Abstract:
A main challenge for UMTS is to convey various types of traffic on the same medium while meeting their different QoS requirements. QoS is an end-to-end concept that has to be satisfied through the inter-working of all the entities that the data is passing through. Major approaches for QoS provisioning in IP backbone network are DiffServ and InterServ. UMTS standardization as performed in 3GPP focuses on QoS signalling for the UMTS part (UE to GGSN) only, and it does not contain the actual QoS provisioning mechanisms for the data transport. This report proposes detailed algorithms and packet treatment strategies for the UMTS core network and analyze them in a large set of simulation experiments. The latter focus on revealing the impact of non-real time services on real-time service in different scenarios with and without full QoS provisioning mechanisms. The implemented algorithms include Call Admission Control, policing, scheduling and queuing schemes.
Preface

This report has been written by group MOB-1122 of the 9th and 10th semester of the International Master of Science (M.Sc.E) in Mobile communication. It constitutes the documentation of the project work done about End-to-End Quality of service provisioning of UMTS network. The report is primarily intended to students and supervisors of Aalborg University.

Reference to literature are in squared brackets with the number that appears in the bibliography, i.e [1].

Figure, Table and Equation Numbering follows the chapter numbering, i.e. Figure 2 of chapter 2 is called Figure 2.2, Table 3 of chapter 5 is called Table 5.3 and Equation 3 of chapter 9 is called Equation 9.3.

Group MOB-1122, Aalborg University, June 16, 2005

__________________________________________
Haibo Wang                                  Devendra Prasad
ACKNOWLEDGEMENTS

We would like to thank our two supervisors, Hans-Peter Schwefel and Oumer Teyeb for their patience during all the phases of this project. We would like to express our gratitude to Dr. Hans to fulfil our dream of doing Masters in a world-class university. He has provided us guidance, support, motivation to succeed, and countless highlighting discussion. We are grateful to Oumer for his advice on our simulator.

Finally, we give a very special thanks to all the professors and PhD students who helped us during the Master’s Program. We would also like to thank our family, back home in India and China, for their unconditional support in all situations.
Contents

1 Introduction 13
  1.1 Motivation .................................................. 13
  1.2 QoS Basic Concept ........................................... 14
  1.3 Problem Statement ........................................... 15
  1.4 Chapter Organization ......................................... 16

2 Overview of UMTS 17
  2.1 UMTS Services ................................................ 17
  2.2 UMTS Architecture ............................................ 18
    2.2.1 User Equipment ......................................... 18
    2.2.2 UTRAN .................................................. 19
    2.2.3 Core Network (CN) ....................................... 19
    2.2.4 Interfaces ............................................... 20
  2.3 UMTS protocol stack .......................................... 21
    2.3.1 UMTS user plane ......................................... 21
    2.3.2 UMTS control plane ...................................... 22
  2.4 UMTS Air Interface ........................................... 23
    2.4.1 The Physical layer ...................................... 24
    2.4.2 The MAC layer .......................................... 24
    2.4.3 RLC layer ............................................... 25
    2.4.4 RRC layer ............................................... 25
    2.4.5 Transport channels ...................................... 25
  2.5 Packet Data Protocol (PDP) Context .......................... 26
    2.5.1 PDP context states ...................................... 27
    2.5.2 PDP context procedures ................................ 27

3 UMTS QoS 29
  3.1 UMTS QoS architecture ...................................... 29
  3.2 UMTS QoS classes ............................................ 30
  3.3 UMTS QoS mechanisms ....................................... 31
    3.3.1 QoS Profile ............................................ 31
    3.3.2 Traffic Flow Template .................................. 31
    3.3.3 PDP context signalling procedure .................... 33

4
### CONTENTS

3.4 UMTS QoS management entities and mechanism .................................. 37
   3.4.1 QoS management in the control plane .................................................. 37
   3.4.2 QoS management in the user plane .......................................................... 38
3.5 Conclusion ........................................................................................................ 39

4 IP QoS ............................................................................................................... 40
   4.1 DiffServ approach ......................................................................................... 40
      4.1.1 Per Hop Behaviors (PHB) and Differentiated Services Code Point (DSCP) .... 40
   4.1.2 Per Hop Behaviors .................................................................................. 42
   4.1.3 Traffic classification and conditioning ...................................................... 44
   4.1.4 Meter and Marker functions - Policy ...................................................... 45
   4.2 IntServ/RSVP Policy .................................................................................... 49
   4.3 Comparison between IntServ and DiffServ ................................................... 50
   4.4 Other IP QoS provisioning mechanisms ....................................................... 52
      4.4.1 Scheduling ......................................................................................... 52
      4.4.2 Queuing - RED .................................................................................. 53
   4.5 Conclusion .................................................................................................... 54

5 NS-2 simulator .................................................................................................... 55
   5.1 Overview of NS-2 ......................................................................................... 55
   5.2 DiffServ module ............................................................................................ 57
      5.2.1 AF PHB Implementation with multiple RED queues .............................. 57
      5.2.2 Policer ................................................................................................. 58
      5.2.3 Scheduler ............................................................................................ 58
   5.3 UTRAN extension .......................................................................................... 59
      5.3.1 Overview of the UMTS model .............................................................. 59
      5.3.2 UTRAN nodes and links ...................................................................... 59
      5.3.3 RLC object ......................................................................................... 60
      5.3.4 Error Model for wireless link .............................................................. 60
      5.3.5 EURANE limitations ........................................................................... 60

6 Delimitation ......................................................................................................... 62
   6.1 State-of-the-Art .............................................................................................. 62
      6.1.1 E2E framework ..................................................................................... 63
      6.1.2 QoS mapping: SAMU project ............................................................... 63
      6.1.3 Call Admission Control (CAC) ............................................................. 65
      6.1.4 Simulation: SEACORN project ............................................................ 66
      6.1.5 Radio Resource Management ............................................................. 66
      6.1.6 Conclusion ......................................................................................... 67
   6.2 Problem Delimitation ...................................................................................... 67
List of Figures

1.1 Entities involved in E2E UMTS QoS Provision ........................................... 15
2.1 UMTS Network Architecture in Release 99 ................................................. 18
2.3 UMTS Protocol Stack: User Plane ............................................................... 21
2.4 UMTS Protocol Stack: Control Plane ........................................................... 22
2.5 PDP Context State Transition ....................................................................... 27
3.1 UMTS End-to-End QoS Architecture ............................................................ 29
3.5 UMTS QoS Management functions in the control plane[1] .............................. 37
3.6 UMTS QoS Management functions in the user plane[1] .................................. 38
4.1 The DS field structure ..................................................................................... 41
4.2 The architecture a DiffServ Domain ............................................................... 42
4.3 Assured Forwarding PHB Group ................................................................. 43
4.4 Possible implementation of Expedited Forwarding PHB ............................... 43
4.5 Logical structure of a packet classifier and traffic conditioner ....................... 44
4.6 An example TSW meter algorithm [16] ......................................................... 46
4.7 TSW3CM Marking Algorithm [16] ............................................................... 46
4.8 Token Bucket algorithm ............................................................................... 48
4.9 IntServ/RSVP Policy ..................................................................................... 49
4.10 WIRR scheduling ......................................................................................... 53
5.1 Simplified User View of NS .......................................................................... 55
5.2 UMTS simulation realization in NS2 .............................................................. 59
5.3 UMTS simulation protocol stack .................................................................... 60
6.1 Layered Service architecture of an E2E QoS chain of UMTS ....................... 63
7.1 E2E Network topology in using Best Effort strategy ..................................... 69
7.2 UTRAN delay simulated in NS2-EURANE .................................................. 71
### LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>7.3</td>
<td>Traffic throughput in a 10% overload scenario</td>
<td>72</td>
</tr>
<tr>
<td>7.4</td>
<td>Droptail queue size evolution</td>
<td>73</td>
</tr>
<tr>
<td>7.5</td>
<td>RED queue size evolution</td>
<td>73</td>
</tr>
<tr>
<td>8.1</td>
<td>Simplified End2End QoS investigate scenario</td>
<td>75</td>
</tr>
<tr>
<td>8.2</td>
<td>Packet data flow in AM [13]</td>
<td>77</td>
</tr>
<tr>
<td>8.3</td>
<td>Policing, buffering and scheduling Implementation</td>
<td>80</td>
</tr>
<tr>
<td>8.4</td>
<td>PRI scheduler</td>
<td>81</td>
</tr>
<tr>
<td>8.5</td>
<td>WRR scheduler</td>
<td>82</td>
</tr>
<tr>
<td>8.6</td>
<td>Implemented QoS algorithm</td>
<td>83</td>
</tr>
<tr>
<td>9.1</td>
<td>Simplified End2End QoS investigate scenario</td>
<td>84</td>
</tr>
<tr>
<td>9.2</td>
<td>Simulation Parameters</td>
<td>85</td>
</tr>
<tr>
<td>9.3</td>
<td>Traffic Parameters in Simulation</td>
<td>86</td>
</tr>
<tr>
<td>9.4</td>
<td>Voice Packet Delay Distribution in Best Effort strategy</td>
<td>87</td>
</tr>
<tr>
<td>9.5</td>
<td>VoIP Packet Delay Distribution in CIR-based QoS algorithm</td>
<td>88</td>
</tr>
<tr>
<td>9.6</td>
<td>Video Packet Delay Distribution in CIR-based QoS algorithm</td>
<td>89</td>
</tr>
<tr>
<td>9.7</td>
<td>VoIP Packet Delay Distribution in proposed QoS algorithm</td>
<td>90</td>
</tr>
<tr>
<td>9.8</td>
<td>Video Streaming Packet Delay Distribution in proposed QoS algorithm</td>
<td>90</td>
</tr>
<tr>
<td>9.9</td>
<td>VoIP Packet Delay Distribution in proposed QoS algorithm</td>
<td>92</td>
</tr>
<tr>
<td>9.10</td>
<td>Video Packet Delay Distribution in proposed QoS algorithm</td>
<td>92</td>
</tr>
</tbody>
</table>
List of Tables

2.1 UTRAN Key Parameters ............................................ 23

3.1 UMTS QoS classes .................................................. 31
3.2 UMTS bearer attributes defined for each traffic class .............. 32
3.3 Valid Packet Filter Attribute Combinations .......................... 33

4.1 A comparison of the IntServ and DiffServ architectures .............. 51

6.1 Mapping of Differentiated Services Classes to UMTS QoS Classes .... 65
6.2 Mapping of Integrated Services Classes to UMTS QoS Classes ....... 65

7.1 Traffic model ......................................................... 70
7.2 Simulation settings ................................................... 70

8.1 Traffic Model Assumptions ........................................... 75
8.2 UMTS QoS requirements of different service classes ................. 76
8.3 UMTS Vs DiffServ Mapping ........................................ 78

9.1 UMTS QoS requirements of each service type ......................... 84
9.2 CIR settings for CIR-based Admission Control ........................ 88
9.3 Simulation result of CIR-based CAC with QoS algorithm ............... 88
9.4 Mean sending rate estimation for Mean-based Admission Control ...... 89
9.5 Simulation result of Mean-based CAC with algorithm .................. 89
9.6 QoS provisioning settings ........................................... 91
9.7 Simulation result of proposed EB-CAC with QoS algorithm .......... 91
## Abbreviation List

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>AF</td>
<td>Assured Forwarding</td>
</tr>
<tr>
<td>AUC</td>
<td>Authentication Center</td>
</tr>
<tr>
<td>AM</td>
<td>Acknowledged Mode</td>
</tr>
<tr>
<td>ARQ</td>
<td>Automatic Repeat Request</td>
</tr>
<tr>
<td>BA</td>
<td>Behavior Aggregate</td>
</tr>
<tr>
<td>BB</td>
<td>BAndwidth Broker</td>
</tr>
<tr>
<td>BE</td>
<td>Best Effort</td>
</tr>
<tr>
<td>BER</td>
<td>Bit Error Rate</td>
</tr>
<tr>
<td>BS</td>
<td>Base Station</td>
</tr>
<tr>
<td>CAC</td>
<td>Call Admission Control</td>
</tr>
<tr>
<td>CIR</td>
<td>Committed Information Rate</td>
</tr>
<tr>
<td>CL</td>
<td>Controlled-Load Service</td>
</tr>
<tr>
<td>CN</td>
<td>Core Network</td>
</tr>
<tr>
<td>CS</td>
<td>circuit switched</td>
</tr>
<tr>
<td>CCTrCH</td>
<td>Coded Composite Transport Channels</td>
</tr>
<tr>
<td>DCH</td>
<td>Dedicated Channel</td>
</tr>
<tr>
<td>DiffServ</td>
<td>Differentiated Services</td>
</tr>
<tr>
<td>DSCP</td>
<td>Differentiated Service Code Point</td>
</tr>
<tr>
<td>EB</td>
<td>Equivalent Bandwidth</td>
</tr>
<tr>
<td>E2E</td>
<td>End-to-End</td>
</tr>
<tr>
<td>EF</td>
<td>Expedited Forwarding</td>
</tr>
<tr>
<td>ETSI</td>
<td>European Telecommunications Standards Institute</td>
</tr>
<tr>
<td>EIR</td>
<td>Equipment Identification Register</td>
</tr>
<tr>
<td>FACH</td>
<td>Forward Access Channel</td>
</tr>
<tr>
<td>FDD</td>
<td>Frequency Division Duplex</td>
</tr>
<tr>
<td>FEC</td>
<td>Forward Error Correction</td>
</tr>
<tr>
<td>FER</td>
<td>Frame Error Rate</td>
</tr>
<tr>
<td>QoS</td>
<td>Quality of Service</td>
</tr>
<tr>
<td>3G</td>
<td>Third Generation</td>
</tr>
<tr>
<td>GMSC</td>
<td>Gateway Mobile services Switching Center</td>
</tr>
<tr>
<td>GSM</td>
<td>Global System for Mobile Communication</td>
</tr>
<tr>
<td>GGSN</td>
<td>Gateway GPRS Support Node</td>
</tr>
<tr>
<td>3GPP</td>
<td>Third Generation Partnership Project</td>
</tr>
<tr>
<td>Acronym</td>
<td>Full Form</td>
</tr>
<tr>
<td>---------</td>
<td>-----------</td>
</tr>
<tr>
<td>3GPP2</td>
<td>Third Generation Partnership Project</td>
</tr>
<tr>
<td>GPRS</td>
<td>General packet radio Service</td>
</tr>
<tr>
<td>GTP</td>
<td>GPRS Tunneling Protocol</td>
</tr>
<tr>
<td>GTP-U</td>
<td>GPRS Tunneling Protocol for the user plane</td>
</tr>
<tr>
<td>GS</td>
<td>Guaranteed Service</td>
</tr>
<tr>
<td>HLR</td>
<td>Home Location Register</td>
</tr>
<tr>
<td>IntServ</td>
<td>Integrated Services</td>
</tr>
<tr>
<td>IETF</td>
<td>Internet engineering Task Force</td>
</tr>
<tr>
<td>IMT-2000</td>
<td>International Mobile Telecommunications 2000</td>
</tr>
<tr>
<td>IP</td>
<td>Internet Protocol</td>
</tr>
<tr>
<td>ISP</td>
<td>Internet Service Provider</td>
</tr>
<tr>
<td>ISDN</td>
<td>Integrated Services Digital Network</td>
</tr>
<tr>
<td>ITU</td>
<td>International Telecommunication Union</td>
</tr>
<tr>
<td>Kbps</td>
<td>kilobits per second</td>
</tr>
<tr>
<td>KPI</td>
<td>Key Performance Indicator</td>
</tr>
<tr>
<td>MAC</td>
<td>Medium Access Control</td>
</tr>
<tr>
<td>MF</td>
<td>MultiField Aggregate</td>
</tr>
<tr>
<td>MMS</td>
<td>Multimedia Message Services</td>
</tr>
<tr>
<td>MSC</td>
<td>Mobile services Switching Center</td>
</tr>
<tr>
<td>NS-2</td>
<td>Network Simulator 2</td>
</tr>
<tr>
<td>NRT</td>
<td>NonReal Time</td>
</tr>
<tr>
<td>NSAPI</td>
<td>Network Service Access Point Identifier</td>
</tr>
<tr>
<td>PCM</td>
<td>Pulse Code Modulation</td>
</tr>
<tr>
<td>PDP</td>
<td>Packet Data Protocol</td>
</tr>
<tr>
<td>PDCP</td>
<td>Packet Data Convergence Protocol</td>
</tr>
<tr>
<td>PDU</td>
<td>Protocol Data Unit</td>
</tr>
<tr>
<td>PIR</td>
<td>Peak Information rate</td>
</tr>
<tr>
<td>PHB</td>
<td>Per Hop Behavior</td>
</tr>
<tr>
<td>PPP</td>
<td>Point-to-point</td>
</tr>
<tr>
<td>PRI</td>
<td>Priority</td>
</tr>
<tr>
<td>PS</td>
<td>Packet Switched</td>
</tr>
<tr>
<td>PHY</td>
<td>Physical layer</td>
</tr>
<tr>
<td>RABs</td>
<td>Radio Access Bearers</td>
</tr>
<tr>
<td>RACH</td>
<td>Random Access Channel</td>
</tr>
<tr>
<td>RANAP:</td>
<td>Radio Access Network Application Protocol</td>
</tr>
<tr>
<td>RED</td>
<td>Random Early Detection</td>
</tr>
<tr>
<td>RLC</td>
<td>Radio Link Control</td>
</tr>
<tr>
<td>RNC</td>
<td>Radio Network Controller</td>
</tr>
<tr>
<td>RR</td>
<td>Round Robin</td>
</tr>
<tr>
<td>RRC</td>
<td>Radio Resource Control</td>
</tr>
<tr>
<td>RRM</td>
<td>Radio Resource Management</td>
</tr>
<tr>
<td>RSVP</td>
<td>Resource Reservation Protocol</td>
</tr>
<tr>
<td>RT</td>
<td>Real Time</td>
</tr>
<tr>
<td>Acronym</td>
<td>Description</td>
</tr>
<tr>
<td>---------</td>
<td>-------------</td>
</tr>
<tr>
<td>SDU</td>
<td>Service Data Units</td>
</tr>
<tr>
<td>SLA</td>
<td>Service layer Agreement</td>
</tr>
<tr>
<td>SMS</td>
<td>Short Message Service</td>
</tr>
<tr>
<td>SGSN</td>
<td>Serving GPRS Support Node</td>
</tr>
<tr>
<td>SPI</td>
<td>Security Parameter Index</td>
</tr>
<tr>
<td>TCA</td>
<td>Traffic Condition Agreement</td>
</tr>
<tr>
<td>TCP</td>
<td>Transmission Control Protocol</td>
</tr>
<tr>
<td>TDD</td>
<td>Time Division Duplex</td>
</tr>
<tr>
<td>TFT</td>
<td>Traffic Flow Template</td>
</tr>
<tr>
<td>TM</td>
<td>Transparent Mode</td>
</tr>
<tr>
<td>TOS</td>
<td>Type of Service</td>
</tr>
<tr>
<td>TSW3CM/2CM</td>
<td>Time Sliding Window Three/Two Color Marker</td>
</tr>
<tr>
<td>UDP</td>
<td>User Datagram Protocol</td>
</tr>
<tr>
<td>UE</td>
<td>User Equipment</td>
</tr>
<tr>
<td>UM</td>
<td>Unacknowledged Mode</td>
</tr>
<tr>
<td>UMTS</td>
<td>Universal Mobile Telecommunications System</td>
</tr>
<tr>
<td>UTRAN</td>
<td>UMTS Terrestrial Radio Access Network</td>
</tr>
<tr>
<td>USIM</td>
<td>UMTS Subscriber Identity Module</td>
</tr>
<tr>
<td>VLR</td>
<td>Visitor location register</td>
</tr>
<tr>
<td>VoIP</td>
<td>Voice over IP</td>
</tr>
<tr>
<td>WCDMA</td>
<td>Wideband Code Division Multiple Access</td>
</tr>
<tr>
<td>CDMA</td>
<td>Code Division Multiple Access</td>
</tr>
<tr>
<td>WIRR</td>
<td>Weighted Interleaved Round Robin</td>
</tr>
<tr>
<td>WRR</td>
<td>weighted Round Robin</td>
</tr>
</tbody>
</table>
Chapter 1

Introduction

1.1 Motivation

The end-to-end (E2E) Quality of Service (QoS) provision in both the standards for Third Generation (3G) mobile radio systems (in Third Generation Partnership Project, 3GPP/3GPP2) and for IP infrastructure (in IETF). The market requires the real-time and quality-assured services provision by the mobile communication world which is going to be a convergence of all-IP backbone [3GPP Release 5] and Universal Mobile Telecommunications System (UMTS) radio access network.

UMTS has been designed to support integrated services such as voice, data, video, etc. The broadband multi-media services is regarded as the key to UMTS success. Some services require UMTS to provide bandwidth on demand, mixed traffic types, efficient network transport and/or guaranteed QoS features. Those services may be divided into real time (RT) and non-real time with quite different demand to the UMTS network.

The QoS provisioning means that the multimedia traffic should get predictable service from the available resources. Typical resources are CPU time (for the communication software to execute) and network bandwidth. The software must also guarantee an acceptable end-to-end delay and maximum delay jitter i.e., maximum allowed variance in the arrival of data at the destination. In most case, the 3-tuple specifies QoS requirements: bandwidth, delay and reliability. The QoS provisioning problem for multimedia traffic in wired network has been extensively studied. The ongoing work in this field mainly concentrates on the problems of bandwidth management and switch based scheduling to provide guarantee on end-to-end delay, throughput and packet losses.

There are few major differences between the wire line and wireless networks due to link characteristics and mobility [1]. The broadband wire-line network transmission link characterized by high transmission rates (in the order of Gbps) and a very low error rates.
In contrast, wireless links have a much smaller transmission rate (Kbps-Mbps). In wire-
less networks the errors are more likely to come in bursts due to multipath dispersion and
Rayleigh fading thus requiring a more complex error detection or forward error correc-
tion. Another major difference between the two networks is the user mobility. In wire-line
networks, the User-networks interface (UNI) remains fixed throughout the duration of a
connection whereas the UNI in a wireless networks keeps on changing throughout the
connection. Therefore, it is necessary to redesign or revise the usual provisioning tech-
niques for wireless networks.

The focus of the project is on investigating the E2E QoS provisioning for Real time
services including both external IP network (Internet) and the packet-switched domain
of UMTS network. Related concepts are reviewed, well-known approaches in both In-
ternet and UMTS system to provide QoS are described, and a new E2E QoS provision
mechanism is proposed.

1.2 QoS Basic Concept

Quality of Service is the collective effectiveness of service performance that determines
the degree of an end user of a given service.

This concept usually leads to the basic idea of QoS, to distinguish traffic into different
types, features, demands to the networks, and to be delivered to the customers on different
charges.

Consequently, certain QoS mechanisms must be implemented to provide/ensure the
E2E QoS features of applications matching their traffic type. We distinguish two main
categories of mechanisms, QoS provision mechanisms and QoS control mechanisms [1],
depending on whether it work before a traffic flow start or during the traffic flow passing
through the communication network:

- QoS provision mechanisms include parameters mapping, admission and resource
  reservations schemes.
- QoS control mechanisms consist of traffic shaping, scheduling, policing and con-
  trol mechanisms.

From application point of view, E2E QoS requirements have to be identified in measurable QoS metrics, which are delay, jitter, loss rate and throughput.[3GPP 1999a]

**Delay** It is the elapsed time for a packet to traverse the network from the source to the
destination. At the network layer, the end-to-end packet latency is the sum of
processing delay, transmission delay, queuing delay and propagation delay.
1.3. PROBLEM STATEMENT

**Jitter** It is defined as the variation in delay encountered by similar packets following the same route through the network. The jitter requirement only affects real-time streaming applications because QoS requirement arises from the continuous traffic characteristics of these class. Jitter is generally included as a performance parameter. since it is very important at the transport layer in packet data systems, due to the inherent variability in arrival times of individual packets. Services intolerant of delay variation will usually try to reduce the delay variation by means of buffering. However, delayed data arrivals make data useless, resulting in receiver buffer underflow. Early arrival can lead to receiver buffer overflow.

**Loss Rate** Loss rate refers to the percentage of data loss among all the delivered data in a given transmission time interval, which can be evaluated in frame level or packet level. Loss rate requirements apply to all classes of applications. The real-time applications might tolerate a limited amount of data loss, depending on the error resiliency of the decoder, and the type of application whereas non-realtime (NRT) applications typically have much more strict requirement on data loss.

**Throughput** It can be defined as the rate at which packets are transmitted in a network. It can be expressed as a peak rate or an average rate.

### 1.3 Problem Statement

End-to-End QoS has to be satisfied through the inter-working of all the entities that the user data is passing through. Figure 1.1 shows the entities involved when a User Equipment (UE) is connected via a UMTS network to a server on the external network. From QoS point of view, the entities can be divided into external IP domain QoS and UMTS domain QoS.

![Diagram of Entities involved in E2E UMTS QoS Provision](image)

Figure 1.1: Entities involved in E2E UMTS QoS Provision

In the UMTS domain, although 3GPP specifies detailed signaling procedure of QoS requirements (in terms of traffic classes, delay parameters, etc.). The standardization
CHAPTER 1. INTRODUCTION

does not provide the mechanisms to support QoS during user-data transfer: QoS parameters need to be mapped to appropriate Radio Resource Management (RRM) strategies in UMTS Terrestrial Radio Access Network (UTRAN) and IP-layer transport mechanisms in the Core Network (CN).

In the IP domain, there are some well-known mechanisms for QoS provisioning, such as Differentiated Services (Diffserv) [2] and Integrated Services (IntServ) [3]. The utilization of these mechanisms in the UMTS architecture is still an open issue.

The first goal of the project is to identify the mechanisms relevant for QoS provisioning in the full E2E transmission chain, both in UMTS network side and in external IP network side. Potential investigation scenarios focus on one or more of the following aspects:

- QoS provisioning in UTRAN via RRM.
- QoS provisioning in the IP based UMTS Core Network, including Gateway GPRS Support Node (GGSN) and Serving GPRS Support Node (SGSN) and/or the routers in UMTS IP backbone.
- QoS mapping between UMTS network and external IP network (Internet).

Once specific investigation aspects are identified, new QoS provisioning algorithm will be proposed and implemented in simulation tool to evaluate its performance.

1.4 Chapter Organization

The report is organized as follows. The first chapter explain the motivation and basic concepts as well as problems in this project. Chapters 2, 3 and 4 review the main characteristics of UMTS network, UMTS QoS and IP QoS, as the background to understand the problem. Chapter 5 analyzes some related projects, their contribution and drawbacks. Chapter 6 delimit the investigation scenarios of this project based on the analysis of Chapter 5 and also the features of the simulation tool. Furthermore, a new E2E QoS provisioning algorithm of UMTS is proposed in Chapter 7, and the implementation is described in Chapter 8 and the simulation result is analyzed in Chapter 8. Finally, conclusion are given in Chapter 9.
Chapter 2

Overview of UMTS

3G Systems are intended to provide a global mobility with wide range of services including telephony, paging, messaging, Internet and broadband data. International Telecommunication Union (ITU) started the process of defining the standard for third generation systems, referred to as International Mobile Telecommunications 2000 (IMT-2000). In Europe, European Telecommunications Standards Institute (ETSI) and later the Third Generation Partnership Project (3GPP) was responsible for Universal Mobile Telecommunications System (UMTS) standardization process.

2.1 UMTS Services

UMTS offers traditional tele-services (such as circuit-switched voice or Short message service (SMS)) and bearer services. It is possible to negotiate and renegotiate the characteristics of a bearer service at session or connection setup stage as well as during ongoing session or connection.

Bearer services have different QoS parameters for maximum transfer delay, delay variation and bit error rate. Offered data rate targets are:

- 144 kbits/s satellite and rural outdoor
- 384 kbits/s urban outdoor
- 2048 kbits/s indoor and low range outdoor

UMTS network services with different QoS classes were divided into four types of traffic:

- Conversational class (voice, video telephony, video gaming)
- Streaming class (multimedia, video on demand, webcast)
- Interactive class (web browsing, network gaming, database access)
- Background class (email, Multimedia Message Services (MMS), file downloading)
2.2 UMTS Architecture

A UMTS network consist of three interacting domains such as Core Network (CN), UMTS Terrestrial Radio Access Network (UTRAN) and User Equipment (UE) [7]. The main function of core network is to provide switching, routing and transit for user traffic. Core network also contains the databases and network management functions. The UTRAN provides the air interface access method for User Equipment. Figure 2.1 depicts the overall architecture of UMTS network, which includes both the UMTS and Global System for Mobile Communication (GSM) network entities. The UTRAN is evolutionary new for UMTS, while the CN can be seen as a smooth evolution from GSM CN. The following subsection will describe the UMTS specific parts in this architecture.

2.2.1 User Equipment

User equipment is the equipment used by the user to access UMTS Services. It is made up of two parts: the UMTS Subscriber Identity Module (USIM) and User Equipment (UE). USIM is a modular IC-card that stores the identity of the subscriber (user), operator and service provider and user service profile. UE is the radio terminal (the mobile phone) for radio communication via the Uu interface.
2.2. UMTS ARCHITECTURE

2.2.2 UTRAN

In UTRAN, Wideband Code Division Multiple Access (WCDMA) technology was selected for radio access, which supports a transmission rate theoretically up to 2Mbit/s (realistic up to about 300kb/s). UMTS WCDMA is a Direct Sequence CDMA system. WCDMA has two basic modes of operation: Frequency Division Duplex (FDD) and Time Division Duplex (TDD). As in Figure 2.2, the UTRAN physical entities are made up of two entities: Node-B and Radio Network Controller (RNC).

**Node-B**

A Node B is a logical node responsible for radio transmission / reception in one or more cells to/from the User Equipment. In UTRAN, it terminates the Iub interface towards the Radio Network Controller (RNC)[9]. It performs Modulation/Demodulation, CDMA Physical Channel coding, Forward Error Correction (FEC), Closed loop power control, and RRM etc.

**RNC**

A RNC controls the usage and the integrity of the radio resources[9]. It controls one or more Node B and act as a access point between UTRAN and CN. Its functions are Radio Resource Control, Admission Control, Channel Allocation, Handover Control, Broadcast Signalling, Open Loop Power Control, etc.

2.2.3 Core Network (CN)

The CN architecture is based on GSM network with GPRS. All equipment has to be modified for UMTS operation and services. The CN is divided in circuit switched (CS) and
packet switched (PS) domains. Some of the circuit switched elements are Mobile services Switching Center (MSC), Visitor location register (VLR) and Gateway MSC(GMSC). Packet switched elements are Serving GPRS Support Node (SGSN) and Gateway GPRS Support Node (GGSN). Some network elements, like Equipment Identification Register (EIR), Home Location Register (HLR), and Authentication Center (AUC) are shared by both domains.

**MSC and GMSC**

The MSC is a switching node that supports circuit-switched connections, user mobility and handover procedures. GMSC is a special MSC which serves as an interface to various external networks, i.e., the Integrated Services Digital Network (ISDN). The MSC is the central network entity of the CS domain.[10]

**HLR and VLR**

The Home Location Register (HLR) is a database in the user’s home system which keeps all the user profile, e.g., roaming areas, allowed services, authentication keys. The VLR is a database similar to HLR which locate the network a user is currently operating. The information in VLR is updated as soon as a user changes location area.

**SGSN and GGSN**

The SGSN carries out routing function in PS domain and manages the authentication and current user position, like MSC/VLR in CS domain. GGSN offers interfaces to external PS networks like the Internet.

According to 3GPP R5 specification, IP will be the main protocol for transporting user packets inside the PS core network. However, user packets are not directly routed via IP routing protocol but based on GPRS Tunneling Protocol (GTP).

**2.2.4 Interfaces**

An interface refers to the connection between different network entities. In this project the most relevant UMTS interfaces are:

- **Uu**: Termed as Uu reference point (UMTS radio interface). It is used as an interface between the user equipment and the UMTS network.

- **IuB**: An interface in between Node-B and RNC.

- **Iu**: This interface connect UTRAN and Core Network. There are two types of Iu, Iu PS for connecting to PS domain and Iu CS connecting to CS domain.

- **Gn**: This one between SGSN and GGSN.
2.3 UMTS PROTOCOL STACK

- **Gi**: Gi is used to interface GGSN with external packet switched network (Internet).

2.3 UMTS protocol stack

The protocols over Uu and Iu interfaces are divided into two structures [11]:

1. **User plane protocols**: The user plane consists of a layered protocol structure providing user information transfer, along with associated information transfer control procedures (e.g., flow control, error detection, error correction and error recovery).

2. **Control plane protocols**: The control plane consists of protocols for control and support of the user plane functions, including requesting the service, controlling different transmission resources, handover, etc.

In the following, we briefly explain these two protocol structures.

2.3.1 UMTS user plane

The UMTS user plane protocol stack is shown in Figure 2.3. The UE and UTRAN include L1, MAC, RLC and PDCP layers. The definition of these layers is as below:

- **Packet Data Convergence Protocol (PDCP)**: This transmission functionality maps higher-level characteristics onto the characteristics of the underlying radio-interface protocols. PDCP provides protocol transparency for higher-layer protocols. PDCP supports IPv4, Point-to-point (PPP) and IPv6. Introduction of new higher-layer protocols shall be possible without any changes to the radio-interface protocols. PDCP provides protocol control information compression. PDCP is specified in 3GPP TS 25.323 [11].
- **Radio Link Control (RLC):** The RLC protocol provides logical link control over the radio interface. There may be several simultaneous RLC links per MS. Each link is identified by a Bearer Id. RLC is defined in 3GPP TS 25.322 [11].

- **Medium Access Control (MAC):** The MAC protocol controls the access signalling (request and grant) procedures for the radio channel. MAC is specified in 3GPP TS 25.321.

- **Physical layer (PHY or L1):** The physical layer offers information transfer services to MAC and higher layers.

UTRAN, SGSN and GGSN include the following layers:

- **GPRS Tunnelling Protocol for the user plane (GTP-U):** This protocol tunnels user data between UTRAN and the SGSN, and between the SGSN and the GGSN in the core network. GTP encapsulates all PDP Protocol Data Units (PDU). The GTP is specified in 3GPP TS 29.060

- **User Datagram Protocol/Internet Protocol (UDP/IP):** These are the backbone network protocols used for routing user data and control signalling.

### 2.3.2 UMTS control plane

The UMTS control plane is illustrated in Figure 2.4. The control plane consists of the following layers:

- **UMTS Mobility Management and Session Management (GMM/SM):** GMM supports mobility management functionality such as attach, detach, security, and routing area update and SM supports PDP context activation and PDP context deactivation.

Figure 2.4: UMTS Protocol Stack: Control Plane
2.4. UMTS AIR INTERFACE

- Short Message Service (SMS) supports the mobile-originated and mobile-terminated short message service.
- Radio Resource Control (RRC): This a signalling protocol for the control and configuration of the radio interface.
- The RLC protocol offers logical link control over the radio interface for the transmission of higher layer-signalling messages and SMS.
- The MAC protocol controls the access signalling (request and grant) procedures for the radio channel.
- The physical layer offers information transfer services to MAC and higher layers.
- Radio Access Network Application Protocol (RANAP): This protocol encapsulates and carries higher-layer signalling, handles signalling between the SGSN and UTRAN, and manages the GTP connections on the Iu interface.

2.4 UMTS Air Interface

UMTS air interface issues are described more specifically in this section because they are highly relevant with most UMTS simulation implementations. The more general elements in the previous two sections can often been simplified.

Wideband CDMA technology was selected for UTRAN air interface. UMTS WCDMA is a Direct Sequence CDMA system. In WCDMA user data is multiplied with quasi-random bits derived from WCDMA Spreading codes. In UMTS, Codes are used for synchronization and scrambling other than channelization. WCDMA has two basic modes of operation: Frequency Division Duplex (FDD) and Time Division Duplex (TDD). In this project only FDD mode is relevant, and its main parameters are shown in Table 2.1.

<table>
<thead>
<tr>
<th>UTRA Mode</th>
<th>FDD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multiple access scheme</td>
<td>W-CDMA</td>
</tr>
<tr>
<td>Carrier spacing</td>
<td>4.4 - 5.2 MHz</td>
</tr>
<tr>
<td>Chip rate</td>
<td>3.84 Mchip/s (Mcps)</td>
</tr>
<tr>
<td>Spreading factor range</td>
<td>4 - 512</td>
</tr>
<tr>
<td>Modulation</td>
<td>QPSK</td>
</tr>
<tr>
<td>Pulse shaping</td>
<td>root raised cosine, roll-off = 0.22</td>
</tr>
<tr>
<td>Frame length</td>
<td>10 ms</td>
</tr>
<tr>
<td>Time-slots per frame</td>
<td>15</td>
</tr>
</tbody>
</table>

Table 2.1: UTRAN Key Parameters
2.4.1 The Physical layer

The physical layer (PHY) is implemented in Node B within the UTRAN. It directly handles the transmission of data over radio. Its tasks related to the air interface include:[10]

- Provisioning of transport channels;
- Mapping of transport channels to physical channels;
- Macro diversity, soft handover(RNC);
- Error protection(FEC, interleaving);
- (De-)Multiplexing of Coded Composite Transport Channels (CCTrCH);
- Synchronization (frequency and time) and power control;
- Measurement of FER, SIR, interference power etc.

The physical layer is controlled by RRC sub-layer. For example, the level of error protection via Forward Error Correction (FEC) can be configured by RRC. The physical layer PHY also report the calculated Frame Error Rate (FER), the Signal to Interference Ratio (SIR) and the measured interference power. The RRC decide the required parameters according to the reported situation.

2.4.2 The MAC layer

The MAC layer coordinates access to the physical medium for data transmission. MAC layer places different data streams to be transmitted in queues, handles priorities/schedule according to the QoS negotiated at connection set-up stage. MAC layer is also associated with RRC sublayer and receive configuration instructions from it. Tasks of MAC is concluded as follow: [10]

- Provisioning of logical channels
- Mapping of logical channels to transport channels
- Selection of appropriate transport formats
- Priority handling/scheduling
- Monitoring of the traffic volume
- Ciphering (for transparent RLC mode)
- No segmentation of data!

The RLC layer delivers the data to MAC layer over logical channels. A logical channel describe which type of data should be transmitted. The MAC layer maps these logical channels to transport channels that represent the interface to the physical layer. The MAC layer thus take charge of multiplexing of several parallel data streams from the logical channels to the transport channels.
2.4. RLC layer

RLC is a sub layer of data link layer. The purpose is to protect data streams from errors caused by radio propagation. To achieve data protection, the RLC make segmentation and reassembly of higher-layer Service Data Units (SDU) into/from smaller RLC payload units (called Protocol Data Unit, PDU). It manage retransmission for the error PDUs. The following mode are configured in RLC layer:

**Transparent Mode (TM)** The RLC doesn’t add a protocol header and erroneous packets are dropped or marked (no error protection). TM is suitable for error-tolerant but delay sensitive data, such as video and audio.

**Unacknowledged Mode (UM)** There are no error correction either, but RLC doesn’t add a sequence number and can ensure the uniqueness of the transmitted data. Segmentation and reassembly is done by adding a header to RLC PDUs. The packets are encrypted.

**Acknowledged Mode (AM)** The Automatic Repeat Request (ARQ) mechanism is used to protect the RLC PDU. The protection level and the delay performance of the RLC can be configured by the RRC. It can be done by changing the maximum retransmissions number. The packets are encrypted.

2.4.4 RRC layer

The Radio Resource Control (RRC) sublayer is in the control phane in layer 3, its main functions include:

- broadcast of system information
- set-up and management of Radio Bearers
- mobility management: location management and handover
- configuration of lower layers
- collection of measurements from lower layers

2.4.5 Transport channels

In UTRA FDD mode, there are logical channels, which are mapped to transport channels, and they are again mapped to physical channels. Logical to Transport channel conversion happens in Medium Access Control (MAC) layer. The transport channels relevant to simulation are:

- Forward Access Channel (FACH)
  FACH is used as down-link Common Packet Channel.
• Random Access Channel (RACH)
  RACH is used for up-link rapid access and transfer of short and small packets.

• Dedicated Channel (DCH)
  The normal data and signalling of a user takes place over the DCH.

2.5 Packet Data Protocol (PDP) Context

The Packet Data Protocol (PDP) context is a data structure present in both the SGSN and the GGSN which contains the subscriber’s session information during subscriber active session. A Mobile needs to acquire and configure itself with a PDP address (i.e. an IP Address when the PDP is IP). Mobile may use single or multiple PDP addresses simultaneously as per requirement. A PDP context must be established (to get PDP address) and activated in the Packet Switched (PS) Core Network (CN) domain and on the mobile, before any user packets destined to or originated from a PDP address can be transported over a 3GPP PS core network.[12]

A PDP context is maintained and used by network nodes to determine the path for user packets destined or originated from a particular PDP address. The PDP context maintained by a mobile, an SGSN, and a GGSN link. The Radio Access Bearers (RABs) and a CN Bearer form a 3GPP Bearer for the mobile.

A PDP context maintained on the SGSN and the GGSN carry the following useful information.

• **PDP address** used by the mobile to and receive PDP Packets.

• **Routing Information** used by the network node to determine where to forward a user packet such as identifier (established between SGSN and GGSN for this PDP context), Access Point Name (APN). APN is a logical name used by SGSN to determine which GGSN is suitable for a mobile and determination of the service request by user or address of an access point in an external packet network to which user packet should be forwarded by GGSN.

• **Quality of Service (QoS) Profile**:

  1. QoS Profile Subscribed: describe QoS characteristic subscribed by mobile user.
  2. QoS profile Requested: Describe QoS characteristic currently requested by mobile user.
  3. QoS profile negotiated; QoS actually provided by the network to the mobile at the current time.
2.5. PACKET DATA PROTOCOL (PDP) CONTEXT

2.5.1 PDP context states

PDP context can be divided into two different states namely ACTIVE and INACTIVE states as shown in Figure 2.5. ACTIVE state contains up-to-date information for forwarding PDP context between the mobile and the GGSN. Active state doesn’t guarantee the RAB establishment over the RAN to transport user packet. RABs may be established only when mobile has a user packet to send to the network or the network user packet for the mobile.

In PDP INACTIVE state may contain a valid PDP address but not valid routing, mapping information needed to determine how to process packets. If GGSN send a user packets to a mobile but the PDP context for the destination PDP address is in INACTIVE state, the GGSN may use network requested PDP context activation procedure to change the PDP context of the destination mobile to ACTIVE state.

2.5.2 PDP context procedures

PDP context is used for user data session management. These session management procedures include:

- **PDP context activation**, in which PDP address, QoS profile are set.
- **PDP context modification**, in which QoS profile is modified.
- **PDP context deactivation**, in which packet data session is released.

A GPRS-attached (GPRS, this terminology is still used by 3GPP to call the packet domain of UMTS CN) MS can initiate the activation, modification, and deactivation functions at any time for a PDP context in the MS, the SGSN, and the GGSN. A GGSN may request the activation of a PDP context to a GPRS-attached subscriber. A GGSN may initiate the
deactivation of a PDP context. The details of these procedures will be explained in the following chapter concern UMTS QoS management issues.
Chapter 3

UMTS QoS

3.1 UMTS QoS architecture

The 3GPP has defined an architecture to supply a certain network QoS between two UEs with several layers of Bearer Services. A bearer service includes all aspects to enable the provision of a contracted QoS. These aspects are among others the control signalling, user plane transport and QoS management functionality. A UMTS bearer service layered architecture is depicted in Figure 3.1, each bearer service on a specific layer offers its individual services using services provided by the layers below. [5]

![Figure 3.1: UMTS End-to-End QoS Architecture](image)

For example, The UMTS Bearer Service offered by service operator provides the UMTS QoS. The UMTS Bearer Service consists of two parts: the Radio Access Bearer (RAB)
CHAPTER 3. UMTS QOS

Service for the cellular wireless network, and the Core Network (CN) Bearer Service for the backbone network.

The RAB service provides confidential transport of signalling and user data between UE and its corresponding SGSN with the QoS adequate to the negotiated UMTS bearer service or with the default QoS for signalling. This service is based on the characteristics of radio interface and is maintained for a moving UE. The radio access bearer service is realized by a Radio Bearer Service and an Iu Bearer Service. The role of the radio bearer service is to cover all the aspects of the radio interface transport. This bearer service uses the UTRA FDD/TDD mechanism. The Iu bearer service provides the transport between UTRAN and CN. The Iu bearer services for packet traffic shall provide different bearer services for a variety of QoS.

The core network bearer service of the UMTS core network connects the SGSN with the GGSN. The role of this service is to efficiently control and utilize the core network in order to provide the contracted UMTS bearer service. The UMTS packet core network shall support different bearer services for a variety of QoS mechanisms. The CN bearer service uses a generic Layer1/Layer2 functionality in order to fulfil the QoS requirements of the core network bearer service. These layers are not specific to UMTS, may reuse an existing standard and is selected according to operator’s choice.

The Bearer Service (BS) concept is important for this project, it allows to divide the E2E QoS problem into a set of sub-problems concerning the QoS provision inside a specific Bearer Service and the QoS profile mapping among them (between two BSs in the same plane or between an upper layer and a lower layer).

3.2 UMTS QoS classes

The UMTS network classify user data flows according to their QoS requirements. The flows that have similar QoS requirements are aggregated together by the network. The 3GPP defined four classes as shown in Table 3.1. Each class is characterized by delay, delay jitter, loss rate and throughput requirements. The main distinguishing factors between the classes are delay sensitivity and loss rate sensitivity. A QoS class is characterized by a set of measurable network parameters values. A specific set of parameters is defined for each bearer service, i.e., the UMTS bearer service has the following attributes:
3.3. UMTS QoS Mechanisms

The main standard UMTS QoS mechanism is the PDP context mechanism. PDP context procedure consists of a series of signalling for the user to set-up a virtual connection with the GGSN and to guarantee its QoS requirements on this connection. A user can have more than one PDP context with different QoS requirements for different applications at the same time. There are two main parts in the PDP context: QoS profile and Traffic Flow Template (TFT). [1]

### 3.3.1 QoS Profile

A QoS profile is a group of user QoS requirements to be used to establish a UMTS bearer service. One QoS profile is associated with one PDP context. The user can negotiate its QoS profile with the network, and the network will provide the QoS level that its resource can support.

### 3.3.2 Traffic Flow Template

When more than one PDP contexts are associated to a PDP address, there shall be a Traffic Flow Template (TFT) for each or all but one context [12]. A TFT is a series of rules for the UMTS networks to detect a flow requiring a certain QoS. It includes one up to eight packet filters, each with a packet filter identifier. If present, the TFT shall be sent transparently via the SGSN to the GGSN to enable packet classification and policing for downlink data transfer (see 3GPP TS 23.060).

A valid packet filter contains at least one of the following attributes:

- Source Address and Subnet Mask.

<table>
<thead>
<tr>
<th>Traffic Class</th>
<th>Examples of Application</th>
<th>Real Time</th>
<th>Delay</th>
<th>Jitter</th>
<th>Low Loss Rate</th>
<th>Guaranteed Bit Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conversational</td>
<td>Voice/Video Telephony</td>
<td>Yes</td>
<td>Stringent</td>
<td>Stringent</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>Streaming</td>
<td>streaming video</td>
<td>Yes</td>
<td>Constrained</td>
<td>Constrained</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>Interactive</td>
<td>Web Browsing</td>
<td>No</td>
<td>Loose</td>
<td></td>
<td></td>
<td>Yes</td>
</tr>
<tr>
<td>Background</td>
<td>Email, Background down- load</td>
<td>No</td>
<td></td>
<td></td>
<td></td>
<td>Yes</td>
</tr>
</tbody>
</table>

Table 3.1: UMTS QoS classes
Table 3.2: UMTS bearer attributes defined for each traffic class
### 3.3. UMTS QOS MECHANISMS

<table>
<thead>
<tr>
<th>Packet filter attribute</th>
<th>Combination I</th>
<th>Combination II</th>
<th>Combination III</th>
</tr>
</thead>
<tbody>
<tr>
<td>Source Address and Subnet Mask</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Protocol Number (IPv4) / Next Header (IPv6)</td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Destination Port Range</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Source Port Range</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Source Port Range</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IPSec SPI</td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>TOS (IPv4) / Traffic Class (IPv6) and Mask</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Flow Label (IPv6)</td>
<td></td>
<td></td>
<td>X</td>
</tr>
</tbody>
</table>

Table 3.3: Valid Packet Filter Attribute Combinations

- Protocol Number (IPv4) / Next Header (IPv6)
- Destination Port Range
- Source Port Range
- IPSec Security Parameter Index (SPI)
- Type of Service (TOS) (IPv4) / Traffic class (IPv6) and Mask
- Flow Label (IPv6)

Some of the listed attributes can be combined in a packet filter but others cannot. Table 3.3 shows the legal combinations, in which the items with an “X” can coexist in one packet filter.

#### 3.3.3 PDP context signalling procedure

Three PDP context procedures are relevant to QoS management, which will be introduced in this subsection.[12]

- The PDP Context Activation Procedure, to establish a PDP context between the MS and the network for a specific QoS on a specific Network Service Access Point Identifier (NSAPI). This context is initiated by the UE or the network.

- The secondary PDP Context activation Procedure, to establish an additional PDP context between the MS and the network for a specific TFT and QoS profile on a specific NSAPI. Each additional PDP context requests a different TFT and QoS profile.
- The PDP Context Modification Procedure, initiated by the network or the MS in order to change the QoS or TFT for a PDP context.

The PDP Context Activation Procedure

Each PDP address may be described by one or more PDP contexts in the UE or the network. The PDP Context Activation procedure as shown in Figure 3.2 is used to activate the first PDP context for a given PDP address and APN, whereas all additional contexts associated to the same PDP address and APN are activated with the secondary PDP context activation procedure.

The PDP Context Activation Procedure will establish the first PDP context for a specific QoS on a specific Network Service Access Point Identifier (NSAPI). It can be initiated by the UE or the network.

![Figure 3.2: PDP Context Activation Procedure for UMTS [11]](image)

1. The UE sends an Activate PDP Context Request message to the SGSN. This message contains a PDP Type, PDP Address and QoS Requested parameters among others. For IP external network, the PDP type and address are IP and IP address. If it request a dynamic PDP(IP) address, this field will be left empty.

2. The SGSN validates the Activate PDP Context Request using the information provided by the UE and the PDP context subscription records. The SGSN has the right to rejects the PDP context activation request if it determined the Activate PDP Context Request is invalid. The SGSN may restrict the requested QoS attributes according to its capabilities, the current load, and the subscribed QoS profile.

3. The SGSN sends a Create PDP Context Request message to the affected GGSN. The message contains PDP Type, PDP Address, QoS Negotiated and Charging Characteristics among other parameters.
4. The GGSN creates a new entry in its PDP context table and generates a Charging Id. The new entry allows the GGSN to route PDP PDUs between the SGSN and the external PDP network, and to start charging. The GGSN then returns a *Create PDP Context Response message* to the SGSN with PDP Address, QoS Negotiated and Charging Id among other parameters. PDP Address is included if the GGSN allocated a PDP address.

5. If the UE has requested a dynamic address, the SGSN inserts the PDP address received from the GGSN in the PDP context. The SGSN returns an *Activate PDP Context Accept message* to the UE with PDP Type, PDP Address and QoS Negotiated. The SGSN is now able to route PDP PDUs between the GGSN and the UE, and to start charging.

For each PDP context a different QoS profile may be requested, depending on the application type associated with it, i.e., a video telephony may require a very strict delay attribute while a E-mail is not sensitive to delay at all. If a QoS requirement is beyond the capabilities of the UMTS network, the network negotiates the QoS profile as close as possible to the requested QoS profile. The UE can either accept the negotiated QoS profile, or deactivate the PDP context. If the PDP Context Activation Procedure fails or if the SGSN returns an Activate PDP Context Reject message, the UE may still set up another activation within a maximum number of attempts.

**The Secondary PDP Context Activation procedure**

The Secondary PDP Context Activation procedure may be used to activate a PDP context with the same PDP address of an already active PDP context, but with a different QoS profile. A major part of the Secondary PDP Context as shown in Figure 3.3 is the TFT. The TFT contains attributes that specify an IP header filter that is used to direct data packets received from the interconnected external packet data network to the newly activated PDP context. The Secondary PDP Context Activation procedure may be executed without providing a TFT to the newly activated PDP context if all other active PDP contexts for this PDP address already have an associated TFT, otherwise a TFT shall be provided.[1]

The Secondary PDP Context Activation procedure is similar with the first PDP Context Activation procedure, as shown in Figure 3.3. The only difference is that UE sends its TFT along with other parameters.

**The PDP Context Modification Procedure**

A PDP Context Modification Procedure can be initiated by a UE, SGSN or GGSN to change the QoS or TFT negotiated during PDP context activation procedure. A GGSN initiated Modification procedure is depicted in Figure 3.4 for illustration. The discussion of this procedure is beyond our investigation scope.
CHAPTER 3. UMTS QoS

Figure 3.3: Secondary PDP Context Activation Procedure for UMTS [11]

Figure 3.4: GGSN-Initiated PDP Context Modification Procedure for UMTS [11]
3.4  UMTS QoS management entities and mechanism

The UMTS QoS management entities and mechanism can be described in the following ways:

3.4.1  QoS management in the control plane

Figure 3.5 depicts the major steps of QoS management in the control plane. These steps are explained below.

1. UMTS BS manager in the UE expresses the QoS profile of the user (UMTS bearer service attributes) to its counterpart in the SGSN through the PDP Context mechanism (Activate PDP Context Request or Activate Secondary PDP Context Request).

2. The UMTS BS manager in the SGSN verifies the subscription data of the user in order to verify if the user has the right to request such a service.

3. The UMTS BS manager of the SGSN makes an admission control in order to check if it has enough resources for the requested service.

4. The UMTS BS manager of the SGSN translates the UMTS bearer service attributes into RAB service attributes, Iu bearer service attributes and core network bearer service attributes. It then requests its Iu BS manager, its CN BS manager and the RAB manager in the UTRAN to provide the required services.

5. The RAB manager verifies with its admission/capability control whether the UTRAN supports the specific requested service and whether the required resources are available.
6. The RAB manager translates the RAB service attributes into radio bearer service and Iu bearer service attributes and requests the radio BS manager and the Iu BS manager to provide bearer services with the required attributes.

7. Once the radio access bearer is successfully established, the RAB manager informs the UMTS BS manager in the SGSN.

8. The UMTS BS manager in the SGSN speaks with its counterpart in the GGSN in order to request the QoS requirements of the user (Create PDP Context Request).

9. The GGSN UMTS BS manager makes an admission control to verify if it has enough resources to handle the new traffic.

10. The GGSN UMTS BS manager translates the UMTS bearer service attributes into core network bearer service attributes and requests its core network BS manager to provide the service.

### 3.4.2 QoS management in the user plane

The major steps of QoS management in the user plane for the downlink direction is depicted in Figure 3.6. These steps are as follows:

1. The GGSN classifies the incoming traffic in order to find out the appropriate PDP context. The classification entity tries to find out a matching TFT for the traffic.

2. The traffic passes a conditioner entity whose role is to verify if the traffic is conforming with the user contract.

3. The traffic passes the mapper function in order to be marked with the network specific QoS indication. In case of IP based network service in the core network
and the Iu interface, this function simply marks the DS bits of the packet according to the local Differentiated Services policy used in these links.

4. The packet passes through the resource manager. This entity decides when the packet can be transmitted to the next node.

### 3.5 Conclusion

Through the overview of the UMTS QoS specifications, we can conclude that the 3GPP has supplied a complete group of standards for the UMTS services and their attributes on different layers of bearer services, as well as the functions/entities and the signalling procedures to achieve these attributes, but how to achieve them was not specified. This is exactly the problem we called "UMTS QoS provisioning". Till this chapter it can be sum up in two scenarios:

- **QoS provisioning on control plane:**
  This function relates to resource reservation protocols at call set-up (Call Admission Control). Call Admission control determines if the call can be accepted by the network while guaranteeing the requested QoS defined by attributes in Table 3.2 and for a given duration. UMTS CAC is based on the concept of equivalent bandwidth. Call Admission Control (CAC) is performed along the path between both terminals throughout the network. It is also performed at lower layers; an automatic mapping function translates representations of QoS between different layers. If the resources are available, then are reserved otherwise a lower QoS is proposed.

  Key words: CAC, Mapping.

- **QoS provisioning on user plane:**
  They act on traffic flow according to requested levels of QoS. As described in the user plane management for downlink, the mechanisms include

  - GGSN specifically: classifier, conditioner, mapper/marker.
  - GGSN, SGSN, UTRAN: Traffic differentiation is performed using different queues or buffers to maintain QoS, using classical IP Differentiated services mechanisms or MPLS/ATM (Multi Protocol Label Switching). This allows flow scheduling, shaping (regulation), policing (control that the user doesn’t exceeds QoS).
4.1 DiffServ approach

DiffServ was designed as a simple architecture of QoS which can provide E2E services across different operator’s domain without using signalling process for every connection. DiffServ does not need extra signalling at the IP level. All necessary QoS is already contains in UMTS specific signalling message (Radio access Network Application Part, GPRS tunneling Protocol Control Plane). The way of DiffServ to achieve this goal is to use Behavior Aggregates (BA) flows and Per-Hop-Behaviors (PHBs) [2]. Individual connection QoS guarantee in DiffServ is not possible due to its aggregation traffic control mechanism[4].

4.1.1 Per Hop Behaviors(PHB) and Differentiated Services Code Point(DSCP)

The DiffServ architecture is used to classify and tag packets into a small number of classes at the edge of the network and to deploy mechanisms inside the network to treat various classes of packet differently. The Type Of Service (TOS) field in IPv4 can be used to define different serving manner for different classes of packet flows, that is, PHB. The TOS field is called as Differentiated Services field (DS). The information related to PHB is carried within DS field and called DiffServ Code Point (DSCP), as shown in Figure 4.1. The different PHB classes within DS field and its associated packets can get a preferential treatment. There are a lot of mechanisms such as queueing, scheduling, and policing to implement PHB, which are described in following two subsections. A DiffServ domain has three main components such as:

- Bandwidth Broker (BB): The resource management, inter-domain negotiation, admission control and dynamic configuration of concerned network elements to support the provisioned QoS services can be controlled in DiffServ using BB within each DS- domain. The BB in the DiffServ network is responsible for the implementation of admission control and resource allocation. The agreement between the user
4.1. DIFFSERV APPROACH

<table>
<thead>
<tr>
<th>DSCP</th>
<th>CU</th>
</tr>
</thead>
<tbody>
<tr>
<td>DSCP (DiffServ Code Point) – 6 bits. CU (Currently Unused) – 2 bits</td>
<td></td>
</tr>
</tbody>
</table>

Figure 4.1: The DS field structure

and the service provider that pertain the type of service required are known as Service Layer Agreement (SLA). The SLAs are used to define the relation between polices and the PHBs, while a service provisioning policy (SPP) indicates the state of traffic conditioners configured at the edge router and the traffic streams mapped to the DS behavior Aggregates. The BB needs both the SLAs and the SPPs in order to achieve a range of services, which are provided to the user. Based on the SLAs the BB decides whether it can provide the allocation and accordingly configures mark and classify the packets at the edge router as decided in the SLA. In the provision of end-to-end QoS, BB negotiates with its neighboring networks, set up bilateral service layer agreement and manages the adequate intra-domain resource allocation.

- **Edge Router:** The edge router are called Boundary nodes that provide a collection of function needed to interconnect a DS-domain to another DS-domain or to a non-DS-capable domain as shown in Figure 4.2. The Edge routers are divided into two categories

  - **Ingress Edge Router:** It handles the incoming traffic streams to a DS domain from an Access Network. The packets arriving at the ingress routers are marked with the corresponding DiffServ code point (DSCP) and transmitted through the core routers. The packets associated with a certain traffic class would get the same DSCP which will be put together in a behavior aggregate (BA) with a specific PHB. PHB is nothing but a dropping precedence. It handles Traffic conditioning Profile and ensure a user traffic adhere to its policy specifications by shaping and policing traffic.

  - **Egress Edge Router:** It handles the outgoing streams from a DS-domain. Egress Edge Router receives packets through the core network.

- **Core Router:** A core router examine incoming packets for the code point marking done on the packet by ingress edge routers and forward them accordingly. Core routers provide a reaction to the marking done by edge router. Core routers have few input links and their service rates are higher. The core router performs Scheduling and buffering for incoming packets from ingress edge router. The buffer size in core router should be lower.
4.1.2 Per Hop Behaviors

There are two well know PHBs: Assured Forwarding (AF) and Expedited Forwarding (EF).

Assured Forwarding (AF)

The AF-PHB group provides N independent PHB classes, each with M dropping precedence level as shown in Figure 4.3. There are N=4 classes and each classes maintain M=3 drop precedence levels as defined in the current specification [2].

Each class must be forwarded independently and different AF classes cannot be aggregated. Each class is configured with separate buffer and bandwidth. One possible approach is to assign Four PHB classes with Four physical queues and 3 virtual queues. Each physical queue will have three drop precedences. Packets within each class are marked with three different dropping precedence. During congestion, highest drop precedence packet should be dropped first. Random Early Detection (RED) can be used to avoid abrupt change in dropping precedence. Each DS node must allocate a certain amount of buffer space and bandwidth to each AF class.

With the AF PHB, user will get reliable services from their service providers, even in time of network congestion. The Service Layer Agreement (SLA) between the user and the Internet Service Provider (ISP) will specify the amount of bandwidth allocated. It is the users responsibility to decide how their applications share total amount of bandwidth.
4.1. DIFFSERV APPROACH

Expeditied Forwarding (EF)

The EF-PHB group was designed for users with fixed peak rate traffic to supply low-loss, low-latency and low-jitter end-to-end services as shown in Figure 4.4. Packet loss, latency, and jitter in wired networks are mainly due to that traffics have to wait in queue while transmitted through networks. Therefore, providing low loss, latency, and jitter for some high priority traffics can be achieved by guaranteeing very small queue.

Queues arise when (short-term) traffic arrival rate exceeds departure rate at some node. Thus, for an EF-PHB traffic stream, the departure rate must equal or exceed a configurable rate. EF traffic should average at least the configured rate when measured over any time interval equal to or longer than a packet time at the configured rate. The SLA specifies a peak rate for a flow or an aggregation of flows. The user is responsible for not exceeding the peak rate; otherwise, excess traffic will be dropped.[1]
4.1.3 Traffic classification and conditioning

Traffic classification and conditioning are the key functions for implementing of a Diff-Serv model as depicted in Figure 4.5. The traffic conditioner contain four logical modules: [2]. The traffic conditioner may not necessarily contain all four element. In case of no effective traffic profile presence, packets may only pass through a classifier and a marker.

![Logical structure of a packet classifier and traffic conditioner](image)

**Classifier** Packet classifiers select packets in a traffic stream based on the content of some portion of the packet header. Two types of classifiers are defined. The Behavior Aggregate (BA) classifier classifies packets based on the DS code point only. The Multi Field (MF) classifier selects packets based on the value of a combination of one or more header fields, such as source address, destination address, DS field, protocol ID, source port, destination port numbers and other information such as incoming interface.

**Meter** Traffic meters measure the temporal properties of the stream of packets selected by a classifier according to a traffic profile specified in a TCA. A meter passes state information to other conditioning functions to trigger a particular action for each packet, which is either in or out of profile (to some extent).

**Marker** Packet markers set the DS field of a packet to a particular code-point, adding the marked packet to a particular DS behavior aggregate. The marker may be configured to mark all packets which are steered to it to a single code-point, or may be configured to mark a packet to one of a set of code-points used to select a PHB in a PHB group, according to the state of a meter. When the marker changes the code-point in a packet it is said to have re-marked the packet.

**Shapper/Dropper** Shapers delay some or all of the packets in a traffic stream in order to bring the stream into compliance with a traffic profile. A shaper usually has a finite-size buffer, and packets may be discarded if there is not sufficient buffer space to hold the delayed packets. Droppers discard some or all of the packets in a traffic stream in order to bring the stream into compliance with its traffic profile. This process is known as policing the stream. Note that a dropper can be implemented as a special case of a shaper by setting the shaper buffer size to zero for a few packets.
4.1.4 Meter and Marker functions - Policy

A policy determines the treatment that a traffic aggregate will receive at the edge device. Edge devices use policy information to determine with what code point to mark packets. Each policy defines a policy type, a target rate, and other policy-specific parameters. As a minimum, each policy defines two code points; and the choice of code point depends on a comparison between the aggregate's target rate and current sending rate [14]. Each traffic aggregate has an associated policy type, meter type, and initial code point. The meter type specifies the method for measuring the state variables needed by the policy. Some well-known meters/markers are explained below.

**Time Sliding Window Three/Two Color Marker (TSW3CM/TSM2CM)**

The TSW3CM has been primarily designed for traffic streams that will be forwarded based on the AF PHB in routers [16]. The TSWTCM meters a traffic stream and marks packets to be either green, yellow or red based on the measured throughput relative to two specified rates: Committed Information Rate (CIR) and Peak Information Rate (PIR).

- The packets contributing to sending rate below or equal to the CIR with be marked with green color.
- Packets contributing to the portion of the rate between the CIR and PIR are marked yellow
- packets causing the rate to exceed PIR are marked with red.

TSW2CM: The marking takes place in the traffic conditioner of the boundary nodes of the DiffServ Domain. A special case of a time sliding window three color marker (TSW3CM) is used, to mark packets with only one of two colors. This is called a time sliding window two color marker (TSW2CM). Time Sliding Window Two Color Marker (TSW2CM) only use only CIR to evaluate traffic stream rate and mark them with either green or red. A CIR is defined for each edge router. As long as the connections rate is below CIR, all packets are marked as high priority. When the rate exceeds CIR, packets are marked probabilistically such that at the average, the rate of packet marked with high priority corresponds to CIR. It serves two purposes:

- Measuring the data rate of the ingress traffic at the boundary node.
- Marking the packets based on the measurement results. TSW2CM requires a CIR to be specified in units of bits/second.

The TSW3/2CM consists of two independent components: a rate estimator, and a marker to associate a colour (drop precedence) with each packet. If the marker is used with the AF PHB, each color would correspond to a level of drop precedence.

**Meter** A TSW meter working procedure is shown in Figure 4.6:
Initially:

\[
\begin{align*}
\text{AVG\_INTERVAL} & = \text{a constant}; \\
\text{avg-rate} & = \text{CTR}; \\
\text{t-front} & = 0;
\end{align*}
\]

Upon each packet's arrival, the rate estimator updates its variables:

\[
\begin{align*}
\text{Bytes\_in\_win} & = \text{avg-rate} \times \text{AVG\_INTERVAL}; \\
\text{New\_bytes} & = \text{Bytes\_in\_win} + \text{pkt\_size}; \\
\text{avg-rate} & = \frac{\text{New\_bytes}}{\text{now} - \text{t-front} + \text{AVG\_INTERVAL}}; \\
\text{t-front} & = \text{now};
\end{align*}
\]

Where:

- \text{now} = \text{The time of the current packet arrival}
- \text{pkt\_size} = \text{The packet size in bytes of the arriving packet}
- \text{avg-rate} = \text{Measured Arrival Rate of traffic stream}
- \text{AVG\_INTERVAL} = \text{Time window over which history is kept}

**Figure 4.6:** An example TSW meter algorithm [16]

\[
\begin{align*}
\text{avg-rate} & = \text{Estimated Avg Sending Rate of Traffic Stream} \\
\text{if} & (\text{avg-rate} \leq \text{CTR}) \\
& \text{the packet is green;}
\text{else if} & (\text{avg-rate} \leq \text{PTR}) \text{ AND } (\text{avg-rate} > \text{CTR}) \\
& \text{(avg-rate - CTR)} \\
& \text{calculate P0} = \frac{\text{(avg-rate - PTR)}}{\text{avg-rate}} \\
& \text{with probability P0 the packet is yellow;}
& \text{with probability (1-P0) the packet is green;}
\text{else} & \text{(avg-rate - PTR)} \\
& \text{calculate P1} = \frac{\text{(PTR - CTR)}}{\text{avg-rate}} \\
& \text{calculate P2} = \frac{\text{(PTR - CTR)}}{\text{avg-rate}} \\
& \text{with probability P1 the packet is red;}
& \text{with probability P2 the packet is yellow;}
& \text{with probability (1-(P1+P2)) the packet is green;}
\end{align*}
\]

**Figure 4.7:** TSW3CM Marking Algorithm [16]
Marker  A TSW3CM marker works as depicted in figure 4.7:
Token Bucket Algorithm

The operation of the Token Bucket in DiffServ can be defined as follow (as depicted in Figure 4.8):

- Arriving packets of L bytes are conforming (immediately processed) if there are at least L tokens in the bucket (one token = one byte).

- If the current number of accumulated tokens \( b' \) is less than the arriving number of packets \( L \), \( L - b' \) packets are nonconforming.

- Packets are allowed to the average rate in bursts up to the burst size, as long as they do not exceed the peak rate, at which point the bucket is drained.

- If there are no packets to be transmitted, tokens can be accumulated up to size of \( b \). The rest of tokens will be thrown away.

![Figure 4.8: Token Bucket algorithm](image-url)

The conforming and non-conforming functions are depends on SLA agreement. If the packets are non-conforming the following actions can be taken.

- The packet may be thrown away

- The packet may be re-marked in a particular way

- The packet can be buffered (by inserting a buffer between the inflow and the decision point) and not released until sufficient number of tokens arrive in the bucket.
4.2 IntServ/RSVP Policy

Integrated Service/Resource Reservation Protocol (IntServ/RSVP) is a flow-oriented QoS approach as depicted in Figure 4.9. This model offers three classes of service:

- **Guaranteed service (GS)** provides a deterministic upper bound on the delay and loss for a given flow. We have bandwidth guarantees for the traffic, and the applications can improve response time by increasing their demands for bandwidth. It is suitable for audio and video applications with strict delay requirements.

- **Controlled-load service (CL)** provides a service equivalent to a best-effort service in an unloaded or lightly loaded network. Though, this service does not provide a deterministic guarantee concerning the end-to-end delay for a given flow. This service class is designed for adaptive real-time applications (e.g. applications that can modify their play-out buffer as the end-to-end delay varies).

- **Best-effort service (BE)** is the fall-back solution providing a level of service similar to the current Internet service, without any bounds on the delay, loss, and jitter.

![Figure 4.9: IntServ/RSVP Policy](image)

Within this model, there is an inescapable requirement for routers to be able to reserve resources in order to provide special QoS for specific user packet streams or flows. This in turn requires flow-specific state in the routers.

IntServ relies on RSVP to reserve resources for the applications. The resources are reserved for every flow requiring QoS at every router hop in the path between a receiver and a transmitter.

Usually, a flow is characterized by a source and destination IP address, and a source and destination port number. The mechanism is the following: the sender emits a PATH message. This message specifies the characteristics of the traffic, and routers between
the sender and the receiver have to forward the PATH message, according to the routing protocol currently used within the network. When receiving a PATH message, the receiver has to send a RESV message to establish resource reservation for this individual flow, so it is a receiver driven mechanism.

When the RESV message flows back to the sender, intermediate routers can accept or reject the reservation. In the case the reservation is rejected by a router, this router has to send an error message to the receiver. Otherwise, the router allocates resources for the flow, according to the reservation, and forwards the RESV message to the next router. The PATH and RESV messages are sent periodically to refresh the reservations, so this is a soft state mechanism.

The weaknesses of the IntServ/RSVP model are:

- The amount of state information is proportional to the number of flows. We can deal with hundred thousands of reservations within the core of the IP network, and this requires processing load and huge resources requirements in the core routers. This approach does not scale well with the number of flows and makes routers very complicated.

- This architecture does not fit well with short-lived flows like web traffic, because the overhead in term of reservation is greater than the processing of all the packets in the flow.

- RSVP is a soft-state mechanism that maintains state information for all flows. This requires a large storage capacity.

- Malicious users could set the timer values in PATH messages to a very high value so that the corresponding state does not time out for a long period and takes up memory resources in network nodes.

### 4.3 Comparison between IntServ and DiffServ

The IntServ and DiffServ has its own advantages and disadvantages as listed in Table 4.1. IntServ supplies a strict E2E QoS guarantee. It requires a complex signalling process among different operators network involved in a connection, like the circuit-switched networks. DiffServ supply a much more flexible solution to only carry QoS related requirements within DSCP field of each IP packet. This maybe always not be sufficient to supply an absolute E2E QoS. Hence, DiffServ approach is used for both external IP network and UMTS IP backbone to investigate End-to-End QoS provisioning.
### 4.3. COMPARISON BETWEEN INTSERV AND DIFFSERV

<table>
<thead>
<tr>
<th>Feature</th>
<th>Integrated services</th>
<th>Differentiated services</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Granularity of service differentiation</strong></td>
<td>Individual flow</td>
<td>Aggregate of flows</td>
</tr>
<tr>
<td><strong>State in routers (e.g., scheduling buffer management)</strong></td>
<td>Per-flow</td>
<td>Per-aggregate</td>
</tr>
<tr>
<td><strong>Traffic classification basis</strong></td>
<td>Several header fields</td>
<td>The DS field (6 bits) of the IP header</td>
</tr>
<tr>
<td><strong>Type of service differentiation</strong></td>
<td>Deterministic or statistical guarantees</td>
<td>Absolute or relative assurances</td>
</tr>
<tr>
<td><strong>Admission control</strong></td>
<td>Required</td>
<td>Required for absolute differentiation only</td>
</tr>
<tr>
<td><strong>Signaling protocol</strong></td>
<td>Required (RSVP)</td>
<td>Not required for relative schemes; absolute schemes need semi-static reservations or broker agents</td>
</tr>
<tr>
<td><strong>Coordination for service differentiation</strong></td>
<td>End-to-end</td>
<td>Local (per-hop)</td>
</tr>
<tr>
<td><strong>Scope of service differentiation</strong></td>
<td>A unicast or multicast path</td>
<td>Anywhere in a network or in specific path</td>
</tr>
<tr>
<td><strong>Scalability</strong></td>
<td>Limited by the number of flows</td>
<td>Limited by the number of classes of service</td>
</tr>
<tr>
<td><strong>Network accounting</strong></td>
<td>Based on flow characteristics and QoS requirement</td>
<td>Based on class usage</td>
</tr>
<tr>
<td><strong>Network management</strong></td>
<td>Similar to circuit-switched networks</td>
<td>Similar to existing IP networks</td>
</tr>
<tr>
<td><strong>Inter-domain deployment</strong></td>
<td>Multilateral agreements</td>
<td>Bilateral agreements</td>
</tr>
</tbody>
</table>

Table 4.1: A comparison of the IntServ and DiffServ architectures
CHAPTER 4. IP QOS

4.4 Other IP QoS provisioning mechanisms

Scheduling and buffering are common mechanisms for IP packets handling and inevitable to implement QoS differentiation between User Equipments.

4.4.1 Scheduling

When multiple queues are sharing the common transmission media, there must be a scheduler to decide how to pick up packets from each queue to send out. The simplest scheduler is Round Robin (RR), which just select packet from each queue circularly, and all the queues have exactly equal chances to send.

Generally a packet scheduler is expected to have the following properties:

- Has low time complexity to select and forward a packet;
- Treat different flow fairly;
- Provide low worst-case delay and delay variation;
- Is simple enough to implement efficiently

The simplicity and time-complexity properties always collides with fairness and delay-bound properties.Schedulers with short term fairness and strict delay bound generally have high time complexity and are hard to implement.

A more useful scheduler for queuing with different classes is Weighted round robin (WRR). It is an extension of RR, which allow the bandwidth between queues to be distributed according to weights. Each time a non-empty queue is examined, it may send as many packets as its weight indicates. The WRR carry several improvements. The most commons are:

- To use bytes, or chunks of bytes, instead of packets. this gives greater granularity on the resource to be distributed.
- Allow queues to be selected several times within short intervals, instead of being selected many times in series. Since the sum of all weights can be fairly big, the improvement experienced can be dramatic. this improvement lowers the burntness of outgoing flows and lowers the average latency. It does however increase the implementation complexity. In WRR, each queue has bandwidth allocated when all other queues have been visited according to their weight.

In a long term, the percentage of each queue to be served is:

\[ P_i = \frac{W_i}{W_{total}} \]
4.4. OTHER IP QOS PROVISIONING MECHANISMS

$W_i$ is the weight of queue $i$; $W_{total}$: the sum of all queue weights.
There could be two implementation methods of WRR: a standard WRR and a Weighted Interleaved round robin (WIRR) implementation. Assume $W_i$ is integer and $W_{total}$ is the total number of packets the scheduler served in one circle, WRR keep serving packets for the selected queue until $W_i$ packets have been sent or the queue is empty. On the other hand, the WIRR works as in Figure 4.10 Another type of scheduler is Priority(PRI). In

begin
for each queue(i), temporaryWeight(i) = queueWeight(i)
then repeat
if queue(i) is not empty or temporaryWeight(i) > 0;
pick 1 packet from each queue in turn;
temporaryWeight(i) = temporaryWeight(i) - 1;
else
skip it, go to pick packet from the next queue;
until all temporaryWeight(i) = 0;
goto begin.

Figure 4.10: WIRR scheduling

PRI the queue with highest priority will always be served first until it’s empty, then the second highest priority queue will be served, and the lowest priority queue can only be picked when all the other queues have been empty.

4.4.2 Queuing - RED

RED is a congestion avoidance algorithm that can be implemented in routers. The simplest queue algorithm for routers is Drop Tail. When there is sufficient buffer space Drop Tail queues accept any incoming packet, and when buffer is full it simply drop any new packet that arrives. [14]

Instead, a RED queue computes a weighted average queue size to detect a congestion before the queue becomes really full, because a sustained long queue is a sign of network congestion. As any packet arrives, a RED gateway checks the weighted average queue size and compare with its minimum and maximum thresholds. If there is congestion, it either drops a packet or sets a bit in a header field of the packet according to a certain probability.
There are three phases for a RED gateways which drops packets:

**Normal operation:** If the average queue size is less than the minimum threshold, no packets are dropped.

**Congestion avoidance:** If the average queue size is between the minimum and maximum thresholds, packets are dropped with a certain probability. This probability is a function of the average queue size, so that larger queues lead to higher drop probabilities.

**Congestion Control:** If the average queue size is greater than the maximum threshold, all incoming packets are dropped.

The effects of RED gateways in a network will be:

- Control the average queue size, and reduce the average queuing delay.
- Act as a low-pass filter, hence the burst traffic get easier to be dropped.
- Avoid synchronization of TCP connections. A TCP based traffic source will reduce its transmitting rate when some of its burst packets will be dropped which cause RTT time-out. But the Drop Tail queue will all the packets in congestion which cause all the TCP source reduce their transmitting rate at the same time and the total throughput suddenly becomes very low after the congestion period. RED drops packets randomly before the buffer get overflow, which prevents this synchronization problem, reduces traffic oscillation effect and hence increase the overall throughput.

### 4.5 Conclusion

In this chapter, the two main standard IP QoS provisioning approaches, DiffServ and InterServ, are described and compared, and some other possible mechanisms to implement IP QoS are also introduced.
Chapter 5

NS-2 simulator

5.1 Overview of NS-2

The Network Simulator version 2 (NS-2) is an event driven network simulator developed at UC Berkeley that simulates variety of IP networks. It implements network protocols such as TCP and UDP, traffic source behavior such as FTP, Telnet, Web, CBR and VBR, router queue management mechanism such as Drop Tail, RED and CBQ, routing algorithms such as Dijkstra, and more. NS also implements multicasting and some of the MAC layer protocols for LAN simulations. The NS project is now a part of the VINT project that develops tools for simulation results display, analysis and converters that convert network topologies generated by well-known generators to NS formats. Currently, NS (version 2) written in C++ and OTcl (Tcl script language with Object-oriented extensions developed at MIT) is available.

As shown in Figure 5.1, in a simplified user’s view, NS is Object-oriented Tcl (OTcl) script interpreter that has a simulation event scheduler and network component object

![Figure 5.1: Simplified User View of NS](image-url)
libraries, and network setup (plumbing) module libraries (actually, plumbing modules are implemented as member functions of the base simulator object). In other words, NS simulation program is written in OTcl script language. To setup and run a simulation network, a user should write an OTcl script that initiates an event scheduler, sets up the network topology using the network objects and the plumbing functions in the library, and tells traffic sources when to start and stop transmitting packets through the event scheduler. The term "plumbing" is used for a network setup, because setting up a network is plumbing possible data paths among network objects by setting the "neighbor" pointer of an object to the address of an appropriate object. When a user wants to make a new network object, he or she can easily make an object either by writing a new object or by making a compound object from the object library, and plumb the data path through the object. This may sound like complicated job, but the plumbing OTcl modules actually make the job very easy. The power of NS comes from this plumbing.

Another major component of NS beside network objects is the event scheduler. An event in NS is a packet ID that is unique for a packet with scheduled time and the pointer to an object that handles the event. In NS, an event scheduler keeps track of simulation time and fires all the events in the event queue scheduled for the current time by invoking appropriate network components, which usually are the ones who issued the events, and let them do the appropriate action associated with packet pointed by the event. Network components communicate with one another passing packets, however this does not consume actual simulation time. All the network components that need to spend some simulation time handling a packet (i.e. need a delay) use the event scheduler by issuing an event for the packet and waiting for the event to be fired to itself before doing further action handling the packet. For example, a network switch component that simulates a switch with 20 microseconds of switching delay issues an event for a packet to be switched to the scheduler as an event 20 microsecond later. The scheduler after 20 microsecond dequeues the event and res it to the switch component, which then passes the packet to an appropriate output link component. Another use of an event scheduler is timer. For example, TCP needs a timer to keep track of a packet transmission time out for retransmission (transmission of a packet with the same TCP packet number but different NS packet ID). Timers use event schedulers in a similar manner that delay does. The only difference is that timer measures a time value associated with a packet and does an appropriate action related to that packet after a certain time goes by, and does not simulate a delay.

NS is written not only in OTcl but in C++ also. For efficiency reason, NS separates the data path implementation from control path implementations. In order to reduce packet and event processing time (not simulation time), the event scheduler and the basic network component objects in the data path are written and compiled using C++. These compiled objects are made available to the OTcl interpreter through an OTcl linkage that creates a matching OTcl object for each of the C++ objects and makes the control functions and the configurable variables specified by the C++ object act as member functions and member
variables of the corresponding OTcl object. In this way, the controls of the C++ objects are given to OTcl. It is also possible to add member functions and variables to a C++ linked OTcl object. The objects in C++ that do not need to be controlled in a simulation or internally used by another object do not need to be linked to OTcl. Likewise, an object (not in the data path) can be entirely implemented in OTcl. Figure 2 shows an object hierarchy example in C++ and OTcl. One thing to note in the figure is that for C++ objects that have an OTcl linkage forming a hierarchy, there is a matching OTcl object hierarchy very similar to that of C++.

5.2 DiffServ module

The DiffServ module in NS-2 implemented the AF PHB with a modified Random Early Detection (RED) queue mechanism. This module was contribute by Nortel Networks. [14]

5.2.1 AF PHB Implementation with multiple RED queues

In the DiffServ approach QoS was provided by dividing traffic into different classes and marking each packet with a code point corresponding to its class, and scheduling packets accordingly. The Assured Forwarding mechanism in NS-2 can include at most 4 physical queues for 4 PHB classes, and 3 virtual queues in each physical one for the 3 drop precedences. Those drop precedences enable differential treatment of traffic within a single class.

Different RED parameters are used for the virtual queues, causing packets from one virtual queue to be dropped more frequently than packets from another. A packet with a lower drop precedence is given better treatment in times of congestion because it is assigned a code point that corresponds to a virtual queue with relatively lenient RED parameters.

For example, one code point might be used for assured traffic and another for best effort traffic. The assured packet virtual queue will have higher minimum and maximum thresholds than those of best effort queue, meaning that best effort packets will enter the congestion avoidance and congestion control phase prior to assured packets.

Core Router

Core routers only have the functionality of performing PHB on incoming packets. In NS-2, core routers maintain at most 4 physical queues in its output link, each of them can have no more than 3 virtual queues or precedences. An incoming packet is placed in a virtual queue within a physical queue based on the DSCP-value in the packet’s IP-header. A physical queue is implemented as a RED-queue and virtual queues within the same
physical queue all use the same buffer to temporarily store packets. However, all virtual queues can be configured with their own RED-parameters (minimum and maximum threshold and dropping probability).

**Edge Router**

Edge routers include the same PHB functionality as core routers, and in addition perform the following functions:

- Policing
- Metering
- Packet marking

### 5.2.2 Policer

The DiffServ module currently supports six policer types.

**TSW2CM (TSW2CMPolicer):** uses a CIR and two drop precedences. The lower precedence is used probabilistically when the CIR is exceeded.

**TSW3CM (TSW3CMPolicer):** uses a CIR, a PIR, and three drop precedences. The medium drop precedence is used probabilistically when the CIR is exceeded and the lowest drop precedence is used probabilistically when the PIR is exceeded, as explained in [16].

**Token Bucket (tokenBucketPolicer):** uses a CIR and a CBS and two drop precedences. An arriving packet is marked with the lower precedence if and only if it is larger than the token bucket.

**Single Rate Three Color Marker (srTCMPolicer):** uses a CIR, CBS, and an EBS to choose from three drop precedences, as explained in [17].

**Two Rate Three Color Marker (trTCMPolicer):** uses a CIR, CBS, PIR, and a PBS to choose from three drop precedences, as explained in [18].

These policers have been explained in the IP QoS chapter.

### 5.2.3 Scheduler

Scheduling modes supported in NS2 are:

- Round Robin (RR)
- Weighted Round Robin (WRR)
- Weighted Interleaved Round Robin (WIRR)
5.3 UTRAN extension

The Enhanced UMTS Radio Access Network extensions to NS-2 (version 2.26) was developed within the SEACORN project for Ericsson Telecommunication B.V.. Detailed information, and updates to code and documentation can be found at the following website: http://www.ti-wmc.nl/eurane.

5.3.1 Overview of the UMTS model

The realization of the UMTS environment in NS2 is shown in Figure 5.2.

![Figure 5.2: UMTS simulation realization in NS2](image)

The major components of the simulation protocol architecture between a UE and a fixed host in the Internet are illustrated in Figure 5.3. Protocols on the Iub, IuPS and Gn interfaces have been abstracted out and are instead modelled by setting adequate link delay for those interfaces.

5.3.2 UTRAN nodes and links

The Enhanced UMTS extensions for NS-2 comprise of an additional three nodes, namely the Radio Network Controller (RNC), Base-station (BS) and the User Equipment (UE), whose functionality allow for the support of the following transport channels:

- FACH
- RACH
The main functionality additions to NS-2 come in the form of the RLC Acknowledged Mode (AM), Unacknowledged Mode (UM) MAC-d/-c/sh support for RACH/FACH and DCH, and MAC-hs support for HS-DSCH, i.e. HSDPA.

5.3.3 RLC object

The main functionality additions to NS-2 is the RLC Acknowledged Mode (AM) and Unacknowledged Mode (UM), which is implemented at RNC and UE. The RLC entity AM is used for the Dedicated Channel (DCH) and the two Common channels (RACH and FACH) in acknowledged mode, correspondingly another RLC entity AM-hs is developed to support HS-DSCH. Unacknowledged mode is also supported for DCH and HS-DSCH by the subsets of AM and AM-hs, namely UM and UM-hs, respectively. Without indication, all channels have the default RLC modes under AM. It needs to be mentioned that in EURANE there is no maximum RLC PDU retransmission limitation, so an error packet will be re-sent until it is correctly received.

5.3.4 Error Model for wireless link

Standard NS-2 error model can be add to DCH and FACH/RACH channels, as shown in Figure 5.2.

5.3.5 EURANE limitations

Some limitation of EURANE may effect the simulation design:
5.3. UTRAN EXTENSION

1. Only 1 UMTS cell can be simulated, and there is no cell capacity limitation for this cell.

2. The wireless link is static, meaning once a DCH has been built, it cannot be released during the simulation.

3. All the DCHs can only be configured with the same parameters, such as down-link/uplink bandwidth, AM/UM mode, etc.
Chapter 6

Delimitation

Based on the investigation of the UMTS and the IP QoS concepts and their mechanisms, the E2E QoS provisioning issues mentioned in Section 1.3 can be identified as follows (taking Figure 6.1 as an layered service architecture model):

1. External IP network: to select IETF standard QoS architecture (DiffServ or Inter-Serv) for the UMTS network to interact with.

2. UMTS network domain:
   - Control plane:
     - QoS attributes mapping between IP QoS classes and UMTS QoS classes.
     - The UMTS Bear Service (BS) attributes mapped to Core Network (CN) BS and Radio Access Bearer (RAB). Furthermore to Radio BS and Iu BS.
     - Call Admission Control in every multiplexing point (GGSN, SGSN, etc) according to Equivalent Bandwidth (EB).
   - User Plane:
     - GGSN policing: check downlink flow and filter it according to TFT.
     - Scheduling and buffering mechanisms in CN transmission
     - QoS differentiation implementation in RRM

3. Other open issues: Application level End2End QoS provisioning, i.e., session Initiation protocol (SIP) signalling

Later in the chapter, several related research projects are reviewed and more specific problems are selected as the focus of our project based on the possibilities supplied by NS-2 simulation tool.

6.1 State-of-the-Art

The following projects or papers are founded relevant to the project.
6.1. STATE-OF-THE-ART

6.1.1 E2E framework


   In this article, the proposed framework, which is an extension to the IST CADE- NUS project, offers wired-wireless QoS translation, QoS control and management, and dynamic SLA policy-based QoS provisioning. Two new functional components, the CUE-SM and CUE-RM are added to the UMTS framework to make up the CADENUS-UMTS Extension (CUE), which enables the dynamic Service Layer Agreement(SLA). The QoS translation was done between UMTS and Diff- Serv Per Hop Behavior (PHB). Simulation as well testbed are mentioned but no experiment result was supplied.

2. "Support of IP QOS over UMTS Networks", M. Ricardo, J. Dias, G. Carneiro, J. Ruela. From ARROW project(2001-2002). A UMTS/InterServ(RSVP) framework was presented and the detailed RSVP and PDP Context signalling process are described, no simulation was mentioned.

6.1.2 QoS mapping: SAMU project

Advanced Services by Mastering UMTS (SAMU)
Duration: July 1999 – September 2002

SAMU project is a collaborative project approved by the French RNRT (Telecommunication Research National Network) and aims at improving the services offered to the users in
an UMTS mobile system. For that purpose, the following topics were under investigation:

1. Defining smart resource allocation strategies for a CDMA mobile system (RRM).
2. Defining an efficient way to map IP over UMTS, to provide the user with an end-to-end QoS (QoS).
3. Complementary to the two previous ones. It aims at using agent software to make the system autonomous (Agent).
4. Demonstrating the feasibility of IP services over UMTS for users in their cars.

http://samu.crm-paris.com/Edocuments.htm

The innovative work in QoS from SAMU includes UMTS/IPv4 QoS mapping architecture and UMTS link layer optimization for TCP. In its report "QoS Architecture, Deliverable SP2-D2", detailed mapping was described in different categories:

- Mapping from IP QoS Attributes to UMTS Bearer Service Attributes:
  - Mapping from Differentiated Services to UMTS QoS Classes
    Besides QoS classes translation, the parameters in each class are also considered. For example, the guaranteed bit rate and maximum SDU size of the UMTS QoS parameters can be mapped to the arrival rate and the maximum burst size in the traffic profile of the SLA in DiffServ.
6.1. STATE-OF-THE-ART

<table>
<thead>
<tr>
<th>DIFFSERV</th>
<th>UMTS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expedited Forwarding</td>
<td>Conversational Class</td>
</tr>
<tr>
<td>Assured Forwarding-Gold Class</td>
<td>Streaming Class</td>
</tr>
<tr>
<td>Assured Forwarding-Bronze Class</td>
<td>Interactive Class</td>
</tr>
<tr>
<td>Best Effort</td>
<td>Background Class</td>
</tr>
</tbody>
</table>

Table 6.1: Mapping of Differentiated Services Classes to UMTS QoS Classes

<table>
<thead>
<tr>
<th>INTSERV</th>
<th>UMTS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Guaranteed Service</td>
<td>Conversational Class</td>
</tr>
<tr>
<td>Controlled Load</td>
<td>Streaming Class-Interactive Class</td>
</tr>
<tr>
<td>Best Effort</td>
<td>Background Class</td>
</tr>
</tbody>
</table>

Table 6.2: Mapping of Integrated Services Classes to UMTS QoS Classes

- Mapping from Integrated Services to UMTS QoS Classes
  The authors associate the peak data rate and the token rate attributes of the Resource Reservation Protocol (RSVP) to the maximum bit rate and the guaranteed bit rate of the UMTS. The maximum packet size parameter of the RSVP is mapped to the maximum SDU size of the UMTS.

- Mapping from UMTS Bearer Service Attributes to Radio Access Bearer Service Attributes:
  - Mapping from Radio Access Bearer Service Attributes to Radio Bearer Service Attributes
  - Mapping from Radio Access Bearer Service Attributes to Iu Bearer Service Attributes

- Mapping from UMTS Bearer Service Attributes to CN Bearer Service Attributes

The mapping analysis in this project is really helpful for our study. But there is no E2E QoS specifically simulation result presented in their report or papers.

6.1.3 Call Admission Control (CAC)

- "End-to-End QoS of Multi-Class traffic in WCDMA and DiffServ Network.", Lei Xiao, Tung Chong Wong, and Yong Huat Chew, PIMRC 2004.

The authors designed a CAC algorithm with UMTS QoS classes in the DiffServ IP network, and then extend this scheme for E2E admission control from the UMTS to the external DiffServ IP networks. Simulations were implemented on Openet for uplink data stream. The proposed CAC take both EB and delay boundary into consideration for each new connection request.
6.1.4 Simulation: SEACORN project

Simulation of Enhanced UMTS Access and Core Networks (SEACORN)  
Duration: March, 2002 – March, 2004  
http://seacorn.ptinovacao.pt/  
The contribution of SEACORN is

- Development, and implementation of resource management algorithms enabling QoS provisioning and differentiation while optimizing resource efficiency, in a multi-service UMTS dynamic network simulator considering both up- and downlink communications;

- Development and implementation of resource management mechanisms for IP-based core networks in a dynamic network simulator.

And a system level simulator was developed based on Network Simulator 2 (NS-2). Related papers from SEACORN project include:

The paper described the implemented UMTS simulation environment and the possible simulation scenarios.

The paper simulate a RRM implementation with CAC in UTRAN part and Power Control schemes.

We build up our simulation based on the UMTS extension developed in SEACORN project.

6.1.5 Radio Resource Management

Some work has been proposed to guarantee QoS for multimedia traffic in wireless cellular networks [23, 24, 25, 26, 27, 28, 29, 30, 31]. Rappaport and Purzynski developed an analytical models for cellular mobile environment consisting of mixed platform types with different classes of channels and resource requirements. Prioritizing hand-off calls over ordinary one and incorporating quotas for each type of resources, various performance and measure like carried traffic, blocking and forced termination probabilities for each platform and call type are numerically computed from analytic model [28]. The other minimum resource requirement criteria provided by the users, Oliveria et al. [26] proposed a bandwidth reservation algorithm for guaranteeing QoS to multimedia traffic. For the real time traffic, the call is admitted only if the requested bandwidth can be reserved in the call.
originating cell and all its neighbors. For a non-real-time call, the requested bandwidth is reserved only in the originating cell. Although this scheme guarantees QoS, the main drawbacks are: 1. bandwidth is reserved redundantly since the user moves only to one of the six neighboring cells (Assuming hexagonal cell geometry) and 2. the stringent call admission procedure might not admit many real-time requests in a highly overload system. The carried traffic in a wireless network can be increased by the graceful degradation of some or all of the existing service in the system [30]. In that, there are two QoS parameters, namely, graceful degradation of service and guarantee of seamless service. With the help of the user-supplied loss profiles, bandwidth usage of the applications that can sustain loss is degraded in situations where user demands exceed the networks capacity to satisfy them. A new transport sub-layer is proposed to implement loss profile by selectively discarding data from special applications like a compressed video stream. Another proposal is to have a virtual connection tree approach [24] for call routing, call admission and resource allocation. In [24], a call admission control algorithm is proposed for QoS provisioning for multimedia traffic, based on an adaptive resource sharing policy among real-time and non-real-time traffic, where the former traffic class has preemptive priority over the latter. A simple analytic model based on Poisson call arrivals and departures and threshold rules for each class of traffic are proposed to capture various QoS parameters like call blocking and dropping probabilities and the probability for minimum bandwidth availability for non-real-time requests which share the available bandwidth equally among themselves. Das et al. [32] addressed QoS provisioning techniques in cellular networks for both real-time and non-real-time data-flows at the link layer and proposed schemes for issues like bandwidth compaction, channel reservation and degradation to satisfy QoS requirements for multimedia traffic.

6.1.6 Conclusion

Contributions and drawbacks of these related works:

1. Many E2E QoS frameworks were proposed but solid implementation and simulation output are still missing.

2. QoS attributes mapping are designed but not all implemented.

3. Some work more focused on the RRM implementation, some on CAC control, no core network QoS user plane mechanisms are investigated in depth.

4. None of the framework gives a clear picture about the impact of mixing real-time and non-realtime traffics on the QoS aspects.

6.2 Problem Delimitation

Based on the above discussion, it is obvious that very few research is taking in Core Network as the bottleneck. However, as the 3G evolution is making the radio link bandwidth higher and higher (via HSDPA, OFDM, etc) and the traffic between the Internet and
the wireless network is getting heavier, the GGSN capacity may becomes the bottleneck. Therefor we chose the following scenarios to investigate:

- Assume both external IP network and UMTS IP backbone support DiffServ approach.
- In UMTS network domain, we select to utilize three mechanisms to achieve QoS provisioning:
  - QoS attributes mapping between External IP QoS classes and UMTS QoS classes at GGSN.
  - Call Admission Control in GGSN, according to Equivalent Bandwidth (EB).
  - Policing, Scheduling and congestion control mechanisms in Core Network transmission.
Chapter 7

Preliminary Simulation

7.1 Simulation including UTRAN

The first preliminary simulation on E2E is based on best effort QoS strategy, meaning different type of traffics (in UMTS traffic classes definition) are treated equally. The simulation topology is depicted in Figure 7.1.

![Figure 7.1: E2E Network topology in using Best Effort strategy](image)

7.1.1 Simulation settings

Traffic model

The traffic model in this simulation is simple and to show the different behavior of RT and NRT traffics and how they will effect each other when aggregated in a common link. Generally, in an IP based network with various traffic source. The link congestion is due to the contemporary traffic overload caused by the independent random sending of these sources. In contemporary congestion period, the total sending rate of all sources will be greater than the link bandwidth (This is unlikely to happen in a long term, since the
network operator will make an estimation and keep transmission capacity redundancy to avoid this situation.), which is the simulated scenario.

To observe the effect of multiplexing varies traffic sources alone, we also excluded any congestion control mechanism and put all the traffic streams on UDP protocol. The realtime traffic is simulated by a const bit rate (CBR) source and the non-realtime traffics are simulated by two exponential on-off source, which can supply a bursty nature in this scenario, as listed in Table 7.1. The bottle-neck link was set on GGSN-SGSN link with bandwidth 1Mbps.

<table>
<thead>
<tr>
<th>Traffic Type</th>
<th>Traffic source</th>
<th>Protocol</th>
<th>Data Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Real-time</td>
<td>CBR</td>
<td>UDP</td>
<td>384kbps</td>
</tr>
<tr>
<td>Non-realtime</td>
<td>Exponential</td>
<td>UDP</td>
<td>300-400kbps(On-time)</td>
</tr>
</tbody>
</table>

Table 7.1: Traffic model

### 7.1.2 Air interface effects

The Simulation setting for air interface is listed in Table 7.2. The air interface effects on E2E simulation are explained as follow:

<table>
<thead>
<tr>
<th></th>
<th>Mobility Model</th>
<th>Error Model</th>
<th>DCH bandwidth per UE</th>
<th>Radio Link Mode</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No</td>
<td>No</td>
<td>2048kbps</td>
<td>AM</td>
</tr>
</tbody>
</table>

Table 7.2: Simulation settings

**RLC PDU error and effects** NS-2 standard error model is implemented in UMTS DCH to represent the RLC PDU loss due to wireless propagation, which is following a Uniform distribution with mean packet loss ratio that can be achieved given fast power control is used. But since the RLC layer is working on AM mode without maximum retransmission limitation, error PDU can be recovered, and only introduce extra E2E delay. In wired network, packet error in transmission is very small and can be neglected. If there is still packet loss, it suppose to be due to queue overflow.

**Error freed UTRAN delay** The air interface introduced delay uncertainty, so the delay RLC PDU level delay between RNC and UEs was measured in NS2-EURANE simulation, as depicted in Figure 7.2. The delay in IuB was set to 15ms, hence air interface delay is about $41ms$ (under all EURANE air interface default settings).
7.1. SIMULATION INCLUDING UTRAN

Figure 7.2: UTRAN delay simulated in NS2-EURANE
7.1.3 Simulation result and Conclusion

Figure 7.3 shows user throughput when the total traffic load is higher than the bandwidth in Core network (SGSN-GGSN). It can be observed that in Best Effort transmission, the real-time traffics will suffered from the mixing with non-realtime Traffics in link congestion. When the realtime packets meet the "burst" of non-realtime packets, both have to wait in queue, which introduces jitter when they are received. Jitter is not a problem for the non-realtime services but really a trouble for realtime services, especially for conversational service. Therefore, real-time and non-realtime should be handled separately to guarantee the QoS requirements of real-time services. This principle can be implemented in differentiated policing, queuing, scheduling technics.

Figure 7.3: Traffic throughput in a 10% overload scenario

7.2 Simulation on DiffServ wired network

The purpose of this part is to evaluate the impact of different IP network QoS provisioning mechanisms when they are implemented in UMTS core network, and their performance under different higher layer protocols and UMTS traffic classes.

It has been explained in chapter 4 that RED queuing can control the average queue size for TCP based traffics by random early dropping. And RED is also available in the DiffServ module of NS2 simulator. For UDP based traffics, RED should have no effects on them since UDP sender cannot interactive with packets dropping. But to be careful we did the following simulation to test.
We made a simple simulation, with exponential traffic and UDP protocol, to compare the queue size evolutions of both RED and DropTail queues. The output is shown in Figure 7.4 and 7.5.

![Figure 7.4: Droptail queue size evolution](image1)

![Figure 7.5: RED queue size evolution](image2)

Averaged queue size:
- DropTail: 14.4 IP packets
- RED: 14.8 IP packets

Therefore in the same simulation scenario the RED queue has a even worse performance than DropTail queue for UDP based traffic.

### 7.2.1 Simulation result and conclusion

The simulation output as well as analysis can be concluded as follow: RED queue is a very efficient congestion control mechanism for TCP based traffics, usually Interactive and Background classes such as HTTP and FTP applications; but it can be worse than DropTail queue for UDP based traffic types, mostly realtime traffics.
Chapter 8

Development of QoS algorithm

The development of E2E QoS algorithm is described in terms of following mechanisms:

1. Mapping
2. Admission Control
3. Policing
4. Buffering and scheduling

However, the investigated scenario and assumptions as well as optimization targets will be described before discussing the algorithm.

8.1 General assumptions and optimization target

The investigated end-to-end scenario for QoS algorithm development is depicted in Figure 9.1. There application servers on an external network and will send a downlink data streams to the UEs in one UMTS cell. The sent IP data packets will be handled by Edge routers, GGSN, SGSN and RNC in turn. Then segmented to RLC PDU in RNC and forwarded to NodeB. In NodeB, PDU will be transmitted via Dedicated Transport Channel (DCH) to UEs. In this scenario application servers and edge routers belong to internet or external IP network domain and GGSN, SGSN, RNC, NodeB belong to UMTS domain. Both external IP network and UMTS IP based core network support DiffServ.

Each server will send only one type of application data corresponding to one UMTS service class. Each UE can use multiple types of services and set up multiple connections simultaneously. The general traffic model are assumed as in Table 8.1, and session of the individual types of traffic arrive according to Poisson process.
8.1. GENERAL ASSUMPTIONS AND OPTIMIZATION TARGET

Both domains support DiffServ

Figure 8.1: Simplified End2End QoS investigate scenario

<table>
<thead>
<tr>
<th>Traffic Type</th>
<th>Application</th>
<th>RT/NRT</th>
<th>Protocol</th>
<th>Application Level traffic model</th>
<th>Holding time model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conversational</td>
<td>Voice</td>
<td>RT</td>
<td>UDP</td>
<td>EXP On/Off</td>
<td>Exponential</td>
</tr>
<tr>
<td>Streaming</td>
<td>Video streaming</td>
<td>RT</td>
<td>UDP</td>
<td>EXP On/Off</td>
<td>Exponential</td>
</tr>
<tr>
<td>Interactive</td>
<td>Web</td>
<td>NRT</td>
<td>TCP</td>
<td>Pareto On/Off</td>
<td>Log-Normal</td>
</tr>
<tr>
<td>Background</td>
<td>FTP</td>
<td>NRT</td>
<td>TCP</td>
<td>CBR</td>
<td>Pareto</td>
</tr>
</tbody>
</table>

Table 8.1: Traffic Model Assumptions
The bottle-neck of this investigation is assumed in the GGSN outlink (GGSN-SGSN), hence the design target of the QoS algorithm will be to improve the bottleneck link utilization (lower session blocking rate, higher link throughput) while keep the E2E QoS requirements of each UMTS service class as boundary conditions. The boundary conditions of each service type are:

<table>
<thead>
<tr>
<th>Service Type</th>
<th>Conversational</th>
<th>Streaming</th>
<th>Interactive</th>
<th>Background</th>
</tr>
</thead>
<tbody>
<tr>
<td>SDU loss rate</td>
<td>$&lt; 10^{-2}$</td>
<td>$&lt; 10^{-1}$</td>
<td>$&lt; 10^{-3}$</td>
<td>$&lt; 10^{-3}$</td>
</tr>
<tr>
<td>E2E delay</td>
<td>$&lt; 100ms$</td>
<td>$&lt; 250ms$</td>
<td>N/A</td>
<td>N/A</td>
</tr>
</tbody>
</table>

Table 8.2: UMTS QoS requirements of different service classes

The radio link is assumed to operate on AM mode (Ack mode in RLC layer), and maximum RLC PDU retransmission time is unlimited. This setting is due to the EURANE limitations. Hence all the error RLC PDU would have been recovered, and the End-to-End SDU error/loss will only introduced by IP packet dropping in bottleneck link queue. An example of the packet data flow in AM mode is given in Figure 8.2.
8.1. GENERAL ASSUMPTIONS AND OPTIMIZATION TARGET

Figure 8.2: Packet data flow in AM [13]
CHAPTER 8. DEVELOPMENT OF QOS ALGORITHM

8.2 Mapping from DiffServ PHB to UMTS QoS classes

A vertical mapping is defined between the UMTS QoS classes and IP DiffServ classes as described in Table 8.3.

<table>
<thead>
<tr>
<th>Application type</th>
<th>UMTS service class</th>
<th>DiffServ DSCP</th>
</tr>
</thead>
<tbody>
<tr>
<td>VoIP</td>
<td>Conversational</td>
<td>EF</td>
</tr>
<tr>
<td>Video streaming</td>
<td>Streaming</td>
<td>AF1X</td>
</tr>
<tr>
<td>Web(http)</td>
<td>Interactive</td>
<td>AF2X</td>
</tr>
<tr>
<td>FTP</td>
<td>Background</td>
<td>BE</td>
</tr>
</tbody>
</table>

Table 8.3: UMTS Vs DiffServ Mapping

EF, which is characterized by low delay, low jitter and low packet loss, emulates a "leased line". To achieve these features Expedited Forwarding (EF) PHB will be implemented in routers with high priority in scheduler and short or even no queue, and drop those packets in an arrival rate exceeding the specified maximum arrival rate in advance, to supply a nearly constant departure rate of each connection. Conversational class like VoIP matches this DiffServ class very well: source traffic rate from a voice call is highly predictable, no burst nature at all, highly delay sensitive but relative loss-rate insensitive.

Assured Forwarding (AF) PHB provides several independent forward classes for IP packets delivery, with different shares of bandwidth as well as different buffering and dropping settings. The AF PHB can also be used to implement a low latency service using an over provisioned AF class. Also in this scheme the maximum arrival rate of the traffic will be specified in advance. So a high priority AF can be mapped to Streaming class since it has higher delay constraint than the interactive class but less constraint than the conversational class. Interactive class with lower traffic handling priority may match a lower AF class, with no explicit delay boundary (or can be several seconds such as in Web browsing, while it’s several hundred ms for Conversational and Streaming classes).

Finally, the best effort is mapped to the background class of the UMTS network.

8.3 Call Admission Control

Generally speaking, call admission control (CAC) is performed in every multiplexing point involved (SGSN, GGSN, routers, etc) in a PDP context. CAC deals with two problems. Can the node accept this new call? Will the node be able to meet the QoS requirements of the new call and of the established calls? UMTS CAC is implemented using the simple and flexible concept of equivalent bandwidth (EB). The philosophy is to estimate the network resources required to provide the requested QoS and to determine whether
8.3. CALL ADMISSION CONTROL

these resources are available. If they are available, the necessary resources are reserved. However, if they are not available, a mechanism is initiated to downgrade the QoS.

In our investigation, to simplify we implement a centralized CAC, and only admit or reject new PDP context request according to the bandwidth limitation of the GGSN out-link capacity. That is, there is no downgrade re-negotiation for existing connections. There are several mathematical models for the estimation of both Exponential on-off traffic model for real-time services [19, 20] and Pareto traffic model for non-realtime services [22], considering transmission issues like packet loss ratio and buffer size in the routers. For example, in an estimation of Exponential On-Off traffic source without buffer is,

\[ EB = m + \alpha \sigma, with, \alpha = \sqrt{-2 \ln(\varepsilon) - \ln(2\pi)} \] (8.1)

where \( m \) is the mean rate of the aggregate sources, \( \sigma \) is the standard deviation and \( \varepsilon \) is the target packet loss ratio. With more sources aggregated, the traffic becomes more Gaussian distribution [19, 20]. Hence in our assumed scenario, given \( N \) sources of one class belongs to the identical exponential On-Off process(means same mean "On" and "Off" time duration, and same sending rate in "On" period), their total equivalent bandwidth is

\[ N \cdot EB = Nm + \alpha \frac{\sigma}{\sqrt{N}} \]

So the estimation of a single EB is

\[ EB = m + \alpha \frac{\sigma}{\sqrt{N}} \]

As \( N \) becomes bigger and bigger, the EB will finally converge to the mean rate, where

\[ m = r_{on} \cdot \frac{OnTime}{OnTime + OffTime}, \]

with \( r_{on} \): on time sending rate.

These mathematical models can be very tricky but basically equivalent bandwidth is an estimated value between the peak rate and the long term average sending rate of a traffic source. The reason to use EB is that if CAC is based on the peak rate of each traffic, the QoS will be absolutely supplied but the average bandwidth utilization will be low, while if reserving bandwidth according to average sending rate of each source, it may severely degrade the QoS of realtime services in congestion period. To balance between resource utilization and QoS guarantee, EB has to be estimated. Therefore, we will simulated with both peak rate and mean rate, and select a proper EB which could be either based on a math model or experiment result.
8.4 Policing, buffering and scheduling

Policing, buffering and scheduling mechanisms are performed each time a packet is sent and/or received (typically microseconds). Policing and buffer acceptance mechanisms are algorithms that decide when a packet arrives, whether or not it can be accepted. Scheduling algorithms decide when and which packet to send first. The full implementation structure of these mechanisms is shown in Figure 8.3.

![Figure 8.3: Policing, buffering and scheduling Implementation](image)

8.4.1 Policing and Multiple queuing

A policer at the edge router is the function to implement DSCP marking, usually it includes two code points (CP) ("red" or "green") for each traffic, which means can be IP packets with different color can be handled later with different physical/virtual queues and different treatment. Since the implemented traffic source in NS2 will not exceed its peak rate (CIR), the two color marking is not necessary. Hence we decide to only mark the packets of each type of traffic with one color, "green", and handle them with multiple queues.

Multiple queues have been designed to provide different treatments to different UMTS classes according to their DSCP field. Since RED queue only effect TCP based traffic streams and cannot help in congestion control of UDP based traffic, two drop-tail queues are designed for UDP based realtime services, Conversational (C) and Streaming (S), while RED queues are for TCP based non-realtime services, Interactive (I) and Background (B). The queue size will be optimized according to their different delay and packet loss requirements simultaneously.
8.4.2 Scheduling

In DiffServ module these schedulers are available:

1. Round Robin (RR)
2. Weighted Round Robin (WRR)
3. Weighted Interleaved Round Robin (WIRR)
4. Priority (PRI)

It is obvious that due to the different delay sensitivity levels, four UMTS services classes C, S, I and B should be assign priority levels from 1 to 4, in which 1 is the highest. RR don’t support priority, so it is out of consideration. For PRI and WRR/WIRR, as depicted in Figure 8.4 and 8.5.

**PRI** PRI serve the highest priority queue first, and once it is empty, PRI turn to serve the second priority queue, and so on. Assume conversational class is voice service, if

\[
VoiceLoad \leq BandWidth_{total}
\]

then the QoS of voice traffic is always guaranteed in a long term. While other traffic types (S, I, B) are all served in a Best Effort manner comparing to voice. And this would be only a best choice for traditional telephony networks.

**WRR** On the other hand, WRR can guarantee each classes given

\[
TrafficLoad(i) \leq \frac{W(i)}{W_{total}}
\]

with a relatively smaller BW for voice traffic. This scheduler serves each queue more fair and the relative importance of each classes can be modified easily according to the change of traffic mix, hence it is more flexible.
Either PRI or WRR/WIRR can be a good choice. The selection between two schedulers depends on the traffic model features. For example, as the mobile operators transfer from 2G to 3G technology, the predictable voice traffic will only maintain a slow or flat increase. Assume the transport network keeps the same capacity, only upgrade from circuit-switch to packet-switch, and the effect will be (shown in table) So, in UMTS 75% backbone transport bandwidth (BW) can be saved for other services. Considering the multiplexing gain of Packet Switch (PS) comparing to Circuit Switch (CS), the BW saving will be actually more. Since the traditional telephone service can at most occupy 25% backbone bandwidth, and UMTS air interface BW are much bigger than what voice need, while neither radio interface nor backbone can be sufficient for all Internet or multimedia services, it will be too conservative to keep voice or conversational services always on the absolute priority, as implemented in [22]. We have to keep in mind that most multimedia traffics are asymmetrical and in streaming class, they are very bandwidth consuming and to guarantee their QoS profile is also very important. In this sense, we consider WRR/WIRR will be more appropriate than PRI.

### 8.5 conclusion

The implemented QoS algorithm is depicted in 8.6.
8.5. CONCLUSION

Figure 8.6: Implemented QoS algorithm
Chapter 9

Enhanced simulation and evaluation

We will investigate the end-to-end scenario with simulation experiments in this chapter, where all the QoS provisioning mechanisms described in the previous chapter is applied. The optimization target in simulation is to achieve a better bottleneck link bandwidth utilization while keeping the packet loss and E2E delay within their boundary as shown in 9.1. So the Key Performance Indicators (KPIs) are:

<table>
<thead>
<tr>
<th>Service Type</th>
<th>Conversational</th>
<th>Streaming</th>
<th>Interactive</th>
<th>Background</th>
</tr>
</thead>
<tbody>
<tr>
<td>SDU loss rate</td>
<td>$&lt; 10^{-2}$</td>
<td>$&lt; 10^{-1}$</td>
<td>$&lt; 10^{-3}$</td>
<td>$&lt; 10^{-3}$</td>
</tr>
<tr>
<td>E2E delay</td>
<td>$&lt; 100ms$</td>
<td>$&lt; 250ms$</td>
<td>N/A</td>
<td>N/A</td>
</tr>
</tbody>
</table>

Table 9.1: UMTS QoS requirements of each service type
9.1 Simulation scenario

As shown in Figure 9.1, there are four application servers located in the Internet domain, and each one of them supplies one type of application: Voice over IP (VoIP), video streaming, web service via HTTP, or FTP download. The IP data packets are sent to the edge router, which is simplified as both the Ingress and Egress router of external DiffServ domain. The edge router marks IP packets with different DSCP according to their application type, and send them to GGSN. GGSN then should attach them with GPRS Tunneling Protocol (GTP) and encapsulated into the UMTS IP transport layer, which is simplified in the simulator. The GGSN outlink differentiate each IP data flow according there DSCP (GGSN doesn’t change their DSCP assigned from external network) and transmits them with the queuing and scheduling schemes described in the previous chapter. Then SGSN receives them and forward to RNC, where the IP header will be compressed into PDCP header, and a whole IP packet are converted into RLC SDU. Hence the RLC layer in RNC segments one RLC SDU into several RLC PDU and further mapped into DCH frames and sent on air interface.

As shown in Figure 9.2, the radio link settings are relatively simple since it is not consid-
error/loss will only introduced by IP packet dropping in bottleneck link queue, which is Early Drop for realtime traffic or queue overflow for all traffic. On the other hand the unlimited RLC PDU retransmission will result in much more uncertainty for the E2E delay, and make the delay control more difficult.

<table>
<thead>
<tr>
<th>Transport Layer Protocol</th>
<th>VolP</th>
<th>Video Streaming</th>
<th>Web/HTTP</th>
<th>FTP</th>
</tr>
</thead>
<tbody>
<tr>
<td>IP Packet size(byte)</td>
<td>UDP</td>
<td>UDP</td>
<td>TCP</td>
<td>TCP</td>
</tr>
<tr>
<td>Traffic source model</td>
<td>Exponential On/Off</td>
<td>Exponential On/Off</td>
<td>Pareto On/Off, shape=1.1</td>
<td>Pareto distributed file size, shape=1.1</td>
</tr>
<tr>
<td>Holding time distribution</td>
<td>Exponential</td>
<td>Exponential</td>
<td>Log-Normal, STD=2.5</td>
<td></td>
</tr>
<tr>
<td>Avg sending rate(on time)(kbps)</td>
<td>30.00</td>
<td>128.00</td>
<td>60.00</td>
<td>120.00</td>
</tr>
<tr>
<td>On-time (sec)</td>
<td>1.00</td>
<td>1.50</td>
<td>1.60</td>
<td>5.00</td>
</tr>
<tr>
<td>Off-time (sec)</td>
<td>1.50</td>
<td>1.50</td>
<td>12.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Session Inter-arrival time (sec)</td>
<td>5.50</td>
<td>44.00</td>
<td>64.00</td>
<td>50.00</td>
</tr>
<tr>
<td>holding time (sec)</td>
<td>120</td>
<td>300</td>
<td>300</td>
<td>5</td>
</tr>
<tr>
<td>bottle-neck BW(kbps)</td>
<td>1000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>traffic load (total average load/BW)(%)</td>
<td></td>
<td></td>
<td>74.32%</td>
<td></td>
</tr>
</tbody>
</table>

Figure 9.3: Traffic Parameters in Simulation

Based on the same network topology settings and traffic model, we investigated four End-to-End QoS provisioning methods:

1. Strict Best Effort: no AC and all types of services were simply equally handled.

2. CIR-based AC with designed QoS provisioning schemes. As introduced in previous chapters, UMTS AC is based on equivalent bandwidth (EB), where the AC admits coming sessions by comparing the available bandwidth with the EB of the service type of the coming session. In this scenario the EB is estimated as CIR, in another word, simply the on-time application layer sending rate in our traffic model. Other QoS schemes, policing, queueing and scheduling are the same as designed in the previous chapter.

3. Mean-based AC with designed QoS provisioning schemes. Here EB is estimated as the mean sending rate of each type of services. Other QoS schemes are the same as described in the previous chapter.

4. Proposed EB based AC with user differentiation. An enhanced EB estimations of each service type are combined with other designed QoS schemes. Both the value of EB and the settings in buffering and scheduling are optimized according to simulation results from the previous three scenarios.
Notes: In all the simulated scenario, the total buffer sizes on bottleneck link (GGSN outlink) are approximately the same.

9.2 Performance evaluation

9.2.1 Best Effort

As depicted in Figure 9.4, the 93% guaranteed E2E delay of VoIP packets is about 250ms, much greater than the 100ms boundary. This result shows that the best effort strategy caused an unacceptable QoS level for conversational service in a high traffic load.
9.2.2 CIR-based CAC with QoS schemes

The CIR(EB) parameters are set as in Table 9.2:

The result shown in Table 9.3 and Figure 9.5 and 9.6 shows that CIR-based AC with de-

<table>
<thead>
<tr>
<th>Application</th>
<th>VoIP</th>
<th>Video</th>
<th>WEB/HTTP</th>
<th>FTP</th>
</tr>
</thead>
<tbody>
<tr>
<td>CIR (kbps)</td>
<td>30</td>
<td>128</td>
<td>60</td>
<td>120</td>
</tr>
</tbody>
</table>

Table 9.2: CIR settings for CIR-based Admission Control

<table>
<thead>
<tr>
<th>UMTS CN(GGSN-SGSN)</th>
<th>VoIP</th>
<th>Video</th>
<th>WEB/HTTP</th>
<th>FTP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Session Blocking Ratio</td>
<td>31.4%</td>
<td>75.8%</td>
<td>54.1%</td>
<td>76.9%</td>
</tr>
<tr>
<td>Packet Loss</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Bandwidth Utilization</td>
<td>31.0%</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 9.3: Simulation result of CIR-based CAC with QoS algorithm

![Figure 9.5: VoIP Packet Delay Distribution in CIR-based QoS algorithm](image)

Figure 9.5: VoIP Packet Delay Distribution in CIR-based QoS algorithm

signed buffering, scheduling and policing schemes easily keep the UMTS QoS boundary but result in very low bottleneck link bandwidth utilization, only 31%, around half of the traffic load, and a extremely high blocking rate. Hence it is too conservative a strategy for our optimization target.
9.2. PERFORMANCE EVALUATION

Figure 9.6: Video Packet Delay Distribution in CIR-based QoS algorithm

9.2.3 Mean-based CAC with QoS schemes

The mean sending rate of each type of service are used as EB and listed in Table 9.4:

<table>
<thead>
<tr>
<th>Application</th>
<th>VoIP</th>
<th>Video</th>
<th>WEB/HTTP</th>
<th>FTP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean (kbps)</td>
<td>12</td>
<td>64</td>
<td>7</td>
<td>120</td>
</tr>
</tbody>
</table>

Table 9.4: Mean sending rate estimation for Mean-based Admission Control

<table>
<thead>
<tr>
<th>UMTS CN(GGSN-SGSN)</th>
<th>VoIP</th>
<th>Video</th>
<th>WEB/HTTP</th>
<th>FTP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Session Blocking Ratio</td>
<td>0.5%</td>
<td>8.8%</td>
<td>0%</td>
<td>0%</td>
</tr>
</tbody>
</table>
| Packet Loss         | $3.2 \times 10^{-2}$ | $1.0 \times 10^{-2}$ | $1.4 \times 10^{-2}$ | 0%
| Bandwidth Utilization | 69.8% |        |          |     |

Table 9.5: Simulation result of Mean-based CAC with algorithm

It’s an aggressive CAC that could achieve the bandwidth utilization of 69.8%, which means it gets a good multiplexing gain. However, to keep staying inside the delay boundary, packet loss ratio boundary of conversational class cannot be met, while the packet loss for interactive class is also out of boundary. So the optimization approach should be to reserve more resource for conversational class and interactive class in both AC and other QoS schemes.
Figure 9.7: VoIP Packet Delay Distribution in proposed QoS algorithm

Figure 9.8: Video Streaming Packet Delay Distribution in proposed QoS algorithm
9.2. PERFORMANCE EVALUATION

9.2.4 EB-based CAC with QoS schemes

In this simulation scenario we proposed a set of EB estimations, and combined it with optimized buffering and scheduling settings.

In the previous chapter we have deduced the estimation for EB of exponential traffic source model:

\[ EB = m + \alpha \frac{\sigma}{\sqrt{N}} \]

Since only EB for interactive class need to be enhanced, we considered \( N \) as the number of independent sessions coming during the simulation duration, then \( EB = 15 \text{kbps} \). But simulation shows its still a too aggressive value which cannot match the QoS requirements in high traffic load (74%) scenario. Besides the EBs for interactive and background services are also modified according to the conclusion from the previous section. The final setting is listed in 9.6.

<table>
<thead>
<tr>
<th>Application</th>
<th>VoIP</th>
<th>Video</th>
<th>WEB/HTTP</th>
<th>FTP</th>
</tr>
</thead>
<tbody>
<tr>
<td>EB setting (kbps)</td>
<td>24</td>
<td>64</td>
<td>10</td>
<td>7</td>
</tr>
<tr>
<td>Queue setting</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Queue type</td>
<td>DropTail</td>
<td>DropTail</td>
<td>RED</td>
<td>RED</td>
</tr>
<tr>
<td>Queue size (pkt)</td>
<td>8</td>
<td>30</td>
<td>120</td>
<td>200</td>
</tr>
<tr>
<td>RED Threshold(Min,Max)</td>
<td>N/A</td>
<td>N/A</td>
<td>60</td>
<td>120</td>
</tr>
<tr>
<td>Early Dropping probability</td>
<td>N/A</td>
<td>N/A</td>
<td>0.01</td>
<td>0.01</td>
</tr>
</tbody>
</table>

Table 9.6: QoS provisioning settings

<table>
<thead>
<tr>
<th>UMTS CN(GGSN-SGSN)</th>
<th>VoIP</th>
<th>Video</th>
<th>WEB/HTTP</th>
<th>FTP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Session Blocking Ratio</td>
<td>11.3%</td>
<td>34.6%</td>
<td>4.0%</td>
<td>0%</td>
</tr>
<tr>
<td>Packet Loss</td>
<td>0</td>
<td>(2.3 \times 10^{-4})</td>
<td>(1.5 \times 10^{-4})</td>
<td>0</td>
</tr>
<tr>
<td>Bandwidth Utilization</td>
<td>60.6%</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 9.7: Simulation result of proposed EB-CAC with QoS algorithm

The outputs are presented in Table 9.7, Figure 9.9 and 9.10, which verified that our enhanced QoS algorithm and settings satisfied the QoS for all the service types and achieved a high bandwidth utilization (60.6%), which is around .
Figure 9.9: VoIP Packet Delay Distribution in proposed QoS algorithm

Figure 9.10: Video Packet Delay Distribution in proposed QoS algorithm
9.3 Conclusion

In this chapter we investigated four different QoS strategies on the same simulator with the same simulation topology and traffic model, in which a high traffic load from the application server side is sent to the UMTS network. Firstly the Best Effort strategy failed to supply End-to-End QoS with an unacceptable large latency for conversational service. Then the designed QoS provisioning mechanisms and ACs with different EB estimation methods are compared: CIR-based estimation certainly supplied the required QoS profile on the price of a low bandwidth utilization; on the contrary, Mean-based estimation achieved a high bandwidth utilization but also failed in guarantee QoS of conversational service; finally, the optimized EB estimation based AC along with QoS provisioning schemes balanced the conflict between QoS and bandwidth utility, and achieved a good performances for all KPIs.
Chapter 10

Conclusion and future work

Conclusion  This Master mainly focus on investigating an optimized UMTS Quality of Service provisioning strategy in an End-to-End scenario to supply QoS guarantee as well as to improve system performance. For this purpose, the main features of UMTS network was studied first. Then the QoS provisioning approaches in both the UMTS domain and the Internat domain were investigated and the open issues were identified. A simulation platform was built up in selected play scenarios and with simplified traffic model and radio link assumptions. An enhanced QoS algorithm was optimized in compare with the Best Effort strategy as well as two other similar QoS schemes with different CAC settings. The simulation results of the proposed strategy was proved to be able to satisfy the UMTS realtime services QoS requirements and achieve a high system bandwidth utility.

To conclude the study throughout this Master project, user perceived QoS is always due to many QoS mechanisms involved in the whole End-to-End transmission chain. Thus the optimization is inevitable a multiple-dimensional process. And the traffic model from the users choices is the key issue to judge which types of QoS mechanisms can be the best strategy.

Future work  Our study of the UMTS QoS provisioning could be extended to cooperate with RRM schemes to enhance the overall performance, especially in case the HSDPA is implemented. It is also possible to improve it by developing a dynamic AC and scheduling, which can adapt to the network traffic load and application mix along with the time.
Bibliography


[8] 3G TS 25.401: "UTRAN Overall Description".


[12] 3GPP TS 24.008 version 6.7.0 (Release 6), "Mobile radio interface Layer 3 specification; Core network protocols; Stage 3", 2005


[22] Lei Xiao, Tung Chong Wong, and Yong Huat Chew, "End-To-End QoS of Multi-Class traffic in WCDMA and DiffServ Network.", PIMRC 2004.


