engineering [29] even go further and assume only aggregate mean arrival rates. Under these circumstances, the application of Hoeffding's results arise naturally. F.P. Kelly in many of his works [23] [24] use these results for charging issues in communication networks. Among many others S. Floyd, Jamin et al, Brichet et al., Gibbens et al. apply the Hoeffding formulae to establish Measurement-based Admission Control algorithms [13] [21] [6] [7] [14] mainly for bufferless multiplexing and using simple traffic characterization. In these works the admission control algorithm relies on a quality of service characterization parameter referred to as the saturation probability, although some considers loss performance as well [26]. Many research papers assuming simple traffic characterization concern the performance of buffered systems [10] [9] [25]. These works for the most part analyze some asymptotic behavior (like many sources or large buffer asymptotics) of buffered communication systems fed by stationary traffic sources. A common feature of the contributions applying simple characterization in traffic engineering is the adoption of the equivalent capacity concept. Turanyi et al [34] provides a family of such type of admission control algorithms for the saturation probability. In his well-received paper [23] F. P. Kelly gives a thorough analysis of a special form of equivalent capacity coined as the effective bandwidth. A related and widely studied research area is the network calculus, where sources are characterized by arrival curves. Between the arrival and departure functions, the constraint curves are used to provide the connection. Network calculus, in its standard form, is a deterministic approach, leaving the effect of statistical multiplexing out of consideration. Recently developed stochastic extensions of the theory are still manipulating bounding curves of traffic processes, concentrating typically to the evaluation of buffered systems.

The ultimate goal of my dissertation is to develop resource assessment techniques using simple traffic characterisation. The basic idea to the results is the Chernoff-method [8]. The input data to all of the formulae presented are the aggregate mean arrival rate and individual peak rates along with the number of flows. The major contribution of the dissertation is to provide easy to use closed form formulae to compute saturation probability, workload loss ratio and equivalent capacity values in traffic situations where the flows (or to be more exact their peak rates) are fairly inhomogenous, homogenous approximations are not reasonable to use. Note that this latter property is important as for the homogenous case, as mentioned above, there exist some (large deviation type) contributions with fairly good properties. The results can be directly applied to estimate bandwidth requirement of transmission links, where the effect of buffering is negligible. Also the development of the probability generating function estimates provided in Chapter 2 are novel contributions having significantly more general consequences than exploited later on in the dissertation. The presented formulae are shown to be (sometimes considerably) better (some, under given circumstances even optimal) than any other previously published ones for the same purpose.

1.5 Research Objectives and Methodology

The objective of the dissertation is threefold. In the first part new parsimonious moment-generating function estimation techniques are introduced. Such investigations turned out to be helpful in the development of QoS descriptors in the case of many loosely dependent/independent traffic sources. Conditions are shown when the presented formulae give optimal results (i.e. no better can be given). The results in this part are quite general and have much deeper consequences than the later part of the dissertation exploits them.

The second part of the dissertation deals with conservative upper-bounds – based on the results of the approximations in the first part – for saturation probability and workload loss ratio. These formulae have significant role in the provision of quality in future QoS networks. A special feature of the algorithms shown is that the exploited amount of information for the computation is rather tight – they are also called parsimonious –, only first order statistics of the state of the network are used. The motivation behind the developed suboptimal formulae in this section is the need for much faster (or real time

computation) of the otherwise quite complicated algorithms.

The purpose of the theorems in the final part of the chapter is to shed considerable light on the relation of probability and corresponding capacity estimation techniques. A much shorter computation time – again – is among the reasons in the first line to develop formulae that give equivalent capacity type results explicitly instead of saturation/packet-loss probability. Typical traffic management algorithms are aware of the link-capacity values in advance and also have the information about the required degree of quality to be guarantied on the managed links (e.g. packet loss probability). When a newcomer flow provides its peak transmission rate, as a first guess it can be summed up with the computed equivalent capacity value and the – e.g. admission – rule is that the sum should not exceed the total capacity of the link. In such a case the equivalent capacity can be computed off-line, and in the critical moment only a summation need to be carried out. Another important reason that makes these results relevant is that related research works in this field prefer the capacity estimation formulae. The methods presented also makes the analytical comparison of previous and new contributions easier. In a few cases of the derived formulae further improvements can be achieved by numerically performed algorithms, of which purpose recursive fixed-point equations are proposed.

The results in my dissertation were obtained mostly through analytical methods. Since the analytical derivation of the formulae contains different simplifying assumptions and approximations, the effect of which analytically are hardly tractable, extensive numerical and simulative investigations were carried out to validate the usefulness of the formulae.

Chapter 2

Parsimonious Approximations for Moment-generating Functions of Sums of Non-negative Valued Random Variables

Parsimonious estimation (ie. estimation that requires few *a priori* parameters) techniques to identify different statistical properties of random events typically rely on the knowledge of the first few moments of the distribution, many times only the mean value is assumed to be known. Such approximations are turned out to be useful when sums of a large number of random variables are considered. Further restrictions may stem from not knowing the individual mean values of a group of monitored events, only their aggregate mean can be observed.

After the introduction of the basic terms and definitions used later in this chapter, I will present two new upper bounds on the moment generating function of the sum of independent random variables. I will put out some important properties of moment generating functions that claimed to be stand also for their reasonable approximations, and prove that they are fulfilled by the presented new conservative bounds. In the last section optimality of the derived bounds is investigated, and indeed shown to be optimal (that is no better one can be given) under certain conditions.

2.1 Terms and definitions

Let $X_1(t), ...X_n(t)$ be stationary stochastic processes that will represent in later chapters the instantaneous information arrival rates of traffic sources as a function of time. Let $X(t) = \sum_{k=1}^{n} X_k(t)$ and $M(t) = E[X(t)]$, where $E[.]$ denotes the expectation value operator. The tail probability of random variable $X(t)$: $P(X(t) > C)$ with C denoting a fix threshold. Because we assume, that the processes are stationary, we can omit the time dependency, so that the tail distribution of the aggregation can simply be $P(X \geq C)$. Let us assume that the random variables X_k are non-negative valued, bounded by p_k as their respective maximum values. Similarly, the individual mean rates are denoted by m_k. Let $G_X(s) \stackrel{def}{=} E[\exp(sX)]$ denote the probability generating function (PGF) of X.

2.2 Upper approximation techniques of PGF for sum of i.i.d. random variables

In this section, we provide conservative bounds for the PGF of sum of n number of random variables, provided, besides n, only the following pieces of information are available: the maximum values of the random variables (p_i) and the aggregate mean value ($M \stackrel{def}{=} E[X]$).

2.2.1 Approximations Based on Hoeffding's results

The following lemma due to Hoeffding (1963), is on the PGF approximation of bounded random variables:

Lemma 2.2.1 ([16]). *Let Y be a random variable with $E[Y] = 0$, $a \leq Y \leq b$. Then for $s > 0$,*

$$E[\exp(sY)] \leq \exp\left(\frac{s^2(b-a)^2}{8}\right) \tag{2.1}$$

Based on this lemma we can easily construct an upper bound on PGF of sums of independent and bounded random variables.

Corollary 1. *Let X_i, $i = 1 \ldots n$ be independent random variables with $X = \sum_{i=1}^{n} X_i$, $M = E[X]$ and $a_i \leq X_i \leq b_i$. Then for $s > 0$,*

$$E[\exp(sX)] \leq \exp(sM) \exp\left(\frac{s^2(b-a)^2}{8}\right) \tag{2.2}$$

Now it is straightforward to apply this result to bound the probability generating function for our case when $0 \leq X_i \leq p_i$.

$$G_X(s) \leq \exp(sM) \exp\left(\frac{s^2}{8} \sum_{i=1}^{n} p_i^2\right) \stackrel{def}{=} G_{X,\text{hoe}} \tag{2.3}$$

Supposing that $p_i \leq 1$, another approximation can be derived form Hoeffding's results:

Theorem 2.2.2. *Let X_i, $i = 1 \ldots n$ be independent random variables with $X = \sum_{i=1}^{n} X_i$, $M = E[X]$ and $0 \leq X_i \leq 1$. Then for $s > 0$,*

$$G_X(s) \leq \left(\frac{M(e^s - 1) + n}{n}\right)^n. \tag{2.4}$$

Furthermore, it can be proved that this result provides the best approximation that can be achieved when $p_i \leq 1$, $\forall i$ [16].

Both previous approximations have their own pros and cons. Despite e.g. the optimal behaviour of (2.4) under given circumstances, if p_i differ substantially, typically $G_{X,\text{hoe}}$ gives better results. The new result presented in the following section mix the advantages of the two Hoeffding formulae inheriting the optimal behaviour of (2.4) even when p_i are different and also $p_i < 1$ is not required.

2.2.2 An Improved Hoeffding-type Approximation

I have obtained the following conservative bound for $E[\exp(sX)]$:

Theorem 2.2.3. *([J6]) Let X_i be independent bounded random variables with $0 \leq X_i \leq p_i$, $X = \sum_{i=1}^{n} X_i$ and $M = E[X]$. Then for $s > 0$,*

$$E[e^{sX}] \leq \prod_{i=1}^{n} \left(\frac{e^{sp_i} - 1}{p_i}\right) \left(\frac{M + \sum_{k=1}^{n} \frac{p_k}{e^{sp_k} - 1}}{n}\right)^n \stackrel{def}{=} G_{X,ih}. \tag{2.5}$$

Proof: From the definition of the PGF:

$$G_X(s) = E(e^{\sum_{k=1}^{n} sX_k}). \tag{2.6}$$

Due to independence, we obtain

$$G_X(s) = \prod_{k=1}^{n} G_{X_i} = \prod_{k=1}^{n} E(e^{sX_k}). \tag{2.7}$$

The exponential function is strictly convex, thus, the expectation value of the random variables e^{sX_k} can be bounded from above such as

$$E(e^{sX_k}) \leq 1 + m_k \frac{e^{sp_k} - 1}{p_k}. \tag{2.8}$$

That is, the PGF of the sum of the random variables can be bounded above by

$$G_X(s) \leq \prod_{k=1}^{n} \left(1 + m_k \frac{e^{sp_k} - 1}{p_k}\right). \tag{2.9}$$

The product on the right-hand side can be reformulated as

$$\prod_{k=1}^{n} \frac{e^{sp_k} - 1}{p_k} \prod_{j=1}^{n} \left(m_j + \frac{p_j}{e^{sp_j} - 1}\right), \tag{2.10}$$

the second term of which can be further approximated by

$$\prod_{j=1}^{n} \left(m_j + \frac{p_j}{e^{sp_j} - 1}\right) \leq \left(\frac{\sum_{j=1}^{n} \left(m_j + \frac{p_j}{e^{sp_j} - 1}\right)}{n}\right)^n, \tag{2.11}$$

because of the relation between the geometrical and arithmetical mean of non-negative real numbers. Now, an upper bound on the PGF can be expressed as

$$G_X(s) \leq \left(\frac{M + \sum_{j=1}^{n} \frac{p_j}{e^{sp_j} - 1}}{n}\right)^n \prod_{k=1}^{n} \frac{e^{sp_k} - 1}{p_k}. \tag{2.12}$$

Q.E.D.

Note that Theorem 2.2.3 can be considered as the generalized version of (2.4), in the sense that for the case of $p_k = 1$, $\forall k$ we get back (2.4). For later reference it is worth noticing the following important observation:

Corollary 2. *The PGF bound on the right-hand side in (2.9) is the exact probability generating function of the sum of heterogeneous on-off random variables with the distribution*

$$P(X_i^{\text{ON/OFF}} = p_i) = \frac{m_i}{p_i}\ ,\ P(X_i^{\text{ON/OFF}} = 0) = 1 - \frac{m_i}{p_i}.$$

This is because

$$E[e^{sX_i^{\text{ON/OFF}}}] = \left(1 - \frac{m_i}{p_i} + \frac{m_i}{p_i} e^{sp_i}\right).$$

The new approximation presented in the next section also has the property of permitting heterogenous p_i, however the main estimation step is based on a substantially different observation.

2.2.3 A New PGF Approximation Based on Stochastic Ordering

In this subsection let us recall the essential definitions and statements of stochastic ordering of random variables needed for PGF approximation.

Definition 2.2.4 ([26]). *Given two random variables X and Y with distribution function F_X and F_Y, respectively. Then, X is said to be smaller then Y with respect to increasing convex ordering, written as $X <_{icx} Y$, if the condition*

$$\int_{-\infty}^{\infty} \phi(x) \mathtt{d}F_X(x) \leq \int_{-\infty}^{\infty} \phi(x) \mathtt{d}F_Y(x) \tag{2.13}$$

holds for all increasing convex function ϕ, for which the integral exists.

An important consequence of this definition for the PGF's of random variables is the following:

Lemma 2.2.5. *Let X and Y be two random variables with the relation $X <_{icx} Y$. Then for $s > 0$, $G_X(s) \leq G_Y(s)$.*

This can be justified by the substitution $\phi(x) = \exp(sx)$.

The following important results presented in [26] leads us to construct a new PGF bound.

Lemma 2.2.6. *Let the random variables $X_1^{\text{ON/OFF}}, \ldots, X_n^{\text{ON/OFF}}$ represent n independent heterogeneous on-off sources with peak rates $p_1, \ldots, p_n$ and mean rates $m_1, \ldots, m_n$. Let $Y_1^{\text{ON/OFF}}, \ldots, Y_n^{\text{ON/OFF}}$ be n_Y independent homogeneous on-off sources with the identical peak rate $p = \max(p_i, i = 1, \ldots, n)$, $n_Y = \lceil \sum_{i=1}^{n} p_i/p \rceil$, and identical mean rate $m = \sum_{i=1}^{n} m_i/n_y$. Then*

$$X_{\text{ON/OFF}} <_{icx} Y_{\text{ON/OFF}}, \tag{2.14}$$

where

$$X_{\text{ON/OFF}} \stackrel{def}{=} \sum_{i=1}^{n} X_i^{\text{ON/OFF}} \text{ and } Y_{\text{ON/OFF}} \stackrel{def}{=} \sum_{i=1}^{n_Y} Y_i^{\text{ON/OFF}}. \tag{2.15}$$

For a proof of Lemma 2.2.6 see [26].

Now, the new PGF bound can be formulated in the following theorem:

Theorem 2.2.7. *Let $X_1, \ldots, X_n$ indicate n independent random variables with $0 \leq X_i \leq p_i$, $X = \sum_{i=1}^{n} X_i$ and $M = E[X]$. Then for $s > 0$,*

$$G_X(s) \leq \left(1 - \frac{M}{n_Y p} + \frac{M}{n_Y p} e^{sp}\right)^{n_Y} \stackrel{def}{=} G_{X,so} \tag{2.16}$$

Proof: By Corollary 2 we have

$$G_X(s) \leq G_{X_{\text{ON/OFF}}}(s) \ , \forall \ s > 0. \tag{2.17}$$

Further, by combining Lemma 2.2.6 and Lemma 2.2.5 the following relation also holds:

$$G_{X_{\text{ON/OFF}}}(s) \leq G_{Y_{\text{ON/OFF}}}(s) \ , \forall \ s > 0. \tag{2.18}$$

The two inequalities above gives the statement of the theorem, because

$$G_{Y_{\text{ON/OFF}}}(s) = \left(1 - \frac{M}{n_Y p} + \frac{M}{n_Y p} e^{sp}\right)^{n_Y}. \tag{2.19}$$

Q.E.D.

2.3 Important properties of PGF approximations

In the previous section I have presented two PGF approximations (conservative bounds) that have several approximation steps throughout their derivation which ignores the generating function nature of the formulae. However, as we will see in the next chapter, it is essential for these formulae to preserve some important properties of moment generating functions of random variables. These properties are proved to be essential to be inherited to the approximations, without which the estimations—in most practical cases—become unreasonable to use. Under current investigation these properties come from the behaviour of the PGF around zero, namely that:

Property 1: $G_X(s)|_{s=0} = 1$

Property 2: $\frac{d}{ds} G_X(s)|_{s=0} = M.$

Second order behaviour of the estimations are not considered, as the given formulae are only based on first order moments of the random variables under investigation.

Theorem 2.3.1. *PGF approximation $G_{X,ih}(s)$ fulfills both Property 1 and Property 2 that is*

$$\lim_{s \to 0} G_{X,ih}(s) = 1 \tag{2.20}$$

and

$$\lim_{s \to 0} \frac{d}{ds} G_{X,ih}(s) = M \tag{2.21}$$

To prove the above theorem, we need the following lemma:

Lemma 2.3.2.

$$\lim_{s\to 0}\frac{\exp(sp_1)-1}{\exp(sp_2)-1}=\frac{p_1}{p_2} \tag{2.22}$$

Applying e.g. the l'Hospital rule the above statement can easily be proved.
Proof of Theorem 2.3.1: Let us take (2.5)

$$G_{\mathrm{ih}}(s)=\prod_{i=1}^{n}\left(\frac{e^{sp_i}-1}{p_i}\right)\left(\frac{M+\sum_{k=1}^{n}\frac{p_k}{e^{sp_k}-1}}{n}\right)^{n}. \tag{2.23}$$

and rearrange the formula such that

$$G_{\mathrm{ih}}(s)=\prod_{i=1}^{n}\frac{e^{sp_i}M-M+\sum_{k=1}^{n}p_k\frac{e^{sp_i}-1}{e^{sp_k}-1}}{p_in}. \tag{2.24}$$

Now taking the $i-th$ factor of the multiplication and performing the limitation as $s\to 0$ applying Lemma 2.3.2 we get

$$\lim_{s\to 0}\frac{e^{sp_i}M-M+\sum_{k=1}^{n}p_k\frac{e^{sp_i}-1}{e^{sp_k}-1}}{p_in}=\frac{M-M+\sum_{k=1}^{n}p_k\frac{p_i}{p_k}}{p_in}=1. \tag{2.25}$$

As all the factors in the product (2.24) are equal to 1 as $s\to 0$ we get (2.20).

To justify (2.21), first we need to derivate (2.24) with respect to s. According to the rule of the derivation of multiplication, for $\frac{d}{ds}G_{X,\mathrm{ih}}(s)$ we get:

$$\sum_{j=1}^{n}\frac{d}{ds}\left(\frac{e^{sp_j}M-M+\sum_{k=1}^{n}p_k\frac{e^{sp_j}-1}{e^{sp_k}-1}}{p_jn}\right)\prod_{i=1,i\neq j}^{n}\frac{e^{sp_i}M-M+\sum_{k=1}^{n}p_k\frac{e^{sp_i}-1}{e^{sp_k}-1}}{p_in}. \tag{2.26}$$

The left side of the above formula is

$$\frac{d}{ds}\left(\frac{e^{sp_j}M-M+\sum_{k=1}^{n}p_k\frac{e^{sp_j}-1}{e^{sp_k}-1}}{p_jn}\right)= \tag{2.27}$$

$$\frac{1}{p_jn}\left(Mp_je^{sp_j}+\sum_{k=1}^{n}p_k\frac{p_je^{sp_j}(e^{sp_k}-1)-p_ke^{sp_k}\left(e^{sp_j}-1\right)}{(e^{sp_k}-1)^2}\right). \tag{2.28}$$

Now let us perform the limitation. According to (2.25), the multiplication on the right-hand side of (2.26) equals to 1, applying also Lemma 2.3.2 (or applying the l'Hospital rule twice) to find the limit of (2.28) at s=0 we get:

$$\lim_{s\to 0}\frac{d}{ds}G_{X,\mathrm{ih}}(s)=\sum_{j=1}^{n}\frac{1}{p_jn}\left(Mp_j+\sum_{k=1}^{n}p_j\frac{p_j-p_k}{2p_k}\right). \tag{2.29}$$

That is

$$\lim_{s\to 0}\frac{d}{ds}G_{X,\mathrm{ih}}(s)=\frac{1}{n}\sum_{j=1}^{n}p_j\frac{M}{p_j}=M. \tag{2.30}$$

Q.E.D.

Theorem 2.3.3. *PGF approximation $G_{X,so}$ fulfills both Property 1 and Property 2 that is*

$$G_{X,so}(s)|_{s=0} = 1 \tag{2.31}$$

and

$$\frac{d}{ds}G_{X,so}(s)|_{s=0} = M \tag{2.32}$$

Proof: As it is shown previously:

$$G_{X,so(s)} = E[e^{sY^{\text{ON/OFF}}}], \tag{2.33}$$

that is there exists a random variable (namely $Y^{\text{ON/OFF}} = \sum_{i=1}^{n_y} Y_{i,\text{ON/OFF}}$, where $Y_{i,\text{ON/OFF}}$ are homogenous random variables defined as $0 \leq Y_{i,\text{ON/OFF}} \leq p$ and $E[Y_{i,\text{ON/OFF}}] = \frac{M}{n_y}$) that has the same PGF as $G_{X,\text{so}}(s)$, holding all the properties of a general PGF. Q.E.D.

Formula (2.33) gives the basis for the investigations carried in connection with the optimality of PGF bound $G_{X,\text{so}}(s)$.

As it is shown, the presented two PGF bounds seem to be quite reasonable approximations, and in the next section I will also prove that under certain conditions they are the optimal choice, that is no better can be given.

2.4 Analytical investigation of PGF bounds $G_{X,\mathbf{ih}}(s)$ and $G_{X,\mathbf{so}}(s)$

Results in recent publications building around an approximation of the distribution of sums of random variables often supported by showing that comparing them to some other well known formulae they perform better in some special cases (e.g. to some representative parameter sets of a simulation scenario). General optimality of them even under some restricted conditions however can rarely be proven. Contrary to this practice both to $G_{X,\text{ih}}(s)$ and $G_{X,\text{so}}(s)$ can be given such criterion .

Theorem 2.4.1. *([C17])*

$G_{X,ih}(s)$, the PGF approximation of a sum of positive real valued random variables is optimal on a S set of s if:

1. *All the information known are n the number of the random variables, the aggregate mean M and the individual maximum values $p_1, p_2, \ldots, p_n$ of the random variables.*

2. The following inequality holds

$$0 \leq \frac{M - \sum_{k=1}^{n} \left(\frac{p_i}{e^{sp_i}-1} - \frac{p_k}{e^{sp_k}-1} \right)}{n} \leq p_i, \ \forall i, \ s \in S. \tag{2.34}$$

Before stepping further to the proof of the above theorem, it is worth taking a closer look at condition (2.34), as the exact meaning of the inequality is not straightforward. In short it tells us that p_i maximum values should not diverge substantially, only between bounds set by the average of the individual mean values (e.g. the inequality becomes apparent when the maximum values are equal). On the other hand the result is also the function of s the often called *space* parameter. As it will be seen in the next chapter, for the purpose of traffic characterisation, the resulted formulae (that will use the PGF approximations) will be the subject of optimization over parameter s, which sets the typical operation interval of it. In practice, set S established by condition (2.34), according also to the results of Chapter 4, turns out to contain the operation point (in several extreme cases, other simplification steps can also be carried out to achieve acceptable results).

Proof of Theorem 2.4.1: To justify the theorem, we should take a closer look at the proof of Theorem 2.2.3. In the derivation process, basically at each step of the proof we subsequently obtain new upper bounds on previous ones till we get to the concluding form, which depends only on known parameters. In the following we attempt to find a particular traffic situation with given M, n and p_i so that at each bounding step throughout the proof the inequalities become equalities. If such a case exists it follows that an upper bound on all the traffic situations cannot be smaller than $G_{X,\mathrm{ih}}(s)$ (since there exists at least one special case when an equation holds).

There are two bounding steps in the derivation that should be considered. At the first one in formula (2.8):

$$E(e^{sX_k}) \leq 1 + m_k \frac{e^{sp_k} - 1}{p_k}. \tag{2.35}$$

the inequality changes to equality simply if the random variables are On/Off types, that is, take only 0 and p_i with probability such that their mean value becomes m_i ($P(X_i = p_i) = \frac{m_i}{p_i}$). The result at (2.9), as it is stated in Corollary 2, is the exact probability generating function of the sum of heterogeneous on-off random variables with the distribution

$$P(X_i^{\mathrm{ON/OFF}} = p_i) = \frac{m_i}{p_i} \ , \ P(X_i^{\mathrm{ON/OFF}} = 0) = 1 - \frac{m_i}{p_i}.$$

The situation is a bit more complicated in the case of bounding step (2.11). It is known that the geometric and arithmetic mean of non-negative real numbers are equal if and only if the numbers are equal, so in our case

$$m_i + \frac{p_i}{e^{sp_i} - 1} = m_j + \frac{p_j}{e^{sp_j} - 1}, \tag{2.36}$$

(where $i = 1, \ldots, n$ and $j = 1, \ldots, n$) should be satisfied. Summing up such equations for $j = 1, \ldots, n$ with i fixed, we get:

$$m_i = \frac{M - \frac{np_i}{e^{sp_i}-1} + \sum_{j=1}^{n} \frac{p_j}{e^{sp_j}-1}}{n}, \quad \forall i, \tag{2.37}$$

since $\sum_{j=1}^{n} m_j = M$. Now, if

$$0 \leq m_i \leq p_i, \ \forall i \tag{2.38}$$

is also satisfied then (2.37) provides us a recipe to construct the random variables, in which case the saturation probability of the aggregate flow cannot be bounded by a smaller value than $G_{X,\mathrm{ih}}(s)$ gives us. To verify that the above variable set with the provided m_i really exists, we should check if $\sum_{i=1}^{n} m_i = M$. Summing up the m_i in (2.37) we get:

$$\sum_{i=1}^{n} m_i = \frac{nM - \sum_{i=1}^{n} \frac{np_i}{e^{sp_i}-1} + \sum_{j=1}^{n} \frac{p_j}{e^{sp_j}-1}}{n} \tag{2.39}$$

that is

$$\sum_{i=1}^{n} m_i = M - \frac{n\left(\sum_{i=1}^{n} \frac{p_i}{e^{sp_i}-1} - \sum_{j=1}^{n} \frac{p_j}{e^{sp_j}-1}\right)}{n} = M. \tag{2.40}$$

If condition (2.38) is not fulfilled at a fixed s, than (2.37) cannot be used to construct the random variables, that is, there is no random variable set with given properties, the sum of which has exactly $G_{X,\mathrm{ih}}(s)$ as its PGF at point s, i.e. in general $G_{X,\mathrm{ih}}(s)$ is not guaranteed to be the tightest one. Q.E.D.
Such optimality criterion can also be given to $G_{X,\mathrm{so}}(s)$ as follows:

Theorem 2.4.2. *$G_{X,so}(s)$ is the optimal PGF bound of the sum of random variables, when all the information known about the set of random variables are n as the number of the variables, M as the aggregate mean, and p_i as the maximum values of the variables with*

$$p_i = p_j, \forall i, j. \tag{2.41}$$

Proof: The justify the theorem, we should follow a very similar thread of thoughts as in the case of $G_{X,\mathrm{ih}}(s)$. If we find a specific case among the possible random variable sets, where equality holds between the PGF and $G_{X,\mathrm{so}}(s)$, we are ready with the proof. Indeed in the derivation of $G_{X,\mathrm{so}}(s)$, at (2) we used the PGF of an existing random variable set. Q.E.D.
Note that according to Theorem 2.4.1 and 2.4.2:

$$G_{X,\mathrm{ih}}(s) = G_{X,\mathrm{so}}(s), \tag{2.42}$$

when $p_i = p_j, \forall i, j$, which is easy to verify with the substitution $p_i = p, \forall i$. This means that PGF approximation $G_{X,\mathrm{ih}}(s)$ gives optimal results for a much wider space of random variable sets than $G_{X,\mathrm{so}}(s)$. However, we will see, that in the application of the approximations, an optimization process should be carried out with respect to s. This process is much easier in the case of $G_{X,\mathrm{so}}(s)$ in many cases resulting in a closed formula. To reach the same result with $G_{X,\mathrm{ih}}(s)$, we need further approximation steps, which often make the outcome less accurate.

In the next chapter, we change the subject from the investigation of PGF bounds to communication resource meters, where the results achieved in this chapter turns out to be quiet beneficial.

Chapter 3

Conservative Upper-bounds and Approximations for Resource Assessment

In QoS aware networks, besides service differentiation, service quality-level guaranties are also to be offered in a reliable manner. To provide such guaranties, efficient measures of available resources should be constructed, which are practical enough to be used in real networks. Current chapter considers such resource usage meters in the context of the bufferless network model.

Bufferless fluid flow multiplexing (*bffm*) is often used in the literature to analyze QoS measures e.g. cell loss probability in a multiplexer. Because this approach assumes no buffer at burst time scales, it is able to provide conservative estimates for the QoS measures under question. For modeling purposes under *bffm*, let us assume that we have n fluid flows to be multiplexed on a communication link with transmission capacity C. Let the instantaneous stationary (that is time dependence is eliminated) arrival rate of flow i be noted by X_i, as a random variable. Because every flow has a peak rate p_i we also have $0 \leq X_i \leq p_i$. Further, let the aggregate flow arrival rate be $X = \sum_{i=1}^{n} X_i$.

A frequently investigated and estimated measure of the quality of data transmission on a link is the saturation probability:

$$P_{\text{sat}} \stackrel{def}{=} P(X > C).$$

This probability reflects the fraction of time when the link is overloaded, that is the combined arrival rate exceeds the link capacity. Although in several papers

the estimates of the saturation probability have been proposed to approximate the loss probability, it also turned out that in numerous traffic situations the loss performance can not be analyzed well through this measure. Nevertheless as resource-based congestion measure, it can still be important, at least from network providers point of view.

Some authors paid considerable attention to the direct estimates of workload loss ratio under *bffm*. The workload loss ratio (or loss probability) can be identified as

$$WLR \stackrel{def}{=} E[(X-C)^+]/E[X],$$

where $E[.]$ stands for the expectation value operator and
$(X-C)^+ = \max(X-C, 0)$. The estimation of this quantity can provide more accurate loss performance analysis, although the analytical formulation of the corresponding estimators, is considerably harder. This measure characterizes the expected loss rate better and could also quantify user satisfaction.

From traffic management (e.g. connection admission control) point of view two important questions can arise. First, whether the ongoing session (possibly together with a newcomer) satisfies a predefined QoS constraint relating to some quality of service measure. In a more formal way, the inequalities

$$P(X > C) \leq e^{-\gamma} \quad , \quad \frac{E[(X-C)^+]}{E[X]} \leq e^{-\gamma} \tag{3.1}$$

represent the fulfillment of the constraint on saturation probability and workload loss ratio, respectively. Second, an important and significant question in practice to identify the minimum capacity to be allocated for the traffic flows, which ensures the satisfaction of the corresponding constraint in (3.1). Again, formally it can be written as

$$C_{\text{equ,sat}} \stackrel{def}{=} \inf\{C : P(X > C) \leq e^{-\gamma}\} \tag{3.2}$$

and

$$C_{\text{equ,wlr}} \stackrel{def}{=} \inf\{C : WLR \leq e^{-\gamma}\}. \tag{3.3}$$

One of the widely used and accepted techniques to approximate P_{sat} and WLR is the Chernoff bounding method. The Chernoff bound of P_{sat} is as follows:

$$P(X > C) \leq \inf_{s>0} \frac{G_X(s)}{e^{sC}} = \inf_{s>0} \exp\left(\Lambda_X(s) - sC\right), \tag{3.4}$$

where $G_X(s) \stackrel{def}{=} E[\exp(sX)]$ and $\Lambda_X(s) \stackrel{def}{=} \log G_X(s)$ are the probability generating function and the cumulant generating function of X, respectively. Generally, the Chernoff bounding method means that Chernoff bounds are to be further bounded or approximated using the available information on X.

Applying this method, using the results of the previous chapter, in the following, efficient conservative bounds on P_{sat}, WLR are presented, and based on them computationally equivalent formulae have been derived for bounding $C_{\text{equ,sat}}$ and $C_{\text{equ,wlr}}$ [J5], [J6]. A polynomial (non-Chernoff-like) bounding method has also been investigated [J6], that has the characteristic feature of dropping the requirement of full independence among the individual traffic sources, providing, however, comparable results for many practical cases. I also show, that an improved approximation technique [C16, C17] (in contrast to strict conservative upper bounds presented previously) on buffer overflow probability known in buffered statistical multiplexer can also be applied to approximate P_{sat} under *bffm*. Here, an open problem is, as to how to apply this improved technique on WLR, and furthermore to provide efficient estimators for $C_{\text{equ,sat}}$ and $C_{\text{equ,wlr}}$ based on this refined approximation. The results presented in this chapter provide estimators of P_{sat}, WLR, $C_{\text{equ,sat}}$ and $C_{\text{equ,wlr}}$, which could be based on any kind of approximations of the underlying probability generating functions.

3.1 Conservative Upper-bounds for P_{sat} and WLR

Using conservative bounds of resource usage measures in traffic management is a possible way of *guaranteeing* a certain level of quality in communication networks. In the following, I will investigate two such measures: the saturation probability and the workload loss ratio, and present some estimation of them constructed by the well-known Chernoff-Hoeffding bounding method. The introduced theorems are directly implied by the results of the previous chapter about PDF approximations. At the end of the section an interesting, unique, polynomial estimation technique will be shown yielding similar results.

Theorem 3.1.1. *If $X_1, X_2, \ldots, X_n$ are independent (and not necessarily identically distributed) random variables, for which $0 \leq X_i \leq p_i$ holds, then*

$$P(X > C) \leq e^{-s^*C} \left(\frac{M + \sum_{j=1}^{n} \frac{p_j}{e^{s^* p_j} - 1}}{n} \right)^n \prod_{k=1}^{n} \frac{e^{s^* p_k} - 1}{p_k}, \qquad (3.5)$$

where s^* is the solution of the following equation.

$$\sum_{k=1}^{n} \frac{e^{sp_k} p_k}{e^{sp_k} - 1} - \frac{n \sum_{j=1}^{n} \frac{e^{sp_j} p_j^2}{(e^{sp_j}-1)^2}}{M + \sum_{j=1}^{n} \frac{p_j}{e^{sp_j}-1}} - C = 0. \tag{3.6}$$

Proof: The starting point is the Chernoff formula:

$$P(X > C) \leq \inf_{s>0} \frac{G_X(s)}{e^{sC}} \leq \inf_{s>0} \frac{\widetilde{G}_X(s)}{e^{sC}}, \tag{3.7}$$

where $G_X(s) \stackrel{def}{=} E[\exp(sX)]$ is the probability generating function of X and $\widetilde{G}_X(s)$ is any kind of conservative upper bound on $G_X(s)$. With $X = \sum_{i=1}^{n} X_i$ and $\widetilde{G}_X(s) = G_{X,\mathrm{ih}}(s)$ (as in Theorem 2.2.3) we get:

$$P(X > C) \leq \inf_{s>0} e^{-sC} \left(\frac{M + \sum_{j=1}^{n} \frac{p_j}{e^{sp_j}-1}}{n} \right)^n \prod_{k=1}^{n} \frac{e^{sp_k} - 1}{p_k}. \tag{3.8}$$

What still remains to be done is to determine the optimal s which minimizes the right-hand side above. Unfortunately, a closed form expression for the optimal s cannot be derived, however, an essentially non-algebraic equation can be formulated, the solution of which serves as the optimal s. The upper bound above is minimized if its logarithm is minimized, that is, the optimal s minimizes the function

$$\sum_{k=1}^{n} \log \frac{e^{sp_k} - 1}{p_k} + n \log \left(\frac{M + \sum_{j=1}^{n} \frac{p_j}{e^{sp_j}-1}}{n} \right) - sC. \tag{3.9}$$

Taking the derivative of this with respect to s, we get (3.6), the solution of which gives the optimal s^*. Q.E.D.

Now using the definition of the WLR:

$$WLR \stackrel{def}{=} E[(X - C)^+]/E[X], \tag{3.10}$$

and observing that

$$E[(X - C)^+] = \int_0^\infty P(X \geq C + x)dx \leq \int_0^\infty \frac{G_X(s)}{e^{s(C+x)}} dx = \frac{G_X(s)}{se^{sC}}, \tag{3.11}$$

we can easily reach a conservative upper bound on the workload loss ratio:

$$WLR \leq \frac{e^{-s^*C}}{s^*M} \left(\frac{M + \sum_{j=1}^{n} \frac{p_j}{e^{s^*p_j}-1}}{n} \right)^n \prod_{k=1}^{n} \frac{e^{s^*p_k} - 1}{p_k}, \tag{3.12}$$

where s^* is the solution of the following equation.

$$\sum_{k=1}^{n} \frac{e^{sp_k} p_k}{e^{sp_k} - 1} - \frac{n \sum_{j=1}^{n} \frac{e^{sp_j} p_j^2}{(e^{sp_j}-1)^2}}{M + \sum_{j=1}^{n} \frac{p_j}{e^{sp_j}-1}} - C - \frac{1}{s} = 0. \tag{3.13}$$

Note that, in this case a further improvement can be reached [C17], if we notice in the derivation at (3.11) that the integral need not be carried out to infinity, it is enough to do it to $\sum_{i=1}^{n} p_i - C$ instead, as $X - C$ cannot get higher than that.

The conservative bound presented in the next theorem applies to the case when $\widetilde{G}_X(s) = G_{X,\mathrm{so}}(s)$:

Theorem 3.1.2. *Let $X_1, \ldots, X_n$ indicate n independent random variables (e.g. transmission rates of communication sources) with $0 \leq X_i \leq p_i$, $X = \sum_{i=1}^{n} X_i$ and $M = E[X]$. Then for $s > 0$,*

$$P(X > C) \leq \left(\frac{M - n_Y p}{C - n_Y p} \right)^{n_Y - \frac{C}{p}} \left(\frac{M}{C} \right)^{\frac{C}{p}}, \tag{3.14}$$

where $p = \max(p_i, i = 1, \ldots, n)$, $n_Y = \lceil \sum_{i=1}^{n} p_i / p \rceil$.

Proof: As in the previous theorem, we start with the Chernoff formula:

$$P(X > C) \leq \inf_{s>0} \frac{G_X(s)}{e^{sC}} \leq \inf_{s>0} \frac{\widetilde{G}_X(s)}{e^{sC}}, \tag{3.15}$$

where $X = \sum_{i=1}^{n} X_i$ and $\widetilde{G}_X(s) = G_{X,\mathrm{so}}(s)$ (as in Theorem 2.2.7) we get:

$$P(X > C) \leq \inf_{s>0} \frac{1}{e^{sC}} \left(1 - \frac{M}{n_Y p} + \frac{M}{n_Y p} e^{sp} \right)^{n_Y}. \tag{3.16}$$

As it is easy to obtain, the optimal s is:

$$s^* = \frac{1}{p} \log \frac{C(M - n_Y p)}{M(C - n_Y p)}, \tag{3.17}$$

which minimizes the right-hand side above. Substituting the s^* into (3.16) we obtain (2.16) directly. Q.E.D.

To construct the corresponding WLR bound for this case, we can use a similar process as above, obtaining:

$$WLR \leq \frac{1}{s^* M e^{s^* C}} \left(1 - \frac{M}{n_Y p} + \frac{M}{n_Y p} e^{s^* p} \right)^{n_Y}, \tag{3.18}$$

where s^* is the solution of the following equation.

$$\frac{n_Y p(n_Y p - M)}{n_Y p - M + Me^{ps}} + n_Y p - \frac{1}{s} - C = 0. \tag{3.19}$$

The remark at the end of the previous theorem also applies here.

The final result in this section uses $G_{X,\mathrm{hoe}}(s)$ as the employed PGF approximation:

Theorem 3.1.3 ([16])**.** *Let $X_1, \ldots, X_n$ indicate n independent random variables (e.g. transmission rates of communication sources) with $0 \leq X_i \leq p_i$, $X = \sum_{i=1}^{n} X_i$ and $M = E[X]$. Then for $s > 0$,*

$$P(X > C) \leq e^{\frac{-2(C-M)^2}{\sum_{i=1}^{n} p_i^2}}. \tag{3.20}$$

Proof: The theorem can easily be justified by using the PGF approximation $G_{X,\mathrm{hoe}}$ introduced at (2.3):

$$P(X > C) \leq \inf_{s>0} \frac{G_{X,\mathrm{hoe}}}{e^{sC}} = \inf_{s>0} \exp\left(\frac{s^2}{8}\sum_{i=1}^{n} p_i^2\right) e^{s(M-C)}.$$

Performing the optimization, we reach our result. Note that the positive coefficient of the term s^2 in the function to be optimized (that is the argument of the exponential function) insure the existence of the minimum. Q.E.D.

Note that in this case of PGF bounds $G_{X,\mathrm{so}}(s)$ and $G_{X,\mathrm{hoe}}(s)$, the P_{sat} bound can be formulated in closed form, which is a computationally more efficient way to use in practice, on the other hand the optimal s of the corresponding WLR bounds are again turned out to be implicit equations, involving s in an essentially non-algebraic way.

3.1.1 A polynomial, non-Chernoff-like bounding method

In this section, I render an approximation technique different from the Chernoff method [J6]. The main advantage of the technique that in some cases it can improve the Chernoff-like approach. The resulting formula gives an upper bound on the saturation probability (modifying to estimate the workload loss ratio, as in previous cases, however, would be a substantially more difficult task).

First, let us consider a closed form bounding technique on aggregate traffic consisting of Bernoulli distributed independent random variables, and show

that this upper bound also holds for arbitrary distributed sources with peak rates of one. As it will be seen, this estimation is at least as accurate as those of obtained by the Chernoff bounding method. After introducing the definition of 't-wise independence', we show that for the obtained bound full independence of the individual traffic sources is not required.

Let $B_1, \ldots, B_n$ be Bernoulli distributed independent random variables, $B = \sum_{k=1}^{n} B_k$ and $M_B = E[B]$. Let us introduce the general class of functions of the variables B_k and the constant vector $\mathbf{a}(a_1, a_2, \ldots, a_n)$ containing non-negative elements as

$$G(B_1, B_2..., B_n, \mathbf{a}) = \sum_{j=1}^{n} a_j \sum_{1 \le k_1 < k_2 ... < k_j \le n} B_{k_1} B_{k_2} \ldots B_{k_j}. \tag{3.21}$$

Because this function is monotone increasing in all its variables, and because the B_i's can take only binary values, for $C > 0$ the following two events are equivalent:

$$B \ge C \Leftrightarrow G(B_1, B_2..., B_n, \mathbf{a}) \ge \sum_{k=1}^{\lfloor C \rfloor} a_k \begin{pmatrix} \lfloor C \rfloor \\ k \end{pmatrix}, \tag{3.22}$$

and so

$$P(B \ge C) = P\left(G(B_1, B_2..., B_n, \bar{a}) \ge \sum_{k=1}^{\lfloor C \rfloor} a_k \begin{pmatrix} \lfloor C \rfloor \\ k \end{pmatrix} \right), \tag{3.23}$$

where $\lfloor C \rfloor$ is the integer part of C. From (3.23) using Markov's inequality we obtain

$$P(B \ge C) \le \frac{E\big[G(B_1, B_2..., B_n, \bar{a})\big]}{\sum_{k=1}^{\lfloor C \rfloor} a_k \begin{pmatrix} \lfloor C \rfloor \\ k \end{pmatrix}}. \tag{3.24}$$

Let X_k, $k = 1, \ldots, n$ be independent (but not necessarily identically distributed) random variables, where $0 \le X_k \le 1$. It can be proved in a straightforward manner ([30] page 8), that when $X = \sum_{k=1}^{n} X_k \ge C$, it follows that

$$\sum_{1 \le k_1 < k_2 ... < k_n \le j} X_{k_1} X_{k_2} ... X_{k_n} \le \begin{pmatrix} \lfloor C \rfloor \\ j \end{pmatrix}, \tag{3.25}$$

provided that $C > 0$ and $j \leq \lfloor C \rfloor$. The application of Markov's inequality gives us:

$$P(X \geq C) \leq \frac{E\big[G(X_1, X_2..., X_n, \bar{a})\big]}{\sum_{k=1}^{\lfloor C \rfloor} a_k \begin{pmatrix} \lfloor C \rfloor \\ k \end{pmatrix}}. \tag{3.26}$$

Exploiting the assumption that the variables X_k are independent, if $E[X_k] = E[B_k]$, $k = 1, 2...n$, we get

$$\begin{aligned} P(X \geq C) \leq \frac{E\big[G(X_1,X_2...,X_n,\bar{a})\big]}{\sum_{k=1}^{\lfloor C \rfloor} a_k \begin{pmatrix} \lfloor C \rfloor \\ k \end{pmatrix}} = \frac{G(E[X_1],E[X_2]...,E[X_n],\bar{a})}{\sum_{k=1}^{\lfloor C \rfloor} a_k \begin{pmatrix} \lfloor C \rfloor \\ k \end{pmatrix}} = \\ = \frac{G(E[B_1],E[B_2]...,E[B_n],\bar{a})}{\sum_{k=1}^{\lfloor C \rfloor} a_k \begin{pmatrix} \lfloor C \rfloor \\ k \end{pmatrix}}, \end{aligned} \tag{3.27}$$

which means that bounding the righthand-side of (3.27) can also be applied for $P(\sum_{k=1}^{n} X_k \geq C)$, where X_k are arbitrarily distributed random variables with $0 \leq X_k \leq 1$ and $E[X_k] = E[B_k]$, $k = 1, 2...n$.

From the relation of the arithmetical and geometrical means of non-negative real numbers, it can be easily seen, that for any $j > 0$ and $y_k > 0$ real numbers, $k = 1, \ldots, n$, $\sum_{1 \leq k_1 < k_2 ... < k_j \leq n} y_{k_1} y_{k_2} ... y_{k_j}$ is maximized by setting $y_{k_1} = y_{k_2} = ... = y_{k_n} = \frac{Y}{n}$, where $Y = \sum_{k=1}^{n} y_k$. This result can be used as a further approximation for the tail probability and avoids using the expectation value of the individual random variables instead only the aggregated one is considered, that is

$$P(X \geq C) \leq \frac{G(E[B_1], E[B_2]..., E[B_n], \bar{a})}{\sum_{k=1}^{\lfloor C \rfloor} a_k \begin{pmatrix} \lfloor C \rfloor \\ k \end{pmatrix}} \leq \frac{\sum_{j=1}^{n} a_j \begin{pmatrix} n \\ j \end{pmatrix} \left(\frac{E[B]}{n}\right)^j}{\sum_{k=1}^{\lfloor C \rfloor} a_k \begin{pmatrix} \lfloor C \rfloor \\ k \end{pmatrix}}. \tag{3.28}$$

Note that on the right-hand side, now, only the sum of the individual means $E[B]$ is present (obviously $E[X] = E[B]$).

To find the tightest upper bound (i.e. the least one) in (3.28) over vector $\mathbf{a}(a_1, a_2...a_n)$, notice that a_i, $i > \lfloor C \rfloor$ should be set to 0 since they contribute

non-negative terms to the numerator and nothing to the denominator. The optimization process is now reduced to the problem of minimizing a general rational function having equal number of terms in both the numerator and the denominator. At this stage, considering that

$$\inf_{a_1,a_2...,a_n} \frac{a_1c_1 + a_2c_2... + a_nc_n}{a_1d_1 + a_2d_2... + a_nd_n} = \min_i \frac{c_i}{d_i}, \tag{3.29}$$

for $a_i \geq 0$ and constants $c_i \geq 0$, $d_i \geq 0$ (for a proof see [J5]), we should calculate

$$\min_j \frac{\begin{pmatrix} n \\ j \end{pmatrix} \left(\frac{E[X]}{n}\right)^j}{\begin{pmatrix} \lfloor C \rfloor \\ j \end{pmatrix}}. \tag{3.30}$$

According to the equation

$$\left(\frac{\begin{pmatrix} n \\ j+1 \end{pmatrix} \left(\frac{E[X]}{n}\right)^{j+1}}{\begin{pmatrix} \lfloor C \rfloor \\ j+1 \end{pmatrix}} \right) / \left(\frac{\begin{pmatrix} n \\ j \end{pmatrix} \left(\frac{E[X]}{n}\right)^{j}}{\begin{pmatrix} \lfloor C \rfloor \\ j \end{pmatrix}} \right) = \frac{(n-j)\frac{E[X]}{n}}{\lfloor C \rfloor - j}, \tag{3.31}$$

which is less than, equal to, or greater than 1 according to as j is less than, equal to, or greater than $\frac{\lfloor C \rfloor - E[X]}{1-\frac{E[X]}{n}}$, our upperbound in (3.28) is minimized at

$$t^* = \lceil \frac{\lfloor C \rfloor - E[X]}{1 - \frac{E[X]}{n}} \rceil. \tag{3.32}$$

Finally, the upperbound

$$P(X \geq C) \leq \frac{\begin{pmatrix} n \\ t^* \end{pmatrix} \left(\frac{E[X]}{n}\right)^{t^*}}{\begin{pmatrix} \lfloor C \rfloor \\ t^* \end{pmatrix}} \tag{3.33}$$

is achieved.

For further investigation let us take the following definition:

Definition 3.1.4. *Let $X_1, X_2..., X_n$ be a collection of random variables. We say that $X_1, X_2..., X_n$ are t-wise independent, if for any $c_1, c_2..., c_n$, and for any distinct indices $1 \leq i_1, i_2..., i_t \leq n$:*

$$P(X_{i_1} = c_1, X_{i_2} = c_2..., X_{i_t} = c_t) = \prod_{j=1}^{t} P(X_{i_j} = c_j). \tag{3.34}$$

Notice that to compute (3.33), in (3.21) we need to sum up products consisting of only t^* number of random variables. Consequently, this bound does not demand full independence of the random variables, t^*-wise independence is enough.

Finally in this section, we show that bound (3.33) at least as good as the bounds derived by the Chernoff method, assuming equal peaks.

First, let us consider the power series expansion of $e^{s\sum_{k=1}^{n} B_k}$:

$$\frac{1}{K} e^{s\sum_{k=1}^{n} B_k} = \frac{1}{K}\left(1 + (s\sum_{k=1}^{n} B_k) + \frac{(s\sum_{k=1}^{n} B_k)^2}{2!} + ...\right), \tag{3.35}$$

where $K = e^{sC}$ is a constant for given s and C. Since $(B_k)^2 = B_k$ for any k, we get $\forall s \geq 0$, $\exists \mathbf{a}(a_1, a_2..., a_n), a_k \geq 0$, $k = 1, 2..., n$, such that

$$e^{s\sum_{k=1}^{n} B_k} = \sum_{j=1}^{n} a_j \sum_{1 \leq k_1 < k_2 ... < k_n \leq j} B_{k_1} B_{k_2} ... B_{k_n}. \tag{3.36}$$

This means that in (3.21), we consider a class of functions that includes all the functions the Chernoff bounding method considers. During the step at inequality (3.28), we restrict the class to be investigated to a smaller one, achieving that we do not need to know the individual mean values, only the sum of them should be considered. It is also true, that in the worst case with the sum of the means fixed (i.e. when the individual means are equal) the bounding functions in (3.28) become equal to the original ones, i.e. they are the best upper bounds that can be achieved, when only the sum of the means is known. Our investigation concerns all the elements of this class, consequently bound (3.33) should be at least as good as any other one derived by the Chernoff method, assuming that it depends only on the sum of the mean values of the random variables and also the peak rates are equal. It is also a consequence that in the case of these latter type of bounds like (2.4), the requirement for t^*-wise independence of the random variables is sufficient as well. Note that (3.36) explains our choice for the class of functions in (3.21) to be investigated.

The disadvantage of bound (3.33) is that it can be generalized to the case of different peak rates only with the previously mentioned 'normalization technique', which implies however further inaccuracy. This is the reason why many Chernoff-type bounds, nevertheless, come out to be better than bound (3.33), in practical cases.

3.2 Computationally feasible $P_{\mathbf{sat}}$ and WLR bounds

In the previous chapter, important resource assessment techniques have been discussed, with the aim of finding the best available bounds (relying on a given set of information). The major consideration was optimality from a theoretical point of view. In practice, however, especially in the case of simple (and as a consequence fast) network management hardwares/softwares with limited processing capacity, computational effectiveness is sometimes almost as important as correctness. Delayed decisions, often result in rerouting or even loss of connection.

The results using the framework of the adopted estimation technique (i.e. the Chernoff bounding method) are concentrated on finding the best available bounds with the aid of optimization with respect to s as a carefully chosen free parameter in the computation. Although the optimal parameter s^* can be numerically computed in a straightforward manner using any standard root-finding algorithm, the optimization may be a time consuming process and can hardly be accomplished in real time.

In current chapter, computationally efficient formulae for resource assessment will be presented based on previous results, providing often suboptimal, but closed form formulae. The fundamental idea to achieve such formulae is to find efficient ways of determining s that are close to s^* but without solving e.g. non-algebraic type recursive equations. The main advantage of such methods is that the conservative nature of the estimations remain unchanged, which is often necessary when guaranteeing a perceived level of quality in traffic management.

Theorem 3.2.1. *Let $X_1, \ldots, X_n$ be n independent random variables with* $0 \leq$

$X_i \leq p_i$, $X = \sum_{i=1}^{n} X_i$ *and* $M = E[X]$.

$$P(X > C) \leq \left(\frac{1}{n} \left(M + \sum_{j=1}^{n} \frac{p_j}{e^{\frac{C-M}{K} p_j} - 1} \right) \right)^n e^{-\frac{(C-M)C}{K}} \prod_{k=1}^{n} \frac{e^{\frac{C-M}{K} p_k} - 1}{p_k}, \tag{3.37}$$

where

$$K = \frac{1}{4} \sum_{k=1}^{n} p_k^2 - \frac{1}{n} \left(M - \frac{1}{2} \sum_{k=1}^{n} p_k \right)^2$$

Proof: To justify the above theorem we will use the approximation in Theorem 3.1.1:

$$P(X > C) \leq e^{-s^* C} \left(\frac{M + \sum_{j=1}^{n} \frac{p_j}{e^{s^* p_j} - 1}}{n} \right)^n \prod_{k=1}^{n} \frac{e^{s^* p_k} - 1}{p_k}, \tag{3.38}$$

where to find s^*, the following formula should be minimized with respect to s:

$$\sum_{k=1}^{n} \log \frac{e^{s p_k} - 1}{p_k} + n \log \left(\frac{M + \sum_{j=1}^{n} \frac{p_j}{e^{s p_j} - 1}}{n} \right) - sC. \tag{3.39}$$

To reach a more tractable formula let us take the first three terms in the Taylor series expansion of the fraction

$$\frac{e^{s p_k} - 1}{e^{s p_j} - 1}$$

with respect to s at $s = 0$, that is:

$$\frac{p_k}{p_j} + s \left(\frac{{p_k}^2}{2 p_j} - \frac{p_k}{2} \right) + s^2 \left(\frac{p_j p_k}{12} - \frac{{p_k}^2}{4} + \frac{{p_k}^3}{6 p_j} \right) + O(s^3).$$

Also using: $e^{s p_k} - 1 \approx s p_k + \frac{1}{2} s^2 {p_k}^2 + O[s^3]$ and $\log[1 + As + Bs^2] \approx As + \left(B - \frac{A^2}{2} \right) s^2 + O[s^3]$ from (3.39) we deduce:

$$sM + \left(\frac{1}{8} P_2 - \frac{1}{2n} \left(M - \frac{1}{2} P \right)^2 \right) s^2 - sC, \tag{3.40}$$

where $M = \sum_{k=1}^{n} m_k$, $P = \sum_{k=1}^{n} p_k$ and $P_2 = \sum_{k=1}^{n} p_k^2$. The expression above should be minimized to obtain a sub-optimal value for s. Thus,

$$\tilde{s}_{t1}^* = \frac{C - M}{\frac{1}{4} P_2 - \frac{1}{n} \left(M - \frac{1}{2} P \right)^2}. \tag{3.41}$$

Now it is possible to express an upper bound for the tail distribution, which does not contain the transformation parameter s. If we consider that for any $\tilde{s}^*_{t1} > 0$:

$$P(\sum_{k=1}^{n} X_k \geq C) \leq \frac{G_{X,\mathrm{ih}}(\tilde{s}^*_{t1})}{e^{\tilde{s}^*_{t1} C}}, \tag{3.42}$$

substituting $\tilde{s}^*_{t1}$ into (3.38) we reach (3.37).

What still remains to be done is to prove that the minimum in (3.41) does exist. To justify that it is enough to show that the quadratic coefficient in (3.40) is non-negative, i.e.

$$\frac{1}{8}P_2 - \frac{1}{2n}\left(M - \frac{1}{2}P\right)^2 \geq 0, \tag{3.43}$$

where $M = \sum_{k=1}^{n} m_k$, $P = \sum_{k=1}^{n} p_k$ and $P_2 = \sum_{k=1}^{n} p_k^2$.

It is known that the quadratic mean is always greater or equal than the arithmetic mean of non-negative real numbers, thus

$$P_2 \geq \frac{P^2}{n}. \tag{3.44}$$

Applying (3.44) to (3.43), we get

$$\frac{1}{8}P_2 - \frac{1}{2n}\left(M - \frac{1}{2}P\right)^2 \geq \frac{1}{8n}P^2 - \frac{1}{2n}\left(M - \frac{1}{2}P\right)^2 = \frac{1}{2n}M\left(P - M\right) \geq 0, \tag{3.45}$$

as obviously $P \geq M$ always. Q.E.D.

The following theorem describes another useful approach of construction computationally feasible formula:

Theorem 3.2.2. *Let $X_1, \ldots, X_n$ be n independent random variables with $0 \leq X_i \leq p_i$, $X = \sum_{i=1}^{n} X_i$, $M = E[X]$, $P = \sum_{i=1}^{n} p_i$ and $P_2 = \sum_{i=1}^{n} {p_i}^2$.*

$$P(X > C) \leq e^{-C\sqrt{\frac{n}{P_2}}\log\frac{C(P-M)}{M(P-C)}} \left(\frac{M + \sum_{j=1}^{n}\frac{1}{L_j}}{n}\right)^n \prod_{k=1}^{n} L_k, \tag{3.46}$$

where

$$L_j = \frac{1}{p_k}\left(\left(\frac{C(M-P)}{M(C-P)}\right)^{p_k\sqrt{\frac{n}{P_2}}} - 1\right)$$

Proof: Again we start the derivation using the results of Theorem 3.1.1. In this case to achieve a suboptimal s, we should first realize that the optimal s

can be given in explicit closed form when for every peak rate $p_k = p$ holds. Now equation (3.6) changes to

$$s^* = \frac{1}{p} \log \frac{C\,(np - M)}{M\,(np - C)}. \tag{3.47}$$

From this we can extrapolate the suboptimal parameter for the case of different peak rates as

$$\tilde{s}^*_{t2} = \sqrt{\frac{n}{\sum_{k=1}^{n} p_k^2}} \log \frac{C\,(\sum_{k=1}^{n} p_k - M)}{M\,(\sum_{k=1}^{n} p_k - C)}. \tag{3.48}$$

This choice still satisfies the requirement, that $\tilde{s}^*_{t2}$ should remain positive, since the argument of the logarithm in (3.48) is less than one only when $M > C$, which case, naturally, is of no relevance in practical cases. Substituting (3.48) into (3.5) we get, what was to be proven. Q.E.D.

An important property of $\tilde{s}^*_{t2}$ is, that as opposed to $\tilde{s}^*_{t1}$, it becomes the original optimal value of s if all the peak rates are equal to each other. For this we can expect that if the deviation of the peak rates is small, this latter solution gives better results than (3.37).

Note that instead of (3.48) the suboptimal parameters

$$\tilde{s}^*_{t2,k} = \sqrt[k]{\frac{n}{\sum_{k=1}^{n} p_k^k}} \log \frac{C\,(\sum_{k=1}^{n} p_k - M)}{M\,(\sum_{k=1}^{n} p_k - C)}, \tag{3.49}$$

for any $k > 0$, would be equally applicable, however according to our extensive numerical investigations, in practical cases $k = 2$ seems to show the best properties. The main reason here that makes it possible to construct an explicit formula was that we found a special case where the optimal s can be given is closed form. This is the property that is missing in several other approaches in later chapters.

To construct closed form WLR formula, instead of deriving it the same way as in the above theorems, we can use the results of Section 3.1 applying the following observation:

$$WLR \leq \inf_{s>0} \frac{1}{sM} e^{\widehat{\Lambda}_X(s) - sC} \leq \frac{1}{s^* M} e^{\widehat{\Lambda}_X(s^*) - s^* C}, \tag{3.50}$$

where $s^* = \operatorname{arginf}_s \widehat{\Lambda}_X(s) - sC$ and $\widehat{\Lambda}_X(s)$ is any kind of conservative bound on the CGF of X. Now, considering Theorem 3.1.2, it can be clearly seen, that the formula for the saturation probability is already in a closed form, which makes this kind of approach a more favorable solution than others in many practical cases.

3.3 Refined Approximations for Saturation Probability and Workload-loss Ratio

In the previous chapters, the discussed resource assessment techniques are presented in the form of conservative bounds. This kind of approach is justified when the measures are used to provide firm guarantees for a certain value. This is the case for example when a virtual circuit-switched connection is to be established. In several cases, however, (e.g. quasi-realtime applications) general approximations are also acceptable. In such situations, the general approximations are much more favorable as the estimation error is significantly smaller, yielding improved resource utilization.

To construct such resource assessment measures the so-called Bahadur-Rao improvement can be used [5], [17] as follows:

$$P_{\text{sat}} \approx \frac{1}{s^* \sqrt{2\pi\sigma^2(s^*)}} \exp\left(\Lambda_X(s^*) - s^* C\right). \tag{3.51}$$

$$WLR \approx \frac{1}{M s^{*2} \sqrt{2\pi\sigma^2(s^*)}} \exp\left(\Lambda_X(s^*) - s^* C\right), \tag{3.52}$$

where

$$\sigma^2(s) = \frac{\partial^2}{\partial s^2}\Lambda_X(s) \quad \text{and} \quad s^* = \text{arginf}_s\{\Lambda_X(s) - sC\}. \tag{3.53}$$

A set of approximations for P_{sat} and WLR could be obtained if the CGF bounds and the related s^* values presented before were used in the formulae above. Although these substitutions result closed-form expressions, these are quite complicated and hence not expressive in case of the improved Hoeffding and stochastic ordering based CGF approximations (2.5), (2.16), due to the presence of the second order derivative of CGF.

But most of all these approximations show relatively bad performance, because the second derivatives of the CGF bounds are not accurate approximations for the second derivatives of the exact CGF functions. This undesirable property motivates the application of a general approximation technique which eliminates the second derivative of CGF. The following theorem shows such an approximation approach.

Theorem 3.3.1. *Based on the Bahadur-Rao approximations presented above the following approximations hold for P_{sat} and WLR:*

$$P_{sat} \approx \exp\left(-I - \frac{1}{2}\log 4\pi I\right) \tag{3.54}$$

$$WLR \approx \exp\left(-I - \frac{1}{2}\log 4\pi I - \log s^* M\right) \tag{3.55}$$

where $-I = \inf_s\{\Lambda_X(s) - sC\}$ *and* $s^* = arginf_s\{\Lambda_X(s) - sC\}$.

Proof of Theorem 3.3.1 :

Assuming that $\Lambda_X(s)$ has the required derivatives (it is true if the arrivals are almost surely bounded) the Taylor's theorem can be applied such that [27]

$$0 = \Lambda(0) = \Lambda_X(s^*) - s^*\frac{\partial\Lambda_X(s^*)}{\partial s^*} + \frac{s^{*2}}{2!}\frac{\partial^2\Lambda_X(s^*)}{\partial s^{*2}} + r, \tag{3.56}$$

where

$$r = \frac{s^{*3}}{3!}\frac{\partial^3\Lambda_X(x)}{\partial x^3} \tag{3.57}$$

for some $x \in (0, s^*)$. Thus, if the remainder r is small (it is exactly zero if X has Gaussian distribution), we have the approximation

$$s^{*2}\sigma^2(s^*) \approx -2(\Lambda_X(s^*) - s^*C) = 2I. \tag{3.58}$$

In this way, the pre-factors before the exponent in the equations (3.51), (3.52) can be approximated as

$$\frac{1}{s^*\sqrt{2\pi\sigma^2(s^*)}} \approx \frac{1}{\sqrt{4\pi I}}\ ,\ \frac{1}{Ms^{*2}\sqrt{2\pi\sigma^2(s^*)}} \approx \frac{1}{Ms^*\sqrt{4\pi I}} \tag{3.59}$$

which gives the statement of the theorem. Q.E.D.

As it will be seen in the last chapter, numerical investigations and also simulations show that many times even these approximations are turn out to be conservative upper bounds, this property in general, however, is not guaranteed.

3.4 Equivalent Capacity Estimators

In general, the purpose of traffic management (or bandwidth management) in quality of service aware networks is to provide methods of sharing the limited amount of resources (above all bandwidth of communication links) of the network among the applications following predefined policies. The successful operation of the process depends on the effectiveness of the monitoring, measurement and decision making tools used. A central issue in the case of all of the three tasks is to construct efficient resource usage measuring techniques.

To perform bandwidth assessment in a given situation the controllable network/traffic/quality parameters are to be fixed. First, let us assume that we have a fixed capacity link, and intend to provide services to an application requiring a certain grade of service e.g. maximum packet loss level, minimum availability, minimum throughput, maximum delay. Under the given circumstances we can estimate the required amount of bandwidth needed and decide if the connection request can be served or not. If the connection is accepted, continuous operation should also be maintained.

In this chapter, two bandwidth estimation techniques are dealt with, based on previous results: bandwidth requirement of a connection request with maintained saturation probability and workload loss (e.g. packet or cell loss) ratio. In both cases, as previously, the bufferless modeling case is considered. The amount of capacity, which is required to fulfill a given condition is generally called the *equivalent capacity*. In the case of saturation probability and workload loss ratio it is formally defined using previously introduced notations as:

$$C_{\text{equ,sat}} \stackrel{def}{=} \inf\{C : P(X > C) \leq e^{-\gamma}\} \tag{3.60}$$

and

$$C_{\text{equ,wlr}} \stackrel{def}{=} \inf\{C : WLR \leq e^{-\gamma}\}. \tag{3.61}$$

An important traffic management process is the Connection Admission Control (CAC), where these traffic measures are essential. To illustrate the usage of the equivalent capacity let us take, as an example, the following general CAC algorithm:

The process begins with a connection request by a newcomer flow, either by demanding a specified amount of bandwidth, or specifying the traffic parameters of the new connection, e.g. average bit rate and maximum bit rate, and defining the required QoS to be fulfilled e.g. maximum packet loss rate.

During the next step, actions are made to determine the amount of bandwidth used on the link under consideration. To accomplish this, depending on former operation, we either sum up the bandwidth allocated by applications previously, or perform specific measurements to decide the consumed portion. In the measurement process, obviously, appropriate traffic differentiation is necessary, to separate flows with various needs. We conclude the step by providing available free bandwidth remaining on the link.

In the final step, investigation is needed to decide if the remaining bandwidth is enough to serve the newcomer flow, e.g. if maximum bit rate of

newcomer flow is given, it is to be compared to the available bandwidth, and decision about the acceptance should be made accordingly.

Obviously, when a new request arrives, and the series of steps discussed above starts again, the measurements will apply to the new system including also the previously added flow. In the algorithm sketched above, the equivalent capacity of the sum of the accepted flows is determined. The resource consumption of this aggregate flow is equivalent with a flow using the computed C_{equ} bandwidth and having the targeted transmission properties, e.g. packet loss. In this sense C_{equ} characterizes resource usage and respective transmission properties together.

Now, let us concentrate to the special type of equivalent capacity computation techniques investigated in this chapter, that is ones that applies to saturation probability and packet loss ratio. Exact solutions for the evaluation of $C_{\text{equ,sat}}$ or $C_{\text{equ,wlr}}$ rarely exist. The main reasons of this include, that in real applications, generally, we do not have fully characterized (e.g. individually measurable) sources with exact distributions, most of the time only the mean and/or the peak rates are known. Furthermore, even if the distributions were well defined, exact results would often still be computationally expensive to obtain. Thus, most of the time, adequate estimation techniques should be developed. The accuracy of these estimates strongly depends on the available information on the traffic, moreover any simplification (e.g. for faster operation) of the algorithms, may have effects also. In recent contributions, many attempts have been made for a wide range of traffic to evaluate such approximations [34], [35], [C27]. For the case, when only little information on the traffic (e.g. only the mean and peak rates) is known, the Chernoff bounding method is the most commonly used technique. The formula in the literature is commonly presented using the effective bandwidth (as opposed to the CGF preferred in this dissertation) in the form:

$$P(X \geq C_{\text{equ}}) \leq \inf_{s,s>0} \frac{E[e^{sX}]}{e^{sC_{\text{equ}}}} = \inf_{s,s>0} e^{s(\alpha(s)-C_{\text{equ}})}, \tag{3.62}$$

where $\alpha(s) = s^{-1} \log E[\exp(sX)]$ referred to as the effective bandwidth of aggregate traffic X [14], [22]. In this case, however, we are looking for C_{equ} in the function of the given saturation probability, or with the help of (3.11), in the modified inequality, the WLR. For the evaluation of (3.62), the exact distribution of X (or at least the logarithmic moment generating function of

it) should be known. Bounding $\alpha(s)$ (or $E[\exp(sX)]$), using only the known descriptors of X, evade this difficulty. The estimates of $\alpha(s)$ will be denoted by $\hat{\alpha}(s)$. In (3.62), the optimization task is performed to get the best upper bound (i.e. the smallest one) with respect to s.

Now, we have two ways to determine C_{equ} e.g. in the case of P_{sat}:

- With a freely chosen initial C, we decide P_{sat} as:

$$P(X \geq C) \leq \inf_{s,s>0} e^{s(\alpha(s)-C)} \leq \inf_{s,s>0} e^{s(\hat{\alpha}(s)-C)} = T(s_t^*, C), \qquad (3.63)$$

  where $T(s_t^*, C)$ is the optimal Chernoff-type upper bound of the tail probability of aggregate traffic X, above capacity C, using the optimizing parameter s as s_t^*. Than comparing the result to the targeted P_{sat}, we continue the process iteratively, till we get C_{equ} with acceptable error.

- Find appropriate techniques to explicitly compute C_{equ} belonging to a given saturation probability:

$$C_{\text{equ}}(\gamma) \leq \inf_{s,s>0} \alpha(s) + \frac{\gamma}{s} \leq \inf_{s,s>0} \hat{\alpha}(s) + \frac{\gamma}{s} = E(s_e^*, \gamma), \qquad (3.64)$$

  where $E(s_e^*, \gamma)$ stands for the smallest equivalent capacity (obtainable by Chernoff-type approximations), which is saturated by the aggregate traffic X only with probability $e^{-\gamma}$. In this case, the optimizing parameter s is denoted by s_e^*.

Whereas both methods (3.63) and (3.64) have their own area to use, many times, when the need for good performance measure arises in today QoS networks, either of them can be applicable. Let us consider, for example, the above mentioned (CAC) algorithm.

Briefly recalling the two possible admission control rules, we accept a new connection request if

$$T(s_t^*, C - p_{n+1}) \leq e^{-\gamma}, \qquad (3.65)$$

or alternatively

$$E(s_e^*, \gamma) + p_{n+1} \leq C, \qquad (3.66)$$

where p_{n+1} is the peak rate of the newcomer connection. Notice that in the moment of acceptance, worst case behaviour of the newcomer flow is assumed, that is it generates traffic at its peak rate.

If the decision is based on the saturation probability, the inequality (3.65) should be the acceptance condition. Similarly, (3.66) is used, when the equivalent capacity is considered for the admission control.

Apparently, inequality

$$e^{s(\hat{\alpha}(s)-C)} < e^{-\gamma} \tag{3.67}$$

is equivalent to inequality

$$\hat{\alpha}(s) + \frac{\gamma}{s} \leq C, \tag{3.68}$$

for every $s > 0$, nevertheless, it does not mean that the inequalities (3.65) and (3.66) are equivalent, since the parameters s_e^* and s_t^*, that minimize the upper bounds are usually different.

To determine, which one is the more stringent condition, we present the following theorem [J6, C17]:

Theorem 3.4.1. *Let us assume, that we have a given (aggregate) traffic characterized so, that a $\hat{\alpha}(s)$ is obtainable. For a given link capacity C and saturation probability γ, it is stated that*

$$\inf_{s,s>0} e^{s(\hat{\alpha}(s)-C)} < e^{-\gamma} \Leftrightarrow \inf_{s,s>0} \hat{\alpha}(s) + \frac{\gamma}{s} < C \tag{3.69}$$

and

$$\inf_{s,s>0} e^{s(\hat{\alpha}(s)-C)} = e^{-\gamma} \Leftrightarrow \inf_{s,s>0} \hat{\alpha}(s) + \frac{\gamma}{s} = C. \tag{3.70}$$

Proof: To proof the first statement of the theorem, let us take the left-hand side of statement (3.69) as our starting point:

$$\inf_{s,s>0} e^{s(\hat{\alpha}(s)-C)} < e^{-\gamma}. \tag{3.71}$$

Let us denote the optimal s in (3.71) with s_t^* so that

$$\inf_{s,s>0} e^{s(\hat{\alpha}(s)-C)} = e^{s_t^*(\hat{\alpha}(s_t^*)-C)} < e^{-\gamma}. \tag{3.72}$$

Now because

$$\inf_{s,s>0} \hat{\alpha}(s) + \frac{\gamma}{s} \leq \hat{\alpha}(s) + \frac{\gamma}{s}, \forall s, \tag{3.73}$$

from (3.72) finally we get

$$\inf_{s,s>0} \hat{\alpha}(s) + \frac{\gamma}{s} \leq \hat{\alpha}(s_t^*) + \frac{\gamma}{s_t^*} < C. \tag{3.74}$$

The proof of the reverse direction in (3.69) and both directions of (3.70) follow along the same lines. Q.E.D.

The immediate consequence of Theorem 3.4.1 is that the CAC mechanism based upon method (3.63) or (3.64) are equivalent regardless of the underlying different optimization tasks. However, it should be emphasized, that the statement holds for the optimal parameters s_t^*, s_e^* only. On the other hand, in practice, often suboptimal solutions are used like the closed forms in Section 3.2, in which case e.g. the admission control algorithms based on the two approaches could provide quite different results.

It is also important to note, that the results of Theorem 3.4.1 do not depend on the estimation technique used at $\hat{\alpha}(s)$. The only necessary condition we need is the monotone decreasing relation between the equivalent capacity and the saturation probability in the Chernoff bound. Note that after some further approximations on the resulting Chernoff-like upper bound [C27], [12] (e.g. for the sake of simplicity) this relation may disappear.

Restricting discussion to the practical case, when the known parameters of a given traffic situation are only the number of sources n, the individual peak arrival rates $p_1, p_2, .., p_n$ of the sources and the aggregate, let us take a look at the result of Theorem 3.1.1 specialized for the case, when the individual peak rates are all equal to one and the aggregate mean arrival rate M is known (e.g. measured):

$$\alpha(s) \leq \frac{n}{s} \log \frac{M(e^s - 1) + n}{n} = \hat{\alpha}(s). \tag{3.75}$$

Putting it into (3.63) and making the optimization, we obtain Hoeffding's result [16]:

$$P(X \geq C) \leq \left(\frac{M}{C}\right)^C \left(\frac{n-M}{n-C}\right)^{n-C}, \tag{3.76}$$

in which case the optimal s becomes in closed form:

$$s^* = \log \frac{C}{M} \frac{n-M}{n-C}. \tag{3.77}$$

Under the same assumptions, the upper bound for the equivalent capacity formula:

$$E(s_t^*, \gamma) = \inf_{s, s>0} \hat{\alpha}(s) + \frac{\gamma}{s} = \inf_{s, s>0} \frac{n}{s} \log \frac{M(e^s - 1) + n}{n} + \frac{\gamma}{s} \tag{3.78}$$

is obtained, which apparently has no closed form solution. This is a typical example supporting that in spite of the result in Theorem 3.4.1, to improve

exactness of the upper bounds, often the tail distribution explicit formulae (providing closed form solutions) should be preferred to use, as opposed to the ones for the equivalent capacity. Several other, simple bounds, apart from the previous ones, have been derived in the literature for the case, when even less information on the traffic is known [12], [C27]. On the other hand, however, as it is pointed out in [31] for the practical case of call admission control in bandwidth management, often more effective decision rules can be designed by checking the inequalities $\widetilde{C}_{\text{equ,sat}} < C$ or $\widetilde{C}_{\text{equ,wlr}} < C$, instead of using decision rules directly based on (3.1).

Using the notation $\widetilde{\Lambda}_X(s) = s\hat{\alpha}_X(s)$ as any kind of conservative upper bound of the CGF of X, let us introduce the definitions of the approximations for equivalent capacities formulated in (3.2) and (3.3).

$$\widetilde{C}_{\text{equ,sat}} \stackrel{def}{=} \inf\{C : \inf_{s>0} \exp(\widetilde{\Lambda}_X(s) - sC) \leq e^{-\gamma}\}, \tag{3.79}$$

$$\widetilde{C}_{\text{equ,wlr}} \stackrel{def}{=} \inf\{C : \inf_{s>0} \exp(\widetilde{\Lambda}_X(s) - sC - \log sM) \leq e^{-\gamma}\}. \tag{3.80}$$

According to Theorem 3.4.1, bound (3.79) and (3.80) can be formulated in a simpler and computationally more efficient way, such as

$$\widetilde{C}_{\text{equ,sat}} = \inf_{s>0} \frac{\widetilde{\Lambda}_X(s) + \gamma}{s} \ . \tag{3.81}$$

and

$$\widetilde{C}_{\text{equ,wlr}} = \inf_{s>0} \frac{\widetilde{\Lambda}_X(s) + \gamma - \log sM}{s} \ . \tag{3.82}$$

The application of the three CGF bounds coming from the PGF approximations (2.3), (2.5) and (2.16) in the equations above yields three conservative upper bounds for $C_{\text{equ,sat}}$ and $C_{\text{equ,wlr}}$ each. When the PGF bound $\widetilde{\Lambda}_{X,\text{hoe}} = \log G_{X,\text{hoe}}$ is used in (3.81) the well known bound on the equivalent capacity can be obtained [13]:

$$C_{\text{equ,sat}} \leq M + \sqrt{\frac{\gamma \sum_{i=1}^{n} p_i^2}{2}} \ . \tag{3.83}$$

Although closed-form bound can not be obtained by the application $\widetilde{\Lambda}_{X,\text{ih}} = \log G_{X,\text{ih}}$, for several traffic scenarios a slightly worse bound has been derived through a suboptimal closed-form optimizer $\tilde{s}^*$. The bound is as follows [34]:

$$C_{\text{equ,sat}} \leq \frac{\widetilde{\Lambda}_{X,\text{ih}}(\tilde{s}^*) + \gamma}{\tilde{s}^*} \ , \tag{3.84}$$

where

$$\tilde{s}^* = \sqrt{\frac{\gamma}{\frac{1}{4}\sum_{i=1}^{n} p_i^2 - \frac{1}{n}\left(M - \frac{1}{2}\sum_{i=1}^{n} p_i\right)^2}}\,. \tag{3.85}$$

The remaining bound $\widetilde{\Lambda}_{X,\mathrm{so}} = \log G_{X,\mathrm{so}}$ for $C_{\mathrm{equ,sat}}$ and all the three bounds for $C_{\mathrm{equ,wlr}}$ can not be expressed in closed-form in a reasonable way.[1]

Now, the question is how to obtain equivalent capacity formulae like in (3.81), (3.82) when the approximations from Section 3.3 are to be used. First, let us consider the saturation probability.

Theorem 3.4.2.

$$\widetilde{C}_{equ,sat} = \inf_{s>0}\left\{\frac{\widetilde{\Lambda}_X(s)}{s} + \frac{\gamma}{s} - \frac{\gamma \log 4\pi\gamma}{s(1+2\gamma)}\right\}. \tag{3.86}$$

Proof: The key step toward an equivalent capacity formula is to solve (approximately) the equation

$$-I - \frac{1}{2}\log 4\pi I = -\gamma, \tag{3.87}$$

for I as a variable. To that end let us first assume that $I = \gamma + \delta$. Combining this with (3.87) we have

$$\delta + \frac{1}{2}\log 4\pi(\gamma+\delta) = 0. \tag{3.88}$$

Taking the first two parts of the power series expansion of $\log 4\pi(\gamma+\delta)$ the following (approximate) equation can be obtained:

$$\delta + \frac{1}{2}\log 4\pi\gamma + \frac{\delta}{2\gamma} = 0, \tag{3.89}$$

which solution to δ is

$$\delta = -\gamma\frac{\log 4\pi\gamma}{2\gamma+1}. \tag{3.90}$$

Using again the assumption $I = \gamma + \delta$ we have

$$\inf_{s>0}\{\Lambda_X(s) - sC\} = \gamma\left(1 - \frac{\log 4\pi\gamma}{2\gamma+1}\right), \tag{3.91}$$

from which the equivalent capacity can be formulated as appeared in (3.86), due to the equivalence between (3.79) and (3.81). Q.E.D.

[1]The approximate solutions for the equations $\frac{\partial}{\partial s}\frac{\tilde{\Lambda}_X(s)+\gamma}{s} = 0$ were highly inaccurate resulting useless and unreasonable bounds.

Utilizing similar approximation technique to the case of workload loss ratio the corresponding equivalent capacity can be formulated as

$$\widetilde{C}_{\text{equ,wlr}} = \inf_{s>0} \left\{ \frac{\widetilde{\Lambda}_X(s)}{s} + \frac{\gamma}{s} - \frac{\gamma(2\log M + \log 4\pi\gamma)}{s(1+2\gamma)} \right\}. \tag{3.92}$$

A corresponding bandwidth requirement estimate for the WLR is obtained as follows:

Theorem 3.4.3.

$$\widetilde{C}_{equ,wlr} \approx \inf_{s>0} \left\{ \frac{\widetilde{\Lambda}_X(s) + \gamma - 1 + \log M + \frac{2\gamma}{1+2\gamma} \log \frac{1+2\gamma}{4M\sqrt{\pi}\gamma^{\frac{3}{2}}}}{-\frac{1}{M} + s} \right\} \stackrel{def}{=} \widetilde{C}^{\text{B-R}}_{equ,wlr} , \tag{3.93}$$

where $\widetilde{\Lambda}_X(s)$ *is any kind of suitable approximation of* $\Lambda_X(s)$.

Before proving this theorem an important property of s^* is highlighted:

Lemma 3.4.4.

$$s^* = \frac{\partial I(C)}{\partial C} . \tag{3.94}$$

Proof of Lemma 3.4.4 : By the definition of $I(C)$ in Theorem 3.3.1

$$\frac{\partial I(C)}{\partial C} = \frac{\partial}{\partial C} \left(s^* C - \Lambda_X(s^*) \right) \tag{3.95}$$

Note that s^* depends on C, hence, the right hand side of the equation above can be formed as

$$\frac{\partial}{\partial C} \left(s^* C - \Lambda_X(s^*) \right) = \frac{\partial s^*}{\partial C} C + s^* - \frac{\partial}{\partial s} \Lambda_X(s)|_{s=s^*} \frac{\partial s^*}{\partial C} = \frac{\partial s^*}{\partial C} \left(C - \frac{\partial}{\partial s} \Lambda_X(s)|_{s=s^*} \right) + s^* . \tag{3.96}$$

Since s^* is the solution of the following equation

$$0 = \frac{\partial}{\partial s} (\Lambda_X(s) - sC) = \frac{\partial}{\partial s} \Lambda_X(s)|_{s=s^*} - C \tag{3.97}$$

the right hand side in equation (3.96) equals to s^*. Q.E.D.

Proof of Theorem 3.4.3 : Combining the equations (3.3) and (3.55) one gets

$$-I - \frac{1}{2} \log 4\pi I - \log s^* M = -\gamma . \tag{3.98}$$

This should be approximately solved with respect to $I(C)$. The presence of s^* makes this task considerably harder than in the previous case, nevertheless,

utilizing the statement of Lemma 3.4.4 the following differential equation of $I(C)$ is obtained:

$$-I - \frac{1}{2}\log 4\pi I - \log M - \log \frac{\partial I(C)}{\partial C} = -\gamma \ , \tag{3.99}$$

which results in a differential equation of $\delta(C)$ with setting $I = \gamma + \delta$

$$\delta + \frac{1}{2}\log 4\pi(\gamma+\delta) + \log \frac{\partial \delta(C)}{\partial C} + \log M = 0 \ . \tag{3.100}$$

Unfortunately, this equation can not be solved with respect to $\delta(C)$ in closed form. However, applying again the approximation $\log 4\pi(\gamma+\delta) \approx \log 4\pi\gamma + \delta/\gamma$ the modified differential equation

$$\log \frac{\partial \delta(C)}{\partial C} + \left(1 + \frac{1}{2\gamma}\right)\delta(C) + \frac{1}{2}\log 4\pi\gamma + \log M = 0 \tag{3.101}$$

is obtained, whose general set of solution is

$$\delta(C, C[1]) = \frac{2\gamma \log\left(-\frac{(1+2\gamma)(-C+C[1])}{4M\sqrt{\pi}\gamma^{3/2}}\right)}{1+2\gamma} \ . \tag{3.102}$$

The constant $C[1]$ is to be determined by some reasonable boundary condition. This could be done by considering the extreme case of $C = 0$, with this setup the rate function becomes

$$I(C)|_{C=0} = -\inf_{s,s>0} \Lambda_X(s) = 0 \tag{3.103}$$

because $\Lambda_X(s)$ is monotone increasing with increasing s. That gives the equation

$$I(C)|_{C=0} = \gamma + \delta(C, C[1])|_{C=0} \tag{3.104}$$

from which the constant $C[1]$ can be determined as

$$C[1] = -\frac{4e^{-\frac{1}{2}(1+2\gamma)} M\sqrt{\pi}\gamma^{\frac{3}{2}}}{1+2\gamma} \ . \tag{3.105}$$

Now, the particular solution for δ in accordance with the boundary condition (3.103) is

$$\delta(C) = \frac{2\gamma \log\left(e^{-\frac{1}{2}-\gamma} + \frac{C(1+2\gamma)}{4M\sqrt{\pi}\gamma^{\frac{3}{2}}}\right)}{1+2\gamma} \ . \tag{3.106}$$

Although $\delta(C)$ already provides an approximate solution of equation (3.98) for $I(C)$, it is still inappropriate to express an equivalent capacity formula.

Towards the direction of having a solvable equation for C , it is also worth approximating $\delta(C)$ in such a way that the negligible term $e^{-\frac{1}{2}-\gamma}$ is skipped and $\log C$ is approximated of the first two parts of its power series expansion around M, $(\log C \approx \log M + (C-M)/M)$:

$$\delta(C) \approx \frac{2\gamma}{1+2\gamma}\log\frac{1+2\gamma}{4M\sqrt{\pi}\gamma^{\frac{3}{2}}} + \log M + \frac{C-M}{M} \stackrel{def}{=} \widetilde{\delta}(C) \ . \tag{3.107}$$

Now, a suitable approximate solution of (3.98) for $I(C)$ as

$$\widetilde{I}(C) \stackrel{def}{=} \gamma + \widetilde{\delta}(C) \tag{3.108}$$

for which the following approximate equivalence holds (similarly to the equivalence in (3.69) and (3.70)):

$$I(C) \approx \widetilde{I}(C) = \gamma + \widetilde{\delta}(C) \Leftrightarrow C \approx \inf_{s>0} \frac{\Lambda_X(s) + \gamma - 1 + \log M + \frac{2\gamma}{1+2\gamma}\log\frac{1+2\gamma}{4M\sqrt{\pi}\gamma^{\frac{3}{2}}}}{-\frac{1}{M}+s} \ . \tag{3.109}$$

Due to the equivalence, the right hand side determines a reasonable approximation for $\widetilde{C}_{\mathrm{equ,wlr}}$. Q.E.D.

Note that the three underlying generating function approximations serve three possible approximations for the equivalent capacity both in the saturation probability and workload loss ratio case.

	$\widetilde{\Lambda}_{X,\mathrm{hoe}}(s)$, [16]	$\widetilde{\Lambda}_{X,\mathrm{ih}}(s)$, [J6]	$\widetilde{\Lambda}_{X,\mathrm{so}}(s)$
$\widetilde{C}_{\mathrm{equ,sat}}$ [J6]	(3.83),[13]	(3.84), [34]	(3.81), (2.16)
$\widetilde{C}_{\mathrm{equ,wlr}}$	(3.82), (2.3)	(3.82), (2.5)	(3.82),(2.16)
$\widetilde{C}^{\mathrm{B-R}}_{\mathrm{equ,sat}}$	(3.86), (2.3)	(3.86), (2.5)	(3.86), (2.16)
$\widetilde{C}^{\mathrm{B-R}}_{\mathrm{equ,wlr}}$	(3.93), (2.3)	(3.93), (2.5)	(3.93), (2.16)

Table 3.1: Summary of the equivalent capacity estimates

In order to summarize the estimates presented, and also clearly highlighted the new contributions Table 3.1 is presented. The citations indicates the previously known generating function approximations and equivalent capacity formulae, while the new contributions are referred only as (combinations) of the corresponding equations.

Chapter 4

Numerical investigations and simulative results

The techniques proposed in the dissertation, provide solutions of calculating the saturation probability, the workload loss ratio and respective equivalent capacity values in both closed form and implicit formulae, using parsimonious traffic characterisation techniques. The results are all based on the Chernoff-bounding method leaving an optimization task in the formulation to be performed. In a few cases, this task can be simplified by solving an explicit equation, in others, it is done through suitable reformulation and approximation, however, in many cases none of these actions can be performed in a reasonable way. In the previous chapter, a number of formulae have been derived for the computation of equivalent capacity values, nevertheless, the attempt to yield upper bounds in closed forms, or even to make appropriate approximations to get one have met several difficulties. In these cases, to reach acceptable results, numerical approximation techniques are inevitable to use, yielding drastically increased processing time, which may be inadmissable in real-time traffic management.

In the first section of this chapter, I propose a novel recursive algorithm to fasten numerical evaluation of the formulae introduced in the dissertation, in a real environment. The focus is on the replacement of the optimization task contained in the formulae with a fixed-point equation, constructed in a way that the number of necessary steps towards a good estimation is reduced significantly. The effectiveness of the algorithm is also demonstrated through numerical examples.

In later parts of this chapter, numerical evaluation of the proposed QoS

measures will be presented in three main sections. First the PGF approximations of Chapter 2 are to be compared to each other and the exact one in a few traffic scenarios. In the comparisons, for better illustration, the logarithms of the PGF, that is the CGF is represented in each case. In the following, the formulae for the estimation of the saturation probability and workload loss ratio are investigated. In this case, obviously, the closed form upper-bounds and approximations are of major interest, as the behavior of the implicit formulae can easily be traced back to that of the underlying PGF approximations already presented previously. In the last section the equivalent capacity estimators are discussed. The comparisons are made according to the results presented in Table 3.1.

4.1 Recursive algorithm for fast equivalent capacity calculation

In the following, I will propose fixed-point algorithms, specifically designed to solve Chernoff-based formulae connected to the saturation probability (that is derived from (3.4)). The main message of this part, in the first place, is to provide means of computing equivalent bandwidth according to (3.82), but without the compromise of finding appropriate closed form approximations as e.g. in Section 3.2. As discussed previously, such type of approximations turned out to be rather difficult to find, due to the nature of the formulation. For saturation probability formulae, like (3.5), computationally feasible (closed-form) solutions are not too difficult to find, and so the proposed fixed-point algorithms are of less significance, nevertheless, in certain cases, the increased exactness of the new methods may have increased importance also. From the large number of existing techniques of obtaining the optimum point in (3.82) I have chosen the fixed-point type algorithm for several reasons. Most importantly this type of method in general provides very fast convergence, which is, as discussed previously, a fundamental requirement in real-time decision making processes. Also an important condition is that due to the approximating nature of the bounds to be optimized, more than one derivation on the objective function would yield unreasonable loss of exactness, for which reason the fixed-point type optimum search algorithm is also suitable. Finally, as it will be seen, for almost all cases to be examined the solvable fixed-point equation comes naturally.

I will present the proposed fixed-point algorithms in two steps: First, the straightforward and more simple versions will be derived. Then, due to the experienced weak convergent properties of the algorithms, next I also propose a modified technique, that is capable of providing the same results significantly faster, and also stability shows substantial improvement.

4.1.1 Fixed-point Equations for the Computation of Equivalent Capacity

Before proceeding to the derivation of the algorithms, first let us take a look at the exact problem to be solved. I have composed, previously, two formulae for the problem of finding equivalent capacity belonging to fixed saturation probability, depending on the used characteristic feature of an aggregated traffic source:

$$\widetilde{C}_{\text{equ,sat}} = \inf_{s>0} \frac{\widetilde{\Lambda}_X(s) + \gamma}{s} \ . \tag{4.1}$$

and

$$\widetilde{C}_{\text{equ,sat}} = \inf_{s>0} \widetilde{\alpha}_X(s) + \frac{\gamma}{s} \ . \tag{4.2}$$

The importance of the distinction stems from the fact that in practice, it is not straightforward as to which of the two functions (ie. $\Lambda_X(s)$ as the CGF or $\alpha_X(s)$ as the effective bandwidth) should or can be measured. Depending on the two choice, as it will be seen, the results are different.

First let us investigate the case of using $\Lambda_X(s)$, the CGF of X:

$$C_{\text{equ}} = \inf_{s} \frac{\Lambda(s) + \gamma}{s}. \tag{4.3}$$

Performing the optimization, we get:

$$\frac{\partial_s C_{\text{equ}}}{\partial s} = \frac{\frac{\partial_s \Lambda(s)}{\partial s} s - \Lambda(s) + \gamma}{s^2} = 0, \tag{4.4}$$

yielding:

$$s_{opt} = \frac{\Lambda(s) + \gamma}{\frac{\partial_s \Lambda(s)}{\partial s}}. \tag{4.5}$$

This last form induces the following fixed-point equation:

$$s_{n+1} = \frac{\Lambda(s_n) + \gamma}{\frac{\partial_s \Lambda(s)}{\partial s} \ |_{s=s_n}}. \tag{4.6}$$

Regarding to the stability of the algorithm above—due to $\Lambda(s)$ being undefined—general statements cannot be derived, only constraints can be shown with respect to $\Lambda(s)$ to be fulfilled. On the other hand using the PGF approximations, proposed in Chapter 2, a thorough investigation can be done through numerical examples using a wide set parameters settings. According to the results, in many cases the (4.6) is proofed to be unstable, however, when stable, the minimum has always been found.

To get an impression of the performance of the algorithm first let us take the example of the aggregation of ten theoretical traffic sources with diverse traffic parameters as given in Table 4.1.

Table 4.1: Data for a 10-source traffic aggregation

	Peak rates (Mbps)	Mean rates (Mbps)
Source 1	2	0.6
Source 2	3	0.9
Source 3	5	1.5
Source 4	7	2.1
Source 5	7	1.4
Source 6	9	2.7
Source 7	9	3.4
Source 8	11	3.0
Source 9	11	3.6
Source 10	15	4.5

Let us use (2.5) as the bases for $\widetilde{\Lambda}_{X,\mathrm{ih}}(s)$. We try to find $C_{\mathrm{equ,sat}}$, when the maximum acceptable saturation probability is fixed at a moderate level of 10^{-3}. In this relatively simple case it is easy to compute that the minimum is reached at $s = 0.1886$ and the corresponding best (i.e. minimum) $C_{\mathrm{equ,sat}} = 56,41$ Mbps. Now if starting the recursion according to (4.6) at an initial good guess of $s_1 = \frac{1}{n}$, we find that the results of the recursion at each steps are:

- $s_2 = 0.1010$
- $s_3 = 0.1021$
- $s_4 = 0.1032$
- $s_5 = 0.1043$
- $s_6 = 0.1054$

Table 4.2: Traffic mix of uncompressed voice and compressed video

	n_1	m_1	p_1	n_2	m_2	p_2
Scenario 2	100	0.0256	0.064	10	2	5

Putting out that iteration stops when the results does not change more than 1%, we get that the over 40 steps are needed for the algorithm.

Now for a more practical scenario let us take the traffic mix of the aggregation of uncompressed voice ($p_1 = 64$ kbit/s, $m_1 = 25.6$ kbit/s) and compressed video ($p_2 = 5$ Mbit/s, $m_2 = 2$ Mbit/s) flows with $n_1 = 100$, $n_2 = 10$. The data are summarized in Table 4.2. Fixing the target saturation probability at 10^{-4} using (2.16) as the PGF approximation we can compute the optimal $C_{\text{equ,sat}} = 46.579$ at $s = 0.350$. The results of the fixed-point algorithm in each steps is illustrated in Fig 4.1. As it is clearly seen, even in this relatively simple case, at least 12 steps were needed to reach the optimal point with acceptable accuracy.

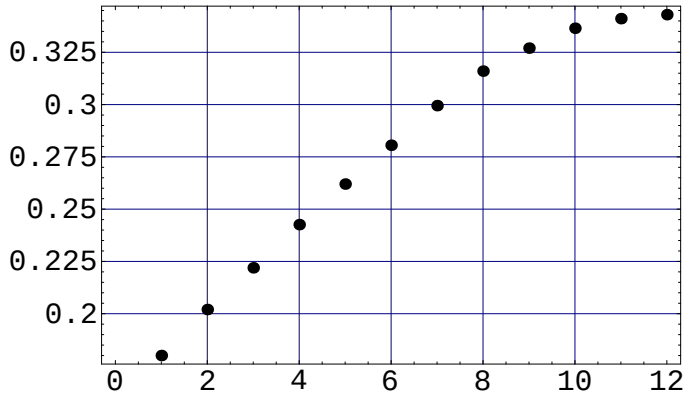

Figure 4.1: Steps of fixed-point algorithm in scenario 2

Unfortunately, as extensive numerical evaluation shows, in general, this behaviour turns out to be typical for the fixed-point algorithm given in (4.6), which property is almost prohibitive for using in real-time environment. In complex cases, often, even a few hundred steps are insufficient to find a good

optimal value. Moreover, the algorithm frequently shows an unstable behaviour, when s_{opt} goes to infinity.

To improve the low convergent rate and instability of the algorithm above I propose the following novel algorithm: We start with the usual C_{equ} formula:

$$C_{\text{equ}} = \inf_s \frac{\Lambda(s) + \gamma}{s}. \tag{4.7}$$

The main idea is to use a recursive technique, where we have the value of $\Lambda(s)$ and $\frac{\partial_s \Lambda(s)}{\partial s}$ at given points and at every step we fit these values onto the cumulant generation function of the gaussian distribution (that has two free parameters: the mean and the variance) and then we perform the optimization process accordingly.

The cumulant generating function of a gaussian distributed random variable (m: mean, σ^2: variance):

$$\Lambda^{\text{Gauss}}(s) = sm + \frac{s^2\sigma^2}{2} \tag{4.8}$$

Derivating the cumulant generating function with respect to s we get:

$$\frac{\partial_s \Lambda^{\text{Gauss}}(s)}{\partial s} = m + s\sigma^2 \tag{4.9}$$

Solving the two-equation system to m and σ^2 (ie.: adding (4.8) to $-\frac{s}{2}$ times (4.9) to get the mean and adding -1 times (4.8) to s times (4.9) to get the variance) we obtain:

$$m = \frac{2}{s}\Lambda^{\text{Gauss}}(s) - \frac{\partial_s \Lambda^{\text{Gauss}}(s)}{\partial s} \tag{4.10}$$

and

$$\sigma^2 = -\frac{2}{s^2}\Lambda^{\text{Gauss}}(s) + \frac{2}{s}\frac{\partial_s \Lambda^{\text{Gauss}}(s)}{\partial s} \tag{4.11}$$

Performing the optimization process in the formula of the equivalent capacity of the gaussian distributed random variable:

$$C_{\text{equ}}^{\text{Gauss}} = \inf_s \frac{\Lambda^{\text{Gauss}}(s) + \gamma}{s} = \inf_s m + \frac{s\sigma^2}{2} + \frac{\gamma}{s} \tag{4.12}$$

$$\frac{\partial_s C_{\text{equ}}^{\text{Gauss}}}{\partial s} = \frac{\sigma^2}{2} - \frac{\gamma}{s^2} = 0 \tag{4.13}$$

The results are:

$$s_{\text{opt}} = \sqrt{\frac{2\gamma}{\sigma^2}} \tag{4.14}$$

and

$$C_{\text{equ}}^{\text{opt}} = m + \sqrt{\frac{2\gamma}{\sigma^2}}\frac{\sigma^2}{2} + \frac{\gamma}{\sqrt{\frac{2\gamma}{\sigma^2}}} = m + \sqrt{2\sigma^2\gamma} \tag{4.15}$$

Now applying the fitting conditions at s_1, the first point of a recursive numerical process

$$\Lambda^{\text{Gauss}}(s_1) = \Lambda(s_1) \tag{4.16}$$

and

$$\frac{\partial_s \Lambda^{\text{Gauss}}(s)}{\partial s} \mid_{s=s_1} = \frac{\partial_s \Lambda(s)}{\partial s} \mid_{s=s_1} \tag{4.17}$$

we obtain the parameters of the gaussian distributed random variable at the first step:

$$m_1 = \frac{2}{s_1}\Lambda(s_1) - \frac{\partial_s \Lambda(s)}{\partial s} \mid_{s=s_1} \tag{4.18}$$

and

$$\sigma_1^2 = -\frac{2}{s_1^2}\Lambda(s) + \frac{2}{s_1}\frac{\partial_s \Lambda(s)}{\partial s} \mid_{s=s_1} . \tag{4.19}$$

For the following steps of the computation, from (4.14) we get our resulting fixed-point type recursive equation:

$$s_{n+1} = \sqrt{\frac{2\gamma}{\sigma_1^2}} = \sqrt{\frac{\gamma s_n^2}{-\Lambda(s_n) + s_n \frac{\partial_s \Lambda(s)}{\partial s} \mid_{s=s_n}}} \tag{4.20}$$

For the investigation of the performance of (4.20), let us take e.g. the 10-source traffic aggregation example set up previously. We find that starting from $s_1 = 0.1$, we get $s_2 = 0.1741$ and $s_3 = 0.1883$, which, according to our stop-rule above, can be considered as the last step of the algorithm as opposed to the result of over 40 steps using (4.6). For a more complex scenario now, we take a traffic mix of 10 traffic classes with widely diverse traffic parameters, beginning with compressed VBR audio at 16kbps and H263 video teleconferencing at 64 kbps to MPEG2 HDTV and Full HDTV at 14.5 and 19.4 Mbps. The configuration data is detailed in Table (4.3). Applying (2.5) and fixing the saturation probability at $\gamma = 4$, the results in Fig (4.2) clearly show that

Table 4.3: Traffic mix consisting 10 traffic classes (bandwidth given in Mbps)

	Cl 1	Cl 2	Cl 3	Cl 4	Cl 5	Cl 6	Cl 7	Cl 8	Cl 9	Cl 10
n_i	200	100	100	50	50	30	20	10	10	10
m_i	0.008	0.0256	0.064	0.7	1	2	3.5	5	7	8.5
p_i	0.016	0.064	0.256	1.2	2.5	5	8	12	14.5	19.4

the novel fixed-point algorithm in (4.20) provide a fairly accurate optimum point even on the third step. On the other hand (4.6) do not even provide a stable result. Now investigating the equivalent capacity formula (4.2) with the

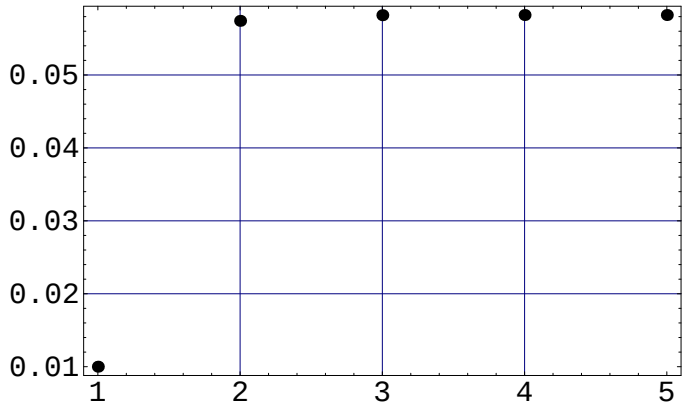

Figure 4.2: Steps of novel fixed-point algorithm on traffic mix of 10 classes

given effective bandwidth ($\alpha(s)$) i.e. the scaled logarithmic moment generating function:

$$C_{\text{equ}} = \inf_s \alpha(s) + \frac{\gamma}{s} \,, \tag{4.21}$$

the resulting fixed-point equation to be solved becomes:

$$s_2 = \sqrt{\frac{\gamma}{\frac{\partial_s \alpha(s)}{\partial s} \mid_{s=s_1}}} \,. \tag{4.22}$$

Making the gaussian distributed traffic substitution as discussed above the

respective equation system is:

$$\alpha(s) = m + \frac{s\sigma^2}{2} \tag{4.23}$$

and

$$\frac{\partial_s \alpha(s)}{\partial s} = \frac{\sigma^2}{2} \ . \tag{4.24}$$

The resulting mean and variance in this case become:

$$m = \alpha^{\text{Gauss}}(s) - s\frac{\partial_s \alpha^{\text{Gauss}}(s)}{\partial s} \tag{4.25}$$

and

$$\sigma^2 = 2\frac{\partial_s \alpha^{\text{Gauss}}(s)}{\partial s} \ . \tag{4.26}$$

Finally, using the effective bandwidth of the gaussian distributed traffic

$$C_{\text{equ}} = \inf_s \alpha(s) + \frac{\gamma}{s} = \inf_s m + \frac{s\sigma^2}{2} + \frac{\gamma}{s} \tag{4.27}$$

and the formula for the optimum parameter s

$$s_{\text{opt}} = \sqrt{\frac{2\gamma}{\sigma^2}} \ , \tag{4.28}$$

we obtain the new fixed-point equation

$$s_{n+1} = \sqrt{\frac{\gamma}{\frac{\partial_s \alpha(s)}{\partial s} \mid_{s=s_n}}} \quad , \tag{4.29}$$

which interestingly turns out to be the same as in the original algorithm. In this case, our advanced method does not yield a different result.

4.1.2 Fixed-point Equations for the Computation of Saturation Probability

We saw that for the formulae targeting the saturation probability of the sum of independent random variables, as presented in Sections 3.1, 3.2 and 3.3, there exists closed-form solution in several cases depending on the PGF approximation used. Also, even, when no closed-form solution can be found appropriate actions can be taken to reach reasonable estimations of it. Nevertheless it is not pointless to investigate numerical evaluation methods for the formulae for two main reasons. Firstly, with numerically computed exact results the accuracy in several occasions can be significantly improved. Secondly, if the

cumulant generating function or the effective bandwidth of the aggregate traffic is directly established/measured (instead of approximating from a few given traffic parameters) numerical evaluation, in general, is obviously necessary. In this section, I briefly discuss fixed-point equations for the computation of the saturation probability for both cases of given $\Lambda(s)$ and $\alpha(s)$.

Considering the first case we have $\Lambda(s)$, and the solvable equation for $P_{\text{sat}} = e^{-\gamma}$ is:

$$\gamma = \inf_s \Lambda(s) - sC. \tag{4.30}$$

It is easy to see that in this case fixed-point equation cannot be established as previously:

$$0 = C - \frac{\partial_s \Lambda(s)}{\partial s} \ . \tag{4.31}$$

On the other hand using the new method proposed in the previous section we obtain

$$s_{\text{opt}} = \frac{C - m}{\sigma^2} \ . \tag{4.32}$$

and using (4.10) and (4.11) we get the resulting recursive equation:

$$s_{n+1} = \frac{{s_n}^2 (C + \frac{\partial_s \Lambda(s)}{\partial s} \mid_{s=s_n}) - 2 s_n \Lambda(s_n)}{2 s_n \frac{\partial_s \Lambda(s)}{\partial s} \mid_{s=s_n} - 2\Lambda(s_n)} \ . \tag{4.33}$$

Following the same considerations, using $\alpha(s)$ in the computation of the saturation probability as

$$\gamma = \inf_s s(\alpha(s) - C) \tag{4.34}$$

we obtain the following fixed-point equation:

$$s_{n+1} = \frac{C - \alpha(s_n)}{\frac{\partial_s \alpha(s)}{\partial s} \mid_{s=s_n}} \ . \tag{4.35}$$

in this case of the result for the gaussian distributed traffic substituted version, using (4.25) and (4.26) would be:

$$s_2 = \frac{C - \alpha(s_1) + s \frac{\partial_s \alpha(s)}{\partial s} \mid_{s=s_1}}{2 \frac{\partial_s \alpha(s)}{\partial s} \mid_{s=s_1}} \ . \tag{4.36}$$

For a summary Table 4.4 provides the obtained fixed-point equations for the different cases discussed in the current chapter. The two columns are for the parameter, which the formula is targeted (that is γ or C_{equ}), and the four rows stand for the type of function (i.e. $\Lambda(s)$ or $\alpha(s)$) the formula uses. $\Lambda_{\text{Gauss}}(s)$ and $\alpha_{\text{Gauss}}(s)$ indicate the gaussian distributed traffic substitution method.

Table 4.4: Summary of the fixed-point equations for γ and C_{equ}

	γ	C_{equ}
$\Lambda(s)$	$0 = C - \frac{\partial_s \Lambda(s)}{\partial s}$	$s_{n+1} = \frac{\Lambda(s_n)+\gamma}{\frac{\partial_s \Lambda(s)}{\partial s}\|_{s=s_n}}$
$\Lambda_{\text{Gauss}}(s)$	$s_{n+1} = \frac{{s_n}^2(C+\frac{\partial_s \Lambda(s)}{\partial s}\|_{s=s_n})-2s_n\Lambda(s_n)}{2s_n \frac{\partial_s \Lambda(s)}{\partial s}\|_{s=s_n}-2\Lambda(s)}$	$s_{n+1} = \sqrt{\frac{\gamma s_n^2}{-\Lambda(s_n)+s_n \frac{\partial_s \Lambda(s)}{\partial s}\|_{s=s_n}}}$
$\alpha(s)$	$s_{n+1} = \frac{C-\alpha(s_n)}{\frac{\partial_s \alpha(s)}{\partial s}\|_{s=s_n}}$	$s_{n+1} = \sqrt{\frac{\gamma}{\frac{\partial_s \alpha(s)}{\partial s}\|_{s=s_n}}}$
$\alpha_{\text{Gauss}}(s)$	$s_{n+1} = \frac{C-\alpha(s_n)+s\frac{\partial_s \alpha(s)}{\partial s}\|_{s=s_n}}{2\frac{\partial_s \alpha(s)}{\partial s}\|_{s=s_n}}$	$s_{n+1} = \sqrt{\frac{\gamma}{\frac{\partial_s \alpha(s)}{\partial s}\|_{s=s_n}}}$

4.2 Numerical Evaluation of PGF, P_{sat}/WLR and C_{equ} Formulae

4.2.1 Comparison of PGF Approximations

In this subsection the PGF approximations presented in (2.3), (2.5) and (2.16) are compared.

First some important characteristics of the formulae mentioned in Section 2.2 are summarized. It is shown in Section 2.3 that the well-known properties of probability generating functions:

$$G_X(s)|_{s=0} = 1 \quad \text{and} \quad \frac{\partial}{\partial s} G_X(s)|_{s=0} = M \ ,$$

which give information on the behaviour of $G_X(s)$ around 0 are true also for all the three approximations. Another interesting observation is that $G_{X,\text{hoe}}(s)$ is exactly the PGF of a Gaussian random variable with mean M and variance $(b-a)^2/4$. Similarly, $G_{X,\text{so}}(s)$ is the probability generating function of sum of on-off random variables ($Y_{\text{ON/OFF}}$) . Although in *Corollary 2* for an intermediate PGF bound a similar property is true, $G_{X,\text{ih}}(s)$ is unlikely to be an exact PGF of any distribution manifesting in closed form (the attempts imposed on symbolic inverse transformation were not successful). It is worth emphasizing again that all three approximations use the same amount of pieces of information (M, n and $p_1, p_2, \ldots, p_n$).

In what follows some numerical examples are presented. For this purpose a simple two-class on-off traffic mix has been defined. The number of sources within the classes are represented by n_1 and n_2, respectively. The mean arrival rate and the peak rate of a source within a class are assumed to be identical

Table 4.5: Traffic Scenarios

	n_1	m_1	p_1	n_2	m_2	p_2
Mix 1	100	0.051	0.064	10	2	5
Mix 2	100	0.051	0.064	1000	0.0048	0.0058

and denoted by m_i, p_i, $i = \{1, 2\}$. Two traffic scenarios are considered in the paper for illustrating the numerical investigations.

The first traffic mix (Mix 1) resembles the aggregation of uncompressed voice ($p_1 = 64$ kbit/s, $m_1 = 51$ kbit/s) and compressed video ($p_2 = 5$ Mbit/s, $m_2 = 2$ Mbit/s) flows with $n_1 = 100$, $n_2 = 10$, while the second one (Mix 2) represents uncompressed and compressed voice traffic ($p_1 = 64$ kbit/s, $m_1 = 51$ kbit/s, $p_2 = 5.8$ kbit/s, $m_2 = 4.8$ kbit/s) with $n_1 = 100$, $n_2 = 1000$. Table 4.5 summarizes the parameters.

In Fig. 4.3 and Fig. 4.4 relative errors of the cumulant generating function approximations to the exact CGF are presented. The curves are drawn by continuous, dotted and dash-dotted lines for Hoeffding, Improved Hoeffding and stochastic ordering based bounds, respectively.

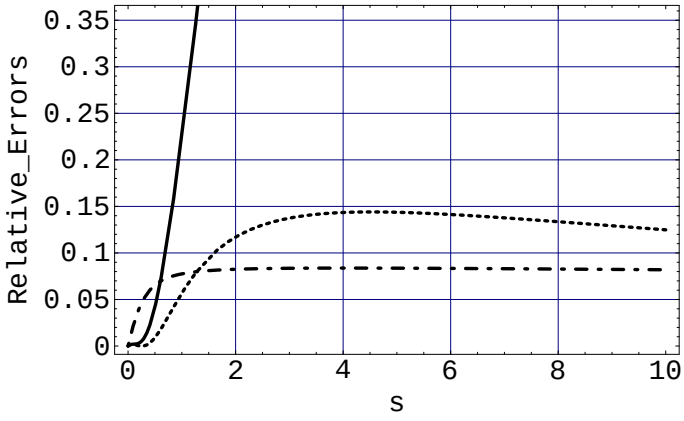

Figure 4.3: CGF Comparisons for Mix 1

Based on extensive numerical analysis the following observations have been made which are partly illustrated by Fig. 4.3 and Fig. 4.4.

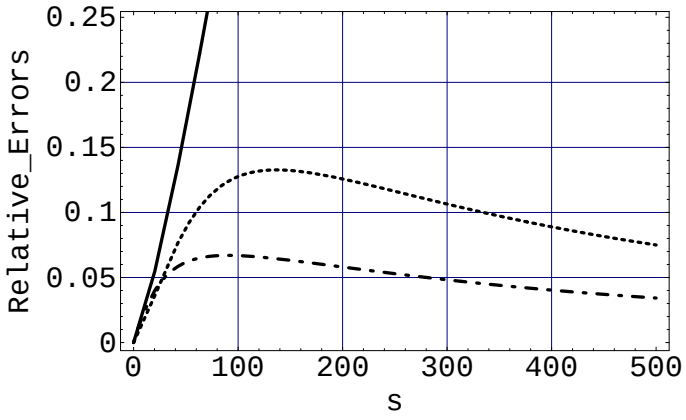

Figure 4.4: CGF Comparisons for Mix 2

For the investigated set of traffic parameters, none of the bounds have been unanimously better than the others in broad range of parameter s. $\Lambda_{X,\mathrm{hoe}}(s)$ often has bad performance, especially for moderate and large s. Moreover, the relative error of $\Lambda_{X,\mathrm{hoe}}(s)$ is quickly (in a quadratic way) increasing with increasing s. When the differences between the peak rates are big (about or more than two order of magnitude) and the ratio of the average aggregate arrival intensity and the sum of the peak rates is also high (greater than 0.6), $\Lambda_{X,\mathrm{ih}}(s)$ usually over-performs $\Lambda_{X,\mathrm{so}}(s)$, especially for small and moderate s. As opposed to this, when the above-mentioned mean to peak ratio is in lower range (smaller than 0.6) and/or the differences between the peak rates is not so high (less than two order of magnitude), $\Lambda_{X,\mathrm{so}}(s)$ has significantly smaller relative errors than the others.

4.3 Performance of Saturation Probability and Workload-loss Ratio Estimations

Resource assessment techniques proposed in the dissertation essentially are based on the PGF approximations presented in Chapter 2. As such, the performance of the derived P_{sat} and WLR formulae are strongly depend on them, the relative exactness of the results usually follows that of the used approximations. Nevertheless, it is worth to take a closer look at the actual performance of the proposed assessment techniques with respect to the achievable

statistical multiplexing gain, when applied in QoS traffic management. The provided investigation also shed light to the relation of the approximations to the theoretically achievable results. For the purpose of the process, we invoke the Measurement-based Call Admission Control algorithm presented in Section 3.4. Briefly recalling the appropriate admission rule, we accept a new connection request if

$$T(s^*, C - p_{n+1}) \leq e^{-\gamma}, \tag{4.37}$$

where $T(s^*, C)$ stands for the saturation probability achieved by each of the investigated Chernoff-type approximations, using C as the link capacity. p_{n+1} is the peak rate of the newcomer connection and s^* is the optimizing parameter.

I have made extensive numerical investigations for traffic mixes both with identical and different peak rates. In these two cases, the typical results from a wide range of traffic types are presented as illustrations in two separate tables. In the first case, the comparison of the different formulae is based on the number of admissible 'users' generating identical traffic flows, while in the second case the statistical multiplexing gain, in bandwidth demand, to the quite conservative peak rate based bandwidth reservation (that is when the sum of the peak rates of the admitted flows cannot exceed the link capacity) is analyzed. In both approaches, the theoretically obtainable results are also presented, which in the first case were generated by general discretely distributed sources,while in the second assuming heterogeneous On/Off fluid sources, by means of probability generating functions. The traffic situations in the tables are characterized by their most important parameters, which have main influence on the results. For all cases, the link capacity is assumed to be 155 Mbit/s (OC-3 link) and the acceptable saturation probability is chosen to $\epsilon = 10^{-6} \Rightarrow \gamma \approx 14$.

In Table 4.6 and 4.7 - five different traffic situations are shown. In the first one, the mean m_i and the peak p_i rates of the identical sources are indicated. Also the traffic intensity – described by $m_i/p_i = M/P$ – is shown for easy reference. These intensity values are chosen from the interval $0.01 - 0.2$, with the quite stringent condition on the acceptable saturation probability. In the next four rows the number of acceptable flows based on the (closed-form) bounds (3.20), (3.14), (3.37) and (3.46) respectively, can be seen. Within parentheses, the ratio of the number of admissible sources using the corresponding bounds and that of the admissible ones by using the conservative peak rate based

reservation technique ($n = \lfloor 155/p_i \rfloor$) is noted. In the next two rows, the calculated maximum number of acceptable sources and the values computed by the peak rate based reservation technique are given. After that, the ratio of the best result among the approximations ((3.20), (3.14), (3.37),(3.46) and the theoretical result is presented.

In our first two traffic examples, the peaks are the same (10 Mbit/s), while the mean rates are 1 Mbit/s and 2 Mbit/s. It can be seen, that in latter case, compared to the former one, the admissible number of sources relative to that of the peak rate reservation technique is significantly smaller. This shows the sensitivity of the multiplexing gain to the traffic intensity under the stringent condition $\epsilon = 10^{-6}$. The same can be experienced in the third and fourth scenario, however in these cases, the relatively large peak rates with respect to the available link capacity results in low achievable multiplexing gain. The low efficiency is also the result of the peak rate-based admission control rule in the forth case. In the fifth traffic situation, a strong effect of statistical multiplexing can be realized due to the small peak rates and mean to peak ratio. Considering the overall performance of the bounds, we can put on, that bounds (3.14) and (3.46) seem to give the best results, however, (3.37) also show good outcomes. The bound (3.20) appeared to be the worst, (in the third and fourth scenario even the worst case scenario overperforms it), probably due to the inappropriate approximation used for the suboptimal parameter s. The best closed form approximations show a quite good performance in the fifth case, with acceptable accuracy.

Table 4.6: The main parameter sets of five different traffic situations with equal peak rates

	1.	2.	3.	4.	5.
m_i	1	2	3	0.3	0.25
p_i	10	10	30	30	5
$\frac{M}{P}$	0.1	0.2	0.1	0.01	0.05
Bound (3.20)	22 (1.47)	18 (1.12)	3 (0.60)	3 (0.80)	92 (2.97)
Bound (3.14)	31 (2.07)	20 (1.33)	5 (1.00)	8 (1.60)	201 (6.48)
Bound (3.37)	30 (2.00)	20 (1.33)	5 (1.00)	8 (1.60)	173 (5.58)
Bound (3.46)	31 (2.07)	20 (1.33)	5 (1.00)	8 (1.60)	201 (6.48)
Best (theoretic)	41 (2.27)	23 (1.53)	5 (1.00)	29 (5.80)	235 (7.58)
Worst (peak rate based)	15	15	5	5	31
Accuracy	0.76	0.83	1.0	0.24	0.856

Table 4.7 illustrates the performance of the bounds for traffic sources with different peak rates. In the first six rows, the most important properties of

the five chosen traffic examples are shown: the aggregate mean arrival rate M, the sum of the peak rates P, the mean to peak ratio $\frac{M}{P}$, the number of flows n and a normalized deviation-like index D_P (for illustrating the variation of the peak rates) defined by

$$D_P = \frac{1}{P}\sqrt{\frac{1}{n}\sum_{k=1}^{n}\left(p_k - \frac{1}{n}P\right)^2}. \tag{4.38}$$

Finally, we also noted the maximum value among the peak rates of the flows in the different cases, which is a necessary input parameter for (3.14). In the next rows, the equivalent bandwidth values computed by formulae (3.20), (3.14), (3.37) and (3.46) are presented and the ratio of the aggregate peak rate P and the computed equivalent bandwidth (i.e. the statistical multiplexing gain) is shown beside them in parentheses. The first of the following two rows indicates the (theoretically) best equivalent bandwidth value (i.e. the minimum one), while the second one shows an accuracy index, which is the ratio of the bandwidth requirement of the best performing closed form bound and the theoretical result.

Table 4.7: The main parameter sets of five different traffic situations with different peak rates

	6.	7.	8.	9.	10.
M	50	60	90	110	200
P	200	200	300	300	500
$\frac{M}{P}$	0.25	0.3	0.3	0.37	0.4
n	50	50	100	100	200
D_P	0.014	0.017	0.020	0.022	0.024
p_{max}	7	9	10	12	12
Bound (3.20)	141.89 (1.41)	175.33 (1.14)	280.87 (1.07)	284.71 (1.05)	485.20 (1.03)
Bound (3.14)	95.12 (2.10)	130.76 (1.53)	214.07 (1.40)	242.24 (1.24)	435.70 (1.15)
Bound (3.37)	110.68 (1.81)	139.03 (1.31)	201.08 (1.49)	225.56 (1.33)	430.70 (1.16)
Bound (3.46)	101.72 (1.97)	134.34 (1.49)	206.55 (1.45)	236.90 (1.27)	400.96 (1.25)
Best (theoretic)	78.74 (2.54)	105.82 (1.89)	174.42 (1.72)	209.00 (1.44)	360.45 (1.39)
Accuracy	0.83	0.81	0.87	0.89	0.90

From the presented results in Table 4.7, we can deduce that as parameter D_P increases, all the bounds get less accurate, however, (3.20) always seems to be the worst one among them. Bound (3.14) performs quite well, while diversity among the peak rates of the sources is not so high, later on as D_P increases, it is overperformed significantly by the (3.37) and (3.46). Bound (3.37) is best at moderate D_P, while (3.46) is the best at high D_P values.

Now let us consider the accuracy indices performed by the different bounds, that is the ratio of the best result among closed form bounds (3.20), (3.14), (3.37), (3.46) and the theoretically achievable bandwidth requirement values. In other words, the accuracy shows the 'performance degradation' caused by using our (best performing but still inaccurate) bounding technique, instead of requiring full characterization of the aggregate traffic and computing the exact values. From the results, we can put on that both in Table 4.6 and Table 4.7, almost all traffic examples, the accuracy is very high at points even reach 0.9 (i.e. the 'utilization loss' of using our techniques compared to the optimal one is around 10-20%), which is remarkable in the light of that the bounds are in closed form, and more importantly that it is enough to know (or measure) the aggregate instantaneous mean and individual peak arrival rates and the number of sources to obtain the results.

To analyze the workability and usefulness of the novel WLR bounds derived, several simulations are carried out by the NS-2 [3] network simulator program. The free object oriented source code of this software made it possible to extend the simulator with extra components, like new MBAC enabled links, without too much difficulties. In the following I will illustrate the results first showing the performance of WLR bound (3.18) in a NS-2 based simulation scenario, and later on direct equivalent bandwidth formulae targeting WLR will also be discussed in Section 4.3.1

The proposed simulation scenario

In the simulations, I have placed great emphasis on using abstract network models, which resemble to a realistic environment and yet are simple and practical. Taking this into consideration, I chose a simulation layout containing two routers connected with a simplex link. This can be regarded for instance as a small segment of a heavily loaded backbone in a large IP network. Since only a small part of the network was examined and the endpoints were not included, the traffic flows were generated explicitly. The incoming flows, according to the CAC algorithm described above, are to provide their peak rates to the MBAC module, which is implemented in the routers. Besides this information, the algorithm also needs the mean rate of the aggregated flow.

To obtain this parameter, a measurement-based estimation method, called *exponential averaging*, is used. This algorithm considers not only present, but also former states of the link. The main idea of this method is to measure

the actual link utilization (b_i) in every S sampling period, then the average bandwidth (AB) usage is calculated with the exponential weighted moving averaging procedure

$$AB = (1 - \omega)AB + \omega b_i \qquad 0 < \omega < 1. \tag{4.39}$$

The algorithm can be fine tuned by ω weight parameter. The higher the ω is set, the more accurately the averaging process will follow the actual bandwidth fluctuations. However, with small ω, past dominates. In the simulations, I have chosen ω, according to the results of [21], to 0.0065.

In the simulation, QoS sensitive stream traffic was sent to the link, hereby putting the CAC algorithm in a more realistic environment. Such traffic types require strict guarantees in terms of delay and delay variation (jitter). These quality requirements can hardly, if at all, be fulfilled in an overloaded network thus the use of call admission control is advantageous. The simulation uses audio and also video streams.

The modeled traffic generators apply the widely used lower bitrate version of G.723.1 [18], G.729 [19] speech and MPEG-4 [1] and H.263 [20] video encoders. Speech traffic is described by practical models, and the video traffic is generated by trace files.

In the case of speech traffic, the packet sizes and inter-arrival times were set according to the specifications. The corresponding parameters are summarized in Table 4.8. Notice that the model also considers silence compression, and the traffic consists of speech and silence sections. During each period constant bit rate (CBR) traffic is generated by the sources with varying packet sizes. The distribution of the holding times of the alternating states are exponential, the mean values of which are set so that 40 percent of the time is adjusted to represent active speech.

Table 4.8: Parameters used in speech traffic modelling

Type of coding	Packet size (byte)	Packet interarrival time (ms)
G.723.1 MP-MLQ	60	30
G.723.1 Silence	44	30
G.729 CS-ACELP	60	20
G.729 Silence	44	20

Regarding the video traffic, I have gained different flows from a single trace file making the simulator occasionally start reading from random locations.

The applied trace files recorded by Fitzek and Reisslein [2] are three different quality instances of the same movie "Silence of the Lambs". These samples are to represent good quality, high bit rate MPEG-4, lower bit rate MPEG-4, and 64 kbit/s bit rate H.263 traffic types.

In setting the bandwidth of the link, two important factors are taken into account. Firstly, high bandwidth channel is necessary in order to admit a high number of flows, so that the aggregate fluctuation is expected to be relatively small. Secondly, low bandwidth is required to reduce the need of calculations and make the simulation run faster. Considering these factors, the selected link capacities resulted in 5 Mbit/s for audio and 100, 50, 25 Mbit/s for high, low and 64 kbit/s bit rate video, respectively.

During the simulations, constant admission decisions planned to be made to admit or reject newcomer flows. The assigned lifetime of the flows is chosen to be independent, exponentially distributed with an average of 5 minutes for speech and 10 minutes for video. The inter-arrival times are modeled with independent, exponentially distributed random variables, whose average values are adjusted to set the bandwidth demand to at least 30 percent higher than the link capacity. The selected values are 0.67 seconds for G.723.1, 0.83 seconds for G.729, and 3, 1, 1.33 seconds for high, low and 64 kbit/s bit rate video traffic, respectively.

Regarding the traffic quality requirements, the strict delay variance of the speech traffic is typically 20 ms. If the buffer size of the routers along a path is selected so that the maximum queuing delay is less than or equal to 1 ms, then it can be guaranteed that the low delay variance criterion is met even if the packet goes through 20 subsequent routers, which considered to be a realistic assumption in the current public Internet.

Simulation results

The simulations ran for an hour and the data collection started after a half an hour. The delay was necessary to eliminate the effects of transient states, which in the beginning could have falsified the results.

Two types of simulations were carried out. The first was performed to evaluate the resource assessment technique and verify whether or not the system works properly. In this case, we tested the resource assessment techniques on a fully loaded link without admission control. The goal was to compare the estimated WLR with the measured one. The measurements provide the

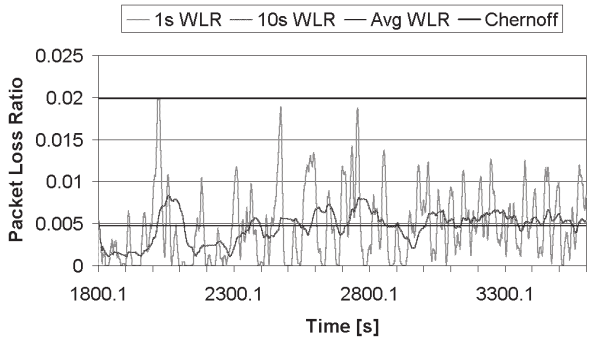

Figure 4.5: Resource assessment using G.729 speech traffic

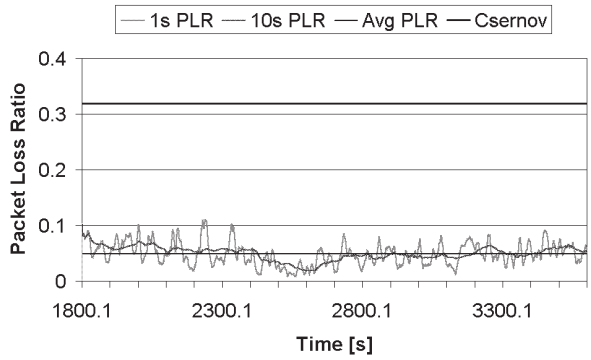

Figure 4.6: Number of admitted flows using G.729 speech traffic

average WLR of the whole simulation, for a one second and a 10 second sliding window. The last two scenarios represent the average WLR for the last 1 and 10 seconds.

The simulative measurements were run with all the mentioned traffic types. Considering the results in Figs. 4.5-4.6, we find that the proposed approximation formula gave us in each case a reliable, strict upper bound for the WLR. It bounds even the much more fluctuating peaks of the 1 second sliding window measurements. The main use of such upper bounds is to provide strict QoS guarantees even in extra low time scales. Comparing the speech and video traffic estimation scenarios, the results of speech traffic proves to be more accurate, which is apparently comes from the stronger fluctuation nature of video streams. Despite this difference, however, in both cases the estimation remained in a generally acceptable range to the actual measured average WLR.

The second type simulations consisted of measurements testing the proposed MBAC algorithm working in scenarios similar to real situations. In these simulations, the average utilization of the link, the average WLR and the number of admitted flows were measured. These values provide us a good view of the efficiency and reliability of the MBAC algorithm. The guaranteed WLR parameter of the algorithm was set to 0.01 suggesting that approximately every hundredth packet might be lost. This value would seem a bit high at first sight and it is surely too much for a reliable data network, however, for general real-time stream traffic it can be often acceptable. Fig. 4.7 shows the degree of utilization for G.729 speech traffic during the simulation. Notice that in the diagram the utilization of the link is quite high. The speech and silence packets differ only slightly in size (60 and 44 bytes) which results in low fluctuation of the aggregate traffic. On the other hand, the utilization at video traffic scenarios in Fig. 4.8 is significantly lower, which comes from the huge oscillation of high speed MPEG-4 traffic (notice that the peak rates are approximately ten times the average). The admitted flow numbers are plotted in Figs. 4.9-4.10. In all the presented simulations, the measured average PLRs (0.0097 for G.729, and 0.0083 for high bit rate MPEG-4) were lower than the targeted 0.01.

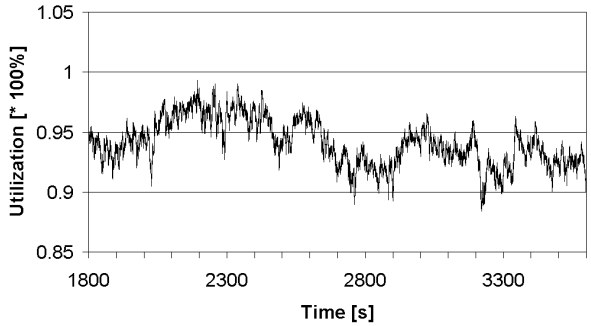

Figure 4.7: Link utilization using G.729 speech traffic

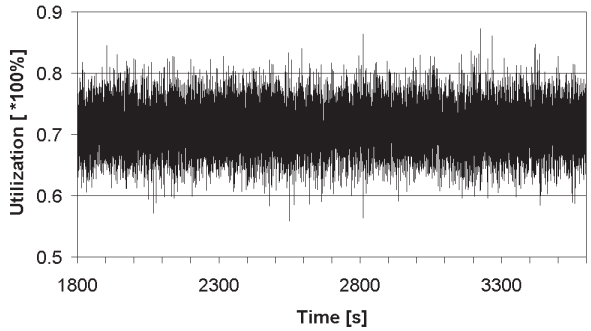

Figure 4.8: Link utilization using G.729 speech traffic

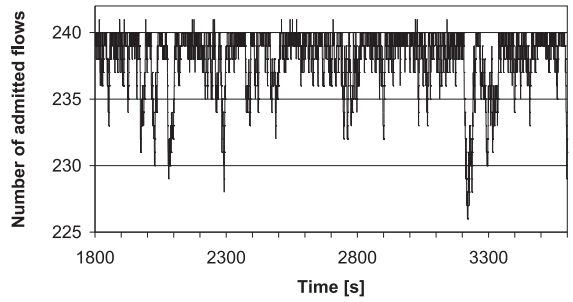

Figure 4.9: Number of admitted flows using G.729 speech traffic

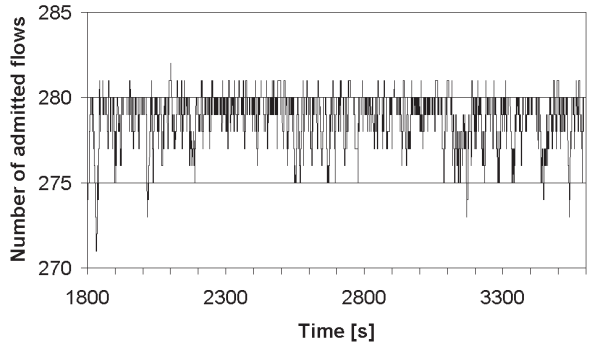

Figure 4.10: Number of admitted flows using G.729 speech traffic

4.3.1 Comparison of Bandwidth Requirement Estimates

In this subsection the performance of the equivalent capacity estimates derived in the dissertation are presented through observations based on extensive analysis and illustrative numerical examples.

The numerical analysis is made in the following way: First, for a given C we compute the exact saturation probability $P(X > C)$ and workload loss ratio $E[(X - C)^+]/E[X]$ where X is the rate distribution in Mix 1 and Mix 2. Then the corresponding γ values are determined from $P(X > C) = e^{-\gamma}$ and $E[(X-C)^+]/E[X] = e^{-\gamma}$ which are substituted together with the above-mentioned CGF bounds based on (2.3), (2.5), (2.16) into the formulae of $\widetilde{C}_{\text{equ,wlr}}$, $\widetilde{C}_{\text{equ,wlr}}$, respectively. In this manner, there could be a direct relation between the exact equivalent capacity (C) and its three approximations. Note that in these comparisons optimized (and thus fixed) values of s have been used.

First let us consider the conservative upper bounds among the equivalent capacity estimators (3.81), (3.82). Embedding the CGF approximation $\widetilde{\Lambda}_{X,\text{hoe}}(s)$ usually results in poor estimates, the accuracy in these cases is not acceptable because the relative error is usually higher than 0.1. It is illustrated with the two traffic mixes for the case of $\widetilde{C}_{\text{equ,sat}}$ in Fig 4.11 and Fig 4.12.

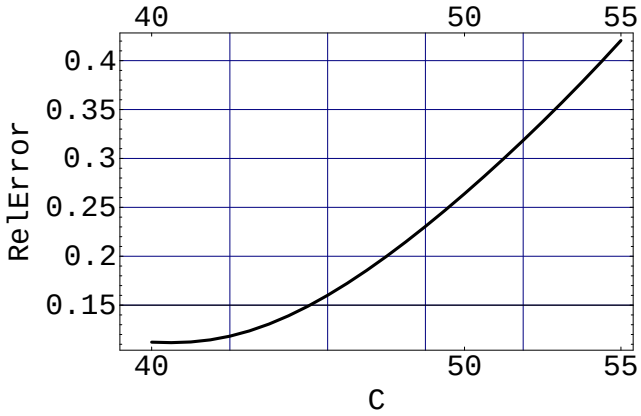

Figure 4.11: $\widetilde{C}_{\text{equ,sat}}$, $\widetilde{\Lambda}_{X,\text{hoe}}(s)$, Mix 1

Applying the improved Hoeffding and stochastic ordering based CGF approximations much better performance can be observed both in case of $\widetilde{C}_{\text{equ,wlr}}$

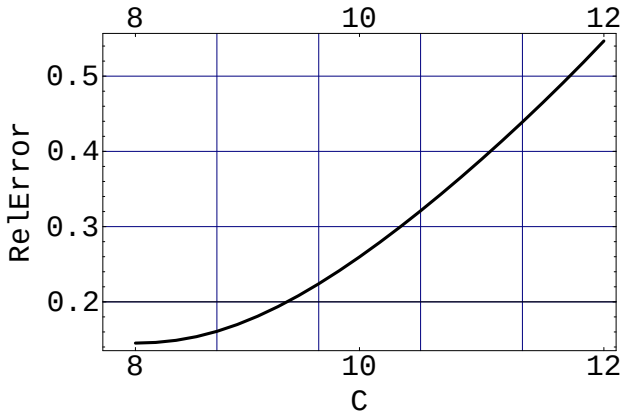

Figure 4.12: $\widetilde{C}_{\mathrm{equ,sat}}$, $\widetilde{\Lambda}_{X,\mathrm{hoe}}(s)$, Mix 2

and $\widetilde{C}_{\mathrm{equ,sat}}$. In Fig 4.13 and Fig 4.14 the relative errors $(\widetilde{C}_{\mathrm{equ,wlr}} - C)/C$ are drawn for the two traffic mixes, respectively. The dotted and dash-dotted curves correspond to the cases of Improved Hoeffding (2.5), and stochastic ordering based (2.16) CGF bound substitutions, respectively.

It can be seen that for both traffic mixes the accuracy of the stochastic ordering based estimate is acceptable, while the improved Hoeffding-based estimator has higher inaccuracy, especially for Mix 2. Another important observation (which is supported by many other examples besides the figures presented) is that both equivalent capacity estimators has a certain value of γ (or corresponding exact equivalent capacity value) from which the relative error starts to increase rapidly when γ is increasing, however, this phase transition starts with considerably higher value of γ (more stringent QoS constraint) in case of the use of $\widetilde{\Lambda}_{X,\mathrm{so}}(s)$ than in case of the use of $\widetilde{\Lambda}_{X,\mathrm{ih}}(s)$.

Continuing the discussion on the refined approximations of the equivalent capacities one can ask why conservative bounds of CGF have been used in the (not necessarily conservative) approximation of the equivalent capacities. Besides the fact that tangible conservative bounds are available for CGF, our analysis has also revealed that the approximations (3.86), (3.93) are likely to approximate the exact value of C_{equ} from above (but it can not always be guaranteed) when embedding suitable CGF bounds in the formulae. On the other hand, such estimates of C_{equ} are still closer to the exact value than the equivalent capacity bounds, due to the underlying refined approximation

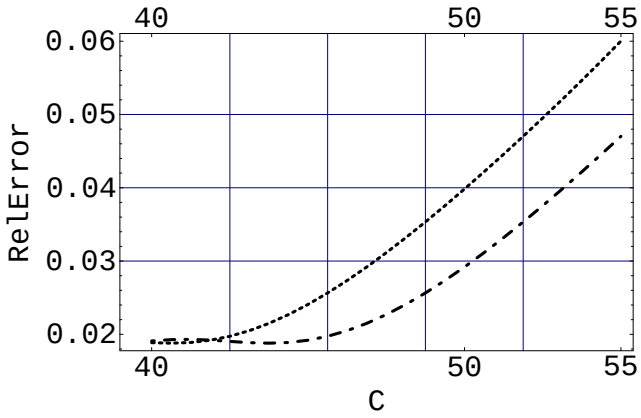

Figure 4.13: Comparison of $C_{\text{equ,wlr}}$ estimates with $\widetilde{\Lambda}_{X,\text{ih}}(s)$ and$\widetilde{\Lambda}_{X,\text{so}}(s)$ Mix 1

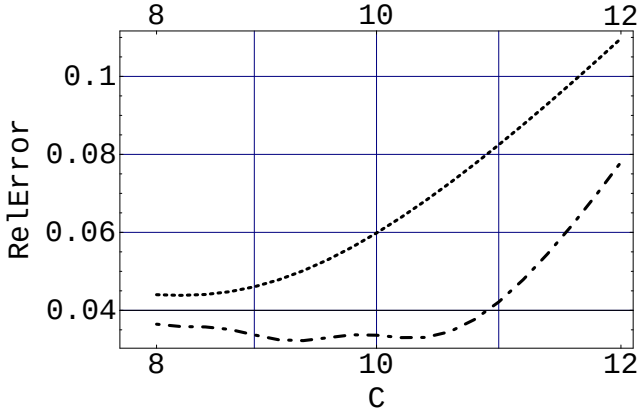

Figure 4.14: Comparison of $C_{\text{equ,wlr}}$ estimates with $\widetilde{\Lambda}_{X,\text{ih}}(s)$ and$\widetilde{\Lambda}_{X,\text{so}}(s)$ Mix 1

technique.

In Fig 4.15 and Fig 4.16 the relative errors $(\widetilde{C}^{\mathrm{B-R}}_{\mathrm{equ,sat}} - C)/C$ are drawn for the two traffic mixes, respectively. The continuous, dotted and dash-dotted curves correspond to the cases of $\widetilde{\Lambda}_{X,\mathrm{hoe}}$, $\widetilde{\Lambda}_{X,\mathrm{ih}}$ and $\widetilde{\Lambda}_{X,\mathrm{so}}$ CGF bound substitutions, respectively. Note that in both cases the $\widetilde{\Lambda}_{X,\mathrm{hoe}}$-based approximations massively under-estimates the exact values. In case of traffic Mix 2 the $\widetilde{\Lambda}_{X,\mathrm{so}}$-based approximation partly under-estimates $C_{\mathrm{equ,sat}}$.

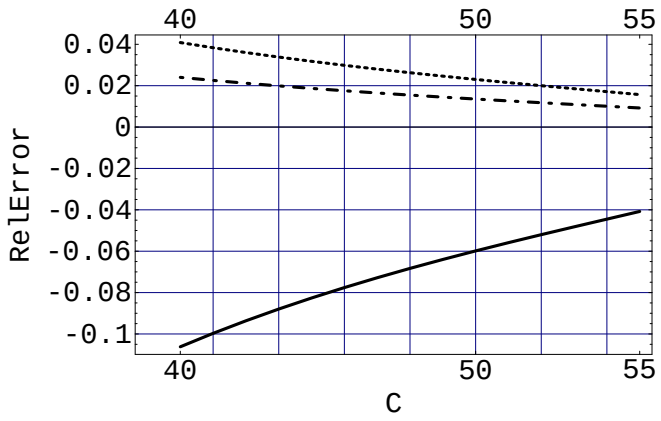

Figure 4.15: $\widetilde{C}^{\mathrm{B-R}}_{\mathrm{equ,sat}}$ comparison for Mix 1

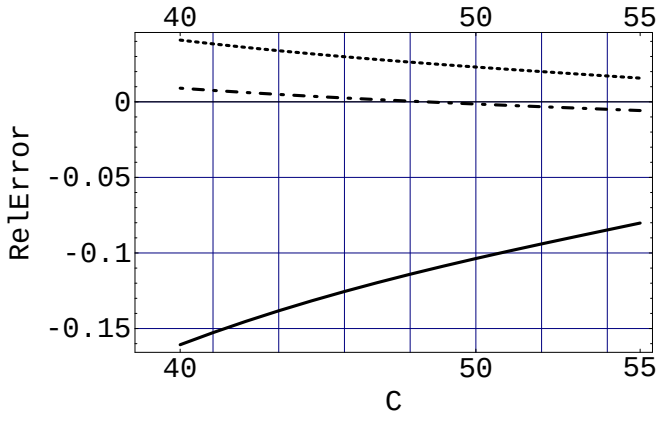

Figure 4.16: $\widetilde{C}^{\mathrm{B-R}}_{\mathrm{equ,sat}}$ comparison for Mix 2

In Fig 4.17 and Fig 4.18 the relative errors $(\widetilde{C}^{\mathrm{B-R}}_{\mathrm{equ,wlr}} - C)/C$ are drawn for the two traffic mixes, respectively. The continuous, dotted and dash-dotted curves correspond to the cases of Hoeffding (2.3), Improved Hoeffding (2.5), and stochastic ordering based (2.16) CGF bound substitutions, respectively.

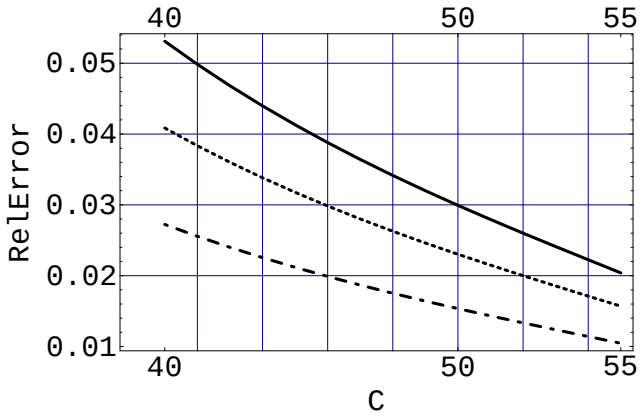

Figure 4.17: $\widetilde{C}^{\mathrm{B-R}}_{\mathrm{equ,wlr}}$ comparison for Mix 1

These figures also illustrate our common and general observations on the refined equivalent capacity approximations based on extensive numerical investigations, as follows:

- The Hoeffding-based bandwidth requirement estimator provide the worst performance (often with unacceptable high inaccuracy) in almost all cases, except some practically insignificant traffic situations. The relatively good performance in Fig 4.17 and Fig 4.18 is an exception.
- The relative error of the stochastic ordering based estimator is usually the smallest one, especially for that traffic mixes in which the mean to peak ratio is high and the differences between the peak rates are more than one order of magnitude.
- It can also be observed that the differences between the relative errors is higher for smaller absolute value of the exact equivalent capacity (i.e. smaller value of γ), as opposed to the case of conservative upper bounds, where the differences between the relative errors are smaller for smaller absolute value of the exact equivalent capacities.

- The refined approximations (3.86), (3.93) are likely to approximate the exact equivalent capacity value from above when embedding the improved Hoeffding and stochastic ordering-based CGF bounds in the formulae.
- The absolute values of the relative errors of the refined approximations (the defined relative error in the refined approximation case can also be negative) are always decreasing with increasing values of γ (with more stringent QoS contraints), as opposed to the case of conservative bounds.
- The refined approximations are tighter for higher values of γ (more stringent QoS constraints), while the conservative bounds are tighter for smaller values of γ (less stringent QoS constraints).

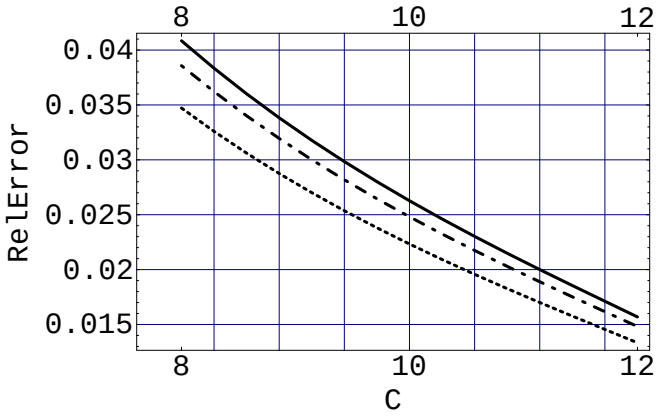

Figure 4.18: $\widetilde{C}^{\mathrm{B-R}}_{\mathrm{equ,wlr}}$ comparison for Mix 2

We also observed that when the QoS constraint is very stringent (very large γ) the estimators could fail to give reasonable values, mainly because of numerical instability caused by the extremely bad characteristics (in terms of optimization) of functions to be minimized.

Chapter 5

Conclusion

Resource assessment techniques play central role in modern high-speed communications networks. They provide the bases for many network and traffic management tasks, from planning and dimensioning, to micro-flow traffic engineering and also network performance evaluation. In real applications the importance of *a priori* characterization of traffic is shifted to the advantage of measurement-based algorithms and traffic monitoring. On the scale of decision time, as time shortens, the set of observable statistical properties of traffic flows gets simplified. In the dissertation I have presented resource assessment formulae, that uses simple traffic characterization. Such techniques in the literature are widely used in research works constructing eg. measurement-based admission control algorithms. The performance of the presented bounds and approximations are thoroughly examined by means of extensive numerical and simulative analysis.

5.1 Research Contributions

The contributions of the dissertation are divided into three main parts. In Chapter 2 parsimonious approximations have been developed for moment-generating functions of sums of non-negative valued random variables. The provided formulae can be used to approximate probability generating functions (PGF) from above using simple characterization of the distributions, that is the formulae exploit only the aggregate mean and individual maximum values of the random variables. The main asset of the results lies in the fact that the variables are not assumed to have identical mean and peak arrival rates. The presented bound even show optimal behavior under reasonable conditions.

In Chapter 3 conservative upper bounds and approximations are presented for resource assessment. The bounds are established by the well-known Chernoff bounding method, and use the previously derived PGF approximations. The resource meters refer to two type of transmission properties: the saturation probability and the workload loss (eg. packet loss) ratio. As the directly derived (Chernoff-type) bounds for the most part are in implicit form, the requirement for fast calculation necessitates to transform them into a closed form. Although, this transformation yields suboptimal results, extensive numerical analysis shows that the bounds are still sharp enough to overperform, previous estimations. Besides the provided conservative bounds, also well-performing approximations are presented for the cases when guarantee is not necessary, but more accurate estimations are needed. Finally, in this part, for many previously discussed cases closed form equivalent capacity formulae are shown, that are claimed to fit more into traffic engineering tasks.

As a final contribution, fixed-point equations are presented in Chapter 4, for the fast computation of some of the previously derived bounds that could not been transformed into a closed form in any reasonable way. The resulting algorithms are shown to be converge quite fast, in most of the cases only 3-4 steps proved to be enough to find optimum with an error of less then 0.001%.

5.2 Future Research Directions

Parsimonious resource assessment techniques are widely used in the area of modern network and traffic management. As such, it is still an extensively researched area, containing a large number of unsolved issues. The simplified assumptions that have been used in the dissertation to establish resource assessment bounds raise quite a few questions to be solved. Also the performance of many of the established results are only examined numerically. Whereas these investigations unambiguously reflect the most important performance issues of the formulae, general statements with few exceptions (like in the case where optimality can be proved) are cannot be made. Similarly in the case of the fixed-point equations in Chapter 4, general stability analysis of the proposed algorithms was not made, however, a large number of numeric examples presume it to be usable in practical cases.

In the case of the approximation of probability generating functions of sums of random variables, the random variables are assumed to be independent.

For homogenous variables there exist some approach that allow some type of dependence (like in Section 3.1.1), however, there is no known method that can be applied directly to the inhomogenous case (and showing any type of optimal property).

The presented formulae for the approximation of the saturation probability and workload loss ratio are all derived by the Chernoff bounding method. Obviously this method does not provide optimal results even, when the underlying PGF approximations possesses some type of optimality property. Although this deficiency in practical cases (eg. for a few hundred sources) for the most part is not significant (as it was clearly seen previously in the numerical investigations) theoretically the issue can still be interesting.

With respect to the traffic measurements made at the characterization phase, interesting question is, as to whether the aggregate mean arrival rate is the most advantageous property that should be monitored. In several approaches (concerning buffered systems) the effective bandwidth or the cumulant generating function of the arrival rate distribution is measured, and the resource assessment is done respectively.

Acknowledgement

I would like to express my gratitude to all those who gave me the possibility to complete this thesis. First I want to thank the Department of Telecommunication and Media Informatics of the Budapest University of Technology and Economics, particulary the HSNLab as one of its research laboratory for giving me the possibility and constant support of doing the necessary research work. I am bound to my Ph.D. supervisor Dr. Tamás Henk the leader of HSNLab for giving me stimulating support, and for inexhaustibly encouraging me to go ahead with my thesis.

I am deeply indebted to my scientific supervisor Dr. József Bíró whose great help, generous ideas, stimulating suggestions and encouragement helped me in all the time of research for and writing of this thesis.

My colleagues from the Department of Telecommunication and Media Informatics supported me in my research work. I want to thank them for all their help, support, interest and valuable hints. Especially I am obliged to János Zátonyi, Mátyás Martinecz and András Gulyás.

This is a prominent opportunity to express my thank to my former teachers. Among them, I am particulary grateful to my math teacher in secondary school Csaba Farkas whose vocation of teaching mathematics gave me considerable impetus to go on with my scientific interest.

Most importantly, none of this would have been possible without the love and patience of my family, especially my mother and father. They have been a constant support of love, concern and strength all these years. I would like to give my special thanks to my wife Eszter whose patient love enabled me to complete this work.

Bibliography

[1] *Moving Picture Experts Group*, http://mpeg.telecomitalialab.com/.

[2] *MPEG-4 and H.263 trace files*, http://www-tkn.ee.tu-berlin.de/ fitzek/TRACE/ltvt.html.

[3] *The Network Simulator - ns2*, http://www.isi.edu/nsnam/ns/index.html.

[4] H. Alnowibet K., Perros, *Nonstationary analysis of circuit-switched communication networks*, Performance Evaluation Journal **63** (2006), 892–909.

[5] R. R. Bahadur and R. Rao, *On deviations of the sample mean*, Ann. Math. Statis **31** (1960), no. 27, 1015–1027.

[6] L. Breslau, S. Jamin, and S. Shenker, *Comments on the performance of measurement-based admission control algorithms*, Proceedings of the Conference on Computer Communications (IEEE Infocom) (Tel Aviv, Israel), March 2000.

[7] F. Brichet and A. Simonian, *Conservative gaussian models applied to measurement-based admission control*, International Workshop on Quality of Service (IWQoS) (Napa, USA), May 1998.

[8] H. Chernoff, *A measure of asymptotic efficiency of tests of a hypothesis based on the sum of observations*, Annals of Mathematical Statistics **23** (1952), 493–507.

[9] C. Courcoubetis, V. A. Siris, and G. D. Stamoulis, *Application of the many sources asymptotic and effective bandwidths to traffic engineering*, Telecommunication Systems **12** (1999), 167–191.

[10] C. Courcoubetis and R. Weber, *Buffer overflow asymptotics for a buffer handling many traffic sources*, Journal of Applied Probability **33** (1996), 886–903.

[11] H. Cramer (ed.), *Mathematical methods of statistics*, Princeton University Press, 1945.

[12] A. Faragó, *Blocking probability estimation for general traffic under incomplete information*, IEEE International Conference on Communications (New Orleans, LA, USA), vol. 3, June 2000, pp. 1547–1551.

[13] S. Floyd, *Comments on measurement-based admissions control for controlled-load services*, submitted to IEEE Computer Communication Review, July 1996.

[14] R. J. Gibbens and F. P. Kelly, *Measurement-based connection admission control*, International Teletraffic Congress (ITC) (Washington, D.C., USA), June 1997, pp. 879–888.

[15] P. Hitczenko and Montgomery-Smith S., *Measuring the magnitude of sums of independent random variables*, Annals of Probability **29** (2001), no. 1, 447–466.

[16] W. Hoeffding, *Probability inequalities for sums of bounded random variables*, Journal of the American Statistical Association **58** (1963), 13–30.

[17] J. Y. Hui, *Resource allocation for broadband networks*, IEEE Journal on Selected Areas in Communications **6** (1988), no. 9, 1598–1608.

[18] ITU-T G.723.1, *Dual rate speech coder for multimedia communications transmitting at 5.3 and 6.3 kbit/s*, recommendation, March 1996.

[19] ITU-T G.729, *Coding of speech at 8 kbit/s using conjugate-structure algebraic-code-excited linear prediction (cs-acelp)*, recommendation, March 1996.

[20] ITU-T H.263, *Video coding for low bit rate communication*, recommendation, 1998.

[21] S. Jamin, S. J. Shenker, and P. B. Danzig, *Comparison of measurement-based admission control algorithms for controlled-load service*, Proceedings of the Conference on Computer Communications (IEEE Infocom) (Kobe, Japan), April 1997, p. 973.

[22] F. P. Kelly, *Effective bandwidths at multi-class queues*, Queueing Systems **9** (1991), 5–16.

[23] F. P. Kelly, *Notes on effective bandwidths*, Stochastic Networks: Theory and Applications **4** (Oxford University Press, 1996), 141–168.

[24] F. P. Kelly, *Charging and accounting for bursty connections*, url: citeseer.ist.psu.edu/kelly96charging.html, 1996.

[25] N. Likhanov and R. R. Mazumdar, *Cell loss asymptotics in buffers fed with a large number of independent stationary sources*, Proceedings of the Conference on Computer Communications (IEEE Infocom) (San Francisco, USA), March/April 1998.

[26] Guoqiang Mao and Daryoush Habibi, *Loss performance analysis for heterogeneous on-off sources with application to connection admission control*, IEEE/ACM Transactions on Networking (TON) **10** (2002), no. 1, 125–138.

[27] M. Montgomery and G. de Veciana, *On the relevance of time scales in performance oriented traffic characterizations*, Proceedings of the Conference on Computer Communications (IEEE Infocom) (San Francisco, USA), vol. 2, March 1996, pp. 513–520.

[28] I. Pinelis, *Optimal tail comparison based on comparison of moments*, High Dimensional Probability **43** (1998), 297–314.

[29] J. Roberts, *Engineering for quality of service*, Draft chapter from the book Self-similar Netwotk Traffic and Performance Evaluation edited by K. Park and W. Willinger., 1998.

[30] J.P. Schmidt, A. Siegel, and A. Srinivasan, *Chernoff-hoeffding bounds for applications with limited independence*, ACM-SIAM Symposium on Discrete Algorithms, 1995.

[31] G. Seres, Á. Szlávik, J. Zátonyi, and J. Bíró, *Quantifying resource usage - a large deviation-based approach*, IEICE Transactions on Communications (Japan) **E85-B** (2002), no. 1, 938–952.

[32] S. G. Shekhovtsov, O. I.; Gorokhov, *Data transmission over nonstationary communication channels*, Tech. report, Izdatel'stvo Leningradskogo Universiteta, 1985.

[33] M. Talagrand, *The missing factor in hoeffding's inequalities*, Annales de l'institut Henri Poincaré (B) Probabilités et Statistiques **31** (1995), no. 4, 689–702.

[34] Z. Turányi, A. Veres, and A. Oláh, *A family of measurement-based admission control algorithms*, PICS'98 (Lund, Sweden), May 1998.

[35] M. Villen-Altamirano and M. F. Sanchez-Canabate, *Effective bandwidth dependent of the actual traffic mix: An approach for bufferless CAC*, International Teletraffic Congress (ITC) (Washington, D.C., USA), November 1997, ITC'15.

[36] A. Weiss, *An introduction to large deviations for communication networks*, Selected Areas in Communications, IEEE Journal on **13** (1995), 938–952.

Publications

Journal papers

[J1] A. Kern, M. Martinecz, Z. Heszberger and J. Bíró. Architecture and Configuration of Broadband Access Networks Supporting Multimedia Applications. *IJCA Journal Special Issue.* pages 34–46, April, 2007.

[J2] Z. Heszberger. Novel Equivalent Capacity Approximation Through Asymptotic Loss Analysis. *Accepted to Computer Communications.* Special Issue on ATM & IP Networks: Performance Modelling and Analysis

[J3] M. Martinecz, J. Bíró and Z. Heszberger. Novel Resource-reservation algorithms in Packet-switched Networks. *Híradástechnika - Telecommunications - Hungarian periodical, Selected Papers.* pages 24–29, June, 2005.

[J4] M. Martinecz, J. Bíró and Z. Heszberger. Újszerû erôforrásigény-becslô módszerek csomagkapcsolt hálózatokban. *Híradástechnika - Telecommunications - Hungarian periodical.* pages 13–18, Sept, 2004.

[J5] Z. Heszberger, J. Zátonyi, and J. Bíró. Performance bounds for rate envelope multiplexing. *Performance Evaluation, Volume 48, Issue 1, May 2002, Pages 87-101.* Special Issue on ATM & IP Networks: Performance Modelling and Analysis

[J6] Z. Heszberger, J. Zátonyi, and J. Bíró. Efficient Chernoff-based resource assessment techniques in multi-service networks. *Telecommunication Systems, Volume 20, Issue 1,2, 2002, Pages 59-80.* Special Issue: Wide Area Networks Design and Analysis

[J7] Z. Heszberger and J. Bíró. An Optimization Neural Network Model with Time-Dependent and Lossy Dynamics. *Neurocomputing, Volume 48, Issues 1-4, October 2002, Pages 53-62.*

[J8] Z. Heszberger, J. Zátonyi, and J. Bíró. Efficient resource usage techniques in multiservice networks. *Periodica Polytechnica Ser. Electrical Engineering, Volume 44, Issue 1, 2000, Pages 91-102.*

[J9] Z. Heszberger. Számlázás ATM-hálózatokban. *Magyar Távközlés - Hungarian Telecommunications Periodical*, Volume 8, pages 44–47, July, 1997.

[J10] Z. Heszberger. Pricing ATM services. *Magyar Távközlés - Hungarian Telecommunications Periodical, Selected Papers II.*, Pages 59–62, 1997.

Conference papers

[C1] J. Bíró, A. Gulyás, Z. Heszberger. A Novel Probabilistic Extension of Network Calculus for Workload Loss Examinations. In Proc. of *The Second Conference on Next Generation Internet Network - EuroNGI 2006*, pages 152–161, April, 2006.

[C2] A. Gulyás, J. Bíró, Z. Heszberger. A Probabilistic Network Calculus for Characterizing Long-run Network Behavior. In Proc. of *IEEE International Conference on Communications - ICC 2006*, Vol. 1., pp. 465–470, June 2006.

[C3] A. Kern, M. Martinecz and Z. Heszberger. Architecture and Configuration of Broadband Access Networks. In Proc. of *ISCC'2005, The Tenth*

IEEE Symposium on Computers and Communications, volume 1, pages 172–181, La Manga del Mar Menor, Cartagena, Spain, June, 2005.

[C4] A. Gulyás, J. Bíró, Z. Heszberger and T. Szénási. Dependency Criteria on Regulated Inputs for Buffer Overflow Approximation. In Proc. of *IEEE International Conference on Communications - ICC 2005*, volume 1, pages 83–87, Seoul, Korea, May, 2005.

[C5] J. Bíró, A. Gulyás and Z. Heszberger. A Novel Direct Upper Approximation for Workload Loss Ratio in General Buffered Systems. *Lecture Notes in Computer Science - LNCS 3042 (Proc of IFIP Networking 2005)*, pages 718–730, Waterloo, Ontario, Canada, May, 2005.

[C6] J.J. Bíró, M. Martinecz, Sz. M. Kis and Z. Heszberger. Novel Equivalent Capacity Approximation Through Asymptotic Loss Analysis. In Proc. of *Second International Conference on the Performance Modelling and Evaluation of Heterogeneous Networks*, pages P5/1-P5/9, Ilkley, West Yorkshire, U.K., July, 2004.

[C7] J.J. Bíró and Z. Heszberger. An Optimization Neural Network Model with Lossy Dynamics and Time-Varying Activation Functions. In Proc. of *IEEE International Joint Conference on Neural Networks 2004 - IJCNN'04*, pages 2245–2249, Budapest, Hungary, July, 2004.

[C8] J.J. Bíró and Z. Heszberger. Analog Neural Networks as Asymptotically Exact Dynamic Solvers. In Proc. of *IEEE International Joint Conference on Neural Networks 2004 - IJCNN'04*, pages 2267–2272, Budapest, Hungary, July, 2004.

[C9] J. Bíró, Z. Heszberger, A. Gulyás and T. Szénási. Distribution-Free Conservative Bounds for QoS Measures. In Proc. of *IEEE International Symposium on Computers and Communications - ISCC'2004*, pages

112–119, Alexandria, Egypt, July, 2004.

[C10] J. Bíró, Z. Heszberger and M. Martinecz. A Family of Performance Bounds for QoS Measures in Packet-Based Networks. *Lecture Notes in Computer Science - LNCS 3042 (Proc of IFIP Networking 2004)*, pages 1108–1119, Athens, Greece, May, 2004.

[C11] J. Bíró, Z. Heszberger, M. Martinecz, N. Felician and Octavian Pop. Towards a Framework of QoS Measure Estimates for Packet-Based Networks. In Proc. of *IEEE International Conference on Communications - ICC 2004*, volume 4, pages 2231–2235, Paris, France, June, 2004.

[C12] J. Bíró, Z. Heszberger, F. Németh, M. Martinecz. Bandwidth Requirement Estimators for QoS Guaranteed Packet Networks. In Proc. of *International Network Optimization Conference - INOC 2003*, pages 41–49, Evry-Paris, France, May, 2003.

[C13] J. Bíró, Z. Heszberger, and M. Martinecz. Equivalent Capacity Estimators for Bufferless Fluid Flow Multiplexing. In Proc. of *IEEE Global Telecommunications Conference - Globecom 2003*, volume 7, pages 3706–3710, San Francisco, CA, USA, Dec. 2003.

[C14] J. Bíró, Z. Heszberger, N. Felician, and M. Martinecz. Bandwidth Requirement Estimators for QoS Guaranteed Packet Networks. In *INOC 2003, International Network Optimization Conference*, volume 1, pages 95–100, Evry/Paris, France, October, 2003.

[C15] J. Bíró, Z. Heszberger, T. Dreilinger, A. Gulyás, and M. Martinecz. Parsimonious Estimates of Bandwidth Requirement in Quality of Service Packet Networks. In Proc. of *First International Working Conference on Performance Modeling and Evaluation of Heterogenous Networks - HET-NETs '03*, pages 69/1-69/9, Ilkley, UK, 21-23 July 2003.

[C16] J. Bíró, Z. Heszberger, G. Kún and M. Martinecz. Advanced QoS Provision for Real-Time Internet Traffic. In *IEEE International Packet Video Workshop - Packet Video 2003*, pages 26–28, Nantes, France, April, 2003.

[C17] J. Zátonyi, Z. Heszberger, and J. Bíró. Packet Loss Based QoS Provision for Real-Time Internet Traffic. In Proc. of *International Symposium on Performance Evaluation of Computer and Telecommunication Systems - SPECTS 2002*, pages 65–72, San Diego, California, July 2002.

[C18] J. Zátonyi, Z. Heszberger, and J. Bíró. Chernoff-based resource assessment techniques in communication networks. In Proc. of *Polish-Czech-Hungarian Workshop 2001 on Circuit Theory, Signal Processing, and Telecommunication Networks*, pages 56–66, Budapest, Hungary, Sept. 2001.

[C19] J. Zátonyi, Z. Heszberger, and J. Bíró. Traffic management research studies supporting QoS network evolution. In Proc. of *9th IFIP Workshop on Performance Modelling and Evaluation of ATM & IP Networks*, pages 245–251, Budapest, Hungary, June, 2001.

[C20] Z. Heszberger, J. Zátonyi, J. Bíró, and T. Henk. Efficient bounds for bufferless statistical multiplexing. In Proc. of *IEEE Global Telecommunications Conference - GLOBECOM 2000*, volume 1, pages 641–646, San Francisco, CA, USA, Nov./Dec. 2000.

[C21] Z. Heszberger, J. Zátonyi, and J. Bíró. Performance bounds for rate envelope multiplexing. In Proc. of *8th IFIP Workshop on Performance Modelling and Evaluation of ATM & IP Networks*, pages 9/1–9/10, Ilkley, England, July, 2000.

[C22] Z. Heszberger, J. Zátonyi, and J. Bíró. Efficient CAC algorithms based on the tail distribution of aggregate traffic. In Proc. of *2nd Conference of PhD Students on Computer Sciences*, page 47, Szeged, Hungary, July

2000.

[C23] Z. Heszberger, J. Bíró. Neural Networks for Global Optimization. In Proc. of *2nd Conference of PhD Students on Computer Sciences*, ext. vol., Szeged, Hungary, July 2000.

[C24] Z. Heszberger, J. Zátonyi, and J. Bíró. Optimization techniques for tail distribution estimation based on the Chernoff bounding method. In Proc. of *17th European Conference on Operational Research - EURO XVII*, page 155, Budapest, Hungary, June, 2000.

[C25] Z. Heszberger, J. Bíró, and T. Henk. Neural Networks for Global Optimization. In Proc. of *17th European Conference on Operational Research - EURO XVII*, page 93, Budapest, Hungary, June 2000.

[C26] Z. Heszberger, J. Bíró, and E. Halász. An Optimization Neural Network Model with Time-Dependent and Lossy Dynamics. In Proc. of *ESANN, 10th European Symposium on Artificial Neural Networks*, pages 287–292, Bruges, Belgium, April 2000.

[C27] Z. Heszberger, J. Bíró, and T. Henk. Comparison of Simple Tail Distribution Estimators. In Proc. of *ICC '99, IEEE International Conference on Communications*, volume 3, pages 1841–1845, Vancouver, CA, June 1999.