



Budapest University of Technology and Economics

Resource Management Problems in Packet Switched Networks

Szabolcs Malomsoky
High Speed Networks Laboratory
Department of Telecommunications and Telematics

Ph. D. Dissertation

Supervisor:
Dr. Edit Halász
High Speed Networks Laboratory
Department of Telecommunications and Telematics

Budapest, Hungary
2003



Budapesti Műszaki és Gazdaságtudományi Egyetem

Erőforrás menedzsment feladatok csomagkapcsolt hálózatokban

Malomsoky Szabolcs
Távközlési és Telematikai Tanszék
Nagysebességű Hálózatok Laboratórium

Ph. D. értekezés

Tudományos vezető
Dr. Halász Edit
Távközlési és Telematikai Tanszék
Nagysebességű Hálózatok Laboratórium

Budapest
2003

Contents

| | | |
|----------|--|-----------|
| 1 | Introduction | 1 |
| 2 | Connection admission control in UTRAN | 4 |
| 2.1 | Introduction | 4 |
| 2.2 | System model | 5 |
| 2.2.1 | UMTS network architecture | 6 |
| 2.2.2 | Traffic model, traffic descriptors | 7 |
| 2.2.3 | QoS requirements | 8 |
| 2.2.4 | ATM/AAL2 transport | 8 |
| 2.2.5 | IP transport | 9 |
| 2.3 | Queueing model | 10 |
| 2.3.1 | Decomposition of burst-level and packet-level effects | 11 |
| 2.3.2 | The proposed model | 13 |
| 2.3.3 | Validation of the proposed model | 14 |
| 2.4 | Connection admission control algorithm | 16 |
| 2.4.1 | Checking delay violation in the non-overloaded system | 17 |
| 2.4.2 | Checking delay violation due to temporary system overload | 18 |
| 2.4.3 | Flow-chart of the algorithm | 21 |
| 2.5 | Validation of the hyper-plane approximation | 21 |
| 2.5.1 | Same TTIs and “always ON” sources | 23 |
| 2.5.2 | Same TTIs but different activity factors | 24 |
| 2.5.3 | Effect of having different TTIs | 26 |
| 2.6 | Numerical examples | 27 |
| 2.6.1 | Comparison of different approximations | 27 |
| 2.6.2 | Example admissible regions | 29 |
| 2.7 | Conclusion | 30 |
| 3 | Mobility and traffic analysis for WCDMA networks | 32 |
| 3.1 | Introduction | 32 |
| 3.1.1 | Overview of the literature | 34 |
| 3.1.2 | Our work | 35 |
| 3.2 | Description of the model | 35 |

| | | |
|----------|---|-----------|
| 3.2.1 | Road systems model and traffic flow estimation | 36 |
| 3.2.2 | Notations and assumptions | 38 |
| 3.2.3 | Soft handover region parameters | 40 |
| 3.2.4 | User-plane traffic estimation on the Iur interface . . . | 47 |
| 3.2.5 | Cell parameters | 48 |
| 3.2.6 | A numerical example | 51 |
| 3.3 | Conclusion | 53 |
| 4 | Real-time VP bandwidth control | 55 |
| 4.1 | Introduction | 55 |
| 4.2 | VP bandwidth control | 56 |
| 4.2.1 | Estimation framework | 58 |
| 4.2.2 | Convergence of the control | 62 |
| 4.2.3 | Practical implications | 64 |
| 4.3 | Simulations | 66 |
| 4.3.1 | Short-range dependent traffic | 66 |
| 4.3.2 | Long-range dependent traffic | 67 |
| 4.3.3 | Actual ATM traffic | 69 |
| 4.4 | Implementation issues | 70 |
| 4.5 | Conclusion | 71 |
| 5 | Summary of the Dissertation | 73 |
| 5.1 | Connection admission control in UTRAN | 73 |
| 5.2 | Mobility and traffic analysis for WCDMA networks | 74 |
| 5.3 | Real-time VP bandwidth control | 74 |
| A | Derivation of the approximation of $Q(x)$ used in Section 2.5.2 | 76 |
| B | Incremental assignment method for Section 3.2.1 | 78 |
| C | State space representation with two-point measurement for Section 4.2.1 | 79 |

A bírálatok és a tézisfüzet a BME Villamosmérnöki és Informatikai Kar Dékáni Hivatalában megtekinthetők.

Acknowledgments

I express my deepest gratitude to all people, who contributed to this thesis.

First of all I wish to thank Dr. Edit Halász, my supervisor and Dr. Tamás Henk, the head of HSN Lab for their kind guidance. Without the financing and encouraging support of Dr. Miklós Boda, the head of Ericsson R&D in Budapest and Hans Eriksson, the head of Traffic Lab, this thesis could have not been completed.

My research work started six years ago at the High Speed Networks Laboratory at the Dept. of Telecommunications and Telematics, Budapest University of Technology and Economics. Special thanks to all members and PhD students of the department and in particular to the members of HSN Lab.

A fruitful and exciting half year was spent at the Multimedia Networks Laboratories, Nippon Telegraph and Telephone Corp., Tokyo, Japan. I would like to acknowledge Dr. Hiroshi Saito for his supervision during this time.

Since 1998 I have been working at the Traffic Analysis and Network Performance Laboratory of Ericsson. I first met here with András Valkó. His professional approach impressed me. I especially enjoyed working with my colleagues András Rác, Árpád Szilávik, István Szabó, Sándor Rác, Szilveszter Nádas, István Maricza and Pál Zarándy on topics presented in this thesis. I am grateful for their effective collaboration. The discussions and the work with Gábor Fodor also meant a lot to me.

My colleague and friend, Attila Vidács deserves special thanks for all the help he provided during the years. He has also thoroughly read through and commented on the text of this thesis before it reached its final shape.

I am very thankful to my family, who provided me a stable and inspiring background during my studies.

Emi, thank you for your love and warmth.

List of Abbreviations

| | |
|--------------|--|
| AAL | ATM Adaptation Layer |
| ATM | Asynchronous Transfer Mode |
| CAC | Call (or Connection) Admission Control |
| CBR | Constant Bit Rate |
| CDMA | Code Division Multiple Access |
| CID | Connection Identifier |
| CLR | Cell Loss Ratio |
| CN | Core Network |
| DHO | Diversity Handover unit |
| FIFO | First In First Out |
| FDDI | Fiber Distributed Data Interface |
| FTP | File Transfer Protocol |
| GSM | Global System for Mobile |
| HTTP | Hypertext Transfer Protocol |
| IDC | Index of Dispersion of Counts |
| IP | Internet Protocol |
| ITU-T | International Telecommunication Union Telecommunication Standardization |
| Iub | The name of the interface between a base station and a radio network controller in UTRAN |
| LAN | Local Area Network |
| LRD | Long-Range Dependent |
| MAC | Medium Access Control |
| PCR | Peak Cell Rate |
| PSTN | Public Switched Telephone Network |
| QoS | Quality of Service |
| RAB | Radio Access Bearer |
| RLC | Radio Link Control |
| RMD | Resource Management in Diffserv |
| RNC | Radio Network Controller |
| RSVP | Resource reSerVation Protocol |
| SHR | Soft Handover Region |
| SRD | Short-Range Dependent |
| TTI | Transmission Time Interval |

| | |
|--------------|--|
| UBR | Unspecified Bit Rate |
| UE | User Equipment |
| UMTS | Universal Mobile Telecommunications System |
| UTRAN | UMTS Terrestrial Radio Access Network |
| VC | Virtual Channel |
| VCH | Virtual Channel Handler |
| VCI | Virtual Channel Identifier |
| VP | Virtual Path |
| VPI | Virtual Path Identifier |
| WCDMA | Wideband Code Division Multiple Access |

Chapter 1

Introduction

As the development of packet switched network infrastructures advances, managing the resources of these networks becomes inevitable. Resource management in packet switched networks differs substantially from that of traditional telephone networks, because a wide range of applications and services with diverse traffic characteristics and quality of service (QoS) requirements have to be transported. The task of resource management typically includes the solution of a complex optimization problem: the fraction of packets that get lost or suffer from unacceptable delays must be kept below a given threshold, while the amount of used resources should be minimized.

In the 80s and 90s, the ATM (Asynchronous Transfer Mode) technology was developed to cope with the new requirements on resource management. ATM has been equipped with the basic tools of resource management, such as resource allocation and service differentiation, for example [1]. On the other hand, the dramatic growth of the Internet in the late 90s made it clear, that developing a general framework of QoS support and resource management for the Internet can result in a more widely used option of multi-service networking.

Although the original goal of ATM, namely to provide a global, multi-service network, did not prove to be viable, ATM is used in the transport infrastructures of backbone networks. One example is the landline transport network of UMTS [2, 3] (Universal Mobile Telecommunication System). UMTS may be regarded as the future successor of today's GSM (Global System for Mobile): the objective with UMTS is to enhance the capabilities of current cellular systems to be able to deliver high quality multimedia content, to become connected to the Internet, etc. [4, 5]. WCDMA (Wideband Code Division Multiple Access) [6] is the radio interface of UMTS. WCDMA essentially operates with so called radio access bearers (RAB), which are basically packet switched radio connections with dedicated resources. Due to requirements of user mobility and radio interface timing, the major QoS requirements on these connections are fast connection set-up and low packet

delay, respectively. To support the bandwidth efficient transmission of low bit-rate, delay-sensitive applications (typically conversational voice over an ATM network), ITU-T has standardized a new ATM Adaptation Layer, AAL2 [7, 8, 9]. Also fast AAL2 connection set-up has been an important goal of AAL2 protocol design, therefore the ATM/AAL2 infrastructure is a suitable transmission technology for the UMTS Terrestrial Radio Access Network (UTRAN) [10, 11]. The IP based transport infrastructure of UTRAN is currently being specified in 3GPP [12].

The objective of this dissertation is to give answers and solutions to three resource management problems: (1) connection admission control in UTRAN transport networks, (2) mobility and traffic analysis in WCDMA systems, and (3) real-time bandwidth control in ATM backbone networks. These problems are motivated by the following practical tasks:

- (1) Each network manufacturer developing transmission infrastructure for UTRAN must solve the problem of connection admission control for both the ATM/AAL2 and the IP based transport options. Admission control algorithms are used in communication systems to check whether a newly arriving connection (of known traffic demand and QoS requirements) can be served by a shared resource such that the QoS requirements of all connections using that resource can be satisfied.
- (2) To estimate the traffic load on UTRAN transport links (e.g., for dimensioning), a connection-oriented model of the system is needed that takes into account the effects of user mobility and the multiple connection rates. In WCDMA, a large amount of traffic (both user plane traffic and signaling) is generated as a consequence of soft handovers for dedicated connections. Therefore, the connection-level model must handle soft handovers.
- (3) In order to efficiently ensure that strict ATM cell-level QoS requirements are met in an ATM backbone network, the cell loss ratio (CLR) should be monitored on-line, and bandwidth control should be applied to keep the CLR under a pre-defined value.

Outline of the dissertation

Chapter 2 deals with connection admission control (CAC) in UTRAN. We focus on the transport links connecting base stations and radio network controllers (i.e., on the Iub interface), because here packet delay and loss requirements are strict and the amount of transmission resources is relatively low. The CAC algorithm on the Iub interface can be parameter-based (no measurements are done), because the traffic on the Iub interface is shaped such that it can be characterized by a few parameters with sufficient accuracy. The method presented in this Chapter is fast, and it is able to satisfy QoS requirements at high system utilization.

In Chapter 3, a connection-level traffic and mobility model for cellular systems, presented originally in [13] for systems with *hard handover* (for example GSM), is extended for WCDMA systems, which use *soft handover*. Soft handover means that a mobile terminal can communicate with more than one base station at the same time.

To model soft handover, we consider overlapping cells in the model, and allow that mobiles moving in a road system can connect to more base stations in overlapping areas. We assume that cell borders divide the analyzed area into *soft handover regions (SHR)*. Closed form solutions of the connection arrival rate and the residence time in an SHR are given and the relations between these parameters are derived. It is shown how these parameters can be used to estimate inter-RNC (radio network controller) traffic, which is a consequence of soft handover, and would not be present if hard handover were used. Then, the probability distribution of the channel occupancy time in a cell is given. By assigning capacity to each cell, we obtain the blocking probability and the offered traffic load on each cell in each connection class by applying a recursive method.

It has been the primary goal of this work to develop an analytic tool for estimating the traffic load on UTRAN transport links. Particularly, the inter-RNC traffic is difficult to estimate, since it strongly depends on user mobility.

Chapter 4 considers an on-line, measurement based Virtual Path (VP) bandwidth control algorithm. In ATM, traffic flows are typically multiplexed into VPs with deterministic (constant bit rate - CBR) bandwidth allocation. We consider traffic flows, for which admission control is not able to guarantee packet-level QoS (i.e., UBR - unspecified bit rate or in other words “best-effort” traffic flows). The VP bandwidth control algorithm dynamically allocates VP bandwidth according to variations in traffic demand.

Measurements are used, because the traffic pattern is rather complex, and an appropriate parametric model is difficult to identify (if possible at all). Furthermore, statistical analysis of a large number of traffic traces taken from a variety of networking environments revealed that the traffic variations are dominant over a wide range of time scales [14, 15]. These variations can be described using the concepts of long-range dependence (LRD) and self-similarity [16, 17]. The presence of LRD in the traffic has a strong impact on queuing behavior [18, 19, 20].

We determine the packet loss performance using periodic buffer measurements, which are valid irrespectively of the presence of LRD, and by applying a rather simple (and thus tractable!) approximate effective bandwidth formula (see also [21]), we build a recursive VP bandwidth estimation algorithm. We show that under weak conditions the method dynamically converges to the optimal VP bandwidth value, which is required to keep the required QoS.

Chapter 2

Connection admission control in UTRAN

On transport links of UMTS Terrestrial Radio Access Networks, but especially on those connecting base stations and radio network controllers (i.e., on the Iub interface), resource allocation is complex, because packet delay and loss requirements are strict and the amount of transmission resources is relatively low. In this chapter a novel connection admission control (CAC) algorithm is provided, which is applicable on the Iub interface with both ATM/AAL2 and IP transport options. The CAC algorithm is validated by mathematical analysis and computer simulation.

2.1 Introduction

With the introduction of 3rd generation mobile systems, such as the Universal Mobile Telecommunications System (UMTS), both equipment vendors and network operators face new challenges in connection with the network roll-out. In contrast with 2nd generation systems, a packet-switched, multi-service transmission network is designed, which should fulfil the specific requirements of the new radio interface technology (Wideband Code Division Multiple Access; WCDMA). Switching and multiplexing technologies used for the first releases of UTRAN are based on Asynchronous Transfer Mode (ATM) in combination with the ATM Adaptation Layer type 2 (AAL2) [7, 10, 22]. Future releases will be deployed using also IP (Internet Protocol) technologies [12, 23].

While performance and resource management of the WCDMA radio interface is thoroughly discussed in the literature (see e.g., [6]), few work is dedicated to the performance evaluation, traffic control and provisioning of the transmission infrastructure of UTRAN. In particular, a CAC algorithm for the Iub interface, which works in the multi-service scenario and fulfils all practical requirements (e.g., on limited complexity and high precision) has not yet been presented.

Admission control algorithms are used in communications systems to check whether a newly arriving connection (of known traffic demand and QoS requirements) can be served by a shared resource such that the QoS requirements of all connections using that resource can be satisfied. The hardship of designing admission control methods stems from the following general requirements:

- *Admission decisions should be conservative:* the QoS requirements have to be always (or at least with very high probability) satisfied. This requirement is utmost important if the perceived quality is very sensitive to overloads (e.g., if the voice coder is not adaptive, observed voice quality deteriorates rapidly upon overloads), or if the quality is a determining criterion in contracts and these contracts can not be violated.
- *System utilization should be high:* the previous requirement should be met such that the possible system utilization is achieved, and resources are not wasted. In other words, a good algorithm will not reject too much call requests in order to be at the safe side.
- *The decisions have to be taken on-line:* if a new call request arrives, the admission decision has to be taken rapidly to ensure that call setup times remain short.

The motivation of our work is

- to establish an adequate model for the UTRAN transport network (compared to [C4], where simple AAL2 connection admission control methods using the Chernoff-bound are evaluated, the Iub specific traffic behavior is modeled),
- to develop a CAC algorithm applicable in the Iub interface, and
- to validate it using and extending methods from the broad literature of ATM performance evaluation [24, 25, 26].

The chapter is organized as follows. Section 2.2 outlines the system model. In Section 2.3 the related queueing model is presented. In Section 2.4 the proposed CAC algorithm is introduced. A validation of the method is given in Section 2.5.

2.2 System model

In this section, the UMTS architecture is shortly described, traffic modeling considerations and QoS requirements are introduced.

2.2.1 UMTS network architecture

The UMTS network architecture [6] is depicted in Figure 2.1. An UMTS network consists of the user equipment (UE), the UMTS Terrestrial Radio Access Network (UTRAN) and the core network (CN).

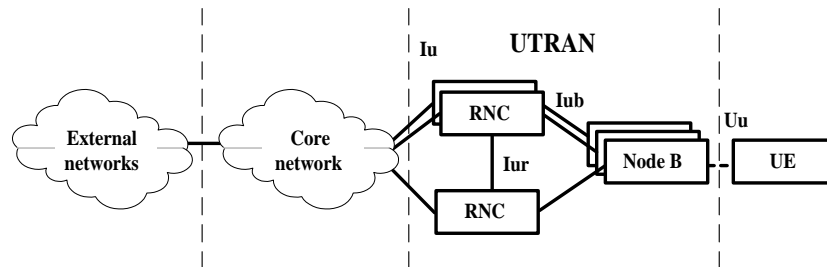


Figure 2.1: UMTS network architecture

UTRAN handles all tasks related to radio access, therefore UTRAN nodes are responsible for radio resource management, handover control, etc. The core network is the backbone of UMTS connecting the access network to external networks (e.g., PSTN, Internet). The mobile (UE) is connected to base stations (Node Bs) over the radio interface (Uu). One mobile can communicate with several base stations at the same time during soft handover (which is an essential interference reduction technique in WCDMA systems [6]). A base station is connected to a radio network controller (RNC) over the Iub interface. RNCs are connected to each other over the Iur interface. The RNC is connected to the core network over the Iu interface.

The three lowest layers of the UTRAN protocol stack are depicted in Figure 2.2. The retransmission mechanism of the radio link control (RLC) protocol ensures reliable transmission of loss-sensitive traffic over the radio interface. The medium access control (MAC) protocol forms radio frames and schedules these periodically according to the timing requirements of WCDMA. This period is called TTI (transmission time interval), and its length can be a multiple of 10 ms. Bit rates of radio connections (so called RABs - radio access bearers) take typical values between 8 kbps and 384 kbps. MAC frame sizes and TTI lengths are RAB-specific. Considering the simplest case, when a user uses a single service, one RLC and one MAC entity are created in the RNC for each actually connected mobile.

On Iub, to decrease the probability of packet congestion in the transmission network queues, the start positions of frames intended for different UEs should not coincide in time. Moreover, on the radio interface it should be avoided to transmit control patterns, such as pilot bits, at the same time for all mobiles, because this would introduce peaks in the interference and,

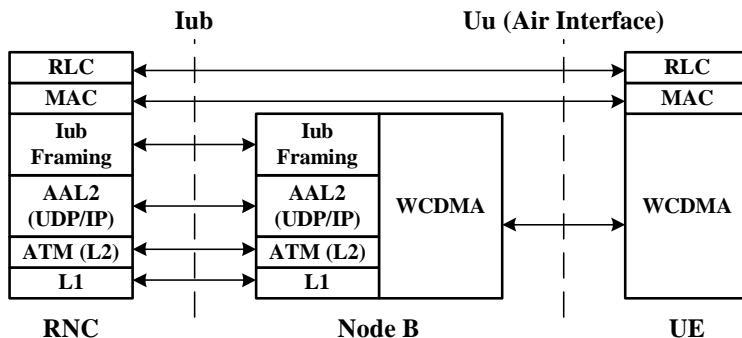


Figure 2.2: UTRAN user-plane protocol stack

thus, might limit the confidence of some measurements. Therefore, phases of the periodic frame flows of different connections are randomly distributed over the TTI [27].

The main role of the transmission network is to transport MAC frames from radio network controllers to base stations (in downlink direction) and to transport MAC frames from base stations to radio network controllers (in uplink direction). In the network, MAC frames are encapsulated into Iub frames. The Iub framing overhead contains information used in base stations to encode the frame into the appropriate radio frame format and to send it out on the radio interface at the right time (t_{out}).

CAC allocates resources for the new connections in the transport network. It makes its decisions based on traffic descriptors and QoS parameters.

2.2.2 Traffic model, traffic descriptors

As we already discussed, the arrival pattern of Iub frames is determined by the MAC scheduler. In other words, the traffic is shaped by the MAC scheduler such that the lowest time-scale behavior is periodic irrespectively of the type of application. The user/application level traffic model is reflected in the UTRAN transport network such that the carried traffic is not a continuous periodic packet flow, but can be modeled by a series of active and inactive intervals. We will refer to these intervals as ON (active) and OFF (inactive) periods. In an ON period MAC frames are sent in each TTI, while in an OFF period, packets are not sent at all. For example, in case of voice traffic the characteristics of the ON and OFF periods are determined by the interaction of the speech process (the speaker behavior) and the voice activity detector in the voice coder.

The traffic descriptors associated with each connection are the following: (Iub) frame size (we will also call it packet size), TTI (the inter-arrival time

of frames), and the so called “activity factor”. In UMTS, the following TTI values are possible: 10, 20, 40 and 80 ms. The activity factor is defined as the average length of ON periods divided by the sum of the average lengths of ON and OFF periods. Note that reliable information on the distributions of the lengths of ON and OFF periods are not available for the CAC.

For example, typical parameters for voice service can be the following: 40 byte long packets arriving in 20 ms periods (TTI) with activity factor 0.6. Note that the activity factor is an effective value, which is set by the operator in order to exploit statistical multiplexing gain.

2.2.3 QoS requirements

In this section it is explained shortly why the packet delay is the most important performance measure in the transport network. We focus on the queueing delay that is the time packets spend waiting in buffers.

If a UE has simultaneous RABs to two or more Node Bs (during soft handover), the radio frames scheduled in downlink have to be sent out from every Node B to the UE at the same time (t_{out}). Therefore, nodes must be synchronized. For the same reason, it has to be ensured that each frame arrives to the Node Bs before t_{out} . This determines a delay requirement on the UTRAN transport network.

For voice traffic (with 20 ms TTI), the queueing delay budget within UTRAN is around 5-7 ms [28]. Queueing delay requirements for other services are *not very* different from that of voice. Since RABs carrying best-effort traffic are also subject to soft handover, delay requirements are defined for them as well. Furthermore, since the round-trip time of RLC packets should be minimized in order to maximize the throughput of best-effort traffic, these delay requirements are strict. (Analytic expressions on the dependence of application-level throughput on the RLC round-trip time can be found in [29].)

The consequence of the strict queueing delay requirements is that short buffers are applied in the system, and therefore only short time-scale traffic fluctuations can be absorbed by buffering. We assume that the queueing delay requirement of a service is typically smaller than (or equal to) the TTI of the RAB carrying the service.

2.2.4 ATM/AAL2 transport

Using this transport option, Iub frames are segmented and packed into AAL2 CPS (Common Part Sublayer) packets, which are multiplexed into ATM cells. AAL2 payload can be of variable length (up to 45 bytes), and the AAL2 header is 3 bytes long. ATM cells are 53 bytes long including a 5 bytes long header. By AAL2 multiplexing, several AAL2 packets from different connections can be carried within an ATM cell. In an ATM network, cells are

transported along a predefined path using the VPI/VCI (Virtual Path and Virtual Channel Identifier) fields in the ATM header. The CID (Connection Identifier) field in the AAL2 header identifies a specific AAL2 connection within an ATM VC.

Fast AAL2 connection set-up has been an important design goal of the AAL2 protocol [10]. In UTRAN, a new AAL2 connection is set up for each new RAB. In network nodes, where AAL2 multiplexing is done, AAL2 CAC allocates resources for the AAL2 connections. Among these nodes, traffic descriptors are sent in AAL2 signaling messages.

The Capability Set 1 (CS1) [7] of the AAL2 signaling protocol does not support service differentiation (or QoS separation) of AAL2 connections. CS2 [8] enables the selection of AAL2 path according to the requested QoS. This way, AAL2 connections having different QoS requirements can be transported over separate ATM VCs.

2.2.5 IP transport

If the IP transport option [12] is applied, QoS is provided by using Differentiated Services (Diffserv) [30], Resource reSerVation Protocol (RSVP) [31] or over-provisioning at IP layer.

Over-provisioning is static and there is no need for admission control. However, it does not take advantage of transport bandwidth efficiency gains that IP can provide. One can combine over-provisioning with admission control. In this case, CAC is done only at the edges of UTRAN (at Node Bs or RNCs), and within the network the resources are provided by over-dimensioning or edge-to-edge resource allocation (for example, using RSVP).

In order to exploit statistical multiplexing gains and provide good QoS, similarly to the ATM/AAL2 solution, distributed admission control is needed. Distributed admission control uses signaling (e.g. RSVP). The admission control function is distributed in the routers and is performed hop-by-hop. RSVP could have scalability problems for large networks if it is used per flow. Therefore, a new QoS framework has been proposed recently, called Resource Management in Differentiated Services (RMD) [32]. It extends the Diffserv architecture with new admission control and resource reservation concepts in a scalable way. RMD applies admission control on resource parameter values included in reservation requests, i. e. signaling messages and available resources per traffic class. The reservation is done in terms of resource units, which may be based on a single parameter, such as bandwidth, or on more sophisticated parameters, such as the traffic descriptors described in this paper. Therefore, assuming for example that UTRAN connections of strict delay requirements exclusively use the Diffserv Expedited Forwarding (EF) class, the admission control algorithm presented in this paper could directly be used for that class.

2.3 Queueing model

According to the system model presented in the previous section, a queueing model is needed that is accurate if the delay requirements are strict (5-15 ms) and the buffer size is small (e.g., smaller than 20 ms). The queueing model of this section is developed for these conditions.

The RAB connections are modeled with independent ON-OFF sources. The ON and OFF periods are bursty, meaning that typically both are many TTI long. The long term correlation characteristics of the arrival process could not be taken into account, because traffic descriptors do not contain any information on the correlation structure of the sources, and it was also not possible to get information on this by measurements. It is not a problem in practice, because the buffer is small enough such that it fills up quickly even during a temporary overload. Therefore we can assume that the ON-OFF burst component of the queue is negligible (see details in Section 2.3.1). The system is temporarily overloaded, if so many connections are temporarily in ON state at the same time, that the server can not serve within one TTI the packets arriving in one TTI. I used the approximation that all packets arriving during a temporary overload situation violate the delay requirement (see Section 2.3.2).

If the server can serve during one TTI all packets that arrive within one TTI (i.e., when the overall arrival rate remains below multiplex capacity), the system behaves like the so called $\sum_i D_i/D/1$ queue: a superposition of independent periodic sources of possibly different periods and independent phases is offered to a multiplexer with a deterministic server. The time between two packet emissions of a class i stream is equal to TTI_i , and its phase with respect to a common time origin is chosen at random between 0 and TTI_i .

We assume that the number of traffic classes (the services with different traffic descriptors) is K . The number of connections present in the system from class i is N_i . Connections within the same class (class i) are characterized by the activity factor α_i , the packet size b_i and the packet inter-arrival time TTI_i . The server capacity is constant and it is denoted by C . Packets are served according to the FIFO (first-in-first-out) scheduling principle. As an example, Figure 2.3 shows a system fed by two connections.

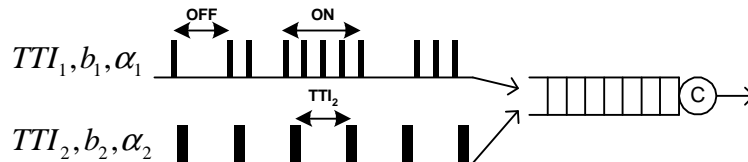


Figure 2.3: A system fed by two periodic ON-OFF connections with different packet inter-arrival times (TTI s) and packet sizes

In this section, a model is provided, which allows us to derive the probability of the packet delay criterion violation: $\Pr(D_i > \tilde{D}_i)$, where D_i is a random variable representing the delay of a packet from class i , and \tilde{D}_i is the delay criterion (or target maximum delay) of packets from traffic class i . The total delay of a packet from class i consists of two parts: $D_i = S_i + Q_i$, where S_i is the service time and Q_i is the waiting time in the queue. Delays of lost packets are considered to be infinite. In the proposed model, we consider delay criterion violations

- due to the ON-OFF behavior, which results in temporary system overloads and
- due to the periodic packet emission during the ON states, where the emission phases of connections are uniformly distributed over the TTIs, which can result in packet congestion.

The allowed delay criterion violation probability for class i is $\tilde{\varepsilon}_i$, meaning that the CAC should ensure that $\Pr(D_i > \tilde{D}_i) \leq \tilde{\varepsilon}_i$ for $i = 1, \dots, K$.

2.3.1 Decomposition of burst-level and packet-level effects

Let $\{A(t), t \geq 0\}$ denote the amount of work arriving to the system in the interval $[-t, 0)$. Define the excess work arriving in $[-t, 0)$ as:

$$W(t) = A(t) - Ct. \quad (2.1)$$

The average input rate to the buffer in $[-t - TTI^{max}, -t)$ ($TTI^{max} = \max_i TTI_i$) is:

$$R(t) = \frac{A(t + TTI^{max}) - A(t)}{TTI^{max}}. \quad (2.2)$$

We define the *accumulated excess work* $W_{acc}(t)$ as the component of $W(t)$ due to the ON-OFF behavior (burst-level fluctuations):

$$W_{acc}(t) = \int_0^t R(u) du - Ct. \quad (2.3)$$

The evolution of $W_{acc}(t)$ depends on the distribution of the ON and OFF periods, the dependency among the sources, etc.

In the considered FIFO queue the waiting time is approximated using the *workload* (or virtual waiting time) [25]. The system is stationary so that $t = 0$ represents an arbitrary time instant. The workload is calculated as:

$$V(0) = \sup_{t \geq 0} W(t). \quad (2.4)$$

The “burst component” of the workload, i.e., the component considering only $W_{acc}(t)$ is:

$$V^{burst}(0) = \sup_{t \geq 0} W_{acc}(t). \quad (2.5)$$

Then, the “packet component” can simply be defined as the difference of the workload and its burst component:

$$V^{packet}(0) = V(0) - V^{burst}(0). \quad (2.6)$$

We are interested in the complementary distribution function of the workload:

$$Q(x) = \Pr\{V(0) > x\}. \quad (2.7)$$

The queueing process can be decomposed as follows:

$$\begin{aligned} Q(x) = & \quad (2.8) \\ & \Pr\{V^{packet}(0) + V^{burst}(0) > x \mid V^{burst}(0) > 0\} \cdot \Pr\{V^{burst}(0) > 0\} + \\ & \Pr\{V^{packet}(0) > x \mid V^{burst}(0) = 0\} \cdot \Pr\{V^{burst}(0) = 0\} \end{aligned}$$

In [24] it is shown that if the burst component is positive, then, in general, the values of the packet component are rather small compared to the value of the burst component $V^{burst}(0)$. Therefore, the following approximation can be used:

$$\begin{aligned} \Pr\{V^{packet}(0) + V^{burst}(0) > x \mid V^{burst}(0) > 0\} \cdot \Pr\{V^{burst}(0) > 0\} \approx & \quad (2.9) \\ \Pr\{V^{burst}(0) > x\}. \end{aligned}$$

Note however, that the traffic descriptors do not include any characterization of the ON and OFF period lengths. It means, that the distribution of the burst component of the workload $\Pr\{V^{burst}(0) > x\}$ can not be evaluated. Since our system has strict delay requirements ($\tilde{D}_i \leq TTI_i$), the waiting time in an overload situation (when $R(t) > C$ during a time interval) reaches very fast the predefined delay criterion. In other words, unless the overload situations are rather short, the queue can not smooth out the temporary overload efficiently, even if considering infinite buffer. Therefore, we take the conservative assumption that the delay of each packet, arriving in an overload situation, is always larger than the delay criterion. Using this assumption, instead of $\Pr\{V^{burst}(0) > x\}$ we will evaluate the probability that a packet arrives at an overload situation, and by this we will approximate the probability of that a packet is delayed due to system overload.

If the burst component is zero, $V^{burst}(0) = 0$, the queue empties periodically (with period TTI^{max}), and the distribution of the workload is determined by the cell emissions in a short interval before considered time $t = 0$.

2.3.2 The proposed model

Applying the assumptions we set up a combined model. Considering packets in the system we can observe two types of delay criterion violation events; some packets are delayed due to temporary system overload (including the ones possibly lost due to buffer overflow) and some packets are exceeding the delay criterion due to temporary packet congestion in the non-overloaded system. We define two measures as

$$\epsilon_i^{overload} = \frac{\# \text{ class } i \text{ packets delayed under overload}}{\# \text{ class } i \text{ packets}}, \quad (2.10)$$

and

$$\epsilon_i^{delayed} = \frac{\# \text{ class } i \text{ packets delayed under non-overload}}{\# \text{ class } i \text{ packets}}. \quad (2.11)$$

Denote the number of active connections (the connections in ON period) at time t of class i by $N_i^{act}(t)$ and let the vector of active connections at time t be $\underline{N}^{act}(t) = [N_1^{act}(t), N_2^{act}(t), \dots, N_K^{act}(t)]$. At a fixed time t_0 , we say that the system is in state \underline{n} if the random vector $\underline{N}^{act}(t_0)$ takes the value \underline{n} (i.e., $N_i^{act}(t_0) = n_i$; $i = 1, 2, \dots, K$). The probability that the number of active connections from class i , $N_i^{act}(t_0)$, is n_i can be obtained with a binomial distribution:

$$\Pi_i(n_i) = \binom{N_i}{n_i} \alpha_i^{n_i} (1 - \alpha_i)^{N_i - n_i}. \quad (2.12)$$

The probability that the system is in state \underline{n} , denoted by $\Pi(\underline{n})$, is calculated using a multi-dimensional binomial distribution as follows:

$$\Pi(\underline{n}) = \prod_{i=1}^K \Pi_i(n_i) = \prod_{i=1}^K \binom{N_i}{n_i} \alpha_i^{n_i} (1 - \alpha_i)^{N_i - n_i}. \quad (2.13)$$

We define the load of an active connection as the packet size divided by the period length: $\rho_i = b_i/TTI_i$. Then the input rate in state \underline{n} is:

$$R(\underline{n}) = \sum_{i=1}^K n_i \rho_i. \quad (2.14)$$

The measure $\epsilon_i^{overload}$ is approximated by the probability that a packet of class i arrives at an overload situation (when $R(\underline{n}) > C$):

$$\epsilon_i^{overload} \approx \epsilon_i^{overload} = \Pr\{\text{packet arrives at overload situation}\} = \frac{\sum_{\underline{n}: R(\underline{n}) > C} n_i \Pi(\underline{n})}{\sum_{\forall \underline{n}} n_i \Pi(\underline{n})}. \quad (2.15)$$

In case of a normal situation ($R(\underline{n}) \leq C$) the waiting time is dominated by the periodic packet emission. Thus the probability that a packet, arriving at time t_0 , is exceeding its delay criterion can be calculated as:

$$\epsilon_i^{delayed} \approx \epsilon_i^{delayed} = \frac{\sum_{\underline{n}: R(\underline{n}) \leq C} n_i \Pi(\underline{n}) \cdot \Pr(D_i > \tilde{D}_i \mid \underline{N}^{act}(t_0) = \underline{n})}{\sum_{\forall \underline{n}} n_i \Pi(\underline{n})}. \quad (2.16)$$

Finally, similarly to the decomposition in Eq.(2.8), the probability of delay criterion violation is the sum of two probabilities:

$$\epsilon_i = \epsilon_i^{overload} + \epsilon_i^{delayed}. \quad (2.17)$$

2.3.3 Validation of the proposed model

The proposed model assumes that each packet that arrives to a temporarily overloaded system violates the delay requirement. This assumption is conservative if the ON periods are short, because in this case a temporary overload may be handled even with a small buffer such that delay violation does not occur. The assumption is less conservative if the ON periods are long and the buffer is small, because in this case the buffer fills up quickly in an overload situation, and most packets will violate the delay requirement, or will be lost. It is possible that the assumption is not conservative if the ON periods are long and the buffer is large. In this case, the large buffers can also fill up, and when the overload situation is over (i.e., the overall input rate gets below the multiplex capacity), serving the whole buffer content takes a long time. It means that the overload situations can be significantly “lengthened” if the buffer is too large. The proposed model does not count with this effect. If the buffer is small, this “lengthening effect” is negligible, and the transition from overloaded to non-overloaded system is fast.

To check the assumptions of the proposed model, we simulated different ON-OFF sources. Figures 2.4, 2.5, 2.6 and 2.7 compare delays of Markov modulated sources with average ON period lengths of 20, 200, 2000 TTI^{max} to the result with the proposed model.

In the simulations, the following two traffic classes are considered: *voice* ($TTI_1 = 20$ ms, $b_1 = 40$ bytes, $\alpha_1 = 0.6$) and *data* ($TTI_2 = 40$ ms, $b_2 = 360$ bytes, $\alpha_2 = 1$). In each simulation, 30 voice and 2 data sources were multiplexed. The buffer sizes were 50 ms (Figures 2.4 and 2.5) and 15 ms (Figures 2.6 and 2.7). The server capacity was $C = 520$ kbps.

Up to $\tilde{D} \approx 10$ ms the delays hardly depend on the length of ON periods, and delay violations are dominated by the periodic packet emission. For larger \tilde{D} values the proposed model is mostly conservative, and the delay violation is mainly caused by too many active sources filling up the buffer. If the buffer is larger (for example 50 ms), then our assumptions are not fully true: the buffer can not get filled up (or emptied) quickly. Therefore, if the ON periods are rather long (2000 TTI^{max}), a slight underestimation

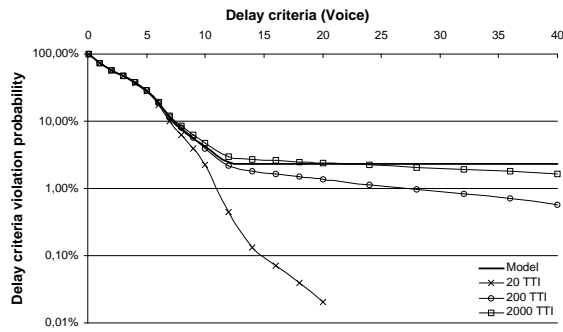


Figure 2.4: Simulation of voice packet delays with 50 ms long buffer

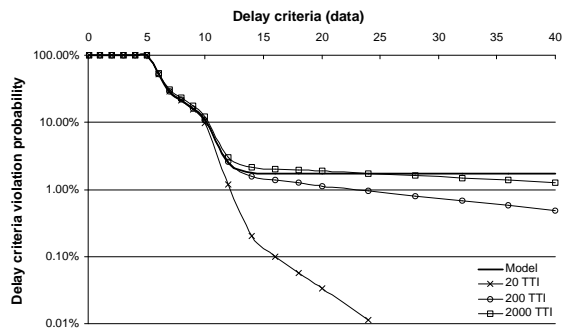


Figure 2.5: Simulation of data packet delays with 50 ms long buffer

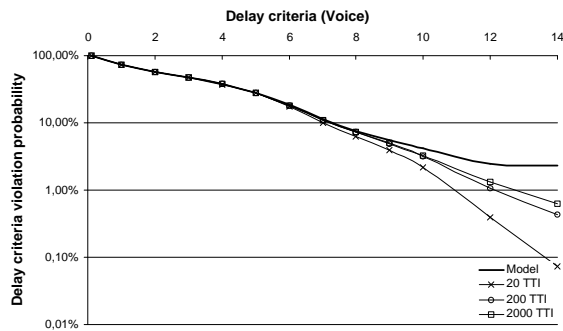


Figure 2.6: Simulation of voice packet delays with 15 ms long buffer

of the delay criterion violation probability is experienced in Figures 2.4 and 2.5.

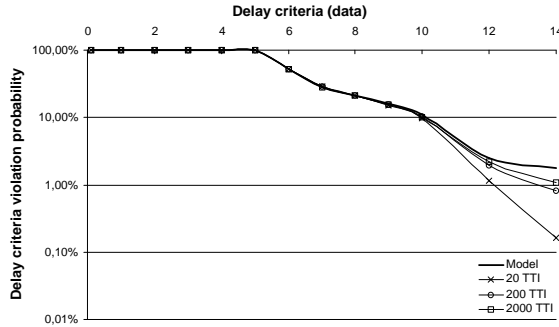


Figure 2.7: Simulation of data packet delays with 15 ms long buffer

2.4 Connection admission control algorithm

In this section, the connection admission control (CAC) algorithm is described. When a new connection arrives, the CAC needs to check:

- the delay violation due to temporary packet congestion in the non-overloaded system (i.e., the “delay limit”), see Section 2.4.1, and
- the delay violation due to temporary system overload (i.e., the “overload limit”), see Section 2.4.2.

In other words, the task of the CAC algorithm is to check on-line whether a certain traffic mix is within the *admissible region*. The admissible region contains the traffic mixes, where the QoS requirements are not violated.

We obtained admissible regions from extensive simulations. Based on these simulation results, we approximate the admissible region by the intersection of K regions with linear borders (also referred to as hyper-planes) and one region with (generally) non-linear border. We will refer to these regions as “delay-limited” and “overload-limited” regions, respectively. The delay-limited region bounded by the i -th hyper-plane contains the mixes, where the delay requirement of class i (target $\varepsilon_i^{delayed}$) is fulfilled. The overload-limited region with the non-linear border contains the mixes, which can temporarily overload the queuing system only with a small probability (target $\varepsilon_i^{overload}$). If the activity factor of each class is 1, the border of the overload-limited region becomes linear. In this case, the overload-limited region contains the mixes, which do not overload the system.

An example is depicted in Figure 2.8, where there are two classes. The first service has significantly stricter delay requirement, therefore the hyper-plane corresponding to the second service (which is not depicted) would be outside both depicted regions. In the shaded area, mixes containing connections from both classes can be accepted. In this figure, some mixes are within the delay-limited region, but outside the overload-limited region,

because it is possible that at 100% utilization the delay-limit is not yet exceeded. If the activity factor of the first class is less than 1, then the maximum number of connections from the first class given that there are no connections in the system from the second class, $N_{1,max}$, increases to $N_{1,max}^*$ and the border of the overload limited region becomes non-linear.

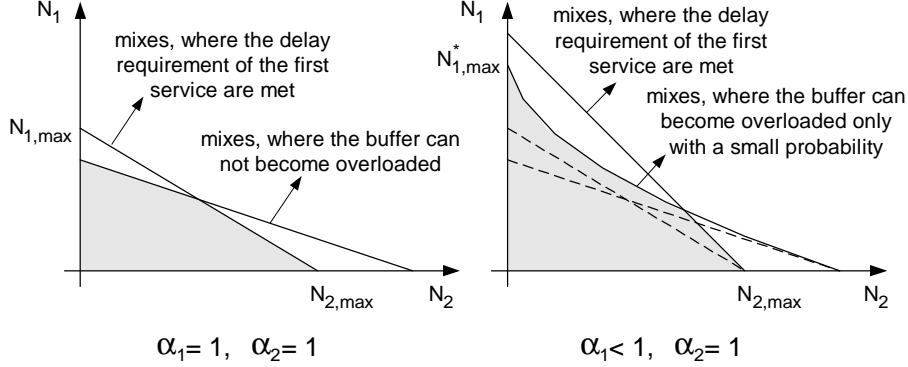


Figure 2.8: Admissible regions constructed of “delay-limited” and “overload-limited” regions

The linear approximation of delay-limited regions (see the validation in Section 2.5) means that taking a single region, the amount of allocated resources for a connection do not depend on the actual traffic mix. The non-linearity of the overload-limited region comes from Eq.(2.15), and its consequence is that the probability of temporary overload must be evaluated individually for each mix.

2.4.1 Checking delay violation in the non-overloaded system

To check the delay violation in the non-overloaded system Eq.(2.16) should be used. However, an exact formula for evaluating $\Pr(D_i > \tilde{D}_i \mid \underline{N}^{act}(t_0) = \underline{n})$ does not exist. Especially those cases are problematic, where classes of different TTIs are mixed [24] (see also Section 2.5.3). The approximation presented in Section 3.3 of [25] is computationally inefficient, therefore it is not applicable in a CAC.

Our approach is the following. We define the “delay-limited” regions with the intersection of K hyper-planes. The i -th hyper-plane defines the region where $\varepsilon_i^{delayed} \leq \tilde{\varepsilon}_i^{delayed}$. We propose a method for the hyper-plane construction, which uses only a single-class calculation for the evaluation of delay violation.

Define TN_{ij} as the maximum number of connections from class i assuming that a *single packet* from class j would fulfill the QoS requirement of

class j ($\varepsilon_j^{delayed} \leq \tilde{\varepsilon}_j^{delayed}$). We approximate by TN_{ij} the maximum number of class i connections if *one additional connection* from class j is present in the system. The proposed formula for determining TN_{ij} values is the following:

$$TN_{ij} = \max \left\{ N_i \left| \sum_{n_i=0}^{N_i} \Pi_i(n_i) \Pr \left\{ D_i > \tilde{D}_j - \frac{b_j}{C} \mid N_i^{act}(t_0) = n_i \right\} \leq \varepsilon_j^{delayed} \right. \right\}, \quad (2.18)$$

where the stability criterion is not needed, because it is included in Eq.(2.29). Note that in Eq.(2.18), we have used the following approximation of Eq.(2.16):

$$\varepsilon_i^{delayed} \approx \sum_{\underline{n}: R(\underline{n}) \leq C} \Pi(\underline{n}) \cdot \Pr(D_i > \tilde{D}_i \mid \underline{N}^{act}(t_0) = \underline{n}), \quad (2.19)$$

which will be used throughout the remaining part of the chapter.

Applying the results in [25], the distribution of the workload can be calculated with the following general formula (when presenting this formula $C = 1$ is assumed to avoid unnecessary notational complexity):

$$\Pr\{V^{packet}(0) > x\} = \sum_{l > x} \Pr\{A(l-x) = l\} \cdot \Pr\{V^{packet}(-(l-x)) = 0 \mid A(l-x) = l\}. \quad (2.20)$$

Using a new time unit $TU = b_j/C$ we introduce $TTI'_i = TTI_i/TU$ and $x = (\tilde{D}_j - b_j/C)/TU$, and then

$$\Pr \left\{ D_i > \tilde{D}_j - \frac{b_j}{C} \mid N_i^{act}(t_0) = n_i \right\} = \sum_{x < l \leq n_i} \binom{n_i}{l} \left(\frac{l-x}{TTI'_i} \right)^l \left(1 - \frac{l-x}{TTI'_i} \right)^{n_i-l} \cdot \frac{TTI'_i - n_i + x}{TTI'_i - l + x}. \quad (2.21)$$

Note that this expression is also valid for a transient system starting from an empty system at time 0 and receiving n_i uniformly distributed arrivals in $[0, TTI']$ (see also [25]). Therefore if $x \geq n_i - TTI'$, then we use Eq.(2.21). If $x < n_i - TTI'$, then $\Pr\{D_i > \tilde{D}_j - b_j/C \mid N_i^{act}(t_0) = n_i\} = 1$.

Applying the hyper-plane approximation, the necessary condition of accepting the traffic mix (N_1, N_2, \dots, N_K) is

$$\sum_{i=1}^K \frac{TN_{jj}}{TN_{ij}} \cdot N_i \leq TN_{jj} + 1 \quad j = 1, 2, \dots, K. \quad (2.22)$$

2.4.2 Checking delay violation due to temporary system overload

To check the delay violation due to temporary system overload Eq.(2.15) can be used. However, it can be too slow to evaluate on-line (it has a

complexity of $\mathcal{O}(N^K)$). Therefore, we propose a method, which exploits the statistical gain only within the different classes, but not among classes (it has a complexity of $\mathcal{O}(KN)$). It is also noted that the latter method can be implemented efficiently by storing the resulting values in memory (the complexity can be reduced to a memory-read). To take into account at least partially the statistical gains among classes, a simple extension is proposed.

To proceed further, we introduce L_i , which is the smallest number of class i connections being in ON state such that the requirement on the temporary overload is met. We will also refer to this value as the ‘‘per-class limit’’ of the number of connections in ON state. Its values are calculated as follows:

$$L_i = \min \left\{ L \left| K_i \sqrt{1 - \tilde{\varepsilon}_i^{overload}} \leq \frac{1}{N_i \alpha_i} \sum_{k=0}^L k \Pi_i(k) \right. \right\}, \quad (2.23)$$

and for the other classes:

$$L_l = \min \left\{ L \left| K_l \sqrt{1 - \tilde{\varepsilon}_l^{overload}} \leq \sum_{k=0}^L \Pi_l(k) \right. \right\}, l = 1, 2, \dots, K_a, l \neq i, \quad (2.24)$$

where K_a is the number of traffic classes with activity factor smaller than one ($\alpha_i < 1$). For always active traffic classes ($\alpha_i = 1$), the per-class limit equals the number of connections in the system, i.e., $L_i = N_i$.

Next, we show how Eq.(2.15) can be decomposed such that in the resulting formula only the terms in Eq.(2.23) and Eq.(2.24) appear. This decomposition is useful, because these terms can be evaluated for the different classes independently. This way, if a class i connection arrives, calculating only the term related to class i is needed. Assuming that approximation Eq.(2.15) is good,

$$1 - \varepsilon_i^{overload} = \frac{\sum_{n_1=0}^{N_1} \sum_{n_2=0}^{N_2} \cdots \sum_{n_K=0}^{N_K} \mathcal{I}_{\{R(\underline{n}) \leq C\}} n_i \prod_{k=1}^K \Pi_k(n_k)}{\sum_{n_1=0}^{N_1} \sum_{n_2=0}^{N_2} \cdots \sum_{n_K=0}^{N_K} n_i \prod_{k=1}^K \Pi_k(n_k)}, \quad (2.25)$$

where $\mathcal{I}_{\{expression\}}$ is the indicator function, which we will use several times throughout the Dissertation. $\mathcal{I}_{\{expression\}}$ equals 1, if the *expression* is true and it equals 0 if the *expression* is false. Using the per-class limit values (L_i values), the above equality is turned into the following inequality:

$$1 - \varepsilon_i^{overload} \geq \frac{\sum_{n_1=0}^{L_1} \sum_{n_2=0}^{L_2} \cdots \sum_{n_K=0}^{L_K} n_i \Pi_1(n_1) \Pi_2(n_2) \cdots \Pi_K(n_K)}{\sum_{n_1=0}^{N_1} \sum_{n_2=0}^{N_2} \cdots \sum_{n_K=0}^{N_K} n_i \prod_{k=1}^K \Pi_k(n_k)}. \quad (2.26)$$

This can be done, because the state space defined by $\underline{n} : R(\underline{n}) \leq C$ is larger than the one defined by $n_i \in \{0, 1, \dots, L_i\}; i = 1, \dots, K$. Since the classes are independent, the following simplification can be done:

$$1 - \varepsilon_i^{overload} \geq \frac{\sum_{n_1=0}^{L_1} \Pi_1(n_1) \cdot \sum_{n_2=0}^{L_2} \Pi_2(n_2) \cdots \sum_{n_i=0}^{L_i} n_i \Pi_i(n_i) \cdots \sum_{n_K=0}^{L_K} \Pi_K(n_K)}{\sum_{n_i=0}^{N_i} n_i \Pi_i(n_i)}. \quad (2.27)$$

Finally, the decomposition is formulated as:

$$1 - \varepsilon_i^{overload} \geq \frac{1}{N_i \alpha_i} \sum_{n_i=0}^{L_i} n_i \Pi_i(n_i) \cdot \prod_{l \neq i} \left(\sum_{n_l=0}^{L_l} \Pi_l(n_l) \right). \quad (2.28)$$

For the sake of simplicity, assume that $\tilde{\varepsilon}_i^{overload} = \tilde{\varepsilon}_j^{overload}$ for each $i, j = 1, \dots, K$. This assumption is typically used in practice. Then, if a new class i connection arrives to the system, and the actual connection mix becomes $(N_1, N_2, \dots, N_i, \dots, N_K)$, one needs to calculate or read from the memory the new L_i value and check the following inequality:

$$\sum_{i=1}^K L_i \rho_i \leq C, \quad (2.29)$$

which is the necessary condition of accepting (N_1, N_2, \dots, N_K) .

It is obvious that using Eq.(2.23) and Eq.(2.24), only the statistical gain of multiplexing sources from the same class is exploited. Keeping the property that L_i values can be obtained independently from each other, but taking into account partially statistical gains from multiplexing different classes, one may proceed as follows. (For the sake of simplicity, we use Eq.(2.23) instead of Eq.(2.24), which is a conservative approximation.)

1. Find L_i^* for all i using Eq.(2.23) with

$$N_i^* = N_i + \sum_{\alpha_k \leq \alpha_i, k \neq i} \min \left(1, \frac{\rho_k}{\rho_i} \right) N_k; \quad k = 1, \dots, K,$$

and calculate the statistical multiplexing gain for class- i as:

$$MG_i = (N_i^* - L_i^*)/N_i^*.$$

Explanation: If we have two classes, and $\alpha_2 \leq \alpha_1$, then the achievable statistical multiplexing gain of class 1 in a system fed by N_1^* connections of load ρ_1 each, is smaller than in the same system fed by N_1 connections of load ρ_1 each, plus N_2 connections of load ρ_2 each.

2. Repeat until MG_i values are no longer increasing:

- Consider the classes with $\alpha_k > \alpha_i$ and $\rho_k < \rho_i, \forall i, k$. If $MG_k > MG_i$, then let $\alpha'_i := \alpha_k$ and calculate MG'_i executing step 1 with the temporary activity factor value α'_i . If the resulting $MG'_i > MG_i$, then let $MG_i := MG'_i$. (At the end of this step reset the original value of α_i .)
- Consider the classes with $\alpha_k > \alpha_i$ and $\rho_k \geq \rho_i, \forall i, k$. If $MG_k > MG_i$, then let $MG_i := MG_k$.

Explanation: If we have two classes, and $\alpha_2 > \alpha_1$ (note that these are different class pairs from the ones considered in step 1), then two cases can occur: (1) either $\rho_2 < \rho_1$, or (2) $\rho_2 \geq \rho_1$. In case of (2), it is clear that class 1 can achieve at least as large statistical multiplexing gain as class 2. In case of (1), class 1 can achieve at least as large statistical multiplexing gain as it could achieve in a modified system, where α_1 were increased such that $\alpha'_1 = \alpha_2$.

3. Finally, $L_i = N_i(1 - MG_i)$ for all i .

Note: The multiplexing gain values, $MG_i, i = 1, \dots, K$, have been modified in steps 1 and 2 such that also statistical multiplexing effects between class-pairs are taken into account. Therefore, L_i values calculated in step 3 can be smaller than the ones obtained without using this extension method.

2.4.3 Flow-chart of the algorithm

A flow-chart of the algorithm is shown in Figure 2.9. In general it is proposed to separate the update (non-real-time) and the decision-making (real-time) parts.

If a connection of a new traffic class (not included in the TN matrix yet) arrives, and the TN update proves to be too slow to perform on-line, a fast update of TN is needed. This fast update can be, for example, using the worst-case peak bandwidths ($TN_{new,i} = b_{new}/\tilde{D}_i, i = 1, \dots, K$, where K already includes the new traffic class), or the fast approximation Eq.(2.48) provided in the next section.

2.5 Validation of the hyper-plane approximation

In this section, a validation of the proposed CAC algorithm is given, and a formula, which can be used for the fast update (see Section 2.4.3) is derived. Since the exact solution to Eq.(2.15) is known, but there is no exact solution to Eq.(2.16) [24], we need to validate only the hyper-plane approximation proposed in Section 2.4.1. We have simulated admissible regions extensively

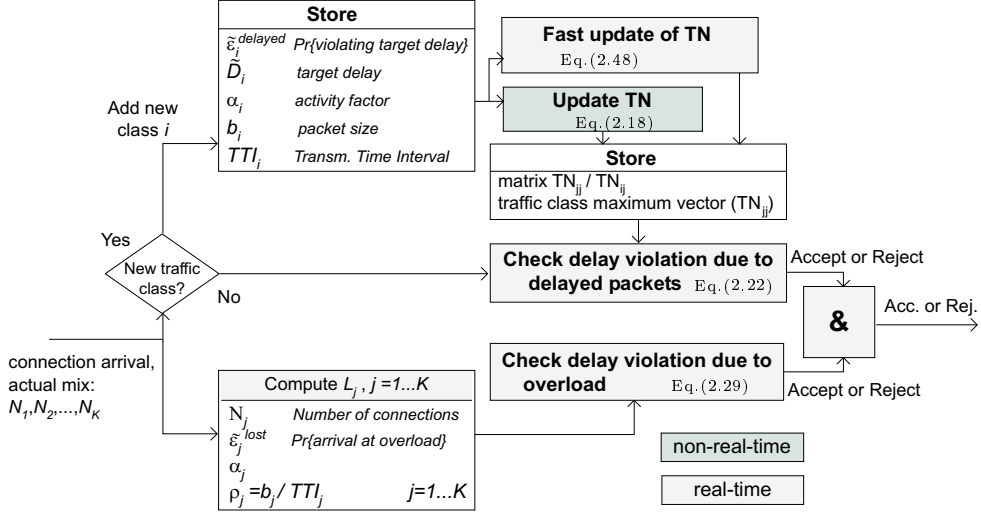


Figure 2.9: Block-diagram of the algorithm. In the figure it is assumed that the L_i values are calculated on-line, or read from tables calculated off-line.

with UTRAN-specific traffic parameters, and found that delay-limited borders of the admissible regions are linear. Below we provide analytic results, which support this observation.

Since we consider only the cases where $R(\underline{n}) \leq C$, which always result in a periodic queue behavior with period TTI^{max} , we need to take into account the queueing process only in $[0, TTI^{max}]$. Let $Q(x)$ be the complementary distribution function of the virtual waiting time in a FIFO queue. $Q(x)$ can be obtained using a general queueing theory result [25]:

$$Q(x) = \Pr \left\{ V^{packet}(0) > C x \right\} = \Pr \left\{ \sup_{\tau \geq 0} (A(\tau) - C \tau) > C x \right\}. \quad (2.30)$$

Denote the arrival process resulting from superimposing several independent periodic sources from class i by $A_i(t)$. The process $A_i(t) - (t/TTI_i) b_i N_i$ can be approximated by the process $\sqrt{N_i} b_i \mathcal{B}(t/TTI_i)$, where $\mathcal{B}(t)$ is a Brownian bridge [26], i.e., the standard Brownian motion conditioned on the event $\mathcal{B}(1) = 0$. In fact, such a superimposed process converges to a Brownian bridge as the number of sources grows [33]. Using the Brownian bridge approximation enables us to obtain the solution of Eq.(2.30) in a closed form, because the following formula is true [33, 26]:

$$\Pr \left\{ \sup_{0 \leq \tau \leq 1} (\mathcal{B}(\tau) - z \tau) \geq a \right\} = e^{-2a(z+a)}, \quad (2.31)$$

where $a > 0, z > 0$. During the analysis, we approximate the arrival process of ON-OFF periodic sources with Gaussian processes, because we conjecture

that closed form solutions, similar to Eq.(2.31) can be obtained. As we can see below, we can deduce the validity of the hyper-plane approximation from closed forms.

2.5.1 Same TTIs and “always ON” sources

The superposition of different traffic classes with the same period $TTI_i = TTI$ and the same activity factor $\alpha_i = 1$ for all classes ($i = 1, \dots, K$) results in [26]:

$$A(t) = \sum_i N_i b_i \frac{t}{TTI} + \sqrt{\sum_i N_i b_i^2} \mathcal{B} \left(\frac{t}{TTI} - \left\lfloor \frac{t}{TTI} \right\rfloor \right). \quad (2.32)$$

Scaling from $t \in [0, TTI]$ to $\tau \in [0, 1]$ we obtain

$$\Pr \left\{ \sup_{0 \leq t \leq TTI} (A(t) - C t) \geq C x \right\} = \quad (2.33)$$

$$\Pr \left\{ \sup_{0 \leq \tau \leq 1} \left(\sum_i N_i \rho_i \tau + \sqrt{\sum_i N_i \rho_i^2} \mathcal{B}(\tau) - C \tau \right) \geq \frac{C x}{TTI} \right\}.$$

Applying Eq.(2.30) and Eq.(2.31);

$$Q(x) = \exp \left\{ - \frac{2 C x}{TTI \sum_i N_i \rho_i^2} \left(\frac{C x}{TTI} + C - \sum_i N_i \rho_i \right) \right\}, \quad (2.34)$$

and the resulting admissible region is a hyper-plane in N_i :

$$Q(x) \leq \tilde{\varepsilon}^{delayed} \iff \sum_i N_i \left(\rho_i + \frac{\rho_i^2}{C} \frac{\gamma TTI}{2 x} \right) \leq C + \frac{C x}{TTI}, \quad (2.35)$$

where $\gamma = -\ln(\tilde{\varepsilon}^{delayed})$, and $\tilde{\varepsilon}_i^{delayed} = \tilde{\varepsilon}^{delayed}$ for $i = 1, \dots, K$.

The accuracy of the Brownian bridge approximation is good, especially for heavy traffic (for more information see [25] and Section 2.6.1). Note that since Eq.(2.31) is exact, the linearity of the true admissible region is justified by the inherent relation between the real arrival process and the Brownian bridge approximation. Therefore, applying the method in Section 2.4.1 means that we accept the validity of the linear admissible region, but calculate the edges with the exact values (i.e., using Eq.(2.18)) instead of the approximation Eq.(2.35).

2.5.2 Same TTIs but different activity factors

In this section we show that having heterogeneous activity factors (but still $TTI_i = TTI$; $i = 1, \dots, K$) does not invalidate the hyper-plane approximation. Furthermore, a fast approximation of the hyper-plane is given, which can be used at the fast update of the TN matrix. Our objective is to find good, closed form approximations of Eq.(2.18) and Eq.(2.19).

For the ease of notation, assume first that there are N_i connections of a single traffic class (class i) in the system. We need to evaluate:

$$\Pr \left\{ \sup_{0 \leq \tau \leq TTI} \left(\sum_{n_i=0}^{N_i} \mathcal{I}_{\{\#active=n_i\}} A_i^{n_i}(\tau) - C \tau \right) > C x \right\}. \quad (2.36)$$

$A_i^{n_i}(t)$ is the arrival process if the number of active connections is n_i :

$$A_i^{n_i}(t) = n_i b_i \frac{t}{TTI} + \sqrt{n_i b_i^2} \mathcal{B} \left(\frac{t}{TTI} - \left\lfloor \frac{t}{TTI} \right\rfloor \right). \quad (2.37)$$

The sum in Eq.(2.36) represents a random sum of Gaussian processes, therefore it is typically not Gaussian. Still, in order to find a closed form solution, we will approximate it as a Gaussian process.

The second moment of process $A_i^{n_i}(t)$ is:

$$\mathbb{E} \left((A_i^{n_i}(t))^2 \right) = n_i b_i^2 f(t) + n_i^2 \rho_i^2 t^2, \quad (2.38)$$

where

$$f(t) = \frac{t}{TTI} \left(1 - \frac{t}{TTI} \right), \quad (2.39)$$

if $t \in [0, TTI]$. Then, the mean and the variance of the arrival process $A_i(t)$, $t \in [0, TTI]$ is:

$$\mathbb{E}(A_i(t)) = \sum_{n_i=0}^{N_i} \Pi_i(n_i) n_i \rho_i t = \alpha_i N_i \rho_i t, \quad (2.40)$$

and

$$\begin{aligned} \mathbf{Var}(A_i(t)) &= \sum_{n_i=0}^{N_i} \Pi_i(n_i) (n_i b_i^2 f(t) + n_i^2 \rho_i^2 t^2) - (\alpha_i N_i \rho_i t)^2 = \quad (2.41) \\ &N_i \left(\alpha_i b_i^2 f(t) + \alpha_i (1 - \alpha_i) \rho_i^2 t^2 \right) = N_i b_i^2 \frac{\alpha_i t}{TTI} \left(1 - \frac{\alpha_i t}{TTI} \right), \end{aligned}$$

which is the same as the variance function of a Brownian bridge over the interval $[0, TTI/\alpha_i]$.

Since we observe the arrival process only in $t \in [0, TTI]$, we consider the auto-covariance function of the arrival process $A_i(t)$ also in this interval

only. Let $s, t \in [0, TTI]$, $|s - t| < TTI$ and $s < t$. When evaluating the auto-covariance function, the term:

$$\mathbb{E} \left(\sum_{n_i=0}^{N_i} \mathcal{I}_{\{\#active=n_i\}} A_i^{n_i}(s) \cdot \sum_{n_j=0}^{N_j} \mathcal{I}_{\{\#active=n_j\}} A_j^{n_j}(t) \right) ; i, j = 1, \dots, K, \quad (2.42)$$

simplifies to:

$$\mathbb{E} \left(\sum_{n_i=0}^{N_i} \mathcal{I}_{\{\#active=n_i\}} A_i^{n_i}(s) \cdot A_i^{n_i}(t) \right) ; i = 1, \dots, K, \quad (2.43)$$

because within one period (within one TTI) we can consider that Eq.(2.42) is not zero only if $i = j$. Therefore, the auto-covariance function can be obtained similarly to the variance function with the following result:

$$\mathbf{Cov}(A_i(t), A_i(s)) = N_i b_i^2 \frac{\alpha_i s}{TTI} \left(1 - \frac{\alpha_i t}{TTI} \right). \quad (2.44)$$

which is the same as the auto-covariance function of a Brownian bridge over the interval $[0, TTI/\alpha_i]$.

Therefore, the arrival process in $t \in [0, TTI]$ having K traffic classes can be approximated by a Gaussian process as:

$$A(t) = \sum_{i=1}^K N_i \alpha_i b_i \frac{t}{TTI} + \sum_{i=1}^K \sqrt{N_i b_i^2} \mathcal{B}_i \left(\frac{\alpha_i t}{TTI} \right). \quad (2.45)$$

For this arrival process, as it is shown in Appendix A, $Q(x)$ can be approximated as:

$$Q(x) \approx \exp \left\{ -\frac{2 C x}{TTI \sum_i N_i \alpha_i \rho_i^2} \left(\frac{C x}{TTI} \frac{\sum_i N_i \alpha_i^2 \rho_i^2}{\sum_i N_i \alpha_i \rho_i^2} + C - \sum_i N_i \alpha_i \rho_i \right) \right\}, \quad (2.46)$$

and the resulting approximation of the admissible region is:

$$Q(x) \leq \tilde{\varepsilon}^{delayed} \iff \sum_i N_i \left(\alpha_i \rho_i + \frac{\alpha_i \rho_i^2}{C} \frac{\gamma TTI}{2 x} \right) \leq C + \frac{C x}{TTI} \frac{\sum_i N_i \alpha_i^2 \rho_i^2}{\sum_i N_i \alpha_i \rho_i^2}. \quad (2.47)$$

This approximation is good (see Section 2.6.1), and this region is practically linear in N_i . The right hand side depends on N_i , but it is easy to check that if the region is not linear, it tends to be convex. Therefore, the hyper-plane approximation is justified. Also note, that Eq.(2.47) can be used for the fast update of the TN matrix as follows:

$$TN_{ij} \approx \frac{C (TTI_i + \alpha_i y)}{\alpha_i b_i} \left(1 - \frac{b_i \ln \left(\tilde{\varepsilon}_j^{delayed} \right)}{2 C y} \right)^{-1}, \quad y = \tilde{D}_j - \frac{b_j}{C}. \quad (2.48)$$

2.5.3 Effect of having different TTIs

In this section we show that having different TTIs does not invalidate the hyper-plane approximation.

Consider two classes with different TTIs, $TTI_1 > TTI_2$ and $TTI_1 = k \cdot TTI_2$, where $k = 2, 4$ or 8 in UMTS. The activity factors of the classes are 1 ($\alpha_1 = \alpha_2 = 1$). In this case Eq.(2.30) takes the following form:

$$Q(x) = \Pr \left\{ \sup_{0 \leq \tau \leq TTI_1} \left(c_1 \mathcal{B}_1 \left(\frac{\tau}{TTI_1} \right) + c_2 \mathcal{B}_2 \left(\frac{\tau}{TTI_2} - \left\lfloor \frac{\tau}{TTI_2} \right\rfloor \right) - c_3 t \geq c_4 x \right) \right\}, \quad (2.49)$$

where $\mathcal{B}_1(t)$ and $\mathcal{B}_2(t)$ are independent Brownian bridges, c_1, c_2, c_3 and c_4 are non-negative, real constants. It can be seen that the arrival process does not have independent increments in $[0, TTI_1]$, because the class with TTI_2 has k periods with exactly the same trajectories within $[0, TTI_1]$. Therefore, this distribution function can not be obtained in a closed form. Instead of evaluating Eq.(2.49), we consider upper and lower bounds.

A lower bound to $Q(x)$ is obtained by taking the supremum in Eq.(2.49) only over $[0, TTI_2]$ instead of $[0, TTI_1]$. This is equivalent to evaluating the modified system where $TTI_1^{low} = TTI_2^{low} = TTI_2$, $\alpha_2^{low} = 1$ and $\alpha_1^{low} = TTI_2/TTI_1$. Note that, if N_1 is small enough, such that all packets from class 1 do not overload the system in TTI_2 (i.e., if $N_1 \cdot b_1/TTI_2 \leq C$), this is not a lower bound, but yields exactly $Q(x)$. It means that for small x values (i.e., strict delay requirements such that the achievable utilization is low) the lower bound equals the exact solution, therefore the hyper-plane approximation is good.

An upper bound to $Q(x)$ can be obtained by evaluating the modified system with $TTI_2^{up} = TTI_1^{up} = TTI_1$, $N_1^{up} = N_1$ and $N_2^{up} = N_2 \cdot TTI_1/TTI_2$.

The left hand side of Eq.(2.47) is the same for the two bounds, and as $x \rightarrow 0$ the difference between the bounds (i.e., the second term on the right hand side of Eq.(2.47)) disappears. Therefore, for strict delay requirements the hyper-plane approximation is good. If $x \rightarrow TTI_2$, then already the overload-limited region corresponding to Eq.(2.15) constrains the admissible region.

In Figure 2.10, the above thoughts are demonstrated using an example. The meaning of the labels are the following: “CAC” is the admissible region calculated by the proposed algorithm, “CAC: check delay only” is the region calculated by checking only the delay violation as proposed in Section 2.4.1, “upper bound” is the hyper-plane approximation calculated using the lower bound to $Q(x)$, and “lower bound” is the hyper-plane approximation calculated using the upper bound to $Q(x)$.

If $\tilde{D} = 5$ ms, then $N_1 \cdot b_1/TTI_2 \leq C$ for all mixes, therefore the upper bound is exact. If $\tilde{D} = 15$ ms or $\tilde{D} = 20$ ms, then already the overload-limited region constrains the admissible region.

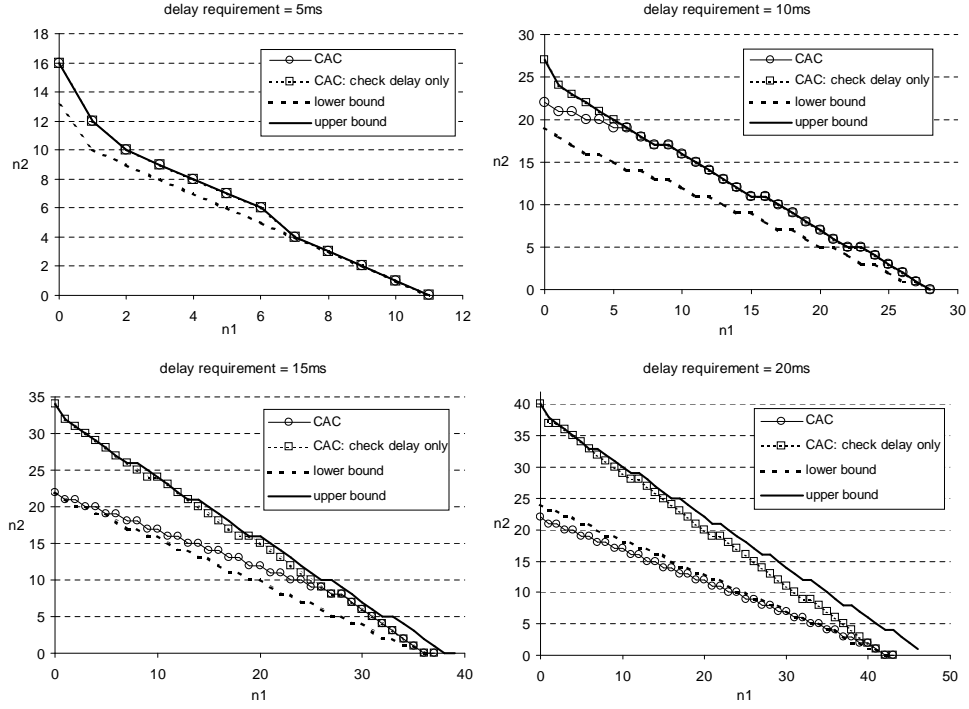


Figure 2.10: Example for the hyper-plane approximation for different TTIs; $C = 1980$ kbps, $TTI_1 = 80$ ms, $b_1 = 3600$ bit, $TTI_2 = 20$ ms, $b_2 = 1800$ bit, $\tilde{D}_1 = \tilde{D}_2 = 5, 10, 15, 20$ ms, $\tilde{\epsilon} = 1$ %

2.6 Numerical examples

In this section, different approximations for calculating the TN matrix are compared by numerical examples, and example admissible regions are presented.

In the examples, three traffic classes are considered with the following traffic descriptors and QoS requirements: $TTI_1 = 20$ ms, $b_1 = 336$ bit, $\alpha_1 = 0.65$, $TTI_2 = 10$ ms, $b_2 = 1512$ bit, $\alpha_2 = 0.85$, $TTI_3 = 40$ ms, $b_3 = 2688$ bit, $\alpha_3 = 1$, $\tilde{D}_1 = 5$ ms, $\tilde{D}_2 = 8$ ms, $\tilde{D}_3 = 20$ ms, $\tilde{\epsilon} = 0.001$. The link capacity is $C = 1920$ kbps.

2.6.1 Comparison of different approximations

The accuracy of three alternatives for the calculation of the TN matrix are compared. The probability of temporary system overload (according to Eq.(2.15)) is computed with the exact method. The three alternatives for calculating the TN values are the following:

- **Alternative A:** TN values are calculated with Eq.(2.18), and the delay violation probability in Eq.(2.18) is calculated with Eq.(2.21). This alternative is the most accurate, but it requires the most computation effort.
- **Alternative B:** TN values are calculated with Eq.(2.18), and the delay violation probability in Eq.(2.18) is calculated with Eq.(2.34). This alternative uses the Brownian bridge approximation, so it is less accurate than **A**, but it also requires less computation effort.
- **Alternative C:** TN values are calculated with Eq.(2.48). This alternative is expected to be less accurate than **A** and **B**, but it is a closed form, so it requires the least computation effort.

Admissible regions using the three alternatives were calculated, and the probabilities of delay criterion violation ($\varepsilon_i, i = 1, \dots, K$) for the traffic mixes on the surface of the admissible region were simulated. Denote by $\varepsilon_i(j)$ the measured delay criterion violation probability for class i having traffic mix j (i.e., traffic mixes are indexed). We constructed the vector $\underline{\varepsilon}^{max}$ such that the j -th element of it is:

$$\underline{\varepsilon}^{max}[j] = \max_i \varepsilon_i(j), \quad j = 1, \dots, M^{alt}, \quad i = 1, \dots, K, \quad (2.50)$$

where M^{alt} is the number of mixes on the surface of the admissible region when using alternative alt ($alt = \mathbf{A}, \mathbf{B}$ or \mathbf{C}). Figure 2.11 shows three histograms created from $\underline{\varepsilon}^{max}$ vectors obtained using alternatives **A**, **B** and **C**.

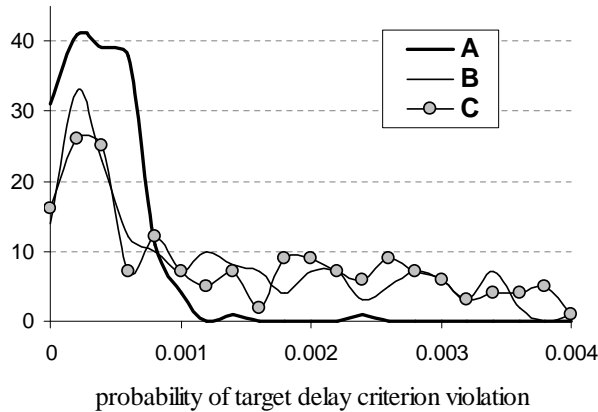


Figure 2.11: Histograms of delay criterion violation probabilities comparing calculation alternatives **A**, **B** and **C**

Denote by $D_i(j)$ ($i = 1, \dots, K; j = 1, \dots, M^{alt}$) the delay value that is exceeded in the simulations only with the target delay criterion violation probability $\tilde{\varepsilon}_i$. For each traffic class, the worst delay value is:

$$D_i^{max} = \max_j D_i(j), \quad i = 1, \dots, K, \quad j = 1, \dots, M^{alt}. \quad (2.51)$$

Table 2.1: D_i^{max} values for the example of Figure 2.11

| | D_1^{max} | D_2^{max} | D_3^{max} |
|----------|-------------|-------------|-------------|
| A | 4.9 ms | 8.4 ms | 13.8 ms |
| B | 6 ms | 8.9 ms | 13.8 ms |
| C | 6 ms | 9 ms | 13.7 ms |

From Figure 2.11 and Table 2.1 we see that alternative **A** is conservative, i.e., with a few exceptions the target delay criterion violation probability ($\tilde{\varepsilon} = 0.001$) is fulfilled. Alternatives **B** and **C** are less accurate than **A**, because they use the Brownian bridge approximation, which is only accurate if the number of sources is large (say above 20). Note that the fastest alternative **C** is not significantly more inaccurate than **B**. In other words, it means that the Brownian bridge approximation (at least for a low number of sources) introduces more approximation error than the Gaussian approximation described in Section 2.5.2.

We propose to use alternative **A** if the number of acceptable sources from a class is less than a certain limit (for example 20). If the number of sources is large (say above 100), the use of **C** is preferred. Also the fast update of the TN matrix (see Section 2.4.3) should be performed using **C**. Otherwise, **B** may be used.

2.6.2 Example admissible regions

In Figure 2.12 and Figure 2.13 admissible regions are presented together with simulations of the traffic mixes on the surface of the regions. These admissible regions have been obtained using the proposed CAC algorithm (using calculation alternative **A**, see Section 2.6.1). For the same examples, admissible regions obtained using simulations are shown in Figure 2.14.

It can be seen that the borders of the admissible region are really hyperplanes. Except some rare cases, the proposed CAC is conservative. In case of small links (for example, Example 2 in Figure 2.13), the calculation of TN as proposed in Section 2.4.1 can be too conservative. In this case, values of TN can be small, and therefore it can be a problem that the elements of TN are integers. For example, if there are large bearers in the system, such as class 2 in this numerical example, $TN_{ij} = 4$ can result in a significantly smaller admissible region than $TN_{ij}=5$. This problem may be solved with

an interpolation method (see [J6] for more details) by allowing elements of TN to take real values. For example, if the algorithm calculates that TN_{ij} is between 4 and 5, then instead of accepting the conservative value 4, one may use interpolation to find a more accurate (but still conservative) value between 4 and 5. If alternative **C** is used, TN_{ij} values are not integers, and thus interpolation is not needed.

2.7 Conclusion

A queuing model for the user traffic on the UTRAN Iub interface has been introduced and verified by simulations. Using the traffic model, analysis methods have been investigated and an efficient connection admission control algorithm has been developed.

The conditions of accepting a newly arriving connection are Eq.(2.22) and Eq.(2.29).

To check the condition given by Eq.(2.22), the hyper-plane approximation has been developed. The hyper-planes can be determined by evaluating Eq.(2.18). For performing this evaluation efficiently, a closed form formula has been given by Eq.(2.48). The approximations used by the algorithm have been validated both by analytical considerations and by simulations. Particularly, the hyper-plane approximation has been validated by analytical means using Eq.(2.47).

The condition given by Eq.(2.29) can be evaluated with the help of inequalities Eq.(2.23) and Eq.(2.24), and with the extension method presented at the end of Section 2.4.2.

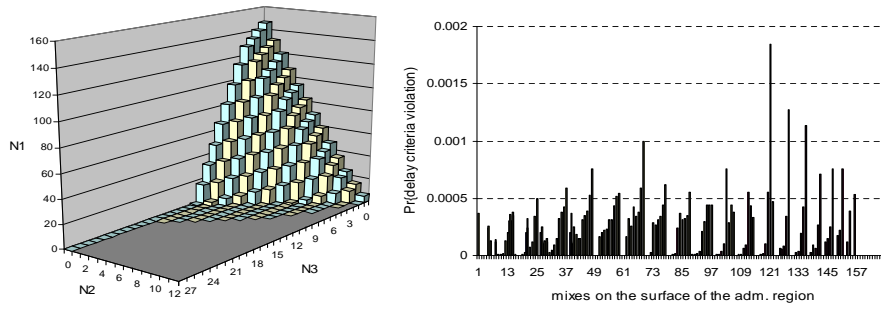


Figure 2.12: Example 1, left: admissible region obtained with the CAC ($C = 1920$ kbps), right: simulated delay violation probabilities for the mixes on the surface

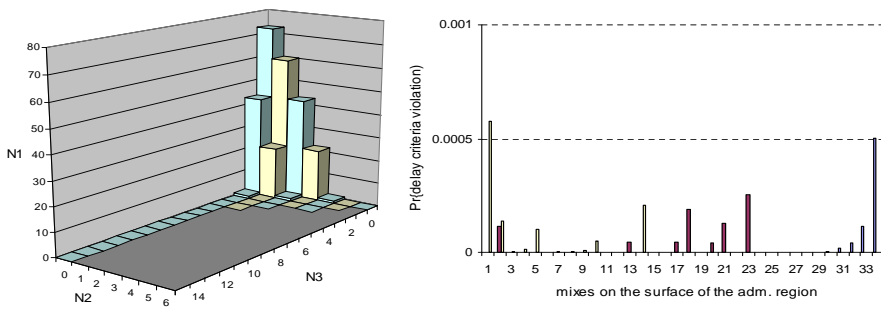


Figure 2.13: Example 2: same as Figure 2.12, but with $C = 1024$ kbps

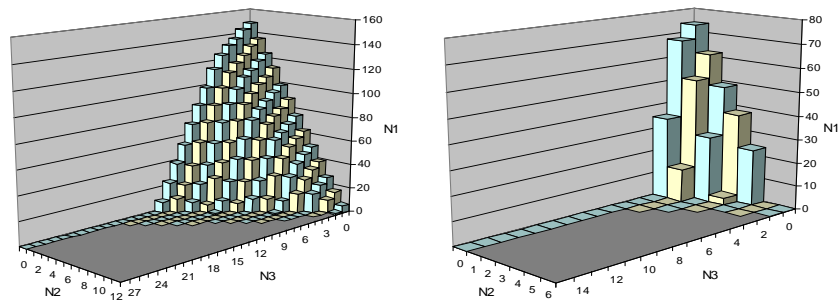


Figure 2.14: Simulated admissible regions for Example 1 and Example 2

Chapter 3

Mobility and traffic analysis for WCDMA networks

In cellular mobile communication systems, the mobility of vehicles affects some important parameters, such as handover rates, channel occupancy times and blocking probabilities. The work presented in this chapter is based on a model proposed by Nakano et al. [13] that suggests a convenient and practical approach to build an analytic traffic model, which also includes the effect of vehicle mobility.

This model is extended for CDMA networks, where the handover is of the so-called “soft handover” type. Performance measures like soft handover rates, soft handover type distributions and the offered communication traffic per cell, etc., are obtained. It is also explained how the model could be used for design and analysis of 3rd generation WCDMA systems.

3.1 Introduction

As today’s cellular operators move to increase the number of services they offer to subscribers – e.g., by integrating wireless access to the Internet – new technologies are required in their systems. Third generation cellular networks [34] are designed to offer:

- increased capacity within their existing spectrum allocation,
- higher capacities and lower system design costs per subscriber,
- new subscriber features and integrated (voice and data) services, which will help the operators to increase their market penetration.

Code Division Multiple Access (CDMA) is regarded as the most suitable multiple access technology to fulfill the above requirements, and Wideband CDMA (WCDMA) [6] is capable to serve the new, high data rate wireless multimedia demand.

The model presented by Nakano et al. in [13] offers a convenient and practical approach to build an analytic traffic model, which also includes the effect of vehicle mobility. In [13], the method was presented for cellular networks, in which the handover is of the so-called “hard handover” type.

In this chapter, the above model is extended to a CDMA cellular network, where the significant part of the handovers is of the “soft handover” type. Soft handover means that a mobile terminal can communicate with more than one base station at the same time. (For a thorough discussion of the advantages and disadvantages of soft handover see [35].) The consequence of this is that the traffic load on the fixed access network will not only (or mostly) depend on the call intensity, but the distribution of the number of legs a mobile is connected to the network with will also have a significant effect. For example, a call that is served by three base stations for some time will be (during that time) carried over three separate RAB connections (see also Chapter 2) in the fixed network up to the so-called Diversity Handover unit (DHO), which combines the information stream on the three connections into one single connection.

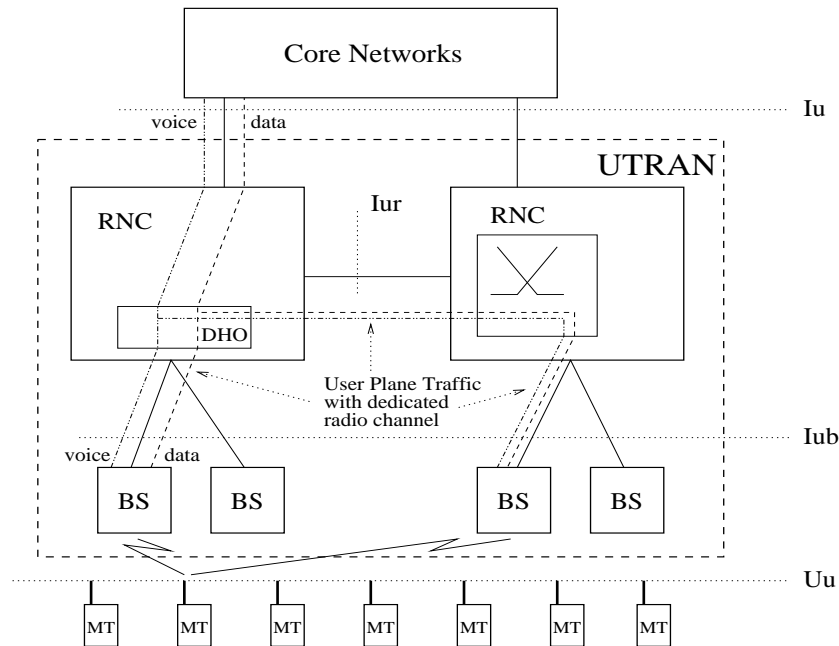


Figure 3.1: UTRAN Architecture

In the model, a road system is defined and covered with overlapping CDMA cells. The vehicle traffic on this road system is described by a vehicle traffic matrix. The vehicle traffic is routed over the road system, and the vehicle traffic load and the vehicle speed on different streets are calculated.

Calls are generated to/from the vehicles as a Poisson process. Different call classes are defined in the model, which makes possible to investigate integrated-services WCDMA systems. Soft handover rates, the offered communication traffic per cell, blocking probabilities and also the distribution of the offered traffic with different number of legs (i.e., different number of connections per call) are obtained. The method also gives a possibility to estimate the offered user-plane traffic between Radio Network Controllers (RNCs) in UMTS access networks (i.e., on the Iur interface, see Figure 3.1).

3.1.1 Overview of the literature

Several papers have been written about mobility modeling and spatial traffic distribution in cellular networks. First we give an overview of the literature that is most relevant to our work, then the objectives and goals of our work are discussed.

In [36], many aspects of mobility modeling in third generation mobile systems are considered, giving an overview and a starting point for the interested reader. In [37] and [38], cells are divided into soft handover regions and calls are uniformly generated in a cell. As a result, the probability distribution function of total sojourn times in the soft handover region of a given cell during a channel holding time is obtained by analytical calculations. In these works, the mobile can move up, down, left and right, changing its speed M times during a call (where M is a geometrically distributed random variable) according to a uniform distribution over $[0, Vmax]$. Applying these distributions, it is difficult to find parameters such that they fit to real measurements. Especially, vehicular traffic is difficult to model, since the effects of a road map (directions, speeds, vehicular traffic hot spots, etc.) are difficult to take into account. Mobility and traffic analysis in a road network model for cellular systems featuring hard handover is presented in [13]. Since this model uses elements of both transportation engineering and teletraffic engineering, it takes a considerable step towards practical application. An analytic traffic model is proposed in [39] to estimate the soft handover rate in CDMA cellular systems. As a limitation, a mobile can have at most two simultaneous connections (to two base stations) at a time. The call generation rate in the different regions (handover or non-handover) depends on the ratio of the areas of regions and cells. The main result is the analysis of the sensitivity of the handover rate when varying the cell radius, the speed of mobiles and the activity of the users. In [40], the authors present a mathematical formulation for systematic tracking of the random movement of a mobile station in a cellular environment. Based on this detailed formulation, a computer simulation is developed to obtain the behavior of different mobility-related parameters (e.g., the handover rate). An analysis of data obtained by simulation shows that the generalized gamma distribution function is a good approximation for the cell residence time distribution.

However, the results in [41] indicate that it is not the residence time distribution, but it is its mean that influences teletraffic results. Therefore, classical Markovian methodology (applied also in our work) has a chance to be valid and useful in practical analysis. In [42], it is emphasized that spatial teletraffic characterization is essential for planning and dimensioning of mobile communication systems. A geographic traffic model is presented, that makes it possible to involve demographical and geographical factors into teletraffic modeling. However, the evaluation of the performance measures related to handovers is not incorporated in this model.

3.1.2 Our work

Our objective is to extend the model presented in [13] for CDMA networks. The model is applied to problems arising in CDMA networks (soft handover modeling), as well as problems arising specifically in WCDMA networks (user-plane traffic estimation on the Iur interface). Our contribution is that we allow overlapping cells, and we consider soft handover instead of hard handover.

The chapter is organized as follows. In Section 3.2.1, a road network model and a method for estimation of the vehicular traffic volume on the road network are explained. Section 3.2.2 gives a summary of all the simplifying assumptions we consider in the model. In Section 3.2.3, closed form solutions of the connection arrival rate and the residence time in a soft handover region (SHR) are given and the relations between these parameters are shown. In Section 3.2.4, it is shown how the parameters achieved that far can be used to estimate inter RNC traffic. The probability distribution of the channel occupancy time in a cell is given in Section 3.2.5. By assigning capacity to each cell, in Section 3.2.5, we obtain the blocking probability and the offered traffic load on each cell in each call class by applying a recursive method. The handover intensity in a cell is also given in Section 3.2.5. Section 3.2.6 presents a simple numerical example. Finally, conclusions are drawn in Section 3.3.

3.2 Description of the model

In [13], Nakano, Saita and Sengoku proposed a method to analyze the mobile communication traffic on a road systems model. We have chosen this model as the basis of our work, because it operates with realistic parameters, and thus results obtained with the solution of this model might be used to draw conclusions in real-life situations. In their work, the road network is covered by non-overlapping hexagonal omnicones. This cell layout assumes adjacent cells with common boundary, therefore it enables to analyze only the handovers of the so-called hard handover type. In our work, we changed this cell structure by applying circle-shaped and overlapping cells, where the

significant part of the handovers is of the soft handover type. This change considerably increases the complexity of the analysis, because a mobile can use the resources of several base stations at the same time, depending on which area it moves. As we will show in the present section, this extension proves to be useful, because many new WCDMA-related performance parameters can be obtained.

3.2.1 Road systems model and traffic flow estimation

From the viewpoint of the mobility, the road network characterizes the geographical area under study. In our model, we consider only one transport mode, let us say vehicular traffic that flows over the specified road network between different traffic sources and absorption points. Pedestrian mobility is not considered, but we believe it is straightforward to include.

The road system can be modeled by a graph, where the links of the graph represent the *streets* and the nodes of the graph can be *junctions* that represent crosses in the road system, or the so-called *centroids* that represent the origin and the destination of the traffic flows on the road system. Figure 3.2 shows an example.

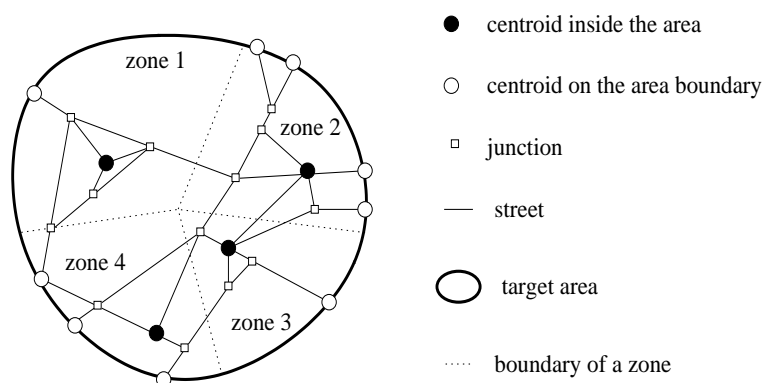


Figure 3.2: A road network in a target area

Traffic flow on the road network of the target area is surveyed in advance. At the beginning of the survey the target area is divided into *zones*. Since the division is based on geographic, population and transport characteristics, these zones usually differ from radio zones such as cells in cellular systems. The vehicular traffic volume flowing from a zone to another zone is measured for each pair of zones. The results of this survey are usually represented by *Origin-Destination tables* (O-D tables) – for example, Table 3.1 in Section 3.2.6.

The location of the centroids can have a notable influence on the accuracy of the model. Teletraffic models of wireless systems typically use a

hierarchical approach, meaning that movements are characterized at different scales, for example: within a metropolitan area, within a national area, and at the international level [36]. Within such a modeling framework, the San Francisco Bay Area has been divided into zones, and a gravity model has been applied to derive the vehicle traffic matrix among the zones [43]. The parameters of the gravity model (the movement attraction and movement generating potentials of a zone, and the gravity constant) were tuned based on measurements. It has been reported that the parameter fitting could be done with a high accuracy.

For our model, the location of the centroids and the respective O-D table could be found using the above mentioned gravity model. A trivial alternative solution would be to put centroids at the ends of each street, and fill the O-D table based on simple per-street vehicle traffic counting.

We distinguish two kind of centroids:

- each zone of the target area has its own centroid placed inside the zone, which represents the origin and the destination of the zone itself (see Figure 3.2),
- other centroids, which exist on the edge of the target area, represent traffic flows from/to the zones out of the target area. (In Figure 3.2 the target area consists of 4 zones.)

Each element OD_{ij} in the O-D matrix (see Table 3.1) is the vehicular traffic volume defined as the number of vehicles moving from centroid ci to centroid cj during a unit time (e.g., rush hour). To determine which routes these vehicles move along towards their destination, we use an *incremental traffic assignment method* (details can be found in Appendix B), but other algorithms could also be used. With this algorithm, the vehicular traffic is routed over the road network and the vehicular traffic load on all the streets are calculated.

The road network is covered by intersecting circle-shaped cells. The intersection of the circles and the links of the graph will be called *imaginary nodes*. We thus consider a new graph representing our road system, where the nodes of the graph can be centroids, junctions or imaginary nodes and the links of the graph are some parts of the original streets (links between junctions and/or centroids) divided by the imaginary nodes. The cells are overlapping, and the overlapped regions are called *soft handover regions* (SHR). For simplicity, in this Chapter we assume that a mobile under soft handover can communicate with *maximum three base stations* at the same time (a soft handover region can be in the intersection of one, two or three cells), see Figure 3.3.

3.2.2 Notations and assumptions

Figure 3.3 shows examples of the notation introduced below. (The meaning of notation $l_i^{z,r,j}$ is explained later, in Section 3.2.5.)

With some routing algorithm used on the O-D table (see Appendix B) we can get the *routes* $r = 1, \dots, R$ and the Q_r traffic volumes for each route r . In Figure 3.3 the vehicles move from left to right (from the Origin to the Destination).

Along route r we have *soft handover regions*

$$SHR_1^r, \dots, SHR_j^r, \dots, SHR_{J_r}^r,$$

listed in order of appearance (some soft handover regions can appear in the list more than once depending on the line of the route). For example, SHR_3^r is the third SHR along route r .

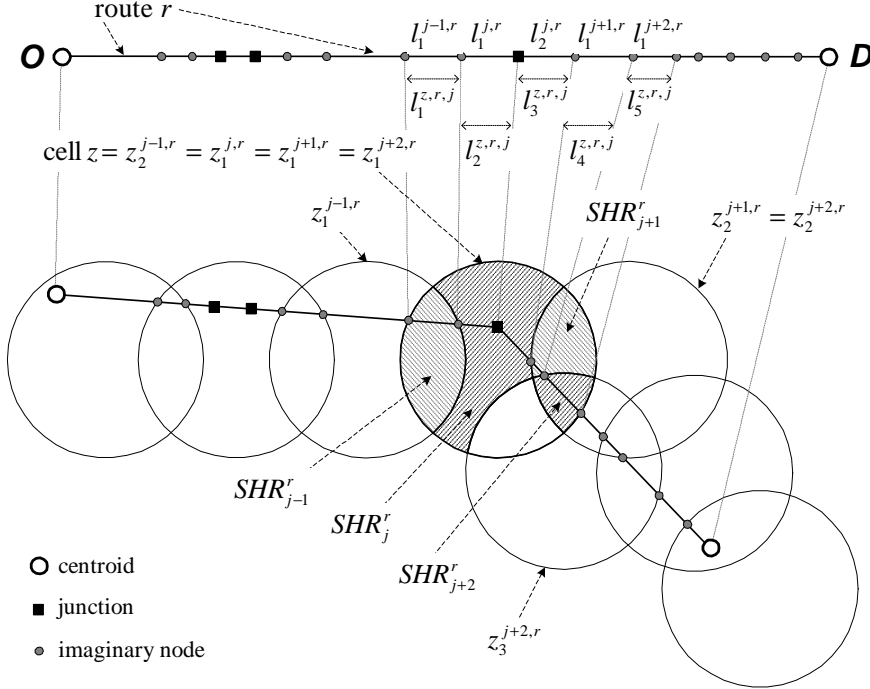


Figure 3.3: Hierarchy of sections on route r

For each SHR_j^r we define the *characteristic set* as:

$$Z_{j,r} = \{\text{cells containing } SHR_j^r\} = \{z_1^{j,r}, \dots, z_{K_{j,r}}^{j,r}\},$$

where SHR_j^r is overlapped exactly by cells $z_1^{j,r}, \dots, z_{K_{j,r}}^{j,r}$, and therefore the number of those cells is $K_{j,r} = 1, 2, \text{ or } 3$. For example, $z_1^{j,r}$ identifies

the first cell that covers SHR_j^r along route r . In Figure 3.3, we note that $z_2^{j-1,r} = z_1^{j,r} = z_1^{j+1,r} = z_1^{j+2,r}$, because the second cell that covers SHR_{j-1}^r along route r is the same as the first cell that covers SHR_j^r along route r , or the first cell that covers SHR_{j+1}^r or the first that covers SHR_{j+2}^r .

Finally, the smallest undivided section of route r is a link, which can begin and end in an imaginary node, a junction or a centroid. On the route r , SHR_j^r contains the *links*

$$l_1^{j,r}, \dots, l_i^{j,r}, \dots, l_{I_{j,r}}^{j,r}.$$

In the example of Figure 3.3, SHR_j^r contains the links $l_1^{j,r}$ and $l_2^{j,r}$, which means that in this particular case $I_{j,r} = 2$.

During the analysis we will apply some set operations. From theory of sets: $x \in X$ ($X \ni x$) means that element x is in set X (set X owns element x), $Y \subseteq X$ ($X \supseteq Y$) means that Y is a subset of set X (set X contains set Y as a subset).

Assumptions

The model parameters and the assumptions are the following:

- The road network in the target area is covered by *circle-shaped* and intersecting *cells* with radius R_0 . (A1)
- The centers of the adjacent cells are placed as the vertices of a *regular triangle*. The distance between the centers of the adjacent cells (the *base stations* are placed in the middle of the cell and have omnidirectional antennas) is D . (A2)
- Consider that we have N_{cell} cells and cell z ($z = 1, \dots, N_{cell}$) has a *capacity* of S_z channels. (A3)
- One transportation mode is considered (vehicular traffic) and the *traffic volume* flowing from a centroid to another is given in form of the *O-D table*. (A4)
- The speed of the vehicles on the streets is given by the *load-speed profile* (see Figure 3.6 later) and is assumed to be *constant* on a street ($v(x)$ stands for the speed on street x). (A5)
- The length of each street and each link is given in advance ($L(x)$ stands for the length of street or link x). (A6)
- The traffic volumes and the speed of the mobile terminals on the links connected by an imaginary node are equal. (A7)
- The distribution of vehicles on a street is *uniform*. (A8)

- We distinguish C different *call classes*. Call class c ($c = 1, \dots, C$) is characterized by the *call arrival rate* λ_c , the *holding time* h_c and the *bandwidth* b_c . (A9)
- If there are n vehicles in a cell, then the arrivals of calls of class c form a *Poisson process* with rate $n \cdot \lambda_c$ (the number of vehicles in the rush hour in a cell is sufficiently larger than the number of channels in the cell). (A10)
- The holding time of a call of class c is an *exponential* random variable with mean h_c . (A11)
- The arrival rate of calls originating *outside* the target area and then entering the target area are given in advance at the centroids on the border ($\lambda_c^0(p)$ for call class c at centroid p). (A12)
- The handover operation is of *soft handover* type (if an active mobile crosses a soft handover region border, we will say its call is “handed over” to the next soft handover region). A soft handover call is not *blocked*, if at least one of its *soft handover legs* is not blocked. (A13)

Assumptions (A1) and (A2) were taken only to simplify implementation and to ease understanding. The model works for all topologies, where the cell and SHR boundaries can be identified.

3.2.3 Soft handover region parameters

From [13] we know that the probability that a new call of class c generated on link $l_i^{j,r}$ successfully reaches the soft handover region boundary between SHR_j^r and SHR_{j+1}^r is:

$$p_c^{new}(SHR_{j+1}^r | l_i^{j,r}) = \frac{h_c}{t_i^{j,r}} e^{-\frac{T_i^{j,r}}{h_c}} (1 - e^{-\frac{t_i^{j,r}}{h_c}}), \quad (3.1)$$

for $1 \leq j < J_r$, where

$$t_i^{j,r} = \frac{L(l_i^{j,r})}{v(l_i^{j,r})} \quad (3.2)$$

is the whole traveling time on $l_i^{j,r}$ ($L(l_i^{j,r})$ is the length of link $l_i^{j,r}$ and $v(l_i^{j,r})$ is the speed on link $l_i^{j,r}$ that equals the speed on the whole street containing this link), and

$$T_i^{j,r} = \sum_{h=i+1}^{J_r} t_h^{j,r} \quad (3.3)$$

is the traveling time from the boundary of $l_i^{j,r}$ and $l_{i+1}^{j,r}$ to the boundary of SHR_j^r and SHR_{j+1}^r .

The proportion of the arrival rate of new calls of class c on $l_i^{j,r}$ that gets connected to the set of base stations $Z \subseteq \mathcal{Z}_{j,r}$ is:

$${}^Z \lambda_c^{new}(l_i^{j,r}) = \lambda_c \frac{Q_r}{v(l_i^{j,r})} L(l_i^{j,r}) {}^Z P_c(SHR_j^r) = \lambda_c Q_r t_i^{j,r} {}^Z P_c(SHR_j^r), \quad (3.4)$$

where Q_r is the vehicle traffic volume on route r (measured in [vehicles/hour]), and

$${}^Z P_c(SHR_j^r) = \prod_{z \in Z} (1 - B_c(z)) \prod_{s \in \mathcal{Z}_{j,r} \setminus Z} B_c(s), \quad (3.5)$$

where $B_c(z)$ is the *blocking probability* of c -type calls in cell z (the proportion of new call intensities $\emptyset \lambda_c^{new}(l_i^{j,r})$ corresponds to the blocked portion of calls, $Z = \emptyset$), see Section 3.2.5 for the calculation of the blocking probabilities.

Remark: The arrival rate of new calls of class c on $l_i^{j,r}$ is:

$$\lambda_c^{new}(l_i^{j,r}) = \sum_{Z \subseteq \mathcal{Z}_{j,r}} {}^Z \lambda_c^{new}(l_i^{j,r}). \quad (3.6)$$

Note that $\lambda_c^{new}(l_i^{j,r})$ is the call arrival rate corresponding to the infinite cell capacity case like in [13] (no blocking, $B_c(z) = 0$ for all cells $z = 1, \dots, N_{cell}$) and also notice that

$${}^Z \lambda_c^{new}(l_i^{j,r}) = \lambda_c^{new}(l_i^{j,r}) {}^Z P_c(SHR_j^r). \quad (3.7)$$

Suppose that a new call of class c originates in SHR_j^r from a vehicle moving along route r . Then the probability that this new call originates on $l_i^{j,r}$ is:

$$p_c(l_i^{j,r}) = \frac{\lambda_c^{new}(l_i^{j,r})}{\sum_{h=1}^{I_{j,r}} \lambda_c^{new}(l_h^{j,r})} = \frac{\lambda_c Q_r t_i^{j,r}}{\sum_{h=1}^{I_{j,r}} \lambda_c Q_r t_h^{j,r}} = \frac{t_i^{j,r}}{T_0^{j,r}}, \quad (3.8)$$

where $T_0^{j,r}$ is the traveling time through SHR_j^r on route r .

This results in the *soft handover probability of new calls* of class c from SHR_j^r (the probability that a call of class c originating in SHR_j^r successfully reaches the boundary of SHR_j^r and SHR_{j+1}^r) to be:

$$\begin{aligned} p_c^{new}(SHR_{j+1}^r | SHR_j^r) &= \sum_{i=1}^{I_{j,r}} p_c^{new}(SHR_{j+1}^r | l_i^{j,r}) p_c(l_i^{j,r}) = \\ &= \frac{h_c}{T_0^{j,r}} \sum_{i=1}^{I_{j,r}} e^{-\frac{T_i^{j,r}}{h_c}} (1 - e^{-\frac{t_i^{j,r}}{h_c}}). \end{aligned} \quad (3.9)$$

Notice that the above expression on the right is a *telescopic sum* (an expanded sum, where the terms have alternating signs following each other

and with the exception of the first and the last term, the others sum up to zero – the middle neighboring terms cancel each other out) with $T_{i-1}^{j,r} = T_i^{j,r} + t_i^{j,r}$ and with $T_{I_j,r}^{j,r} = 0$, thus we get

$$p_c^{new}(SHR_{j+1}^r | SHR_j^r) = \frac{h_c}{T_0^{j,r}}(1 - e^{-\frac{T_0^{j,r}}{h_c}}), \text{ for } 1 \leq j < J_r. \quad (3.10)$$

Next, consider the *soft handover probability of soft handover calls* of class c from SHR_j^r , i.e., the probability that while moving on route r the call of class c enters SHR_j^r (so it is “handed over” from SHR_{j-1}^r to SHR_j^r) and successfully reaches the boundary of SHR_j^r and SHR_{j+1}^r (it is “handed over” to SHR_{j+1}^r) and that is:

$$p_c^{ho}(SHR_{j+1}^r | SHR_j^r) = e^{-\frac{T_0^{j,r}}{h_c}}, \text{ for } 1 < j < J_r, \quad (3.11)$$

because of the assumptions (A9),(A10),(A11).

We still need the boundary conditions for the soft handover probabilities. Consider an ongoing call that enters the target area at the centroid in SHR_1^r and moves from this centroid to the destination along route r . The problem is that the centroid is usually found inside the area rather than on the boundary of the cells of $\mathcal{Z}_{1,r}$, thus the boundary conditions for the soft handover calls entering these cells are not straight-forward. Therefore, considering these types of handover calls, we assume that the first link of route r is “lengthened backwards” to reach the boundary of one of the cells of $\mathcal{Z}_{1,r}$. If the lengthened part is d_r long, the traveling time of the new first street is:

$$T_{0'}^{1,r} = \frac{d_r}{v(l_1^{1,r})} + T_0^{1,r}. \quad (3.12)$$

Thus we have the boundary condition:

$$p_c^{ho}(SHR_2^r | SHR_1^r) = e^{-\frac{T_{0'}^{1,r}}{h_c}}. \quad (3.13)$$

On the other hand, the calls in $SHR_{J_r}^r$ are not “handed over” to further soft handover regions, therefore

$$p_c^{new}(SHR_{J_r+1}^r | SHR_{J_r}^r) = 0, \quad (3.14)$$

$$p_c^{ho}(SHR_{J_r+1}^r | SHR_{J_r}^r) = 0, \quad (3.15)$$

and thus the formulas for the soft handover probabilities are completed.

Let us define the *Z-set call arrival rate of new calls* of class c in SHR_j^r on route r as the proportion of the call arrival rate of new calls of class c in SHR_j^r on route r that get connected to the base station set $Z \subseteq \mathcal{Z}_{j,r}$. The *Z-set* is similar to the “active set” in CDMA systems. The active

set contains the radio cells the mobile could be connected to, because the measured pilot signal strength is sufficient. The Z -set contains the cells the mobile is actually connected to. The Z -set call arrival rate of new calls can be calculated as:

$${}^Z\lambda_c^{new}(SHR_j^r) = \sum_{i=1}^{I_{j,r}} \lambda_c Q_r t_i^{j,r} {}^Z P_c(SHR_j^r). \quad (3.16)$$

Therefore we have:

$${}^Z\lambda_c^{new}(SHR_j^r) = \lambda_c Q_r T_0^{j,r} {}^Z P_c(SHR_j^r), \quad \text{for each } 1 \leq j \leq J_r. \quad (3.17)$$

The *call arrival rate of new calls* of class c in SHR_j^r is:

$$\lambda_c^{new}(SHR_j^r) = \sum_{Z \subseteq \mathcal{Z}_{j,r}} {}^Z\lambda_c^{new}(SHR_j^r) = \lambda_c Q_r T_0^{j,r}. \quad (3.18)$$

Remark: We have the following simple relationship for the Z -set call intensities of the newly initiated calls:

$${}^Z\lambda_c^{new}(SHR_j^r) = \lambda_c^{new}(SHR_j^r) {}^Z P_c(SHR_j^r). \quad (3.19)$$

Let us define the *Z -set call arrival rate of soft handover calls* of class c entering SHR_j^r (denoted by ${}^Z\lambda_c^{ho}(SHR_j^r)$) as the proportion of the call arrival rate of soft handover calls of class c entering SHR_j^r that get connected to the base station set $Z \subseteq \mathcal{Z}_{j,r}$. It consists of two parts, namely the newly initiated calls of class c in SHR_{j-1}^r and then “handed over” to SHR_j^r , and the calls of class c “handed over” from the previous soft handover region to SHR_{j-1}^r and then “handed over” further to SHR_j^r for $1 < j \leq J_r$.

While reading the next four paragraphs, keep in mind that set Z is the base station set, and by definition $Z \subseteq \mathcal{Z}_{j,r}$. (For example if $\mathcal{Z}_{j,r} = \{1, 2\}$, then $Z \subseteq \mathcal{Z}_{j,r}$ means that $Z \in \{\emptyset, \{1\}, \{2\}, \{1, 2\}\}$.) Furthermore, note that we try to describe the calculation of the Z -set call arrival rate of soft handover calls as general as possible, because it is useful if one wants to use different assumptions than our (A1) and (A2).

If SHR_j^r is covered by more cells than SHR_{j-1}^r , then the Z -set can *remain the same*, if the new cells block the connection setup request, or it can be *extended* with new cells, if at least one of the new cells does not block the request. Using the notations, for $Z \not\subseteq \mathcal{Z}_{j,r} \setminus \mathcal{Z}_{j-1,r}$, if $\mathcal{Z}_{j-1,r} \subseteq \mathcal{Z}_{j,r}$, ${}^Z\lambda_c^{ho}(SHR_j^r)$ can be calculated as:

$$\begin{aligned} {}^Z\lambda_c^{ho}(SHR_j^r) &= p_c^{new}(SHR_j^r | SHR_{j-1}^r) {}^{Z \cap \mathcal{Z}_{j-1,r}}\lambda_c^{new}(SHR_{j-1}^r) \times \\ &\times \prod_{z \in \mathcal{Z}_{j,r} \setminus \mathcal{Z}_{j-1,r}} \{(1 - B_c(z))\mathcal{I}_{\{z \in Z\}} + B_c(z)\mathcal{I}_{\{z \notin Z\}}\} + \\ &+ p_c^{ho}(SHR_j^r | SHR_{j-1}^r) {}^{Z \cap \mathcal{Z}_{j-1,r}}\lambda_c^{ho}(SHR_{j-1}^r) \times \\ &\times \prod_{z \in \mathcal{Z}_{j,r} \setminus \mathcal{Z}_{j-1,r}} \{(1 - B_c(z))\mathcal{I}_{\{z \in Z\}} + B_c(z)\mathcal{I}_{\{z \notin Z\}}\}, \quad (3.20) \end{aligned}$$

because for the $Z \subseteq \mathcal{Z}_{j,r}$ that contains a cell from the tighter characteristic set $\mathcal{Z}_{j-1,r}$, the Z -set soft handover calls consist of exactly the $Z \cap \mathcal{Z}_{j-1,r}$ -set newly initiated and soft handover calls of SHR_{j-1}^r that successfully reach SHR_j^r and get blocked exactly by the required new cells in the looser characteristic set $\mathcal{Z}_{j,r}$.

The example in Figure 3.4 refers to the above case and helps to understand notations.

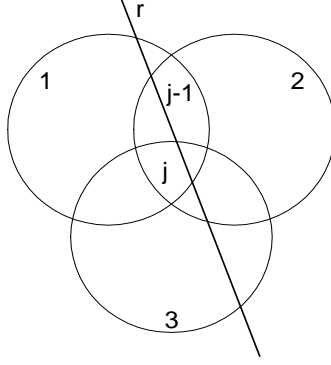


Figure 3.4: Example explaining the notations used in Eq.(3.20)

In this example, the notations used in Eq.(3.20) have the following meaning:

- $\mathcal{Z}_{j,r} = \{1, 2, 3\}$,
- $\mathcal{Z}_{j-1,r} = \{1, 2\}$,
- $\mathcal{Z}_{j,r} \setminus \mathcal{Z}_{j-1,r} = \{3\}$,
- $Z \in \{\{1\}, \{2\}, \{1, 2\}, \{1, 3\}, \{2, 3\}, \{1, 2, 3\}\}$ and
- $Z \cap \mathcal{Z}_{j-1,r} \in \{\{1\}, \{2\}, \{1, 2\}\}$.

The Z -set can not be *entirely changed* to a new set, i.e., for $Z \subseteq \mathcal{Z}_{j,r} \setminus \mathcal{Z}_{j-1,r}$, if $\mathcal{Z}_{j-1,r} \subseteq \mathcal{Z}_{j,r}$, we get:

$${}^Z\lambda_c^{ho}(SHR_j^r) = 0, \quad (3.21)$$

because for the $Z \subseteq \mathcal{Z}_{j,r}$ that does not contain any cell from the tighter characteristic set $\mathcal{Z}_{j-1,r}$, the Z -set soft handover calls do not exist (there are no calls in the previous SHR that can be “handed over” to SHR_j^r in such way).

If SHR_j^r is covered by less cells than SHR_{j-1}^r , then the Z -set can be reduced or emptied. If the Z -set is *reduced but not emptied*, i.e., for $Z \neq \emptyset$,

if $\mathcal{Z}_{j-1,r} \supseteq \mathcal{Z}_{j,r}$, we get:

$$\begin{aligned} {}^Z\lambda_c^{ho}(SHR_j^r) &= p_c^{new}(SHR_j^r | SHR_{j-1}^r) \sum_{Z \subseteq S \subseteq \mathcal{Z}_{j-1,r}} {}^S\lambda_c^{new}(SHR_{j-1}^r) + \\ &+ p_c^{ho}(SHR_j^r | SHR_{j-1}^r) \sum_{Z \subseteq S \subseteq \mathcal{Z}_{j-1,r}} {}^S\lambda_c^{ho}(SHR_{j-1}^r), \end{aligned} \quad (3.22)$$

because for the $Z \subseteq \mathcal{Z}_{j,r}$ not empty sets, the Z -set soft handover calls consist of all the S -set newly initiated and soft handover calls of the previous SHR (of looser characteristic set $\mathcal{Z}_{j-1,r}$) for which S is looser than Z that successfully reach the boundary of SHR_{j-1}^r and SHR_j^r (no blocking can take place for calls “handed over” to SHR_j^r this way).

If the Z -set is *emptied*, i.e., for $Z = \emptyset$, if $\mathcal{Z}_{j-1,r} \supseteq \mathcal{Z}_{j,r}$, we get:

$$\begin{aligned} {}^\emptyset\lambda_c^{ho}(SHR_j^r) &= p_c^{new}(SHR_j^r | SHR_{j-1}^r) \times \\ &\times \sum_{\emptyset \neq S \subseteq \mathcal{Z}_{j-1,r} \setminus \mathcal{Z}_{j,r}} {}^S\lambda_c^{new}(SHR_{j-1}^r) + p_c^{ho}(SHR_j^r | SHR_{j-1}^r) \times \\ &\times \sum_{\emptyset \neq S \subseteq \mathcal{Z}_{j-1,r} \setminus \mathcal{Z}_{j,r}} {}^S\lambda_c^{ho}(SHR_{j-1}^r). \end{aligned} \quad (3.23)$$

because for the $Z = \emptyset$, the blocked soft handover calls consist of all those non-blocked S -set newly initiated and soft handover calls arriving from the previous SHR, which successfully reach SHR_j^r (no new blocking again), such that S does not contain a cell from the tighter characteristic set $\mathcal{Z}_{j,r}$.

For the completion of this recursive formula we need to give the initial condition and that can be:

$${}^Z\lambda_c^{ho}(SHR_1^r) = {}^Z\lambda_c^0(SHR_1^r) \frac{Q_r}{\sum_{\text{route } h \text{ starts in } SHR_1^r} Q_h}, \quad (3.24)$$

for $Z \subseteq \mathcal{Z}_{1,r}$, where ${}^Z\lambda_c^0(SHR_1^r) = \lambda_c^0(p_1^r) {}^ZP_c(SHR_1^r)$ and p_1^r is the first centroid of route r . Here I have assumed that the arrival rate of calls coming from outside the target area to SHR_1^r is divided among the routes starting in the particular soft handover region in the proportion of the traffic volumes on these routes.

Remark: The call arrival rate of soft handover calls of class c entering SHR_j^r is:

$$\lambda_c^{ho}(SHR_j^r) = \sum_{Z \subseteq \mathcal{Z}_{j,r}} {}^Z\lambda_c^{ho}(SHR_j^r), \quad (3.25)$$

and it does not provide us with such a simple relation like Eq.(3.19) for the newly initiated call intensities (an analogous to Eq.(3.19) for the soft handover call intensities does not hold, for example because of Eq.(3.21)).

Remark: We derive the Z -set call arrival rates to be able to calculate inter RNC traffic (we need the 2-leg and 3-leg proportion of the call arrival intensities in the soft handover regions to calculate the user-plane traffic on the Iur interface, see Section 3.2.4). We also need to know the proportion of the blocked calls in some soft handover regions when calculating the offered traffic load for a cell, see Section 3.2.5.

We define the *mean SHR residence time of call class c* as the mean value of the holding time of the call of class c being in a specific SHR until the call is terminated or until it reaches the border of (is “handed over” to) the next SHR. The mean SHR residence time of call class c originated in SHR_j^r (from [13]) is:

$$h_c^{new}(SHR_j^r) = h_c - \frac{h_c^2}{T_0^{j,r}} \sum_{i=1}^{I_{j,r}} e^{-\frac{T_i^{j,r}}{h_c}} (1 - e^{-\frac{t_i^{j,r}}{h_c}}), \quad (3.26)$$

and since it includes a telescopic sum with:

$$T_{i-1}^{j,r} = T_i^{j,r} + t_i^{j,r}, \quad (3.27)$$

we get:

$$h_c^{new}(SHR_j^r) = h_c - \frac{h_c^2}{T_0^{j,r}} (1 - e^{-\frac{T_0^{j,r}}{h_c}}), \quad \text{for } 1 \leq j < J_r. \quad (3.28)$$

The *mean SHR residence time of call class c “handed over”* from SHR_{j-1}^r to SHR_j^r is:

$$h_c^{ho}(SHR_j^r) = h_c (1 - e^{-\frac{T_0^{j,r}}{h_c}}), \quad \text{for } 1 < j < J_r. \quad (3.29)$$

The boundary conditions are:

$$h_c^{ho}(SHR_1^r) = h_c (1 - e^{-\frac{T_0^{1,r}}{h_c}}), \quad (3.30)$$

because of the “backward lengthening” in the first cells of the route r , and

$$h_c^{new}(SHR_{J_r}^r) = h_c, \quad (3.31)$$

$$h_c^{ho}(SHR_{J_r}^r) = h_c, \quad (3.32)$$

because the calls in the last cell of the route are not “handed over” to further SHRs.

Remark: Note that there is a simple relationship between the soft handover probability and the mean SHR residence time:

$$h_c^{new}(SHR_j^r) = h_c (1 - p_c^{new}(SHR_{j+1}^r | SHR_j^r)), \quad (3.33)$$

$$h_c^{ho}(SHR_j^r) = h_c (1 - p_c^{ho}(SHR_{j+1}^r | SHR_j^r)), \quad (3.34)$$

where Eq.(3.33) stands because of Eq.(3.10) and Eq.(3.28); Eq.(3.14) and Eq.(3.31) and Eq.(3.34) stands because of Eq.(3.11) and Eq.(3.29); Eq.(3.13) and Eq.(3.30); Eq.(3.15) and Eq.(3.32), respectively.

We can calculate the Z -set traffic load for $Z \subseteq \mathcal{Z}_{j,r}$ for the call class c for each soft handover region SHR_j^r (defined as the proportion of the traffic load that is induced by the calls that get connected to the base station set $Z \subseteq \mathcal{Z}_{j,r}$) by:

$$\begin{aligned} {}^Z Load_c(SHR_j^r) &= {}^Z \lambda_c^{new}(SHR_j^r) h_c^{new}(SHR_j^r) + \\ &+ {}^Z \lambda_c^{ho}(SHR_j^r) h_c^{ho}(SHR_j^r), \end{aligned} \quad (3.35)$$

and thus the l -leg offered traffic load for any area (some soft handover regions together) can be calculated as the sum of those Z -set offered traffic loads of soft handover regions in the area for which $|Z| = l$, for $l = 0, 1, 2, 3$.

3.2.4 User-plane traffic estimation on the Iur interface

Let $RNC_1 \subseteq \{1, \dots, N_{cell}\}$ and $RNC_2 \subseteq \{1, \dots, N_{cell}\}$ represent two disjoint cell sets that contain cells of two RNC-s ($RNC_1 \cap RNC_2 = \emptyset$ and the corresponding cells are connected to RNC_1 and RNC_2 respectively). If an SHR is covered by two cells, which are in different cell sets, the traffic in that SHR appears on the Iur interface. The three-leg traffic gives one or two-legged traffic on the Iur interface depending on the route structure in the corresponding SHR. Using Eq.(3.35), we can get a lower bound for the traffic generated on the Iur interface between the two RNC-s:

$$\begin{aligned} I_{ur}^{lower}(RNC_1, RNC_2) &= \\ &= \sum_{z_1 \in RNC_1, z_2 \in RNC_2} \sum_{\mathcal{Z}_{j,r} \ni z_1, z_2} \sum_{Z \subseteq \mathcal{Z}_{j,r}} \sum_{c=1}^C {}^Z Load_c(SHR_j^r) \times \\ &\times \left\{ \mathcal{I}_{\{z_1, z_2 \in Z, |Z|=2\}} + 1 \cdot \mathcal{I}_{\{z_1, z_2 \in Z, |Z|=3\}} \right\}, \end{aligned} \quad (3.36)$$

and an upper bound

$$\begin{aligned} I_{ur}^{upper}(RNC_1, RNC_2) &= \\ &= \sum_{z_1 \in RNC_1, z_2 \in RNC_2} \sum_{\mathcal{Z}_{j,r} \ni z_1, z_2} \sum_{Z \subseteq \mathcal{Z}_{j,r}} \sum_{c=1}^C {}^Z Load_c(SHR_j^r) \times \\ &\times \left\{ \mathcal{I}_{\{z_1, z_2 \in Z, |Z|=2\}} + 2 \cdot \mathcal{I}_{\{z_1, z_2 \in Z, |Z|=3\}} \right\}, \end{aligned} \quad (3.37)$$

where we summed up all those Z -set traffic loads of the soft handover regions, which are both in RNC_1 and RNC_2 (see the first two summations) for sets $Z \subseteq \mathcal{Z}_{j,r}$ that contain a cell of each RNC ($z_1 \in RNC_1, z_2 \in RNC_2$), so the proper two and three-leg traffic loads (see the third summation and the indicator functions) for each call class (see the fourth summation). Note

that we need the blocking probabilities here implicitly in the Z -set traffic loads through the Z -set call arrival rates.

Notice that we could determine the exact value of the traffic on the Iur interface, but here we omit it because it is contagious to formulate it.

3.2.5 Cell parameters

Distribution of the channel occupancy time.

Consider the mean value of the cell residence time that is usually called *channel occupancy time* (it is the time during which an active call holds a channel in a cell).

Similarly to Eq.(3.28), the *mean channel occupancy time* of c class calls originating in cell z on route r is:

$$h_c^{new}(z^{(r,j)}) = h_c - \frac{h_c^2}{T_0^{z,r,j}}(1 - e^{-\frac{T_0^{z,r,j}}{h_c}}), \quad (3.38)$$

where $z^{(r,j)}$ is the cell identified by the cell identifier z , and (r, j) means that the computation is done on route r over the section, which is covered by cell z and contains SHR_j^r . (The index j is needed, because it is possible that route r goes through the area of cell z more than once.) $T_0^{z,r,j}$ is the traveling time through cell z on the section of route r that contains SHR_j^r :

$$T_0^{z,r,j} = \sum_{i=1}^{I_{z,r,j}} t_i^{z,r,j} = \sum_{i=1}^{I_{z,r,j}} \frac{L(l_i^{z,r,j})}{v(l_i^{z,r,j})}, \quad (3.39)$$

(see Figure 3.3). Similarly to Eq.(3.29), the *mean channel occupancy time* of c class calls handed over to cell z is:

$$h_c^{ho}(z^{(r,j)}) = h_c(1 - e^{-\frac{T_0^{z,r,j}}{h_c}}). \quad (3.40)$$

We can also derive the *distribution of the channel occupancy time* in the cell for the newly initiated calls of class c as follows:

Let $\xi_{(z)}^{r,j}$ be the random variable of time required for a trip from the call originating point to the boundary of cell z on the section of route r that contains SHR_j^r . From assumption (A8) and denoting the random variable of the holding time of a class c call by τ_c , we can get the probability of $\{\xi_{(z)}^{r,j} < t\}$ given that $\{\tau_c > t\}$ and given that the class c call is generated on $l_i^{z,r,j}$ as:

$$p_c(\xi_{(z)}^{r,j} < t \mid \tau_c > t, l_i^{z,r,j}) = \begin{cases} 0 & , t \leq T_i^{z,r,j} \\ \frac{t - T_i^{z,r,j}}{t_i^{z,r,j}} & , T_i^{z,r,j} < t \leq T_{i-1}^{z,r,j} \\ 1 & , T_{i-1}^{z,r,j} < t, \end{cases} \quad (3.41)$$

where

$$T_i^{z,r,j} = \sum_{h=i+1}^{I_{z,r,j}} t_h^{z,r,j}, \quad (3.42)$$

for each $i = 1, \dots, I_{z,r,j} - 1$ and $T_{I_{z,r,j}}^{z,r,j} = 0$.

On the other hand (similarly to the soft handover regions), the probability that a class c call is generated on $l_i^{z,r,j}$ is equal to

$$p_c(l_i^{z,r,j}) = \frac{t_i^{z,r,j}}{T_0^{z,r,j}}, \quad (3.43)$$

and thus unconditioning, we get

$$\begin{aligned} Pr\{\xi_{(z)}^{r,j} < t \mid \tau_c > t\} &= \sum_{i=1}^{I_{z,r,j}} p_c(\xi_{(z)}^{r,j} < t \mid \tau_c > t, l_i^{z,r,j}) p_c(l_i^{z,r,j}) = \\ &= \frac{t}{T_0^{z,r,j}} - \sum_{i=1}^{I_{z,r,j}} \mathcal{I}_{\{T_i^{z,r,j} < t \leq T_{i-1}^{z,r,j}\}} \frac{T_i^{z,r,j}}{T_0^{z,r,j}} + \sum_{i=1}^{I_{z,r,j}} \mathcal{I}_{\{T_{i-1}^{z,r,j} < t\}} \frac{t_i^{z,r,j}}{T_0^{z,r,j}} = \frac{t}{T_0^{z,r,j}}. \end{aligned} \quad (3.44)$$

Introducing the density function of the above distribution, $g_c^{z,r,j}(t) = \frac{1}{T_0^{z,r,j}}$, we can get for the $\tau_c^{z,r,j}$ random variable (the channel occupancy time in cell z on route r):

$$Pr\{\tau_c^{z,r,j} > t\} = \int_0^{T_0^{z,r,j}-t} g_c^{z,r,j}(s) ds \int_t^\infty \frac{1}{h_c} e^{-\frac{h}{h_c}} dh = \frac{T_0^{z,r,j} - t}{T_0^{z,r,j}} e^{-\frac{t}{h_c}}. \quad (3.45)$$

Thus the probability distribution of $\tau_c^{z,r,j}$ is:

$$Pr\{\tau_c^{z,r,j} < t\} = 1 - \frac{T_0^{z,r,j} - t}{T_0^{z,r,j}} e^{-\frac{t}{h_c}}, \quad (3.46)$$

and the $f_c^{z,r,j}$ probability density function of $\tau_c^{z,r,j}$ is:

$$f_c^{z,r,j}(t) = \frac{T_0^{z,r,j} - t + h_c}{T_0^{z,r,j} h_c} e^{-\frac{t}{h_c}}. \quad (3.47)$$

The *mean channel occupancy time* for the new c class calls in cell z on route r is then:

$$h_c^{new}(z^{(r,j)}) = \int_0^{T_0^{z,r,j}} t f_c^{z,r,j}(t) dt = h_c - \frac{h_c^2}{T_0^{z,r,j}} (1 - e^{-\frac{T_0^{z,r,j}}{h_c}}), \quad (3.48)$$

and this coincides with Eq.(3.38).

Blocking probability and offered traffic load

The *offered traffic load* for a cell is composed of all the newly initiated call intensities in the cell, the newly initiated and handover call intensities in the SHR preceding the cell on some route and successfully reaching the boundary of the cell and finally the call arrival rates from outside the target area to the cell each multiplied by the corresponding mean channel occupancy times:

$$\begin{aligned}
Load_c(z) = & \sum_{Z_{j,r} \ni z} h_c^{new}(z^{(r,j)}) \lambda_c^{new}(SHR_j^r) + \sum_{Z_{j,r} \ni z \notin Z_{j-1,r}} h_c^{ho}(z^{(r,j)}) \times \\
& \times \{p_c^{new}(SHR_j^r | SHR_{j-1}^r) (\lambda_c^{new}(SHR_{j-1}^r) - \emptyset \lambda_c^{new}(SHR_{j-1}^r)) + \\
& + p_c^{ho}(SHR_j^r | SHR_{j-1}^r) (\lambda_c^{ho}(SHR_{j-1}^r) - \emptyset \lambda_c^{ho}(SHR_{j-1}^r))\} + \\
& + \sum_{Z_{0,r} \ni z} h_c^{ho}(z^{(r,j)}) \lambda_c^0(p_1^r), \tag{3.49}
\end{aligned}$$

Having the offered traffic load, we can calculate the *blocking probabilities* of each call class c in each cell z by the *multirate Erlang B formula* (see [44]):

$$B_c(z) = \frac{\sum_{s=0}^{S_z} q(s)}{\sum_{s=S_z-b_c+1}^{S_z} q(s)}, \tag{3.50}$$

where b_c is the equivalent power of the calls of class c , S_z is the capacity of the cell z and the auxiliary function $q()$ is given by the following recursion:

$$q(s) = \begin{cases} 1 & , \text{ for } s = 0 \\ s^{-1} \sum_{c=1}^C Load_c(z) b_c q(s - b_c) & , \text{ for } 0 < s \leq S_z. \end{cases} \tag{3.51}$$

Unfortunately, the offered traffic load and the blocking probability are not independent of each other, therefore we can not get them explicitly. We have a simultaneous system of equations for the parameters of interest that we are going to solve numerically using the following straight-forward *iterative algorithm*:

Step # 0: Calculate the *soft handover probabilities* for the new and the soft handover calls for each call class $c = 1, \dots, C$ in each soft handover region SHR_j^r , $j = 1, \dots, J_r$ on each route $r = 1, \dots, R$ by Eq.(3.10), Eq.(3.11), Eq.(3.13), Eq.(3.14) and Eq.(3.15) and the *mean SHR residence time* by equations Eq.(3.28), Eq.(3.29), Eq.(3.30), Eq.(3.31) and Eq.(3.32) or by equations Eq.(3.33) and Eq.(3.34) knowing the soft handover probabilities already – all these will not change any more.

Step # 1: Calculate the *call intensities* for each call class $c = 1, \dots, C$ in each soft handover region SHR_j^r , $j = 1, \dots, J_r$ on all routes $r = 1, \dots, R$ by Eq.(3.17), Eq.(3.20), Eq.(3.21), Eq.(3.22), Eq.(3.23) and Eq.(3.24) (the very first step is calculated assuming infinite cell capacities that is each blocking probability $B_c(z)$ equals 0 for all cells $z = 1, \dots, N_{cell}$ and for each call class $c = 1, \dots, C$).

Step # 2: Calculate the *Z-set traffic load* for each $Z \subseteq \mathcal{Z}_{j,r}$ for call class $c = 1, \dots, C$ in each soft handover region SHR_j^r , $j = 1, \dots, J_r$ on each route $r = 1, \dots, R$ by Eq.(3.35) and the *offered traffic load* for each cell $z = 1, \dots, N_{cell}$ by Eq.(3.49) respectively.

Step # 3: Calculate the *blocking probabilities* for each call class $c = 1, \dots, C$ and for each cell $z = 1, \dots, N_{cell}$ by using Eq.(3.51) and Eq.(3.50).

Repeat: Steps #1, #2 and #3 until a predefined stopping condition is not satisfied (some stopping conditions: given number of iterations; soft handover probabilities and/or offered loads and/or blocking probabilities do not change more than some predefined small positive real number(s)).

Remark: The simultaneous system of equations define an $f : \mathbb{R}^M \rightarrow \mathbb{R}^M$ continuous function, where M is the number of parameters (as a function of the system parameters). We can consider function f as a continuous, bounded function from a bounded M -dimensional space (the probabilities are bounded and other parameters can be normalized for example by the sum of the corresponding parameters). Therefore, this function f has a fixed point $f(x) = x$ (this is exactly the claim of the *Brouwer's fixed point theorem*, see [45]) and this makes our iterative algorithm reasonable.

Soft handover intensities

We can derive a very important cell parameter, the *soft handover intensity* (mean number of handover requests in the cell (to the cell)) that consists of the non-blocked newly initiated calls in the SHR region preceding the cell on some route, the non-blocked soft handover calls in the same SHR that reach the boundary of the cell and the call arrival rates from outside the target area (if the cell is on the boundary of it):

$$\begin{aligned}
{}^{in} \lambda_c^{ho}(z) = & \sum_{\mathcal{Z}_{j,r} \ni z \notin \mathcal{Z}_{j-1,r}} \{ p_c^{new}(SHR_j^r | SHR_{j-1}^r) (\lambda_c^{new}(SHR_{j-1}^r) - \\
& - \varnothing \lambda_c^{new}(SHR_{j-1}^r)) + p_c^{ho}(SHR_j^r | SHR_{j-1}^r) (\lambda_c^{ho}(SHR_{j-1}^r) - \\
& - \varnothing \lambda_c^{ho}(SHR_{j-1}^r)) \} + \sum_{\mathcal{Z}_{0,r} \ni z} \lambda_c^0(p_1^r). \quad (3.52)
\end{aligned}$$

3.2.6 A numerical example

In this section an example is provided, which demonstrates an application of the model. When deciding which cell should be served by which RNC, one

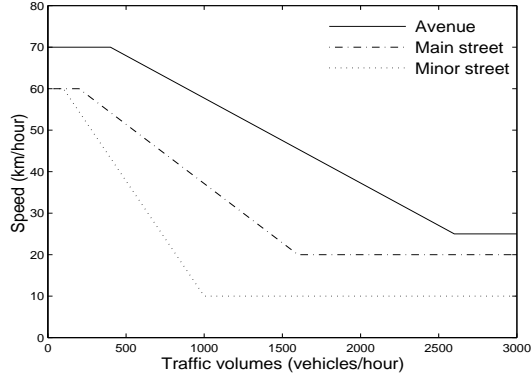


Figure 3.6: Load-speed profiles

Table 3.2: Class parameters

| | h_c (hour) | λ_c (call/hour/user) | b_c (voice equivalent) |
|-------------------|--------------|------------------------------|--------------------------|
| voice ($c = 1$) | 0.025 | 0.8 | 1 |
| data ($c = 2$) | 0.075 | 0.2 | 3 |
| video ($c = 3$) | 0.25 | 0.002 | 8 |

We have two scenarios for calculating the traffic on the Iur interface (two pairs of RNC-s):

- *Scenario 1*: $RNC_1 = \{0, 1, 2, 3\}$, $RNC_2 = \{4, 5, 6, 7, 8, 9\}$ (vertical cut),
- *Scenario 2*: $RNC_1 = \{0, 3, 4, 7\}$, $RNC_2 = \{1, 2, 5, 6, 8, 9\}$ (horizontal cut).

The Iur traffic estimators for the two scenarios are:

- Scenario 1: $6.55 \leq I_{ur}(RNC_1, RNC_2) \leq 8.06$,
- Scenario 2: $I_{ur}(RNC_1, RNC_2) = 0.49$.

Not surprisingly, the offered user-plane traffic on the Iur interface is much larger in the case of the “vertical cut” scenario (the vertical cut goes through the streets $[c0, c3]$ and $[c1, c2]$ with heavier traffic flowing on them). In Table 3.3, we have summarized the offered traffic load, blocking probabilities and soft handover intensities for the cells.

3.3 Conclusion

The traffic analysis method presented in [13] has been derived for a model considering soft handovers by introducing more detailed parameters (Eq.(3.17), Eq.(3.20), Eq.(3.21), Eq.(3.22) and Eq.(3.23)) for finite capacity cells.

The distribution and density of the channel occupancy time for a cell in Eq.(3.46) and Eq.(3.47) has been derived. An iterative algorithm has also been detailed to solve our system of simultaneous non-linear equations (the

Table 3.3: Cell parameters of the example

| cells | $L_2(z)$ | $B_2(z)$ | ${}^{in}\lambda_2^{ho}(z)$ |
|---------|----------|----------|----------------------------|
| $z = 0$ | 3.68 | 0.54 | 15.45 |
| $z = 1$ | 0.48 | 0.01 | 2.06 |
| $z = 2$ | 3.34 | 0.48 | 13.24 |
| $z = 3$ | 3.39 | 0.58 | 10.02 |
| $z = 4$ | 2.01 | 0.33 | 15.43 |
| $z = 5$ | 2.02 | 0.31 | 12.58 |
| $z = 6$ | 3.36 | 0.56 | 7.82 |
| $z = 7$ | 3.56 | 0.52 | 15.59 |
| $z = 8$ | 0.31 | 0.003 | 2.52 |
| $z = 9$ | 3.80 | 0.54 | 13.64 |

offered traffic load, needed for calculating the blocking probabilities of the cells, has been expressed with the detailed parameters – Eq.(3.49)). The traffic load for soft handover regions has been derived enabling the calculation of the overhead caused by the soft handover traffic in the transport network. The traffic load for soft handover regions were used to calculate the offered traffic on the Iur interface (Eq.(3.35), Eq.(3.36) and Eq.(3.37)).

We have shown on a simple numerical example that there can be significant difference in terms of inter RNC traffic among different RNC structures. Therefore, it is useful to solve a clustering problem for the grouping of cells to RNC-s optimizing (minimizing) the inter RNC traffic. For example, this way our results and our software could be a part of a network planning procedure.

Chapter 4

Real-time VP bandwidth control

Vidács has developed a buffer monitoring method, which is able to accurately measure the cell loss ratio (CLR) in the output buffer of an ATM switch for both short and long-range dependent traffic [21].

In this chapter the above buffer monitoring method is applied to Virtual Path (VP) bandwidth control. For this purpose, the notion of state-space representation of a single server queue is introduced, and Bayesian regression analysis is applied to estimate the state variable of that system. Our approach does not require any models describing the statistics of the traffic other than the asymptotic behavior of the CLR. The applicability of the VP bandwidth control is thoroughly discussed and the method is evaluated by extensive simulations.

4.1 Introduction

In Asynchronous Transfer Mode (ATM) networks, cell loss ratio and cell delay variation are considered to be the major quality of service (QoS) factors. Cell loss and cell delay mainly occur in the output buffers of the network nodes. Traditional queuing analyses based on parametric models have the drawback, that since the traffic pattern of ATM streams may be quite complex (e.g., when large number of heterogeneous sources are multiplexed), the appropriate statistical model seems to be difficult to identify. In order to avoid this difficulty, traffic control methods based on real time measurements have been proposed recently [46, 47, 48, 49]. In particular, the flexibility of VP bandwidth is a characteristic of ATM [50], and network operators try to use this characteristic to enforce the adaptability to traffic fluctuations.

If the actual cell loss performance of an ATM output buffer could be determined in real time, the rate of the server (that is, the VP bandwidth) could be adjusted such that the cell loss would be smaller than a pre-determined threshold. In [51], Shioda and Saito presented a method for

estimating the cell loss ratio in real time, and applied it to VP bandwidth estimation and call admission control. They utilized the large deviation result of Glynn and Whitt [52], that the CLR decays exponentially as the buffer size increases. In this case the CLR (in the regime of large buffers) is characterized by two parameters, the so called asymptotic constant β and the asymptotic decay rate η [53]. As a result, an algorithm for estimating these coefficients on an on-line basis from buffer measurements was proposed.

The objective of this chapter is to develop an algorithm (by reformulating the method proposed in [51]) for real time VP bandwidth estimation, which works also in the case of long-range dependent traffic. The most relevant parts of reference [51] are copied to Appendix C.

The inherent correlations of a long-range dependent (LRD) stochastic process decay hyperbolically as the lag increases. As a result, the autocorrelation function is nonsummable. This nonsummability captures the intuition behind long-range dependence, namely, that while high-lag correlations are all individually small, the cumulative effect is of importance and gives rise to features which are drastically different from those of the more conventional, i.e., short-range dependent (SRD) processes. The latter are characterized by an exponential decay of the correlations, resulting in a summable autocorrelation function. LRD is characterized by the Hurst parameter H : a stationary process $\{Y_i, i \in \mathbb{Z}\}$ exhibits long-range dependence if $0.5 < H < 1$ (for details see Chapter 1. in [21]).

For long-range dependent traffic, the asymptotics for the queue length distributions are no longer exponential [17, 19, 20, 54, 55]. This phenomenon is used in [21], where the exponential approximation of the queue length distribution (used in [51]) is replaced with a Weibullian approximation.

First, we show that when the queue length distribution is asymptotically exponential, by identifying only one specific point of it, a dynamic VP bandwidth control method can be built. Then, we introduce the state space representation of a single server queue, and apply an error term in the state equation. Next, changing the exponential approximation to Weibullian (as proposed in [21]), we show that the estimated bandwidth also converges to its optimal value when the asymptotic queue length distribution is non-exponential.

4.2 VP bandwidth control

In this section, a simple formula for effective bandwidth estimation is derived using existing results from earlier work.

Consider a single server queue that has an infinite buffer with stationary and ergodic cell arrivals from a single source. Let $A(t)$ denote the number of arrivals in $(-t, 0]$, and the workload process is defined by $W(t) = A(t) - Ct$, where C is the number of cells that can be served by the VP in unit time

(it corresponds to the peak cell rate of a constant bit rate VP). The number of customers in the queue in the stationary state is given by [56]:

$$Q = \sup_{t \geq 0} W(t). \quad (4.1)$$

For SRD traffic, the probability that the queue contents exceed a given value k (in the regime of queues) is characterized by two parameters, the so called asymptotic constant β and the asymptotic decay rate η [53]:

$$p_k \stackrel{\text{def}}{=} P(Q \geq k) \approx \beta e^{-\eta k}. \quad (4.2)$$

It is known that the CLR of a single server queue with a finite buffer size, K , is less than p_K [57]. Hence, we have

$$CLR \leq \beta e^{-\eta K}. \quad (4.3)$$

From this relation, the QoS objective of the CLR, CLR_{obj} , is satisfied if

$$\eta \geq \eta_{obj} \stackrel{\text{def}}{=} \frac{\log \beta - \log CLR_{obj}}{K}. \quad (4.4)$$

The large deviation result of Glynn and Whitt [52] yields the necessary and sufficient condition for meeting the QoS objective of the CLR:

$$\alpha(\eta_{obj}) \stackrel{\text{def}}{=} \frac{M_A(\eta_{obj})}{\eta_{obj}} \leq C, \quad (4.5)$$

where $M_A(\cdot)$ is the cumulant generating function of $A(t)$. The function $\alpha(\eta)$ is called the effective bandwidth function of the traffic subjective to the condition that the tail distribution of the queue length has the decay rate η . In particular, $\alpha(\eta_{obj})$ is simply called the *effective bandwidth*, given the decay rate objective η_{obj} .

The simplest realization of VP bandwidth control is to replace the VP bandwidth (C) by the estimated effective bandwidth of the cell stream transferred in the VP. Our objective is to estimate the effective bandwidth of the traffic using real time measurements.

In [21], the following simple approximation has been derived:

$$\frac{1 - \rho}{\rho} \approx d \log p_K, \quad (4.6)$$

where ρ is the link utilization ($\rho = A/C$, A is the long-term average rate of a stationary process), and d is a coefficient, which is constant given that the utilization and the actual p_K probability do not change with time. Once we know the coefficient, d , using Eq.(4.6) the effective bandwidth of the traffic can be estimated by:

$$\alpha(\eta_{obj}) = A(1 + d \log CLR_{obj}). \quad (4.7)$$

To obtain coefficient d , we only need to measure the link utilization ρ and the probability p_K . (Note, that the finer details of the complementary queue length distribution are not important for us. We only need to know one specific point of it.)

Next, a method is proposed for estimating the coefficient d in real time from buffer measurements.

4.2.1 Estimation framework

Consider a buffer of size K cells, where the server utilization and the buffer occupancy probability p_K are measured periodically. The measured values in period n are denoted by $\rho(n)$ and $p_K(n)$, respectively. Having these measurements, we can observe coefficient d in each measurement period (i.e., $d(n)$ in period n) through Eq.(4.6).

The measurement of $\log p_K(n)$ is not trivial. In [21], it has been proposed to measure periodically the buffer occupancy probabilities p_{k_1}, p_{k_2} and p_{k_3} at three different thresholds, k_1, k_2 and k_3 , respectively. Doing this, in measurement period n we obtain:

$$\log p_K(n) = \frac{\log p_{k_2}(n) - \log p_{k_1}(n)}{k_2^{\gamma(n)} - k_1^{\gamma(n)}} K^{\gamma(n)} + \frac{k_2^{\gamma(n)} \log p_{k_1}(n) - k_1^{\gamma(n)} \log p_{k_2}(n)}{k_2^{\gamma(n)} - k_1^{\gamma(n)}}, \quad (4.8)$$

where

$$\gamma(n) = \frac{\log(\log p_{k_2}(n) - \log p_{k_3}(n)) - \log(\log p_{k_1}(n) - \log p_{k_2}(n))}{\log c}, \quad (4.9)$$

with $c = k_2/k_1 = k_3/k_2$. Note that for $\gamma(n) = 1$ Eq.(4.8) gives back the result for the short-range dependent case (see Eq.(C.3) in Appendix C). As a result, using the three-point measurement method, the same calculation can be used for short-range and long-range dependent traffic, which is useful from implementation point of view.

We introduce a measurement error $\sigma(n)$ and assume the following relation:

$$\frac{1 - \rho(n)}{\rho(n) \log p_K(n)} = d(n) + \sigma(n), \quad (4.10)$$

where $\sigma(n)$ is Gaussian white noise with mean 0, representing the measurement error. To take into account traffic pattern variation and to ensure the convergence of the bandwidth to the objective, even when the effective bandwidth formula Eq.(4.7) is not valid for actual traffic characteristics (for a discussion of convergence see Section 4.2.2), suppose that the coefficient d can change with time according to the following equation:

$$d(n) = d(n-1) + \omega(n), \quad (4.11)$$

where $\omega(n)$ is also a Gaussian white noise with mean 0.

In the terminology of the control system Eq.(4.10) and Eq.(4.11) are called the “state space representation” of our single server queue, and the variable d is called the state variable of the system. In Appendix C, the state space representation used in [51] is presented. Since we need only one specific point of the queue length distribution (p_K) to measure the single state variable d , our state space representation is significantly simpler than the one in [51]. One important consequence of this is that the recursive estimation method, presented below, is also simpler, and therefore it is easier to tune its parameters.

Bayesian regression analysis (for a good introduction, see [58]) can be applied to estimate the state variable d . The state variable can be recursively estimated at the end of every measurement epoch by the following Kalman Filter formulation:

[State Renewal]

$$\begin{aligned}\hat{d}(n) &= \hat{d}(n|n-1) + K_n \left\{ \frac{1 - \rho(n)}{\rho(n) \log p_K} - \hat{d}(n|n-1) \right\}, \\ K_n &= \frac{D(n|n-1)}{D(n|n-1) + \Sigma(n)}, \\ D(n) &= (1 - K_n)D(n|n-1), \\ \Sigma(n) &\stackrel{\text{def}}{=} \text{Var}\{\sigma(n)\}.\end{aligned}\tag{4.12}$$

[Projection]

$$\begin{aligned}\hat{d}(n|n-1) &= \hat{d}(n-1), \\ D(n|n-1) &= D(n-1) + \Omega(n), \\ \Omega(n) &\stackrel{\text{def}}{=} \text{Var}\{\omega(n)\}.\end{aligned}\tag{4.13}$$

Here $\hat{d}(n)$ and $\hat{d}(n|n-1)$ are respectively estimates of the state variable $d(n)$ at the n th and $(n-1)$ th measurement epochs, and $D(n)$ and $D(n|n-1)$ are respectively the variances of $\hat{d}(n)$ and $\hat{d}(n|n-1)$. Initial value of the state variable can, for example, be given under the assumption that $A(t)$ is a stationary process, and we have [51]:

$$d = -\frac{1}{2K} \lim_{t \rightarrow \infty} \frac{\text{Var}\{A(t)\}}{E\{A(t)\}}.\tag{4.14}$$

In practice, to implement the above Kalman Filter formulation it is not a trivial question how to set and/or measure the noise variances $\Sigma(n)$ and $\Omega(n)$. These settings are typically based on engineering judgment and simulation experience.

It might be possible to estimate the variance of the measurement noise directly from the measurements, but because that would require a large number of measurements, we take another approach.

From Eq.(4.10) we can see that the measurement contains $\rho(n)$ and $\log p_K(n)$. We assume that $\rho(n)$ can be measured accurately in each measurement epoch. Since small probabilities can not be measured accurately within reasonably long measurement intervals, the uncertainty of the estimation of $\log p_K(n)$ has a major effect on the measurement noise. Our approach is, that we try to bound the range of the measurement error. We know that after the method has found the optimal bandwidth, the value of $\log p_K(n)$ takes its values around $\log CLR_{obj}$. By choosing an appropriate a in the equation below, we request that the range of the error of $\log p_K(n)$ should be within a certain region.

$$|\log p_K(n) - \log CLR_{obj}| < a |\log CLR_{obj}|. \quad (4.15)$$

Using parameter a , ($0 < a < 1$), we determine the size of the interval, in which $\log p_K(n)$ is likely to take its values. According to our experience, $a = 0.25$ is an acceptable value in practice. (Naturally, if the accuracy of the estimation of $\log p_K(n)$ is high, a could be chosen smaller. The value $a = 0.25$ can be regarded as a worst case setting.) From Eq.(4.15) we can derive the bound on the range of the error of the measurement $(1 - \rho)/(\rho \log p_K)$:

$$\frac{1 - \rho(n)}{\rho(n)} \left| \frac{1}{\log p_K(n)} - \frac{1}{\log CLR_{obj}} \right| < \frac{1 - \rho(n)}{\rho(n) \log CLR_{obj}} \left(\frac{a}{a - 1} \right). \quad (4.16)$$

Note, that the left hand side of the inequality is the measurement error around the optimal point, where $p_K \approx CLR_{obj}$. We can choose the term on the right side in Eq.(4.16) to be the standard deviation of the measurement noise in the Kalman Filter. Since more than $\sim 70\%$ of the probability mass of a centered Gaussian density function is in $[-\sigma, +\sigma]$, where σ is the standard deviation of the distribution, and the measurement error is smaller than the bound with high probability, our choice is reasonable. Applying our method, we determine the measurement noise variance as follows:

$$\Sigma(n) = \left(\frac{a(1 - \rho(n))}{(a - 1)\rho(n) \log CLR_{obj}} \right)^2. \quad (4.17)$$

Let us consider the variance of the state noise, $\Omega(n)$. By applying a noise term in the state equation, we allow the state (and thus the bandwidth) to change. Our principle to set the state noise variance is the following: since the observed cell loss ratio is less accurate when it is small, the change of bandwidth should be limited when the observed cell loss ratio is close to CLR_{obj} . But in this case the utilization is also close to the value which is appropriate for the traffic pattern (we will reference this utilization as the “desired utilization”). Since $\Sigma(n)$ depends on $\rho(n)$, its value is also requested to get stabilized, and according to our principle the state noise variance will also become stable in this region. According to the Kalman

Filter formulation, $\hat{d}(n)$ can be written in the form

$$\hat{d}(n) = K_n \left(\frac{1 - \rho(n)}{\rho(n) \log p_K(n)} \right) + (1 - K_n) \hat{d}(n - 1). \quad (4.18)$$

Eq.(4.18) reveals how the Kalman Filter works. Since $0 < K_n < 1$, the n th value of the state variable is calculated as a weighted sum of its previous value and the new measurement. As K_n is the function of $\Omega(n)$ and $\Sigma(n)$, the values of the noise variances determine its transient behavior and limit value. According to the recursive formulas in Eq.(4.12) and Eq.(4.13), the following holds when the noise variances reached their stable values (i.e., $\Sigma(n) \approx \Sigma$ and $\Omega(n) \approx \Omega$):

$$K_\infty \stackrel{\text{def}}{=} \lim_{n \rightarrow \infty} K_n = \frac{\sqrt{\Omega^2 + 4\Sigma\Omega} - \Omega}{2\Sigma}. \quad (4.19)$$

We will use this formula to determine the state noise variance based on our principle by setting the following rule: at any desired $\rho(n)$, if (due to measurement uncertainties) we estimate such a small $\log p_K(n)$ as $\log p_K(n) = b \cdot \log CLR_{obj}$ for some $b > 1$, the relative change in the VP bandwidth should be less than s (for example, $s = 5\%$). Note that we set the state noise variance for the desired $\rho(n)$, because we intend to regulate the behavior of the system when it is stabilized, i.e., the variations of the state are due to measurement uncertainties. According to our experience, $b = 2$ is a possible choice. This is because the third point (corresponding to the largest threshold) in our three-point measurement algorithm is measured sometimes so inaccurately, that the fitted Weibullian distribution results in a very small $\log p_K(n)$ value (e.g. $2 \log CLR_{obj}$).

The above rule is formulated in Eq.(4.20), which can be derived using Eq.(4.22), Eq.(4.19), the first equation in Eq.(4.12) and that the desired value of the state is $\hat{d}(n) \simeq (1 - \rho(n))/(\rho(n) \log CLR_{obj})$.

$$\frac{\Delta C(n)}{C(n-1)} = \left(1 - \frac{1}{b} \right) \frac{\sqrt{\Omega^2(n) + 4\Sigma(n)\Omega(n)} - \Omega(n)}{2\Sigma(n)} (1 - \rho(n)) < s, \quad (4.20)$$

where $\Delta C(n) = C(n) - C(n-1)$. Thus, after measuring $\rho(n)$ and determining $\Sigma(n)$, $\Omega(n)$ is computed as follows:

$$\Omega(n) = \frac{s^2 \Sigma(n)}{(1/b - 1)^2 (1 - \rho(n))^2 - s(1/b - 1)(1 - \rho(n))}. \quad (4.21)$$

Note, that the state noise variance is artificially introduced in order to obtain a model that agrees with our objectives regarding the VP bandwidth allocation. Thus, it is a reasonable approach to tune the value of $\Omega(n)$ based on our expectations about the system behavior (i.e., based on the above rule).

4.2.2 Convergence of the control

If we combine the estimate of coefficient d , described in the previous section, with the effective bandwidth result Eq.(4.7), VP bandwidth control can be expected to be performed through the following control equation for both short-range dependent and long-range dependent traffic:

$$C(n+1) = \hat{A}(n+1)(1 + \hat{d}(n) \log CLR_{obj}), \quad (4.22)$$

where $\hat{A}(n+1)$ is the estimated average traffic in the $n+1$ -th control interval predicted in the n -th control interval. (For example, assuming stationarity, the simplest estimate is $\hat{A}(n+1) = A(n)$.) The average has to be estimated because of two reasons (see e.g., [59]): Firstly, although LRD is assumed to be second-order stationary, the average cannot be measured accurately in short control intervals when the traffic is LRD. Secondly, within finite time it is difficult to distinguish between traffic fluctuations caused by long-term correlations and those caused by non-stationarity. Therefore, one can capture the non-stationarity by a given daily profile (estimated in a certain network environment). It means that the average is predicted (for example) every hour. Deviations from this given average are assumed to result from the correlation structure of the arrivals process. Because the control intervals may be too short to measure the average accurately, smoothing (e.g., by using a Kalman Filter or a moving average) may be applied.

For example, assuming simple moving-average smoothing:

$$\hat{A}(n+1) = \frac{\hat{A}_D(n+1)}{\hat{A}_D(n)}(\alpha A(n) + (1-\alpha)\hat{A}(n)), \quad (4.23)$$

where $A(n)$ is the measured average in the n -th control interval, $\hat{A}_D(n)$ is the daily profile-based estimation of the one-hour relative average, and α is the parameter of the moving-average method ($0 < \alpha < 1$, depending on the length of the control interval).

Since our primary interest is not related to the estimation of the actual traffic volume, in the rest of this section we will assume that the average does not change with time and it can be estimated accurately, thus we use the notation $\hat{A}(n+1) = A$.

The problem is that the notion of effective bandwidth can not be applied for long-range dependent traffic, because Eq.(4.5) was derived under the assumption of short-range dependence (or linear scalings). What we can do is to show that the proposed Kalman Filter formulation with the three-point measurement method works well, despite the fact that we lack the equations leading us to the direct solution. To do this, we can rewrite Eq.(4.6) as:

$$d(C, p_K(C)) \approx \frac{C - A}{A \log p_K(C)}. \quad (4.24)$$

Here, the notation $d(C, p_K)$ emphasizes that (for the given traffic $A(t)$) d is the function of the bandwidth C and the probability p_K , which probability also depends on C .

For a given CLR_{obj} to be satisfied, Eq.(4.7) gives us the required VP bandwidth (C_{obj}). From Eq.(4.7), we get:

$$d(C, CLR_{obj}) = \frac{C - A}{A \log CLR_{obj}}. \quad (4.25)$$

For the short-range dependent case (where Eq.(4.5) is valid), assuming that the error of approximations on the way to get Eq.(4.6) is negligible, $d(C, p_K(C))$ *does not* depend on C and remains constant for various bandwidths. Therefore, once we can evaluate d by using Eq.(4.24) for a given initial bandwidth (C_0), d is used in Eq.(4.25) and we can derive C_{obj} , which is the bandwidth to be determined (see Figure 4.1). In this case, C_{obj} is at the point where $d(C, p_k(C))$ meshes $d(C, CLR_{obj})$ and it can be calculated *in one step*.

Next, we consider the case when the traffic is long-range dependent (the notion of effective bandwidth can not be derived), or Eq.(4.6) involves non-negligible approximation errors. The function $d(C, CLR_{obj})$ remains the same as previously, but $d(C, p_K(C))$ becomes dependent on C (see Figure 4.2 for an example). Our algorithm works as previously. Calculate $d(C, p_K(C))$ for C_0 and set the bandwidth to C_1 such that $d(C_1, CLR_{obj}) = d(C_0, p_K(C_0))$. Measuring p_K at C_1 gives $d(C_1, p_K(C_1))$, different from $d(C_0, p_K(C_0))$, and that is the reason why the parameter $\omega(n)$ was introduced to the state equation Eq.(4.11) in the Kalman Filter. This parameter makes it possible for \hat{d} to change its value successively. Next, the VP bandwidth is adjusted appropriately and the same step is repeated again. In our example, in Figure 4.2, the VP bandwidth converges to its target value (C_{obj}). For such an iterative method the question of convergence emerges. The sufficient condition for that—assuming accurate measurements—is that the slope of $d(C, p_K(C))$ can not exceed the slope of $d(C, CLR_{obj})$, namely,

$$\left| \frac{\partial d(C, p_K(C))}{\partial C} \right| < \left| \frac{\partial d(C, CLR_{obj})}{\partial C} \right|. \quad (4.26)$$

As an example, Figure 4.3 shows the simulated curve of $d(C, p_K(C))$ for long-range dependent traffic trace with Hurst parameter $H = 0.7$ and $A = 10Mbps$ (see Section 4.3.2 for more details, where the same model with the same parameters is used). In this example, the sufficient condition of convergence (Eq.(4.26)) is satisfied around C_{obj} . Note, that in practice, when the parameter K_n is sufficiently small in the Kalman Filter, the sequence of the determined bandwidth C_i , ($i = 0, 1, \dots$) can approach C_{obj} , even if Eq.(4.26) does not hold. That is, by introducing the Kalman Filter or the state space model, our control method becomes more robust in practice.

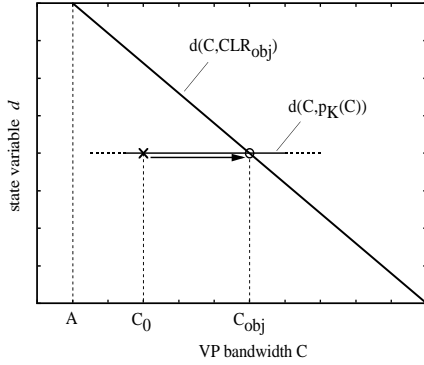


Figure 4.1: Transient behavior of the VP bandwidth control algorithm for the short-range dependent case.

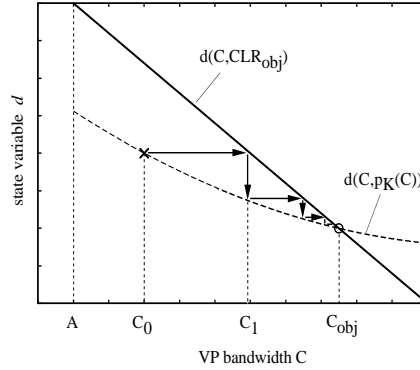


Figure 4.2: Transient behavior of the VP bandwidth control algorithm in general case.

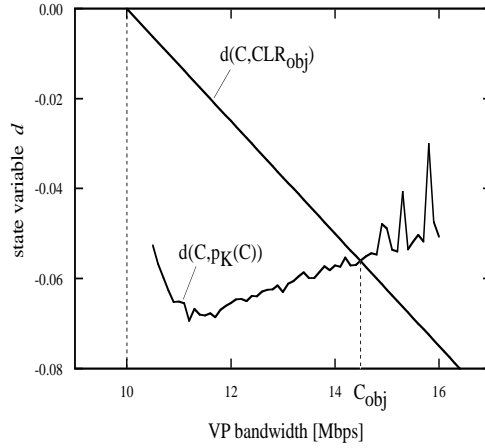


Figure 4.3: Simulated plot of $d(C, p_K(C))$ for long-range dependent traffic.

4.2.3 Practical implications

In this section, we briefly consider practical questions concerning applicability and implementation.

The greatest advantage of our algorithm comes from its simplicity (see Eq.(4.10) and Eq.(4.22)): it requires only simple processing on a limited number of parameters during each control interval.

In our algorithm, the notion of self-similarity or long-range dependence (LRD) is not taken into account in the control equations; only the applied buffer measurements are based on an LRD assumption (Weibullian queue length asymptotic). Even so, the method should also work for traffic exhibiting LRD because the value of $\hat{d}(n)$ is always adjusted such that for the

resulting bandwidth $\log p_K(n) \approx \log CLR_{obj}$.

Responsiveness

We can say that because the applied effective bandwidth formula is not always a proper approximation, it does not yield the required bandwidth after a single measurement, but rather after a series of measurements (see also Section 4.2.2). Therefore, the dynamic behavior and accuracy of our algorithm basically depend on the Kalman Filter.

The responsiveness of the algorithm is determined by three factors:

- the values of $\Sigma(n)$ and $\Omega(n)$ in the Kalman Filter,
- the length of the control intervals, and
- if the average is estimated like in Eq.(4.23), then the value of the α parameter (but this effect is not significant if the control intervals are large enough).

In our experience, the above parameters can be determined based on practical considerations, such that the smoothing is effective and the responsiveness is acceptable for practical application.

Scaling properties

When independent and bursty sources are multiplexed on a single link, the independence in the statistical variations of the individual sources makes it possible to reduce the bandwidth required for the combined stream. Multiplexing more such connections results in a higher potential for multiplexing gain.

Both the asymptotic constant β and the Hurst parameter H correspond to the burstiness of the arrivals process. This means that the algorithm could be improved by incorporating β and H in the state variable, d . This, however, may decrease the adaptability of the method, and also deteriorate its simplicity.

In [21] it is discussed how these two parameters might be incorporated into the algorithm.

Buffer thresholds, length of control intervals

The actual buffer-monitoring thresholds depend on the buffer length, the CLR objective, and the length of the control interval, as well as on the rate and burstiness of the traffic. In Section 4.2.1 of [21], simple formulas are proposed for calculating good thresholds.

4.3 Simulations

Here we describe simulation experiments, when the VP bandwidth control using the proposed three-point measurement was applied. First, the aggregated traffic offered to the VP is short-range dependent with low burstiness. In this case, the queue length distribution can be well approximated by an exponential distribution with asymptotic constant β close to one [60]. Second, the offered traffic is short-range dependent, but more bursty, implying $\beta \ll 1$. Finally, the traffic is long-range dependent.

4.3.1 Short-range dependent traffic

Consider N identical but independent on-off sources (VCs) with alternating activity periods (T_{on}) and silence periods (T_{off}). In the on-states, cells were offered from the VC at a constant rate, while in the off-states the source remained silent. The duration of each on-state and each off-state were exponentially distributed. (We refer to this model as the ‘on-off source’.) The cells from the multiplexed on-off sources arrived at the output buffer dedicated to the VP.

The simulation conditions were as follows: the VP bandwidth was initially 1.2 Mbps, and the buffer size was 128 cells. 10 sources were multiplexed, and in the on-states cells were offered from each VC at a rate of 1 Mbps. The average rate of each VC was 0.1 Mbps, and the mean duration of the on-state was 1 ms in the ‘non-bursty’ case (Figure 4.4), and 5.43 ms in the ‘bursty case’ (Figure 4.5). At three different thresholds, the frequencies in which the number of cells in the buffer was greater than or equal to the thresholds were measured every 20 minutes. The thresholds were set at 5, 10 and 20 cells. Every 20 minutes the unknown parameter, d , in Eq.(4.10) was estimated, and the VP bandwidth was adjusted by the proposed formula (Eq.(4.22)). (The initial value of d was set approximately using the results from the first measurement epoch.) The CLR objective was 10^{-8} .

Figure 4.4 and Figure 4.5 show the dynamic behavior of the VP bandwidth under the proposed control. The actually required VP bandwidth was calculated through an analytical method [57, 61], given complete information about the cell arrival process.

In the case of ‘bursty sources’, the asymptotic constant β is in the order of 10^{-2} , and therefore the queue length distribution slowly approaches its asymptote, implying larger decay rate for small buffer sizes than the asymptotical one. The fitted Weibullian distribution slightly overestimates the CLR, that results in larger VP bandwidth than theoretically required (Figure 4.5).

To make the difference between the non-bursty and bursty case more clear, we plotted the IDC plot for both traffic traces (see Figure 4.7). The

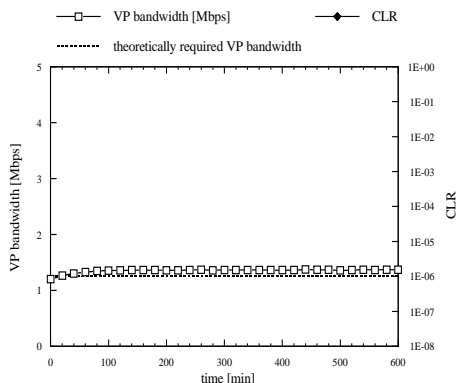


Figure 4.4: Sample path of VP bandwidth under the proposed control (non-bursty case).

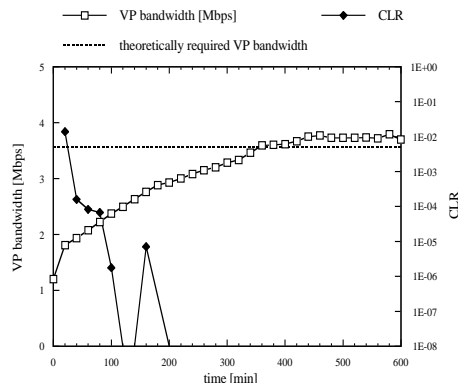


Figure 4.5: Sample path of VP bandwidth under the proposed control (bursty on-off sources).

IDC value for the arrival process $A(t)$ is given by:

$$IDC(t) \stackrel{\text{def}}{=} \frac{Var\{A(t)\}}{E\{A(t)\}}, \quad (4.27)$$

for $t > 0$, and the IDC plot is obtained by plotting $\log IDC(t)$ against $\log t$. For short-range dependent processes the IDC is bounded. Higher IDC value corresponds to higher degree of burstiness, and the positive slope of the curve indicates the presence of positive correlations. (For long-range dependent processes the IDC increases monotonically with slope $2H - 1$, where H is the Hurst parameter of the process.)

4.3.2 Long-range dependent traffic

Markovian source models with finite state space are only capable of generating short-range dependent traffic. If we want to generate long-range dependent traffic, we need to find an appropriate method to do that. In [62], six different such methods are discussed. For our simulator the on-off model originally proposed by [14] was suitable.

To introduce long-range dependence into the on-off model, assume a heavy tail for the distribution of the activity periods T_{on} [20]. Here we use the (translated) Pareto distribution for T_{on} , where

$$P(T_{on} \leq t) = 1 - \left(\frac{\theta}{t + \theta} \right)^{3-2H} \quad (4.28)$$

for some $\theta > 0$, $t > 0$ and $0.5 < H < 1$, thus confirming that heavy tailed activity periods generate long-range dependence and identifying H as the Hurst parameter. (We refer to this source model as the ‘Pareto-type

source'.) In the limit of a large number of such sources and high load, it is shown in [14] that the aggregated traffic, properly normalized, converges to an exactly self-similar Gaussian process, and in [20] Brichet *et al.* showed that the tail of the queue length distribution is Weibullian.

In our numerical example the number of sources was set to 100. The mean duration of the on period was set to 0.1 ms to assure the high load, and the parameters of the Pareto distribution were set with Hurst parameter $H = 0.7$ for each source. In Figure 4.7, the IDC plot of the aggregated traffic is shown. The monotonically increasing curve reveals the presence of long-range dependence at the given time scales. The measured Hurst parameter agrees accurately with the desired value ($H = 0.7$). The VP bandwidth was initially set to 12 Mbps, and all the other parameters were kept as they had been in the previous example.

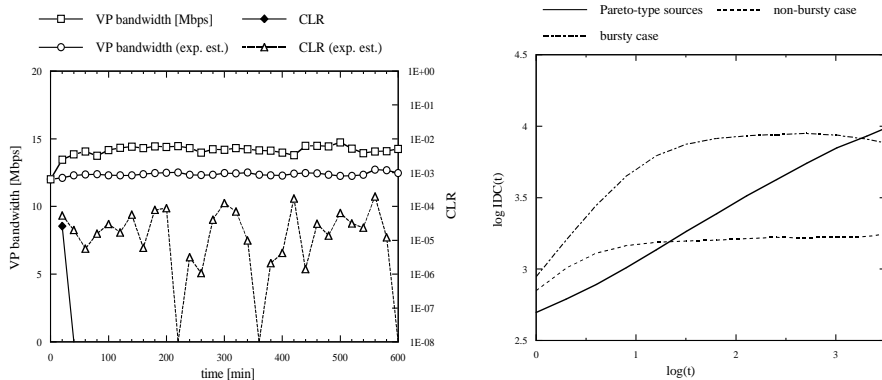


Figure 4.6: Sample path of VP bandwidth for Pareto-type sources under the proposed control and also for the linear estimate method. Figure 4.7: IDC plot for the simulated traffic traces.

The simulation result is shown in Figure 4.6. What we can see is that the estimated VP bandwidth reaches a certain value and (except in the initial phase) no cell loss occurs during the simulation time. Since we gave up the Markovian property in our source model, the actually required bandwidth can not be calculated. To verify the results, we rejected the assumption of Weibullian queue length distribution, and using only the first and third thresholds and fitting an exponential curve, reran the simulation. As can be seen from Figure 4.6, the estimated VP bandwidth is smaller by a few Mbps than previously, and the experienced CLR is in the order of 10^{-4} – 10^{-5} . This means that our proposed control shows good CLR performance and its derived bandwidth is not excessive. At most the excessive bandwidth is less than a few Mbps.

The above simulation results were obtained with a short buffer size (128 cells). We note that at larger buffer sizes the VP bandwidth is easier to

control, according to our experiences.

4.3.3 Actual ATM traffic

For the actual ATM traffic, we analyzed the aggregated traffic on the Swedish University Network (SUNET) for the summer of 1996. The LAN traffic of universities in the northern region of Sweden, around Uppsala, is connected to a FDDI backbone network connected to the ATM backbone network in Stockholm. This network joins the northern LANs of SUNET to the international Internet backbone and to the southern university networks around Göteborg. The measurements reported here were for traffic between Uppsala and Göteborg over a CBR connection with a cell rate of 90,000 cps (38.16 Mbps). The ATM traffic stream was duplicated and routed on dedicated links to the Telia Research Center in Haninge, where almost 100 traffic traces were collected, with more than 8 million cell arrivals in each trace, using a non-commercial custom-built measurement instrument developed in the RACE Parasol project [63]. A good assumption is that the traffic was an ordinary mix of common Internet traffic types, such as HTTP, FTP, telnet, chat, and IPphone.

Vidács *et al.* presented an LRD analysis of the measured traffic traces [64]. To estimate the Hurst parameter, H , R/S and variance-time analysis [14] were performed for 45 data sets. The results showed the analyzed data to be rather bursty, with an H of about 0.9.

In our simulation, 14 consecutively measured data sets were used as the input; they contained more than 10^7 ATM cells altogether. (Note that this file merging is somewhat artificial because there was actually a time gap between the consecutive measurements. After the recording of the 8 million cell arrivals, the collected data was saved, which took some time.) The traffic on the link was rather bursty, with a mean rate of 6.12 Mbps. Figure 4.8 shows the results of applying our bandwidth control method. (The control interval was set to 400 s; the buffer size was 1024 cells with monitoring thresholds of 50, 100, and 200, with $\hat{A}_D(n) \equiv 1$ and $\alpha = 1$ in Eq. (4.23)) The controlled bandwidth fluctuated between 20 and 30 Mbps, implying a utilization rate of about 25% – confirming that the traffic was highly bursty. In spite of high burstiness, cell loss occurred in only one control interval, when the bandwidth dropped to 20 Mbps because of the relatively low traffic during the previous control interval. Again, to verify the result the simulation was repeated with a fixed VP bandwidth of 20 Mbps and our control turned off. The experienced CLR turned out to be as high as 10^{-4} , which shows that the desired bandwidth was well above 20 Mbps. This means that to satisfy our CLR objective, the bandwidth must be higher than 20 Mbps, at least during busy periods.

The dashed line in Figure 4.8 shows the VP bandwidth when the dynamic behavior of the algorithm was modified to get a more robust estimate of the

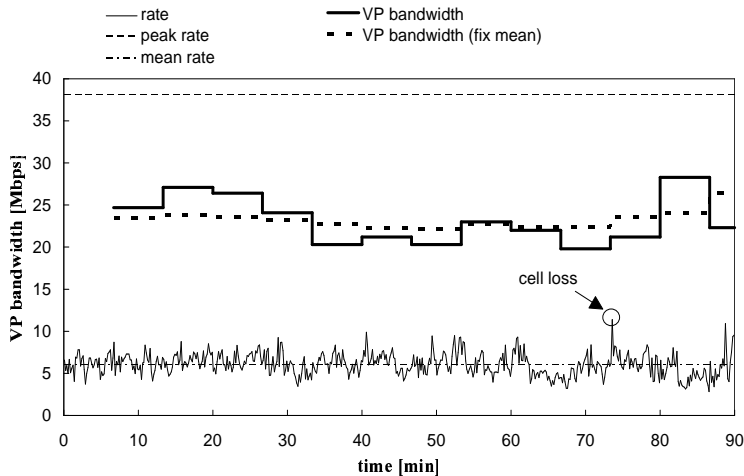


Figure 4.8: Sample paths of VP bandwidth under the proposed control for measured ATM traffic with $\hat{A}_D(n) \equiv 1$; first (solid line) with $\alpha = 1$ and second (dashed line) with $\hat{A}(n) \equiv A$ (fix mean) and $\alpha = 0$.

bandwidth. First, we assumed that the average traffic rate was predicted accurately in advance (e.g, from the daily profile: $\hat{A}(n) \equiv A$ with $\alpha = 0$ in Eq. (4.23)); then the noise parameters of the Kalman Filter were adjusted to increase the robustness while decreasing the adaptability. The resulting VP bandwidth was much smoother, but cell loss occurred in the same control interval as before. The traffic peak in that interval was so pronounced that even the 2 Mbps extra bandwidth was still not enough to prevent loss.

4.4 Implementation issues

We assume that the VP is deterministic, or a constant bit rate (CBR) one. That is, the VP bandwidth is fixed and the cell transmission at the origination point is scheduled according to the VP bandwidth. (Typically the VP is used by the intranetwork of a company, and we focus on a physical link that includes the VP in a public network. Then, the spare capacity on this physical link is not used by other companies. In this sense, the allocation of this physical link, is not work-conserving.)

We also assume that each virtual channel handler (VCH) (i.e., switching node, see Figure 4.9) has an output buffer, and that cell loss due to buffer overflow may occur at the output buffer of the VCHs where the VPs originate. (We call such a handler the originating VCH.) Best-effort VCs (i.e., UBR VCs) are multiplexed in the VP. If one wants to provide a good cell-level QoS, it must be measured and VP bandwidth control must be

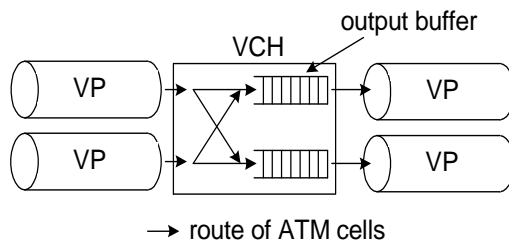


Figure 4.9: Virtual channel handler.

applied.

Virtual path handlers are assumed to have sufficient buffer capacity and a high-capacity transmission path. Thus, the cell loss in the VP handlers is negligible. (If this cell loss is not negligible, the VP bandwidth must be calculated for each link composing a VP. The maximum bandwidth calculated for each VP link should be used for the VP bandwidth. For simplicity, we consider the case in which the cell loss in the VP handlers is negligible.)

Traffic is measured at the originating VCH. The bandwidth is calculated by a processor on the line-interface card accommodating the VP or in the operation systems for traffic and network management. If the bandwidth is calculated in the processor, the result is transmitted to the operation systems. If the bandwidth is calculated in the operation systems, the traffic measurement is transmitted from the originating VCH to the operation systems. The operation systems judges whether the calculated bandwidth can be handled by determining how much bandwidth remains of the transmission path accommodating the VP. If it is sufficient, the calculated bandwidth is assigned to the VP.

4.5 Conclusion

In this chapter, a VP bandwidth control algorithm applicable to long-range dependent traffic as well as short-range dependent traffic was developed.

The VP bandwidth control is governed by Eq.(4.22), in which the state variable (\hat{d}) is estimated by a Kalman Filter. The state space representation of the queueing system is described by Eq.(4.10) and Eq.(4.11), and the Kalman Filter works according to Eq.(4.12) and Eq.(4.13). The variance parameters in the Kalman Filter are tuned with Eq.(4.17) and Eq.(4.21). The condition of the convergence of the control is presented by Eq.(4.26).

The effective bandwidth formula, Eq.(4.7), was derived from a linear scaling assumption or a short-range dependence assumption, while the measurement of the buffer occupancy probability (p_K) was based on a nonlinear scaling assumption or a long-range dependence assumption. Our goal was

to show that despite this difference in assumptions, a recursive estimation framework can integrate them under a nonlinear scaling, and the target VP bandwidth can be found under weak conditions. Numerical examples for short-range dependent traffic, long-range dependent traffic and real ATM traffic were shown. For all cases, our method showed good CLR performance with small excessive bandwidth.

Chapter 5

Summary of the Dissertation

In this chapter, the contributions of the dissertation are summarized. Publications corresponding to the work presented in the dissertation, as well as applications of the contributions are presented.

5.1 Connection admission control in UTRAN

In Chapter 2, I have investigated connection admission control in UTRAN. UTRAN is a connection oriented radio access network, where radio frames are transported in packets, using ATM or IP technologies. Due to the specific characteristics of the WCDMA radio interface, requirements on packet delay and jitter are strict. Therefore, connection admission control must be solved.

- I have given a connection admission control method, which is applicable in the Iub interface of UTRAN. The admissible region is approximated by the intersection of delay-limited regions with linear borders (also referred to as hyper-planes) and one overload-limited region with (generally) non-linear border. This construction makes it possible to fulfil all practical requirements (e.g., on limited complexity and high precision).
- I have validated the hyper-plane approximation by using the Brownian bridge approximation of the packet arrival process. Based on the results of the validation, I have also proposed a closed formula for the fast approximation of the edges of the hyper-planes.

This work was carried out at the Traffic Analysis and Network Performance Laboratory of Ericsson Research. It has been presented in [J1]. Further related publications are [J2], [J3] and [J6]. (Earlier work on UTRAN CAC with my participation can be found in [C4].) One patent application has been filed on this method [P2], and two other patent applications are directly related to this algorithm ([P4] and [P6]).

The method is implemented in Ericsson's UTRAN transport node products that contain AAL2 multiplexers.

5.2 Mobility and traffic analysis for WCDMA networks

The work presented in Chapter 3 is based on a model proposed in [13], which suggests a convenient and practical approach to build an analytic traffic model of cellular systems, including also the effect of vehicle mobility. I have extended this model for WCDMA systems, where typically soft handover is applied for user-plane traffic instead of hard handover.

- I have introduced to the model the notion of soft handover regions (SHR) and derived the following measures: the soft handover probabilities and the Z-set call arrival rates of connections generated in a certain SHR or handed-over from a certain SHR, the mean SHR residence time, and the Z-set traffic load in a certain SHR. Using these parameters, I have given an estimation of user-plane traffic on the Iur interface.
- Based on the SHR parameters, I have obtained the following cell parameters: the distribution of the channel occupancy time, the blocking probability, the offered traffic load, and the soft handover intensities.

This work was carried out at the Traffic Analysis and Network Performance Laboratory of Ericsson Research. It has been presented in [C1]. The calculations are implemented in a software tool at Ericsson. This work has been translated to German language and it has been used at RWTH (Rheinisch-Westfälische Technische Hochschule, Aachen) in a seminar in the winter semester of 2001 as teaching material.

5.3 Real-time VP bandwidth control

In Chapter 4, the buffer monitoring method presented in [21] is applied to Virtual Path (VP) bandwidth control. I assumed that the VP is deterministic, or a constant bit rate (CBR) one, and best-effort VCs (i.e., UBR VCs) are multiplexed in the VP. If one wants to provide a good ATM cell-level QoS, it must be monitored and VP bandwidth control must be applied.

- I have developed a recursive bandwidth control algorithm, and showed that the control converges in case of both short and long-range dependent traffic.

- I have introduced a state description of the system, and set up a Kalman filter for the estimation of the state. I have given simple formulas for tuning the parameters of the Kalman filter.
- I have applied the bandwidth control algorithm for both artificial and real traffic.

This work was done at NTT Multimedia Networks Laboratories. It has been presented in [J4], [J5], [C5] and [C6]. The proposed methods are subject to a patent application [P8].

Appendix A

Derivation of the approximation of $Q(x)$ used in Section 2.5.2

Denote by $\mathcal{W}(t)$ and $\mathcal{W}_i(t)$ independent Wiener processes. To obtain $Q(x)$, we need to evaluate the following expression:

$$\Pr \left\{ \sup_{0 \leq \tau \leq TTI} \left(\sum_i N_i \alpha_i b_i \frac{\tau}{TTI} + \sum_i \sqrt{N_i b_i^2} \mathcal{W}_i \left(\frac{\alpha_i \tau}{TTI} \right) - C \tau \right) \geq C x \mid \mathcal{W}_i(1) = 0; \forall i \right\}.$$

To simplify the notation, we introduce $A_i = N_i \alpha_i \rho_i$, and $B_i = N_i \alpha_i \rho_i^2$. Knowing that $\mathcal{W}_i(\alpha_i t) = \sqrt{\alpha_i} \mathcal{W}_i(t)$, we obtain:

$$\Pr \left\{ \sup_{0 \leq \tau \leq 1} \left(\sum_i A_i \tau + \sum_i \sqrt{B_i} \mathcal{W}_i(\tau) - C \tau \right) \geq \frac{C x}{TTI} \mid \mathcal{W}_i \left(\frac{1}{\alpha_i} \right) = 0; \forall i \right\}.$$

Conditioning on the events $\{\sqrt{B_i} \mathcal{W}_i(1) = y_i\}$ results in:

$$\int_{\mathbb{R}} \dots \int_{\mathbb{R}} \Pr \left\{ \sup_{0 \leq \tau \leq 1} \left(\sum_i \sqrt{B_i} \mathcal{W}_i(\tau) - (C - \sum_i A_i) \tau \right) \geq \frac{C x}{TTI} \mid \sqrt{B_i} \mathcal{W}_i(1) = y_i, \mathcal{W}_i \left(\frac{1}{\alpha_i} \right) = 0; \forall i \right\} dF_1(y_1) \dots dF_K(y_K),$$

where F_i is the normal distribution function with zero mean and variance $\sigma_i^2 = N_i \rho_i^2 \alpha_i (1 - \alpha_i)$, which corresponds to the distribution of $\{\sqrt{B_i} \mathcal{W}_i(1)\}$ with $\mathcal{W}_i(1/\alpha_i) = 0$.

Since both process $\mathcal{X}_i(t) = \sqrt{B_i} \mathcal{W}_i(t)$ and its reverse version $\mathcal{X}'_i(t) = \sqrt{B_i} \mathcal{W}_i(1/\alpha_i - t)$, where $\alpha_i \in (0, 1]$, are Markov processes and we consider

only $t \in [0, 1]$, we can cancel the condition $\mathcal{W}_i(1/\alpha_i) = 0$:

$$\int_{\mathbb{R}} \dots \int_{\mathbb{R}} \Pr \left\{ \sup_{0 \leq \tau \leq 1} \left(\sum_i \sqrt{B_i} \mathcal{W}_i(\tau) - (C - \sum_i A_i) \tau \right) \geq \frac{C x}{TTI} \mid \sqrt{B_i} \mathcal{W}_i(1) = y_i; \forall i \right\} dF_1(y_1) \dots dF_K(y_K).$$

We can modify the condition as follows:

$$\int_{\mathbb{R}} \dots \int_{\mathbb{R}} \Pr \left\{ \sup_{0 \leq \tau \leq 1} \left(\sum_i \sqrt{B_i} \mathcal{W}_i(\tau) - (C - \sum_i A_i - \sum_i y_i) \tau \right) \geq \frac{C x}{TTI} \mid \mathcal{W}_i(1) = 0; \forall i \right\} dF_1(y_1) \dots dF_K(y_K),$$

which enables us to apply $\sum_i c_i \mathcal{W}_i(t) = \sqrt{\sum_i c_i^2} \mathcal{W}(t)$:

$$\int_{\mathbb{R}} \dots \int_{\mathbb{R}} \Pr \left\{ \sup_{0 \leq \tau \leq 1} \left(\sqrt{\sum_i B_i} \mathcal{W}(\tau) - (C - \sum_i A_i - \sum_i y_i) \tau \right) \geq \frac{C x}{TTI} \mid \mathcal{W}(1) = 0; \forall i \right\} dF_1(y_1) \dots dF_K(y_K).$$

Using Eq.(2.31) and introducing the variable $z = \sum_i y_i$ the K integrals can be substituted by a single integral as follows:

$$\int_{-\infty}^a \exp \left\{ -\frac{2 C x}{TTI \sum_i B_i} \left(\frac{C x}{TTI} + C - \sum_i A_i - z \right) \right\} dF(z) + \int_a^{\infty} dF(z), \quad (\text{A.1})$$

where $a = C + \frac{C x}{TTI} - \sum_i A_i$ and F is the normal distribution function with zero mean and variance $\sigma^2 = \sum_i N_i \rho_i^2 \alpha_i (1 - \alpha_i)$.

After evaluating this integral a closed form expression can be obtained. Eliminating the negligible terms from that closed form, one obtains the approximation Eq.(2.46).

If one looks back to Eq.(2.19), it is obvious that Eq.(A.1) is rather similar to a continuous approximation of Eq.(2.19), where the binomial distribution is approximated by a normal distribution, $\Pr(D_i > \tilde{D}_i \mid \underline{N}^{act} = \underline{n})$ is approximated by Eq.(2.34), and the summation is substituted with an integral. For a single service this integral is the following:

$$\frac{1}{\sqrt{2\pi N \alpha (1 - \alpha)}} \int_0^N \exp \left\{ -\frac{2 C x}{TTI n \rho^2} \left(\frac{C x}{TTI} + C - n \rho \right) - \frac{(N \alpha - n)^2}{2 N \alpha (1 - \alpha)} \right\} dn \quad (\text{A.2})$$

Eq.(A.2) does not yield such a simple closed form solution as Eq.(2.46). The basic difference between Eq.(A.1) and Eq.(A.2) (resulting from the Gaussian approximation of the arrival process) is that in the former, $2Cx/(TTI \sum_i B_i)$ is constant, while in the latter, $2Cx/(TTI n \rho^2)$ depends on the random variable n .

Appendix B

Incremental assignment method for Section 3.2.1

The incremental assignment method is based on a multistep shortest route routing algorithm, which assumes that each subscriber always selects the shortest route between two centroids.

Similarly to [13], we define the transportation *usage cost of a street* as the travel time of vehicles moving along the street as:

$$\text{UsageCost}(j) = \frac{L(j)}{v(j)},$$

where $L(j)$ is the length of street j and $v(j)$ is the speed of vehicles on street j . $L(j)$ is obtained from the road systems model and $v(j)$ is given by the load-speed profile of street j , which represents the relation between the traffic volume and the speed of vehicles in the street. Streets are classified into different types such as avenue, main street, minor street etc., according to the scale of the traffic they can carry. Each type of street has its own load-speed profile. Figure 3.6 shows an example of these profiles.

Let OD_{ij} be the traffic volume from centroid i to centroid j , which is given by the O-D table. The incremental assignment method approximately estimates the amount of traffic volume flowing along each possible alternative route from centroid i to j . For the estimation, the method divides the traffic volume into m parts and in m number of steps it assigns OD_{ij}/m amount of traffic volume to the shortest route between centroids i and j . The well-known *Dijkstra-algorithm* is used for the calculation of the shortest route (see in [65] for example), and the cost of a route is the sum of the usage costs of the streets along the route between centroids i and j . As the traffic volume on each link changes, the usage cost of each street also changes. Therefore in each step, the vehicular traffic gradually adapts to the street network environment, consequently the shortest route can change as well. The larger the value of m , the more accurate approximation is achieved.

Appendix C

State space representation with two-point measurement for Section 4.2.1

In [51], exponential queue length distribution was assumed, and a two-point buffer measurement method was proposed. The following relations among the asymptotic constant β , the asymptotic decay rate η and the link utilization ρ were derived:

$$\frac{1-\rho}{\rho} = \sum_{k=1}^{\infty} d_k^{(2)} \eta^k, \quad (\text{C.1})$$

and

$$\log \beta = \sum_{k=1}^{\infty} d_k^{(1)} \eta^k, \quad (\text{C.2})$$

where

$$\begin{aligned} d_k^{(1)} &= - \lim_{t \rightarrow \infty} \left\{ \frac{\langle A(t)^k \rangle_c}{k!} - \left(\lim_{t \rightarrow \infty} \frac{\langle A(t)^k \rangle_c}{k!t} \right) t \right\}, \\ d_k^{(2)} &= \lim_{t \rightarrow \infty} \frac{\langle A(t)^{k+1} \rangle_c}{(k+1)! \langle A(t)^1 \rangle_c}, \end{aligned}$$

and $\langle A(t)^k \rangle_c$ is the k -th cumulant of the arrivals process $A(t)$.

Suppose that, at two different thresholds, k_1 and k_2 , the buffer occupancy probabilities and the link utilization are periodically observed. Denote the measured values in the n -th measurement period by $p_{k_1}(n)$, $p_{k_2}(n)$ and $\rho(n)$. Then, the values of η and $\log \beta$ in period n can be obtained as follows:

$$\begin{aligned} \eta(n) &= \frac{\log p_{k_1}(n) - \log p_{k_2}(n)}{k_2 - k_1}, \\ \log \beta(n) &= \frac{k_2}{k_2 - k_1} \log p_{k_1}(n) - \frac{k_1}{k_2 - k_1} \log p_{k_2}(n). \end{aligned} \quad (\text{C.3})$$

Based on Eq.(C.1) and Eq.(C.2) the following relations were assumed:

$$\begin{aligned}\log \beta(n) &= \sum_{i=1}^{I_1} d_i^{(1)}(n) \eta(n)^i + \sigma_1(n), \\ \frac{1 - \rho(n)}{\rho(n)} &= \sum_{i=1}^{I_2} d_i^{(2)}(n) \eta(n)^i + \sigma_2(n),\end{aligned}\tag{C.4}$$

where $\sigma_j(n)$ is Gaussian white noise, representing the measurement error. To take into account traffic pattern variation, the coefficients $\{d_i^{(1)}\}$, $\{d_i^{(2)}\}$ were assumed to change with time according to the following equations:

$$\begin{aligned}d_i^{(1)}(n) &= d_i^{(1)}(n-1) + \omega_i^{(1)}(n), \quad i = 1, \dots, I_1, \\ d_i^{(2)}(n) &= d_i^{(2)}(n-1) + \omega_i^{(2)}(n), \quad i = 1, \dots, I_2,\end{aligned}\tag{C.5}$$

where $\omega_i^{(1)}(n)$ and $\omega_i^{(2)}(n)$ are also Gaussian white noise with mean 0. Equations Eq.(C.4) and Eq.(C.5) were written in vector form as:

$$\begin{aligned}\underline{b}(n) &= H(n)\underline{x}(n) + \underline{\sigma}(n), \\ \underline{x}(n) &= \underline{x}(n-1) + \underline{\omega}(n).\end{aligned}\tag{C.6}$$

Note that matrix $H(n)$, which is used to relate the state vector \underline{x} to the measurements $\underline{b}(n)$, is also measured. This may cause difficulties in practical implementation (e.g., when setting the noise variances).

Bibliography

- [1] ATM Forum. *ATM Forum Traffic Management Specification Version 4.0*, 1996.
- [2] Göran Eneroth and Martin Johnsson. ATM transport in cellular networks. In *proc. International Switching Symposium (ISS'97)*, Toronto, Canada, September 1997.
- [3] Hiroshi Nakamura, Hisakazu Tsuboya, Masatomo Nakano, and Akihisa Nakajima. Applying ATM to mobile infrastructure networks. *IEEE Communications*, January 1998.
- [4] *Introduction to 3G Mobile Communications*. Artech House, Norwood, USA, 2001.
- [5] *Third Generation Mobile Systems*, <http://www.ericsson.com/3G>.
- [6] Tero Ojanpera and Ramjee Prasad. *Wideband CDMA for Third Generation Mobile Communications*. Artech House, Norwood, USA, 1998.
- [7] ITU-T. *AAL Type 2 Signalling Protocol (Capability Set 1)*, 1999.
- [8] ITU-T. *AAL Type 2 Signalling Protocol (Capability Set 2)*, 2000.
- [9] John H. Baldwin, Behram H. Bharucha, Bharat T. Doshi, Subrahmanyam Dravida, and Sanjiv Nanda. AAL-2 — a new ATM Adaptation Layer for small packet encapsulation and multiplexing. *Bell Labs Technical Journal*, April 1997.
- [10] G. Eneroth, G. Fodor, G. Leijonhufvud, A. Rácz, and I. Szabó. Applying ATM/AAL2 as a switching technology in 3rd generation mobile networks. *IEEE Communications Magazine*, 37(1), 1999.
- [11] 3GPP Technical Specification TR 25.426. *UTRAN Iur and Iub Interface Data Transport & Transport Signalling for DCH Data Streams*, 1999.
- [12] 3GPP. *IP Transport in UTRAN*, March 2002.

- [13] K. Nakano, K. Saita, and M. Sengoku et al. Mobile communications traffic analysis on a road system model. *Performance and Management of Complex Communication Networks, International Federation for Information Processing (IFIP), Kluwer Academic Publishers*, 1998.
- [14] W.E. Leland, M.S. Taqqu, W. Willinger, and D.V. Wilson. On the self-similar nature of ethernet traffic (extended version). *IEEE/ACM Trans. on Networking*, 2(1), 1994.
- [15] J. Beran, R. Sherman, M.S. Taqqu, and W. Willinger. Long-range dependence in variable-bit-rate video traffic. *IEEE Trans. on Communications*, 43(2/3/4), 1995.
- [16] W. Willinger, M.S. Taqqu, R. Sherman, and D.V. Wilson. Self-similarity in high-speed packet traffic: Analysis and modeling of Ethernet traffic measurements. *Statistical Science*, 10(1), 1995.
- [17] I. Norros. On the use of fractional Brownian motion in the theory of connectionless networks. *IEEE JSAC*, 1995.
- [18] N.G. Duffield. Economies of scale in queues with sources having power-law large deviation scalings. *J. Appl. Prob.*, 33, 1996.
- [19] A. Erramilli, O. Narayan, and W. Willinger. Experimental queueing analysis with long-range dependent packet traffic. *IEEE/ACM Trans. on Networking*, 4(2), 1996.
- [20] F. Brichet, J. Roberts, A. Simonian, and D. Veitch. Heavy traffic analysis of a storage model with long range dependent On/Off sources. *Queueing Systems*, 23, 1996.
- [21] A. Vidács. *Fractal characterization of network traffic: from parameter estimation to application, PhD dissertation*. BUTE-DTT, 2000.
- [22] 3GPP. *UTRAN Iur and Iub Interface Data Transport and Transport Signalling for DCH Data Streams*, March 2000.
- [23] M. Menth. Analytical performance evaluation of low-bitrate real-time traffic multiplexing in UMTS over IP-networks. *International Journal of Interconnection Networks*, 2(1), 2001.
- [24] I. Norros, J. W. Roberts, A. Simonian, and J. T. Virtamo. The superposition of variable bit rate sources in an ATM multiplexer. *IEEE Journal on Selected Areas in Communications*, 9(3), 1991.
- [25] *Methods for the performance evaluation and design of broadband multi-service networks, The COST 242 Final Report, Part III: Traffic models and queueing analysis, Seminar version - June 4-5*. 1996.

- [26] F. P. Kelly. Notes on effective bandwidths. *Stochastic Networks: Theory and Applications*, 4, 1996.
- [27] 3GPP. *Synchronisation in UTRAN (Stage 2)*, March 2002.
- [28] 3GPP. *Delay Budget within the Access Stratum*, May 2001.
- [29] J. Peisa and M. Meyer. Analytical model for TCP file transfers over UMTS. In *proc. 3G Wireless 2001*, 2001.
- [30] IETF RFC 2475. *An Architecture for Differentiated Services*, 1998.
- [31] IETF RFC 2205. *Resource reSerVation Protocol (RSVP)*, 1997.
- [32] L. Westberg et. al. Resource Management in Diffserv (RMD): A Functionality and Performance Behavior Overview. In *proc. PfHSN*, Berlin, Germany, 2002.
- [33] B. Hayek. A queue with periodic arrivals and constant service rate. *Probability, Statistics and Optimisation, F.P. Kelly (ed.)*, 1994.
- [34] 3GPP. *UMTS Phase 1.*, April 1999.
- [35] Daniel Wong and Teng J. Lim. Soft handoffs in CDMA mobile systems. *IEEE Personal Communications*, December 1997.
- [36] J. G. Markoulidakis et al. Mobility modeling in third-generation mobile telecommunications systems. *IEEE Personal Communications*, August 1997.
- [37] S. Park et al. Modeling and analysis of CDMA soft handoff. In *proc. IEEE 46th Vehicular Technology Conference*, New York, USA, 1996.
- [38] J. K. Kwon and D. K. Sung. Soft handoff modeling in CDMA cellular systems. In *proc. IEEE 47th Vehicular Technology Conference*, New York, USA, 1997.
- [39] Moo-Ho Cho et al. The handoff rate of two-way soft handoff scheme in DS-CDMA cellular systems. *IEICE Transactions of Communications*, E80-B(8), August 1997.
- [40] M. M. Zonoozi and P. Dassanayake. User mobility modeling and characterization of mobility patterns. *IEEE JSAC*, 15(7), September 1997.
- [41] E. Chlebus et al. Analysis of channel holding time in wireless mobile systems: Does the probability distribution of cell residence time matter? In *proc. 16th International Teletraffic Congress*, Edinburgh, UK, June 1999.

- [42] K. Tutschku et al. A framework for spatial traffic estimation and characterization in mobile communication network design. In *proc. 16th International Teletraffic Congress*, Edinburgh, UK, June 1999.
- [43] D. Lam, D. Cox, and J. Widom.
- [44] A. Nilsson and M. J. Perry. Multirate blocking probabilities: Numerically stable computations. In *proc. 15th International Teletraffic Congress*, Washington DC, USA, June 1997.
- [45] L. E. J. Brouwer. über abbildung von mannigfaltigkeiten. *Mathematische Annalen*, 71, 1910.
- [46] N.G. Duffield, J.T. Lewis, N. O'Connell, R. Russell, and F. Toomey. Entropy of ATM traffic streams: A tool for estimating QoS parameters. *IEEE JSAC*, 13(6), 1995.
- [47] Z. Dziong, M. Juda, and L.G. Mason. A framework for bandwidth management in ATM networks—aggregate equivalent bandwidth estimation approach. *IEEE/ACM Trans. on Networking*, 5(1), 1997.
- [48] H. Saito. Dynamic resource allocation in ATM networks. *IEEE Communications Magazine*, 5, 1997.
- [49] H. Saito et al. Innovations of circuit/path operations in ATM networks—Self-sizing network. *NTT Rev.*, 8(1), 1996.
- [50] S. Ohta and K. Sato. Dynamic bandwidth control of the virtual path in an asynchronous transfer mode network. *IEEE Trans. on Communications*, 40(7), 1992.
- [51] S. Shioda and H. Saito. Real-time cell loss ratio estimation and its application to ATM traffic controls. In *proc. IEEE INFOCOM*, Kobe, Japan, 1997.
- [52] P.W. Glynn and W. Whitt. Logarithmic asymptotics for steady-state tail probabilities in a single-server queue. *J. Appl. Prob.*, 31A, 1993.
- [53] N.G. Duffield and N. O'Connell. Large deviations and overflow probabilities for the general single-server queue, with applications. *Proc. Cam. Phil. Soc.*, 118, 1994.
- [54] B. Tsybakov and N.D. Georganas. On self-similar traffic in ATM queue: Definitions, overflow probability bound, and cell delay distribution. *IEEE/ACM Trans. on Networking*, 5(3), 1997.
- [55] N. Likhanov and B. Tsybakov. Analysis of an ATM buffer with self-similar("fractal") input traffic. In *proc. IEEE INFOCOM*, Boston, USA, 1995.

- [56] L. Takács. *Introduction to the Theory of Queues*. Oxford University Press (New York), 1962.
- [57] N.G. Duffield. Exponential bounds for queues with markovian arrivals. *Queueing Systems*, 17, 1994.
- [58] G. Welch and G. Bishop. An introduction to the Kalman Filter, TR 95-041, <http://www.cs.unc.edu/~welch/kalman/kalmanintro.html>, Department of Computer Science, University of North Carolina at Chapel Hill, NC 27599-3175.
- [59] S. Molnár and A. Vidács A. Nilsson. Bottlenecks on the way towards fractal characterization of network traffic: estimation and interpretation of the Hurst parameter. In *proc. PMCCN*, Tsukuba, Japan, 1997.
- [60] G.L. Choudhury, D.M. Lucantoni, and W. Whitt. Squeezing the most out of ATM. *IEEE Trans. on Communications*, 44(2), 1996.
- [61] D.D. Botvich and N.G. Duffield. Large deviations, the shape of the loss curve, and economies of scale in large multiplexers. <ftp://stp01.stp.dias.ie/DAPG/dapg9412.ps>, 1995.
- [62] V. Paxson. Fast approximation of self-similar network traffic. <ftp://ftp.ee.lbl.gov/papers/fast-approx-selfsim.ps.Z>, 1995.
- [63] O. Melteig. Introduction to the PARASOL project. In *proc. of the 9th Nordic Teletraffic Seminar*, August 1990.
- [64] A. Vidács, S. Molnár, and I. Cselényi. The impact of long range dependence on cell loss in an ATM wide area network. In *proc. IEEE GLOBECOM*, Sydney, Australia, 1998.
- [65] R. K. Ahuja, T. L. Magnanti, and J. B. Orlin. *Network Flows: Theory, Algorithms, and Applications*. Prentice-Hall, 1993.

Publications

Journal Papers

- [J1] **Sz. Malomsoky**, S. Rácz, Sz. Nádas. Connection admission control in UMTS radio access networks. accepted to *Elsevier Computer Communications*, June, 2002.
- [J2] **Sz. Malomsoky**, Sz. Nádas, B. Sonkoly. UMTS hozzáférési hálózatok teljesítőképesség vizsgálata. *Híradástechnika*, Augusztus, 2002.
- [J3] **Sz. Malomsoky**, Sz. Nádas, B. Sonkoly. Performance Evaluation of UMTS Terrestrial Radio Access Networks. *Journal on Communications*, July, 2002.
- [J4] **Sz. Malomsoky**, A. Vidács, H. Saito. Real time VP bandwidth control for long range dependent traffic. *International Journal of Communications Systems*, (12):229-247, 1999.
- [J5] A. Vidács, **Sz. Malomsoky**, H. Saito. A simple adaptive bandwidth control for real traffic. *Advances in Performance Analysis*, 2(1):21-44, 1999.
- [J6] Sz. Nádas, S. Rácz, **Sz. Malomsoky** and S. Molnár. Connection Admission Control in All-IP UTRAN. *Submitted to IEEE Journal on Selected Areas in Communications, special issue on All-IP wireless networks*, date of submission: February, 2003.

Conference Papers

- [C1] **Sz. Malomsoky**, A. Szlávik. Mobility and traffic analysis for WCDMA networks. *International Conference on the Performance and QoS of Next Generation Networking*, Nagoya, November 2000.
- [C2] G. Fodor, G. Malicskó, **Sz. Malomsoky**. A joint radio-IP resource reservation scheme in ALL-IP 3rd generation networks. *IEEE Wireless Communications and Networking Conference*, Chicago, September 2000.

- [C3] **Sz. Malomsoky**, G. Tóth, Sz. Nádas, P. Zarándy. Simulation based GPRS network dimensioning. *ITC Specialist Seminar on Mobile Networks*, Lillehammer, March 2000.
- [C4] G. Fodor, G. Leijonhufvud, **Sz. Malomsoky**, A. Rácz. Comparison of call admission control algorithms in ATM/AAL2 based 3rd generation mobile access networks. *IEEE Wireless Communications and Networking Conference*, New-Orleans, September 1999.
- [C5] A. Vidács, **Sz. Malomsoky**, H. Saito. Real-time cell loss ratio estimation for bursty and self-similar traffic. *International Conference of the Performance and Management of Complex Communication Networks (PMCCN97), Workshop 2*, Tsukuba, November 1997.
- [C6] **Sz. Malomsoky**, A. Vidács, H. Saito. Bandwidth control and its applicability based on queue length monitoring. *International Conference of the Performance and Management of Complex Communication Networks (PMCCN97), Workshop 2*, Tsukuba, November 1997.
- [C7] A. Faragó, T. Cinkler, V.T. Hai, **Sz. Malomsoky**. Joint planning of the physical and logical configuration for ATM networks. *Networks'96*, Sydney, November 1996.

Presentations

- [R1] **Sz. Malomsoky**. Traffic planning in a WCDMA network. *The 3GSM World Congress*, Cannes, February 2001.

Patents

- [P1] **Sz. Malomsoky**, Sz. Nádas, S. Rácz. Efficient Traffic Concentrator. Patent Application filed in September 2002.
- [P2] **Sz. Malomsoky**, Sz. Nádas, S. Rácz. Connection admission control in packet-oriented, multi-service networks. Patent Application filed in March 2002.
- [P3] **Sz. Malomsoky**, Sz. Nádas, S. Rácz. Protocol multiplexing. Patent Application filed in March 2002.
- [P4] **Sz. Malomsoky**, I. Szabó, S. Rácz. Facilitating reliable connection admission control for telecommunications system using AAL2 signaling. Patent Application filed in March 2001.
- [P5] Pál Zarándy, **Sz. Malomsoky**. Method to transfer parallel TCP connections of an UMTS subscriber. Patent Application filed in December 2000.

- [P6] **Sz. Malomsoky**. Randomized packet arrival's process in UTRAN. Patent Application filed in October 2000.
- [P7] F. Máthé, **Sz. Malomsoky**. Transcoding data in a packet switched communication network supporting radio interfacing by selecting a transcoding processor and/or network portion. Patent Application filed in August 1999, US patent granted in 2003, Pat. No. 6,512,918.
- [P8] **Sz. Malomsoky**, A. Vidács, H. Saito. Virtual path bandwidth control apparatus and virtual path bandwidth dimensioning method. Patent Application filed in December 1997.
- [P9] W. Holender, **Sz. Malomsoky**. Adaptive virtual path dimensioning method especially for paths defined on telecommunications network using entropy rate function as blocking measure and balancing loads on links by equalizing blocking probabilities and determining allocation of physical resources. Patent Application filed in July 1995, two US patents granted in 2001 and 1999 with the numbers 6,304,639 and 5,872,918, respectively.