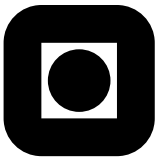
Quality of Service, Traffic Conditioning and Resource Management in Universal Mobile Telecommunication System (UMTS)

Frank Yong Li

A DISSERTATION SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF

DOKTOR INGENIØR



Department of Telematics
Norwegian University of Science and Technology
January 2003

Norwegian University of Science and Technology
Department of Telematics
N-7491 Trondheim
Norway

Doktor ingeniøravhandling 2003:6
Rapportnr. 1:2003

ISBN 82-471-5553-2
ISSN 0809-103X

Printed by NTNU-trykk, 2003.

To my family

Abstract

The successful deployment of Universal Mobile Telecommunication System (UMTS) is heavily dependent on Quality of Service (QoS) to be achieved. This thesis addresses a few facets of the QoS issues in UMTS, including traffic shaping and policing, conformance consistency, Call Admission Control (CAC) and resource allocation.

The main focus of this thesis is traffic conditioning related issues for QoS provisioning in UMTS. Assuming an end-to-end QoS scenario supported by IntServ or/and Diffserv architectures, the thesis initially presents an all nodes traffic conditioning-enabled framework in UMTS. Under this framework, the traffic generated at each User Equipment (UE) is regulated by a traffic shaper in the form of a token bucket, and the conformance of the traffic flow is policed at the policing node. The performance of imposing traffic shaping at the UE is studied and compared with the case without shaping. Next, having observed that the performance of the traffic conditioned system is sensitive to the values of the token bucket parameters, the thesis proposes a heuristic approach for searching local and global QoS-aware token bucket parameters. By tuning the system operating at the obtained 'optimal' shaping parameters, the requirements for all concerned QoS attributes are guaranteed. Furthermore, the thesis studies conformance consistency in a traffic conditioned multi-hop network, by monitoring the conformance status of a traffic flow using an identical token bucket for both traffic shaping and policing. In the presence of variable packet size, the thesis gives a quantitative result, for a simple case, on how much percent of the originally conformant packets may misbehave at further policing node(s). The performance of the aggregated traffic flows and the measures to minimize the effect of conformance deterioration are also studied in the thesis.

Another facet of the QoS issues in UMTS, CAC together with resource allocation, is also studied in the thesis. A priority-oriented framework for QoS management of multimedia services in UMTS is proposed. Based on a traffic class priority definition, the framework is implemented through a priority-oriented CAC, channel congestion control and adaptive bandwidth allocation.

Preface

This dissertation is submitted in partial fulfillment of the requirements for the degree of *doktor ingeniør* (Dr. Ing) at the Department of Telematics, Norwegian University of Science and Technology (NTNU). The work presented here has been carried out at the Department of Telematics, NTNU, during the period from July 1999 to December 2002, under the supervision of Associate Professor Norvald Stol.

Three years of my Dr.ing studies has been funded by Telenor R&D and 6 months by the Department of Telematics, NTNU. During these years, I have being involved, as a main participant, in a Telenor-NTNU collaboration under *TURBAN* project, which stands for TURBo codes, Access and Networking Technology related to UMTS.

Production note: This thesis, as well as all my papers produced over the past few years, has been written using LaTeX. All simulation results were obtained from self-made models using OPNET Modeler simulator.

Acknowledgements

First of all, I would like to express my sincere acknowledgement to my supervisor, Associate Professor Dr.ing Norvald Stol, at the Department of Telematics, NTNU. Norvald has been acting not only as an enthusiastic tutor but also as an always available discussion partner of my studies through these years. I do appreciate our stimulating discussions and his insightful advises on various aspects of my work.

I am grateful to all colleagues at the Dept. of Telematics, NTNU. Particularly I would like to thank Prof. Bjarne E. Helvik. His not so often but always to the point comments make my work more valuable than I thought. I am indebted to Prof. Steinar Andresen for initiating my PhD study and continuous encouragement. Special thanks to Randi S. Flønes, Jarle Kotsbak and Otto Witter for various kinds of help, to Pål S. Sœther and Asbjørn Karstensen for keeping my machine(s) running under Unix, Linux and NT platforms, and

to Hein Meling for much more than providing the LaTeX template of this thesis to me.

Besides a general gratitude to Telenor R&D for generously sponsoring my research fellowship, I would like to say thank you very much to my contact person Stein-Wegard Svaet at Telenor R&D for various kinds of help during this period, to Thor Gunnar Eskedal and Zaw-sing Su for their suggestions on my potential thesis topics at the early stage of my study, and to Dr. Terje Jensen and Dr. Josef Noll for valuable discussions during TURBAN meetings. Furthermore, I would like thank our project leader, Prof. Geir Øien at the Dept. of Telecommunications, NTNU, for his very positive comments on my research work, and to Associate Professor Andrew Perkis for initiating the TURBAN project which gave me a chance to pursue my PhD studies. Good luck to my TURBAN fellow students Tung Pham Phan and Ola Jutlund in pursuing their PhDs.

My acknowledgement also goes to Prof. Arne Svensson from Chalmers University of Technology for hosting my pre-PhD research in Sweden.

All my publications as a PhD candidate have been reviewed at international level. I would like to thank all anonymous reviewers who have commented and proposed modifications to my papers. I also enjoy numerous face-to-face discussions with people I met during the conferences I participated, and on-line discussions with researchers from various countries. Some of these discussions have been very constructive to my research work.

Last, but not least, I would like to thank my wife, Yue Li, for her patience and full time *'work'* at home in supporting my PhD studies. I wish my daughter and my son would understand what their father has done, sooner or later. I feel indebted to my mother for endless love and understanding. My father would feel proud of my PhD if he knew I finally get it. I thank my two sisters in China for decades-lasting understanding and support. I devote this thesis to the whole family of mine.

Trondheim, December 2002
Frank Yong Li

The author has published the following papers, where the papers with highlighted sequence numbers constitute part II of this dissertation.

[1] Frank Y. Li, “Local and Global QoS-aware Token Bucket Parameters Determination for Traffic Conditioning in 3rd Generation Wireless Networks”, *ACM/Kluwer Wireless Networks Journal,* revised version submitted, December 2002.

[2] Frank Y. Li and Norvald Stol, “A Study on Traffic Shaping, Policing and Conformance Deterioration for QoS Contracted Networks”, in *Proceedings IEEE Global Telecommunications Conference (GLOBECOM),* Taipei, Taiwan, November 17-21, 2002, paper CQRS-03-5.

[3] Frank Y. Li and Norvald Stol, “QoS Provisioning using Traffic Shaping and Policing in 3rd-generation Wireless Networks”, in *Proceedings IEEE Wireless Communications and Networking Conference (WCNC),* Orlando, FL. USA, March 17-21 2002, pp. 139-143.

[4] Frank Y. Li, “Local and Global QoS-aware Token Bucket Parameters Determination for Traffic Conditioning in 3rd Generation Wireless Networks”, in *Proceedings European Wireless (EW),* Florence, Italy, February 25-28, 2002, pp. 362-368.

[5] Frank Y. Li and Norvald Stol, “Providing Conformance of the Negotiated QoS Using Traffic Conditioning for Heterogeneous Services in WCDMA Radio Access Networks”, in *Proceedings Norwegian Informatics Conference (NIK),* Tromsø, Norway, November 26-28, 2001, pp. 117-128.

[6] Frank Y. Li, Norvald Stol, Tung Thanh Pham and Steinar Andresen,“A Priority-oriented QoS Management Framework for Multimedia Services in UMTS”, in *Proceedings The 4th International Symposium on Wireless Personal Multimedia Communications (WPMC),* Aalborg, Denmark, September 9-12, 2001, pp. 681-686.

[7] Frank Y. Li, “Embodying Traffic Conditioning at Radio Access Network for QoS Provision in Next-Generation Wireless Networks”, in *Proceedings IFIP Workshop on IP and ATM Traffic Management (WATM & EUNICE),* Paris, France, September 3-5, 2001, pp. 25-32.

[8] Tung Pham Thanh, Andrew Perkis and Frank Y. Li, “Call Admission Control Algorithm for Multichannel Users in Hierarchical Cellular Systems”, in *Proceedings IEEE International Conference on Third Generation Wireless and Beyond (3Gwireless),* San Francisco, CA. USA, May 30-June 2, 2001, pp. 974-979.

[9] Frank Y. Li and Norvald Stol, “A Priority-oriented Call Admission Control Paradigm with QoS Re-negotiation for Multimedia Services in UMTS,” in *Proceedings IEEE Vehicular Technology Conference (VTC),* Rhodes, Greece, May 6-9, 2001, pp. 2021-2025.

[10] Yong Li, Arne Svensson and Sorour Falahati, “Hybrid Type-II ARQ Schemes with Scheduling for Packet Data Transmission over Rayleigh Fading Channels”, in *Proceedings The 3rd ITG Conference on Source and Channel Coding,* Munich, Germany, January 17-19, 2000, pp. 293-299.

List of Abbreviations

3G	3rd Generation
3GPP	3rd Generation Partnership Project
AF	Assured Forwarding
ATM	Asynchronous Transfer Mode
BS	Bearer Service
CAC	Call Admission Control
CDMA	Code Division Multiple Access
CL	Controlled Load
CN	Core Network
CTB	Complex Token Bucket
DiffServ	Differentiated Service
EF	Expedited Forwarding
FIFO	First In First Out
FlowSpec	Flow Specification
GGSN	Gateway GPRS Serving Node
GoS	Grade of Service
GPRS	General Packet Radio Service
GS	Guaranteed Service
IETF	Internet Engineering Task Force
IntServ	Integrated Service
IP	Internet Protocol
ISDN	Integrated Services Digital Network
JPEG	Joint Photographic Experts Group
LB	Leaky Bucket

MPEG	Moving Picture Experts Group
MTU	Maximum Transmission Unit
PDF	Probability density function
PDP	Packet Data Protocol
PHB	Per-Hop Behavior
PSTN	Public Switched Telephone Network
QoS	Quality of Service
RFC	Request for Comment
RNC	Radio Network Controller
RNS	Radio Network Subsystem
RRM	Radio Resource Management
RSVP	Resource ReSerVation Protocol
RTP	Real-time Transport Protocol
SDU	Service Data Unit
SIR	Signal-to-Interference Ratio
SLA	Service Level Agreement
SLS	Service Level Specification
STB	Simple Token Bucket
TB	Token Bucket
TBC	Token Bucket Counter
TCP	Transmission Control Protocol
TE	Terminal Equipment
TS	Technical Specification
UDP	User Datagram Protocol
UE	User Equipment

UMTS	Universal Mobile Telecommunication System
UTRAN	UMTS Terrestrial Radio Access Network
VBR	Variable Bit Rate
VoIP	Voice over Internet Protocol
WCDMA	Wideband Code Division Multiple Access
WWW	World Wide Web

Contents

Part I

Introduction

Part I: Introduction

Third Generation (3G) mobile networks, especially Universal Mobile Telecommunication System (UMTS), have been an extremely hot topic in recent years, both academically and commercially. In addition to its up to 2 Mbps high bitrate services, UMTS also promises to provide Internet Protocol (IP)-based multimedia services. While 3G standardization work is still undergoing intensively by 3rd Generation Partnership Project (3GPP), the novel trend is the convergence of traditional mobile telecommunications and IP technologies. As one of the most important issues in 3G networks, end-to-end Quality of Service (QoS) plays a vital role to the success of UMTS.

This thesis addresses some of the issues within the area of QoS in UMTS. The thesis is composed of two parts, an introduction chapter (Part I) and four included papers (Part II). The introductory part presents background information to the issues dealt with in Part II of this thesis. It functions as a guidance for the reader to understand what I have done during the past three and half years. Part II deals with, more specifically, four aspects within this area.

The chapter starts with a general overview of QoS in UMTS in Section 1. The architecture, end-to-end QoS scenarios for QoS provisioning are outlined first. Then the section explains how the work in this thesis falls into the umbrella of QoS in UMTS. As an important functionality for QoS provisioning, traffic conditioning is presented in Section 2, followed by a discussion on the concept of traffic shaping and policing, as well as shaping/policing schemes. Employing Token Bucket (TB) algorithm as the *default* shaping algorithm, the difficulty of determining suitable TB parameters is addressed in Section 3, which leads to a heuristic approach on TB parameters determination. Next, in order to evaluate the performance of the shaped traffic flow, we investigate what happens at other intermediate node(s) along the path to the peer node in Section 4 and Section 5. Another aspect of the thesis work, Call Admission Control (CAC) and radio resource management, is addressed in Section 6. Later on, Section 7 goes briefly through two interesting potential

research topics. Finally, Section 8 summarizes the contributions of each included paper in Part II.

I have to stress here that it is not the objective of this thesis to provide a comprehensive overview of the QoS issues in UMTS. For authoritative descriptions on this issue, please refer to the latest version of 3GPP Technical Specifications (TSs), for instance TS 23.107 [3] and TS 23.207 [2]. What is showing in this chapter intents to present a perspective of a big picture of QoS in UMTS, and then subsequently presents deeper insight into some of the relevant issues within this scope, which probably still touches only partially the topic of QoS in UMTS.

1 Overview

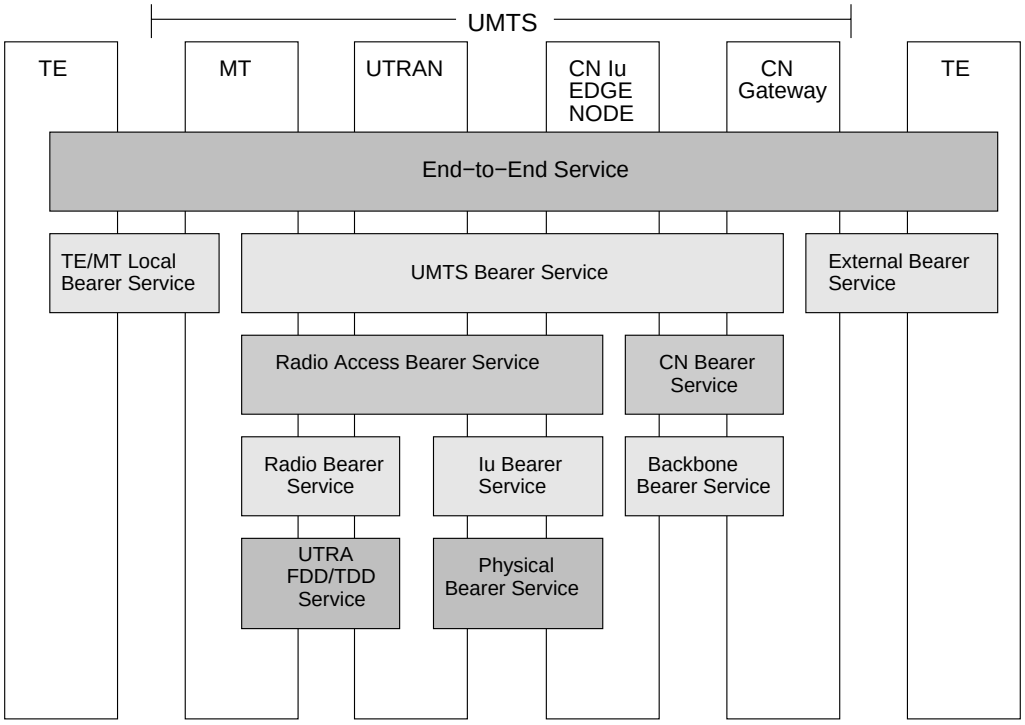
A significant contribution of 3GPP Release 99 is the definition of a layered architecture for the provision of end-to-end QoS in UMTS. The contracted QoS is supported through interaction of bearer services at different layers, which means each bearer service on a specific layer offers its individual services using services provided by the layers below.

1.1 QoS Architecture in UMTS - a 3GPP Perspective

The objective of the UMTS services is to provide appropriate end-to-end QoS guarantees to the end users. Here end-to-end means from a Terminal Equipment (TE) to another. On its way from one TE to another TE, the traffic is conveyed over several underlying networks (not only UMTS). Figure 1 shows a layered service architecture defined by 3GPP [3], which promises guaranteed QoS for multimedia services.

The architecture provides mapping of end-to-end QoS to services provided by the User Equipment (UE), UMTS Terrestrial Radio Access Network (UTRAN), Core Network (CN), and external networks. The external networks could be traditional Public Switched Telephone Network (PSTN), Integrated Services Digital Network (ISDN), or IP-based network, depending on type of the service.

Based on this architecture, 3GPP has defined four traffic classes with associated QoS attributes, as well as QoS management function in UMTS [3]. The traffic classes are distinguished based on their delay sensitivity and the defined four classes are, conversational class, streaming class, interactive class and background class. The QoS attributes are maximum bitrate, delivery order, maximum Service Data Unit (SDU) size, SDU format information, SDU error ratio, residual bit error ratio, delivery of erroneous SDUs, transfer



UMTS OoS Architecture

Figure 1: QoS architecture in UMTS defined by 3GPP.

delay, guaranteed bit rate, traffic handling priority and allocation/retention priority. The QoS management functions are implemented in two planes, i.e., the control plane and the user plane.

1.2 End-to-End QoS Scenarios in UMTS

Even though the QoS architecture shown in Figure 1 covers both Circuit Switched (CS) and Packet Switched (PS) domains, more recent 3GPP releases (Release 5 onwards) focus mainly on packet switched networks. Furthermore, the novel trend in 3G networks is to integrate more and more Internet Engineering Task Force (IETF) standards, Request For Comments (RFCs), into 3GPP TSs. Consequently, six all-IP end-to-end QoS scenarios have been proposed by 3GPP [1], as tabulated below in Table 1 [2, 26]:

Basically, the scenarios can be categorized according to the capability of each network device and the location of the IP bearer service manager. Each network device could be a

[1]Reduction to fewer scenarios are currently under discussion, but there are still 6 scenarios in Version 5.3.0 of TS 23.207.

UE, a Gateway GPRS Support Node (GGSN) or an external IP network. The symbol **X** in each scenario specifies where the IP Bearer Service (BS) manager and Packet Data Protocol (PDP) context are implemented, and furthermore whether the network device supports one of the two IP QoS architectures presented in the next subsection or not, that is, Integrated Services (IntServ) [11] architecture/Resource ReSerVation Protocol (RSVP) [12] and Differentiated Services (DiffServ) architecture. For example, scenario 1 does not support IntServ/RSVP at the UE, but from GGSN to external IP network, it is DiffServ capable.

Table 1: Summary of the 3GPP end-to-end all-IP QoS scenarios.

Scenario 1	UE	Gateway (GGSN)	External IP Net
PDP Context	**X**	**X**	
IP BS Manager		**X**	
RSVP (IntServ)			
DiffServ		**X**	**X**

Scenario 2	UE	Gateway (GGSN)	External IP Net
PDP Context	**X**	**X**	
IP BS Manager	**X**	**X**	
RSVP (IntServ)			
DiffServ	**X**	**X**	**X**

Scenario 3	UE	Gateway (GGSN)	External IP Net
PDP Context	**X**	**X**	
IP BS Manager	**X**	**X**	
RSVP (IntServ)	**X**		
DiffServ	**X**	**X**	**X**

Scenario 4	UE	Gateway (GGSN)	External IP Net
PDP Context	**X**	**X**	
IP BS Manager	**X**	**X**	
RSVP (IntServ)	**X**	**X**	
DiffServ		**X**	**X**

Scenario 5	UE	Gateway (GGSN)	External IP Net
PDP Context	**X**	**X**	
IP BS Manager		**X**	
RSVP (IntServ)			
DiffServ		**X**	**X**

Scenario 6	UE	Gateway (GGSN)	External IP Net
PDP Context	**X**	**X**	
IP BS Manager		**X**	
RSVP (IntServ)		**X**	
DiffServ		**X**	**X**

1.3 IntServ and DiffServ Architectures from IETF

To evolve traditional best effort only IP network to a modern end-to-end QoS capable Internet, a fundamental change on Internet architecture is necessitated. Among many service models and mechanisms proposed by IETF, two distinct approaches, IntServ/RSVP and DiffServ architectures, are most notable solutions.

Aiming at providing end-to-end QoS, the IntServ architecture is characteristic of resource reservation. The application must set up paths and reserve resources before data transmission. The architecture uses explicit setup mechanism, e.g., RSVP, to convey information to routers so that they can provide requested resources to the flows. In addition to best effort service, IntServ is mainly focusing on two new services, i.e., the Controlled Load (CL) service [75] and the Guaranteed Service (GS) [66].

DiffServ intends to meet the need for simple and coarse methods of providing differentiated classes of services over Internet. The DiffServ architecture is characterized by a relative-priority scheme, which marks packets into predefined service classes with different codepoints. Packets in different classes receive different services. By aggregating packets with the same codepoint, the flow receives a particular forwarding treatment, or Per-Hop-Behavior (PHB), at each network node. To enable QoS, preferential treatments on, for example, buffer management and scheduling mechanisms, are given to flows marked with Expedited Forwarding (EF) [38] or Assured Forwarding (AF) [34] classes.

The most salient distinction between these two approaches is their different treatment of packet streams. IntServ emphasizes on guaranteeing QoS on a per-flow basis. It requires explicit signaling to reserve network resources along the path to the other end. Having difficulty in monitoring and processing possibly millions of flow states on a per-flow basis, IntServ is harassed by the problem of scalability. On the other hand, instead of focusing on per-flow treatment, DiffServ prioritizes the flows on an aggregated basis, i.e., a set of micro-flows with similar service requirements are treated the same based on its codepoint. By different treatment of a limited number of classes of services, the performance of the aggregated flow is guaranteed based on its PHB. However, DiffServ applies only to large scale networks and thus cannot provide end-to-end QoS. As a solution, a framework combining both IntServ and DiffServ has been proposed by IETF [5], where the DiffServ domains are viewed as a network element in the total IntServ end-to-end path. Based on this approach, a promising framework for end-to-end QoS could be embodying IntServ in the edge clouds and DiffServ in the core network.

1.3.1 SLA and SLS

Since QoS is meaningful over the whole Internet (end-to-end), the service is provided based upon a set of technical parameters that both the customers and the service providers will have to agree upon. This is known as a Service Level Agreement (SLA). The SLA is a service contract between a customer and a service provider that specifies the forwarding service a customer should receive [6]. An SLA may include traffic conditioning rules which constitute a traffic conditioning agreement in whole or in part. The SLA contains both technical and non-technical terms and conditions.

The technical specification of the IP connectivity service is given in Service Level Specifications (SLSs). An SLS is a set of technical parameters and their values, which together define the service, offered to a traffic stream by a DiffServ domain [30]. The SLS is associated with several attributes, for example, ingress and egress interfaces, flow identification, traffic envelop and traffic conformance parameters. The traffic conformance parameters include data flow specifications described below.

1.3.2 FlowSpec

In order to reach a service level agreement, the data flow must be represented by a set of parameters. For example, source and destination addresses, available path bandwidth, minimum path latency, path MTU, token bucket specification etc. [70, Chapter 15]. The TOKEN_BUCKET_TSPEC (see description below), which is part of the sender traffic specification and takes the form of a token bucket specification r, b plus a peak rate p, maximum packet size M and minimum policed unit m, is one of these parameters [67]. It is used for setting up a negotiable contract [76].

To represent a data flow, different notations, for instance, Traffic Specification (Tspec), TOKEN_BUCKET_TSPEC, or FlowSpec, are used in the literature. They might have slightly different meanings. But to simplify our notation, the thesis purposely does not distinguish the difference among them in the context. Therefore, we claim that the information about a data source's generated traffic is explicitly represented by the following 5-tuples (p, r, b, M, m) [67], [2, Annex C], often referred to as FlowSpec in this thesis.

- Peak traffic rate p, measured in bytes of IP datagrams per second. Values of this parameter may range from 1 byte per second to 40 terabytes per second;
- Token bucket rate r, measured in bytes of IP datagrams per second permitted by the token bucket. Values of this parameter may range from 1 byte per second to 40

terabytes per second;

- Token bucket size b, measured in bytes. Values of this parameter may range from 1 byte to 250 gigabytes;
- Maximum packet size M, measured in bytes, the biggest packet unit allowed to enter the network. Any packets of larger size sent into the network may not receive QoS-controlled service, since they are considered not to meet the traffic specification;
- Minimum policed unit m, measured in bytes. This size includes the application data and all protocol headers at or above the IP level (IP, Transmission Control Protocol (TCP), User Data Protocol (UDP), Real-time Transport Protocol (RTP), etc.). All IP datagrams less than size m are treated as being of size m for purposes of resource allocation and policing.

As a summary of this subsection, the end-to-end QoS in Internet is provided based on either IntServ/RSVP, DiffServ architecture, or a combination of both. The service is achieved according to the contract negotiated between the customer and the service provider, based on an agreed upon SLA. The FlowSpec, represented by 5-tuples (p, r, b, M, m), constitutes the basis for SLA negotiation. Even though the SLA encompasses more contents, it is assumed, for the sake of simplicity in this thesis, that the SLA contract is established in terms of these 5-tuples.

1.4 About Thesis Topic

According to 3GPP definitions, the QoS management functions cover both control plane and user plane. The control plane is responsible for coordinating overall management procedures, such as signaling, DiffServ edge function or RSVP function. The user plane is responsible for transport of user data traffic within the limits defined by QoS attributes. The QoS management functions of the UMTS bearer service in the user plane are summarized as follows [3, 2]:

- **Classification.** The classification function assigns user data units received from external or local services for appropriate treatment, usually according to the header information.
- **Traffic conditioning.** The traffic conditioner, which is optional in the UE and mandatory at the Gateway node, provides conformance of the user data traffic with the QoS attributes of the relevant UMTS bearer service.

– **Mapping.** The mapping function marks each data unit with the specific QoS indicated to the bearer service performing the transfer of the data unit. The function includes mapping among different protocol layers and mapping between the UMTS domain and the external domains.

– **Resource management.** Each of the resource managers of a network entity is responsible for a specific resource. They perform scheduling, queueing management, and power control for the radio bearer, in order to distribute the available resources among the established services appropriately.

Having in mind the big picture of QoS in UMTS, we are solely interested in the user plane functions in this thesis. In other words, the signaling in the control plane is not considered. More precisely, among the QoS management functions in the user plane discussed above, the thesis addresses such issues as call admission control, resource allocation, traffic shaping, traffic policing, and conformance consistency. Furthermore, adopting QoS scenario 4 shown in Table 1, we face a framework of encompassing IntServ in UTRAN and DiffServ in CN. Therefore we have a traffic conditioning-enabled network, which constitutes the fundamental framework for our more detailed discussions in the subsequent sections.

Before proceeding the descriptions in the following sections, we initiate briefly the basic concepts of traffic shaping, admission control, and policing here. *Traffic shaping* is a technique to control the volume of traffic entering the network, along with controlling the rate at which it is transmitted. *Admission control* is a measure to discriminate which traffic is admitted to the network at the connection set-up phase. *Policing* is a measure to determine, on a hop-by-hop basis within the network beyond the ingress point, whether the traffic being presented is compliant with pre-negotiated traffic shaping policies or not. Typically, traffic shaping and admission control need to be implemented at the network edge node, and traffic policing , on the other hand, is employed on intermediate nodes on the way to the other end of the network. The reader should also note that traffic shaping is not always necessary in a network. In fact it depends on what type of applications we are considering. For instance, a real-time application with strict jitter demands may not be traffic shaping applicable, since traffic shaping would introduce more delay.

Admission control and traffic shaping can be used as stand-alone technologies, or used integrally with other technologies, such as in IntServ architecture [25]. In this thesis, we intent to treat call admission control and traffic shaping/policing separately. In the following descriptions, the traffic conditioning and conformance evaluation related issues are

presented in Sections 2, 3, 4 and 5, and the admission control and resource management issues are addressed later in Section 6.

2 Traffic Conditioning in UMTS

No matter which approach, IntServ, DiffServ, or IntServ over DiffServ is employed, traffic conditioning always plays an important role in an end-to-end QoS model. Traffic conditioning has three major functions [56]. First, a flow can regulate its traffic according to a specification so that the network knows the characteristic of the traffic flow to be expected. Second, the network can allocate its resources, based upon the traffic specification of the flow, to reach a service agreement at setup phase. Third, the network can monitor whenever necessary the flow's behavior and ensure that the flow is performing as it promised.

2.1 Traffic Shaping and Policing: a Reference Model

A generic reference model for a traffic conditioning capable network with policing and shaping functions, similar to the one shown in [54, Chapter 6], is shown in Figure 2. Here a Traffic shaper regulates packets within a traffic flow to make sure it satisfies a predefined traffic specification. A traffic policer verifies whether incoming traffic flow conforms to a specific contract. The primary objective of traffic policing is to prevent users from violating the contract so that other users suffer from this misuse. Traffic conditioning is a control function which performs rules, including metering, marking, shaping, and policing to the traffic flows.

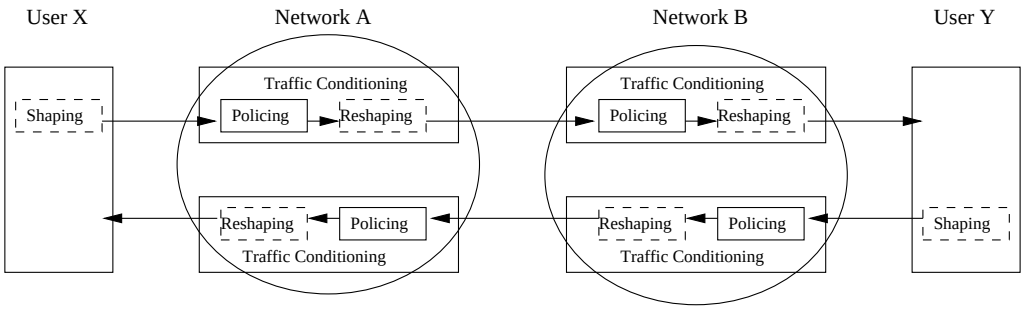


Figure 2: Reference network for placement of shaping and policing functions.

As depicted in the figure, a traffic conditioner, which may contain meter, marker, dropper and shaper, is usually located at the network edge node. In Figure 2, the traffic conditioner encompasses only a traffic shaper at each source node (User X or User Y), and both traffic policer and (re)shaper at each network node (Network A or Network B). The reshaping function may be performed to the traffic flow at the intermediate node(s), when necessary.

Throughout the context, we differentiate the terminologies *shaping* and *reshaping* by implementing shaping function *only* at the source node and reshaping function *only* at the intermediate network node(s). In general terms, a flow may be shaped at the source node, and policed or reshaped into profile again in case of misbehaving at other node(s).

2.2 Traffic Shaping and Policing in UMTS

Applying the above generic reference model into the UMTS architecture, we obtain an IntServ for UTRAN and DiffServ for CN framework as a traffic conditioning-enabled UMTS network. With this framework, we can employ traffic shaping at the UE(s) and traffic policing at the Gateway (GGSN) [2]. Moreover, referring to Figure 2, we have UE as User X, UTRAN as Network A, and CN as Network B. The studies in Part II of this thesis cover nodes UE, UTRAN in Papers B, C and D, and nodes UE, UTRAN and CN in Paper A.

Now we have a traffic conditioning-enabled UMTS network which provides conformance between the negotiated QoS for an application and the data unit traffic by performing shaping or policing at the UE or the Gateways node(s) respectively. Usually in this thesis, it is assumed that the UE conditioner has only traffic shaper embedded and the Gateway node conditioner has both traffic policer and reshaper implemented.

Among all issues relating to this framework, the thesis is mainly interested in the performance of the traffic conditioned system in terms of several concerned QoS attributes [46]. The shaping and policing schemes, among others, deserve further discussions below. For shaping scheme, the token bucket algorithm is further discussed in Subsection 2.3. For traffic policing, there are two kinds of policing functions applied in this thesis. In Paper B and C, we consider channel congestion status as the criterion for policing function. In Paper A, we employ a token bucket with the same (r, b) parameters as the policer. The former scheme is actually not a '*recommended*' policing scheme described in RFC 2212 [66], but a specific measure employed in this thesis. The latter case will be further addressed in

[2] We hypothesize in this thesis that the RNC is also traffic conditioning-enabled.

Sections 4 and 5.

2.3 Traffic Shaping/Policing Algorithms

The first paper on traffic shaping was published in early 1980's [62], where the output process was regulated by not permitting interdeparture times less that a specific value. Later on the Leaky Bucket (LB) algorithm and the TB algorithm were adopted as the default shaping algorithms for Asynchronous Transfer Mode (ATM) and IP networks respectively.

This thesis deals with only TB algorithm together with its variants. It can be adopted either as a traffic shaper or as a traffic policer, depending on the location and the implementation of the algorithm.

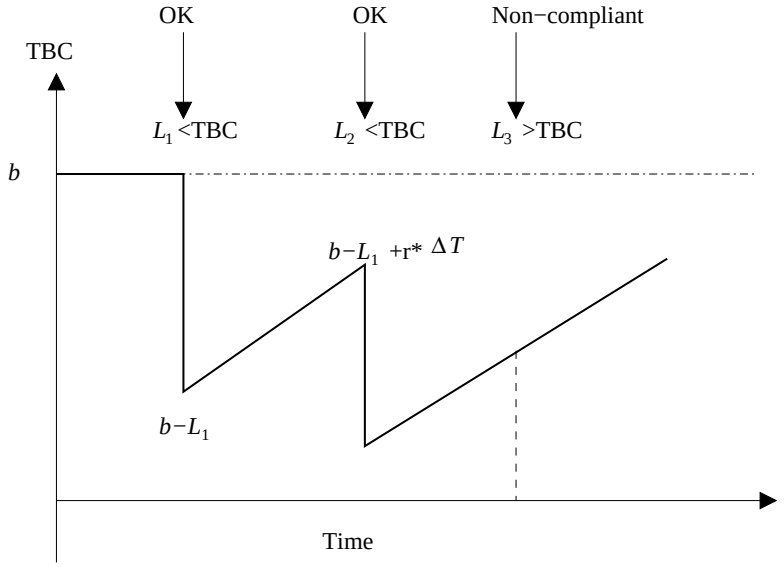


Figure 3: Token bucket algorithm.

2.3.1 Token Bucket Algorithm

To verify the conformance of the traffic, a well known token bucket algorithm has been described in RFC 2215 [67]. A token bucket traffic specification consists of two parameters: a token replenishment rate r and a bucket size b. The token rate r specifies the continually sustainable data rate, i.e., the number of bytes of IP datagrams per second permitted by the token bucket. The bucket size b specifies the amount by which the data rate can exceed r

for short periods of time. One internal variable Token Bucket Counter (TBC) is used to record the number of the remaining tokens at any time.

The algorithm, adopted by 3GPP in [3, Annex B], operates as follows :

- TBC is usually increased by r in each small time unit. However, TBC has upper bound b and the value of TBC shall never exceed b;
- When a packet $\#i$ with length L_i arrives, the receiver checks the current TBC. If the TBC value is equal to or larger than L_i, the packet arrival is judged compliant, i.e., the traffic is conformant. At this moment tokens corresponding to the packet length is consumed, and TBC value decreases by L_i.
- When a packet $\#j$ with length L_j arrives, if TBC is less than L_j, the packet arrival is non-compliant, i.e., the traffic is not conformant. *At this moment, no tokens are consumed and the value of the TBC continues to increase by $r*\Delta T$ since last arrival* [3].

2.3.2 Variants of the Token Bucket Algorithm

There exists several variants of the token bucket algorithm in the literature. For example, *Send-Now* and *No-Delay* do not allow any delay on arriving traffic, while *Send-Smooth* allows a packet to be delayed up to the point when the next packet arrives [71, 57]. *Token bucket with leaky bucket rate control* has the capacity of limiting the peak rate p of a source [56]. This variant is also referred to as *Complex Token Bucket (CTB)* in [29].

In this thesis, we refer to the *standard* TB algorithm depicted in Figure 3 as *traffic tagging* where a packet not finding enough tokens upon arrival is simply deemed and tagged as non-conformant. Another variant, referred to as *traffic conforming* in this thesis, delays a packet up to its conformance point if there are not enough tokens upon arrival. In Paper B of part II, this version is also referred to as *Simple Token Bucket* (STB) algorithm. These two forms of the token bucket algorithm are frequently used in this thesis.

The reason why we employ both traffic tagging and traffic conforming in this thesis is explained as follows. We adopt TB algorithm as both traffic shaper and traffic policer. When a TB is placed at the source node, it functions as a traffic shaper. With traffic conforming, it shapes the traffic so that all packets not larger than bucket size are conformant.

[3] The last sentence of this paragraph in 3GPP TS 23.107v5.5.0 [3] is: *In this case, the value of TBC is not updated*. But the author of this thesis insists on re-writing this sentence as above because the value of the TBC is indeed updated by $r*\Delta T$ at this moment, since last arrival. The author has proved through simulation that, without this modification, it may lead to wrong result.

With traffic tagging, it only acts as a marker, which marks the packets as conformant or not. When a TB is placed at the intermediate node, it functions as a traffic policer. Traffic tagging is in this case employed for traffic policing in the thesis. Traffic conforming may also be used at any intermediate node as a reshaper if we want to reshape the traffic flow. Table 2 summaries the functions and locations of a TB as a traffic shaper or policer used in the thesis.

Table 2: Summary of the TB functions and locations.

Variant of TB Algorithm	Location	
	Source node	Intermediate node
Traffic tagging	Shaper (Marker)	Policer
Traffic conforming	Shaper	Reshaper

A fundamental assumption on packets generation in this thesis is that no segmentation happens to the generated packets. With this assumption, packets larger than the bucket size will be judged as non-conformant. If we set bucket size at least as large as maximum packet size (Paper A), all packets are conformant after traffic conforming. However, if we allow smaller-than-MTU bucket size (Paper B and C), packets larger than b are still non-conformant after traffic conforming. With traffic tagging, there always exists both conformant and non-conformant packets, independent of whether b is larger or smaller than M.

On the other hand, if we allow segmentation at the source node, we can truncate the packets at b so that all truncated segments are not larger than the bucket size. In this case, all packets can be kept as conformant. But the shaping delay to a packet is more complex because it is then an accumulation of the shaping delays to all data segments belonging to the same packet. Nevertheless, this case is beyond the consideration of this thesis.

3 How to Determine TB Parameters

The token bucket traffic shaper consists of two parameters, r and b. In many cases, the token rate r is specified as the average data rate of a flow. But it could also be sensible to specify r as the peak rate of a flow [3]. The bucket size b decides the maximum allowable burstiness of the flow.

As described in Section 1, the values of r and b specified by the RFCs are intentionally at a large range, in order to support capacities achievable in future networks. However, one

needs to determine the specific (r, b) values when designing a traffic conditioning capable network, since the performance of the system is heavily dependent upon the values of the TB parameters.

3.1 Problem Identification

The motivation of this part of the thesis work is trying to find a sensible method for determining TB parameters in a traffic conditioned system. To identify the problem, let's first summarize four methods found in the literature.

Method I:

The RFCs do not specify how (r, b) values are determined. However, according to RFC 2212 [66], for guaranteed service in IntServ architecture, the end-to-end delay D_{ete} can be expressed as:

$$D_{ete} = \frac{b - M}{R} \times \frac{p - R}{p - r} + \frac{M + E_{tot}}{R} + D_{tot} \qquad p > R \geq r \tag{1}$$

where M is the maximum packet size, R is the bitrate of the link connected to the traffic shaper, and p is the peak rate of the source. E_{tot} and D_{tot} correspond to total accumulated values of a rate dependent error term E and a rate independent error term D [66, 28].

Setting a targeted end-to-end delay D_{ete}, we can re-write the above equation by putting r on the left hand side of the equation as:

$$r = \frac{(b - M)(p - R) + (M + E_{tot})p - (D_{ete} - D_{tot})pR}{(M + E_{tot}) - (D_{ete} - D_{tot})R} \tag{2}$$

Formula (1) implies that the token bucket size should not be smaller than the maximum packet size M. Formula (2) shows that there is a strong dependency among the bucket size, the token rate, and the link bitrate. Given D_{ete}, R and b, we can obtain the corresponding token rate r which guarantees the end-to-end delay from Formula (2) [4].

Method II:

Most straightforwardly, one can simply set r as the peak rate or the average rate of the traffic flow, and bucket size b as [3]:

$$b = N \times M \qquad (N = 1, 2, 3, ...) \tag{3}$$

[4] If we put R on the left hand side of the equation, we can also determine the required bitrate for R to satisfy D_{ete} requirement in an IntServ/RSVP framework.

where M is the maximum packet size and N is an integer. For Release 99 of 3GPP TSs, $N = 1$ is recommended.

Method III:

The bucket size decided by Methods I and II is at least as large as the maximum packet size M [5]. But as bucket size b specifies the burstiness tolerable to the network, we could also set a comparatively smaller b to have a 'smoother' traffic flow injected into the network when lower bitrate flow is concerned.

Alternatively, 3GPP [2, Annex C] has recommended to use the peak bitrate p for token bucket b calculation, and the average bitrate for token rate calculation. The sampling interval of the source data δ_T and a protocol header L_h are included in the calculation, as follows:

$$b = p \cdot \delta_T + L_h \tag{4}$$

Consequently the b value decided by Formula (4) could be smaller than M for certain flows. An example in [2, Annex C] shows that an H.263 video flow with peak rate 40 kbps ends up with a bucket size of 373 bytes.

Method IV:

Empirically we argued in [45] that the bucket size defined by equation (4) was too conservative since the bursty property of the traffic flow had not been considered. Therein we proposed to define the bucket size, denoted by $\hat{b}$ there, using the following equation:

$$\hat{b} = b \cdot \beta = b \cdot \frac{p}{a} = (p \cdot \delta_T + L_h) \cdot \frac{p}{a} \tag{5}$$

where the burstiness of a flow is defined by the ratio between the peak rate p and the average rate a, as $\beta = \frac{p}{a}$. We have also compared the performance of a traffic conditioned system using the definitions in Eqs. (4) and (5) in Paper C.

As a summary, any one of the above four described methods could be used to obtain a pre-designed (r, b) pair. But these methods are neither *standardized* calculation, nor generally true for arbitrary traffic patterns. This leads to the problem of identifying suitable TB parameters.

[5] Normally M = MTU in Internet = 1500 bytes, and M = SDU in UMTS = 1502 or 1500 octets, depending on type of connection.

3.2 Approaches to TB Parameters Determination

Generally speaking, the determination of the sensible (r, b) parameters is application-dependent or case sensitive. In addition to the above four methods, various approaches have been proposed in the literature [71, 29, 64, 52, 4, 13], from different perspectives. We have classified these approaches for determining TB parameters into two categories in [43]: deterministic or measurement based.

But still, an analytical solution to this problem could not be feasible for general traffic patterns. Therefore, a heuristic method is proposed in Paper B of this thesis. The problem is expressed in a way that the 'optimal' TB values can be obtained by exhaustive search. In particular, after the determination of a set of possible values for token bucket parameters accomplished locally at the UE, the RNC selects, in the sets provided by all UEs, the pair of (r, b) parameters that should be used to limit the packet loss ratio. We claim that this approach is QoS-aware, in a sense that the designer knows explicitly to what degree the QoS will be achieved using the obtained TB parameters.

4 Traffic Policing and Conformance Deterioration

In a traffic conditioning capable network, it must be possible to check whether a flow is indeed keeping its traffic in the way it promised when the flow was set up. This is done by traffic policing. The policer monitors the behavior of a flow at the policing node and enforces the flow to be in profile again by proper means when necessary.

4.1 Traffic Policing

According to RFC 2212 [66], there are two forms of policing. One form is simple policing in which arriving traffic is compared against a TB specification. This form is executed by traffic tagging algorithm in this thesis. The other form is reshaping, where an attempt is made to restore possibly distorted traffic's shape to conform to the TB specification. This form is executed by traffic conforming algorithm in this thesis.

The location of the traffic policer is often situated at the gateway entry of the network. Obviously, to monitor the behavior of a traffic flow, a policer should be employed at every intermediate node along the path to the other end.

In order to verify conformance consistency of a flow, it is quite straightforward to apply the same TB algorithm with identical (r, b) parameters for traffic policing, and reshape the

traffic if necessary. This is the second form of traffic policing mentioned above, and it will be further discussed in the next subsection. Traffic policing can also be jointly considered together with CAC policies, as discussed in [78].

4.2 Conformance Deterioration

The motivation of this part of my thesis work is to verify conformance consistency in a traffic conditioning-enabled multi-hop network. A system model for verifying conformance consistency between two nodes is depicted in Figure 4, where two token buckets with identical (r, b) parameters are employed as the shaper and the policer respectively. Throughout the context, the shaping node and the policing node are also referred to as ingress node and egress node respectively. Accordingly, the ingress node could be a UE and the egress node could be the RNC. With another case, where reshaping is done at the RNC, the RNC is regarded as the ingress node and the next hop to the RNC is regarded as the egress node. This does not mean any modifications to any IETF terminologies, but is only used for describing the concerned system model in this thesis, as shown in Figure 4. Furthermore, we assume that the ingress is directly connected to the egress node, which implies that other network components between them are neglected here.

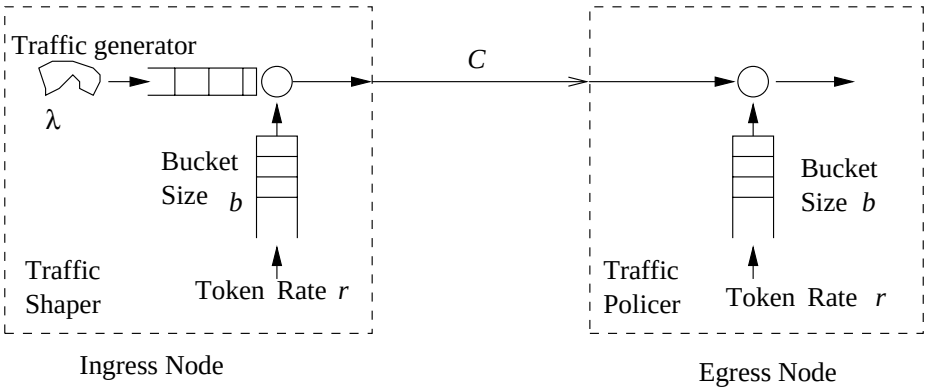


Figure 4: System model for conformance deterioration – the same token bucket with identical (r, b) parameters at both ingress and egress nodes.

As shown in the figure, the ingress node and the egress are connected using a link of constant speed of C bps. The shaping function guarantees that all packets are conformant at the departure instant from the ingress node. The policing node checks the performance of the shaped traffic flow at the egress node by monitoring whether the incoming traffic

conforms to the same TB specification.

Referring to this system model, it has been demonstrated that, in the presence of variable packet length, an originally conformant packet may become non-conformant at the egress node [32]. A terminology, *conformance deterioration*, is introduced in Paper A of this thesis for describing this non-intuitive phenomenon. The objective of this part of the thesis work is to provide better *quantitative* understanding of this phenomenon and derive proper means to minimize the effect without introducing much burden for the traffic reshaper. As a quantitative measurement, the probability of conformance deterioration, denoted by η, is introduced and defined as the probability that an originally conformant packet at the ingress becomes non-conformant at the egress node.

Briefly, the study answers the following questions:

- Is an originally conformant packet still conformant after transmission? If not,
- How much percent of the traffic flow may become non-conformant?
- Can these deteriorated packets become conformant again? By which means?

4.2.1 Static TB Status: a Simple Case

Considering a simple case, static TB status, we have deduced a closed form expression for the probability of conformance deterioration in Paper A of this thesis [47]. In this subsection, we describe a few more facets of conformance deterioration which are not covered there. To better understand the rest of this section and Section 5, the reader should probably read Paper A, or at least the summary of Paper A in Section 8 first.

Before we proceed, we have to clarify that the analytical result obtained in Paper A is obtained under the following three direct or implicit fundamental assumptions:

Assumption I The token bucket is full when the first packet arrives at both the ingress and the egress nodes;

Assumption II The two consecutive packets concerned are sent back-to-back;

Assumption III The packets are independent and identically distributed (i.i.d.), and more precisely, uniformly distributed.

4.2.1.1 Another Example of Conformance Deterioration An example of how conformance deterioration happens has been shown in Figure 2 of Paper A, where the second packet arrives *exactly* at the same instant when the first packet arrives at the egress. In fact, there are three possibilities regarding the time spacing between two consecutive packets. The second packet could arrive either before, at , or after the instant when the first packet arrives at the egress node. With the first two cases, two packets are sent back-to-back. Figure 2 of Paper A has already covered these two cases. The third case, the second packet arrives after the first one has finished its transmission, is depicted here in Figure 5.

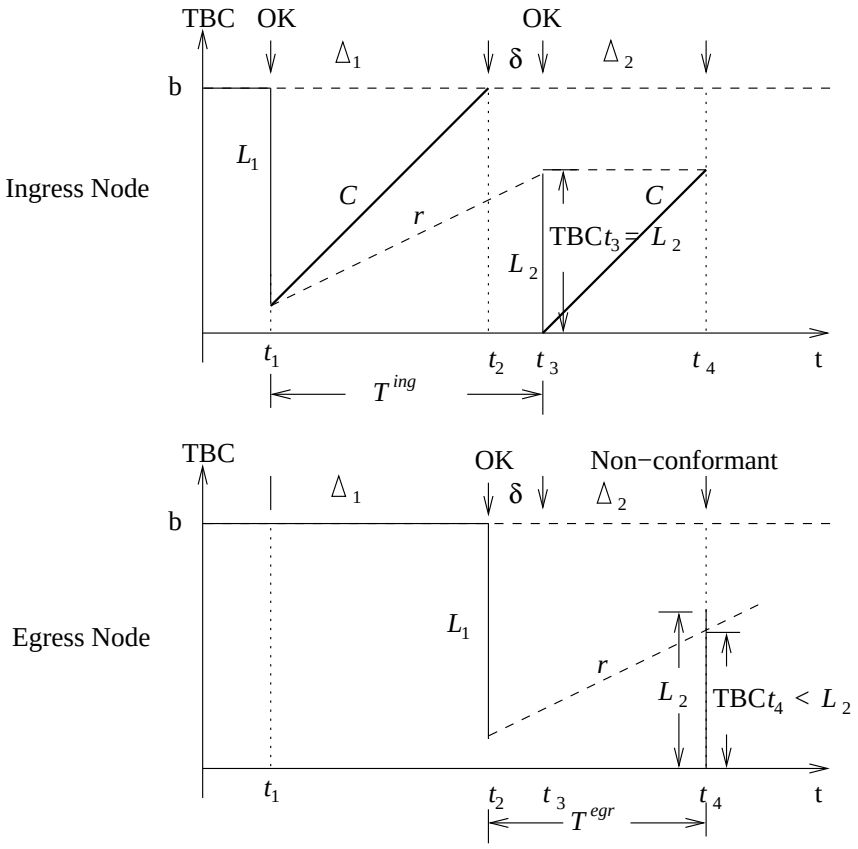


Figure 5: Another example of conformance deterioration: the packets are not sent back-to-back.

The same as in Paper A, two packets of length L_1 and L_2 bits are denoted as L_1 and L_2 respectively. Note now the packets are not sent back-to-back, but with a time spacing of δ seconds. That is, L_2 arrives at the ingress δ seconds after L_1 has reached the egress node. At the ingress, the interarrival time between L_1 and L_2 is $T^{ing} = t_3 - t_1 = \Delta_1 + \delta$. At the egress, the interarrival time between L_1 and L_2 becomes $T^{egr} = t_4 - t_2 = \delta + \Delta_2$. Here $\Delta_1 = L_1/C$ and $\Delta_2 = L_2/C$ represent transmission time for L_1 and L_2 respectively. The

difference between T^{ing} and T^{egr} makes it possible that L_2 becomes non-conformant at the egress, as the example shown in the figure.

The careful reader may notice that the size of the second packet in this example, even with the same notation, is larger [6] than the size of L_2 in Figure 2 of Paper A. It is therefore less likely that L_2 becomes deteriorated in this case, as δ increases. In fact, the interarrival time of the flow has impact on the probability of conformance deterioration. For further details, read Section 5.

4.2.1.2 Comments on Accuracy of the Analytical Result We have illustrated in Paper A that the simulation results *roughly* meet the analytical result. This roughness is indeed caused by the three fundamental assumptions described above. In addition to the status of the token bucket, interarrival time between two consecutive packets and the distribution of the packet length are two other factors affecting the accuracy of the analytical result. A separate section, Section 5, is devoted to further analyze the impact of the packet length and interarrival time distributions to the probability of conformance deterioration.

4.2.1.3 Deterioration Elimination with Larger Bucket As a measure to eliminate conformance deterioration, we have shown in Paper A that, by doubling the policing TB bucket size, the intensity of deterioration has been dramatically reduced. We illustrate here in Figure 6 that, by further enlarging the policing bucket size to $b = 4$ MTUs, the probability of conformance deterioration η has reached a quite low value of $\eta = 10^{-3}$. In fact, identifying the size of the reshaping TB needed to reduce the egress non-conformance to an acceptable level is also an interesting aspect of this study. It has been shown in [32] that a relatively small bucket size ($b = 4 \sim 9$ MTUs) is sufficient to ensure a conformance deterioration probability of less than 10^{-5}. However, since large bucket size leads to very bursty traffic, we still recommend not to facilitate too large bucket size at the policing node in a QoS ensured network.

4.2.2 Stochastic TB Status: a More Complex Case

The first fundamental assumption on static TB status is obviously also too idealistic since in practice it is impossible to control the TBC value at any given instant. The TB status is thus a stochastic variable due to the fact that the number of available tokens at any instant

[6] More accurately, by $r\delta$ bits.

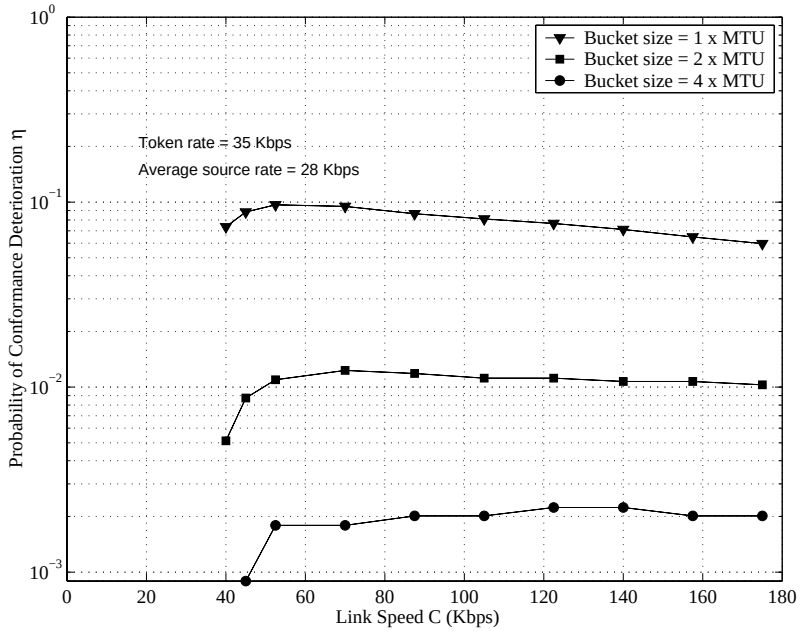


Figure 6: Probability of conformance deterioration with different bucket sizes.

is affected by the packet arrival process which is stochastic. In Appendix, we formulate the problem of conformance deterioration in a more realistic model and present a stochastic approach to this more sophisticated case.

5 Analyses on Conformance Deterioration: The Static Case

The discussions in this section are probably more detailed than just an introduction, compared with other sections in Part I. This is because this part of the thesis work has not yet been formed as a paper for further publication.

To continue our analysis on conformance deterioration, we first remove the 3rd fundamental assumption on Page 20 and the corresponding results are described in Subsection 5.1. Then both the 2nd and the 3rd assumptions are lifted, and the results are presented in Subsection 5.2. But in this section the status of the TB is still assumed to be static and the bucket is full when the first packet arrives at both nodes. The assumption is of course naive, but it allows us to proceed the analyses below. An approach where all three assumptions are lifted is later presented in the appendix.

Furthermore, in order to verify the analytical results in this section, it is possible to emulate the first fundamental assumption by observing *only the first two packets of the*

whole flow in simulation. By doing so, it is guaranteed that the bucket is full at both nodes when the first packet arrives. If we repeat this procedure to a statistically large enough number, we can compare the analytical results on the probability of conformance deterioration shown below with the simulation results, for given packet length and interarrival time distributions.

5.1 Arbitrary Packet Length Distribution

Using the same notation as in Paper A, we denote C for link speed, r for token rate, b for bucket size and $\alpha = r/C$ for the ratio between r and C. Packet length L_1 and L_2 are hereby denoted by two random variables U and V respectively. U and V are arbitrarily distributed within $(0, b]$ with Probability density functions (PDFs) $f_u(x)$ and $f_v(x)$. The joint PDF of U and V is $f_{uv}(x)$. Generally we have the distributions of U and V as $F_u(x) = P(U < x) = \int_0^x f_u(u)\,\mathrm{d}u$ and $F_v(x) = P(V < x) = \int_0^x f_v(v)\,\mathrm{d}v$.

5.1.1 Probability of Conformance Deterioration for Arbitrarily Distributed Packets

Following the same reasoning on conformance deterioration in Section III of Paper A, the probability of conformance deterioration η can be re-written as

$$\begin{aligned} \eta &= \frac{P(\frac{b-L_1}{1-\alpha} < L_2 \leq b + (\alpha-1)L_1)}{P(L_2 \leq b + (\alpha-1)L_1)} \\ &= \frac{P[(L_2 + \frac{L_1}{1-\alpha} > \frac{b}{1-\alpha}) \cap (L_2 + (1-\alpha)L_1 \leq b)]}{P(L_2 + (1-\alpha)L_1 \leq b)} \end{aligned} \tag{6}$$

Substituting L_1 and L_2 by U and V, we can simply obtain the probability of conformance deterioration for arbitrary packet length distribution η^{arb} as

$$\begin{aligned} \eta^{arb} &= \frac{P[(V + \frac{U}{1-\alpha} > \frac{b}{1-\alpha}) \cap (V + (1-\alpha)U \leq b)]}{P(V + (1-\alpha)U \leq b)} \\ &= \frac{\iint\limits_{(V+\frac{U}{1-\alpha} > \frac{b}{1-\alpha}) \cap (V+(1-\alpha)U \leq b)} f_{uv}(u,v)\mathrm{d}u\mathrm{d}v}{\iint\limits_{V+(1-\alpha)U \leq b} f_{uv}(u,v)\mathrm{d}u\mathrm{d}v} \end{aligned} \tag{7}$$

Now let us assume again that U and V are i.i.d. random variables. Hence we have $f_{uv}(u,v) = f_u(x)f_v(x)$ and $f_u(x) = f_v(x) \equiv f(x)$. The probability of conformance

deterioration for i.i.d. packet length distribution η^{iid} can be obtained from Equation (7) as

$$\eta^{iid} = \frac{\int_{\frac{b}{2-\alpha}}^{b} f(u)[\int_{\frac{b-u}{1-\alpha}}^{b-(1-\alpha)u} f(v)\mathrm{d}v]\mathrm{d}u}{\int_{0}^{b} f(u)[\int_{0}^{b-(1-\alpha)u} f(v)\mathrm{d}v]\mathrm{d}u} \tag{8}$$

Note the upper and lower limits of the integrations in Equation (8) are obtained from the conditions of the integrations, $V + \frac{U}{1-\alpha} \geq \frac{b}{1-\alpha} \cap V + (1-\alpha)U \leq b$ and $V + (1-\alpha)U \leq b$.

Equations (7) and (8) give the analytical results on the probability of conformance deterioration for arbitrarily distributed and i.i.d. packets, respectively. In the following subsubsections, three examples are given for further illustrating this result.

5.1.2 Example I: Uniformly Distributed i.i.d. Packets

Assuming uniformly distributed i.i.d packet sizes within $(0, b]$, we have PDF

$$f^{uni}(x) = \begin{cases} \frac{1}{b} & \text{for} \quad 0 \leq x \leq b \\ 0 & \text{for} \quad x < 0, x > b \end{cases} \tag{9}$$

Inserting $f^{uni}(x)$ into Equation (8), we can easily get

$$\begin{aligned} \eta^{uni} &= \frac{\int_{\frac{b}{2-\alpha}}^{b} \frac{1}{b}[\int_{\frac{b-u}{1-\alpha}}^{b-(1-\alpha)u} \frac{1}{b}\mathrm{d}v]\mathrm{d}u}{\int_{0}^{b} \frac{1}{b}[\int_{0}^{b-(1-\alpha)u} \frac{1}{b}\mathrm{d}v]\mathrm{d}u} = \frac{\int_{\frac{b}{2-\alpha}}^{b} \frac{1}{b}\frac{1}{b}[b-(1-\alpha)u-\frac{b-u}{1-\alpha}]\mathrm{d}u}{\int_{0}^{b} \frac{1}{b}\frac{1}{b}[b-(1-\alpha)u]\mathrm{d}u} \\ &= \frac{\frac{\alpha(1-\alpha)}{2(2-\alpha)}}{\frac{1+\alpha}{2}} = \frac{\alpha-\alpha^2}{2+\alpha-\alpha^2} \end{aligned} \tag{10}$$

This result is exactly the same as our conclusion in Equation (8) of Paper A.

5.1.3 Example II: Exponentially Distributed i.i.d. Packets

Now we have $f_u(x) = f_v(x) = f^{exp}(x)$ as

$$f^{exp}(x) = \lambda e^{-\lambda x}, \quad \lambda > 0, x \geq 0 \tag{11}$$

The numerator of Equation (8) becomes

$$\begin{aligned}\int_{\frac{b}{2-\alpha}}^{b} \lambda e^{-\lambda x}[\int_{\frac{b-u}{1-\alpha}}^{b-(1-\alpha)u} \lambda e^{-\lambda x}\mathrm{d}v]\mathrm{d}u &= \int_{\frac{b}{2-\alpha}}^{b} \lambda e^{-\lambda x}[-e^{-\lambda b}e^{\lambda(1-\alpha)u} + e^{-\frac{\lambda b}{1-\alpha}}e^{\frac{\lambda u}{1-\alpha}}]\mathrm{d}u \\ &= \frac{1}{\alpha}e^{-(1+\alpha)\lambda b} - \frac{2-\alpha}{\alpha}e^{-\frac{2\lambda b}{2-\alpha}} + \frac{1-\alpha}{\alpha}e^{-\lambda b} \quad (12)\end{aligned}$$

The denominator of Equation (8) becomes

$$\begin{aligned}\int_{0}^{b} \lambda e^{-\lambda x}[\int_{0}^{b-(1-\alpha)u} \lambda e^{-\lambda x}\mathrm{d}v]\mathrm{d}u &= \int_{0}^{b} \lambda e^{-\lambda u}[-e^{-\lambda b}e^{(1-\alpha)\lambda u} + 1]\mathrm{d}u \\ &= \frac{1}{\alpha}e^{-(1+\alpha)\lambda b} - \frac{1+\alpha}{\alpha}e^{-\lambda b} + 1 \quad (13)\end{aligned}$$

Thus by putting them together, we obtain the probability of conformance deterioration for exponentially i.i.d. packets η^{exp} as

$$\eta^{exp} = \frac{\frac{1}{\alpha}e^{-(1+\alpha)\lambda b} - \frac{2-\alpha}{\alpha}e^{-\frac{2\lambda b}{2-\alpha}} + \frac{1-\alpha}{\alpha}e^{-\lambda b}}{\frac{1}{\alpha}e^{-(1+\alpha)\lambda b} - \frac{1+\alpha}{\alpha}e^{-\lambda b} + 1} \quad (14)$$

Different from our result with uniform i.i.d. packets, the probability of conformance deterioration for exponential i.i.d. packets is not only a function of α, but also related to the ration between bucket size and average packet length, λb. Figure 7 illustrates η^{exp} in three dimensions as a function of α and λb, with α ranging between (0,1] and λb ranging between (0,5].

Looking at the convex curve from x-axis, it is shown theoretically that the peak value for η^{exp} could reach as high as about 10% and it is achieved when α is around 0.5 and λb close to 0. This is quite similar to the result with uniformly i.i.d. packet in Paper A. The exact peak value of η^{exp} with corresponding α and λb values can be obtained by solving the joint partial differentiation equations $\frac{\partial \eta^{exp}}{\partial \alpha} = 0$ and $\frac{\partial \eta^{exp}}{\partial \lambda b} = 0$, and inserting the results into Equation (14). But in reality, we have $\lambda b > 1$ because all generated packets are not larger than b. The maximum value for η^{exp} is therefore about 6.8% at $\lambda b \simeq 1$.

Observing Figure 7 from y-axis, it is shown that the larger the λb is, the smaller the η^{exp} becomes. Figure 3 in Paper A and the lower integral limits in the numerator of Equation (8) show that the first packet in a back-to-back packet pair must be larger than certain value for conformance deterioration. The traffic is hence more seriously deteriorated when there are more large packets in a flow.

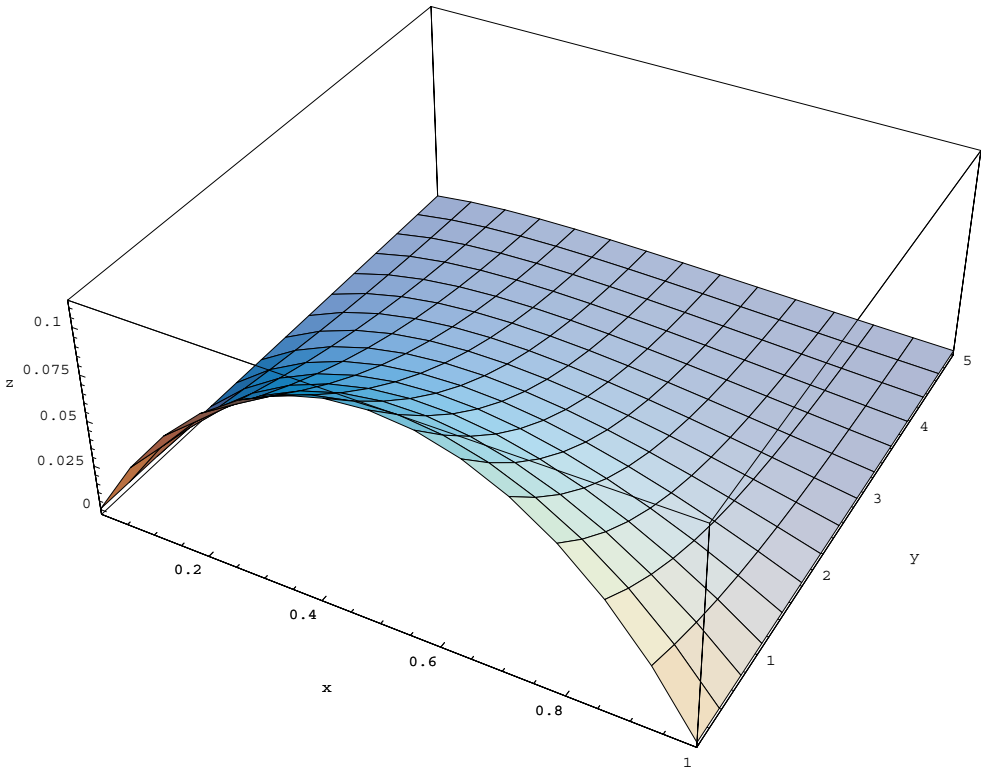


Figure 7: Probability of conformance deterioration for exponentially i.i.d. packets η^{exp} as a function of α and λb, where x-axis = α, y-axis = λb and z-axis = η^{exp}.

5.1.4 Example III: Beta Distributed i.i.d Packets

Beta distribution is used to describe continous random variable whose values have a lower bound a and an upper bound b. The PDF of the beta distribution is

$$f^{beta}(x) = \begin{cases} \frac{(x-a)^{p-1}(b-x)^{q-1}}{(b-a)^{p+q-1}B(p,q)} & \text{for} \quad a \leq x \leq b \\ 0 & \text{for} \quad x < a, x > b, p > 0, q > 0, b > a \end{cases} \tag{15}$$

where the beta function is defined as $B(p,q) = \int_0^1 x^{p-1}(1-x)^{q-1}\mathrm{d}x$.

In our case where the packet size is restricted to $(0, b]$, we have $a = 0$. Inserting Equation (15) into Equation (8) gives the probability of conformance deterioration for beta distributed i.i.d. packets

$$\eta^{beta} = \frac{\int_{\frac{b}{2-\alpha}}^{b} \frac{x^{p-1}(b-x)^{q-1}}{b^{p+q-1}B(p,q)} [\int_{\frac{b-u}{1-\alpha}}^{b-(1-\alpha)u} \frac{x^{p-1}(b-x)^{q-1}}{b^{p+q-1}B(p,q)} \mathrm{d}v] \mathrm{d}u}{\int_0^b \frac{x^{p-1}(b-x)^{q-1}}{b^{p+q-1}B(p,q)} [\int_0^{b-(1-\alpha)u} \frac{x^{p-1}(b-x)^{q-1}}{b^{p+q-1}B(p,q)} \mathrm{d}v] \mathrm{d}u} \tag{16}$$

The above integral cannot generally be solved in closed form for arbitrary p and q. But for some given p and q values, it is possible to express η^{beta} explicitly. For the simplest case where $p = 1, q = 1$, we can easily get

$$\eta^{beta}_{(1,1)} = \frac{\int_{\frac{b}{2-\alpha}}^{b} \frac{x^0(b-x)^0}{bB(1,1)}[\int_{\frac{b-u}{1-\alpha}}^{b-(1-\alpha)u} \frac{x^0(b-x)^0}{bB(1,1)}\mathrm{d}v]\mathrm{d}u}{\int_{0}^{b} \frac{x^0(b-x)^0}{bB(1,1)}[\int_{0}^{b-(1-\alpha)u} \frac{x^0(b-x)^0}{bB(1,1)}\mathrm{d}v]\mathrm{d}u} = \frac{\alpha - \alpha^2}{2 + \alpha - \alpha^2} \tag{17}$$

This is again exactly the same result as in Equation (10), because the beta distribution becomes a uniformly continuous distribution when $p = 1, q = 1$.

Any other expressions for $\eta^{beta}_{(p,q)}$ are pretty lengthy. Table 3 lists calculated η^{beta} values from Equation (16) for three simple examples, ($p = 1, q = 1$), ($p = 3, q = 2$) and ($p = 2, q = 3$). Note here α cannot be 1 for $\eta^{beta}_{(3,2)}$ and $\eta^{beta}_{(2,3)}$ because $(1 - \alpha)$ appears as a denominator in Equation (16).

From the property of the beta distribution, we know that ($p = 3, q = 2$) represents much higher probability of large packets compared with the case of ($p = 2, q = 3$). This leads to a much higher η^{beta} for ($p = 3, q = 2$). A similar trend has been observed with exponential distribution, where smaller λb leads to higher η^{exp}.

Table 3: Comparison of η^{beta} for three simple cases.

$\alpha = r/C$	0	0.2	0.4	0.5	0.6	0.8	0.99	1
$\eta^{beta}_{(1,1)}$ (%)	0	7.40	10.71	11.11	10.71	7.41		0
$\eta^{beta}_{(3,2)}$ (%)	0	10.18	15.18	15.19	13.54	6.17	0.02	
$\eta^{beta}_{(2,3)}$ (%)	0	2.51	2.73	2.22	1.53	0.30	0	

5.2 Arbitrary Interarrival Time Distribution

Recall that the arrival of a packet means the arrival of the last bit of the packet. The interarrival time here represents the time difference between the arrival instant of the last bit from two consecutive packets. As depicted in Figure 5, the interarrival time between L_1 and L_2 at the ingress TB shaper is $T^{ing} = \Delta_1 + \delta$ where Δ_1 is the transmission time for Packet L_1. With $\delta \leq 0$, packets are sent back-to-back and these two cases have been analyzed in the above subsection. We are now interested in the third case, $\delta > 0$, where two packets are not sent back-to-back, but with a gap $\delta > 0$ between them.

Refer to Figure 5 for the following descriptions and still follow the reasoning on conformance deterioration in Section III of Paper A. Now packet L_2 is conformant at the ingress if

$$L_2 \leq b - L_1 + (\frac{L_1}{C} + \delta)r \tag{18}$$

i.e.,

$$L_2 + (1-\alpha)L_1 - \delta r \leq b \tag{19}$$

It becomes non-conformant if

$$b - L_1 + (\Delta_2 + \delta)r = b - L_1 + (\frac{L_2}{C} + \delta)r < L_2 \tag{20}$$

i.e.,

$$L_2 + \frac{L_1}{1-\alpha} - \frac{\delta r}{1-\alpha} > \frac{b}{1-\alpha} \tag{21}$$

Denote δ by a random variable W with PDF $f_w(x)$, and substitute L_1 and L_2 by U and V. The probability of conformance deterioration for arbitrary interarrival time η^{int} is therefore decided by the following equation, where $f_{uvw}(x)$ is the joint PDF of U, V and W.

$$\begin{aligned} \eta^{int} &= \frac{P[(V + \frac{U}{1-\alpha} - \frac{Wr}{1-\alpha} > \frac{b}{1-\alpha}) \cap (V + (1-\alpha)U - Wr \leq b)]}{P(V + (1-\alpha)U - Wr \leq b)} \\ &= \frac{\iiint_{(V+\frac{U}{1-\alpha}-\frac{Wr}{1-\alpha}>\frac{b}{1-\alpha})\cap(V+(1-\alpha)U-Wr\leq b)} f_{uvw}(u,v,w)\mathrm{d}u\mathrm{d}v\mathrm{d}w}{\iiint_{V+(1-\alpha)U-Wr\leq b} f_{uvw}(u,v,w)\mathrm{d}u\mathrm{d}v\mathrm{d}w} \end{aligned} \tag{22}$$

Note now even though interarrival time T^{ing} is independent of packet length U and V, the third concerned random variable W is dependent of U, because of $W = \mathcal{T}^{ing} - \Delta_1 = \mathcal{T}^{ing} - \frac{U}{C}$, where $\mathcal{T}^{ing}$ is the random variable representing T^{ing}. This means that Equation (22) can not be further simplified because of $f_{uvw}(u,v,w) \neq f_{uv}(u,v)f_w(w)$.

Ultimately, by combining Equations (8) and (22), we conclude that the probability of conformance deterioration for static TB status, η^{static}, given interarrival time and packet

length are arbitrarily distributed, can be expressed as

$$\eta^{static} = \begin{cases} \dfrac{\iint\limits_{(V+\frac{U}{1-\alpha}>\frac{b}{1-\alpha})\cap(V+(1-\alpha)U\leq b)} f_{uv}(u,v)\mathrm{d}u\mathrm{d}v}{\iint\limits_{V+(1-\alpha)U\leq b} f_{uv}(u,v)\mathrm{d}u\mathrm{d}v} & \text{for} \quad \delta \leq 0 \\ \dfrac{\iiint\limits_{(V+\frac{U}{1-\alpha}-\frac{Wr}{1-\alpha}>\frac{b}{1-\alpha})\cap(V+(1-\alpha)U-Wr\leq b)} f_{uvw}(u,v,w)\mathrm{d}u\mathrm{d}v\mathrm{d}w}{\iiint\limits_{V+(1-\alpha)U-Wr\leq b} f_{uvw}(u,v,w)\mathrm{d}u\mathrm{d}v\mathrm{d}w} & \text{for} \quad \delta > 0 \end{cases} \tag{23}$$

6 Radio Resource Management

Radio Resource Management (RRM) is of paramount importance for provisioning of 3G services, since the wireless bandwidth is the bottleneck of the UMTS network. The RRM module is responsible for distributing the available resources among all users sharing the same resource. The available resources are distributed according to the required QoS.

Functionalities for RRM include handover control, power control, CAC, load control and bandwidth allocation, packet scheduling and code management [3, 35, 40]. Typically, power control and handover control commands are implemented at the UE, and Node B is responsible for code generation in addition to power control. All other functions are performed at the RNC, based on information gathered from the UEs and the Node B. Figure 8 illustrates the locations of the RRM functions in UTRAN [35] [40].

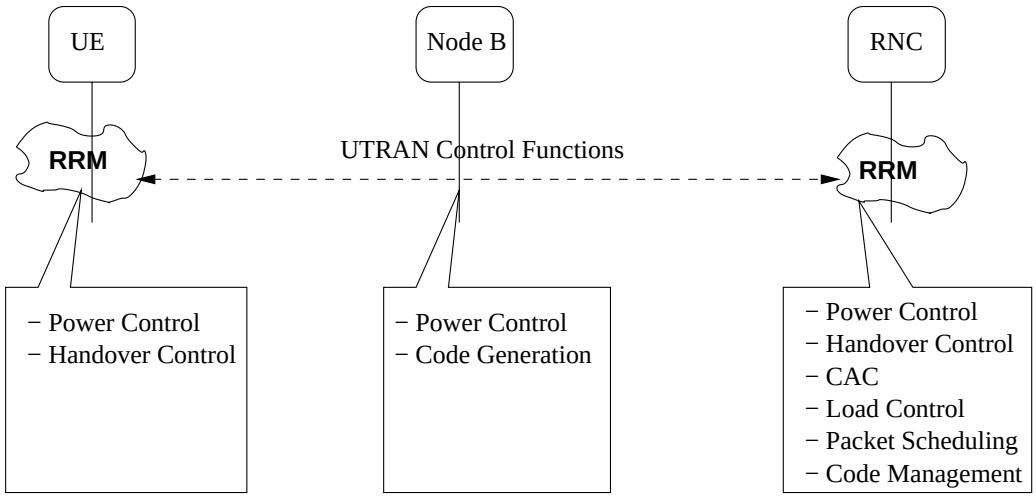


Figure 8: Typical locations of RRM algorithms in UTRAN.

Usually, connection requests are initiated by UEs, either through a new call or a handover call. The negotiation of a service contract is conducted between the RNC and the UE,

based on available resources within the system. Various functionalities are co-ordinated via UTRAN control functions.

This part of my thesis work is restricted to the UTRAN part of UMTS which is Wide-band Code Division Multiple Access (WCDMA)-based. The network resource considered is actually bandwidth on wireless channel. Among various RRM functionalities mentioned above, only CAC and load control/bandwidth allocation are addressed in this thesis. More detailed discussions on power control and handover control can be found in [35, Chapter 9] and [40, Chapter 4], and on packet scheduling in [22, 49]. For general descriptions on resource management in UMTS, see [24] and [3].

6.1 Call Admission Control

As an essential function of traffic control, CAC has been intensively investigated from both data communication and telecommunication communities. The objective of CAC schemes is to accommodate, in an optimal way, a maximum number of connection requests, and at the same time, maintain the agreed QoS for existing connections. This is done by managing the available network resources and allocating them according to a particular strategy, among the users [23].

Conventional CAC policies in wireline networks are complete sharing, complete partitioning and threshold. In the complete sharing policy, calls of all classes share the bandwidth resource. Whereas, in the complete partitioning policy, bandwidth for each class is exclusively reserved. In the threshold policy, a newly arriving call is blocked if the number of calls of each class is greater than a predefined threshold [17]. A survey and comparison of CAC schemes in ATM networks can be found in [58]. In an IP network, CAC could also be jointly considered with policing, shaping and resource allocation [78].

For a CDMA-based system like UMTS, the CAC strategies can be roughly classified into two approaches [37]:

- Number-based CAC: a new call is rejected if the number of ongoing connections has already reached system limit.
- Interference-based CAC: a new call is rejected if the observed interference level exceeds a predefined threshold.

The interference-based CAC uses Signal-to-Interference Ratio (SIR) to ensure that the interference created after adding a new call does not exceed a pre-specified threshold [73,

51], and it is more often adopted in the literature. Usually, the interference from other existing connections within the same cell and the neighboring cells are considered in SIR calculation. One can also consider CAC in a system comprising of hierarchical cells [59].

In order to verify the performance of various CAC strategies, blocking probability for new calls and outage probability of existing calls are commonly used as the criteria for decision making. Since dropping a call is more annoying for a user than being blocked as a new call, Grade of Service (GoS), which is equal to the sum of the blocking probability plus ten times of the dropping probability [74, Chapter 9], is utilized for performance evaluation. More recent work defined a more complex GoS by weighting the blocking probability, handoff failure and QoS loss in the summation [60]. The ultimate result of CAC, either calculated or measurement-based [31], is usually a hard-decision.

In a practical system, the CAC function involves both uplink and downlink interference calculation and decision making. A new call should only be admitted if it passes both downlink and uplink admission algorithm [61, Chapter 7]. While most CAC schemes appeared in the literature are focusing on uplink, research results on downlink CAC schemes are relatively fewer [77, 55]. Like many other studies, the CAC scheme proposed in this thesis also considers only uplink traffic.

As UMTS services are characteristic of high bandwidth, multiple QoS requirements and asymmetric traffic, we have to face a heterogeneous traffic environment. The delay tolerance could also be considered in a CAC scheme. Basically, delay tolerable traffic class(es) can be treated differently from delay stringent traffic class(es). Lots of approaches have been proposed regarding CAC for multimedia/heterogeneous traffic, see, for example, [16, 65, 41, 18, 50, 33].

Based on traffic class definition by 3GPP in [3], we have proposed a so-called priority-oriented CAC paradigm in [44]. In practice, the priority definition could be more flexible as described in [78]. In our approach, we hypothesize that the certain traffic classes have multiple rate connections and the lower priority users are willing to release their bandwidth occupation to more important users whenever necessary. Under this assumption, we have achieved quite low overall blocking probability for new calls. But overload may happen if we don't have resource management after CAC phase. This triggered our work in Paper D [48], where a general framework combining CAC, load control and bandwidth allocation is studied.

6.2 Load Control and Bandwidth Allocation

As an important functionality of RRM, load control, or congestion control, is a measure to ensure that the system is stable and not overloaded. At the same time, bandwidth allocation is a measure to maximize the usage of the air interface resources. With these measures, the system is optimally utilized, thus guaranteeing QoS from various service requirements.

As a well-planned system with CAC and packet scheduling, overload situation is exceptional. However, once congestion happens, load control and bandwidth allocation should be able to react efficiently until the targeted load is reached. Means for load control could include, but not exhaustive [35, 39], the following:

- Bandwidth reservation, part of the resources are reserved for certain connections;
- Reduce the E_b/N_o targets for certain users;
- Handover traffic to other cells or systems;
- Decrease bitrates of certain users;
- Drop calls in a controlled fashion.

Among various means, bandwidth reservation can guarantee the on-demand occupation of the resources for some 'important' users. But, it can overestimate the amount of reserved bandwidth thus lowering bandwidth utilization. Reducing the E_b/N_o targets of 'less important' users can release certain amount of bandwidth, but it is generally annoying for those 'unlucky' users. Handover traffic to other cells or systems could be another solution, but it depends on the network architecture of an operator.

In included Paper D, we employed the last two means on the above list for our load control and bandwidth re/allocation scheme. There are four traffic applications in consideration, i.e., voice, video, World Wide Web (WWW) and email. Basically, we assume that video and WWW traffic have multiple bitrates and are delay tolerable. When the channel is overloaded, the video/WWW users are requested to reduce their bitrates but still connected. In the worst cases, the connection is interrupted for a time period, but it may not be noticeable to the end users. If the waiting time in the buffer is too long, the call is dropped. One distinction of our work from other literature is that we have also observed the amount of time when a user is satisfied with the service he/she receives.

7 Future Work

The following issues are also quite interesting, but left for future work.

- **QoS parameters mapping**

  QoS attribute mapping is also an important issue in UMTS [3, Clause 8]. Basically, this problem is twofold. On the one hand, the QoS parameters defined within UMTS should be mapped vertically among different levels. For example, how transfer delay for radio access bearer service is meant to UMTS bearer service (See Figure 1). On the other hand, the QoS parameters defined in UMTS domain should be translated horizontally to CN domain. For example, maximum bitrate in UMTS could easily be mapped to peak data rate in RSVP, but how about SDU loss ratio?

  According to 3GPP discussions, part of the QoS mapping problems is an operator choice or implementation issue, but the remaining issues are far from being well studied. A framework on end-to-end QoS mapping was presented in [36]. Paper [21] studied how cell loss ratio and delay can be mapped between two protocol layers vertically. Another paper studied how packet loss ratio and delay can be interpreted horizontally between IntServ and DiffServ services [14]. A more recent paper discussed the mapping between UMTS traffic classes and DiffServ AF/EF PHB classes [53]. However, more research effort is needed on this topic.

- **Network calculus issues**

  One powerful solution to the problem of QoS mapping between IntServ and DiffServ is a tool known as *network calculus* [14]. But the beauty of the network calculus is far more than dealing with QoS mapping. Pioneered by the work in [19, 20], the theory of network calculus has been extensively studied in the literature [10, 15].

  Network calculus is a theory of deterministic queueing system, which provides a set of rules and models of network entities for determining performance of the packet data networks, for instance end-to-end delay. The mathematical foundation of network calculus is min plus algebra. With the concept of arrival curve and service curve, the deterministic bounds on delay and backlog in a lossless network can be achieved, thus guaranteed QoS is provided.

  The theory of network calculus can be applied to for instance the guaranteed service of ATM Variable Bit Rate (VBR), the guaranteed service networks [8, 72], and the

hybrid IntServ and DiffServ framework [7]. However, very little literature can be found on how network calculus applies to UMTS networks.

8 Contributions and Summary

The remainder of this thesis is composed of four papers already submitted to or published in journal or international conferences. As each publication is written as a self-contained paper, all relating to my thesis topic, redundancy among papers is unavoidable. However, the overlapping has been minimized by the selection process itself.

The main contributions of this dissertation, in the order of the included papers appearing in Part II of the thesis, are outlined as follows:

☞ **A Study on Traffic Shaping, Policing and Conformance Deterioration for QoS Contracted Networks (Paper A)**

The objective of this paper is to better understand conformance consistency in a traffic conditioned multi-hop network. After reiterating a recent finding that an originally conformant packet may become non-conformant even transmitted at a constant bitrate [32], the probability of this non-intuitive phenomenon, which is referred to as *conformance deterioration*, is analyzed for a simple case. We demonstrate that up to 11% of the originally conformant packets may be deteriorated in the worst case, which occurs exactly when the link speed is twice as high as the token rate. In order to eliminate conformance deterioration, two measures, enlarging bucket size or reshaping at the policing node, are investigated by simulation. Conformance deterioration for aggregated traffic flows is also studied and compared with the results from an individual traffic flow.

The major contribution of this paper is the quantitative result on probability of conformance deterioration. Various means to minimize conformance deterioration and the performance of the aggregated flows are also studied.

☞ **Local and Global QoS-aware Token Bucket Parameters Determination for Traffic Conditioning in 3rd Generation Wireless Networks (Paper B)**

Due to the difficulty of determining TB parameters analytically, a heuristic approach for searching 'optimal' (r, b) values is proposed in this paper. We refer to this method as local and global QoS-aware token bucket parameter determination technique for

traffic conditioning in 3G wireless networks, in a sense that the system is aware of the QoS level to be achieved. The local QoS-awareness is achieved by bounding either shaping delay or out-of-profile probability at the UE, and the global QoS-awareness is obtained by the tradeoff between a local attribute and the global attribute, packet loss ratio for conformed packets, at the RNS level. By tuning the system operating on the obtained global 'optimal' shaping parameters, the requirements for all concerned QoS attributes are guaranteed.

The major contribution of this paper is the heuristic approach for TB parameters determination for designing a QoS-aware traffic conditioned system.

☞ **QoS Provisioning using Traffic Shaping and Policing in 3rd-Generation Wireless Networks (Paper C)**

A framework of traffic conditioning with QoS provisioning in 3G radio access network is proposed in this paper. The main idea of our traffic conditioning approach is to employ traffic shaping at each UE and traffic policing at the RNC. The traffic generated at each UE is regulated by a traffic shaper in the form of a token bucket, and the conformance of the traffic is policed at the RNC according to traffic policing policies. A system model based on the proposed framework is implemented. The simulation results regarding the impact of traffic shaping on packet discarding probability, the tradeoff between probability of non-compliance and shaping delay are presented.

The major contribution of this paper is a framework of having all network nodes traffic conditioned. Our conclusion is that it is advantageous to implement traffic shaping at the UE, as a required functionality.

☞ **A Priority-oriented QoS Management Framework for Multimedia Services in UMTS (Paper D)**

Triggered by 3GPP traffic class and QoS ranking definition in [3], we propose a priority-oriented QoS management framework for multimedia services provision in this paper. There are three basic components in our framework, i.e., the priority-oriented CAC, CDMA channel congestion control, and adaptive bandwidth allocation. Our approach ensures uninterrupted service provision to 'more important' traffic class(es) while trying to accommodates negotiated QoS parameters to other classes. To investigate the performance of the proposed framework, we also present

our observations on such parameters as forced delay, dropping probability, time satisfactoriness and transmission status in this paper, based on our system level simulation. The joint consideration of our framework with traffic control mechanism envisages a paradigm for QoS provisioning in UMTS.

The major contribution of this paper is a priority-oriented QoS management framework. The performance of the proposed system is manipulated through CAC, congestion control and bandwidth (re)allocation.

Bibliography

[1] 3GPP, "UTRA(BS) FDD; Radio Transmission and Receiption," *TS 25.104v3.3.0, http://www.3gpp.org*, June 2000.

[2] 3GPP, "End-to-End QoS Concept and Architecture," *TS 23.207v5.4.0, http://www.3gpp.org*, June 2002.

[3] 3GPP, "QoS Concept and Architecture," *TS 23.107v5.5.0, http://www.3gpp.org*, June 2002.

[4] M. F. Alam, M. Atiquzzaman, and M. A. Karin, "Traffic Shaping for MPEG Video Transmission over the Next Generation Internet," *Computer Communications*, pp. 1336–1348, Vol. 23 2000.

[5] Y. Bernet, P. Ford, R. Yavatkar, F. Baker, L. Zhang, M. Speer, R. Braden, B. Davie, J. Wroclawski, and E. Felstaine, "A Framework for Integrated Services Operation over Diffserv Networks," *RFC 2998, IETF*, Nov. 2000.

[6] S. Blake, D. Black, M. Carlson, E. Davies, Z. Wang, and W. Weiss, "An Architecture for Differentiated Serives," *RFC 2475, IETF*, Dec. 1998.

[7] N. Blefari-Melazzi, D. D. Sorte, M. Femminella, and G. Reali, "Resource Allocation Rules to Provide QoS Guarantees to Traffic Aggregates in a DiffServ Environment," in *Proc. IST Mobile Communications Summit*, (Barcelona, Spain), Sept. 2001.

[8] J. Y. Le Boudec, "Application of Network Calculus to Guaranteed Service Networks," *IEEE Transactions on Information Theory*, vol. 44, pp. 1087–1096, May 1998.

[9] J. Y. Le Boudec, "Some Properties of Variable Length Packet Shapers," *IEEE/ACM Transactions on Networking*, vol. 10, pp. 329–337, June 2002.

[10] J. Y. Le Boudec and P. Thiran, *Network Calculus*. Springer Verlag LNCS 2050, 2001.

[11] R. Braden, D. Clark, and S. Shenker, "Integrated Services in the Internet Architecture: an Overview," *RFC 1633, IETF*, June 1994.

[12] R. Braden, L. Zhang, S. Berson, S. Herzog, and S. Jamin, "Resource ReSerVation Protocol (RSVP) – Version 1 Functional Specification," *RFC 2205, IETF*, Sept. 1997.

[13] R. Bruno, R. G. Garroppo, and S. Giordano, "Estimation of Token Bucket Parameters of VoIP Traffic," in *Proc. IEEE Conference on High Performance Switching and Routing*, (Heidelberg, Germany), pp. 353–356, May 2000.

[14] T. Chahed, G. Hebuterne, and C. Fayet, "On Mapping of QoS Between Intergrated Services and Differentiated Services," in *Proc. IEEE/IFIP International Workshop on Quality of Service*, (Pittsburgh, USA), pp. 173–175, May 2000.

[15] C. S. Chang, *Performance Guarantees in Communication Networks*. Springer Verlag, 2000.

[16] C. Chao and W. Chen, "Connection Admission Control for Mobile Multiple-Class Personal Communications Networks," *IEEE Journal on Selected Areas in Communications*, vol. 15, pp. 1618–1626, Oct. 1997.

[17] J. Choi, T. Kwon, Y. Choi, and M. Naghshineh, "Call Admission Control for Multimedia Services in Mobile Cellular Networks: A Markov Decision Approach," in *Proc. IEEE International Symposium on Computers and Communications*, (New York, USA), pp. 282–286, Aug. 2000.

[18] S. Choi and K. G. Shin, "An Uplink CDMA System Architecture with Diverse QoS Guarantees for Heterogeneous Traffic," *IEEE/ACM Transactions on Networking*, vol. 7, pp. 616–628, Oct. 1999.

[19] R. L. Cruz, "A Calculus for Network Delay, Part I: Network Elements in Isolation," *IEEE Transactions on Information Theory*, vol. 37, pp. 114–131, Jan. 1991.

[20] R. L. Cruz, "A Calculus for Network Delay, Part II: Network Analysis," *IEEE Transactions on Information Theory*, vol. 37, pp. 132–141, Jan. 1991.

[21] L. A. DaSilva, "QoS Mapping along the Protocol Stack: Discussion and Preliminary Results," in *Proc. IEEE International Conference on Communications*, (New Orleans, USA), pp. 713–717, May 2000.

[22] Ö. Gürbüz and H. Owen, "Dynamic Resource Scheduling Schemes for W-CDMA Systems," *IEEE Communications Magazine*, vol. 38, pp. 80–84, Oct. 2000.

[23] N. Dimitriou, G. Sfikas, and R. Tafazolli, "Quality of Service for Multimedia CDMA," *IEEE Communications Magazine*, vol. 38, pp. 88–94, July 2000.

[24] S. Dixit, Y. Guo, and Z. Antoniou, "Resource Management and Quality of Service in Third-Genration Wireless Networks," *IEEE Communications Magazine*, vol. 39, pp. 125–133, Feb. 2001.

[25] P. Ferguson and G. Huston, *Quality of Service, Delivering QoS on the Internet and in Corporate Networks*. Wiley Computer Publishing, John Wiley & Sons, Inc., 1998.

[26] G. Fodor, B. Olin, F. Persson, C. Roobol, and B. Williams, "Providing Diffentiated- and Integrated Services for IP Applications over UMTS Access Networks," in *Proc. The 4th International Symposium on Wireless Personal Mutimedia Communications*, (Aalborg, Denmark), pp. 17–31, Sept. 2001.

[27] A. Folkestad, *Traffic Analysis of Resource Usage and Load Control in Distributed Systems*. PhD thesis, Department of Telematics, Norwegian University of Science and Technology, 1998.

[28] L. Georgiadis, R. Guerin, V. Peris, and R. Rajan, "Efficient Support of Delay and Rate Guarantees in an Internet," in *Proc. ACM Annual Conference of the Special Interest Group on Data Communication*, Aug. 1996.

[29] J. Glasmann, M. Czermin, and A. Reidl, "Estimation of Token Bucket Parameters for Videoconferencing System in Corporate Networks," in *Proc. International Conference on Software, Telecommunications and Computer Networks*, (Rijeka-Venice, Croatia-Italy), Oct. 2000.

[30] D. Goderis, S. V. D. Bosch, Y. T'joens, O. Poupel, C. Jacquenet, G. Memenios, G. Pavlou, R. Egan, D. Griffin, P. Georgatsos, L. Georgiadis, and P. V. Heuven, "Service Level Specification Semantics, Parameters and Negotiation Requirements," *draft-tequila-sls-02, Internet draft, Work in Process*, Feb. 2002.

[31] M. Grossglauser and D. N. C. Tse, "A Framework for Robust Measurement-Based Admission Control," *IEEE/ACM Transactions on Networking*, vol. 7, pp. 293–309, June 2002.

[32] R. A. Guerin and V. Pla, "Aggregation and Conformance in Differentiated Service Networks: A Case Study," *ACM Computer Communication Review*, vol. 31, pp. 21–32, Jan. 2001.

[33] Y. Guo and H. Chaskar, "Class-Based Quality of Service over Air Interface in 4G Mobile Networks," *IEEE Communications Magazine*, vol. 40, pp. 132–137, May 2002.

[34] J. Heinanen, F. Baker, W. Weiss, and J. Wroclawski, "Assured Forwarding PHB Group," *RFC 2597, IETF*, June 1999.

[35] H. Holma and A. Toskala, *WCDMA for UMTS Radio Access for Third Generation Mobile Communications*. London: John Wiley & Sons, Ltd, 2001.

[36] J.-F. Huard and A. A. Lazar, "On End-to-End QoS Mapping," in *Proc. IEEE/IFIP International Workshop on Quality of Service*, (New York, USA), May 1997.

[37] Y. Ishikawa and N. Umeda, "Capacity Design and Performance of Call Admission Control in Cellular CDMA Systems," *IEEE Journal on Selected Areas in Communications*, vol. 15, pp. 1627–1635, Oct. 1997.

[38] V. Jacobson, K. Nichols, and K. Poduri, "An Expedited Forwarding PHB," *RFC 2598, IETF*, June 1999.

[39] L. Jorguseski, E. Fledderus, J. Farserotu, and R. Prasad, "Radio Resource Allocation in Third-Generation Mobile Communication Systems," *IEEE Communications Magazine*, vol. 39, pp. 117–123, Feb. 2001.

[40] H. Kaaranen, A. Ahtiainen, L. Laitinen, S. Naghian, and V. Niemi, *UMTS Networks Architecture, Mobility and Services*. London: John Wiley & Sons, Ltd, 2001.

[41] S. Kim, T. Kwon, and Y. Choi, "Call Admission Control for Prioritized Adaptive Multimedia Services in Wireless/Mobile Networks," in *Proc. IEEE Vehicular Technology Conference*, (Tokyo, Japan), pp. 1536–1540, May 2000.

[42] R. Koodli and M. Puuskari, "Supporting Packet-Data QoS in Next-Generation Celluar Networks," *IEEE Communications Magazine*, vol. 39, Feb. 2001.

[43] F. Y. Li, "Local and Global QoS-aware Token Bucket Parameters Determination for Traffic Conditioning in 3rd Generation Wireless Networks," in *Proc. European Wireless*, (Florence, Italy), pp. 362–368, Feb. 2002.

[44] F. Y. Li and N. Stol, "A Priority-oriented Call Admission Control Paradigm with QoS Re-negotiation for Multimedia Services in UMTS," in *Proc. IEEE Vehicular Technology Conference*, (Rhodos, Greece), pp. 2021–2025, May 2001.

[45] F. Y. Li and N. Stol, "Providing Conformance of the Negotiated QoS Using Traffic Conditioning for Heterogeneous Services in WCDMA Radio Access Networks," in *Proc. Norwegian Informatics Conference*, (Tromsø, Norway), pp. 117–128, Nov. 2001.

[46] F. Y. Li and N. Stol, "QoS Provisioning using Traffic Shaping and Policing in 3rd-generation Wireless Networks," in *Proc. IEEE Wireless Communications and Networking Conference*, (Orlando, USA), pp. 139–143, Mar. 2002.

[47] F. Y. Li and N. Stol, "A Study on Traffic Shaping, Policing and Conformance Deterioration for QoS Contracted Networks," in *Proc. IEEE Global Telecommunications Conference*, (Taipei, Taiwan), Nov. 2002.

[48] F. Y. Li, N. Stol, T. T. Pham, and S. Andresen, "A Priority-oriented QoS Management Framework for Multimedia Services in UMTS," in *Proc. The 4th International Symposium on Wireless Personal Mutimedia Communications*, (Aalborg, Denmark), pp. 681–686, Sept. 2001.

[49] Y. Li, A. Svensson, and S. Falahati, "Hybrid Type-II ARQ Schemes with Scheduling for Packet Data Transmission over Rayleigh Fading Channels," in *Proc. The 3rd ITG Conference on Source and Channel Coding*, (Munich, Germany), pp. 293–299, Jan. 2000.

[50] T-K. Liu and J. A. Silvester, "Joint Admission/Congestion Control for Wireless CDMA Systems Supporting Integrated Services," *IEEE Journal on Selected Areas in Communications*, vol. 16, pp. 845–857, Aug. 1998.

[51] Z. Liu and M. E. Zarki, "SIR-Based Call Admission Control for DS-CDMA Cellular Systems," *IEEE Journal on Selected Areas in Communications*, vol. 12, pp. 638–644, May 1994.

[52] A. Lombardo, G. Morabito, and G. Schembra, "Traffic Specification for MPEG Video Transmission over the Internet," in *Proc. IEEE International Conference on Communications*, (New Orleans, USA), pp. 853–857, June 2000.

[53] S. I. Maniatis, E. G. Nikolouzou, and I. S. Venieris, "QoS Issues in the Converged 3G Wireless and Wired Networks," *IEEE Communications Magazine*, vol. 40, pp. 44–53, Aug. 2002.

[54] D. McDysan, *QoS & Traffic Management in IP & ATM Networks*. N. Y.: McGraw-Hill, 2000.

[55] C. Mihailescu, X. Lagrange, and P. Godlewski, "Radio Resource Management for Packet Transmission in UMTS WCDMA System," in *Proc. IEEE Vehicular Technology Conference*, (Amsterdam, The Netherlands), pp. 573–577, Sept. 1999.

[56] C. Partridge, *Gigabit Networking*. Addison-Wesley Publishing Company, 1993.

[57] C. Partridge, "Manual Page of TB Program." BBN Systems and Technologies, Unpublished, 1995.

[58] H. G. Perros and K. M. Elsayed, "Call Admission Control Schemes: A Review," *IEEE Communications Magazine*, vol. 34, pp. 82–91, Nov. 2001.

[59] T. T. Pham, A. Perkis, and F. Y. Li, "Call Admission Control Algorithm for Multichannel Users in Hierarchical Cellular Systems," in *Proc. IEEE International Conference on Third Generation Wireless and Beyond*, (San Francisco, USA), June 2001.

[60] V. Phan-Van and S. Glisic, "Radio Resource Management in CDMA Cellular Segements of Multidemia Wireless IP Networks," in *Proc. The 4th International Symposium on Wireless Personal Mutimedia Communications*, (Aalborg, Denmark), pp. 57–73, Sept. 2001.

[61] R. Prasad, W. Mohr, and W. Konäuser (ed.), *Third Generation Mobile Communication System*. London: Artech House, 2000.

[62] L. A. Rønningen, "Analysis of a Traffic Shaping Scheme," in *Proc. The 10th International Teletraffic Congress*, (Montreal, Canada), 1983.

[63] M. Schwartz, *Broadband Integrated Networks*. New Jersey: Prentice-Hall, Inc., 1996.

[64] T. Shan and O. W. W. Yang, "Improving Resource Utilization for the Rate-Controlled Traffic Flows in High Speed Networks," in *Proc. IEEE International Conference on Communications*, (Vancouver, Canada), pp. 864–868, June 1999.

[65] D. Shen and C. Ji, "Admission Control of Multimedia Traffic for Third Generation CDMA Network," in *Proc. IEEE Conference on Computer Communications*, (Tel-Aviv, Israel), pp. 1077–1086, Mar. 2000.

[66] S. Shenker, C. Patridge, and R. Guerin, "Specification of Guaranteed Quality of Service," *RFC 2212, IETF*, Sept. 1997.

[67] S. Shenker and J. Wroclawski, "General Characterization Parameters for Integrated Service Network Elements," *RFC 2215, IETF*, Nov. 1997.

[68] M. Sidi, W-Z. Liu, I. Cidon, and I. Copal, "Congestion Control Through Input Rate Regulation," *IEEE Transactions on Communications*, vol. 41, pp. 471–477, Mar. 1993.

[69] ETSI SMG, "Selection Procedures for the Choice of Radio Transmission Technologies of the UMTS." http://www.etsi.fr, UMTS 30.03v3.2.0, April, 1998.

[70] W. Stallings, *High-speed Networks: TCP/IP and ATM Design Principles*. New Jersey: Prentics-Hall,Inc., 1998.

[71] P. P. Tang and T-Y. C. Tai, "Network Traffic Charcterization Using Token Bucket Model," in *Proc. IEEE Conference on Computer Communications*, vol. 1, (New York, USA), pp. 51–62, Mar. 1999.

[72] O. Verscheure, P. Frossard, and J. Y. Le Boudec, "Joint Smoothing and Source Rate Selection for Guaranteed Service Networks," in *Proc. IEEE Conference on Computer Communications*, (Anchorage, USA), pp. 613–620, Apr. 2001.

[73] A. M. Viterbi and A. J. Viterbi, "Erlang Capacity of a Power Controlled CDMA Sytem," *IEEE Journal on Selected Areas in Communications*, vol. 11, pp. 892–900, Aug. 1993.

[74] B. H. Walke, *Mobile Radio Networks Networking and Protocols*. London: John Wiley & Sons, Ltd, 1999.

[75] J. Wroclawski, "Specification of the Controlled-Load Network Element Service," *RFC 2211, IETF*, Sept. 1997.

[76] J. Wroclawski, "The Use of RSVP with IETF Integrated Services," *RFC 2210, IETF*, Sept. 1997.

[77] W-B. Yang and E. Geraniotis, “Admission Policies for Integrated Voice and Data Traffic in CDMA Packet Radio Networks,” *IEEE Journal on Selected Areas in Communications*, vol. 12, pp. 654–664, May 1994.

[78] R. Yavatkar, D. Pendarakis, and R. Guerin, “A Framework for Policy-based Admission Control,” *RFC 2753, IETF*, Jan. 2000.