Contrôle de trafic et gestion de la qualité de service basée sur les mécanismes IP pour les réseaux LTE

William David Diego Maza

UNIVERSITE
BRETAGNE
LOIRE

THÈSE / Télécom Bretagne
sous le sceau de l'Université Bretagne Loire

pour obtenir le grade de Docteur de Télécom Bretagne
En accréditation conjointe avec l'Ecole Doctorale Matisse
Mention : Informatique

présentée par
William David Diego Maza
préparée dans le département Réseaux, Sécurité et Multimédia (RSM)
Laboratoire Irisa

Contrôle de trafic et gestion de la qualité de service basée sur les mécanismes IP pour les réseaux LTE

Thèse soutenue le 03 octobre 2016
Devant le jury composé de :

César Viho
Professeur, Université de Rennes 1 / président

Lila Boukhatem
Maître de conférences (HDR), Université Paris-Sud / Rapporteur

André-Luc Beylot
Professeur, INP-ENSEEIHT / Rapporteur

Tijani Chahed
Professeur, Télécom SudParis / Examinateur

Catherine Rosenberg
Professeur, University of Waterloo – Canada / examinatrice

Xavier Lagrange
Professeur, Télécom Bretagne / Directeur de thèse

N° d'ordre : 2016telb0406

Sous le sceau de l'Université Bretagne Loire

Télécom Bretagne

En accréditation conjointe avec l'Ecole Doctorale Matisse

Contrôle de trafic et gestion de la qualité de service basée sur les mécanismes IP pour les réseaux LTE

Thèse de Doctorat

Mention : Informatique

Présentée par **William David Diego Maza**

Département : Réseaux, Sécurité et Multimédia (RSM)

Laboratoire : Orange Labs

Directeur de thèse : Xavier Lagrange

Soutenue le 3 octobre 2016

Jury :

Mme. Lila Boukhatem, Maître de conférences (HDR), Université Paris-Sud (Rapporteur)
M. André-Luc Beylot, Professeur, INP-ENSEEIHT (Rapporteur)
M. Xavier Lagrange, Professeur, Télécom Bretagne (Directeur de thèse)
Mme. Catherine Rosenberg, Professeur, University of Waterloo (Examinateur)
M. Tijani Chahed, Professeur, Télécom SudParis (Examinateur)
M. César Viho, Professeur, Université de Rennes 1 (Examinateur)
Mme. Isabelle Hamchaoui, Ingénieur de Recherche Senior, Orange Labs (Encadrant)

Acknowledgements

I would like to thank my thesis supervisor, Dr. Isabelle Hamchaoui, for having given me the opportunity to join Orange Labs as PhD student and for her confidence and advices in my investigations. I am and will always be grateful to her for having been a friend, a mentor and a great source of support during these three years.

I would also like to thank my thesis Director, Professor Xavier Lagrange, from Télécom Bretagne for his invaluable help and guidance throughout my thesis

I also like to express my admiration for Dr. Fabrice Guillemin for his always available advices and useful suggestions.

I would like to thank the jury members, Dr. Lila Boukhatem, Prof. André-Luc Beylot, Prof. Catherine Rosenberg, Prof. Tijani Chahed, and Prof. Cesar Viho, for reviewing my dissertation and discussing it thoroughly.

I want to thank all my colleagues of iTEQ team, where I passed these three years with daily interaction with them, thanks for their help and for all those good moments spent together around a coffee. I would particularly like to thank Jean-Marc Corolleur for his support and encouragement during my thesis.

I would also like to thank my colleagues and friends Djamel, Lida, Edoardo and all other Orange's PhD students and interns, for their friendship and good times in Lannion. I want to especially thank Veronica for her great support, encouragement and motivation during the final stage of my PhD. I wish to thank also all the friends I made during my stay in Lannion. All of them made easier and enjoyable my stay in this lovely and peaceful region of France.

Last but not least, I want to thank my family for their support throughout the years. I cannot express enough gratitude and thanks to my parents, my sister and my nieces for their love and support to whom I dedicate these three years of work.

Abstract

The mobile data landscape is changing rapidly and mobile operators are today facing the daunting challenge of providing cheap and valuable services to ever more demanding customers. As a consequence, cost reduction is actively sought by operators as well as Quality of Service (QoS) preservation.

Current 3GPP standards for LTE/EPC networks offer a fine-tuning QoS (per-flow level), which inherits many characteristics of legacy telco networks. In spite of its good performance, such a QoS model reveals costly and cumbersome and finally, it remains very rarely deployed, thereby giving way to basic best-effort hegemony.

This thesis aims at improving QoS in mobile networks through cost-effective solutions; To this end, after an evaluation of the impact and cost of signaling associated with the standard QoS model, alternative schemes are proposed, such as the IP-centric QoS model (per aggregate) inspired from the DiffServ approach widely used in fixed IP networks. This model provides a simple, efficient and cost-effective IP level QoS management with a performance level similar to standardized solutions. However, as it requires enhancements in the eNB, this scheme cannot be expected in mobile networks before a rather long time.

Thus, we introduce Slo-Mo, which is a lightweight implicit mechanism for managing QoS from a distant point when the congestion point (e.g. eNB) is not able to do it. Slo-Mo creates a self-adaptive bottleneck which adjusts dynamically to the available resources taking advantage of TCP native flow control. Straightforward QoS management at IP level is then performed in the Slo-Mo node, leading to enhanced customer experience at a marginal cost and short term.

Keywords: *Quality of Service, Mobile Networks, Protocol Design, Performance Analysis, DiffServ, Active Queue Management, TCP protocol*

Acronyms

3GPP 3rd Generation Partnership Project xv
ACK ACKnowledgment 102
ADSL Asymmetric Digital Subscriber Line
AF Application Function 42
AM Acknowledged Mode 81
AMC Adaptive Modulation and Coding 16
AMBR Aggregated Maximum Bit Rate 38
AMR-NB Adaptive Multirate Narrow Band 149
API Application Program Interface 79
APN Access Point Name 18
AQM Active Queue Management 103
ARED Adaptive RED 103
ARP Allocation and Retention Priority 37
ATM Asynchronous Transfer Mode
BLER Block Error Rate 16
CAPEX CApital EXpenditure 1
CDF Cumulative Distribution Function 92
CDMA Code Division Multiple Access 37
CDN Content Delivery Network 99
CoDel Controlled Delay 103
CS Circuit Switched 23

CQA Channel and QoS Aware 89
CQI Channel Quality Indicator 10
DASH Dynamic Adaptive Streaming over HTTP 101
DiffServ Differentiated Services 23
DPI Deep Packet Inspection 76
DSCP Differentiated Services Code Point 31
DRB Data Radio Bearer 8
E2E End-to-End 71
eNB evolved Node B 3
ECM EPS Connection Management
EDGE Enhanced Data for GSM Evolution 5
EMM EPS Mobility Management 7
EPA Extended Pedestrian A 85
EPC Evolved Packet Core 5
EPS Evolved Packet System 5
ETSI European Telecommunications Standards Institute 5
E-RAB E-UTRAN Radio Access Bearer 18
E-UTRAN Evolved Universal Terrestrial Access Network 5
ESM Session Management 7
FIFO First In, First Out 33
FQ Fair Queuing 33
FTTH Fiber to the Home 27
FPI Flow Priority Index 76
GBR Guaranteed Bit Rate 7
GGSN Gateway GPRS Support Node 75
GPRS General Packet Radio Service 5
GTP GPRS Tunneling Protocol 9
GSM Global System for Mobile 5
HARQ Hybrid Automatic Repeat reQuest 12
HAS HTTP Adaptive Streaming 46
HO HandOver 7
HOL Head of Line 91
HSPA High Speed Packet Access 76

HSS Home Subscriber Server 18
IETF Internet Engineering Task Force 24
IMS IP Multimedia Subsystem 64
IMSI International Mobile Subscriber Identity 17
IntServ Integrated Services 26
IoT Internet of Things
IP Internet Protocol 23
ITU International Telecommunication Union 24
KPI Key Performance Indicators 121
LA Link Adaptation 13
LMA Local Mobility Anchor 7
LTE Long-Term Evolution xii
MAC Medium Access Control xii
MBR Maximum Bit Rate 9
MOS Mean Opinion Score 86
MCC Mobile Country Code 21
MCS Modulation and Coding Scheme 14
MiFi Mobile Wi-Fi 71
MIMO Multiple-Input-Multiple-Output 13
MME Mobility Management Entity 7
MNC Mobile Network Code 21
NAS Non Access Stratum 7
NFV Network Functions Virtualization
NP Network Performance 25
NRSPCA Network Requested Secondary PDP Context Activation 76
OCS Online Charging System 42
OFCS Offline Charging System 42
OFDM Orthogonal Frequency Division Multiple 13
OFDMA Orthogonal Frequency Division Multiple Access 13
OPEX Operational EXpenditure 1
OSI Open Systems Interconnection
OTT Over-The-Top 1

P-GW Packet Data Network Gateway 7
PCC Policy and Charging Control 39
PCEF Policy and Charging Enforcement Function 42
PCRF Policy and Charging Rules Function 42
PDCP Packet Data Convergence Protocol xii
PDN Packet Data Network 8
PDP Packet Data Protocol 35
PF Proportional Fair 74
PHB Per-Hop Behavior 31
PHY Physical xii
PIE Proportional Integral controller Enhanced 103
PLR Packet Loss Ratio 86
PLT Page Load Time 94
PoC Proof of Concept 117
PQ Priority Queuing 33
PRB Physical Resource Block 13
PS Packet Switched 10
PSS Priority Set Scheduler 89
QCI QoS Class Identifier 37
QoS Quality of Service iii
QoE Quality of Experience 25
QUIC Quick UDP Internet Connections
RAB Radio Access Bearer 18
RAN Radio Access Network 7
RB Resource Block 13
RBG Resource Block Group 89
RLC Radio Link Control xii
RNC Radio Network Controller 75
ROHC Robust Header Compression 12
RRC Radio Resource Control 10
RRM Radio resource management 7
RSVP Resource Reservation Protocol 29
RTT Round-Trip Time 100

S-GW Serving Gateway 7
SDF Service Data Flow 39
SDN Software Defined Networking 84
SGSN Serving GPRS Support Node
SINR Signal to Interference plus Noise Ratio 85
SMS Short Message Service
SLA Service Level Agreements 31
SPI Scheduling Priority Indicator 75
SPR Subscription Profile Repository 42
SPUD Session Protocol Underneath Datagrams
TA Tracking Area 21
TAC Tracking Area Code 21
TAI Tracking Area Identifier 21
TAL Tracking Area List 19
TAU Tracking Area Update 21
TC Traffic Class 31
TEID Tunnel Endpoint ID 10
TFT Traffic Flow Template 7
ToS Type of Service 31
TTI Time Transmission Interval 13
UE User Equipment 6
UM Unacknowledged Mode 81
UMTS Universal Mobile Telecommunications System 5
UPCON User Plane CONgestion 76
VAD Voice Activity Detection
ViLTE Video over LTE 47
VoD Video on Demand 147
VoLTE Voice over LTE 47
VoIP Voice over IP 35
V2V Vehicle-to-vehicle
WebRTC Web Real-Time Communication 46
WFQ Weighted Fair Queuing 33
WRR Weighted Round Robin 33

Contents

Acknowledgements **i**

Abstract **iii**

List of Acronyms **v**

Contents **xvi**

Introduction **1**

1 Context and Objectives 1

2 Contribution 2

3 Thesis Outline 3

I Overview of the LTE/EPC Mobile Network Architecture and Protocols **5**

1 Introduction 5

2 3GPP LTE/EPC Architecture 6

2.1 Evolved Packet Core 6

2.2 Evolved Universal Terrestrial Access Network 7

2.3 EPS bearer 8

2.4 3GPP LTE/EPC Protocol Stack 9

3 The Long-Term Evolution (LTE) Radio Interface 11
3.1 Radio Protocol Stack Overview 11
3.1.1 The Packet Data Convergence Protocol (PDCP) Sub Layer . . . 12
3.1.2 The Radio Link Control (RLC) Sub Layer 12
3.1.3 The Medium Access Control (MAC) Sub Layer 12
3.1.4 The Physical (PHY) Layer 13
3.2 Scheduling in LTE System 14
3.2.1 Radio Modulation and Coding Schemes 14
3.2.2 Radio scheduling algorithms 14
3.2.3 Opportunistic Scheduling 15
3.2.4 Link Adaptation in LTE system 16
3.3 LTE State Machines 16
4 LTE-EPC relevant procedures 17
4.1 The Attach and Default Bearer Setup: 17
4.2 Service Request procedure 19
4.2.1 Service Request triggered by the UE 19
4.2.2 Service Request triggered by the network 19
4.3 Handover procedure 20
4.4 Tracking Area Update 21
5 Conclusions 22

II QoS Management from Fixed to Mobile Networks 23

1 Introduction 23
2 Quality of Service in Packet Networks 24
2.1 Definition of QoS 24
2.2 QoS Approaches 26
2.2.1 Over-dimensioning 27
2.2.2 Per Flow 27
2.2.3 Per Aggregate 28

3 Overview of QoS in IP Networks 28
3.1 QoS Approaches in IP Networks 29
3.1.1 IntServ Model 29
3.1.2 DiffServ Model 30
3.2 DiffServ vs IntServ 35
4 Overview of QoS in 3GPP Mobile Networks 35
4.1 Where does 3GPP QoS come from? 35
4.2 3GPP LTE/EPS QoS Architecture 37
4.2.1 3GPP QoS Procedures 39
4.2.2 Discussion on Operators' Point of View 41
4.3 Introduction to 3GPP Policy and Charging Control 42
4.4 Identification of Critical Elements in the QoS Provisioning 44
4.4.1 Evolved Packet Core 44
4.4.2 Evolved Universal Terrestrial Access Network 44
5 Conclusions 46

III The Cost of QoS Management in 3GPP Mobile Networks 47

1 Introduction 47
1.1 Related Work 48
2 Cost Factor Analysis 48
2.1 Processing Load 48
2.2 Context Load 49
2.3 Memory Access Rate 49
2.4 Radio Signaling Overhead 49
3 Analytical Model Description 49
3.1 Application Model 51
3.2 User Equipment Model 51
3.3 Context Time Duration 53
3.4 System Model 53

3.5 Dedicated Bearer Model 54
3.6 Mobility Model 55
4 Analytical Model for Estimating the Cost Factors 56
4.1 Context Load Evaluation 56
4.2 Processing Load, Memory Access Rate and Radio Signaling Overhead Evaluation 56
4.2.1 Service Request procedure 57
4.2.2 Dedicated Bearer 59
4.2.3 Handover 60
4.2.4 Tracking Area Update 60
4.2.5 Summary of Processing Load and Memory Access Rate Evaluation 61
5 Numerical Results and Analysis 62
5.1 Network Parameters 62
5.2 Traffic Description 63
5.3 Scenarios Description 63
5.4 Impact of the Inactivity Timer 64
5.5 Impact of the Sessions Arrival Rate 66
5.6 QoS Signaling Impact in LTE Radio Segment 68
6 Conclusions 70

IV IP-centric QoS Model for 3GPP Mobile Networks 71

1 Introduction 71
2 Discussion on the 3GPP QoS management 72
2.1 3GPP QoS management 72
2.2 Today's mobile networks QoS policies 73
3 DiffServ approaches for QoS provisioning in Mobile Networks 74
3.1 Problem Statement and Solution Strategies 75
3.2 Related Work 75
4 IP Centric model for 3GPP Mobile Networks 77
4.1 IP-centric Architecture 77

4.1.1 Intra-bearer arrangements . . . 78
4.1.2 Inter-bearer arrangements . . . 78
4.2 Business models . . . 79
4.3 Cross-layer design for IP-aware eNB . . . 80
4.3.1 Design challenges . . . 80
4.3.2 Design description . . . 83
4.4 IP-centric QoS model and SDN mobile networks . . . 84
5 Simulation and Performance evaluation . . . 85
5.1 Network Parameters . . . 85
5.2 Traffic Description . . . 86
5.2.1 VoIP Traffic . . . 86
5.2.2 WEB Traffic . . . 86
5.2.3 HAS traffic - YouTube . . . 86
5.2.4 FTP Traffic (Background Traffic) . . . 89
5.3 Radio Schedulers description . . . 89
5.3.1 Proportional Fair Scheduler . . . 89
5.3.2 3rd Generation Partnership Project (3GPP) Schedulers . . . 89
5.4 Scenarios description . . . 91
5.5 Performance analysis . . . 92
6 Conclusions . . . 97

V Slo-Mo: An Implicit Mechanism to Improve QoS in Mobile Networks 99

1 Introduction . . . 99
2 TCP flow control in mobile networks . . . 100
2.1 TCP on radio medium . . . 100
2.2 Cross-layer solutions . . . 101
2.3 The Bufferbloat phenomenon . . . 102
3 Slo-Mo main features . . . 103
3.1 Motivations and principle . . . 104

3.2 Rate tracking . . . 105
3.3 QoS Differentiation in the P-GW . . . 106
3.4 TCP interactions with Slo-Mo . . . 106
4 Slo-Mo possible implementation . . . 107
4.1 QoS Differentiation through Priority Queueing . . . 107
4.2 Rate tracking . . . 107
4.2.1 Refresh on CoDel . . . 108
4.2.2 CoDel-like implementation . . . 109
4.2.3 Closed-loop-system . . . 111
5 Simulation and Performance Evaluation . . . 112
5.1 Network Parameters . . . 112
5.2 Traffic Description . . . 114
5.3 Scenarios Description . . . 114
5.4 Performance analysis . . . 114
6 Conclusion and Future Work . . . 117

Conclusions and Perspectives **119**

List of Publications **123**

Bibliography **139**

Résumé en Français **141**

Appendix A **147**

Appendix B **157**

List of Figures **163**

List of Tables **166**

Introduction

1 Context and Objectives

For some time now, the mobile ecosystem is undergoing major transformations, as evidenced by the skyrocketing explosion of mobile data traffic reported for example by Cisco [1] and Ericsson [2]. Mobile devices as smartphones and tablets have also evolved. Today they have better connectivity capabilities, close to or even better than fixed devices. Usages and traffic patterns have also changed, as video and music streaming are common today but were almost nonexistent some years ago.

This evolution resuts from increased network capacities thanks to huge investments (i.e. LTE). Furthermore, the mobile ecosystem has been completely transformed with the arrival of new actors called Over-The-Top (OTT), who have completely disturbed the existing balance of forces. This new reality breaks the old telephony business model in which the mobile operators had a completer control on the services they offered on their networks. Indeed, OTTs have introduced web oriented interfaces as well as services and models from the fixed Internet world, which have led to evolution of mobile services to the detriment of mobile operators [3]. Hence, OTTs have became today key actors in this new mobile ecosystem. On the same time, cheap data plans are now flourishing, especially in Europe. All these factors lead to an important loss of value for mobile operators. As a consequence, CApital EXpenditure (CAPEX) and Operational EXpenditure (OPEX) reduction is actively sought by mobile operators together with QoS preservation. Note that this question is particularly crucial on the radio segment due to high costs and subsequent frequent bottlenecks.

The mobile standards have not yet fully integrated this expectation of open, cheap and flexible web oriented mobile Internet access. Current 3GPP QoS management for LTE/EPC networks [4, 5] offer a highly granular QoS management (per flow level), which

inherits many characteristics of legacy telco standards. Namely, LTE QoS management is based on a circuit-oriented model, relying essentially on virtual circuits called "bearers", which provide for end-to-end transport service with specific QoS attributes. Though its good QoS performances, such a scheme reveals costly and cumbersome. As a matter of fact, 3GPP QoS features remain very poorly deployed in commercial mobile networks, thereby giving way to basic best-effort (no QoS) hegemony. As a result, bottlenecks often appear in today mobile networks, mainly on radio segments.

Therefore, the 3GPP QoS model is probably not the right answer to the somehow contradictory expectations for cheap, efficient and flexible QoS in the mobile Internet. In order to cope with this issue, this thesis has two major objectives:

i) First, it investigates the standardised QoS model in order to identify and analyse cost factors related to it and finally to propose tools to assess these cost factors taking into consideration the complex characteristics of the current mobile Internet traffic.

ii) Second, it investigates alternative approaches in order to provide a cost-effective QoS management better adapted to the current and future mobile data traffic. Those alternative QoS models are mainly inspired of different mechanisms existing in the domain of QoS and traffic management in the fixed Internet world and must improve cost-factors identified in the first part.

2 Contribution

The main contributions of this thesis are as follows:

◊ Identifying the main weaknesses of current mobile standard for QoS provisioning and proposing an analytical model in order to evaluate the impact of current mobile standard for QoS, in terms of Processing Load, Memory access and Context Load.

◊ Characterizing and modeling of the main traffic sources for current mobile ecosystem (Web and HAS). These models improve the reliability of our research findings, inasmuch as that reflect the real mobile traffic behaviour.

◊ Introducing an IP-centric QoS model for mobile networks called "IP-Aware". This model solves previously identified weaknesses of current QoS standard. Current and proposed QoS models are completely compatible and may be deployed in parallel.

◊ Developing a cross-layer architecture for IP-centric model, which requires minimal modifications of current QoS standard in order to simplify its implementation, deployment and operation.

◊ Introducing an implicit cross-layer mechanism called "Slo-Mo" for improving QoS on mobile networks. Slo-Mo implements a lightweight mechanism based on an innovate queue management strategy.

3 Thesis Outline

In this thesis we tackle the End-to-End QoS in LTE/EPC in exploring the state-of-the art and an in-depth study of the weaknesses of current QoS to finally propose two complement solutions to improve QoS in mobile networks. This thesis is organized as follows:

In chapter I we provide a background information about LTE/EPC networks by reviewing its architecture and protocols, which is detailed by 3GPP standards. This include the review of protocol stacks and main procedures, which are mainly focused on the radio segment.

In chapter II we provide a general panorama of the QoS management in packet networks and its approaches, which have been extensively studied in the field of fixed networks. Subsequently, these are compared to the current mobile QoS approach, in order to figure out their main differences and identify possible improvements.

In chapter III we present a cost factor analysis of 3GPP QoS model. We develop an analytical model and define cost analysis metrics in order to investigate the weak points of the standardized model. At this stage, we carry out comparative cost analysis of different scenarios where the 3GPP QoS model is deployed, which pave the way for proposing enhancements or even novel schemes.

In chapter IV we introduce an IP-centric model, which is mainly inspired from DiffServ architecture widely used in fixed IP networks. Our proposed IP-centric cross-layer scheme, called IP-Aware, aims to provide a cost-effective QoS, while it meets QoS requirements demanded by current and future mobile data traffic. The IP-aware scheme is designed under the constraint of performing minimum modifications of current standard in order to make its implementation, deployment and operation easy. Performances of this proposal compared to various implementations of the 3GPP QoS model is evaluated using the ns-3 simulator in realistic scenarios. However, as our IP-Aware proposal requires some enhancements in the evolved Node B (eNB), this scheme will not be implemented by all vendors nor standardised in short term. In the meanwhile, lightweight solutions are greatly hoped for.

In this respect, in chapter V we introduce Slo-Mo, which is an implicit cross-layer mechanism for managing QoS from a distant point when the congestion point (e.g. eNB) is not able to do it. Slo-Mo enhances the customer experience at marginal cost and does not require any major modification in the LTE/EPC elements, which allows a deployment in short term. Performances of Slo-Mo is evaluated using the ns-3 simulator in realistic scenarios and some good properties of Slo-Mo are also brought into evidence.

Chapter I

Overview of the LTE/EPC Mobile Network Architecture and Protocols

1 Introduction

The beginning of digital mobile communication has been marked by the Global System for Mobile (GSM), which was launched early in the 90s. GSM is a standard developed by the European Telecommunications Standards Institute (ETSI) and originally was developed to transmit voice. GSM is considered a technology of second generation or 2G because the first generation were the earlier analogical mobile systems. In order to be able to transmit data GSM specifications were evolved towards General Packet Radio Service (GPRS) and then Enhanced Data for GSM Evolution (EDGE) by the 3GPP, which is a global organization of groups.

Subsequently, the 3GPP developed third-generation (3G) or Universal Mobile Telecommunications System (UMTS) standards, followed by fourth-generation (4G) or Long-Term Evolution (LTE) standards. The LTE aims to propose technical specification for mobile system evolution. It seeks to improve mobile technology by increasing downlink and uplink peak data rates, proposing a scalable bandwidth, improving spectral efficiency, using an all-IP architecture. In this sense the 3GPP specify the Evolved Packet System (EPS), which is composed by the new access network called Long-Term Evolution (LTE) or Evolved Universal Terrestrial Access Network (E-UTRAN) and a new core network called Evolved Packet Core (EPC). In order to avoid confusion we will use only the term EPC to refer to the core network of EPS and E-UTRAN to refer to the access network.

We can found in [6] and [7] the following definition: *"The network architecture of LTE is based on the functional decomposition principle, where required features are decomposed into functional entities without specific implementation assumptions about physical network entities. This is why 3GPP specified a new packet core, the EPC, network*

architecture to support the E-UTRAN through a reduction in the number of network elements, simpler functionality, improved redundancy, and most importantly allowing for connections and hand over to other fixed line and wireless access technologies, giving the service providers the ability to deliver a seamless mobility experience".

In this chapter we present an overview of LTE/EPC mobile networks, focusing on LTE protocol stack and relevant signaling procedures, which direct relation with the QoS architecture and procedures. Signaling procedures addressed in this chapter are relevant for a better understanding of chapter III, where a study of the impact related to those procedures is presented. Regarding the LTE protocol stack, it is important to understand its architecture and main functions in order to identify possible improvements in the QoS management.

2 3GPP LTE/EPC Architecture

The LTE/EPC architecture is an All-IP network, which means that real time services and non-real time services are transported by the IP protocol. Figure I.1 shows the EPS and its interfaces, which will be described in following sections.

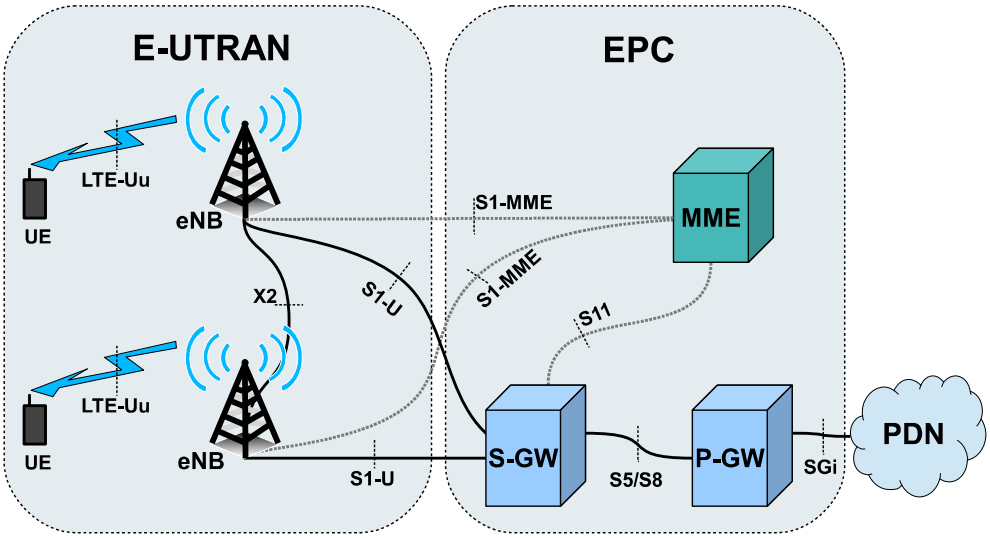


Figure I.1 – 3GPP LTE Architecture

2.1 Evolved Packet Core

The Evolved Packet Core (EPC) is the core network of EPS, which is responsible for the overall control of the User Equipment (UE) and establishment of the bearers.The main logical nodes are shown in Figure I.1 and are described in [8], below a brief description of them:

- **Packet Data Network Gateway (P-GW)**: is responsible for UE IP address allocation, as well as QoS enforcement. Therefore, the P-GW performs marking and filtering of downlink IP packets based on Traffic Flow Templates (TFTs), placing it into its corresponding bearers. Furthermore, The P-GW performs QoS enforcement for Guaranteed Bit Rate (GBR) bearers. It also serves as the mobility anchor for inter-working with non-3GPP technologies. Moreover, it is a point used for lawful interceptions.
- **Serving Gateway (S-GW)**: acts as the Local Mobility Anchor (LMA) for bearers when the UE moves through the E-UTRAN (S1 HandOver (HO)). In case of mobility between LTE and other 3GPP technologies, the S-GW acts as mobility anchor. It is also a point used for lawful interceptions.
- **Mobility Management Entity (MME)**: is the control node that processes the signaling between the UE and the EPC. Such signaling is related to the Non Access Stratum (NAS) protocols. The MME performs functions related to bearer management as establishment, maintenance and release. It is also charged of the connection management and security between the network and UE.

The NAS is a set of protocols in the EPS, which are specified in [9]. It is the highest stratum of the Control Plane between UE and MME at the E-UTRAN. Main functions are the EPS Mobility Management (EMM) and Session Management (ESM). The EMM refers to procedures related to mobility over the E-UTRAN as well as the authentication and security. The ESM refers to session management procedures to establish and maintain IP connectivity between the UE and the P-GW.

2.2 Evolved Universal Terrestrial Access Network

The E-UTRAN consists of a network of eNBs, as illustrated in Figure I.1. The aim of a Radio Access Network (RAN) composed of only one type of network element (i.e. eNB) is to reduce latency of all radio interface operations. The eNBs are connected to Evolved Packet Core (EPC) through the S1 interface (S1-U for User Plan and S1-MME for Control Plan) and are also connected to each other via the X2 interface.

The E-UTRAN is responsible for all radio-related functions, in case of the eNB, its main functions can be summarized as:

- Radio resource management (RRM) is relative to radio bearer management such as radio bearer control, radio admission control, connection mobility control and dynamic allocation of resources to UEs in both Uplink and Downlink (scheduling);
- IP header compression and encryption of user data stream;
- Routing of User Plane data towards the corresponding S-GW;

- Measurement and measurement reporting configuration for mobility and scheduling;
- Scheduling and transmission of paging messages and broadcast information.

Figure I.2 summarizes the functional splits between E-UTRAN and EPC. Grey boxes depict the logical nodes, turquoise boxes depict the functional entities of the Control Plane and blue boxes depict the radio protocol layers.

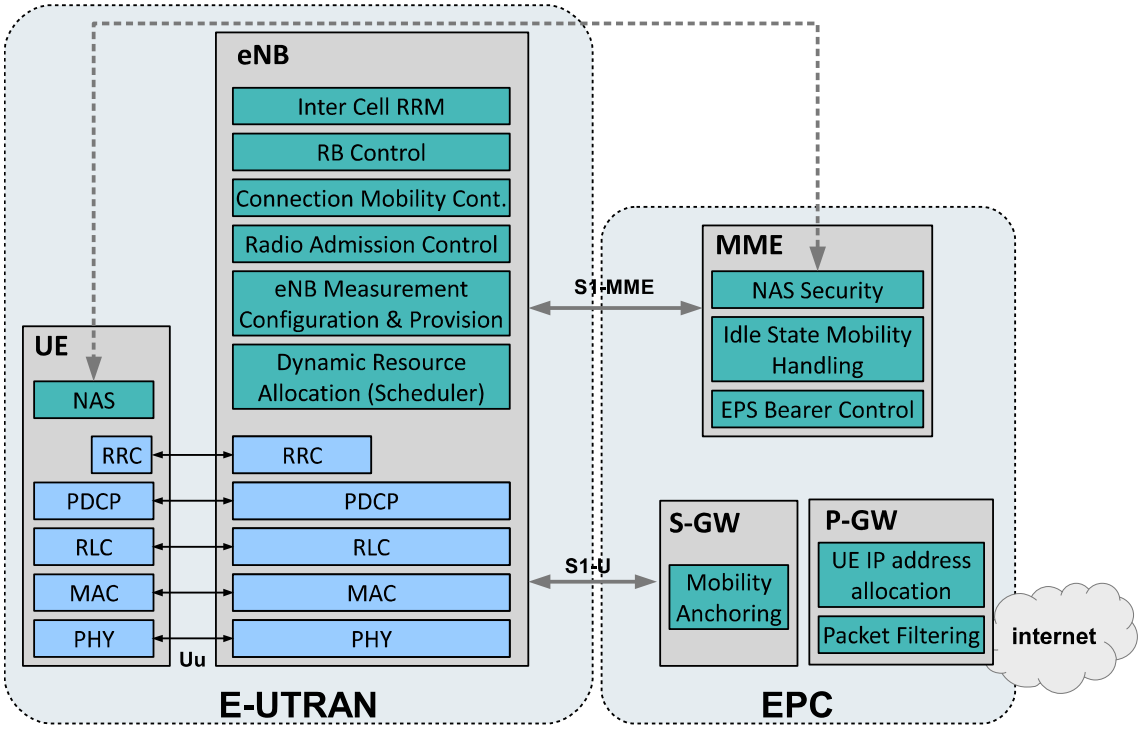


Figure I.2 – Functional Split between E-UTRAN and EPC [10]

2.3 EPS bearer

The LTE/EPS architecture requires the establishment of a logical connection between the end points: P-GW and UE. This logical connection is called EPS bearer, which is set up between UE and the P-GW before any user traffic can be exchanged.

In [11], the authors define the EPS bearer as the representation of the level of granularity for QoS control in E-UTRAN/EPC and that provides a logical transmission path with well-defined QoS properties between UE and the Packet Data Network (PDN).

It should be noted that this EPS bearer is constituted of several local bearers established between neighbors network elements. A Data Radio Bearer (DRB) is set up between the UE and the eNB, a S1 bearer is set up between the eNB and the P-GW and a S5 or S8

bearer is set up between the P-GW and the S-GW. In case of S1 bearer and S5/S8 bearer, the GPRS Tunneling Protocol (GTP)-U is used. The EPS bearer service architecture is depicted in Figure I.3.

For a unique terminal, multiple bearers can be established, one per required QoS level. For each bearer establishment, control plane signaling protocols are used in order to communicate the bearer information to each LTE network element (i.e. eNB, P-GW, S-GW). When an UE is attached to the network, a default bearer with a "Best Effort" QoS is established. Other EPS bearers could be further set up, one per required QoS level.

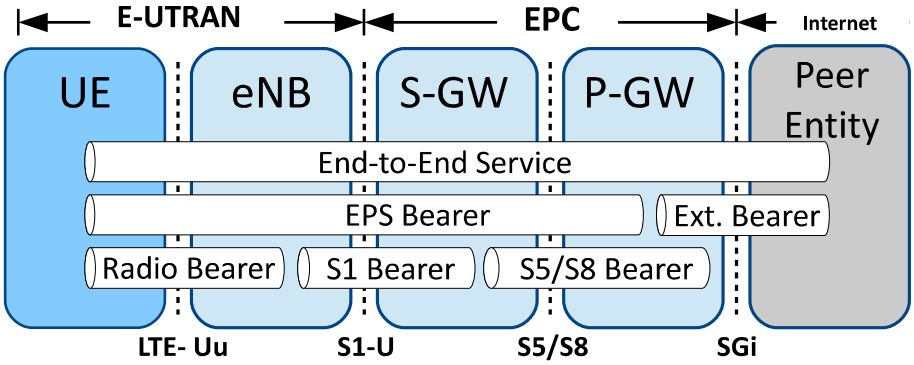


Figure I.3 – EPS bearer architecture

Bearers can be classified into two categories:

- Guaranteed Bit Rate (GBR) bearers have permanently allocation of dedicated network resources when they are established or modified. In cases of availability of resources, higher bit rate than GBR may be allowed with a maximum limit defined by the parameter Maximum Bit Rate (MBR).
- Non-Guaranteed Bit Rate (GBR) is referred to those bearers that not have allocation of dedicated network resources such as the default bearer.

2.4 3GPP LTE/EPC Protocol Stack

The protocol stack structure of LTE/EPC system is illustrated in Figure I.4. It is divided into two main groups according to the final purpose service: User Plan protocols and Control Plan protocols. User Plane protocols are responsible to carry user data through access stratum, whereas Control Plane protocols are responsible of connection control between the UE and the network through the establishment, modification and release of bearers.

In Figure I.4 turquoise boxes depict the Control Plane protocol stack and blue boxes depict the User Plane protocol stack. The Control Plane is used to control the radio

access bearers and the connection between the UE and the network (i.e. signaling between E-UTRAN and EPC). The control Plane consists of:

- The network access connection (attaching and detaching procedures);
- The attribution of an established network access connection, such as activation of an IP address;
- The user mobility (i.e routing);
- The resources allocation in order to meet users demands, respecting operator policies

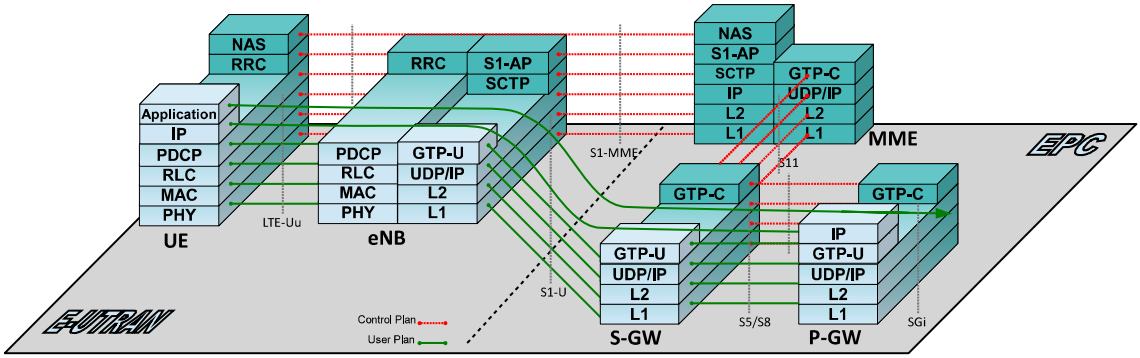


Figure I.4 – User Plane and Control Plane protocol stack

Furthermore, the NAS protocol is also used for control such as network attach and identification, bearers establishments and mobility management. It should also be noted that all NAS messages are ciphered and their integrity is checked by MME and UE. On the other hand, the Radio Resource Control (RRC) layer is used by the eNB in order to makes decision about handover process, pages for the UE over the E-UTRAN, controls UE measurement reporting (i.e. Channel Quality Indicator (CQI)) and allocates temporary identifiers to active UE on the cell-level. RRC is also involved in the set up and maintenance of bearers, and configuration of all the lower layers (i.e. PHY, MAC, RLC and PDCP). In this respect, RRC signaling is used between the eNB and the UE.

In E-UTRAN, the radio access uses the protocols MAC, RLC and PDCP. The interface S1-U (S1 User Plane) is based on the GTP. GTP consists in a virtual tunnel, which ensure the correct delivery of IP packets destined to a given UE. In this sense, GTP encapsulates the IP packet into another IP packet, which is addressed to the eNB, where the UE is currently attached. GTP is a set of protocols within 3GPP Packet Switched (PS) core network (i.e. GPRS, UMTS, EPC), which is composed by GTP-C and GTP-U. The GTP-C is used to Control Plane and the GTP-U is used to User Plane. A GTP tunnel is identified in each node with a Tunnel Endpoint ID (TEID) (4 bytes field in the GTP header), an IP address and a UDP port number. TEID values are exchanged between the tunnel end-points.

3 The LTE Radio Interface

As described previously, the P-GW provides connectivity from external packet data networks to UEs. Downstream IP packets are then carried through the LTE network to the eNB via a GPRS Tunneling Protocol (GTP) tunnel (named S1 bearer); the eNB removes the GTP header of each IP packet before delivering it to the radio interface (Uu). Figure I.5 shows the eNB protocol stack for the downstream radio interface (Uu). 3GPP specifications stipulate the creation of independents PDCP and Radio Link Control (RLC) entities for each EPS bearer. A brief summary of the radio protocol stack and its main functions can be found below.

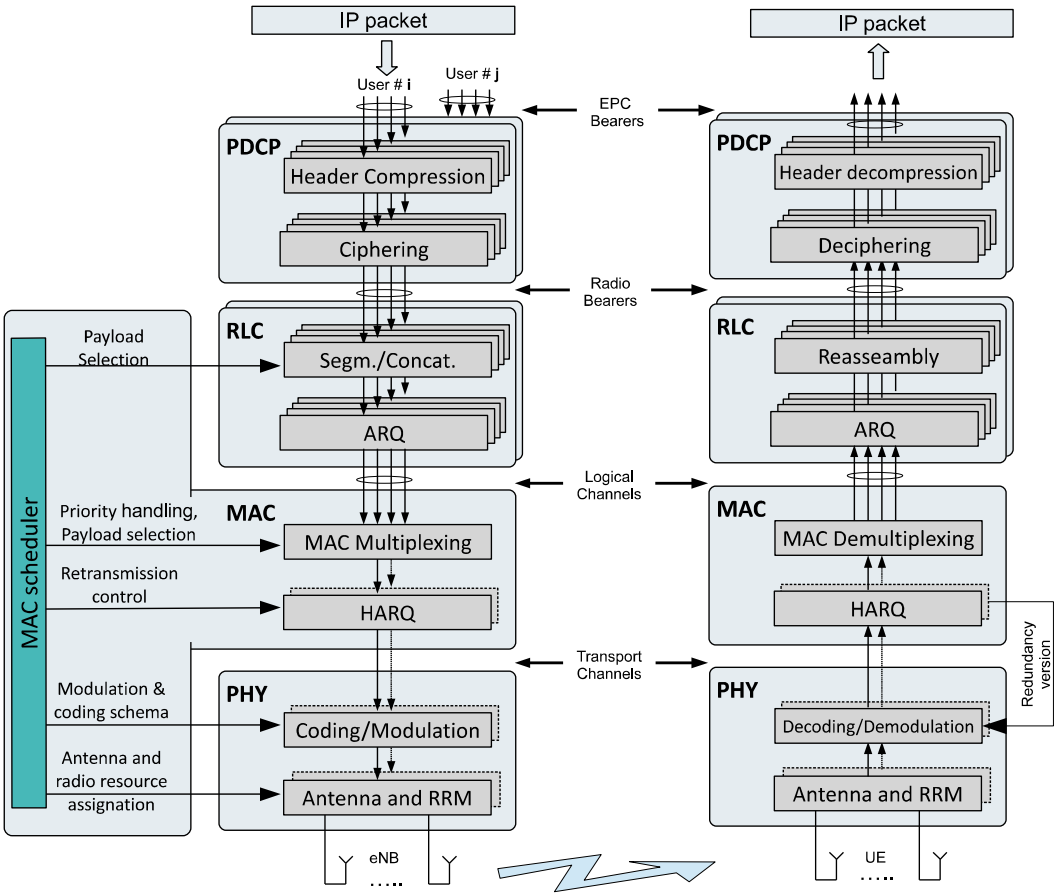


Figure I.5 – eNB: Uu interface Downlink protocol stack (simplified - *source: [12]*)

3.1 Radio Protocol Stack Overview

On the Air interface, Layer 2 is split into the following sub-layers: Medium Access Control (MAC), Radio Link Control (RLC) and Packet Data Convergence Protocol (PDCP). We give a description of the Layer 2 sub-layers of eNB in terms of services and functions.

The figures below depict the MAC/ RLC/PDCP/MAC architecture for downlink . In order to transport data across the LTE radio interface, various "channels" are defined, and may be grouped into:

- **Logical Channels** define *what type* of information is transmitted over the air interface (e.g. traffic channels, control channels, system broadcast). User and control messages are carried on logical channels between the RLC and MAC layers.
- **Transport Channels** define *how data units are* transmitted over the air interface (e.g. what are encoding, interleaving options used to transmit data). User and control messages are carried on transport channels between the MAC and the PHY layer.
- **Physical Channels** define *where data units are* transmitted over the air interface (e.g. number of symbols in the Downlink frame). User and control messages are carried on physical channels between the different levels of the PHY layer.

3.1.1 The Packet Data Convergence Protocol (PDCP) Sub Layer

According to 3GPP specifications [10], PDCP sub-layer provides data transfer, header compression using the Robust Header Compression (ROHC) algorithm, ciphering (User and Control planes), and integrity protection for the control plane.

3.1.2 The Radio Link Control (RLC) Sub Layer

According to 3GPP specifications [13], RLC sub-layer performs segmentation/concatenation and reassembly. It can be operated in three modes: Transparent Mode (TM), Unacknowledged Mode (UM) and Acknowledged Mode (AM). However TM is used only for control plane signaling. The choice of the RLC transmission mode depends on whether delay or integrity is favoured at radio level. Each PDCP entity is associated with one or two RLC entities (i.e. uni-directional or bi-directional), which depends on the radio bearer characteristics and the RLC transmission mode (TM/UM or AM). Each RLC entity implements a buffer in order to store packets coming from the PDCP sub-layer and reports periodically the amount of buffered data to the Medium Access Control (MAC) sub-layer.

3.1.3 The Medium Access Control (MAC) Sub Layer

According to 3GPP specifications [10], MAC sub-layer multiplexes logical channels (radio bearers) onto a MAC SDUs transport blocks on transport channels, performs scheduling of resources, error correction using Hybrid Automatic Repeat reQuest (HARQ), and transport format selection. It guarantees the QoS for each radio bearer by instructing the RLC sub-layer about the amount of data to be transmitted from each radio bearer, based on the

scheduler strategy. Every Time Transmission Interval (TTI), the radio scheduler distributes the available radio resources among all UEs according to the scheduler strategy. The scheduling strategy is mainly based on the QoS requirements, the RLC buffers status and the Link Adaptation (LA) algorithm.

3.1.4 The Physical (PHY) Layer

The PHY Layer is based on Orthogonal Frequency Division Multiple Access (OFDMA) technology, combined with a high performing of modulations (i.e. 64QAM), large bandwidths (up to 20 MHz) and spatial multiplexing based on Multiple-Input-Multiple-Output (MIMO) technology in the Downlink (up to 4x4).

In Orthogonal Frequency Division Multiple (OFDM) systems, many radio resources can be allocated every TTI, which corresponds to the sub-frame duration, to each UE. In LTE, the Resource Block (RB) is the fundamental unit being used for resource allocation. The numbers of RBs available depend on the system bandwidth (e.g. for 20 MHz of bandwidth, 100 RBs are available). The LTE physical layer supports a subset of 6 different system bandwidth (1.4, 3.5, 10, 15 and 20 MHz), according to the current LTE specifications. Figure I.6 shows the Physical Resource Block (PRB) structure in time and frequency dimensions.

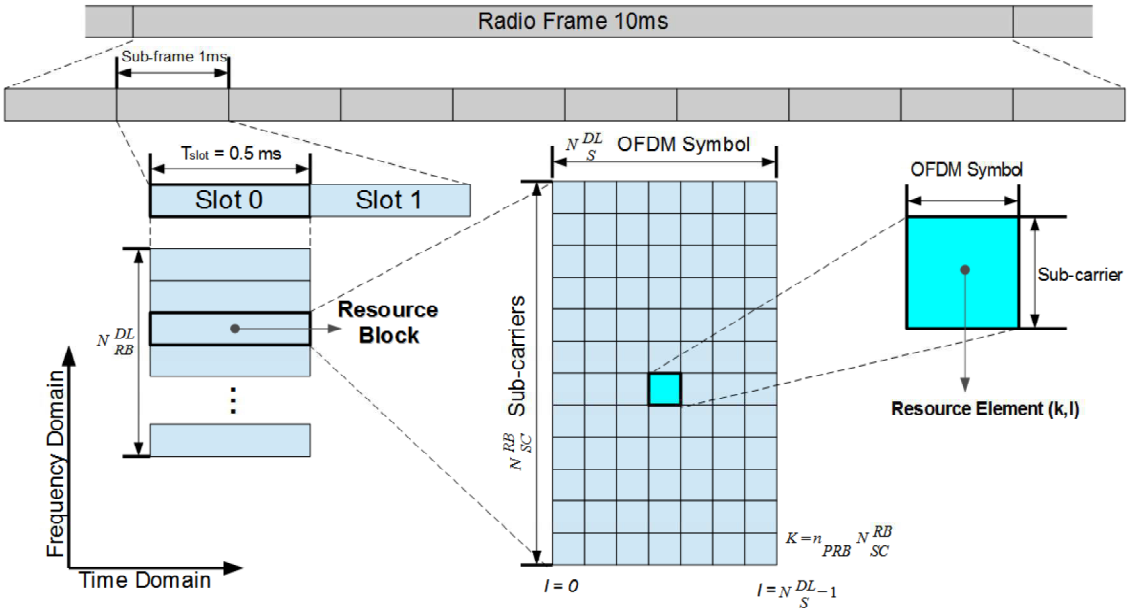


Figure I.6 – Physical Resource Block (PRB) structure

Another key aspect is the reduction of TTI to 1ms compared to 10ms specified in Release 99 and 2ms in Release 5 [14]. As a result, high data rates with low delays can be achieved. The theoretical peak data rate on the transport channel is 75 Mbps in the Uplink and 300 Mbps in the Downlink, using spatial multiplexing.

3.2 Scheduling in LTE System

3.2.1 Radio Modulation and Coding Schemes

The quality of the radio channel of a given UE is a key parameter determining its achievable throughput. Indeed, sophisticated radio Modulation and Coding Scheme (MCS) can be used when the UE is in very good radio conditions, leading to a higher throughput per radio resource. On the contrary, in poor radio conditions, an UE requires more robust MCS and experiences lower bit rates for the same amount of radio resources allocated. In other words, in poor radio conditions, an UE will need much more radio resources to reach the same throughput when compared to an UE in good radio conditions. For example, 64QAM modulation allows 3 times more throughput than QPSK modulation for the same amount of radio resources allocated.

3.2.2 Radio scheduling algorithms

The packet scheduling plays a key role for QoS provisioning in the current communication networks. Its basic function is to determine UEs who will be served and the transmission order of their packets. This function becomes critical in congestion points, because resources will not be enough to satisfy all QoS needs. Therefore, some services will be degraded for the benefit of priority services according to packet scheduler strategy.

The principal purpose in designing of packet schedulers is to satisfy both user requirements (e.g. delay, throughput or packet loss) and the network requirements (e.g. efficiency and simplicity of implementation). Therefore we can summarize the packet schedulers design goals as following properties:

- **Efficiency**: it should provide the same QoS under any network state (e.g. traffic load and users load).
- **Flow isolation**: it should provide a QoS level to a flow with a minimal impact to others flows.
- **Flexibility**: it should be able to support users with a broad range of services and their corresponding QoS levels.
- **Low complexity**: it should have a reasonable computational implementation complexity. Nowadays the technology of the communication equipments allow reaching higher and higher data rates. The packets processing rate is directly related to the scheduling algorithm complexity, which has become a critical parameter to be aware.

Many packets scheduling for wireless networks proposed were inspired of packet scheduling algorithms that demonstrated excellent results in wired networks. But, due

to fundamental differences in physical layer, special considerations should be taken into account in wireless packet scheduling design.

In LTE system, every Time Transmission Interval (TTI), radio resources are dynamically allocated to the active UEs of the cell according to a scheduling algorithm. Various examples of radio scheduling algorithms can be found in the literature [15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26].

Mobile network vendors generally implement adaptations of the Proportional Fair algorithm. This algorithm proposes a trade-off between cell throughput optimization and fairness (see [27]). This type of scheduler prioritizes users which offer the highest ratio of achievable instantaneous throughput normalized by its mean throughput. For example, a user in good radio conditions will frequently be scheduled at the beginning. But after a certain amount of time, its mean throughput will increase, and the associated ratio will become lower than the ratio of another user with poorer channel quality which has still not transmitted. In the long term, fairness in terms of radio resources is therefore ensured.

3.2.3 Opportunistic Scheduling

The opportunistic Scheduling takes advantage of instantaneous channel variations for distributing resources in a wireless network. This can be performed for example by giving priority to the users with favorable channel conditions. Opportunistic scheduling tackles the issue of scheduling from different aspects. Therefore we can classify Opportunistic Schedulers in two main groups:

Best-Effort schedulers We can define as best-effort scheduler, those that do not provide any QoS guarantees to each UE or services. The main objective of the best-effort schedulers is to provide a set of requirements (e.g. maximum throughput, minimum delay, etc) to the overall system. The Round Robin, Maximum C/I and Proportional Fair are some examples of best-effort schedulers.

QoS-Aware schedulers QoS-Aware schedulers those that provides a QoS level to each UE or service. For example, some schedulers propose to assign to certain UEs a minimum GBR on the radio segment, almost independently of their radio conditions. In other schedulers, a weight is assigned to each UE and is further used in the Proportional Fair algorithm. In this case, the priority of UEs is governed by the weights of favoured UE against those of non-favoured UEs. These QoS aware schedulers are generally vendor specific. QoS-Aware scheduler will be addressed in next chapter.

3.2.4 Link Adaptation in LTE system

In LTE, Link Adaptation (LA) is based on the Adaptive Modulation and Coding (AMC). AMC is the appropriate choice of the MCS to the conditions on the radio link, thanks to the information provided by the system. To perform channel estimation in LTE system, reference signals are embedded in each PRB both downlink and uplink. (there are four reference OFDM symbols within one PRB).

Each UE sends a CQI to the eNB, this CQI reports the downlink signal strength based on the channel estimation performed by the UE. The conversion from UE power measures of reference signal to a CQI value is not specified by LTE standards. The only requirement is that the MCS associated to CQI value be able to guarantee that the Block Error Rate (BLER) does not exceed 10%.

3.3 LTE State Machines

The EPS Mobility Management (EMM) protocol provides procedures related to mobility over the E-UTRAN like access, authentication and security (e.g. Attach/detach, Tracking Area Update). There are two main EMM states described in the specifications [9], EMM-DEREGISTERED and EMM-REGISTERED. These EMM states determine the reachability of a UE as well as the capability of a UE to exchange user traffic with the network.

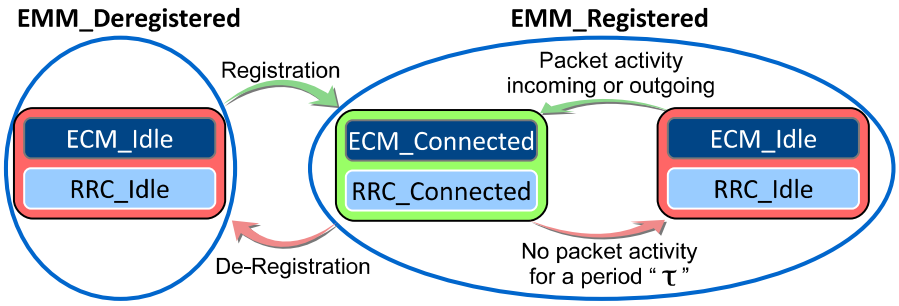


Figure I.7 – EMM-ECM-RRC states

Once a UE is registered in an LTE/EPC network (EMM-REGISTERED), the ECM states describes the signaling connectivity between the UE and the EPC [8, 9, 28]. A UE can be either in CONNECTED state (*ECM-Connected / Radio RRC-Connected*) or in IDLE state (*ECM-Idle / RRC-Idle*). In the CONNECTED state, the UE has a data connectivity in the E-UTRAN (UE↔eNB), and a signalling connectivity in the EPC (UE↔MME). After an inactivity period (RRC inactivity timer), the UE switches to IDLE state and its corresponding radio resources are released in the E-UTRAN via a specific signalling procedure illustrated in Fig. I.8. Thus, only the resources allocated in the EPC are kept

active. Fig. I.7 illustrates EMM, ECM and RRC states associated with the User-plan and Control-plan status.

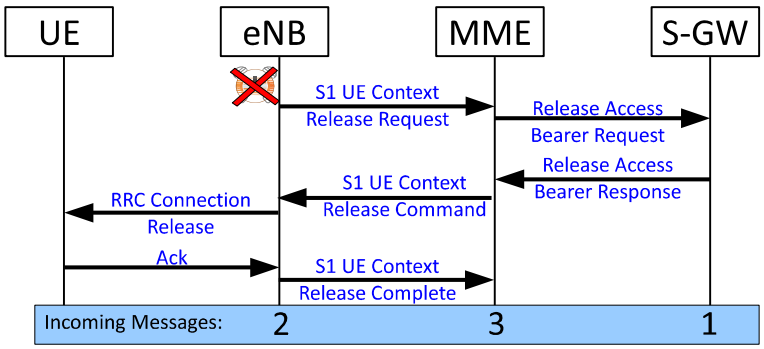


Figure I.8 – Switch from CONNECTED state to IDLE state

Fig. I.9 illustrates the bearer states in each LTE/EPC network segment after and before the Service Request procedure (IDLE / CONNECTED states).

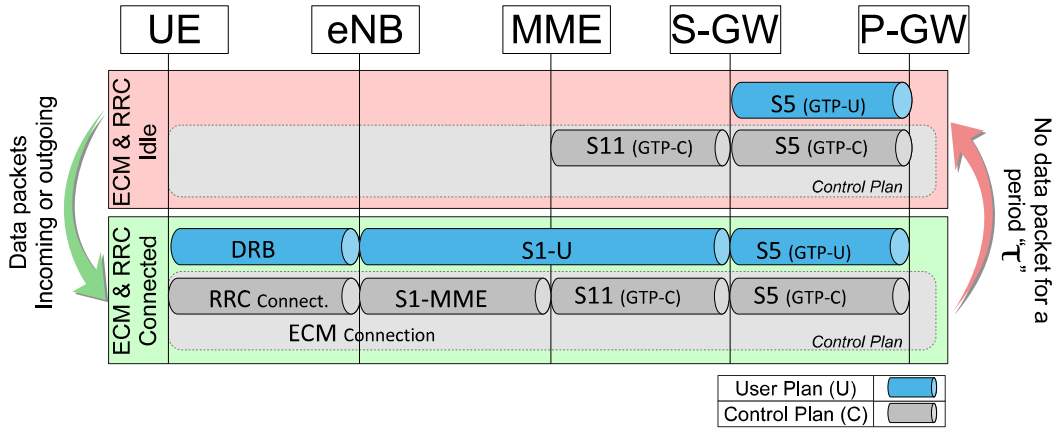


Figure I.9 – Bearer states before/after Service Request procedure

4 LTE-EPC relevant procedures

4.1 The Attach and Default Bearer Setup:

Typical UE attach procedure is specified in [8] and its simplified procedure is showed in Figure I.10. The procedure begins when a user sends an Attach Request message to an MME, and ends when the MME returns an Attach Accept message to the UE. The Attach and Default Bearer Setup Messaging are shown in Figure I.10 .

1. *Initial UE Message* is the first message sent to the MME to establish a connection, it includes the UE-IDs (i.e. International Mobile Subscriber Identity (IMSI)).

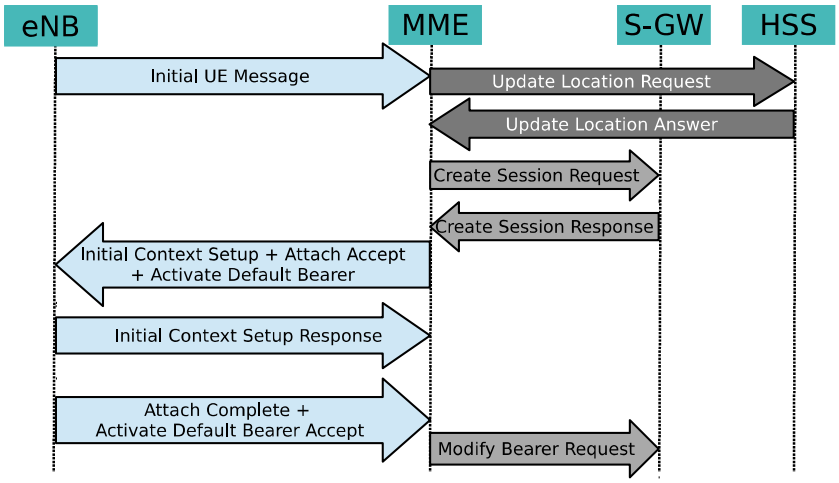


Figure I.10 – LTE Attach and Default Bearer Setup Messaging

2. MME updates the UE location in the Home Subscriber Server (HSS), which supports the database containing all the user subscription information.

3. The HSS accesses the database and responds with user information to the MME (the Access Point Name (APN) configuration is also included).

4. The Default Bearer Establishment procedure is started, the MME initiates the default route establishment by asking the selected S-GW to create a tunnel GTP. The S-GW creates a new entry in its EPS bearer table and sends a *Create Session Request* message to the P-GW, which provides UE IP address.

5. The next message from the MME is *Initial Context Setup Request*, which contains: *Initial Context Setup Request, NAS Attach Accept and Activate Default Bearer Request* messages. *Initial Context Setup* message contains a request to establish a context between the MME and eNB, and also contains S-GW tunneling information. *NAS Attach Accept* message acknowledges the successful Attach to the UE, eNB re transmits this message to the UE. *Activate Default Bearer Request* message initiates the default bearer setup on the UE and eNB also re transmits this message to the UE.

6. The eNB sends the *Initial Context Setup Response* message to the MME, which confirms the establishment of the GTP tunnel on the S1-U interface and also contains information about the Radio Access Bearers (RABs) (i.e. E-UTRAN Radio Access Bearer (E-RAB) ID, transport layer IP address, the eNB GTP TEID).

7. eNB transports*Attach Complete* and *Activate Default Bearer Accept* messages, which are received from the UE.

8. Finally, MME informs S-GW about the eNB's User Plane IP address and GTP TEID.

4.2 Service Request procedure

4.2.1 Service Request triggered by the UE

When data traffic is emitted by a UE in IDLE state, the UE performs the procedure illustrated in Fig. I.11. The UE sends the MME a Service Request message to establish an ECM connection. This message is delivered through a RRC connection established over the radio segment between the UE and the eNB, and then through the S1 signaling connection established between the eNB and the MME. After receiving the Service Request message from the UE, the MME sends a Initial Context Setup Request to eNB in order to establish a DRB and a downlink S1 bearer. Thus, connectivity is setup in the control plan and user plane, allowing the UE to receive and send data traffic.

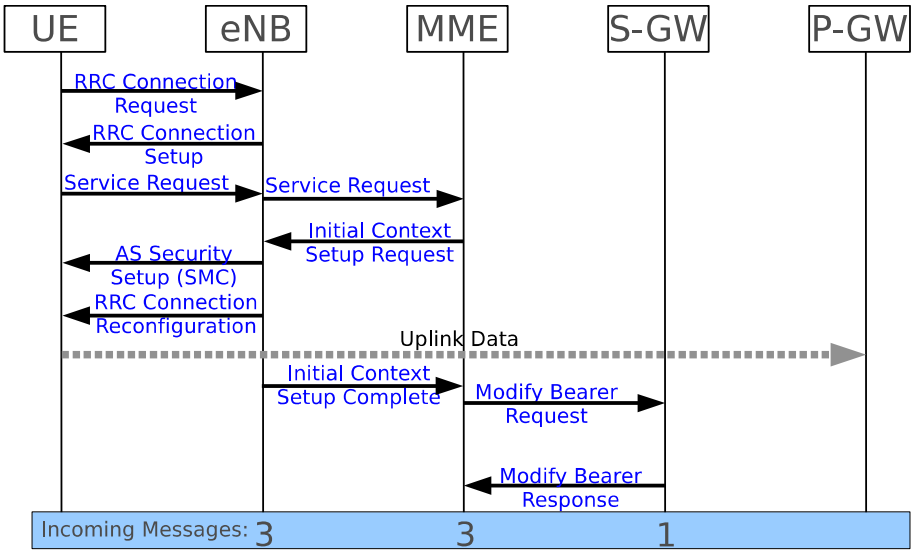


Figure I.11 – UE triggered Service Request procedure

4.2.2 Service Request triggered by the network

If the network has data to send to a UE in IDLE state, the procedure illustrated in Fig. I.12 is performed. First of all, the S-GW buffers the downlink packets and identifies which MME is serving the UE, then the S-GW sends a message to the MME in order to trigger the paging procedure to find and activate the UE. The UE is paged in all eNBs belonging to the Tracking Area List (TAL) in which it is currently registered. Once the UE is located and becomes aware of incoming traffic, it triggers the *Service Request* procedure as was described previously.

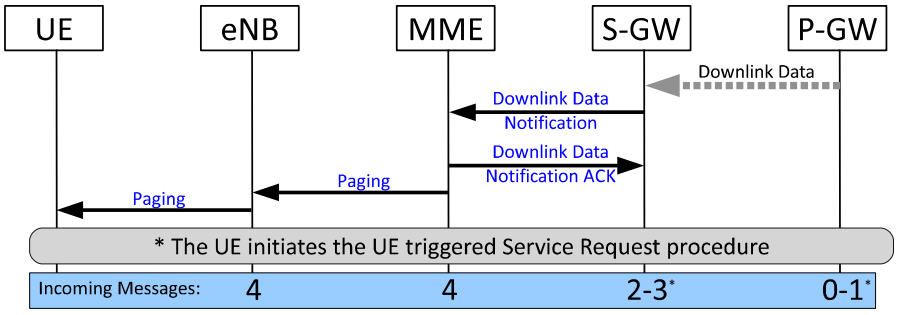
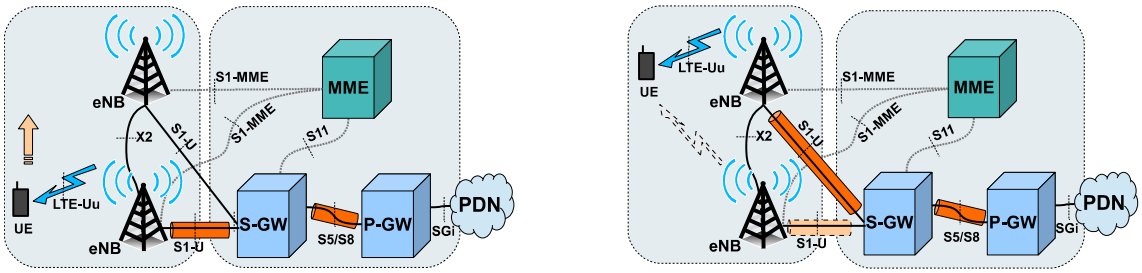


Figure I.12 – Network triggered Service Request procedure



(a) Before Handover **(b)** After Handover

Figure I.13 – Handover procedure

4.3 Handover procedure

The handover procedure handles mobility when a UE is in the CONNECTED state (i.e. the UE has a communication in progress), as shown in Figure I.13. There exist several scenarios to trigger this procedure, but assuming that X2 interfaces are available on every eNB, we can list two relevant scenarios:

1. Handover without S-GW relocation
2. Handover with S-GW relocation

The handover preparation, execution and completion phases are performed as specified in [4]. Fig. I.14 shown the call flow of handover with [1 - 7] and without [1 - 9] S-GW relocation. As part of handover execution, downlink packets are forwarded from the source eNB to the target eNB via the X2 interface. Uplink data from the UE can be delivered via the S-GW or source S-GW to the P-GW, depending on handover scenario. In both handover scenarios, the preparation and execution phases are identical. In handover without S-GW relocation, the source S-GW not exist and the target S-GW will be called just S-GW.

(1) In handover completion phase, the target eNB sends a *Path Switch Request* message to MME to inform that the UE performs a handover, this message contain the list of EPS

bearer used by the UE. (2) The MME sends a *Modify Bearer Request* message to the (target) S-GW, which contains the list of all validated EPS bearers accepted by the target eNB. (3) The (target) S-GW informs the P-GW about the handover event and sends enough information as the list of EPS bearers. (4,5) Then a *Modify Bearer Response* message is sent back to the (target) S-GW, who sends back *Modify Bearer Response* message to the MME. (6,7) The MME confirms previous received message with a *Path Switch Request Ack* message to target eNB, who informs to source eNB the success of the handover and triggers the release of resources. (8,9) In the case of a handover with S-GW relocation, the MME releases the bearer(s) in the source S-GW by sending a *Delete Session Request* message, which is acknowledged by the source S-GW by the *Delete Session Response* message.

It is important to highlight that the number of signaling messages used in handover procedures does not depend on the QoS levels used by a UE (i.e. number of EPS bearers per UE). On the contrary, the number of context modifications in the various involved LTE/EPC elements depends linearly on the number of EPS bearers - thus of QoS levels - per UE.

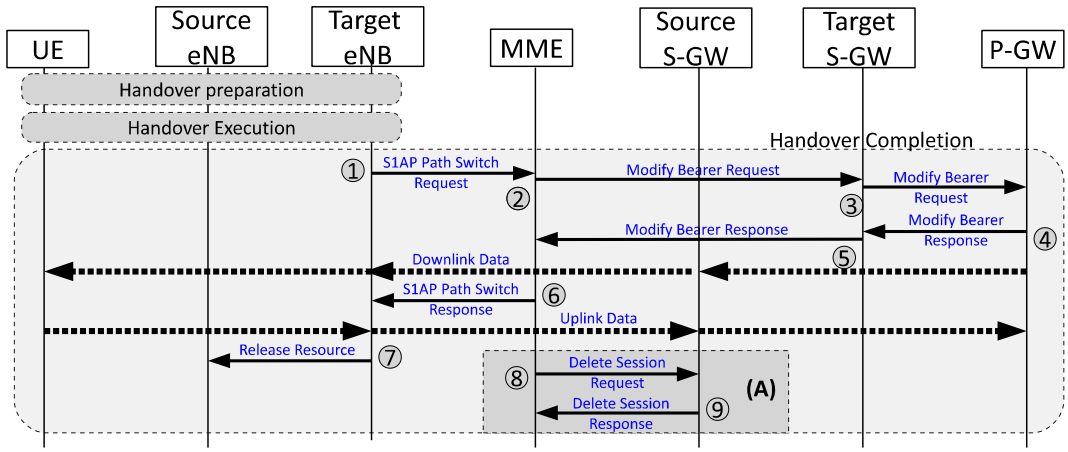


Figure I.14 – X2 Handover procedure

4.4 Tracking Area Update

While the UE is in CONNECTED state, its location is known by the LTE network at cell level. However, in IDLE state the UE location is only known at TAL level, which is a group of Tracking Area (TA). A TA is a group of neighbor eNBs, which are defined by the operator. The TA concept permits the reduction of signaling and delay, when an UE in IDLE state notifies the LTE network of its current TAL by sending a Tracking Area Update (TAU) message every time that it moves to another TAL. Each TA is identified by a Tracking Area Code (TAC), which is a unique assigned by the operator. A Tracking Area Identifier (TAI) is also defined, which consists of Mobile Country Code (MCC), Mobile Network Code (MNC) and TAC.

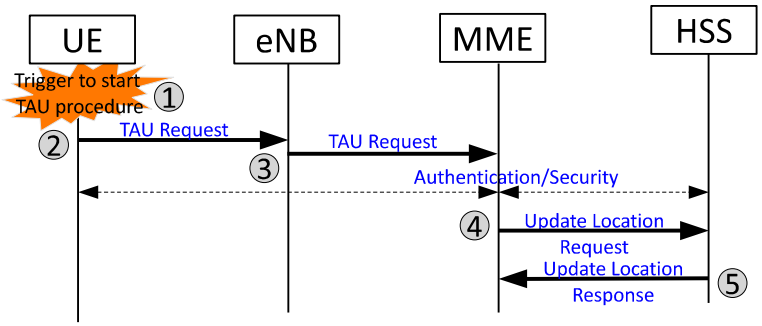


Figure I.15 – Tracking Area Update procedure

The UE in IDLE state notifies the LTE network of its current TA location by sending a Tracking Area Update (TAU) message every time it moves between TAs. When a TAU is triggers, it could involve a MME change but it has a low probability. The Call flow of the case of TAU without MME change is shown in Fig. I.15.

5 Conclusions

In this chapter we presented an overview of LTE/EPC mobile network. We focused on the E-UTRAN architecture and its protocol stack since it includes the radio interface which is a critical segment due to high variability and subsequent frequent bottleneck. In this chapter we also detailed some procedures related to fundamental characteristics of mobile networks, as they can have a non negligible impact in the QoS provisioning.

The LTE/EPC system is an all-IP network, which introduces challenges in terms of QoS, especially for real time services. Therefore, mechanisms to manage the QoS should be implemented, to allow for a graceful coexistence between applications with various QoS requirements. In this regard, in the next chapter we are going to present the most important QoS approaches for communication networks as well as discuss about the QoS approach of 3GPP mobile system in order to understand the major issues of this approach.

Chapter II

QoS Management from Fixed to Mobile Networks

1 Introduction

At the beginning, the communication networks were designed as several separate physical networks in which each one carried specific type of traffic. An example of this is the voice services, traditionally transported in Circuit Switched (CS) infrastructure. Over time, the packetized transmissions have demonstrated to be more interesting en terms of cost compared to conventional Circuit Switched (CS) communications. Nowadays, the convergence is the trend of communication networks. Therefore, the current communication networks are a packet-based networks which can provide all services and where the Internet Protocol (IP) is commonly used. Then, the term all-IP is used today to define it.

However, this heterogeneity where each type of traffic have different requirements (e.g. delay, data rate), lead to significant challenges to guarantee a minimum QoS. Today, the QoShas become in an inherent need in communication networks, especially in those where the physical medium condition is fluctuating, which is the case of wireless medium (e.g. mobile networks, Wi-Fi, satellite, etc). The importance of QoS has increased as users and applications increasingly consume data across the communication networks. The over-provisioning of the network is a way to meet the QoS requirements, but is very expensive and hardly ever the way taken. Hence the interest in developing mechanisms to improve traffic management through the introduction of intelligent control schemes, especially for wireless networks.

This chapter presents a brief description about the QoS in packet networks and its approaches (over-dimensioning, per flow and per aggregating). Then, an overview about the QoS in current IP networks is presented, which is based on Differentiated Services (DiffServ) approach. Furthermore, an overview about the QoS in 3GPP mobile networks

is detailed. Finally, works based on the DiffServ approach for QoS provisioning in Mobile Networks are presented and analyzed.

2 Quality of Service in Packet Networks

Evolution of telecommunication networks is characterized by a great heterogeneity as is shown in Figure II.1. This heterogeneity implies almost all layers of OSI model, L1/L2 (wireline and wireless technologies).

Current Internet ecosystem experiences a huge diversification of services and applications wich have different requirements in terms of QoS. Therefore, guaranteeing an appropriate level of QoS has become a challenge to telco operators.

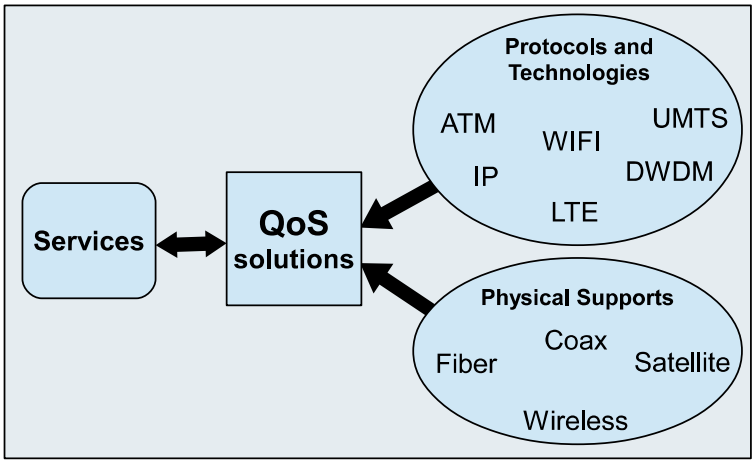


Figure II.1 – QoS in Heterogeneous Network

But, what QoS means? and how operators can guarantee an acceptable QoS level to their customers?. Following is a resume of main QoS definitions , which were proposed by standard bodies as International Telecommunication Union (ITU), ETSI and Internet Engineering Task Force (IETF).

2.1 Definition of QoS

In order to clarify the QoS definition W.C. Hardy defines three notions of QoS in [29], which has become a general model. These three notions are:

(a) **Intrinsic QoS** measures the satisfaction of arbitrary delivery criteria, in terms of packet delay, jitter, packet loss etc, result of technical choices adopted in the network;

(b) **Perceived QoS** measures the satisfaction of the user experience. Thus, it reflects a subjective component influenced by the user expectation. Which means that same services with same intrinsic QoS may be perceived differently by different users. And this is sometimes defined as Quality of Experience (QoE).

(c) **Assessed QoS** defines the willingness of a user to continue to use a service. This decision can be influenced by several factors (i.e price, perceived QoS, customer service). None organization as ITU, ETSI or IETF defines this notion of QoS.

The definition of the QoS of ITU [30] and ETSI [31] is basically the same. They define the QoS as *"the collective effect of service performance which determine the degree of satisfaction of a user of the service"*. This definition mainly involves the perceived QoS and slightly the intrinsic QoS. It also introduces the notion of Network Performance (NP), which is part of the intrinsic QoS in general model. In ITU/ETSI definition, the NP has a relationship with the QoS but not is part of it. In this definition QoS cover the following four points of view as is also shown in Figure II.2:

- QoS requirements of the customer
- QoS perceived by the customer
- QoS offered by the provider
- QoS achieved by the provider

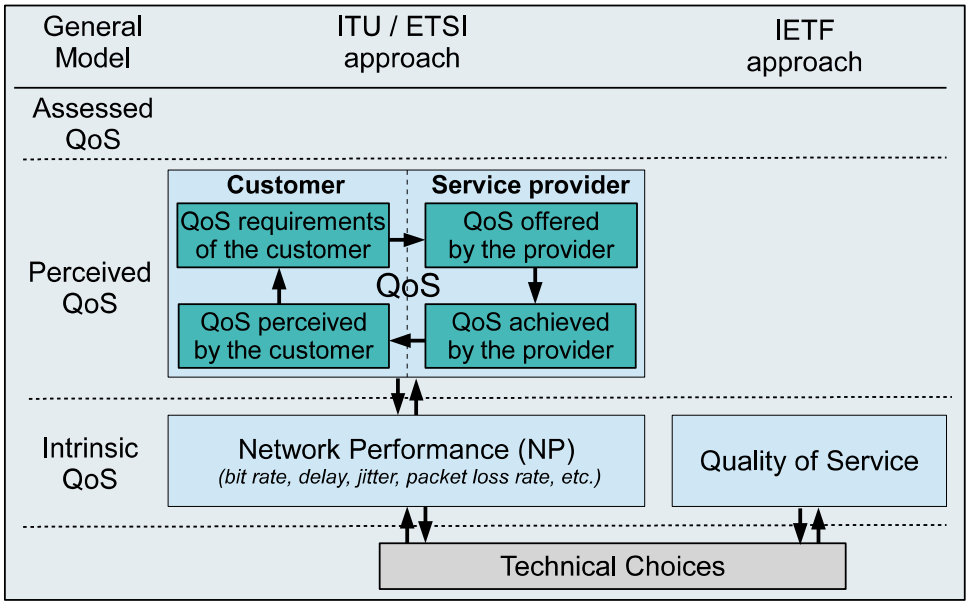


Figure II.2 – ITU/ETSI and IETF approaches and general QoS model [32]

Furthermore, the IETF [33] only focuses on the intrinsic QoS and defines it as *"A set of service requirements to be met by the network while transporting a flow"*. This because the main objective of this organization is the definition of the Internet architecture and its evolution. It must underline that this definition is equivalent to the notion of Network Performance (NP) proposed by ITU/ETSI. IETF has worked in two approaches for QoS provisioning and proposes their respective network architectures, Integrated Services (IntServ) and Differentiated Services (DiffServ), which are described below in this chapter.

Intrinsic QoS for the Internet have been widely developed mainly by IETF. Architectures and mechanisms have also been proposed, in which, many important aspects should be determined by the operators for a correct implementation and a efficient exploitation. Intrinsic QoS can be expressed by the following set of metrics:

- **Bit rate** is the measure the number of bits that are conveyed across a segment of the network
- **One-way delay** is the time taken for a packet to be transported from the sender to the destination
- **Jitter** measures the variation of the packet latency belonging to the same packet flow that cross a network
- **Packet loss rate** expresses the ratio between the total of successful delivered packets and those sent

Parameters of perceived QoS are more difficult to define. They depend not only on the network choices. It exist many works that study the relationship between intrinsic and perceived QoS, but in this chapter we focus on describing intrinsic QoS, presenting its different generic approaches.We will focus on current models for QoS provisioning in fixed IP networks and 3GPP mobile networks, presenting their QoS architectures and relevant procedures.

2.2 QoS Approaches

Different generic approaches to QoS provisioning are considered, in this section we discuss three generic approaches: over-dimensioning approach, per flow approach and per aggregation approach. We describe strategies adopted of each approach and some examples.

2.2.1 Over-dimensioning

The most common approach to QoS provisioning is to Over-dimensioning. This approach is the most simple but efficient for ensuring the QoS, the main idea is to dimension the network capacity based on peak traffic load estimation. Over-dimensioning the network has been often used to avoid congestion, packet delays and loss without traffic distinction. Nevertheless, this approach imply high CAPEX and OPEX so today is mainly used in critical parts of the network. The monitoring of the traffic evolution is critical to ensure that traffic load does not exceed network capacity. Therefore, this requires careful and continuous engineering work, taking into account the likely effects of the equipment failures and demand forecasting.

However, over-dimensioning is only useful in network segments with predictable peak loads. This approach is not recommendable over all network, particularly in segments with high load variability as the access network. In fact, network dimensioning and investment are lengthy processes and in case of a sudden increase in traffic or in network equipment failure, over-dimensioning is not a viable solution.

In some cases over-dimensioning the network is a feasible QoS choice for the operator (e.g. Core Network). Core Network is characterized by very high capacity links receiving aggregated traffic flows from a multitude of sources. This huge traffic aggregation allows to have a statistical knowledge of traffic behavior, which produce an almost constant load (law of large numbers). An over-dimensioned network might occasionally suffer overload in failure conditions, but it is almost always anticipated with a redundant architecture due to critical impact that failures could have. In this case, load depends on ratio between peak rate of individual flows and link capacity. In order to have an optimal utilization, engineering rules specify that this ratio should be lower than 1%. For example, 100 Mbps Fiber to the Home (FTTH) access for a link of 10 Gbps. The network would need to be sized so that the given utilization level is not exceeded even in anticipated failure conditions. It is important to mention that the consequence of overload can be very serious and impact many users and services.

2.2.2 Per Flow

The IETF Request for Comments (RFC) 3697 [34] defines traffic flow as *"a sequence of packets sent from a particular source to a particular unicast, anycast, or multicast destination that the source desires to label as a flow. A flow could consist of all packets in a specific transport connection or a media stream. However, a flow is not necessarily 1:1 mapped to a transport connection"*. In IP networks a flow is identified by its source/destination IP address and source/destination port number.

In this approach the QoS provisioning is based on allocating resources individually to flows manifested as signaled connections. Individual flows should be identified in order to perform a QoS management more flexible and precise (per flow). This identification

allows to enable rules on each intermediate node in order to specify how a specific flow should be treated. In this case we can talk about fine tuning QoS because it would be possible prioritize flows according to user preferences on the basis of the end-user as well as the application. For example, if a user wants establish a communication path, first it signals the traffic characteristics and performance requirements of the flow (in terms of intrinsic QoS metrics), the network performs admission control and allocates the necessary resources.

In case of ATM technology QoS provisioning was based on this approach. Five service categories were defined. The QoS on ATM was mainly built based on its connection-oriented nature. IETF also proposes a QoS architecture based on this approach called Integrated Services (IntServ) [35]. IntServ architecture will detailed in next section.

2.2.3 Per Aggregate

This approach is used in order to simplify the QoS management, it does not provide a fine tuning of QoS as per flow approach. This approach is built on the principle of the flows aggregation, in which a giving number of flows having the same QoS requirements (i.e data rate, delay, etc) are treated in the same way. This model provides a "soft" QoS provisioning but more scalable than per flow approach. Scalability is achieved since it not policing individual application flows but by the flows aggregation (e.g. class of service). QoS provisioning of a given class mainly depends on the characteristics of the flows composing that class and on the overall volume of traffic that it generates. It also depends up to which point the characteristics and volume can be controlled. Classes of service may be defined to distinguish the different types of traffic. They can additionally be defined to distinguish traffic of particular services or users.

One example of aggregation QoS is the provision of a class based service differentiation. An operator may provide multiple levels of service to clients for example Gold, Silver and Bronze. Clients with a gold profile will be provided with a "better" QoS than Silver and Bronze ones. This "better" level of service may be expressed in terms of higher data rates, lower delay and/or lower packet loss rates.

IETF proposes a QoS architecture based on this approach called Differentiated Services (DiffServ) [36]. DiffServ was an alternative QoS architecture developed by IETF. The basic assumption of this model was to achieve scalability. DiffServ architecture will detailed in next section.

3 Overview of QoS in IP Networks

IP network was designed to be as simple as possible. The main function of the IP network was to forward packets from one node to the next without any guarantee. All packets

are treated the same way and stored in a single buffer and forwarded in FIFO manner. Furthermore, most of intelligence is placed in terminals, which is typically a computer. For example, if a packet arrived at its destination with error, it could be re-transmitted using a simple ACK/NACK mechanism. The capability of retransmitting lost or errored packets was placed in the terminal. Because the early IP network carried basically only one kind of information, non-real time data, for this reason the network was designed to operate in the "best effort" mode. Which means all packets were treated equally.

In 90's the idea to create a single "converged" network to carry both voice and data started in with the aim to have a more cost-efficient network. With this convergence, new technical challenges have appeared. The best effort operation of the first IP network is no longer good enough to meet diverse performance requirements. The QoS provides solutions to this technical problem.

3.1 QoS Approaches in IP Networks

As was mention at the above section, there exist three main approaches to QoS provisioning in IP networks. In case of over-provisioning approach, is mainly used in core network due its high CAPEX/OPEX and its statistical advantages due to the traffic aggregation.

On the other hand, in access and transport networks an over-provisioning approach is unworkable given the high costs that it would imply. The IETF has defined two models to address the fundamental problem of providing QoS in IP networks, IntServ model (IETF RFC 1633 [35]) and DiffServ model (IETF RFC 2475 [36]). The following describes the previously mentioned model of QoS management in IP networks.

3.1.1 IntServ Model

IntServ model is based in per-flow QoS approach, where each flow is identified by its source/destination IP address and its source/destination transport layer port number. IntServ aims to provide individualized QoS guarantees to individual sessions. IntServ is based on building a virtual circuit in an IP network using the resources (i.e. delay and bandwidth) reservation technique as shown in Figure II.3. A Call Setup is performed before to establish a virtual circuit in order to reserve sufficient resources at each network element on path to ensure QoS requirements are met. When resources are reserved for a certain service, they cannot be reassigned for another service. In order to perform the resource reservation across an IP network the Resource Reservation Protocol (RSVP) is used by IntServ. RSVP is a signaling protocol aims a dynamic resources reservation in IP networks, which is specified in IETF RFC 2205 [37].

The IntServ architecture defines three main types of services:

- **Guaranteed Service [38]:** guarantees that datagrams will arrive within the guaran-

teed delivery time and will not be discarded due to queue overflows, provided the flow's traffic stays within its specified traffic parameters

- **Controlled Load service:** provides a QoS closely approximating to a QoS that a flow would receive from a lightly loaded network element
- **Best Effort service:** does not provide any guarantee of QoS

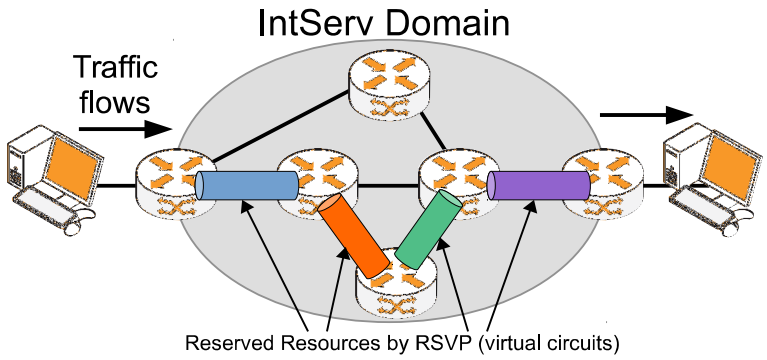


Figure II.3 – IntServ model

Technically, it could be possible to provide the requisite QoS for every flow in an IP network, as long as a resource reservation protocol (e.g. RSVP) is used and the resources are available. But the task of reserving resources in order to create virtual circuits would be very tedious in a busy network such as the Internet. Therefore, there are several drawbacks to this approach:

- Every network element along the path of a flow, including the end elements (e.g. PC, server) must implement RSVP protocol and be capable of signaling the required QoS.
- Maintaining contexts in each network element and increased memory requirements to support a large number of reservations (contexts), adds to the complexity of each network element along the path.
- Because a context for each reservation needs to be maintained at every network element, the scalability with large number of flows through a large IP network becomes a serious issue.

3.1.2 DiffServ Model

In the DiffServ architecture, traffic entering to a network is classified and conditioned at the ingress node of the network only, and assigned to different behavior aggregates. The

8-bit Type of Service (ToS) field in the IP header has been included to support packet classification. The ToS byte is divided into 6 bit Differentiated Services Code Point (DSCP) field and a 2 bits unused field. A DiffServ behavior aggregate is a set of flows with the same DSCP, crossing a link in a particular direction. This causes QoS assurance to be applied only in one direction of data transmission as in the IntServ model. The number of services is limited to 64.

In the DiffServ model independent flows select one of the predefined class of services and are served in the same way as other flows that choose the same service level. Flows served by the same service are aggregated and experience the same QoS level.

DiffServ Architecture: The DiffServ architecture consists of a number of functional elements that are implemented in network nodes as routers. These include per hop forwarding behaviors and traffic classification and marking. Network traffic is divided into a small number of forwarding classes and resources are allocated to each class. Traffic first enters at the ingress nodes of the network, where it is classified and assigned into a forwarding classes.

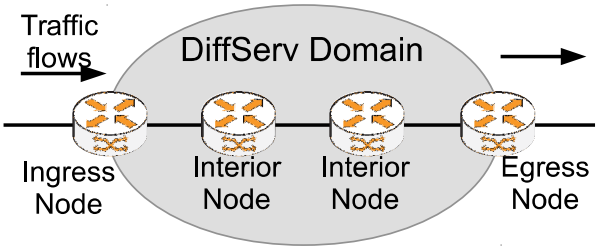


Figure II.4 – DiffServ domain

Therefore the DiffServ architecture consists of two types of nodes which have different responsibilities. The edge (ingress) node is responsible for traffic classification and conditioning and the interior or interior node is responsible for Per-Hop Behavior (PHB) based forwarding of packets as shown in Figure II.4.

Packets Classification and Marking: Classification involves mapping selected packets to a particular forwarding class by marking them with the corresponding Differentiated Services Code Point (DSCP). DiffServ uses the ToS field of the IPv4 header and the Traffic Class (TC) field of the IPv6 header for marking packets. At the ingress nodes, the ToS (IPv4) and the TC (IPv6) fields are overridden and redefined as the DiffServ field. In other words the ToS and TC fields, are used for DiffServ packet marking. It also ensures that the traffic flows from a customer are in conformance with their Service Level Agreementss (SLAs). This involves also taking action for non conformant packets; dropping them for example as shown in Figure II.5.

The interior nodes are responsible for identifying the forwarding class from the DSCP

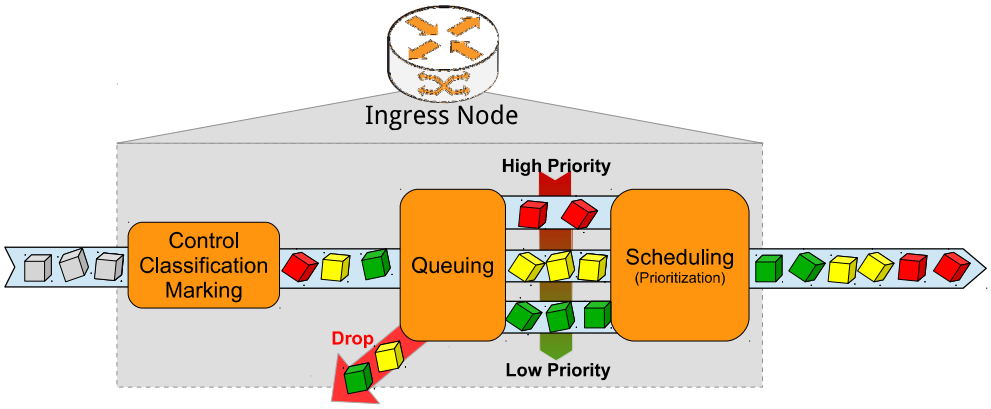


Figure II.5 – Packets Classification and Marking (Ingress Node)

of the packet IP header and applying the corresponding Per-Hop Behavior (PHB). It must be noted that ingress nodes must be capable of doing this as well. In this way the DiffServ architecture becomes more scalable as instead of managing some individual flows (as IntServ), it only manages fewer aggregate flows. Scalability is also achieved by only implementing the classifying and marking functions at network ingress nodes.

Per-Hop Behavior: Aggregated flows processing by a network node is called Per-Hop Behavior (PHB) and is defined formally on RFC2475 [36] as *"externally observable forwarding behavior applied at a DiffServ compliant node to a DiffServ aggregate"*. DiffServ is fundamentally per-hop based and is based on defining PHB's for individual nodes. However, to provide DiffServ across a DiffServ domain, PHB's for each individual node may be designed in such a way that the overall end-to-end QoS is best provided.

A DiffServ domain consists of interconnected DiffServ complaint nodes that implement the same PHBs, and share a common service provisioning policy. This service provisioning policy defines how the traffic conditioners located on ingress nodes are configured as well as how traffic streams are mapped to behavior aggregates. A DiffServ domain normally consists of networks under the same administration (e.g. telco operator network). The service provider is responsible to ensure that the domain is provisioned with sufficient resources to support the SLAs offered by the domain.

Queuing and Scheduling: Queuing and scheduling allows traffic to be split into multiple queues in order that the scheduler can treat the traffic inside each queue in a different way, for example according its priority. In IntServ model, IP packets to each queue belongs to a specific class of service. Then, scheduler can apply differentiated behavior to different classes of service, as illustrated in Figure II.7.

The queuing and scheduling strategy determines how the resources are allocated. However, queuing and scheduling system is only useful in presence of congestion. Which

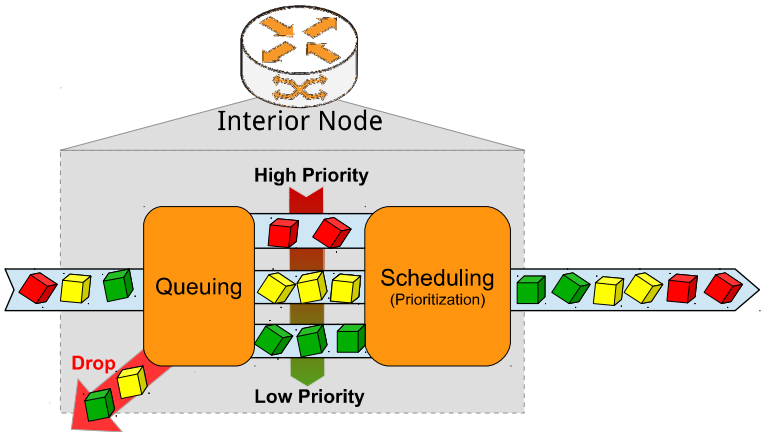


Figure II.6 – Per-Hop Behavior (PHB) (Interior Node)

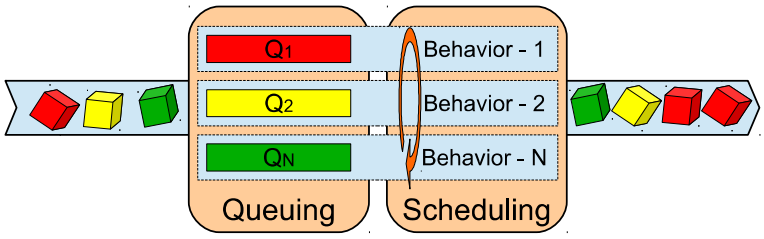


Figure II.7 – Queuing and Scheduling strategy

means that if resources are sufficient and there is no competition for resources, there is no need for queuing. Congestion is generated when traffic that has to be sent by an interface is higher than the outgoing line speed it can support. Otherwise, congestion can also be created artificially, by applying a shaping rate to the interface that imposes a speed limit lower than the maximum interface line speed. Then, overflow traffic is placed in corresponding queues, then these queues are visited by the scheduler, which is responsible for the rate at which packets from each queue are transmitted.

Over a number of years, many works have led to a variety of queuing algorithms, but describing all existing queuing mechanisms is impossible. However, some major well-known disciplines regarding scheduling are detailed below. However, the four main scheduling mechanisms are discussed in the literature: First In, First Out (FIFO) queuing, Fair Queuing (FQ), Priority Queuing (PQ), Weighted Fair Queuing (WFQ) and Weighted Round Robin (WRR) queuing are described below.

a) **First In, First Out (FIFO) queuing:** is the most basic queuing scheduling discipline, where all packets are treated equally. Basically, there is one queue and the scheduler only serves this queue and packets are serviced in the same order in which they were placed into the queue, as illustrated in Figure II.8-a.

b) **Fair Queuing (FQ):** is a scheduling algorithm that addresses the basic limitation of FIFO queuing, which does not separate flows. FQ classifies packets belonging to

the same flow or classes of service into multiple queues, offering a fair scheduling scheme for the flows to access the link, as illustrated in Figure II.8-b.

c) **Priority Queuing (PQ):** is a scheduling that provides a simple method of supporting differentiated service classes. PQ classifies packets belonging to the same flow or classes of service into queues representing different priority levels. The PQ algorithm schedule packets following a priority-based model, which means that a lower priority queue is scheduled if all queues of higher priority are empty, as illustrated in Figure II.8-c.

d) **Weighted Fair Queuing (WFQ):** is a straightforward extension of FQ, where instead of giving each class an equal access to resources, the WFQ scheduling assigns resources to each class according to a weight that is a proportion of the interface rate, as illustrated in Figure II.8-d.

e) **Weighted Round Robin (WRR):** is a scheduling that may seem very similar to WFQ , but the difference between them is that WFQ can be seen as a "bit-by-bit" Round Robin scheduler, whereas WRR handles packets in each scheduling turn, as illustrated in Figure II.8-e.

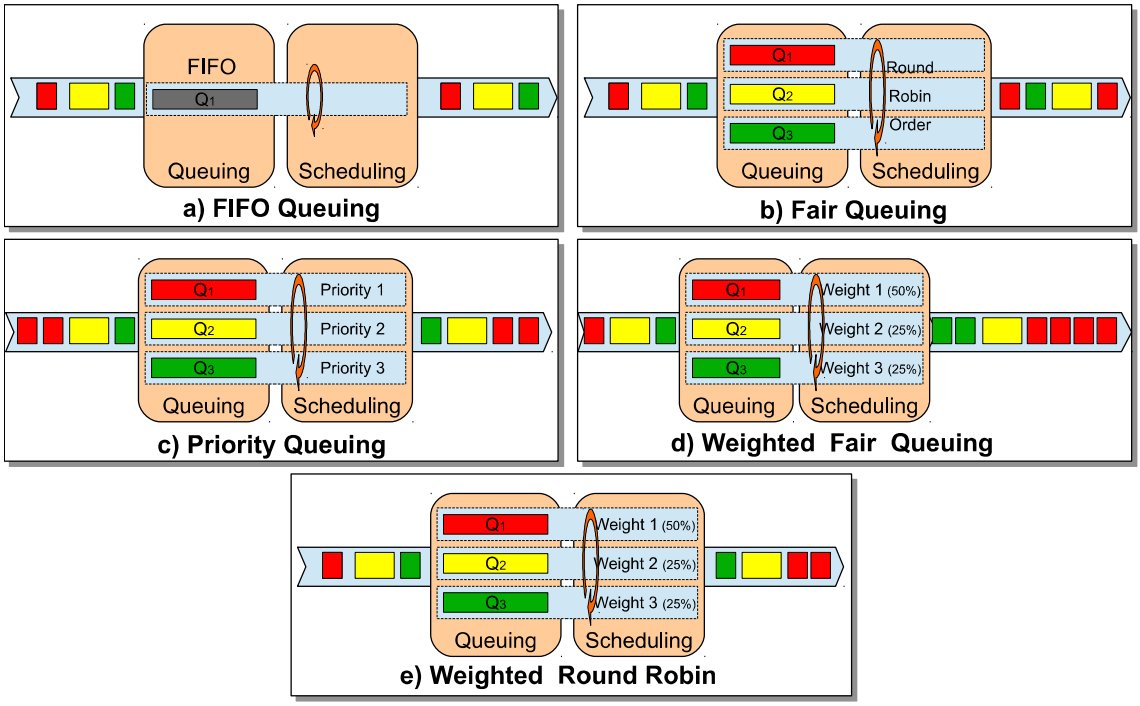


Figure II.8 – Examples of main queuing algorithms

3.2 DiffServ vs IntServ

The QoS management in the backhaul and access network is more classically performed on the basis of DiffServ model, which uses DSCP marking to determine the class service. Although more flexible and precise QoS management can be performed using a IntServ approach where individual flows are identified in order to allow resource reservation for flows based on a precise specification of their traffic characteristics and requirements.

While IntServ provides per-flow guarantees building a virtual circuit for each of them, DiffServ follows the philosophy of mapping multiple flows into a few service levels, as illustrated in Figure II.9.. The main issue of DiffServ model is how the traffic classes are defined. The reliability of DiffServ marking is an issue for any traffic from an unknown and uncontrolled source and this qualification counts for the large majority of downstream Internet traffic. Therefore, class of service differentiation can be more easily performed for the operator's own managed services or their partners. For example, when this definition is under operator control, the tendency is to give priority to managed services like Voice over IP (VoIP) and streaming TV with unidentified traffic to and from the Internet confined to residual capacity.

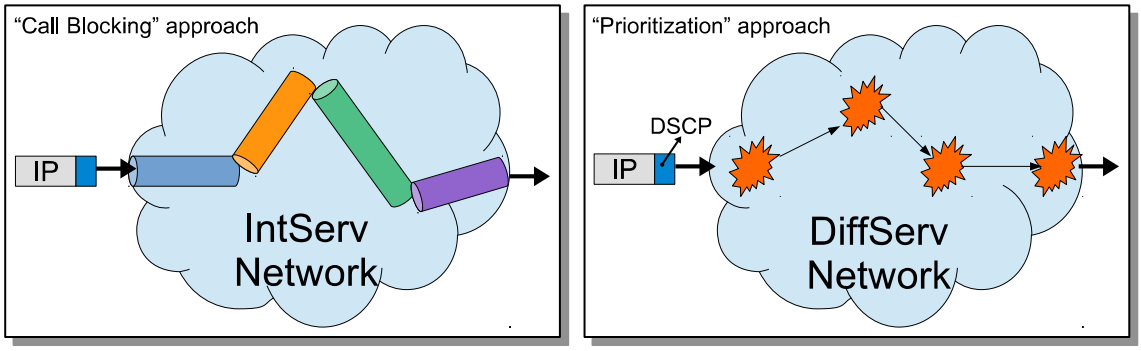
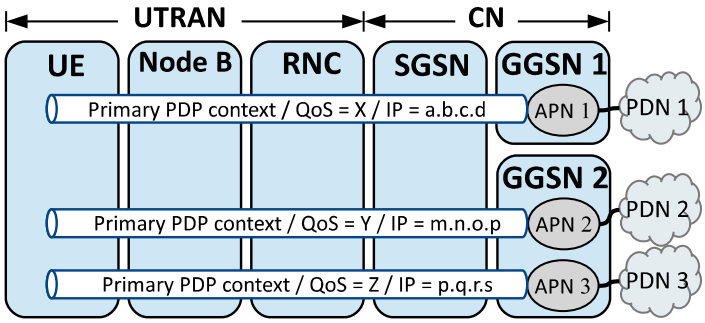


Figure II.9 – IntServ vs DiffServ

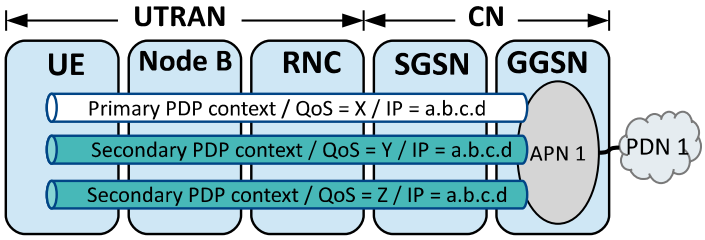
4 Overview of QoS in 3GPP Mobile Networks

4.1 Where does 3GPP QoS come from?

In 2000, 3GPP UMTS (Release 99) made his apparition, proposing a high data rate compared to previous standard. Which consists of a CS domain for voice and a PS domain for data. This was the first standard that promised QoS provisioning in PS. In this sense a user is assigned dedicated radio resources for PS data that are permanently available through a virtual tunnel called bearer. This bearer is a virtual tunnel established from the terminal to the PS core (i.e. SGSN and GGSN) and is called Packet Data Protocol (PDP) context.PDP context procedures and specifications are described in [39].



(a) Multiple primary PDP contexts



(b) Secondary PDP contexts

Figure II.10 – Packet Data Protocol (PDP) context

When a terminal sets up a PDP context, the terminal must specify its characteristics as End Point and a QoS profile. In that case this PDP context is called Primary PDP context. The End Point is a particular GGSN identified by the APN, which provides access to a particular PDN (i.e. Internet access). One GGSN may provide different services, which can be accessed by different APNs. Each PDP context request contains a QoS profile. However, the resulting QoS is negotiated according to the operator policies.

When a terminal wants to perform a data transfer, first a PDP context should be activated. During the PDP context activation, terminal is allocated a PDP address (e.g. IPv4 or IPv6). This address remains unchangeable for the duration of the PDP context. Once the PDP context is activated dedicated, radio resources are assigned to the terminal. Those resources are used for transmission of both user traffic and signaling. Furthermore, GTP tunnels are established through all nodes between the RNC to the PS core network (i.e. GGSN).

3GPP specifies the Secondary PDP Context [39], which is always associated with a primary PDP context and use the same IP address and same End Point (i.e. same APN). The only difference lies in the QoS profile. It is also important to highlight that QoS of any active primary or secondary PDP context can be modified with the PDP context modification procedure specified in [39]. Specifications also describe that multiple PDP contexts can be established by a terminal:

- **Multiple primary PDP contexts**: provides connections to different APNs (Figure II.10a)

- **Secondary PDP contexts**: provides connections to the same APNs but with different QoS profiles (Figure II.10b)

Nevertheless, UMTS was limited due its performance in terms of data rate, delay and capacity due to radio access technology based in Code Division Multiple Access (CDMA) in addition to its complex architecture. In 2004, 3GPP launches the task group for a new mobile system called Long-Term Evolution (LTE), with the promise of higher data rate, less delay and more capacity; all this based in a simple and scalable architecture. This standard evolution of mobile system is based on all-IP architecture with a unique PS domain for voice and data both, in contrast to the older standards. Moreover, current 3GPP specification aims to offer a fine-tuning QoS management. Nevertheless, this new QoS model inherits many characteristics of legacy standards, which is based in a virtual tunnel called EPS bearer. In fact, this EPS bearer is equivalent to PDP contexts of precedent UMTS system. The only real difference between both virtual tunnels is that in LTE the number of signaling needed to establish an EPS bearer have been reduced. The following is a more detailed description of LTE and its QoS architecture. You can also find more information about early 3GPP QoS standards in [40, 41].

4.2 3GPP LTE/EPS QoS Architecture

As mentioned in chapter I, the main element of 3GPP LTE QoS architecture is a virtual circuit called "bearer". It is a virtual circuit established between the P-GW adn the UE. Bearers provide an end-to-end transport service with specific QoS attributes, predetermining how user traffic flows are treated when it traverses the LTE network. This architecture has not significantly changed from its definition in Release 99 [39]. 3GPP describe the QoS concept and architecture for LTE/EPC system in release 8 [42].

In order to support multiple QoS requirements, different bearers can be set up by each UE. Each EPS bearer (GBR or Non-GBR) is associated with a QoS profile defined mainly by the following parameters [10]:

- **QoS Class Identifier (QCI)** is scalar parameter, which is characterized by priority, packet delay budget and acceptable packet loss rate. 3GPP specify 9 QCIs in order to ensures that operators can expect an uniform traffic handling throughout the network regardless of the equipment manufacturer. QCIs information such as the priority and packet delay budget from can be used to determined the RLCs mode configuration. The set of standardized QCIs [43] and their characteristics is provided in Table II.1.
- **Allocation and Retention Priority (ARP)** is used for call admission control in case of radio congestion. It also used in case of new bearer establishment request when the eNB is congested, in order to prioritization with respect to a established bearer. It is important to note that ARP does not have any impact on the prioritization of packet treatment over an established bearer.

- **Aggregated Maximum Bit Rate (AMBR)** is the total amount of bit rate of a group of non-Guaranteed Bit Rate (GBR) bearers. 3GPP defines 2 types: APN-AMBR and UE-AMBR. APN-AMBR limits the aggregated bit rate that can be provided to all non-Guaranteed Bit Rate (GBR) bearers of the same APN. UE-AMBR limits the aggregated bit rate that can be provided to all non-Guaranteed Bit Rate (GBR) bearers of a UE.

QCI	Resource Type	Priority	Packet Delay Budget	Packet Error Loss Rate	Example Services
1	GBR	2	100 ms	10e-2	Conversational voice
2		4	150 ms	10e-3	Conversational video (live streaming)
3		3	50 ms	10e-3	Real-time gaming
4		5	300 ms	10e-6	Non-conversational video (buffered streaming)
5	Non-GBR	1	100 ms	10e-6	IMS signaling
6		6	300 ms	10e-6	Video (buffered streaming) TCP-based (e.g., www, e-mail, chat, FTP, P2P file, sharing, progressive video, etc.)
7		7	100 ms	10e-3	Voice, video (live streaming), interactive gaming
8		8	300 ms	10e-6	Video (Buffered Streaming) TCP-based (e.g., www, e-mail, chat, FTP, P2P file, sharing, progressive video, etc.)
9		9			

Tableau II.1 – Standardized QCI characteristics

LTE/EPC system specify two kinds of bearers: default EPS bearer and dedicated EPS bearer [8]. Default EPS bearer, as previously mentioned, is establish when an UE is attached to the network and provides a best effort service. On the other hand, dedicated EPS bearer provides a dedicated virtual tunnel to one or a set of flows. Dedicated bearer is an additional bearer on top of default bearer. It does not require separate IP address due to the fact that the IP address allocated for the default bearer shall also be used for the dedicated bearers within the same PDN connection, similar to second PDP context of UMTS. Therefore, a dedicated bearer is always linked to one default bearer established previously. Every dedicated EPS bearer is associated with a packet filtering based on TFT to give special treatment to specific flows. The TFTs uses IP header information (i.e. source and destination IP addresses, source and destination port range, ToS) to filter packets and to put them on their corresponding bearers. In Uplink TFT is performed by the UE and in Downlink TFT is performed by the P-GW.

An EPS bearer has to cross multiple interfaces, across each of them it is mapped to its respective local bearer, which count with a unique ID as is shows in Figure II.11. Each node must keep track of the correspondence between the bearer IDs across its different interfaces. S1 and S5/S8 bearers are identified by the GTP tunnel endpoints and the IP address (source/destination TEID, source/destination IP address). The S-GW stores a

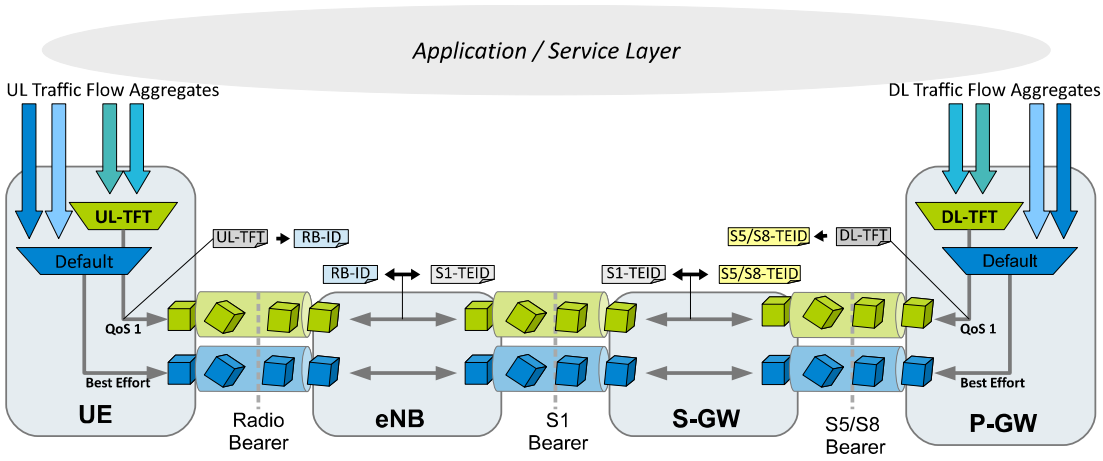


Figure II.11 – Two EPS bearers across the different interfaces *(Source: [44])*

one-to-one mapping between an S1 bearer and an S5/S8 bearer, in its turn, the eNB stores a one-to-one mapping between a radio bearer ID and its corresponding S1 bearer.

The eNB maps the EPS bearer QoS to the radio bearer QoS and then use RRC signaling to setup the radio bearer to UE. The RRC signaling contain all the parameters needed for the radio interface configuration. The RRC signaling contain all the needed parameters for the radio interface configuration, mainly for the configuration of the layer 2 (i.e. PDCP, RLC and MAC parameters). RRC signaling also contain the layer 1 parameters required for the UE to initialize the PHY protocol stack. Note that the radio scheduler plays a crucial role on the QoS over the radio segment, and consequently, it dramatically impacts the end-to-end customer experience. However, 3GPP specifications do not define any scheduling algorithm, leaving open its design and implementation. Various examples of radio scheduling algorithms can be found in the literature [45] [19]. However, vendors generally implement Proportional Fair (PF) algorithms. Such algorithms propose a trade-off between cell throughput optimization and fairness between the UEs, which is detailed in [46]. Following section details the bearer procedures.

The EPS also supports the transport of traffic flow aggregate(s), consisting of one or more Service Data Flows (SDFs). The concept of SDF is defined in the context of Policy and Charging Control (PCC) [43], which will be addressed later on.

4.2.1 3GPP QoS Procedures

Dedicated Bearer Setup: As was described at the beginning of this section, the QoS model in LTE has inherited many characteristics of UMTS. In this sense, the dedicated bearer is quite similar to secondary PDP context, in fact the only real difference between both is that LTE system reduces the number of signaling messages needed over the air interface. Dedicated bearer setup is always initiated by EPC, who triggers setup signaling

due to incoming packet connection. The dedicated bearer establishment procedure is detailed below:

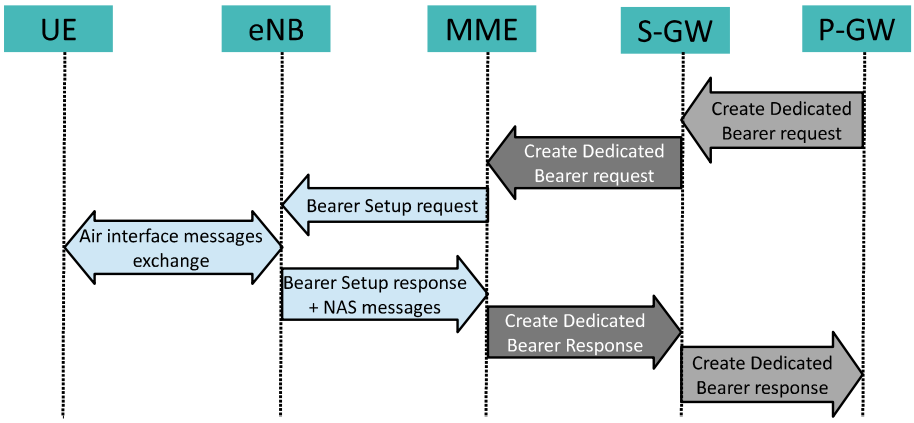


Figure II.12 – Dedicated Bearer Setup Messaging

1. The P-GW uses the QoS Policy values (i.e. QCI, ARP, GBR and MBR) to assign to the EPS Bearer QoS. The P-GW sends a *Create Bearer Request* message to the S-GW with the necessary information as IMSI, EPS Bearer QoS, TFT, P-GW S5/S8 TEID, Charging ID, default EPS Bearer Identity, etc.

2. The S-GW re transmits the Create Bearer Request message to the MME with adding the S-GW S1 TEID.

3. The MME assigns a unique EPS Bearer Identity to the UE and builds a Session Management Request (used at NAS layer), which is encapsulated into a *Bearer Setup Request message*. This message includes EPS Bearer ID, EPS Bearer QoS, *Session Management Request* and S1 TEID. Then, it is sends to the eNB.

4. Air interface messages exchanges. The eNB maps the EPS Bearer QoS to the Radio Bearer QoS (this mapping is defined by operators) and signals an *RRC Connection Reconfiguration* message to the UE which includes received information (i.e. EPS Radio Bearer ID, Radio Bearer QoS, Session Management Request). On NAS Layer level, the UE stores the EPS Bearer ID and links the dedicated bearer to the default bearer. The UE makes use of received Uplink TFT parameters in order to determine the mapping of traffic flows to the corresponding radio bearer. Then, the UE acknowledges the radio bearer activation to the eNB using an *RRC Connection Reconfiguration Complete* message.

5. The eNB acknowledges the bearer activation to the MME using the *Bearer Setup Response* message which includes the EPS Bearer ID and S1 TEID. The eNB also sends an Uplink NAS Transport (*Session Management Response*) message to the MME.

6. The MME acknowledges the bearer activation to the S-GW by sending a *Create Dedicated Bearer Response* (EPS Bearer Identity, S1-TEID) message.

7. Finally, the S-GW acknowledges the bearer activation to the P-GW by sending a *Create Dedicated Bearer Response*.

Dedicated Bearer Deactivation Once a flow transition through a dedicated bearer is finished, a procedure of bearer deactivation is trigger by the P-GW as is shown in Fig. II.13. This procedure has the same messages exchanged during the dedicated bearer activation, which contain necessary information to release resources allocated to the dedicated bearer.

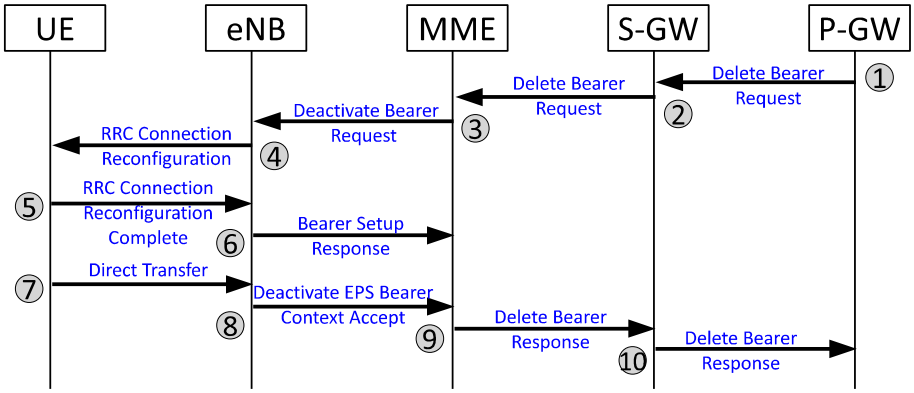


Figure II.13 – P-GW: Dedicated Bearer Deactivation Call Flow

Furthermore, in case where the UE triggers the bearer deactivation, he should request a dedicated bearer deactivation to the MME. Then, the MME requests a Dedicated Bearer deactivation to the S-GW. The S-GW sends the Bearer Resource Command to the corresponding P-GW as is shown in Fig. II.14. Finally, the P-GW use the procedure described above.

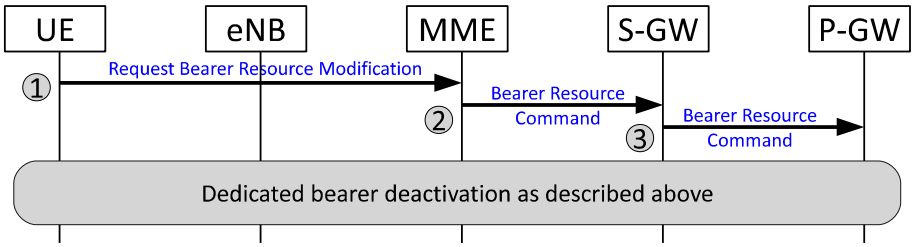


Figure II.14 – UE: Dedicated Bearer Deactivation Call Flow

4.2.2 Discussion on Operators' Point of View

The most precious resource in wireless networks is the bandwidth (spectrum). For operators, bandwidth represents a relevant part of CAPEX. In this respect, from the point of

view of operators an optimal radio scheduling algorithm must aim to maximise revenues. In other words, the monetization of each radio resources must be maximized.

For example, if it is more profitable for an operator to sell mobile contracts or provide a paid service guaranteeing/needing a minimum data rate, it is logical that the best radio scheduling algorithm is one which meet this objective.

On the other hand, the QoS can be seen as a competitive differentiation and not as a product. In this context, it is important to operators to invest in a cost-effective infrastructure to provide QoS at the lower price. According to [47], network quality and coverage is rated as the most important form of competitive differentiation by 75% of operators.

In this sense and face to current reality where telcom infrastructure costs are rising and operators profits are plummeting it seems urgent and necessary to explore alternative QoS models. These new QoS models aim to provide a competitive differentiation in markets where monetization of QoS is not possible (e.g. most of european countries)

4.3 Introduction to 3GPP Policy and Charging Control

LTE can make use of an extensive policy management architecture that provides operator with fine-grained control over users and services. This is integrated, via standardized interface, to online and offline charging system specified initially in R7 [43, 48] by 3GPP and called Policy and Charging Control (PCC). This architecture consists mainly of the Policy and Charging Enforcement Function (PCEF), the Policy and Charging Rules Function (PCRF), the Application Function (AF), the Online Charging System (OCS), the Offline Charging System (OFCS) and the Subscription Profile Repository (SPR); and is shown in Figure II.15.

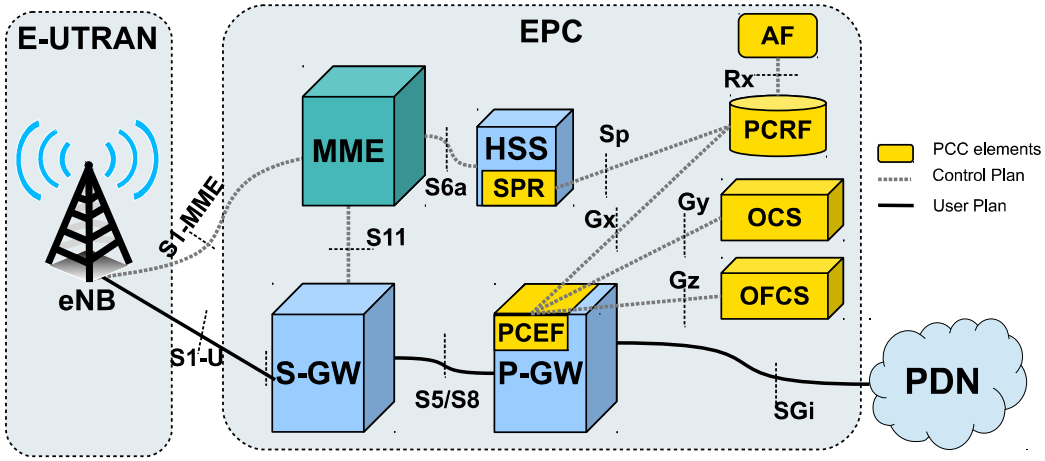


Figure II.15 – Policy and Charging Control (PCC) logical architecture R8

On PCC architecture, user traffic is classified into SDF traffic and EPS bearer traffic.

An SDF refers to a group of IP flows associated with a service that a user is using, while an EPS bearer refers to IP flows of aggregated SDFs that have the same QoS class. The SDF and EPS bearer are detected by matching the IP flows against the packet filters (SDF templates for SDFs or TFTs for EPS bearers). An SDF aggregate refers to a group of SDFs which have the same QCIs and ARPs and belong to one EPS session. In addition to these two basic parameters, there are other QoS parameters, such as GBR, MBR and AMBR that specify the bandwidth characteristics of SDFs and EPS bearers. In resume, SDF and EPS bearer QoS parameters are as follows:

- **SDF QoS parameters:** QCI, ARP, GBR and MBR
- **EPS bearer QoS parameters:** QCI, ARP, GBR, MBR, APN-AMBR and UE-AMBR

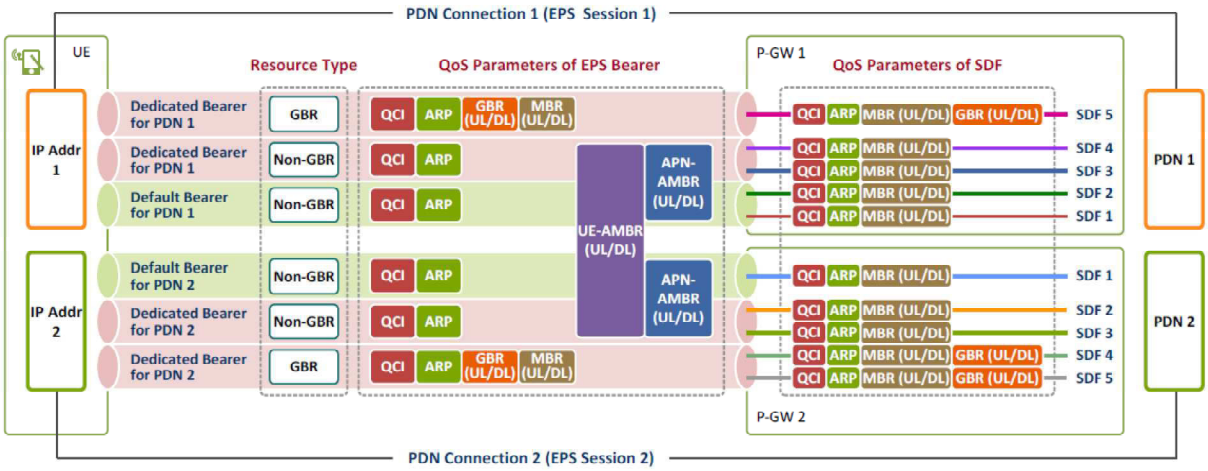


Figure II.16 – QoS Parameters for SDF and EPS Bearer *(Source: [49])*

PCC enables a centralized control to ensure that the service sessions are provided with appropriate bandwidth and QoS. PCC also provides a means to control charging on a per-service basis. When an EPS session is established or modified, PCRF determines a PCC rule for each SDF based on the operator's policy (e.g. QoS policy, charging methods). PCEF (P-GW) detects an SDF, and applies a PCC rule that is specific to the particular SDF to the user packets in it. It also binds the SDF QoS and bearer QoS, and applies the bearer QoS to the EPS bearer. That is, EPS bearer contexts are set or modified at LTE entities (i.e. UE, eNB, S-GW, P-GW, and MME). The purpose of the PCC rule is listed below and there exist two types. Dynamic PCC rules are dynamically provisioned by PCRF to PCEF, and pre-defined PCC rules are pre-configured in the PCEF.

- Detect a packet belonging to an SDF to map that packet to proper EPS bearer in downlink and uplink direction;

- Identify the service;
- Provide appropriate applicable charging;
- Provide policy control

This approach for fine tuning of the QoS management provides network operators tools to create much richer policy sets for charging and enforcement solutions that go beyond of the 3GPP standards solutions , but at the same time this increases complexity of the network and require an non negligible investment.

4.4 Identification of Critical Elements in the QoS Provisioning

As described in Chapter I, the LTE/EPC network is composed of a Core Network (i.e. EPC) and a Radio Access Network (i.e. E-UTRAN). The identification of the critical elements in the QoS provisioning in both Core Network and Radio Access Network are outlined below.

4.4.1 Evolved Packet Core

In case of the EPC, the P-GW plays a key role on QoS management since it is responsible for bearer establishments and QoS enforcement with the help of PCRF on PCC architecture. The P-GW acts as the interface between the LTE/EPC network and other PDNs, which means that P-GW is the entry point to all established EPS bearers in the LTE/EPC network. Therefore, the P-GW is charged to place arrived IP packets into their corresponding EPS bearer, according to predefined rules.

In spite of critical function of P-GW, generally, all EPC network elements (i.e P-GW, S-GW, MME) are extremely robust and reliable. Furthermore, operators almost always over-dimensioning EPC [50]. Consequently, the EPC is rarely a serious problem on QoS provisioning.

4.4.2 Evolved Universal Terrestrial Access Network

It is very evident that providing QoS in wireless networks is much harder than providing QoS in wireline networks. Indeed one of the biggest difference between wireline and wireless networks is the transmission link variability, which in case of most of wireless networks technologies is often a bottleneck.

This is largely due to that the wireless channels suffer radio propagation phenomenons as interference, fading and shadowing. As a result, the capacity of a wireless links have

very high variability. In addition, the reality of a shared and often congested radio network, make it a real challenge guaranteeing a minimal level of QoS [51].

In case of LTE/EPC, its radio interface is part of E-UTRAN. And besides the time/location dependent problem of all wireless networks, the LTE radio capacity is also users-dependent. Therefore, at a particular time instance, an eNB can communicate with more than one UE simultaneously, which means that the eNB should share available radio resources between all connected UEs (i.e. radio scheduling).

Consequently, we can identify two main factors that have relevant impact in QoS in mobile networks, the quality of radio link and the competition for radio resources among UEs [52].

Furthermore, LTE/EPC network is expected to support a wide range of services (e.g. multimedia, real-time, web) in different scenarios (e.g. urban, high mobility). Therefore, multi-user scheduling is one of the key feature in LTE because, according 3GPP QoS strategy, radio scheduler is in charge to satisfy different requirements of QoS levels through the distribution of available radio resources among active users. The radio scheduling design is a very wide topic and has been studied in depth, various examples can be found in the literature [15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 45].

In this sense and in line with the objective of this thesis, we study the role and strategies of radio schedulers in QoS provisioning. Therefore, we can classify radio schedulers according its strategy into two large groups:

1. **Meeting Collective QoS Requirements:**

   This category of schedulers aims to meet collective QoS requirements, such as maximize throughput, reduce mean delay or both. Some examples of this category are:

   a) Maximum Throughput (MT)

   b) Proportional Fair (PF)

   c) Throughput to Average (between MT and PF)

2. **Meeting Individual QoS Requirements:**

   This category of schedulers aims to meet individual QoS requirements, such as guaranteed a minimum throughput to a group of users or a class of service. Some examples of this category are:

   a) Weighted Fair Queuing (WFQ)

   b) QoS-aware schedulers listed in [19]

5 Conclusions

In the PCC architecture, flows (or services comprising several flows) are explicitly signalled. Flow set up and control rely on a series of operations and functions whose detailed realization is specified in the standard (see Figure II.15). For QoS control, flows declare explicit traffic characteristics and performance requirements and functions within the PCC architecture have the task of performing admission control and reserving sufficient network resources. Admission control and the evaluation of required capacity for each flow are not defined in the standards. This is straightforward for flows with constant data rate but problematic when the data rate is variable (e.g. Web Real-Time Communication (WebRTC), HTTP Adaptive Streaming (HAS)).

IntServ is the QoS architecture initially proposed by IETF. IntServ has not been implemented mainly because it is considered too complex, as described above. In fact, IntServ is not more complex than PCC (or 3GPP QoS model generally speaking) and could be implemented in the backhaul and access networks. However, the traffic controls envisaged for IntServ as well as 3GPP QoS model are arguably not well adapted to the nature of current Internet flows.

The following chapters aim to evaluate the cost of 3GPP QoS based in study of signaling, context cost and others cost factors in order to propose an alternative QoS model. This QoS model is called IP-centric because QoS is managed at IP level. Finally, a mechanism called Slo-Mo is proposed, which aims be a solution for the transition from 3GPP QoS model to IP-centric, when many constructors equipments are deployed by an operator and not all of them support IP-centric feature.

Chapter III

The Cost of QoS Management in 3GPP Mobile Networks

1 Introduction

The disruptive evolution of mobile usage and services (e.g. Voice over LTE (VoLTE), Video over LTE (ViLTE)) results in an important challenge for operators who struggle to differentiate themselves from competitors. In this respect, the QoS seems to be the best way forward, but at what cost?

The management of QoS in LTE/EPC systems is clearly connection-oriented as it is based on virtual circuits called EPS bearers. Let us recall that when a UE attaches to the network, a default bearer with a "Best Effort" QoS is established. Other bearers can be further set up, one per QoS level. Bearers are operated in connected mode, that is, established, or disconnected via signalling protocols, as described in chapter II .

In this chapter we present a cost analysis of QoS in LTE/EPC mobile networks. We develop an analytical model and define different types of cost as *Context Load*, *Processing Load*, *Memory Access Rate* and *Radio Signaling Overhead* in order to evaluate the impact of the standardized QoS features. These four metrics are defined below. Simulation results based on realistic values obtained through measurements on Orange networks are also presented.

Through carrying out different analyses, we can investigate the weak points of the 3GPP QoS model. We are then able to draw out conclusions and perspectives, in order to propose enhancements of the standardized scheme or even to introduce a novel QoS model. Much of the content of this chapter has been published [53, 54].

1.1 Related Work

Many studies have proposed evaluations of signaling load, together with various analytical models.

For example, in [55] the authors evaluate the signaling load associated to various MME architectures (distributed and centralized) and propose a multicast paging procedure to alleviate the MME signaling load.

In [56, 57], the authors propose a mathematical framework for analyzing and evaluating the signaling load due to authentication. They finally propose an analytical method to optimize the frequency of security-key updates due to handover procedures.

In [58] the authors evaluate the signaling load in order to quantify its impact on the energy consumption. In [59] the authors propose an analytical model for comparing the mobility management performance in terms of signalling cost.

In [60] the authors compare the signaling loads of 3GPP Release 99 and Release 5 network architectures. And show that the signaling loads for Release 5 are in general much higher than the correspondent loads for Release 99.

Nevertheless, none of these studies addresses the impact of the QoS model defined by the 3GPP standards. Consequently, in this chapter we present a wide cost factor analysis of the 3GPP QoS model in LTE/EPC mobile networks.

2 Cost Factor Analysis

The QoS model in LTE [8] has inherited many characteristics of UMTS. In fact the only real difference between them is that LTE systems reduce the number of signaling messages needed over the air interface.

In order to quantify the cost of QoS in LTE/EPC networks, we define the *Processing Load*, the *Context Load*, the *Memory Access Rate* and *Radio Signaling Overhead*, in the following sections. The defined costs aim to evaluate the 3GPP QoS scheme from the point of view of the Control Plan since it is a relevant drawback of the circuit oriented model. Therefore, we only take into consideration of mobile data traffic information which have impact on Control Plan as the session duration and arrival rate from Orange network statistics; and not the amount of exchanged data traffic.

2.1 Processing Load

The *Processing Load* (S) is defined as the average number of incoming signalling messages per unit time on an LTE/EPC equipment x. This cost factor is referred to as Processing

Load as each incoming message generates processing on the related LTE/EPC equipment x. For example, when a LTE/EPC equipment generates a new message or creates, modifies or releases a context. It is measured per unit time.

2.2 Context Load

The *Context Load* (D) is defined as the average number of simultaneous active bearers on an LTE/EPC equipment x. This metric represents the measure of memory occupancy as each active bearer is associated with an entry for its context. As the context load grows, memory overflow issues may appear together with increased latency of memory's lookups.

2.3 Memory Access Rate

In addition to the loads defined above, we consider the *Memory Access Rate* (L), which evaluates the average number of memory accesses per unit time due to creation, modification or release of contexts, which is evaluated on an LTE/EPC equipment x. The number of memory's lookup per unit time increases with the number of contexts (default and dedicated) and this metric should be carefully considered in the design and sizing of LTE equipment.

2.4 Radio Signaling Overhead

An other relevant cost factor presented in this section is the *Radio Signaling Overhead* (Z). It is defined as the amount of signaling messages in bytes exchanged at the radio interface per unit time, in both Uplink and Downlink. The number of established dedicated bearers strongly contributes to the *Radio Signaling Overhead* when the handover procedure is performed since information on each dedicated bearer must be exchanged. In short, the *Radio Signaling Overhead* of "n" dedicated bearers is equivalent to "n" times the *Radio Signaling Overhead* of a default bearer.

3 Analytical Model Description

In order to facilitate the readability of this section a List of Symbols is available in Table III.1, which summarize the models parameters defined hereafter.

From this point forward, it is assumed that all UEs are already registered in the LTE/EPC network (EMM-REGISTERED), which is the case of most of terminals in real world. As is detailed in chapter I, when a user turns on his UE, it establish a RRC

Symbol	Description
A_{c}	Average coverage area of eNBs
A_t	Area of the evaluated region
A_x	Area served by the LTE/EPC equipment x
B_d^y	Amount of data generated by signaling messages exchanged on the radio interface during procedure y in d direction (i.e. UL or DL)
C_{e}	Total number of eNBs in the evaluated region
C_{ta}	Number of eNBs per Tracking Area List (TAL)
D_x	Context Load for an LTE/EPC equipment x
I_x^y	Number of context creation, modification or release in element x during procedure y
l	Perimeter length of an enclosed region (e.g. eNB, S-GW)
L_x^y	Average number of context creation, modification or release due to procedure y in element x
m	Number of applications using the default bearer
M_x^y	Number of incoming signaling messages to element x during procedure y
n	Total number of applications running in a UE
N_x^{enb}	Average number of eNBs served by an equipment x
N_x^{ta}	Number of TAs served by an equipment x
P_{ue}	Probability that a session is originated by the UE
P_{rel}	Probability that a handover event involves a S-GW relocation
r	eNB cell radius
S_x^y	Average number of incoming signals per unit time due to procedure y in element x
T_0	Duration of a period of total inactivity
T_1	Duration of a period of total activity
$\overline{T}_i^{\mathrm{on}}$	Average time where a dedicated bearer used by type-*i* application is enabled
$\overline{T}_0$	Average of T_0
$\overline{T}_1$	Average of T_1
V	Average speed of UEs
Z_d^y	*Radio Signaling Overhead* in d direction (i.e. UL or DL) due to procedure y
α_i	Average arrival rate of type-*i* session
β	Average number of transitions from CONNECTED to IDLE states per unit of time
γ	Overlapping factor
λ_i^{-1}	Average duration of the OFF state of type-*i* session
λ_{ue}	Arrival rate of UEs in an eNB area
μ_i^{-1}	Average type-*i* session duration (ON state)
$\pi_{\mathrm{connected}}$	Probability that a UE is in CONNECTED state
π_{idle}	Probability that a UE is in IDLE state
$\pi_{\mathrm{s},i}$	Probability that the type-*i* session state is s (s=ON or s=OFF)
ρ	UE density in number of UEs per unit area
σ_{c}	Arrival rate of a context in an eNB
τ	RRC Inactivity timer
υ	Rate of border crossings per UE for a given eNB
ϕ	Duration that the dedicated bearer waits to receive a data packet before timing out
χ_i	Probability that a type-*i* session is originated by a UE
Ω_j	Average number of transition from ON to OFF states per unit time of a dedicated bearer carrying application type-*j*

Tableau III.1 – List of Symbols

connection to network in order to subscribe to LTE/EPC network, which means that it transits from EMM-DEREGISTERED to EMM-REGISTERED state.

We assume that each UE is a multitask terminal, capable of supporting n different applications (e.g. voice, video streaming, web, etc.), which can be either originated by itself or by its peer (another UE or a server). Let χ_i be the probability that a type-*i* session is originated by a UE.

3.1 Application Model

Each application is modeled as a random ergodic ON-OFF process as shown in Figure III.1. The average type-*i* session duration (ON state) is denoted by μ_i^{-1} and the average duration of the OFF state of type-*i* session is denoted by λ_i^{-1}. The average arrival rate α_i of type-*i* sessions is thus:

$$\alpha_i = 1/\left(\lambda_i^{-1} + \mu_i^{-1}\right) \tag{III.1}$$

Let $\pi_{s,i}$ be the probability that the type-*i* session state is s ($s =$ ON or $s =$ OFF). The stationary probability of an application state is independent of time and thus satisfies the global balance conditions $\lambda_i \pi_{\text{off},i} = \mu_i \pi_{\text{on},i}$ and $\pi_{\text{off},i} + \pi_{\text{on},i} = 1$ (ergodic markov chain). We have thus:

$$\begin{cases} \pi_{\text{off},i} = \frac{\mu_i}{\lambda_i + \mu_i} \\ \pi_{\text{on},i} = \frac{\lambda_i}{\lambda_i + \mu_i} \end{cases} \tag{III.2}$$

3.2 User Equipment Model

We now consider that all applications are running on the same UE. Let T_0 be the duration of a period of total inactivity (state 0 = no active application) and T_1 be the duration of a period of activity (state 1 = at least one application active) on this terminal. We only assume that the duration of the OFF state for application i denoted by X_i is an exponential random variable. Hence, $P(X_i > t) = e^{-\lambda_i t}$ and we have:

$$\Pr(T_0 > t) = \prod_{i=1}^{n} \Pr(X_i > t) = e^{-(\sum_{i=1}^{n} \lambda_i)t} \tag{III.3}$$

Let $\overline{T}_0$ and $\overline{T}_1$ be respectively the average of T_0 and T_1. We have:

$$\overline{T}_0 = 1/\sum_{i=1}^{n} \lambda_i \tag{III.4}$$

The stationary probability π_0 that all applications are inactive can be expressed as follows:

$$\pi_0 = \frac{\overline{T}_0}{\overline{T}_0 + \overline{T}_1} \tag{III.5}$$

Based on the assumption that considered ON/OFF processes are ergodic (see Appendix B), the probability that all applications are inactive can be expressed as:

$$\pi_0 = \prod_{i=1}^{n} \pi_{\text{off},i} \tag{III.6}$$

and combining equations (III.6) and (III.2) we get:

$$\pi_0 = \prod_{i=1}^{n} \frac{\mu_i}{\lambda_i + \mu_i} \tag{III.7}$$

Let τ be the RRC Inactivity timer, which is the max period of data inactivity before to switch a UE from CONNECTED state to IDLE state. Figure III.1 shows the applications states together with the UE states (IDLE/ CONNECTED).

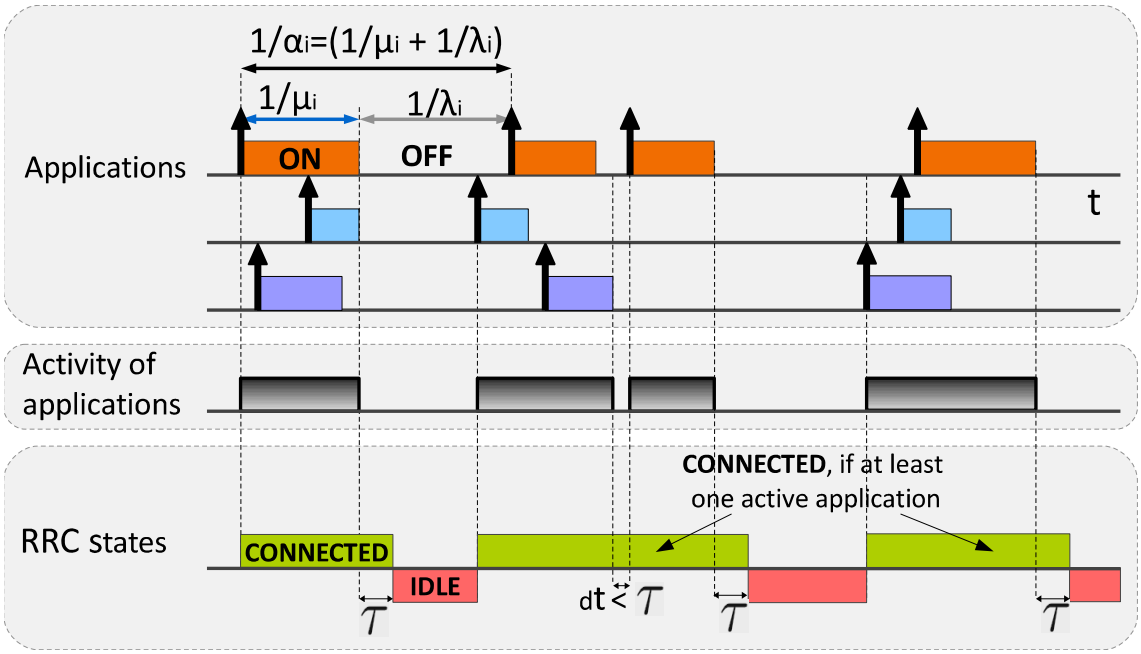


Figure III.1 – Modeled RRC States

The probability π_{idle} that a UE is in IDLE state is $\pi_{\text{idle}}=\Pr(T_0 > \tau)\pi_0$. Combining

equations (III.3) and (III.7) we get:

$$\pi_{\text{idle}} = \prod_{i=1}^{n} \frac{\mu_i}{\lambda_i + \mu_i} e^{-(\sum_{i=1}^{n} \lambda_i)\tau} \tag{III.8}$$

3.3 Context Time Duration

Let T_{idle} be the time during which a UE is in IDLE state and $T_{\text{connected}}$ be the time during which a UE is in CONNECTED state. We have:

$$\Pr(T_{\text{idle}} > t) = \Pr(T_0 > t + \tau \mid T_0 > \tau) \tag{III.9}$$

Due to the memoryless propriety of exponential random variable T_0, we thus have $\Pr(T_{\text{idle}} > t)$=$\Pr(T_0 > t)$. From (III.3) we have thus:

$$\overline{T}_{\text{idle}} = 1/\sum_{i=1}^{n} \lambda_i \tag{III.10}$$

The stationary probability π_{idle} that a UE is in IDLE state can be also expressed as π_{idle}=$\overline{T}_{\text{idle}}/(\overline{T}_{\text{idle}} + \overline{T}_{\text{connected}})$, we thus have:

$$\overline{T}_{\text{connected}} = \overline{T}_{\text{idle}} \frac{1 - \pi_{\text{idle}}}{\pi_{\text{idle}}} \tag{III.11}$$

3.4 System Model

The average number of transitions from state 1 (at least one active application) to state 0 (no active application) is $\frac{1}{\overline{T}_0 + \overline{T}_1}$ per unit time. Let β be the average number of transitions from CONNECTED to IDLE states per unit of time, we thus have:

$$\beta = \frac{\Pr(T_0 > \tau)}{\overline{T}_0 + \overline{T}_1} \tag{III.12}$$

Combining equations (III.3), (III.4), (III.5) and (III.7) we get:

$$\begin{aligned} \beta &= \frac{e^{-(\sum_{i=1}^{n} \lambda_i)\tau}}{\overline{T}_0 + \overline{T}_1} = \frac{\pi_0}{\overline{T}_0} e^{-(\sum_{i=1}^{n} \lambda_i)\tau} \\ &= \sum_{i=1}^{n} \lambda_i \prod_{i=1}^{n} \frac{\mu_i}{\lambda_i + \mu_i} e^{-(\sum_{i=1}^{n} \lambda_i)\tau} \end{aligned} \tag{III.13}$$

Let P_{ue} be the probability that a session is originated by the UE. It may be estimated as the ratio between the average arrival rate of sessions originated by a UE ($\sum_{i=1}^{n} \alpha_i \chi_i$) and average arrival rate of all sessions ($\sum_{i=1}^{n} \alpha_i$) of a UE. We have thus:

$$P_{\text{ue}} = \frac{\sum_{i=1}^{n} \alpha_i \chi_i}{\sum_{i=1}^{n} \alpha_i} \tag{III.14}$$

3.5 Dedicated Bearer Model

As mentioned in chapter II.3, the bearer-inactivity timer ϕ is the maximum duration of data inactivity where the dedicated bearer is keeping connected before releasing it. [61, 62]. After timer ϕ expires, the P-GW takes down the dedicated bearer following the dedicated bearer deactivation procedure. Figure III.2 shows the modeled dedicated bearer Activation/Deactivation.

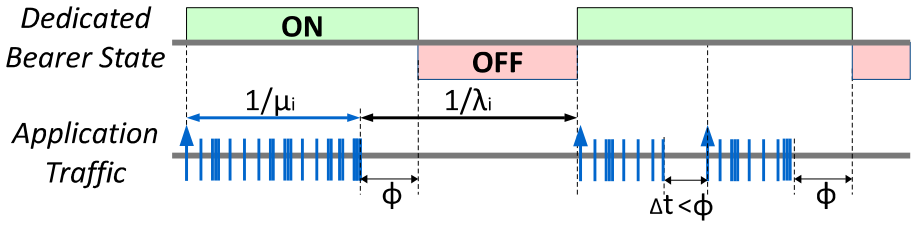


Figure III.2 – Modeled Dedicated Bearer Activation/Deactivation

Let j be an application using a dedicated bearer, which follows the defined applications model. Let Ω_j be the average number of transition from ON to OFF states per unit time of a dedicated bearer carrying application type-*j*. Using the same procedure as in the equation (III.12), we thus have:

$$\Omega_j = \frac{\lambda_j \mu_j}{\lambda_j + \mu_j} e^{-\lambda_j \phi} \tag{III.15}$$

Let $\overline{T}_j^{\text{on}}$ be the average time where a dedicated bearer used by type-*j* application is enabled. We can state as for equation (III.11) that $\overline{T}_j^{\text{on}} = \overline{T}_j^{\text{off}} \left(\frac{1-\pi_{\text{off},j}}{\pi_{\text{off},j}} \right)$, as for equation (III.10) that $\overline{T}_j^{\text{off}} = 1/\lambda_j$ and as for equation (III.8) that $\pi_{\text{off},j} = \frac{\mu_j}{\lambda_j+\mu_j} e^{-\lambda_j \phi}$. We have thus:

$$\overline{T}_j^{\text{on}} = \frac{e^{\lambda_j \phi}}{\mu_j} + \frac{e^{\lambda_j \phi} - 1}{\lambda_j} \tag{III.16}$$

3.6 Mobility Model

Let C_e be the total number of eNBs in the evaluated region, A_t be its total area, A_c be the coverage area of an eNB. In order to simplify our analysis, each eNB is represented by a cell, which is assumed to be a disk. Assuming uniform circular cells with an overlapping factor γ ($\gamma \geq 1$), the required cell radius, r, to cover the entire area is $r = \gamma\sqrt{A_t/(C_e\pi)}$. Let C_{ta} be the number of eNBs per Tracking Area List (TAL). In order to compute the messages load due to handover events we use the fluid-flow model [63] to determine the mobile crossing rate out of an enclosed region with perimeter length l. We assume that UEs have an average speed V. Based in [64], we estimate the rate of border crossings per UE for a given eNB coverage area as follows:

$$\upsilon = \frac{Vl}{A_c\pi} \tag{III.17}$$

Let P_{rel} be the probability that a handover event involves also a S-GW relocation. We assume that each S-GW serves a TAL [65], then P_{rel} can be well approximated by $1/\sqrt{C_{ta}}$.

Let ρ be the UE density (number of UEs per unit area). Then, the average number of customers in the eNB area can be computed as ρA_c.

Let λ_{ue} be the arrival rate of UEs in an eNB area. Each eNB area can be seen as an infinite capacity system with random arrival of customers (UEs) with rate λ_{ue} and service rate υ. Hence, using Little's law, we can write:

$$\lambda_{ue} = \upsilon\rho A_c \tag{III.18}$$

When a UE has active sessions, contexts are created in the eNB to which it is attached, one per DRB as shown in Figure I.9. In handover scenario, the related contexts of the EU must be transferred to its new eNB in order to have the continuity of the active sessions.

In steady-state, the arrival rate of UEs in the eNB area is also the departure rate of UEs from the eNB area. The probability that a UE has at least one new session (created in current eNB) at any time is $(1 - \pi_{idle})$. This is the same probability with which a UE carries at least one new session while departing from the eNB area. Let σ_c be the arrival rate of a context on an eNB, which can be computed as $\sigma_c = \lambda_{ue}(1 - \pi_{idle})$. Therefore, from (III.8) and (III.18) we have:

$$\sigma_c = \upsilon\rho A_c \left(1 - \prod_{i=1}^{n} \frac{\mu_i}{\lambda_i + \mu_i} e^{-(\sum_{i=1}^{n}\lambda_i)\tau}\right) \tag{III.19}$$

4 Analytical Model for Estimating the Cost Factors

We define the following symbols related to each proposed cost factor in order to measure each one of them in the related LTE/EPC elements.

- **Context Load Evaluation** D_x represents the average number of active context per unit time in an equipment x,
- **Processing Load** S_x^y represents the average incoming signal messages per unit time to element x due to procedure y,
- **Memory Access Rate** L_x^y represents the average number of context creation, modification or release per unit time in element x due to procedure y,
- **Radio Signaling Overhead** Z_d^y represents the average Radio Signaling Overhead in d direction (i.e. UL or DL) due to procedure y (measurable in the eNB).

4.1 Context Load Evaluation

Let A_x be the area served by the LTE/EPC equipment x. Let n be the total number of applications running in a UE and m the number of applications using the default bearer ($m \leqslant n$). Therefore, the Context Load for an LTE/EPC equipment x is computed as follows:

$$D_x = A_x \rho \left[\beta \overline{T}_{\text{connected}} + \sum_{i=m+1}^{n} \Omega_i \overline{T}_i^{\text{on}} \right] \tag{III.20}$$

In case of the P-GW and MME A_x=$A_{\text{c}}C_{\text{e}}$, for the S-GW A_x=$A_{\text{c}}C_{\text{ta}}$ and for the eNB A_x=A_{c}.

4.2 Processing Load, Memory Access Rate and Radio Signaling Overhead Evaluation

Concerning the *Processing Load* (S) evaluation, we complete the analysis proposed in [53], adding the mobility and related procedures. This analysis is based on some elements described in [55], as its signaling compute method, which have been enriched introducing our machine state model described above. We take into consideration mechanisms described in Section I and II, whose signaling call flows are detailed in [8, 9, 28, 66] and summarized in Table III.3.

We define variables related to Processing Load evaluation. Let M_x^y be the number of incoming signaling messages and I_x^y be the number of context creation, modification or release addressed to element x (e.g. MME, S-GW, eNB) during procedure y. Abbreviations of procedures that are take in consideration in our analysis are shown in Table III.2.

Abbreviation	Procedure description
ci	Switch to IDLE state
db	Dedicated Bearer Activation/Deactivation
db-net	Dedicated Bearer Activation/Deactivation initiated by the network
db-ue	Dedicated Bearer Activation/Deactivation initiated by the UE
ho	Handover
ho-nsr	Handover without S-GW relocation
ho-sr	Handover with S-GW relocation
sr	Service Request
sr-net	Service Request initiated by the network
sr-ue	Service Request initiated by the UE
tau	Tracking Area Update

Tableau III.2 – Abbreviations of LTE/EPC procedures

Concerning the *Processing Load* (S) and *Memory Access Rate* (L), the number of incoming signaling messages and context creation, modification or release needed for their evaluation are summarized in Table III.3.

Finally, we investigate the impact of the 3GPP QoS model on the LTE radio segment and evaluated it at eNB side.. For this purpose, we study the *Radio Signaling Overhead* based on [67].

Let B_d^y be the amount of data generated by signaling messages exchanged on the radio interface during procedure y. RRC messages size of Default/Dedicated bearer activation/deactivation and Handover procedures are shown in Table III.4.

4.2.1 Service Request procedure

The *Processing Load* on node x due to Service Request procedure is computed as follows:

$$S_x^{\text{sr}} = \beta A_x \rho \left[M_x^{\text{sr-ue}} P_{\text{ue}} + M_x^{\text{sr-net}} (1 - P_{\text{ue}}) + M_x^{\text{ci}} \right] \tag{III.21}$$

Furthermore, the *Memory Access Rate* is given by:

$$L_x^{\text{sr}} = \beta A_x \rho \left(I_x^{\text{sr}} + I_x^{\text{ci}} \right) \tag{III.22}$$

Procedures	Events	eNB	MME	SGW	PGW
Service Request *(sr-net/ue)*	Context Creation Context Release Context Modification	1	 1	 1	
	Incoming Messages	3/4*	3/4*	1/3*	0
Switch to IDLE state *(ci)*	Context Creation Context Release Context Modification	 1	 1	 1	
	Incoming Messages	2	3	1	0
Dedicated Bearer Activation *(db-net/ue)*	Context Creation Context Release Context Modification	1	1	1	1
	Incoming Messages	3/4*	3/4*	2/3*	1/2*
Dedicated Bearer Deactivation *(db-net/ue)*	Context Creation Context Release Context Modification	1	1	1	1
	Incoming Messages	3/4*	3/4*	2/3*	1/2*
Handover without S-GW relocation *(ho-nsr)*	Context Creation Context Release Context Modification	k k	 k		
	Incoming Messages	7	2	2	0
Handover with S-GW relocation *(ho-sr)*	Context Creation Context Release Context Modification	k k	 k	 k	
	Incoming Messages	7	3	3	1
Tracking Area Update *(tau)*	Context Creation Context Release Context Modification	 1	 1		
	Incoming Messages	1	2	0	0

k: number of active bearers in current time

** If communication is initiated by the UE*

Tableau III.3 – Context and signaling events details of relevant LTE/EPC procedures

Let B_x^{sr} be the amount of signaling messages exchange during the Default Bearer Activation/Deactivation. The *Radio Signaling Overhead* due to Service Request procedure for d direction is computed as follows:

$$Z_d^{\text{sr}} = \beta A_{\text{c}} \rho B_d^{\text{sr}} \tag{III.23}$$

DL/UL	Message	MAC PDU Size (bytes)	
		UL	DL
Default Bearer Activation/Deactivation			
DL	Random Access Response	–	8
UL	RRC Connection Request	7	–
DL	Contention Resolution CE	–	7
DL	RRC Connection Setup	–	30
UL	RRC Connection Setup Complete	20	–
DL	RLC Status PDU	–	3
DL	RLC Status PDU	–	3
DL	RRC Connection Reconfiguration	–	45
UL	RRC Connection Reconfiguration Complete	19	–
DL	RLC Status PDU	–	3
DL	RRC Connection Release	–	10
UL	RLC Status PDU	3	–
Total Bytes (B_d^{sr})		**49**	**109**
Dedicated Bearer Activation/Deactivation			
DL	RRC Connection Reconfiguration	–	118
UL	RRC Connection Reconfiguration Complete	10	–
UL	UL Information Transfer	13	–
DL	RRC Connection Release	–	10
Total Bytes (B_d^{db})		**23**	**128**
Handover Procedure			
UL	Buffer Status Report	2	–
UL	Measurement Report	2	–
DL	RLC Status Report	–	3
DL	RRC Connection Reconfiguration	–	87 + 73×k
DL	Random Access Response	–	7
UL	RRC Connection Reconfiguration Complete	13	–
DL	RLC Status Report	–	3
Total Bytes (B_d^h)		**34**	**100 + 73×k**

k: number of dedicated bearers in current time

Tableau III.4 – RRC Signaling Overhead *(Source: [67])*

4.2.2 Dedicated Bearer

Once a data transmission through a dedicated bearer has been completed, the release procedure is triggered after an inactivity time ϕ. Therefore, the *Processing Load* due to Dedicated Bearer Activation and Deactivation requested by the type-*i* application for x can

be computed as follows:

$$S_x^{\text{db}}(i) = \Omega_i A_x \rho \Big[M_x^{\text{db-ue}} \chi_i + M_x^{\text{db-net}} (1 - \chi_i) \Big] \tag{III.24}$$

Furthermore, the *Memory Access Rate* is given by:

$$L_x^{\text{db}}(i) = \Omega_i A_x \rho I_x^{\text{db}} \tag{III.25}$$

Let B_d^{db} be the amount of signaling messages exchange during the Dedicated Bearer Activation/Deactivation. The *Radio Signaling Overhead* due to Dedicated Bearer procedure for d direction is computed as follows:

$$Z_d^{\text{db}}(i) = \Omega_i A_{\text{c}} B_d^{\text{db}} \tag{III.26}$$

4.2.3 Handover

We use the mobility fluid-flow model described previously. Let N_x^{enb} be the number of eNBs served by an equipment x. The *Processing Load* due to handover events for x is given by:

$$S_x^{\text{h}} = \sigma_{\text{c}} N_x^{\text{enb}} \Big[M_x^{\text{h-sr}} (1 - P_{\text{rel}}) + M_x^{\text{h-nsr}} P_{\text{rel}} \Big] \tag{III.27}$$

For the MME and P-GW N_x^{enb}=C_e, for the S-GW N_x^{enb}=C_{ta} and for the eNB N_x^{enb}=1. Furthermore, the *Memory Access Rate* is given by:

$$L_x^{\text{h}} = \sigma_{\text{c}} N_x^{\text{enb}} \Big[I_x^{\text{h-sr}} (1 - P_{\text{rel}}) + I_x^{\text{h-nsr}} P_{\text{rel}} \Big] \tag{III.28}$$

Let B_d^{h} be the amount of signaling messages exchange during the Handover procedure. The *Radio Signaling Overhead* due to Handover procedure for d direction is computed as follows:

$$Z_d^{\text{h}} = \sigma_{\text{c}} B_d^{\text{h}} \tag{III.29}$$

4.2.4 Tracking Area Update

We assume a centralized MME architecture which only involves intra-MME TAU. Let N_x^{ta} be the number of TAs served by an equipment x and $\lambda_{\text{ue}} \sqrt{C_{ta}}$ be the crossing rate out of a

TA. The *Processing Load* due to TAU events can be approximated by:

$$S_x^{\text{tau}} = N_x^{\text{ta}} M_x^{\text{tau}} \lambda_{\text{ue}} \sqrt{C_{\text{ta}}} \tag{III.30}$$

where, in case of the MME and P-GW N_x^{ta}=$C_{\text{e}}/C_{\text{ta}}$, for the S-GW N_x^{ta}=1 and for the eNB N_x^{ta}=$1/C_{\text{ta}}$.

Furthermore, the *Memory Access Rate* is given by:

$$L_x^{\text{tau}} = N_x^{\text{ta}} I_x^{\text{tau}} \lambda_{\text{ue}} \sqrt{C_{\text{ta}}} \tag{III.31}$$

Furthermore, the amount of signaling messages exchanged during the Tracking Area Update procedure can be considered as negligible compared to previous procedures. Consequently, we do not take in consideration this procedure in the *Radio Signaling Overhead* evaluation.

4.2.5 Summary of Processing Load and Memory Access Rate Evaluation

Finally, let n be the total number of applications running on a UE, m the number of applications using the default bearer and $n - m$ the number of applications supported by a dedicated bearer. From equations (III.21) to (III.31) the total *Processing Load* (S) of element x can be computed as follows:

$$S_x = S_x^{\text{sr}} + S_x^{\text{h}} + S_x^{\text{tau}} + \sum_{i=m+1}^{n} S_x^{\text{db}}(i) \tag{III.32}$$

The total *Memory Access Rate* (L) of the element x can be computed as follows:

$$L_x = L_x^{\text{sr}} + L_x^{\text{h}} + L_x^{\text{tau}} + \sum_{i=m+1}^{n} L_x^{\text{db}}(i) \tag{III.33}$$

And finally, the total *Radio Signaling Overhead* (Z) in direction d can be computed as follows:

$$Z_d = Z_d^{\text{sr}} + Z_d^{\text{h}} + \sum_{i=m+1}^{n} Z_d^{\text{db}}(i) \tag{III.34}$$

5 Numerical Results and Analysis

We present and discuss in this section the numerical results and a performance evaluation of different scenarios. Based on the analytical models defined before, FigureIII.3 illustrates the simulated mobile network where UEs are moving according to a fluid-flow mobility model.

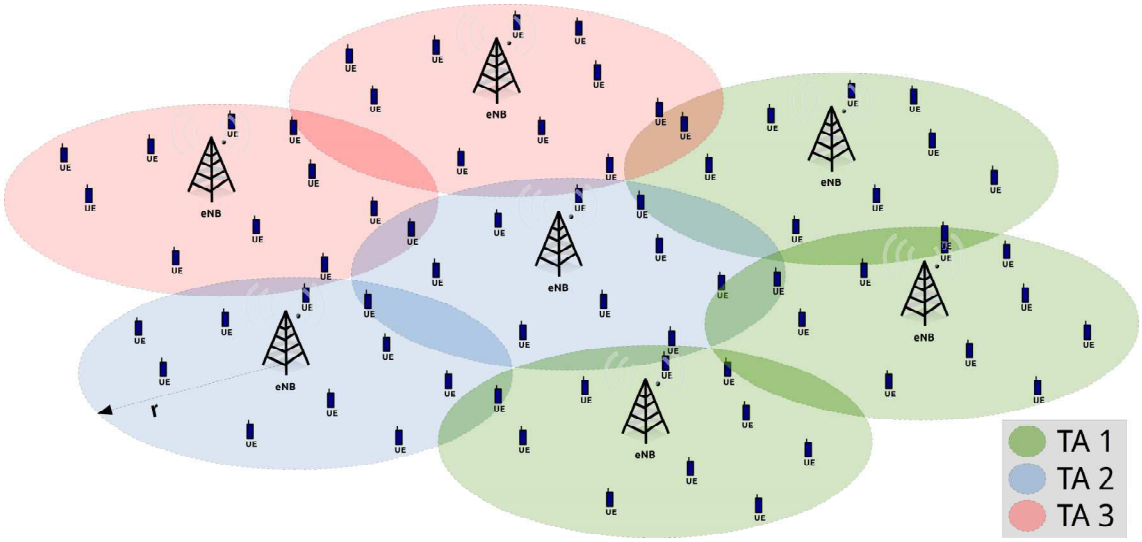


Figure III.3 – Simulated LTE/EPC mobile network

5.1 Network Parameters

The proposed scenarios are based on real statistics from a high user density area in Paris region and its suburbs which are presented in Table III.5.

Parameter	**Value**
Area Size (A_t)	$1300km^2$
User Density (ρ)	2300 UEs/km^2
Mean User Speed (V) [68]	$5km/h$
Number of eNBs in studied region (C_e)	2800 eNBs
Number of eNBs in a TAL	300 eNBs
Overlapping Factor 20% (γ)	1.2

Tableau III.5 – Scenario Parameters

5.2 Traffic Description

Based on Orange statistics we propose four main application types: voice, media streaming (i.e. YouTube, Dailymotion, Deezer), social networks (i.e. Facebook, Tweeter) and Background with their associated busy-hour parameters, which are detailed in Table III.6.

Application	**Session arrival rate per hour** (α_n)	**Session duration (s)** (μ_n^{-1})	χ_i
(A) Voice	0.67	180	0.5
(B) Streaming	5	180	1
(C) Social Network	20	30	0.5
(D) Background	40	10	0.8

Tableau III.6 – Traffic Parameters

5.3 Scenarios Description

Table III.7 shows the five analysed scenarios; the first one is a "Best Effort" deployment, which is currently the most frequent case. The second one assumes a Voice over LTE (VoLTE) offer, which is currently being deployed by some operators (we assume that a unique PDN is used by all services including VoLTE). The last three scenarios represent multi-level QoS deployments. In the third scenario, 10% of streaming traffic is supported by dedicated bearers in addition to VoLTE. In the fourth scenario, 5% of Social Network traffic is supported by dedicated bearers in addition to VoLTE and 10% of Streaming traffic. In the fifth scenario, all streaming and VoLTE traffic is supported by dedicated bearers (worst case).

Scenarios	**App. Using Default Bearer**	**App. Using a Dedicated Bearer**
BE	A + B + C + D	-
QoS_1	B + C + D	A
QoS_2	90% B + C + D	A , 10% B
QoS_3	90% B + 95% C + D	A , 10% B , 5% C
QoS_4	C + D	A , B

Tableau III.7 – Scenarios

5.4 Impact of the Inactivity Timer

The bearer-inactivity timer is usually around 20 seconds [69]. VoLTE dedicated bearers are deactivated via the IP Multimedia Subsystem (IMS) signaling procedure described in [70, 71]. This means that the bearer-inactivity timer for VoLTE dedicated bearer is 0. We also assume that the default Access Point Name (APN) is the IMS-APN, therefore only an extra dedicated bearer is needed to establish a VoLTE call.

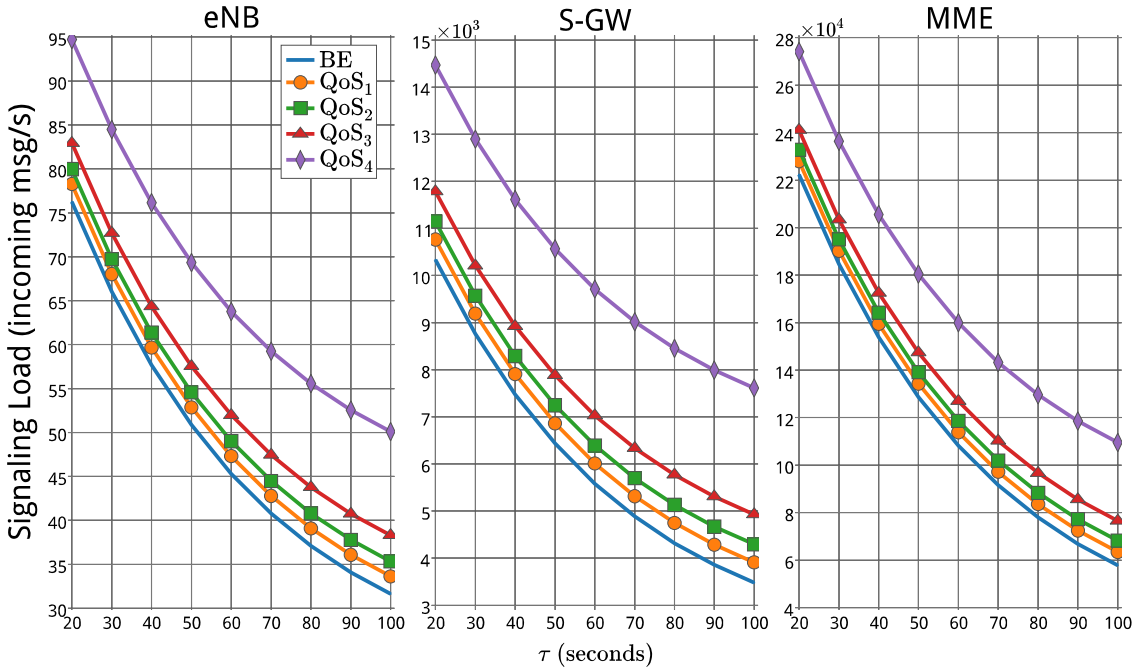


Figure III.4 – Impact of the inactivity timer (τ) on eNB, S-GW and MME *Processing Load*

We also vary the inactivity timer τ from 20 to 100 seconds. A common value used by mobile operators for dense areas [67] is an inactivity timer (τ) equal to 40 seconds.

Figure III.4 shows the impact of the inactivity timer (τ) on MME, S-GW and eNB *Processing Loads*. The *Processing Loads* decrease exponentially when τ increases in relation with the decreasing number of transitions from CONNECTED to IDLE states. The total *Processing Load* is obviously affected when QoS management is offered; the difference in load between the "Best Effort" scenario (BE) and the "QoS scenarios" (i.e. QoS_1, QoS_2, QoS_3 and QoS_4) is about constant, irrespective of τ. The relative increase in *Processing Load* with respect to a "Best Effort" scenario (BE) is showed in Figure III.5.

Figure III.6 shows the impact of the inactivity timer (τ) on MME, S-GW and eNB *Context Load*. The *Context Loads* increase when τ increases in relation with the increasing number of context in various LTE/EPC elements.

Figure III.7 shows the impact of the inactivity timer (τ) on MME, S-GW and eNB *Memory Access Rate*. The *Memory Access Rate* decrease exponentially when τ increases in

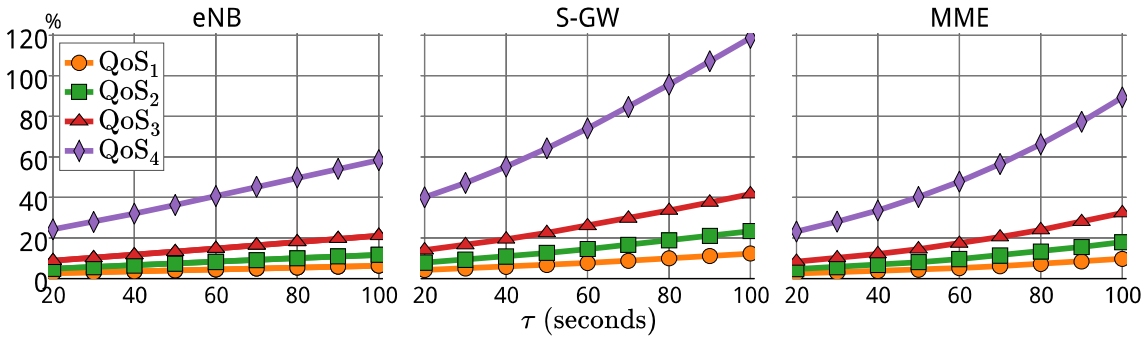


Figure III.5 – *Processing Load* Increase compared to BE scenario

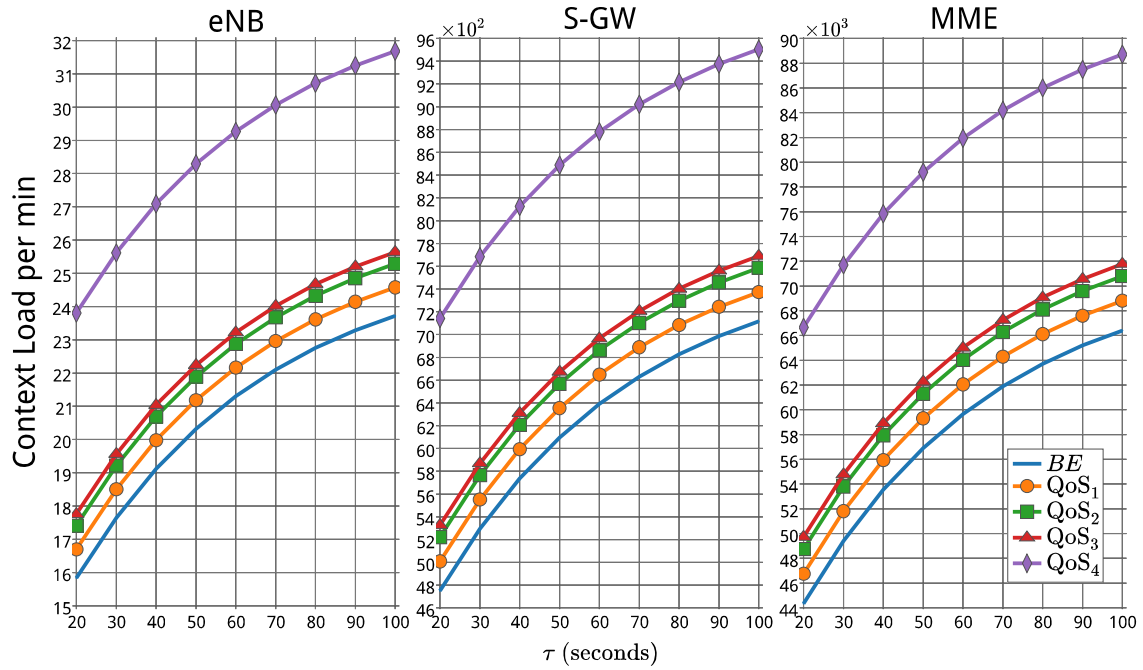


Figure III.6 – Impact of the inactivity timer (τ) on eNB, S-GW and MME *Context Load*

relation with the decreasing number Context creation, release and modification. Figure III.5 and III.8 show the percentage increase in *Processing Load* and *Memory Access Rate* respectively, relatively to the BE scenario.

Simulations show that the increase of *Processing Load* and *Memory Access Rate* due to multi-bearer deployment for QoS management is relatively moderate, less than 60% in usual configuration (τ=40s). However, it is more perceptible in rather centralized equipments such as the MME and the S-GW. Nevertheless, scenario QoS_4 shows a major impact in *Context Load*, which is increased by around 200%.

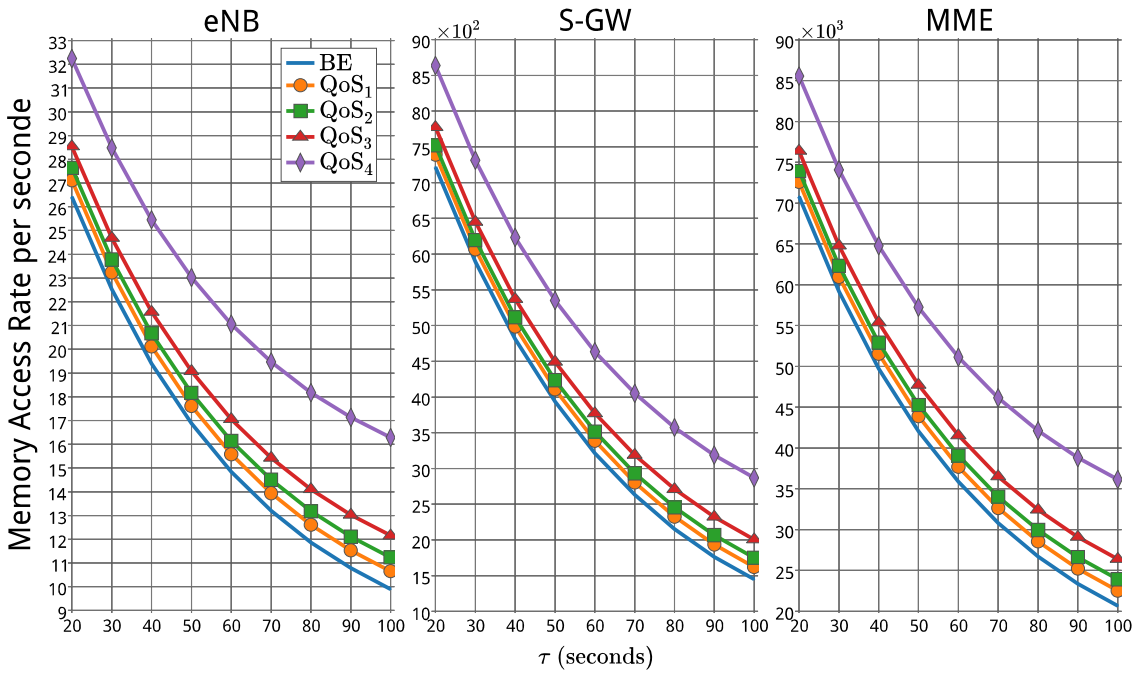


Figure III.7 – Impact of the inactivity timer (τ) on eNB, S-GW and MME *Memory Access Rate*

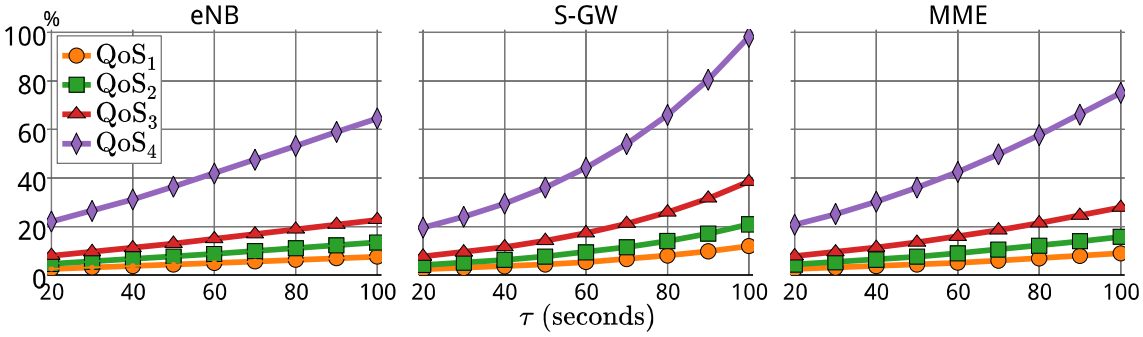


Figure III.8 – Memory Access Rate compared to BE scenario

5.5 Impact of the Sessions Arrival Rate

Now we examine the impact of the VoLTE sessions arrival mean rate. Here α is the session arrival rate of VoLTE. The inactivity timer τ is fixed at 40 seconds, which is the common value in urban deployments [67]. Then, we consider all the default values for other parameters and we vary α from 0 to 10 (sessions arrival rate per hour). As α increases, the average number CONNECTED UEs also increase, which means that the average duration of UEs period of activity increases.

Figure III.9 shows the impact of the VoLTE sessions arrival (α) on MME, S-GW and eNB *Processing Loads*. In case of BE scenario, as α increases, the *Processing Load* decrease as well as the *Memory Access Rate* (Figure III.10). This is because in BE

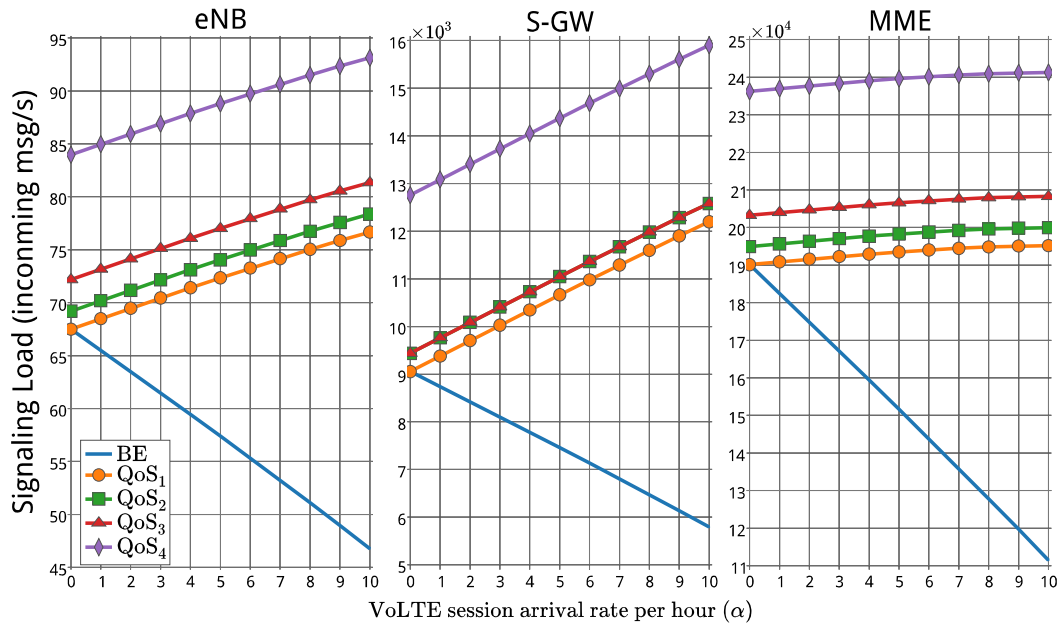


Figure III.9 – Impact of the VoLTE sessions arrival rate (α) on eNB, S-GW and MME *Processing Load*

scenario, as VoLTE session arrival rate increases, the average duration of a period of activity (CONNECTED state) also increases, as a result the average "life time" of a default bearer also increases. Therefore, *Processing Load* and *Memory Access Rate* decrease. By contrast, in case of QoS scenarios, the *Processing Load* and the *Memory Access Rate* increase, due to QoS procedures associated with the VoLTE Dedicated bearer.

In our current proposed configuration (scenario and traffic parameters), the *Memory Access Rate* of the S-GW decreases with α in all scenarios (BE and QoS scenarios). This situation is due to that the *Memory Access Rate* related to non-QoS procedures (left side of the equation III.33) decrease more rapidly and much larger extent than the QoS procedures (right side of the equation III.33) increase, as shown in Figure III.9. One of the reasons for this is that the S-GW is not implied in the TAU and the Handover (without S-GW relocation) procedures, which significantly contributes to the *Memory Access Rate* in the others LTE/EPC equipments.

Figure III.11 shows the impact of the VoLTE sessions arrival rate (α) on MME, S-GW and eNB *Context Loads*. In contrast to *Processing Load* and *Memory Access Rate*, the *Context Loads* increase in all scenarios. This is because the average duration of UEs activity increases.

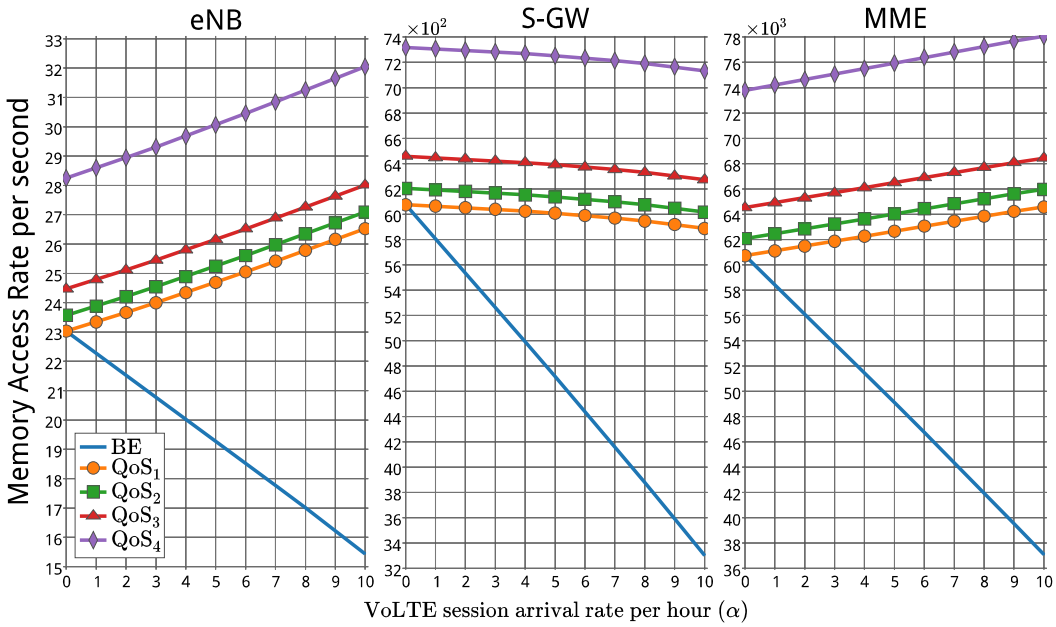


Figure III.10 – Impact of the VoLTE sessions arrival rate (α) on eNB, S-GW and MME *Memory Access Rate*

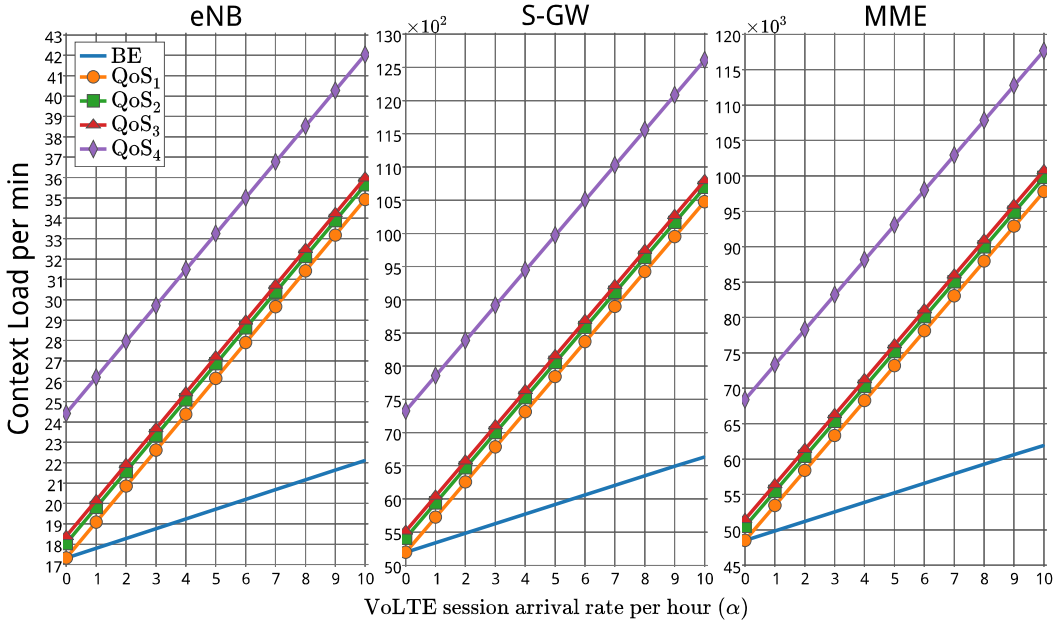


Figure III.11 – Impact of the VoLTE sessions arrival rate (α) on eNB, S-GW and MME *Context Load*

5.6 QoS Signaling Impact in LTE Radio Segment

Now we examine the impact of the QoS signaling in LTE radio segment. This evaluation is inspired of Ref. [72], which evaluates RRC signaling overhead of diverse data applications.

It should be remembered that each Dedicated Bearer is always established via the Control Plan procedures, which means that on each network segment of LTE/EPC system, a local dedicated bearer is established, including the radio segment (i.e. Data Radio Bearer (DRB)). These Control Plan procedures in the radio segment are managed by the RRC protocol layer [73].

Application	**Mean DL Data Rate (kbps)**	**Mean UP Data Rate (kbps)**
(A) Voice [74]	10.2	10.2
(B) Streaming	540	53
(C) Social Network	217	16
(D) Background	423	43
Total	**1190.2 kbps**	**122.2 kbps**

Tableau III.8 – Mean Applications Data Rate

In order to evaluate the *Radio Signaling Overhead* we vary the inactivity timer τ from 20 to 100 seconds, we keep constant the bearer-inactivity timer to 20 seconds and the VoLTE sessions arrival mean rate to 0.67 per hour. We measure the ratio of the *Radio Signaling Overhead* to the mean Applications Data Rate. The mean Applications Data Rate is shown in Table III.8, which is based on Orange France statistics.

Figure III.12 shows the ratio of the *Radio Signaling Overhead* to the mean Applications Data Rate in both Uplink (UL) and Downlink (DL). As expected, the 3GPP QoS scenarios enlarges the *Radio Signaling Overhead* compared to a BE scenario in UL and DL, and as a result the ratio of the *Radio Signaling Overhead* to the mean Applications Data Rate is higher in QoS scenarios.

Furthermore, the *Radio Signaling Overhead* (DL/UL) in decreases in all scenarios when τ increases in relation with the decreasing number of exchanged signaling messages, due to the decreasing number of transitions from CONNECTED to IDLE states. As a result the ratio of the *Radio Signaling Overhead* to the mean Applications Data Rate also decreases. It should also be noted that the handover procedure strongly contributes to the *Radio Signaling Overhead*, since information on each established dedicated bearer must be exchanged (see Table III.4).

Finally, in our current scenario the *Radio Signaling Overhead* can be considered negligible in most cases. Nevertheless, the *Radio Signaling Overhead* can become relevant as the mean applications Data Rate decreases. This is the case of LTE networks with a low radio configuration (low bandwidth) and high traffic density.

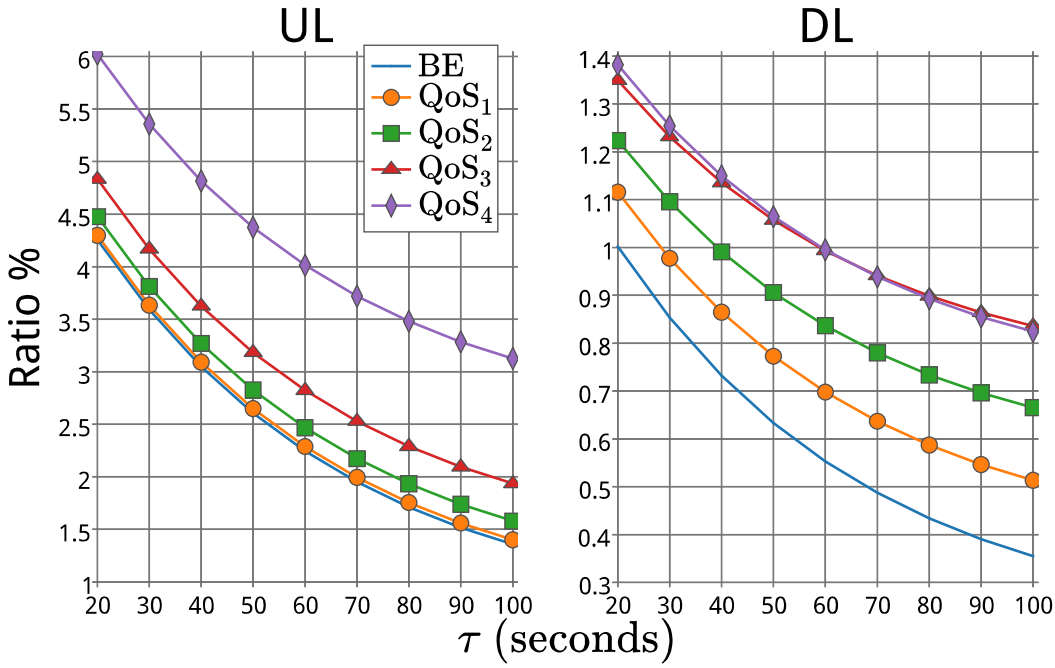


Figure III.12 – Ratio of the *Radio Signaling Overhead* to the mean Applications Data Rate

6 Conclusions

In this chapter we have presented a novel analytical model to evaluate the impact of the standard LTE/EPC QoS model in terms of *Context Load*, *Processing Load*, *Memory Access Rate* and LTE radio segment *Radio Signaling Overhead*. Note that the cost factor of the 3GPP QoS model has not been analysed until today in the literature; in this extend, this thesis provides for a valuable contribution to enlighten this model.

It has been shown that the deployment of the VoLTE/VoLTE and other premium services using dedicated bearers could have a significant impact on the performances of LTE/EPC elements. A proper dimensioning of equipment processing/memory capacity and appropriate engineering rules (i.e. τ value) are therefore essential. It is also important to take into account the traffic behaviour (ON-OFF cycles) of premium services, since it could be detrimental to the performances of LTE/EPC elements. Furthermore, the RRC Inactivity timer can have a relevant impact on the battery life of UEs [75]. A tradeoff between user satisfaction and network optimisation should then be found.

As an alternative to this 3GPP QoS model, the QoS may be managed as in fixed IP networks, which we will call the "IP-centric" approach. In this case, QoS is managed at IP level without dedicated bearers nor signaling. This "IP-centric" model has already drawn interest amongst some major actors of the mobile industry. The "IP-centric" approach is considered in the next Chapter and will be widely investigated.

Chapter IV

IP-centric QoS Model for 3GPP Mobile Networks

1 Introduction

New mobile devices and increased mobile data traffic have dramatically altered the mobile ecosystem, which should now accommodate to rapidly changing customer behaviours. These changes in usage patterns will undoubtedly lead to a huge traffic increase, as was shown in chapter III. Even if LTE systems provide higher data rates and lower latency, the radio segment will probably remain a potential bottleneck due to coverage problems or bad network dimensioning to cope with peak hours.

In order to face up with this mobile data traffic increase interest has been growing for improve QoS mechanisms on the radio segment. In this sense the 3GPP has anticipated the need for differentiated QoS with R8 release. These mechanisms enable the operator to give a higher priority services or even to some called "premium users".

However, the associated QoS model took into account the mobile ecosystem of that time (2008). Since then, a vast number of new devices, usages (i.e. LTE dongles, Mobile Wi-Fi (MiFi), and tethering applications) and applications have come to market. For example, video and music streaming services are widely used today, but did not even exist just a few years ago. This disruptive evolution of mobile usages and services (e.g. VoLTE) results in an important challenge for operators who struggle to differentiate themselves from competitors. As a matter of fact, mobile standards have not yet really addressed this expectation of open, cheap and flexible web oriented Internet access.

In a nutshell, there is a need to improve QoS in mobile networks. One approach to provide an appropriate End-to-End (E2E) QoS in this multi-service environment is to over-provision the mobile network on the wireless segment and the wireline segment both. However, over provisioning the entire network is very costly, rarely practical

or even impossible. As a result, bottlenecks oftenly appear in the network, mainly on wireless segments. This requires cost-effective QoS mechanisms in order to support service differentiation. This cost-effectiveness is crucial considering the expected decrease of revenue per transported bit. Consequently, these QoS mechanisms should have a low complexity of implementation to decrease costs without degrading the QoE, in the same way that it is done in the fixed Internet world.

In this sense, in this chapter we propose an IP-centric QoS model mainly inspired by the DiffServ architecture commonly found in fixed networks. We further describe a possible cross-layer design for this model. Some implementation challenges have been highlighted, together with possible solutions implying only minor modifications in eNB, and none in terminals. Performances of this proposal compared to various implementations of the 3GPP QoS model have been evaluated using the ns-3 simulator in realistic scenarios. Some good properties of our IP-centric proposal compared to the standardized QoS model have been brought into evidence.

2 Discussion on the 3GPP QoS management

2.1 3GPP QoS management

As seen in the previous chapters, the QoS management in 3GPP mobile networks (i.e. 3G, 4G) is clearly connection-oriented, which aims to offer a fine-tuning QoS.

Bearers are operated in connected mode, that is, they are established, modified or disconnected via mobile signaling protocols. Note that the radio scheduler plays a crucial role on the 3GPP QoS model, and consequently, it dramatically impacts the end-to-end customer experience. However, 3GPP specifications do not define any scheduling algorithm, leaving its design and implementation open. Previous examples of radio scheduling algorithms can be found in the literature [19, 45]. But many of them propose purely improve network capacity regardless of QoS and others pursue QoS objectives (i.e. throughput , delay, packet loss) and are designed to support a single traffic type (e.g. VoIP, video). Some of them as [76] and [77], aim to satisfy the QoS requirements to the detriment of global performance or the other UEs. However, vendors generally implement Proportional Fair (PF) algorithms. Such algorithms propose a trade-off between cell throughput optimization and fairness between the UEs, which is detailed in [46].

The 3GPP QoS model may often reveal tricky and complex to use. In particular, it raises issues in terms of:

- **Scalability:** In a centralized architecture such as the LTE/EPC network, the P-GW manages a huge number of bearers, as evidenced in chapter III. It should be recalled that each UE uses an independent bearer per QoS level, in addition to its default bearer.

- **Efficiency:** The establishment, modification, release and mobility management of each bearer generate a non negligible signaling load in the network [78]. Furthermore, the creation and destruction of PDCP and RLC entities imply an important processing load on the eNB.

- **Performance:** Bearers' establishment implies a processing time which results into network latency. This delay is not negligible and may severely impact applications.

- **Costs:** QoS-aware radio scheduler are the cornerstone of the 3GPP QoS model. However, they break the virtuous trade-off of proportional fair algorithms as they allocate more resources to some UEs regardless of their radio conditions. This may severely impact the overall cell capacity, particularly when prioritized UEs are in poor radio conditions. This is phenomenon is emphasized when prioritized UE are strongly favoured. As a consequence, tuning QoS-aware schedulers so as to offer a satisfying QoS level without starving the other UEs reveals a harsh task.

In fact, keeping up a fine-tuning QoS management, makes insurmountable the issues listed above and at the end of the day, mobile operators often regard the 3GPP QoS scheme as too complex and costly in view of its expectable revenues in 3G and 4G networks, except for VoLTE.

Furthermore, the multi-bearer 3GPP QoS model is very similar to access architectures proposed in the late 90s for residential fixed services on ADSL. As a matter of fact, in less than a decade, Internet QoS paradigms (packet-oriented QoS / DiffServ) have gradually replaced connexion-oriented QoS schemes inherited from (virtual) circuit-based networks (synchronous or ATM networks) in the vast majority of fixed networks, including those supporting voice.

Finally, 3G/4G traffic is often supported through a unique best-effort default bearer per customer, or through two bearers when VoLTE is offered. Note that some performance issues have been reported in the 2014 ITU Workshop on "Monitoring Quality of Service and Quality of Experience of Multimedia Services in Broadband / Internet Networks" (April 2014) [79, 80] for VoLTE. The pointed performance's issues are mainly due to incompatibility of 3GPP standards or its wrong implementation by constructors. In case of VoLTE, Orange's tests have highlighted many issues with the voice quality, which is stable until the cell edge, where it brutally decreases.

2.2 Today's mobile networks QoS policies

Nowadays, the first LTE deployments generally do not implement the full 3GPP QoS policies and only operate a mono-bearer Best-Effort architecture. Note that nowadays only little deployment of multi-bearer QoS architecture is observed in order to support. This is due to the VoLTE deployment, which require a specific QoS level (i.e. low latency).

According to GSMA [81], until now only 40 of 432 LTE operators (10%) all over the world have launched VoLTE. As a consequence, the vast majority of mobile terminals are only connected through a unique default bearer, their whole traffic being transported on a best effort basis as show in Figure IV.1. Furthermore, only few mobile terminals are compatible with it (today only 75 models are compatibles with VoLTE [81]).

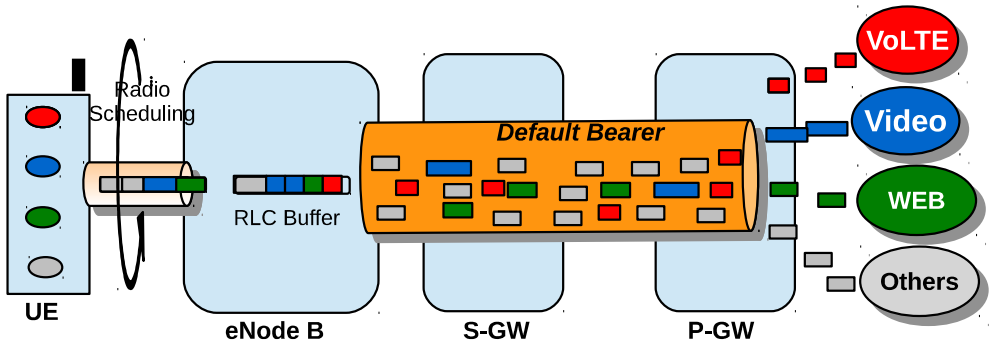


Figure IV.1 – Today's mobile networks QoS policies

In this mono-bearer context, it then appears clearly that increasing bearer resource allocation through GBR or weighted Proportional Fair is not always the most appropriate solution to improve customer experience. For example, when several applications are running on the same UE (this situation is expected to be fairly common with tethering, MiFi, etc.), it may be much more efficient to properly schedule the user flows without modifying his global resources allocation. This implies intra-bearer QoS differentiation, where sensitive flows belonging to the same UE are favored against less sensitive flows, as proposed in [82]. Note that this may be performed without modifying the user radio resource allocation. This allows for cell capacity preservation as replacing basic Proportional Fair by QoS-aware scheduling (e.g. GBR or weighted Proportional Fair (PF)) statistically reduces this capacity.

Consequently, the 3GPP QoS model is probably not the right answer to the somehow contradictory expectations for cheap, efficient and flexible QoS in the mobile Internet.

3 DiffServ approaches for QoS provisioning in Mobile Networks

Based on the foregoing, it makes sense to investigate more cost-effective QoS mechanisms, which should have a low level of complexity in terms of implementation to decrease costs without degrading too much the QoE. In this sense, we can take advantage of the all-IP architecture of LTE in order to use IP mechanisms widely studied and which are used on the fixed Internet, such as the DiffServ approach described in chapter II . These mechanisms could be adapted to fit the mobile ecosystem. Indeed, these IP mechanisms

have proven to be flexible, cost-effective, scalable, easy-to-configure and well adapted to open ecosystems.

The DiffServ approach, what we call the "IP-centric" architecture, can be envisaged in order to build an alternate QoS architecture for mobile networks. IP-centric approach for QoS management on mobile networks has already drawn interest amongst some major actors of the mobile industry such as Huawei [83] , Nokia [84], Alcatel-Lucent and AT&T [85].

3.1 Problem Statement and Solution Strategies

In the IP-centric architecture, the QoS provisioning can be performed by using only information contained in each packet's header. Thus, the QoS is managed packet by packet and all flows belonging to the same user should be transported on a unique multi-QoS bearer.

An important aspect of IP centric model is the packet marking (i.e. DSCP), which should be performed by a trusted entity, for example the P-GW or an upstream trusted equipment outside of the LTE network.

3.2 Related Work

In this section, we discuss some prior research related work for IP-centric QoS architecture. These works are inspired of DiffServ approach.

An early IP-centric model for QoS provisioning on mobile networks was introduced by Soldani in [86]. This model implements a pseudo IP-centric approach for UMTS mobile networks (3G networks). At Radio Network Controller (RNC) level, the PDCP sublayer behaviour is modified, which implements priority packet buffering for each QoS level, in this paper 3 QoS levels are proposed (Gold, Silver and Bronze). Packets are classified according to DSCP marking without regard the UE destination. The DSCP marking is performed by the Gateway GPRS Support Node (GGSN). The modified PDCP sublayer uses the queue weight (i.e. DSCP value) to service prioritisation. At Node B level, the radio scheduler uses the Scheduling Priority Indicator (SPI) to UEs prioritisation. The SPI is a key scheduling parameter, which is computed for each bearer by the RNC according its QoS parameters provide by the Core Network via signaling procedure. A key element in this approach is the radio scheduler, which aims to guarantee the GBR for those UEs implemented it. Radio scheduler allocates the remaining resources to UEs according to associated SPI weight.

This solution has been implemented by one of Orange industrial partners and we have tested it in our lab. Tests showed that the throughput is negatively impacted due to radio scheduler nature (i.e. GBR bearers prioritization).

In [84] the authors present two alternative IP-centric mechanisms. The first one, called "in-bearer prioritization" is proposed for High Speed Packet Access (HSPA) systems. This work is equivalent to the previous proposition and there are not relevant differences. The second mechanism is called Network Requested Secondary PDP Context Activation (NRSPCA) for a HSPA network. The NRSPCA implements a secondary PDP context per QoS level detected via and Deep Packet Inspection (DPI) implemented at GGSN level.This approach gives a greater flexibility in choosing a RLC transmission mode and granularity in the scheduling process. The NRSPCA is equivalent to dedicated bearer procedure of LTE/EPC networks.

In [87] the authors present one of the first IP-centric models. In this work an opportunistic radio scheduler for UMTS systems called TCP-aware scheduler is proposed. It aims to improve TCP user perception. It proposes the implementation of a set of buffers per user at Node B level. Furthermore, each buffer is associated to a TCP service (i.e. DSCP marking). It also propose a set of hierarchical channel-dependent and traffic-aware scheduler at the MAC and IP layer.

Note that, all works mentioned above use the DiffServ approach in order to manage the QoS in 3G mobile networks. The main idea is to use the DSCP marking in order to place packets in different buffers, which are scheduled according to priority level. Nevertheless, neither of these address the IP-centric approach in deep way (i.e. implementation propositions, study of its impact in current services - e.g. voice, streaming), which is the aim of current chapter.

In November 2012, 3GPP initiated a study item on mobile User Plane CONgestion (UPCON). The associated Technical Report [88] is intended to study scenarios and use cases leading to user plane traffic congestion in the RAN, and to propose system enhancements for managing this congestion.

The initial proposal was to notify the congestion status of the RAN to the Core Network, which may apply policies in order to reduce the traffic sent to this saturated RAN. It then clearly appears as the continuation of the connection-based model, where QoS management is performed from a centralized point located in the core network (e.g. P-GW) on a per connection basis.

However, a joined contribution between ALU and AT&T [85] has been submitted during the January 2013 meeting. This contribution claims that 2 different ranges of solutions should be studied, one based on "reactive" solutions where information about RAN congestion is notified to the core, and the other one based on "proactive" solutions where RAN locally deals with congestion. While "reactive" solutions are perfectly in line with the original UPCON spirit, the "proactive" solution is quite compatible with the IP-centric approach. That is the case of the proposition called Flow priority-based traffic differentiation on the same QCI or Flow Priority Index (FPI), based in Soldani's work described above. In FPI proposition the P-GW marks each user plane data packet delivered in the downlink direction with a FPI. Then, the RAN node applies the FPI for

differentiating the treatment of the packet compared to other packets mapped to the same QCI during congestion situations.

In order to carry FPI information to RAN, two solutions are proposed. The FPI could be defined as a new GTP-U extension header or it could be encoded as a DSCP value in the IP header.

4 IP Centric model for 3GPP Mobile Networks

Based on the foregoing, it makes sense to investigate more cost-effective QoS mechanisms, which should have a low level of complexity in terms of implementation to decrease costs without degrading too much the Quality of Experience (QoE). In this sense, we can take advantage of the all-IP architecture of LTE in order to use IP mechanisms widely studied and which are used on the fixed Internet. These mechanisms could be adapted to fit the mobile ecosystem. Indeed, these IP mechanisms have proven to be flexible, cost-effective, scalable, easy-to-configure and well adapted to open ecosystems.

In hereafter, we present a complete analysis of our IP centric model, which was developed on [78, 89, 90, 91, 92]. This model is based on IP paradigms in order to address QoS issues in mobile networks and aims to satisfy the QoS requirements for different traffic types on a heterogeneous traffic environment while guaranteeing the fairness and without degrading the global throughput. It is also important to mention that this model offers an End-to-End QoS technology-agnostic.

4.1 IP-centric Architecture

In this IP-centric model, see Figure IV.2, a unique EPS bearer (default bearer) per UE is established so as to manage the connectivity and other specific features of mobile networks. All the flows belonging to a given user are transported on this unique multi-QoS bearer.

QoS is further managed at packet level, according to the DSCP field for example. An IP multiplexing stage should then be added before the radio scheduling in the eNB. Furthermore, a set of priority queues per UE should further be implemented in the eNB.

For a given UE, dowstream packets arriving at the eNB through the aforementioned bearer are placed onto the relevant queue based on their DSCP marking. A queuing system is located in the eNB, which can be schedule at a rate taking into account very accurately the lower layers status (e.g. radio conditions, cell load or available radio resources) that generally change every TTI.

Hence, we distinguish then two main types of behavior for this IP-aware eNB:

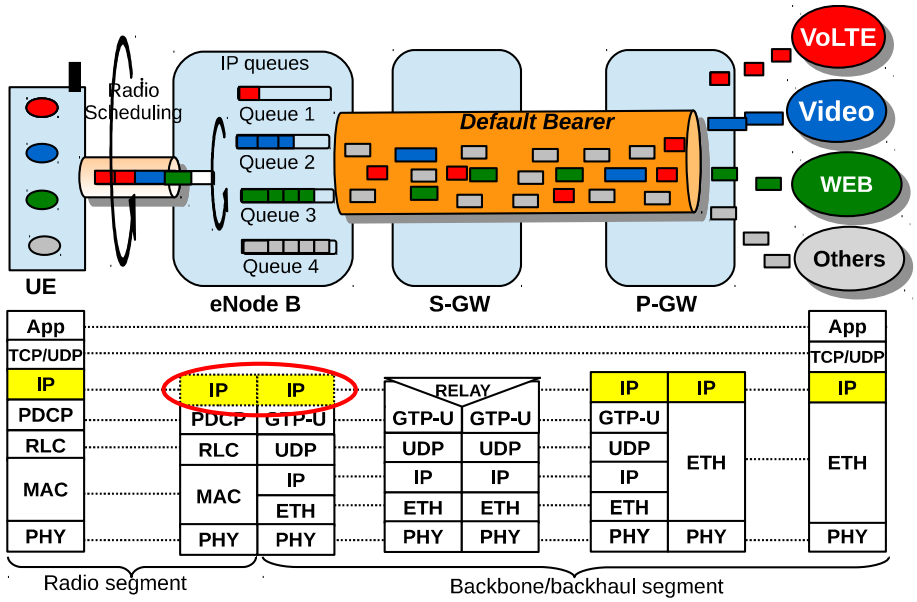


Figure IV.2 – IP-centric Architecture

4.1.1 Intra-bearer arrangements

The radio resource allocation is independent of the traffic mix waiting for transmission at IP layer; in this case, radio resources allocated to a given UE are determined by a basic radio scheduler (e.g. Proportional Fair), taking into account the radio conditions of the UE as usual (i.e. based on 3GPP specifications - RLC buffer size) as shown in Fig.IV.3. It should be noted that in our model, each RLC sublayer entity has full knowledge of the queues size of IP layer, which is communicated to lower layers as the RLC buffer size. The addition of an IP priority queuing system per UE before the radio scheduler (without influencing it) allows for prioritization of the sensitive flows of a given user against his own other flows when populating the radio frame (i.e. LTE Transport Block). This model is described in more detail below.

4.1.2 Inter-bearer arrangements

The allocation of the radio resources depends on the traffic mix waiting for transmission; in this case, the radio resources allocated to one UE depend not only on the radio conditions of the UE, but also on the traffic mix offered to the IP queuing system as show in Fig.IV.4. In this case, the scheduling algorithm should know the queues state of the UE. For this purpose, several approaches are possible (e.g. weighting the allocation according to the prioritized traffic volume / priority queue backlog, ensuring a maximum latency for specific classes).

The following is a discussion about the cross-layer design of the "intra-bearer arrange-

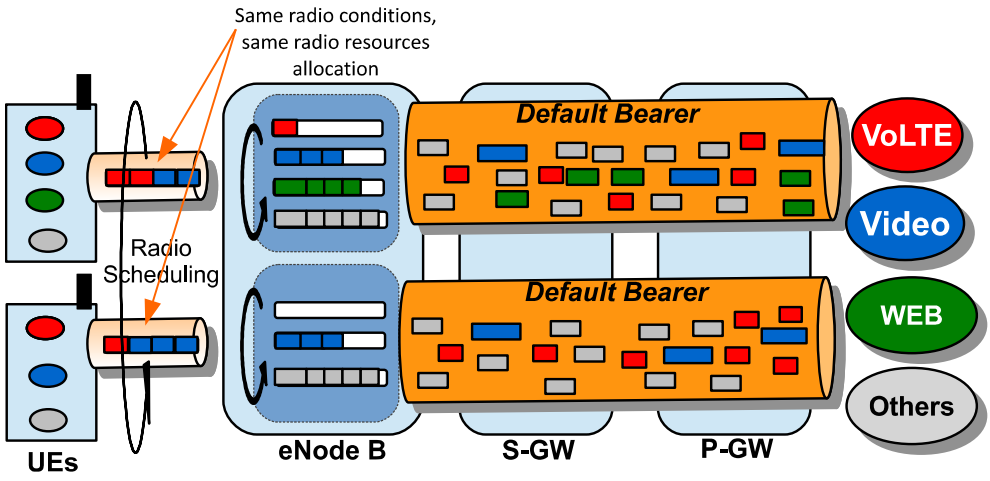


Figure IV.3 – Intra-bearer arrangements

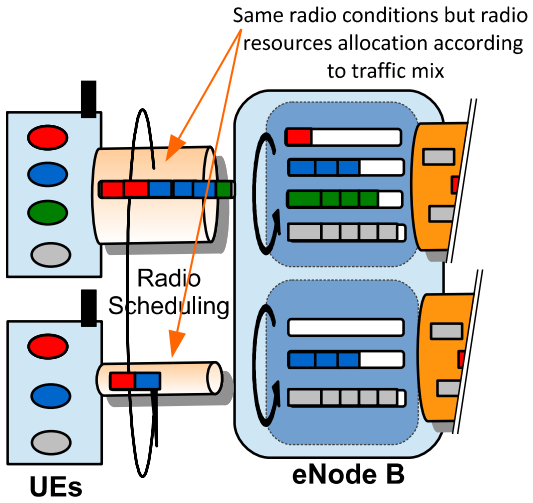


Figure IV.4 – Inter-bearer arrangements

ment" model, which provides an interesting tradeoff between fairness and efficiency in heterogeneous traffic scenarios such as today's.

4.2 Business models

IP-centric QoS management is in line with typical OTT business models: Instead of the flow by flow Application Program Interface (API) suggested by 3GPP typical architectures, third parties may interact with mobile operators through static SLAs at interconnection point. The SLA should define the marking and the maximum rate for each QoS class at interconnection point, as depicted on Fig.IV.5.

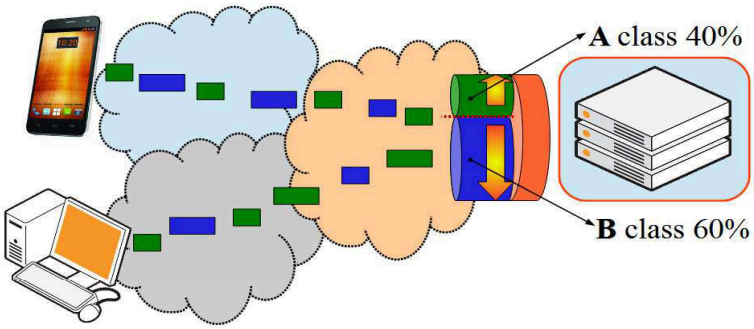


Figure IV.5 – Possible business model with IP-centric QoS approach

Note that this model is totally consistent with fixed IP networks QoS architecture, and paves the way to seamless services and access agnostic applications. This opens the door to valuable multilevel QoS agreements between content providers and operators.

Furthermore, the IP-centric is quite compatible with 3GPP multi-bearer scheme through nested configurations. For example, managed services may be supported by dedicated bearers according to 3GPP model when web oriented OTT services would be differentiated between themselves through IP aware management in the default bearer.

4.3 Cross-layer design for IP-aware eNB

4.3.1 Design challenges

As explained above, downstream IP packets destined for a given UE arrive at the eNB via a unique GTP tunnel (S1 bearer). Then the IP-aware eNB removes the GTP header of each packet and places them in the IP queuing system associated with this UE. The eNB should then implement a pseudo IP layer over the LTE radio protocol stack, the main features of which come down to packet classification and queueing. Many different designs may then be considered in order to address the interactions between this new IP layer and lower radio layers.

At this point, in order to address the interaction between the IP layer and lower radio layers, two different design ways can be considered:

1. **Virtual radio bearer:** at the radio interface, local DRBs between the eNB and the UE may be used, these are called virtuals because them are associated to a single EPS bearer. This approach can be see as a generalization of the NRSPCA proposed in [84] to LTE/EPS networks. The aim is to exploit a fine granularity in the radio scheduling process. Local signaling procedures should be used in order to establish each virtual bearer. Based on the virtual DRB solution two alternatives are possible.The first one is to use a virtual DRB associated to each QoS level (i.e. each queue), which gives a greater flexibility in choosing a RLC transmission mode

and granularity in the radio scheduling process. The second one is to use only two virtual DRB per UE, for example one GBR and the other non-GBR or one for each RLC transmission mode (RLC Unacknowledged Mode (UM) for delay-sensitive applications (i.e. real time applications) and a RLC Acknowledged Mode (AM) for loss-sensitive applications (e.g. HTTP and FTP). Virtual radio bearer approach can be implemented in downlink and uplink, nevertheless the main disadvantage is that majors modifications in the terminal and the eNB are required. In the case of use a virtual bearer per QoS level, a modification in the control plan is also required in order to establish and release the necessary virtual bearers. It is also important to mention that a appropriated radio scheduler is needed, which takes into account the priority of the virtual bearers (e.g. Weighted Proportional Fair scheduler)

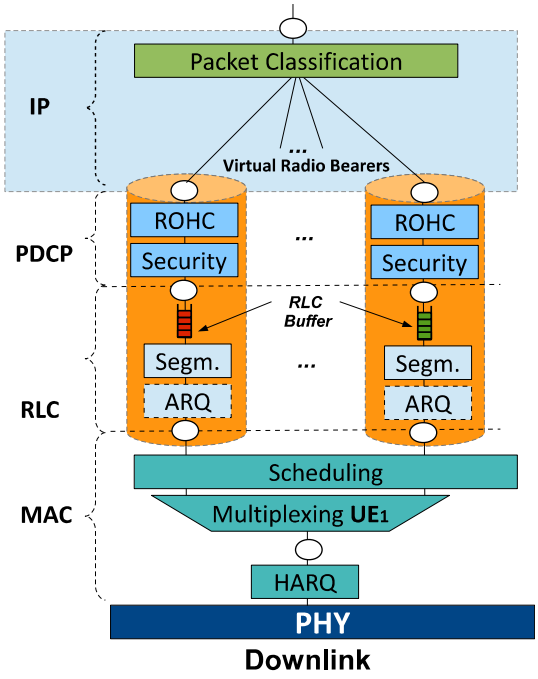


Figure IV.6 – Virtual radio bearer design

2. **Mono radio bearer:** in this approach, we are based on current bearer architecture, one DRB associated to default EPS bearer in which all packets are conveyed according to its priority (i.e. DSCP marking). The main advantage of this approach is that no relevant modification is needed on the eNB or the terminal. Furthermore, it can be implemented in downlink and uplink. In this case, the Proportional fair scheduler is the best choice because it provides a good compromise between fairness and efficiency.

In Table IV.1, we summarise some key characteristics of each IP-centric design, respect to the 3GPP solution. In 3GPP solution and Multi DRB, packet prioritisation is performed by the MAC scheduler, which involves the implementation of QoS-Aware schedulers. The

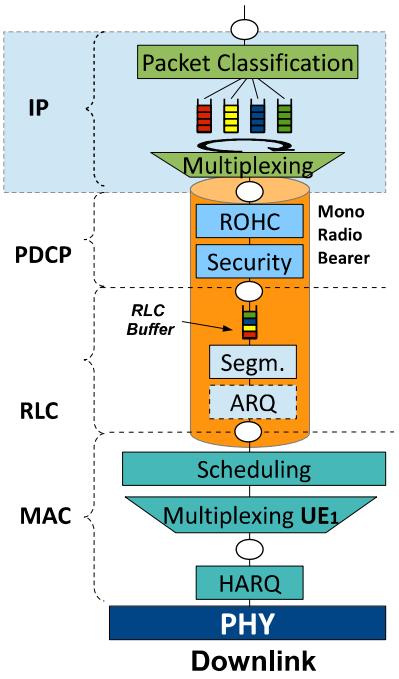


Figure IV.7 – Mono radio bearer design

Criteria	3GPP QoS	IP-centric QoS	
		Mono DRB	Multi DRB
QoS Prioritization Level	MAC	IP	MAC
Implementation Complexity	High	Medium	High
Fixed–mobile QoS convergence	Complex	IP based	IP based
Congestion control mechanism	Few propositions [85]	Those proposed for IP	Those proposed for IP
Terminals compatibility	Few	All	Few

Tableau IV.1 – Key Characteristics of 3GPP and IP-centric Approaches

fixed–mobile QoS convergence is almost native in case of IP-centric solution, as well as the implementation of all existing Congestion Control mechanisms for IP networks. Finally, in case of Mono DRB, all terminals are fully compatibles, in case of downlink implementation.

We describe hereafter a solution that implies only minor modifications in the eNB and none in the UE (i.e. Mono radio bearer). In this case, a basic PF scheduler can be used in the eNB , and only minor changes in the RLC sublayer are required in addition to our pseudo IP layer. The main challenges regarding the RLC sublayer are described hereafter:

1. **RLC buffer management:** The RLC buffer should contain no more data than the exact amount required by the MAC sub-layer at each TTI. Otherwise, on the next scheduling cycle (next TTI), low priority data that remain in the RLC buffer would be served before high priority packets arrived meanwhile.

2. **Data volume waiting for transmission:** In 3GPP standard implementations, the radio scheduler should know the amount of data queued in the RLC buffer in order to allocate no more radio resources than needed. In our IP-aware eNB, the RLC buffer is managed so as to be generally empty (see RLC buffer management above), as data are mainly queued in the IP Layer queuing system. The MAC sublayer should then take into account all data waiting for transmission, either buffered in the RLC or in the IP layers.

4.3.2 Design description

Figure IV.8 shows our IP-aware eNB. Our model implements an IP layer with an IP queueing system per UE, then a unique DRB per UE is used.

At the MAC sublayer, users are chosen each TTI, together with their respective radio resources and coding schemes depending upon the radio scheduler strategy. This strategy takes also into account a queueing status provided by the RLC sublayer, so that resources allocated to a UE do not exceed its real needs. In our IP-centric model, this queueing status should take into account not only the RLC queue - which is supposed to be empty most of the time thanks to the IP/RLC synchronization described above - but mainly the IP queues devoted to this UE.

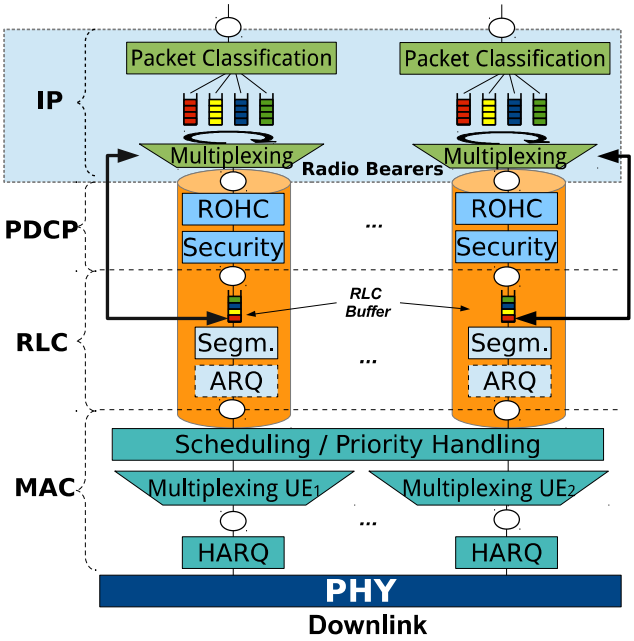


Figure IV.8 – eNB-Uu interface - IP-aware downstream protocol stack (simplified)

Then, packets are picked up from the IP queueing system by the RLC layer in accordance with the credits allocated to this UE by the MAC sublayer ("Transmission opportunity") at each scheduling cycle. Based on this number of credits, each IP entity

provides packets to its corresponding PDCP entity. Then, the PDCP sublayer adds a sequence number. After that, it compresses headers using ROHC protocol and ciphers the packet, finally it adds a PDCP header before passing it to its corresponding RLC entity.

The RLC entity segments or concatenates packets according to the data credits and a RLC header is added containing, amongst others, the corresponding sequencing number. As only one DRB per UE is implemented, the RLC transmission mode is the same for all packets destined for this UE. A possible option is to set a fixed RLC transmission mode (i.e. AM or UM), but we do not take advantage of the variety of RLC transmission modes. Another option (more complex), is to develop a dynamic RLC transmission mode based on packet loss, radio conditions or even the application type as suggested in [93]. Both solutions are entirely compatible with our IP-centric model. Finally the RLC entity sends required data to the MAC sublayer, which performs usual treatments described in [94].

4.4 IP-centric QoS model and SDN mobile networks

As detailed above, the IP centric model needs a unique EPS bearer in order to be compatible with current 3GPP model (connected mode). Nonetheless, this IP-centric QoS model is also completely compatible to all IP based architectures.

In that regard, the recent mobile network architectures based on the Software Defined Networking (SDN) are fully compatible with the IP-centric QoS model. SDN is a new network approach where a separation between the control and data planes is done. It considers network nodes as dummy packet forwarding devices, which are controlled by a centralized entity (SDN controller). State-of-the-art SDN architectures for mobile networks are listed in [95].

Examples of these are [96, 97, 98], where the EPS bearer concept of current 3GPP mobile networks is not needed and they propose a full SDN architecture, where the IP protocol functionalities are used to fill the gap left by the 3GPP model as the mobility and the QoS. In [99], a general SDN architecture for wireless networks (i.e. LTE, WIFI, Wimax) is presented, with the aim of a global convergence. Unfortunately, in this work the QoS is not addressed, but our IP-centric QoS model can be perfectly implemented. In [100], a SDN architecture for the EPC is proposed but the bearer concept of current mobile networks is retained, with the aim of the QoS management, in a 3GPP way as in [101].

On all above mentioned works, the IP-centric QoS model may be used without any relevant modification on the proposed architectures.

5 Simulation and Performance evaluation

5.1 Network Parameters

We have evaluated the proposed cross-layer design thanks to the ns-3 simulator (version 3.21). We have modified the LENA module [102] in order to implement an IP-Aware eNB, as described in Section 4.3 . We have considered a typical outdoor scenario with 20 UEs attached to a single eNB, thus inter cell interference is not taken into account. The eNB is equipped with an omnidirectional antenna and UEs experience varying channel conditions. The RLC sublayer in the eNB is configured in UM. We used a realistic channel model with path loss and fading, descri. The path loss was simulated as described in COST-231 [103] and the fast fading as described by the Extended Pedestrian A (EPA) model using a Rayleigh multi-path fading model [104]. Given the path loss model and the other network parameters, we obtained a wide range of Signal to Interference plus Noise Ratio (SINR) values, which provided CQI values in range of $[1, 15]$.

Number of terminals	20 (random positions)
Users mobility model	RandomWalk (3 km/h)
Bandwidth	50 RB (10 MHz)
Cell coverage radius	500 m
Pathloss Model	Cost231
eNB TX Power / Noise Figure	46 dBm / 5 dB
UE TX Power / Noise Figure	24 dBm / 5 dB
Fading loss model	EPA 3 km/h (urban scenario)
AMC model	PiroEW2010
DL/UL carrier frequency	2120 / 1930 MHz
RLC Transmission Mode	UM (Unacknowledged Mode)
RLC Buffer Size	100 kbytes (70 FTP packets)

Tableau IV.2 – Simulation parameters

At the beginning of each run, UEs are placed randomly in a disc representing the cell within a distance range of 30-500m. Then, UEs move within the disc according to a Random Walk Model, at a fixed speed of 3 km/h. The simulation parameters are shown in Table IV.2 and the system configuration is as follows: The cell is connected via the P-GW to the internet. Two servers are implemented, one for VoIP and the other for FTP, Youtube and Web services. Both servers are connected to the P-GW via an over-provisioned point-to-point link in order to avoid congestion on this segment of the network. In the IP-centric scenario, each server performs the DSCP marking depending on the application.

5.2 Traffic Description

In order to have a representative scenario of current mobile data traffic we define following configuration: Each UE uses one, two, three or four applications as described in Table IV.3. One UE uses all applications, which corresponds to a tethering configuration.

Traffic Mix	Nb of UEs	**VoIP**	**YouTube**	**Web**	**FTP**
VoIP only	1	1	-	-	-
YouTube only	1	-	1	-	-
Web only	1	-	-	1	-
FTP only	1	-	-	-	1
FTP + Web	4	-	-	4	4
FTP + YouTube	4	-	4	-	4
FTP + VoIP	4	4	-	-	4
FTP + VoIP + Web	3	3	-	3	3
Tethering (all app)	1	1	1	1	1
Total	**20**	**9**	**6**	**9**	**17**

Tableau IV.3 – Traffic distribution

5.2.1 VoIP Traffic

The arrivals of new voice calls are assumed to follow a Poisson process with an average inter-arrival time λ = 7.1 seconds [105]. The call duration (seconds) follows a LogNormal distribution with parameters $\mu = 3.9$ and σ = 1.0 as was estimated in [106]. VoIP traffic is based on AMR-NB codec [107]. In order to measure the QoE of VoIP, we use the Mean Opinion Score (MOS) which is specified in ITU G.107 [33]. The codec fitting parameters to compute the MOS for AMR-NB are specified in [108], which is based in delay, Packet Loss Ratio (PLR) and jitter of each voice communication.

5.2.2 WEB Traffic

We use a realistic HTTP traffic model, as proposed by [109], which is also implemented on ns-3 simulator. We use 5 top french web sites based on Alexa's ranks in order to simplify and harmonize the performance measurements in terms of Page Load Time, see Table IV.4.

5.2.3 HAS traffic - YouTube

Most popular video streaming services (i.e. YouTube, Netflix) use HAS. HAS, split up media (i.e. video) into a series of small files called chuncks, which are then encoded

	Web Site	Avg. Size (Bytes)	Total Objs
1	boutique.orange.fr	2,366,380	129
2	Leboncoin.fr	728,592	78
3	Facebook.com	1,047,690	185
4	Lefigaro.fr	1,054,458	50
5	pole-emploi.fr	1,205,991	320

Tableau IV.4 – Web sites details

as shown in Figure IV.9 [110]. Each chunk is transmitted individually as a single web object via plain HTTP. In the course of playout of the video, the client continuously assess available bandwidth and requests successive chunks for the data rate that can be supported. Typically, the client keeps a buffer of chunks to deal with eventual network issues (e.g. latency, packet loss, connection loss).

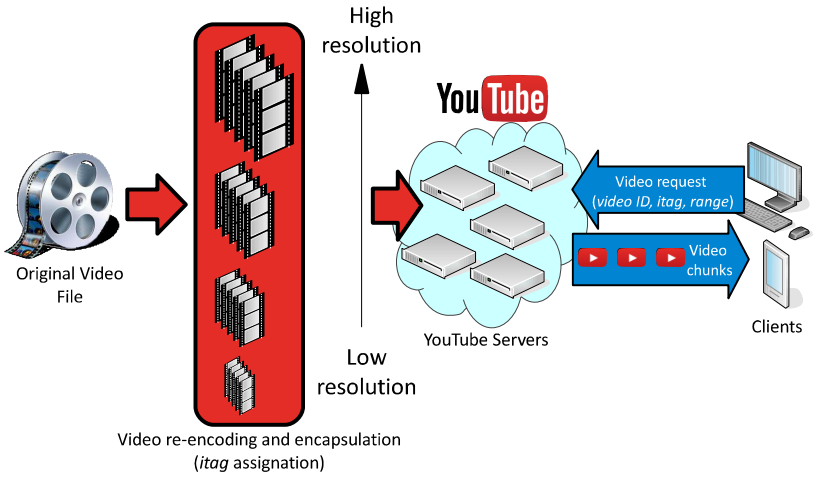


Figure IV.9 – YouTube video service

A wide range of encoding for every video file is available for clients, which can be selected according needed. Besides, a numerical identifier named "$itag$" is used in order to identify different encoding schemes of a video. The $itag$ " information is included in the HTTP requests, moreover each chunk is transmitted individually as a single web object via plain HTTP. During the Play-out of the video, the client continuously estimates available bandwidth and requests chunks for the data rate that can be supported as shown in Fig. IV.10. The client try to keep a buffer of video data to deal with eventual network issues (e.g. latency, packet loss, connection loss), in order to perform it, a buffering strategy is necessary.

We have implemented a specific module in ns-3 in order to emulate the delivery of YouTube traffic in HAS mode. For this purpose, we have taken advantage of the traffic model described on [110] and [111]. In HAS, the video object is broken out into chunks

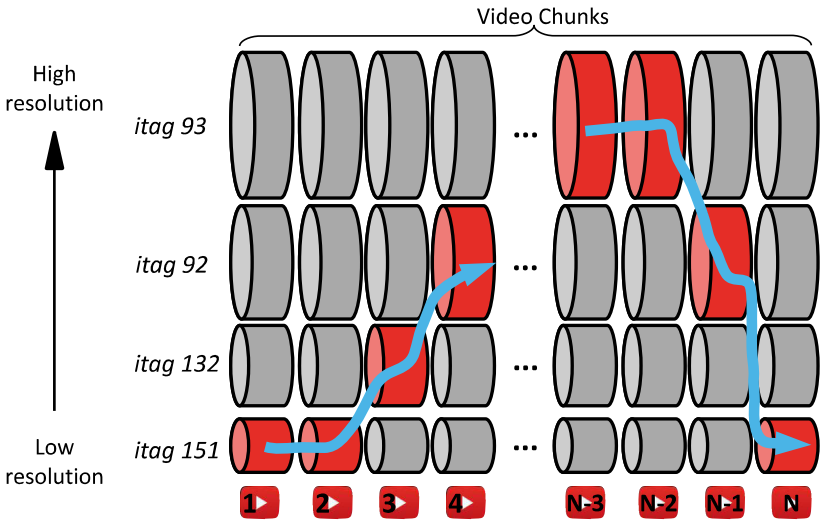


Figure IV.10 – YouTube video streaming strategy

and the server maintains several profiles for each chunk, one per video quality, that is, encoding rate. In our simulation, the encoding duration of each chunk is 5 seconds and four profiles are defined and identified by their *itag* values (see Table A.5) depending on their encoding rate. Each time the client requests a chunk from the server, it selects the best possible profile with respect to the network conditions.

Our implementation is based on a dual-threshold buffering strategy using parameters estimated in [110] in order to select the chunk profiles according to the client playback queue size. We have also set a maximum timeout for chunk request to avoid blocking situations during simulation, and we have fixed this parameter at 100 seconds. When the timeout is exceeded, the video session is stopped and a new video session starts. The duration of the whole video is fixed at 150 seconds with an exponential inter-video interval with 10 seconds of mean. You can find more details about our implementation in [112] or in Appendix A.

itag	Resolution	Encoding rate
151	128 x 72	64 kbps
132	426 x 240	266 kbps
92	426 x 240	395 kbps
93	640 x 360	758 kbps

Tableau IV.5 – YouTube video quality information

5.2.4 FTP Traffic (Background Traffic)

Some UEs support a background best effort traffic (see Table IV.3), which is represented by FTP sessions. The size of these files follows a uniform law between $[1, 5]$ Mbytes. The arrivals of new FTP sessions follow a Poisson process with an average inter-arrival time λ = 10 seconds.

5.3 Radio Schedulers description

In order to evaluate and compare performance our IP-aware QoS model to 3GPP QoS model, we use different radio schedulers. It is important to remember that in order to guarantee an End-to-End QoS, the 3GPP QoS model needs the implementation of QoS-aware radio schedulers. Therefore, we propose two QoS-aware radio schedulers (Priority Set Scheduler (PSS) and Channel and QoS Aware (CQA) scheduler) in case of the 3GPP QoS Scenarios and the PF scheduler in the Best-Effort 3GPP and IP-centric scenarios. The following is a description of proposed radio schedulers.

5.3.1 Proportional Fair Scheduler

As a reminder, a brief description of the well-known PF algorithm is outlined below.

Let k be a resource unit defined as a RB or a Resource Block Group (RBG) such as defined in [113]. Let R_i be the past average throughput computed every TTI for the UE i and $D_{i,k}$ be the estimated achievable throughput in the resource unit k. The PF priority metric $M_{i,k}$ is computed as follows:

$$M_{i,k} = \frac{D_{i,k}}{R_i}$$

For resource unit k, the PF scheduler selects UE i that maximizes $M_{i,k}$. The MAC sub layer then requests data from the RLC sublayer for each selected UEs, in accordance with the Modulation and Coding Scheme resulting from its radio conditions.

5.3.2 3GPP Schedulers

To support dedicated bearers, we selected two QoS-aware schedulers available in the ns-3 simulator:

- Priority Set Scheduler (PSS) described in [114].

- Channel and QoS Aware (CQA) scheduler described in [17].

Both algorithms are QoS-aware and perform joint time and frequency scheduling according to a two level design, as shown on Figure IV.11; CQA is more particularly adapted to real time applications. In the Time Domain (TD) a set of UEs are chosen according to TD scheduling criteria. In the Frequency Domain (FD), radio resources are distributed between the set of chosen UEs according to FD scheduling criteria.

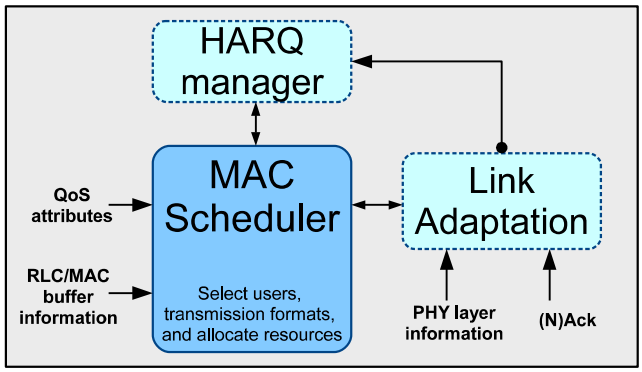


Figure IV.11 – Two levels QoS aware schedulers

We consider the PF version of each scheduler, PSS_{PF} scheduler and CQA_{PF} scheduler, which have shown better performance than the other QoS-aware schedulers implemented on the ns-3 simulator [17].

Priority Set Scheduler (PSS) - PF This scheduler aims on the one hand to control the fairness among users and on the other hand to guarantee a predefined bit rate to GBR bearers.

1. **Time Domain scheduler**: PSS defines two sets of schedulable UEs: a high priority set TD_1 including UEs that operate GBR bearers and a low priority set TD_2 including the rest of the UEs.

   The PSS Time Domain scheduler selects then N_{UE} UEs from the high priority set TD_1 as described in [114]. If the number of high priority UEs turns out to be less than N_{UE}, the PSS Time Domain scheduler selects UEs from the low priority set TD_2. Afterward, the PSS Time Domain scheduler passes on the list of the chosen UEs to the FD scheduler.

2. **Frequency Domain scheduler**: FD scheduler uses the PF metric in order to distribute the available radio resources between the previously chosen UEs. The PF metric is described in previous section 5.3.1.

Channel and QoS Aware (CQA) Scheduler - PF The CQA Scheduler is particularly intented towards real time services. It is based on an algorithm which considers the Head of Line (HOL) delay, the GBR parameters, as well as the channel quality over different resource units. The HOL delay refers to the waiting time of the first packet of the queue. The CQA scheduler is also based on joint TD and FD scheduling.

1. **Time Domain scheduler**: At each TTI the CQA scheduler selects preferentially UEs that have not yet reached their GBR and orders them by HOL delay.

   Let d_i^{HOL} be the current value of HOL delay of flow i. CQA selects then the N_{UE} UEs with the highest d_i^{HOL} and forwards the list of chosen UEs to the FD scheduler.

2. **Frequency Domain scheduler**: The FD scheduler uses the PF metric in order to allocate the available resource units. Let R_i be the past average throughput of UE i and $\widehat{d}_{i,k}$ its estimated achievable throughput in the resource unit k. For each user i operating a GBR bearer with a value of GBR_i, we define $M_{i,k}^{\text{GBR}} = \frac{GBR_i}{R_i}$. Moreover, $M_{i,k}^{\text{PF}} = \frac{\widehat{d}_{i,k}}{R_i}$ is the classical PF metric. Thus, $M_{i,k}^{\text{FD}}$ for a UE i is defined in the following way:

   $$M_{i,k}^{\text{FD}} = d_i^{\text{HOL}} M_{i,k}^{\text{GBR}} M_{i,k}^{\text{PF}}$$

   Therefore, on the resource unit k, the FD scheduler selects the UE i that maximizes the metric $M_{i,k}^{\text{FD}}$.

At the end of the day, QoS-aware schedulers implemented in ns-3 are rather limited as they do not support multi-service UEs : If several dedicated bearers are established from a given UE, the scheduler will apply the same traffic parameters (QCI, GBR) to any of them. To get around this problem, we emulate a unique UE which use "N" QoS levels through "N" UEs with the same mobility pattern. This approach was validated by the ns-3 (LENA) community.

5.4 Scenarios description

In order to evaluate the performance of our IP-centric model, we consider four scenarios:

- ▷ **Scenario 1** - Mono bearer, Best Effort scheme, PF scheduler. This scenario addresses the case of current deployments in LTE networks supported by a single bearer in Best Effort with a basic PF scheduler.
- ▷ **Scenario 2, 3** - Multi-bearer, CQA_{PF} or PSS_{PF} scheduler. The two following scenarios simulate standard 3GPP multi-bearer configurations; in these cases, the UEs

establish dedicated EPS bearers for non Best-Effort applications (i.e. VoIP, Youtube and Web). Whether in scenario 2 or scenario 3, VoIP bearers are affected with a QCI equal to 1 and a GBR of 44 kbps. YouTube bearers are assigned a QCI equal to 4 and a GBR equal to 1,5 Mbps, and finally, Web is supported by non-GBR dedicated bearers with a QCI equal to 9.

▷ **Scenario 4**- Mono-bearer, IP-centric model, PF scheduler. This scenario simulates an IP-aware eNB as described in the previous sections.

MAC Schedulers	**3GPP Scenarios:** (a) Best Effort (PF) (b) CQA_{PF} (c) PSS_{PF} **IP-Aware scenario:** (d) PF
QoS 3GPP dedicated bearer parameters (CQA and PSS)	
VoIP: QCI = 1, GBR = 44 kbps, conversational voice	
YouTube: QCI = 4, GBR = 1.5 Mbps, non conversational video	
Web: QCI = 9, Non-GBR, TCP-based	

Tableau IV.6 – Scenario parameters

In case of the scenarios 2, 3 and 4, four levels of QoS are implemented. In scenarios 2 and 3 the QCI parameter determines the QoS level that is specified in [115]. In scenario 4, an IP strict priority queuing system is implemented. Thus, scenarios 2, 3 and 4 implement four levels of QoS, which in descending order of priority are: VoIP, YouTube, Web and FTP. Table IV.6 summarizes the description of the four scenarios proposed in this section.

5.5 Performance analysis

We present here performance results related to the scenarios described previously. Each simulation run lasts for 1000 seconds, with a warm-up time of 5 seconds where statistics are not collected, and is replicated three times with different seeds. Applications are started at a random time uniformly distributed in $[1, 5]$ seconds.

We show the performance of each service and give in the following simple conclusions. We then consider all the services together and go deeper in the analysis.

Figure IV.12 depicts the cumulative Cumulative Distribution Functions (CDFs) of cell throughput in the analysed scenarios (Best Effort, CQA_{PF}, PSS_{PF} and IP-aware). Note that the cell throughput in the CQA_{PF} scenario is particularly degraded, because the CQA algorithm strongly favors GBR bearer, even in bad radio conditions, at the expense of

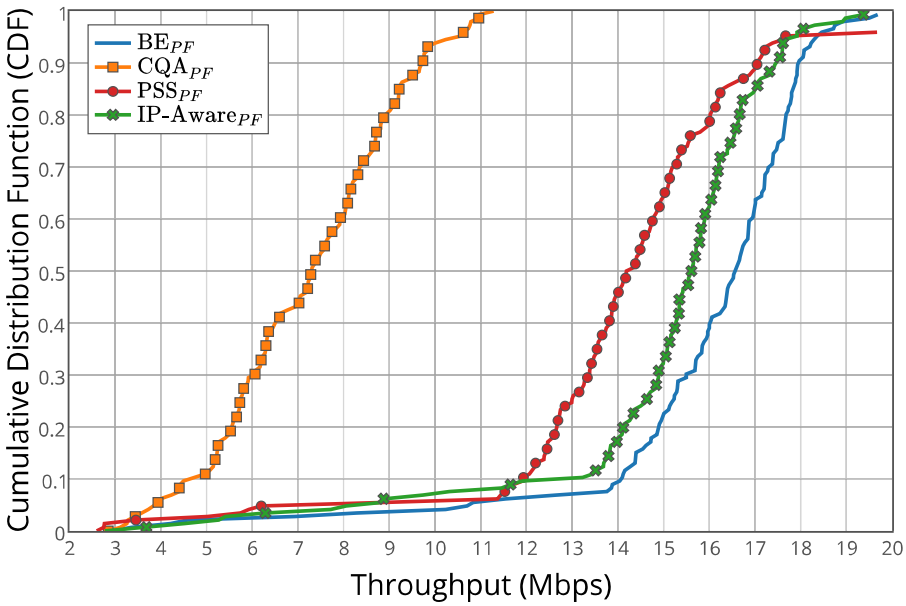


Figure IV.12 – Cell Throughput performance

others bearers [17]. On the contrary, the Best Effort PF scheme achieves the highest cell throughput, closely followed by the IP-centric scenario; this is because neither of them modify the PF resource allocation strategy. Otherwise, the cell throughput of the PSS scenario is only moderately degraded as the PSS_{PF} algorithm is less aggressive than CQA_{PF}.

Figure IV.13 provides the boxplots of the VoIP MOS. The median is given by the central mark and the borders of the box are the 0.25 and 0.75 percentiles. It can be seen from this figure that the MOS for the Best Effort scheme is poor as its median is around 1.8 (Poor Quality), whereas the CQA, PSS and IP-aware schemes have similar MOS, around 4.1 (Good Quality).

Regarding YouTube traffic (Figure IV.14), almost 70% of chunks are delivered in the highest quality in the CQA scenario; the other schemes obtain a rather mixed distribution in chunk quality.

Figure IV.16 provides a boxplot display of the YouTube's flows throughputs. While the PSS and the IP-Aware schemes obtain a median throughput around 1.2 Mbps, the Best Effort median throughput is limited around 400 kbps. The CQA scheme is far ahead with an impressive median throughput of 3.8 Mbps.

Figure IV.15 shows the mean of the initial buffering time for YouTube videos for each scenario. The worst case is performed by the Best Effort scheme that has a first buffering time around 9.87 s. The CQA scenario has the lowest mean buffering time, that is around 3,88 seconds. PSS and IP-Aware schemes also improve the buffering times, which are around 4.94 seconds and 5.76 seconds respectively. The scenarios implementing QoS

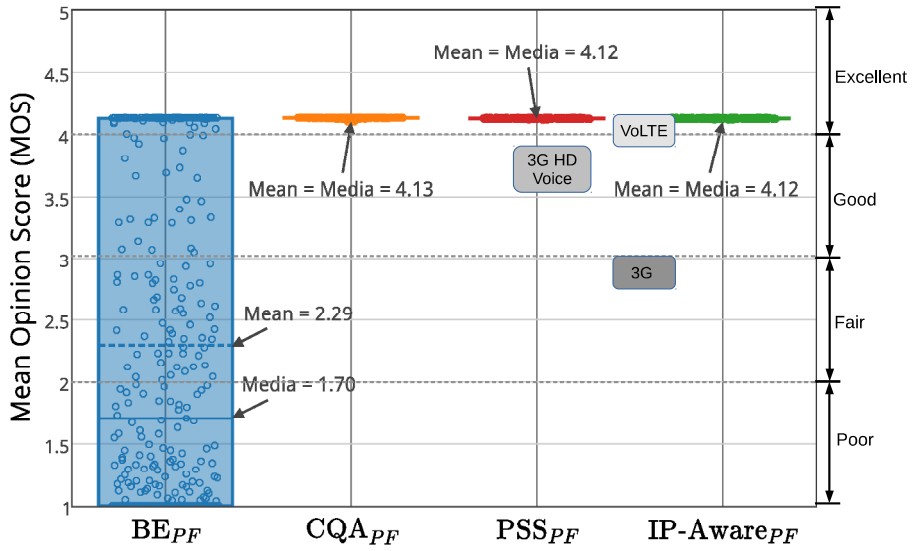


Figure IV.13 – VoIP QoE in terms of Mean Score Opinion

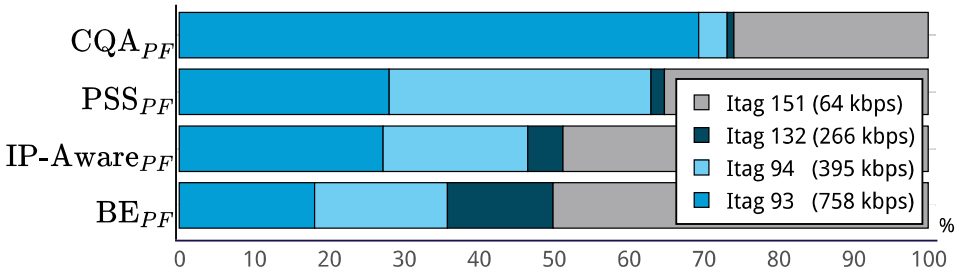


Figure IV.14 – YouTube chunks distribution

mechanisms (i.e. CQA, PSS and IP-Aware) reduce by around half the mean buffering time performed by the Best Effort scheme. It should be noted that this Best Effort scheme is widely used today by the majority of operators.

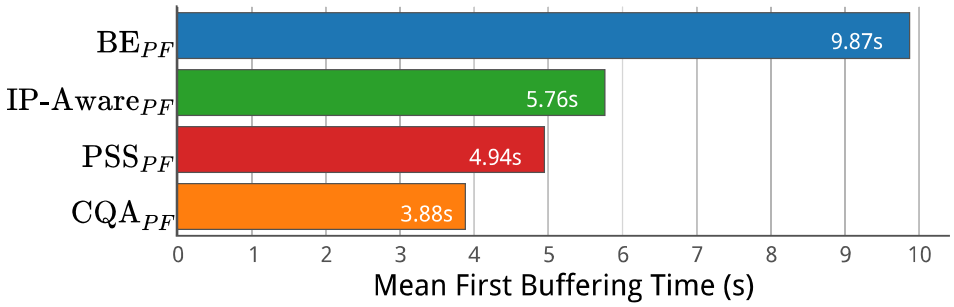


Figure IV.15 – YouTube First Buffering Time

Figure IV.17 provides the Page Load Time (PLT) of the web sites listed in Table IV.4. As mentioned previously, CQA algorithm strongly favors GBR bearers at the expense of

web traffic, supported by non-GBR bearers. Figure IV.17 shows that the PLT systematically exceed the timeout fixed at 30 seconds in the CQA scenario. Note also that the Best Effort scheme obtains the longer Page Load Time, but no timeout, whereas the IP-Aware scenario achieves better performance than the PSS scheme in most cases, except for the largest web site (boutique.orange.fr) where the PSS scheme has a advantage of 1 second in PLT.

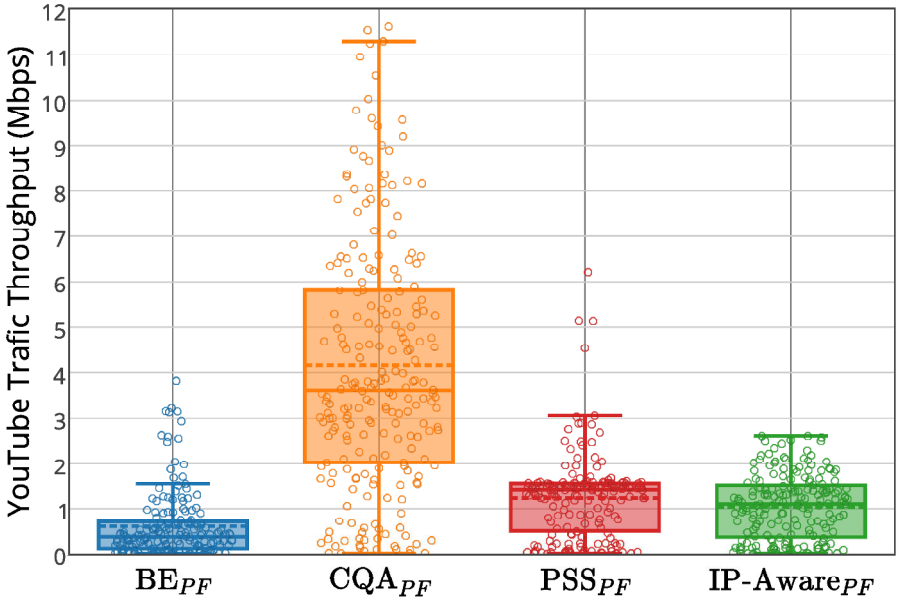


Figure IV.16 – YouTube's Flows Throughput

In a nutshell, the simulation results show that the 3GPP scenarios' QoS performances are directly related to the aggressiveness of the radio scheduler. The most aggressive scheduler (CQA) obtains outstanding QoS for flows supported by GBR dedicated bearers, at the expense of flows supported by non-GBR bearers. Note that this behaviour is strongly detrimental to the global cell capacity, which is dramatically reduced for 3GPP schemes: The better the QoS for GBR flows, the more severely affected the cell capacity is.

From an operational point of view, this means that such schemes would certainly required eNB densification, with associated costs and issues.

Regarding VoIP, our IP-centric model shows performances similar to 3GPP QoS schemes. VoIP generates a moderate bit rate, even for UEs in bad radio conditions, the scheduler can allocate enough resources to provide the required throughput. Hence, the mere prioritization at IP level is sufficient to meet the VoIP delay requirements. This control on packet delay explains the dramatic gap between IP-Aware and Best Effort scenarios regarding VoIP. Note that the BE and the IP-centric models allocate the same amount of resources to UEs (thanks to the PF scheduler), and only differ in packet interleaving on the radio interface.

The IP-Aware scheme shows medium performances for YouTube flows in our simulation framework, but still outperforms the BE model. Indeed, when UEs are in poor

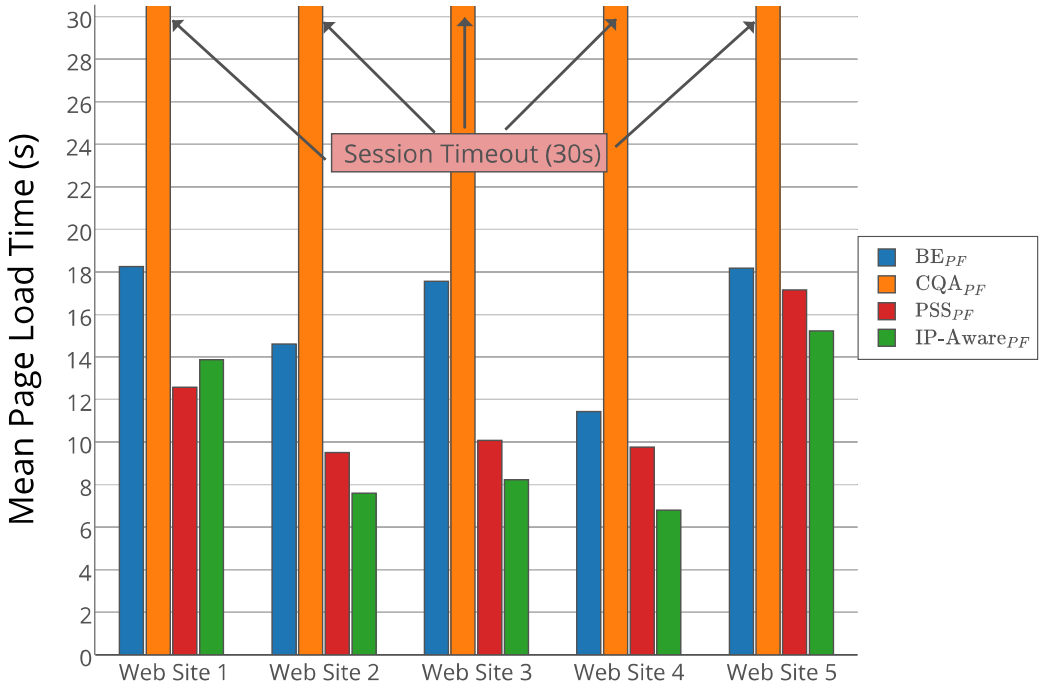


Figure IV.17 – Web Page Load Time

radio conditions, they do not obtain enough resources to support high definition video chunks. The impacts of this medium QoS level should be evaluated in the light of customer experience (i.e. depending on screen size). It has also to be seen against the global cell capacity preservation and the excellent performances of web services obtained in the IP-Aware scenario.

Traffic	Best-Effort	3GPP CQA	3GPP PSS	IP-Aware
VoIP	NOK	OK	OK	OK
YouTube	?	OK	?	?
Web	OK	NOK	OK	OK
FTP	OK	NOK	OK	OK

Tableau IV.7 – Performance analysis summary

Table IV.7 summarises performance analysis. Concerning YouTube traffic, is very hard to conclude if the IP-aware performance can satisfy customer expectations, since it will depend on terminals specifications as the screen size. In case of a small terminal screen, the differences between chunks types (video quality) will not be evident and the key parameter to evaluate the QoE will be the first buffering time. On the other hand, IP-aware provides a good performance for the other traffic sources and has quite similar performance to the 3GPP QoS schemes. Hence, we can conclude that IP-Aware is efficient when multiple flows with different QoS requirements are simultaneously supported by a terminal.

6 Conclusions

Evidence shows that mobile operators have made very little use of the 3GPP connection oriented QoS model, and the current trend is a "Best Effort for all" model for data traffic, except maybe for the VoLTE. This low revenue model - based on a unique bearer per customer – would probably lead to poor customer experience when carrying multiple applications within the same bearer.

The IP-centric QoS management described in this chapter allows for QoS differentiation within such a mono-bearer scheme. This solution is easy to deploy and operate and above all, it is perfectly in line with usual Internet paradigms, based on connectionless packet-oriented QoS management. However, an IP multiplexing stage should be added before the radio scheduling in the eNB. This makes possible not only the re-use of standard IP functions (e.g. DiffServ), but also the activation of advanced IP mechanisms (active queue management, flow aware management, etc) in the eNB.

Simulations have highlighted the efficiency of the IP-aware scheme notably when multiple flows with different QoS requirements are simultaneously supported by a terminal. In particular, it has been shown that the VoIP MOS obtained in the IP-aware scenario is not significantly different from those obtained with standardized multi-bearer scenarios. Moreover, the IP-aware scheme preserves the global throughput of the cell, contrary to multi-bearer scenarios with GBR bearers, which may strongly degrade the global cell capacity.

For these reasons, we believe that the IP-centric model offers a promising scheme for mobile networks as it provides for cost-effective QoS management with a performance level similar to those of standardized solutions.

Finally, note that the underlying IP-centric model is quite in line with QoS management in fixed internet networks. In this extend, the IP-centric model allows for a unified solution in a fixed/mobile convergent world, which opens the door to access-agnostic applications.

Chapter V

Slo-Mo: An Implicit Mechanism to Improve QoS in Mobile Networks

1 Introduction

Customer experience on mobile networks has significantly improved for the last few years thanks to huge deployment of radio capacities and advanced content delivery architectures such as Content Delivery Network (CDN). However, access networks still constitute the bottleneck in most mobile networks and congestion, if any, typically occurs on the radio segment, as shown in Figure V.1.

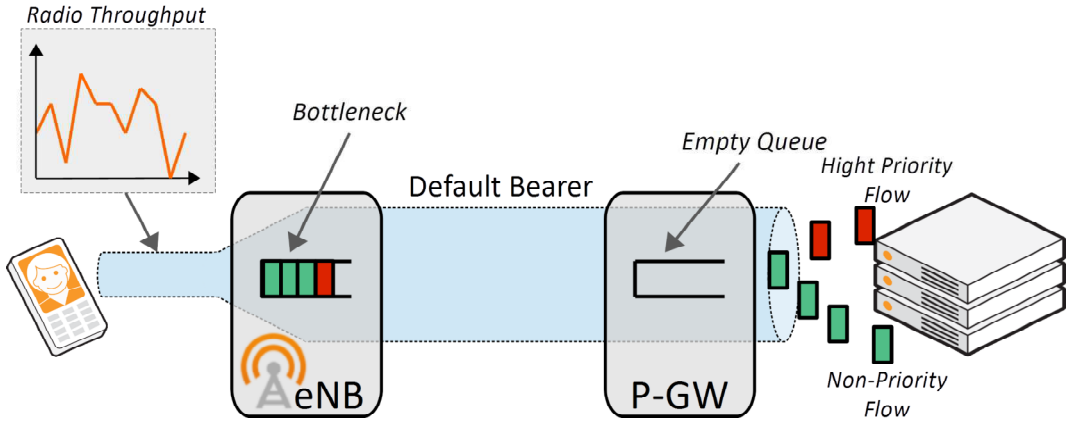


Figure V.1 – Traffic bottlenecks in LTE/EPC networks

As a matter of fact, standardized QoS features remain sparsely deployed in mobile networks, thereby giving way to basic best-effort hegemony due to mono-bearer architectures. In such a scheme, congestion of the radio segment leads to poor customer experience, as sensitive flows cannot benefit from specific treatments. Alternative solutions, such as the

IP-centric QoS model we presented in chapter IV addresses properly this issue through packet differentiation. However, as it requires enhancements in eNB and probably prior standardization, this scheme can not be expected in mobile networks before a rather long time. In the meanwhile, lightweight QoS management solutions are greatly hoped for.

For this purpose, we introduce in this chapter an implicit mechanism called Slo-Mo that aims at improving QoS in mobile networks without any modification in RAN nodes nor added protocols. Slo-Mo have been developed on our ns-3 simulator and its performances have been assessed thanks to this tool. The parameters of the algorithm have then been tuned so as to improve QoS for sensitive flows while preserving the cell capacity.

2 TCP flow control in mobile networks

Contents are delivered today through HTTP-based robust applications on TCP transport, which provides congestion control thanks to its usual algorithms (i.e. slow start, congestion avoidance, fast retransmit, fast recovery). As a matter of fact, in the current Best Effort mono-bearer scheme commonly encountered in mobile networks, QoS relies mainly on the supposed TCP flow control good behaviour; indeed, this control aims at matching the source emissions with the available bandwidth, thanks to TCP careful monitoring of losses and delays.

2.1 TCP on radio medium

Unhapilly, TCP congestion control was not designed in view of radio support. Similarly, Radio Access Networks have mainly been designed and evaluated regardless of the atop applications. This leads sometimes to discrepancies between these two layers and results in poor performances: When the TCP driven source emissions exceed consistently the radio resource allocation, latency and losses increase in the RLC buffer and the resulting customer experience degrades, as evidenced in several papers, such as [51, 116, 117]. Such event typically occurs when the Round-Trip Time (RTT) of the TCP connection is consistently larger than the radio resource allocation slot. This is commonly encountered in cellular networks of today, where the RTT is on the order of several tens of milliseconds compared to the one millisecond TTI of the radio scheduler. This means that the radio allocation changes significantly faster than the TCP adaptation mechanism. Degradation of the experience for sensitive flows is not the only damaging effect of this discrepancy : dramatic impact on TCP achievable load has also been mentionned for example in [51], where the authors mention an utilization rate of radio links inferior to 50% when using TCP for large flows on LTE networks. This is clearly not acceptable in terms of network optimization, nor customer experience.

In [118], the authors pointed out other various side effects of usual radio features -

such as IDLE/CONNNECTED transitions - on TCP performances. In addition, concurrent retransmissions on RLC and TCP layers contribute to the degradation of the system efficiency too.

2.2 Cross-layer solutions

The scientific community has recently emitted several cross-layer proposals to tackle this issue. Some works proposed that the network be able to send explicit information on the radio link state to the TCP source. The source must adapt the TCP flow control according to the radio characteristics. In [119], the authors propose a Mobile Throughput Guidance Exposure mechanism where a throughput guidance information per user is sent to the TCP video server by the network. This information can be used to assist TCP congestion control decisions and also to ensure that the application level coding matches the estimated capacity at the radio downlink (e.g. HAS).

Alternatively, it is also proposed that the customer terminal be able to predict the radio resource allocation and transmit this information to the TCP source or other transport protocol/application. In [120], the authors propose a cross-Layer congestion control protocol for cellular networks called CQIC. It estimates the cellular link capacity exploiting information about previous data rates (based on CQI observation) to predict the future allocated data rates. Then, the predicted link capacity is communicated to the source in order to optimise the data transfer and avoid congestion.

In [121], the authors propose a cross-Layer bandwidth estimator for adaptive HTTP video streaming (e.g. Dynamic Adaptive Streaming over HTTP (DASH)) over LTE called piStream. Unlike CQIC, piStream not only explores its allocated PRBs' information but also explores unused resource in order to have an overall view of available resources in the cell and thus to have a highly reliable data rate estimation. For that purpose, piStream implements a complex Radio Resource Monitor (RMon) that is not yet feasible in the most current LTE hardware.

In [122], the authors propose an end-to-end transport protocol for interactive applications over mobile networks called Sprout. Unlike TCP congestion control mechanism, Sprout implements a stochastic forecast framework to infer the uncertain dynamics of the network path. This framework is based on the packets arrival time observation, which allows to forecast how many bytes may be sent by the sender, and hence avoid congestion.

Finally, source emissions may also be implicitly controlled via distant shaping/pacing or other action on their traffic profiles; this approach is currently under study at Orange Labs, which was initiated in the context of this thesis, and it is presented in this chapter.

2.3 The Bufferbloat phenomenon

Bufferbloat has been subjected to an abundant literature since the late 90s [123, 124] and has acquired significant relevance since late 2010 after studies performed by Gettys and Nichols [125]. Behind this colorful term is hidden a phenomenon very frequently encountered in fixed Internet networks of today. More precisely, the late 90's witnessed an unprecedented growth in the buffer size of network elements in fixed IP networks. As a consequence, loss rates became very low, as the occurrence of buffer overflows plummeted in such over-buffered networks. In return, latency dramatically increased, because TCP congestion control algorithms tends to fill these network buffers permanently as depicted in [125, 126]. This phenomenon is called "Bufferbloat", that is a perpetual large backlog in the congestion point. The related excessive end-to-end delay is considered as one of the major cause of the QoE degradation of the Internet today [127].

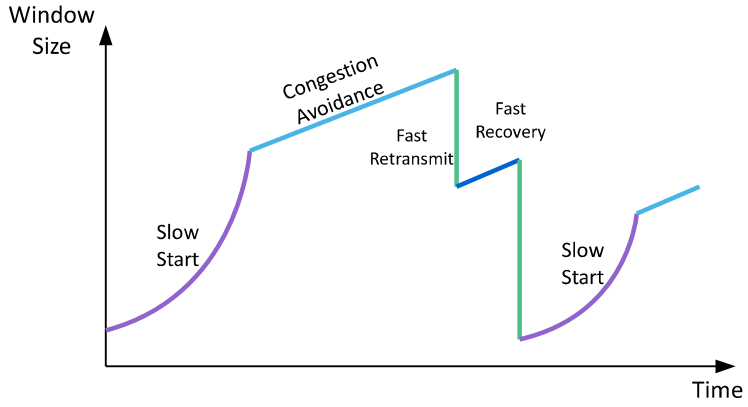


Figure V.2 – TCP congestion control strategies

Indeed, the TCP congestion control mainly consists in well known mechanisms such as slow start, congestion avoidance, loss recovery as shown in Figure V.2. These algorithms are based on a congestion window (cwnd) [128], which controls the number of packets allowed to be sent without waiting for ACKnowledgment (ACK). Packet loss is also taken into account [129, 130] to determine the congestion window size, that is the source emission rate.

In order to understand the Bufferbloat phenomenon let's look at the following illustration based on [131]. Figure V.3 shows a TCP connection at the steady-state, one RTT after the TCP startup. When a TCP session starts, the sender launches its window of "n" packets (TCP's initial congestion window), which means that TCP sends "n" packets continuously, then stops and waits for ACKs. When these packets arrive to the bottleneck (bandwidth reduction), the transmission time for each packet is longer than in the previous segment. Thus, the bottleneck spaces out the packets, and they this spacing (ϕ) pattern is retained till the receiver. When a packet arrives to the receiver, it generates an ACK, so the ACK stream reflects the bottleneck spacing (ϕ) on the return path too. If no loss occurs, each ACK arrival at the sender increases the congestion window by one and it triggers the emission

of data packet following once again the spacing (ϕ). Therefore, after just one RTT, the packet arrival rate at the bottleneck exactly equals its departure rate, and its queue turns into a persistent large backlog. As a consequence, latency dramatically increases, and may badly affect customer experience for sensitive applications.

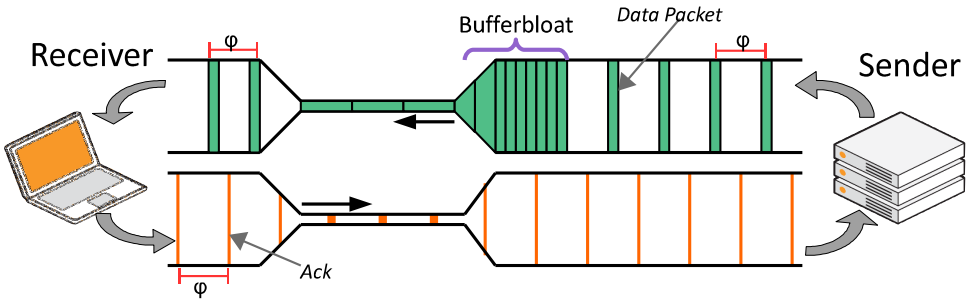


Figure V.3 – Bufferbloat illustration

Ironically, packet loss, which has been considered as the scourge of early IP networks, was then almost desirable; indeed, losses induce TCP data rate reduction, thus shrinking queues backlogs and limiting latency.

Active Queue Managements (AQMs) aim at tackling the Bufferbloat phenomenon by dropping carrefully selected packets. More precisely, according to [132], AQMs' goals are to manage queues in order to absorb packet bursts while avoiding the generation of standing queues and prevent flow synchronization;

The most well-known AQMs algorithms are Controlled Delay (CoDel) [131, 133], Proportional Integral controller Enhanced (PIE) [134], Adaptive RED (ARED) [135] and their variants. These AQMs are mainly based on queue state information as queue size, packets' sojourn time, etc. and trigger some actions (e.g. packet drop) when a threshold (fixed or dynamic) is exceeded in order to induce a TCP data rate reduction.

3 Slo-Mo main features

Slo-Mo is a patented mechanism (Orange/B-com[1] co-innovation [136]) which aims at mitigating degradations of customer experience in mobile networks, especially when they are operated in Best Effort through a one-bearer-for-all scheme. Slo-Mo was proposed, designed and evaluated as part of the work carried out in this thesis, and which is detailed below.

[1] B-com is a french Research Institute of Technology

3.1 Motivations and principle

As already mentioned, core (e.g. EPC) and backhaul networks are not usually subject to congestion, and the bottleneck of a mobile communication is typically located on the radio interface. When a mono-bearer architectures is used, customers flows are multiplexed in a unique bearer and all the packets sent to a given UE are supported through one FIFO buffer in the eNB whatever their QoS needs.

As soon as the buffer dedicated to this UE in the eNB is non-empty, then all the flows it supports, including the sensitive ones, undergo QoS degradation in terms of delay or loss. Only a very short backlog in the eNB could allow QoS preservation for sensitive flows. Slo-Mo aims at maintaining almost empty queues in the eNB through queuing de-location in the P-GW.

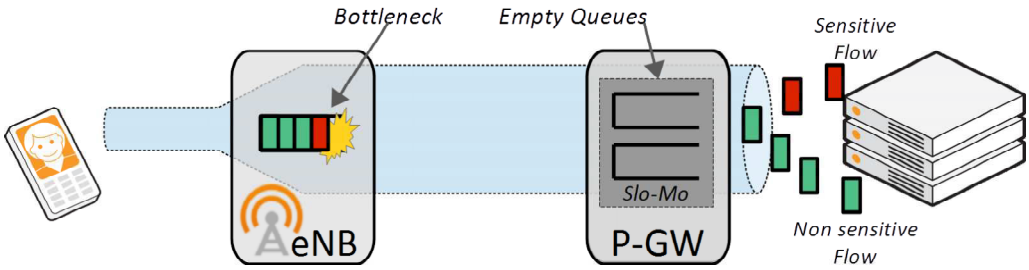


Figure V.4 – Slo-Mo Mechanism: initial bottleneck location

More precisely, Slo-Mo operates according to the following principles, as shown in Figures V.4 and V.5:

1. A bottleneck is created in the P-GW, and its rate is dynamically controlled so as to reflect the rate really available on the radio segment. In this way, the natural bottleneck is de-located from the eNB to the P-GW, preventing then damaging congestions in the eNB. Indeed, keeping the lowest possible backlog in eNB queues ensures a negligible contribution of this node to loss and latency.

2. Typical IP QoS management mechanisms are activated in the P-GW, such as packet prioritisation based on their header as defined in [137]. Contrary to eNBs, the P-GW is an IP level node which typically implements DiffServ oriented mechanisms allowing differentiated treatments, for example based on the DSCP bits of each packet. These mechanisms, typical of fixed IP networks, are not specific to Slo-Mo. However, none of them are available in the eNB as it is not an IP node.

The main contribution of Slo-Mo is the remote implicit tracking of the radio rate (item 1 above). This feature is rather innovative and will be described in details in the following section.

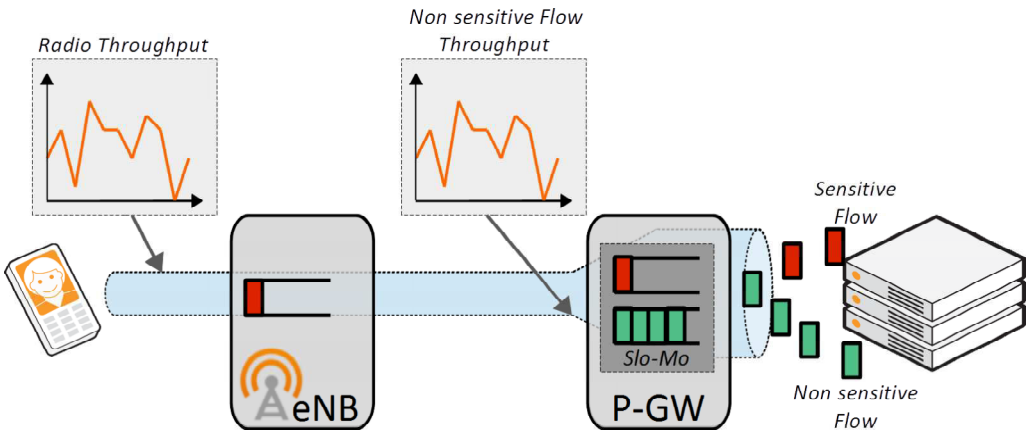


Figure V.5 – Slo-Mo Mechanism: bottleneck de-location

3.2 Rate tracking

Slo-Mo is typically implemented in the P-GW, which constitutes an end-point of each bearer. As a consequence, when a mono-bearer scheme is deployed the P-GW, as the eNB, has got a per bearer view on traffic. Slo-Mo aims at adapting the bearer rate in the P-GW so as to reflect the available resources allocated to this specific bearer in the eNB. As already said, Slo-Mo does not require any explicit communication between the eNB and the P-GW. It has been designed as a local mechanism implemented in the P-GW, based on the careful observation of its queues. Roughly, Slo-Mo rate tracking operates as follows :

1. Rate decrease in the P-GW: As long as the queue related to a given bearer in the P-GW does not build up, the rate of this bearer is reduced. Indeed, if the P-GW buffer is about empty, then it means that the bottleneck is located elsewhere, probably in the eNB. As we want to avoid QoS degradations in the eNB, Slo-Mo reduces the rate of the corresponding bearer in the P-GW, and dries up consequently the supposed downstream eNB queue. Due to this rate decrease, the eNB queue tends to disappear, and conversely the corresponding P-GW queue tends to build up. The bottleneck has been de-located from the eNB to the P-GW.

2. Rate increase: When a bufferbloat is locally detected in the P-GW, then the bearer rate is increased. Indeed, if a permanent large queue build up in the P-GW, this means that this node constitutes a long-lasting and significant bottleneck for this communication, which is sub-optimal. To avoid wasting valuable radio resources, the bearer rate is then increased in the P-GW so as to relax this bottleneck. The bufferbloat phenomenon is detected using an AQM algorithm, such as CoDel, PIE or ARED.

3.3 QoS Differentiation in the P-GW

Slo-Mo relies on the de-location of the queues from the eNB to the P-GW. It is then convenient to manage these queues in the P-GW thanks to typical IP QoS management mechanisms. Indeed, most of these mechanisms are implemented by default in any IP nodes according to the IETF DiffServ specification.

In a DiffServ domain, packets benefit from differentiated treatments in DiffServ nodes according to the DSCP value set in their header. These differentiated treatments are typically implemented through priority queueing disciplines, for example, several queues with Head of Line (HOL) priorities. Other scheduling policies are also possible, such as Weighted Fair Queueing (several queues with a specific weight per queue). A mix of the previous policies is also very commonly encountered in IP networks, for example a first queue in strict priority and the subsequent ones served in a Weighted Fair Queueing manner. In any case, IP nodes have at their disposal a great number of QoS mechanisms for packet/flow differentiation. Whatever the chosen IP level mechanism, Slo-Mo is compatible with it.

3.4 TCP interactions with Slo-Mo

Figures V.6 and V.7 represent the adaptation of the TCP source rate to the bottleneck capacity, which is de-located from the eNB to the P-GW thanks to Slo-Mo. Note that the entire mechanism relies on the TCP elastic behaviour, which tends to align the source emission rate on the bottleneck rate.

On Figure V.6, the radio segment is the bottleneck on the communication path and acts as a physical spacer for downstream packets; If the RLC buffer is large enough, no losses occur and the source emission rate matches the radio bottleneck located in the eNB, thanks to the regular spacing of upstream ACK. Consequently, the P-GW buffer is about empty. Slo-Mo then reduces the P-GW rate for this particular UE; As long as the P-GW bearer rate exceeds the radio bottleneck, the P-GW buffer remains about empty. When the bearer rate in the P-GW finally falls behind the one in the eNB, then the bottleneck is de-located in the P-GW as depicted on figure V.7. From that moment, the TCP emissions pattern is no more guided by the eNB bottleneck, but by the Slo-Mo bottleneck located in the P-GW.

A bufferbloat appears then gradually in the P-GW, as it is now the most stringent bottleneck. Note that the P-GW buffer depicted here may be very large, as it only supports delay-tolerant flows. Indeed, sensitive flow are supported by a separate high priority queue, not shown on these figures for the sake of clarity.

When a large bufferbloat builds up in the P-GW, then Slo-Mo reacts by increasing the bearer rate in the P-GW, as it is now supposed that the Slo-Mo bottleneck is too stringent, and then probably smaller than the resources available at radio level.

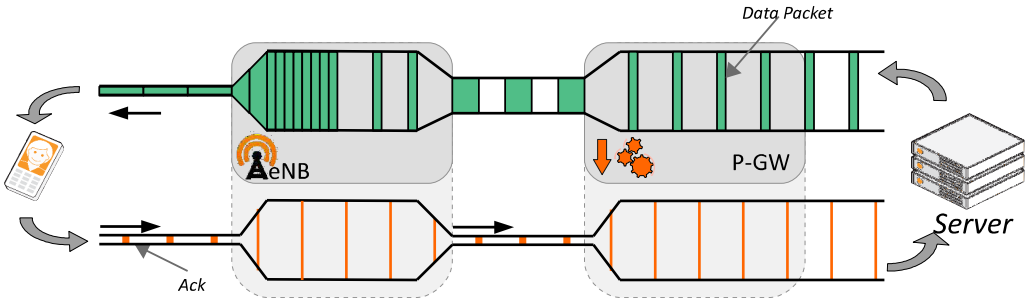


Figure V.6 – Slo-Mo: bit-rate decrease *(High priority queue not depicted)*

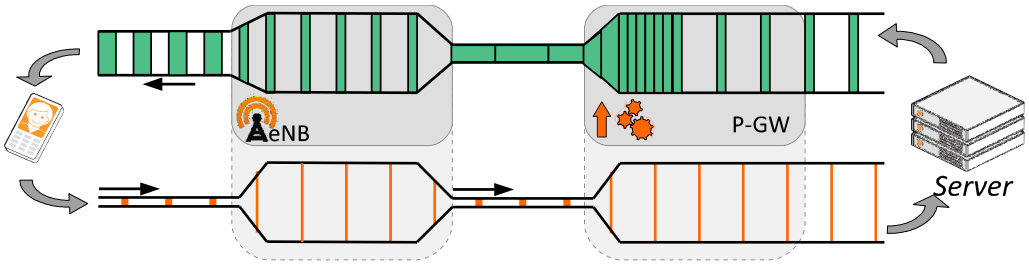


Figure V.7 – Slo-Mo: bit-rate increase

4 Slo-Mo possible implementation

Based on the previous principles, we propose a straightforward design for Slo-Mo, based on a closed loop for the rate tracking feature, and on a class-based queuing system for the QoS differentiation function. For each UE, a Slo-Mo entity is instantiated, made of a queueing system served at a rate depending on its own state.

4.1 QoS Differentiation through Priority Queueing

Incoming packets are sent to the relevant Slo-Mo entity according to their IP destination address (i.e. their UE). If this address is not known, a Slo-Mo entity is created. Each Slo-Mo entity is composed of a queueing system made of a set of two queues, one for high priority flows and the other for low-priority flows; The high priority queue has got a strict priority on the low priority queue. The packet priority, and then the queue it is sent to, is indicated by the DSCP marking of its header.

4.2 Rate tracking

Based on the already exposed principles, we propose a specific implementation for the Slo-Mo rate tracking inspired by the CoDel AQM algorithm. This first implementation aims at validating the Slo-Mo's concept and lays the foundations of future studies and

improvements.

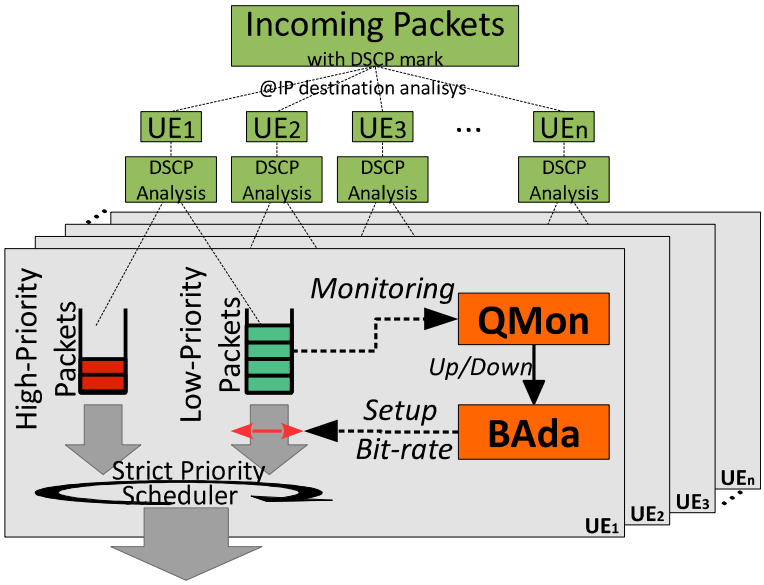


Figure V.8 – Slo-Mo System Architecture (n Slo-Mo entities for n UEs)

4.2.1 Refresh on CoDel

In the chosen solution, the server rate of the queueing system attached to a given UE is controlled by a modified CoDel algorithm. Remember that CoDel doesn't change the packets scheduled times, but drops some packets when the queue is subject to bufferbloat so as to induce a TCP rate fallback [133]. More precisely, CoDel detects bufferbloat in a queue through the monitoring of the packet sojourn times. For this purpose, each incoming packet is tagged with its arrival time. When a packet is de-queued, its packet delay (D_p) is computed as the difference between the current time and the tagged arrival time.

When the packet delay exceeds a predefined target delay D_t, the algorithm enters a "dropping state" and a timer is set so as to determine the next drop time, N_d. At time N_d, a packet is discarded if the system is still in dropping state.

The system exits the dropping state with the first packet denoting a sojourn time under the target delay; When the system is not in the dropping state, packets are then simply forwarded.

This behaviour aims at filtering transient phenomenons where a few number of packets are affected with long sojourn times. Indeed, bufferbloat is diagnosed only when the target delay is consistently exceeded for a significant number of packets, thus denoting a permanent state. Note that in case of persistent bufferbloat after a first drop (that is, persistent dropping state), the timer triggering the next drop time is reduced in order to make the algorithm more aggressive.

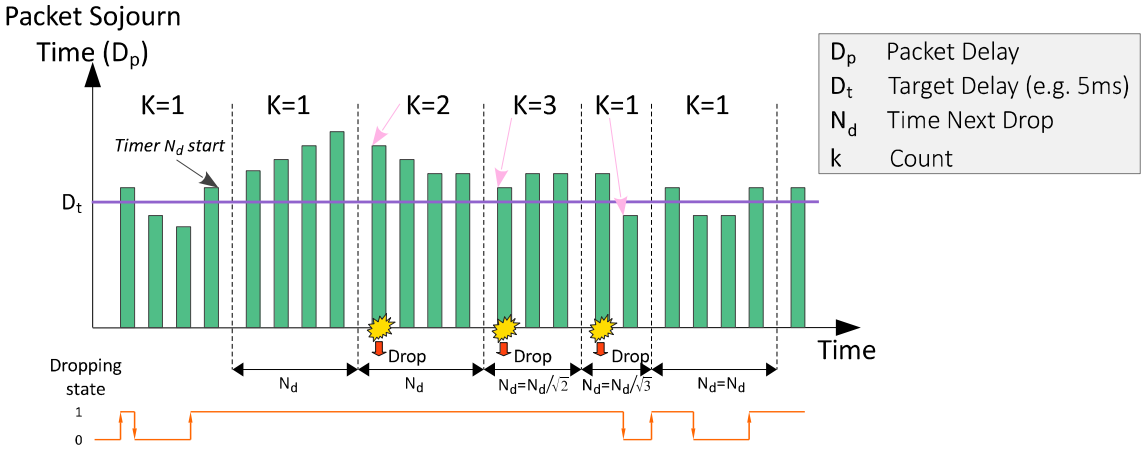


Figure V.9 – CoDel example

For example, on Figure V.9, a packet is dropped after the initialisation of N_d if and only if every packet monitored during this interval exceeds the Target Delay D_t. During the first cycle, packet delays are mainly under D_t, then no packet is discarded. During the second cycle, D_t is exceeded by every packets, then a packet is discarded at the end of the cycle. During the third cycle, D_t is once again consistently exceeded, then a packet is dropped and the next cycle is reduced to $N_d/\sqrt{2}$. During the fifth cycle, one packet delay is under D_t; all parameters are then reset.

4.2.2 CoDel-like implementation

Derived from CoDel, Slo-Mo CoDel controls the server rate of each queueing system (i.e. UE) in the P-GW in order to :

- slow down the server when sojourn times in the low priority queue do not denote bufferbloat, as it suggests that the eNB hosts an unwanted bottleneck.
- speed up the server when sojourn times in the low priority queue denote a bufferbloat so as too relax the artificial bottleneck we have just created.

The aim is to "turn around" the bufferbloat phenomenon in the P-GW without the typical packet-drop feature of CoDel as described above. Figure V.10 shows the block diagram of our algorithm based on CoDel in order to control the server rate of an UE queueing system. As with CoDel, each time a packet is dequeued, its packet delay (D_p) is computed and compared to the target delay (D_t).

When the packet delay is short (under the target delay), it denotes empty queues in the P-GW; As SloMo aims at maintaining a bottleneck in this node, the server rate should be decreased. The UE server rate is then decreased for each forwarded packet (Rate DOWN).

When the packet delay is longer (beyond the target delay), the system enters a state equivalent to the Codel "dropping state". The timer Time Next Drop (N_d) is first set, and when it expires, the server rate is increased (rate UP). In case of persistent dropping state, N_d is reduced at each cycle K ($N_d = Now + \frac{\text{Inter}}{\sqrt{k}}$).

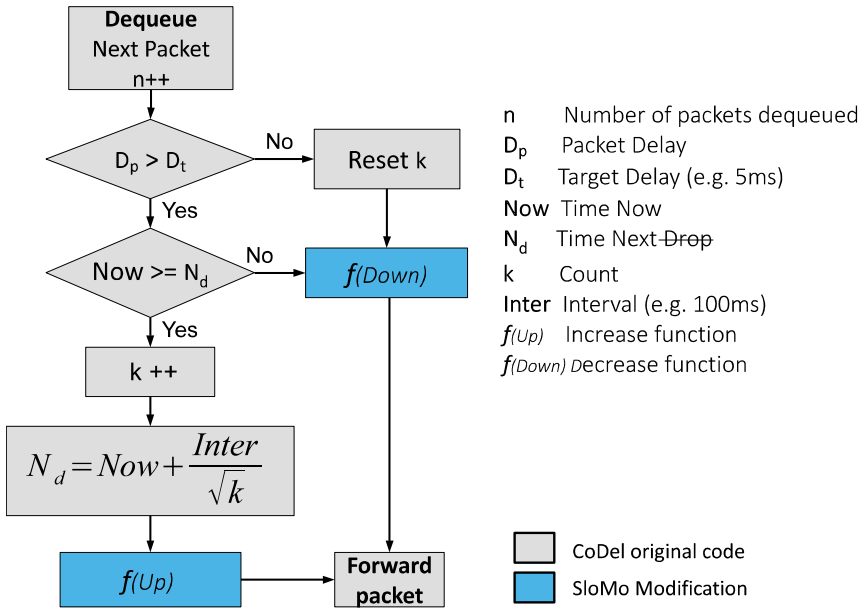


Figure V.10 – QMon simplified block diagram based on CoDel criteria

The system exits the dropping state with the first packet with a sojourn time under the target delay; As said before, the server rate is then reduced at each forwarded packet (rate DOWN) when not in the dropping state, as shown in Figure V.11.

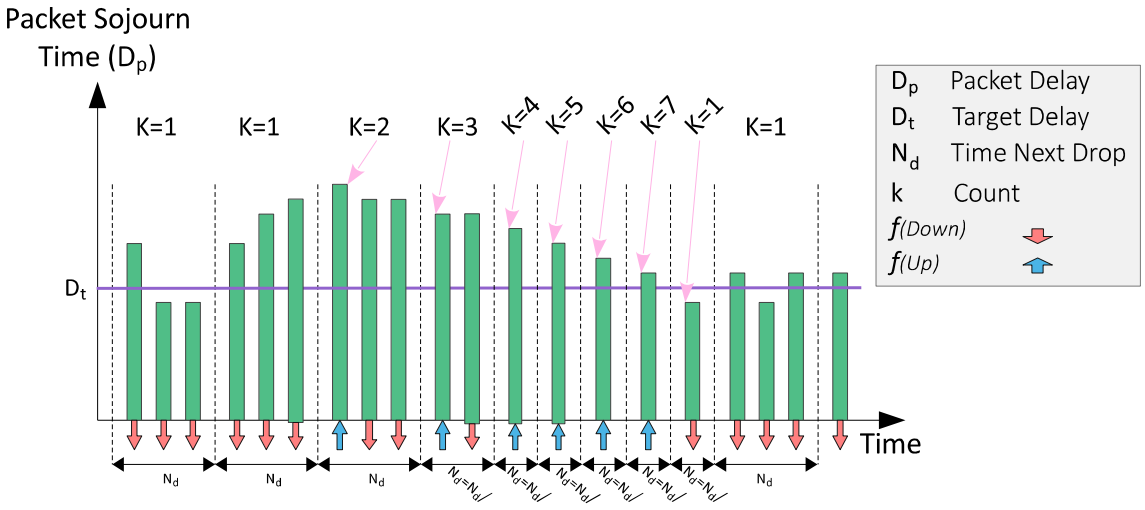


Figure V.11 – Slo-Mo example

Increase/decrease function: BAda is the function which increase/decrease the server rate depending on whether or not a bufferbloat has been detected. In order to validate the Slo-Mo concept and evaluate its performance we propose a basic linear increase/decrease function, as described below. Due to the burstiness behaviour of the mobile traffic [110, 138] we define the server rate regulation based on the packet arrival time.

Let n be a dequeued packet counter, which is reset at each change from rate UP to rate DOWN (or vice versa). Let C_0 be the initial rate and let α be the incremental/decremental factor, negative or positive for increasing or decreasing the bit-rate respectively. The server rate at point n is computed as follows:

$$C(n) = C_0 + \alpha n \tag{V.1}$$

Let N_{paq} the amount of packets required to achieve the max or min server bit-rate (C_{max} or C_{min}). Thus, α is computed as follows:

$$\alpha = \frac{C_{\text{max/min}} - C_0}{N_{paq}} \tag{V.2}$$

4.2.3 Closed-loop-system

The Slo-Mo rate tracking feature described in section 3.2 is implemented for each Slo-Mo entity (i.e. for each UE) through two main components as shown on Figure V.8. The *Queue Monitor* (QMon) block estimates the queue state through typical indicators (e.g. packet queueing delay or queue size). Thanks to these measurements, the QMon block issues a boolean - increase or decrease the server rate. The *Bit-rate Adapter* (BAda) block defines then how the queue server rate should increase/decrease (e.g. exponential, linear, etc).

Therefore, the Slo-Mo rate tracking feature can be modeled as a closed-loop control system [139], as shown in Figure V.12. Let $Q_{stat} = \{q_1, q_2, ..., q_n\}$ be the set of parameters reflecting the queue's state, such as delay, average delay, queue size, average queue size, etc. Let $f(Up)$ and $f(Down)$ be the queue server rate increase/decrease functions respectively.

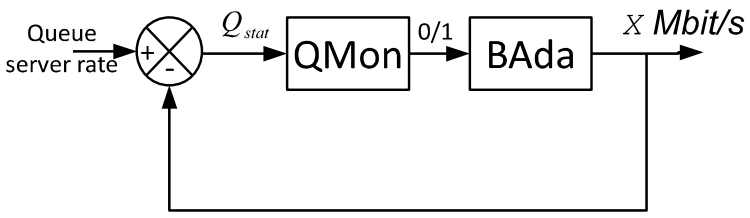


Figure V.12 – Slo-Mo rate tracking simplified block diagram

The *Queue Monitor* (QMon) block monitors the queue and takes the decision to increase or decrease its server rate (boolean output). The QMon algorithm proceeds as

summarized in Algorithm 1, where queue's monitored parameters are represented by Q_{stat} as defined above.

Algorithm 1: *Queue Monitor* (QMon) algorithm

```
Input  : Q_stat (base on queue's monitored parameters)
Output : Queue server rate Up or Down
while a packet is Dequeued by the scheduler do
    get input function Q_stat;
    val ← BufferbloatCriteria(Q_stat);
    // val:  Boolean
    if val==true then
        f(Up);
        // UP queue server rate
    else
        f(Down);
        // DOWN queue server rate
    end
end
```

QMon implements a key function called *BufferbloatCriteria*, which defines if the queue system meets with the defined Bufferbloat requirements based on input function (e.g. Q_{stat}). This input function provides an system overall indicator on the basis of queue's monitored parameters. Finally, *BufferbloatCriteria* function returns a boolean value, which is used to transmit to BAda component the order to increase or decrease queue server rate.

The *Bit-rate Adapter* (BAda) controls the server rate. Therefore, it implements two functions, one for increasing the server rate and another one to decrease it. The BAda block has a main impact on the system stability, since it acts as the "control function".

5 Simulation and Performance Evaluation

In this section, we present and discuss the numerical results and performance evaluation of Slo-Mo. Figure V.13 illustrates the simulated mobile network where Slo-Mo is implemented in the P-GW.

5.1 Network Parameters

The proposed Slo-Mo design has been evaluated thanks to the ns-3 simulator (version 3.21). We have modified the ns-3 source code in order to implement the Slo-Mo mechanism.

A typical outdoor scenario has been considered with 10 UEs attached to a single eNB, thus inter cell interference is not taken into account. The eNB is equipped with an

omnidirectional antenna and UEs experience varying channel conditions. The RLC layer in the eNB is configured in UM. We used a realistic channel model with path loss and fading. The path loss was simulated as described in COST-231 [103] and the fast fading as described by the EPA model using a Rayleigh multi-path fading model [104]. Given the path loss model and the other network parameters, we obtained a wide range of SINR values, which provided CQI values in range of $[1, 15]$.

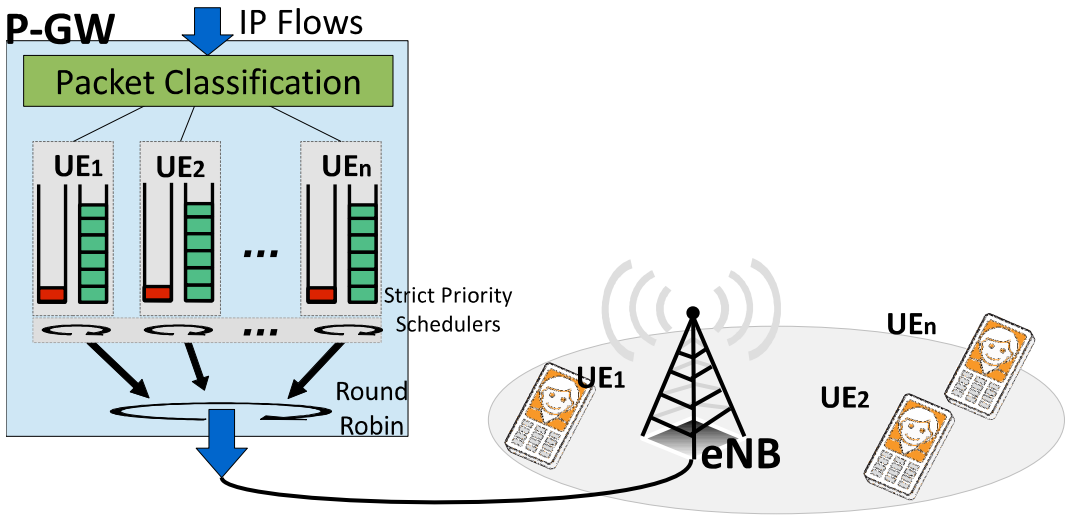


Figure V.13 – Simulated scenario

Number of terminals	10 (random positions)
Users mobility model	RandomWalk (3 km/h)
Bandwidth	25 RB (5 MHz)
Cell coverage radius	500 m
Pathloss Model	Cost231
eNB TX Power / Noise Figure	46 dBm / 5 dB
UE TX Power / Noise Figure	24 dBm / 5 dB
Fading loss model	EPA 3 km/h (urban scenario)
AMC model	PiroEW2010
DL/UL carrier frequency	2120 / 1930 MHz
RLC Transmission Mode	UM (Unacknowledged Mode)
RLC Buffer Size	100 kbytes (70 FTP packets)

Tableau V.1 – Simulation parameters

At the beginning of each run, UEs are placed randomly in a disc representing the cell within a distance range of 30-500m. Then, UEs move within the disc according to a Random Walk Model, at a fixed speed of 3 km/h. The simulation parameters are shown in Table V.1 and the system configuration is as follows: The cell is connected

to the EPC, which is composed of a P-GW and a router implementing Slo-Mo. Two servers are connected to EPC through the Slo-Mo router, one for the VoIP service and the other supports an FTP service. Both servers are connected to the Slo-Mo router via an over-provisioned point-to-point link in order to avoid congestion on this segment of the network. In the Slo-Mo scenario, each server performs the DSCP marking depending on the application.

5.2 Traffic Description

In order to simplify the analysis and reduce the computation time, each UE supports the canonical traffic mix described below:

- **VoIP Traffic:** simulated as described in Chapter 5.2.1.
- **FTP Traffic:** today, most of mobile applications use TCP protocol. Consequently in order to simplify our analysis, we represent all applications (e.g. web, streaming, social network) by a FTP session. A unique FTP session is performed by each UE during all the simulation time.

5.3 Scenarios Description

We consider 2 main scenarios:

▷ **Scenario 1** - Mono bearer, Best Effort (BE) scheme, PF scheduler. This scenario addresses the case of current deployments in LTE networks supported by a single bearer in Best Effort with a basic PF scheduler.

▷ **Scenario 2** - Mono-bearer, Slo-Mo mechanism, PF scheduler. In this scenario the Slo-Mo mechanism is implemented in the core network (i.e. P-GW) as shown in Figure V.13. We have tested many parameters combination for Slo-Mo, the Table V.2 summarises parameters that have shown better results.

5.4 Performance analysis

We present here performance results related to the scenarios described previously. Each simulation run lasts for 150 seconds, with a warm-up time of 5 seconds where statistics are not collected. VoIP calls are started at a random time uniformly distributed in $[1, 5]$ seconds.

QMon Parameters	*Value*
Target Delay (D_t)	250ms
Interval (Inter)	1ms
BAda Parameters	*Value*
N_{paq}	50
C_{min}	100kbps
C_{max}	20Mbps

Tableau V.2 – Simulated Slo-Mo parameters

We show the performance of each service and give in the following simple conclusions. We then consider all the services together and go deeper in the analysis.

Figure V.14 depicts the CDFs of cell throughput in the analysed scenarios (Best Effort and Slo-Mo). The Best Effort scheme achieves the highest cell throughput, but is closely followed by the Slo-Mo scenario; this is because neither of them modify the PF resource allocation strategy. Otherwise, in current scenario, we have configured the cell capacity (in number of PRB) about half respect of the simulated scenario in chapter IV , where the cell throughput of the CQA_{PF} (Figure IV.12) is particularly degraded and its cell throughput is fallen by nearly half respect to the Best Effort scheme.

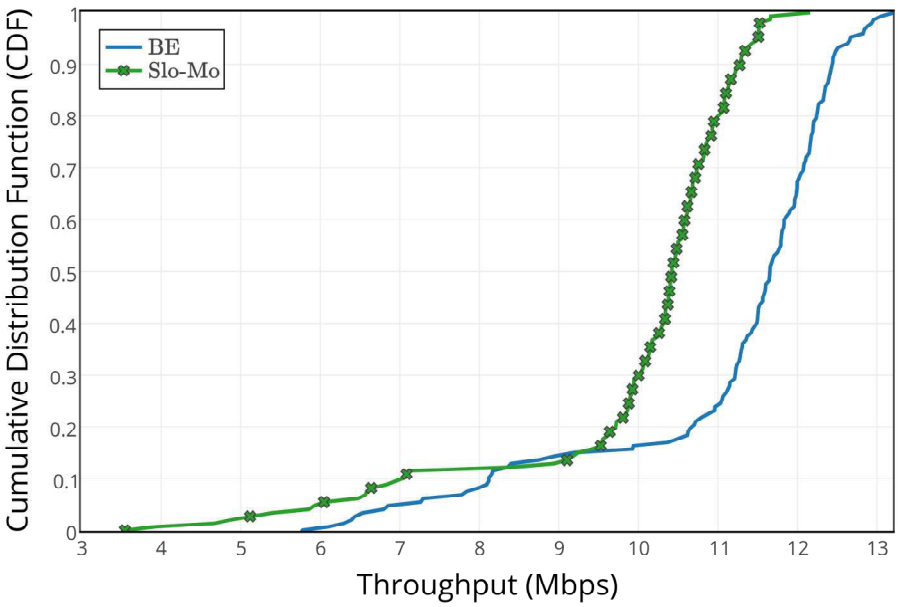


Figure V.14 – Cell Throughput performance

Figure V.15 provides the boxplots of the VoIP MOS. It can be seen from this figure that the MOS for the Best Effort scheme is relatively poor as its median is around 3 (close to Fair Quality) and has a min value close to 0 and its Q1 is close to 2.5 (Bad Quality). On the other hand, the Slo-Mo scheme has a MOS around 3.7 (Good Quality), an smaller

standard deviation and its Q1 is close to 3.5. It is noticeable, however, that a smaller VoIP communications using Slo-Mo scheme have a MOS equivalent to 3G mobile networks.

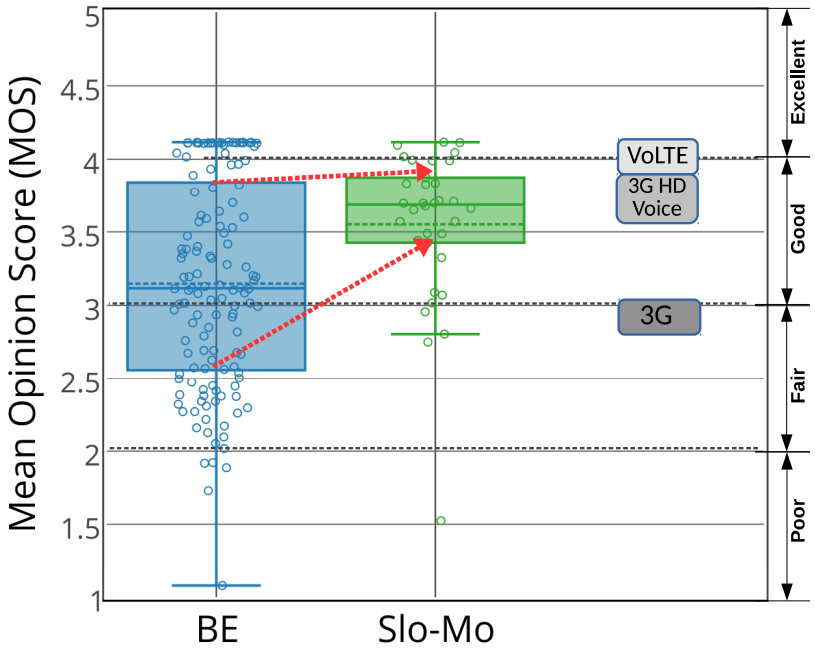


Figure V.15 – VoIP QoE in terms of Mean Score Opinion

BE

Ethernet · 20 | IPv4 · 20 | IPv6 | TCP · 10 | UDP · 24

Address A	Port A	Address B	Port B	Packets	Bytes	Packets A → B	Bytes A → B	Packets B → A	Bytes B → A	Bits/s A → B	Bits/s B → A
20.2.0.2	35065	7.0.0.2	21	13 544	13 M	9102	13 M	4442	262 k	733 k	14 k
20.2.0.2	46846	7.0.0.3	21	12 033	12 M	8126	12 M	3907	230 k	655 k	12 k
20.2.0.2	37108	7.0.0.4	21	26 383	26 M	17546	26 M	8837	515 k	1415 k	27 k
20.2.0.2	48485	7.0.0.5	21	19 640	21 M	13761	20 M	5879	337 k	1109 k	18 k
20.2.0.2	47000	7.0.0.6	21	25 170	25 M	16568	24 M	8602	511 k	1336 k	27 k
20.2.0.2	55940	7.0.0.7	21	27 424	27 M	18052	27 M	9372	552 k	1455 k	29 k
20.2.0.2	37204	7.0.0.8	21	24 112	24 M	15829	23 M	8283	492 k	1276 k	26 k
20.2.0.2	33361	7.0.0.9	21	2 689	2998 k	1974	2958 k	715	40 k	209 k	2851
20.2.0.2	57808	7.0.0.10	21	1 916	2002 k	1313	1967 k	603	35 k	191 k	3434
20.2.0.2	53807	7.0.0.11	21	48 826	50 M	33373	50 M	15453	868 k	2691 k	46 k

UEs IP@ — 200 M

Slo-Mo

Ethernet · 20 | IPv4 · 20 | IPv6 | TCP · 10 | UDP · 24

Address A	Port A	Address B	Port B	Packets	Bytes	Packets A → B	Bytes A → B	Packets B → A	Bytes B → A	Bits/s A → B	Bits/s B → A
20.2.0.2	35065	7.0.0.2	21	15 218	15 M	10074	15 M	5144	309 k	812 k	16 k
20.2.0.2	46846	7.0.0.3	21	15 028	15 M	9972	14 M	5056	301 k	804 k	16 k
20.2.0.2	37108	7.0.0.4	21	25 064	25 M	16401	24 M	8663	522 k	1322 k	28 k
20.2.0.2	48485	7.0.0.5	21	19 232	20 M	13265	19 M	5967	348 k	1070 k	18 k
20.2.0.2	47000	7.0.0.6	21	23 542	23 M	15257	22 M	8285	495 k	1230 k	26 k
20.2.0.2	55940	7.0.0.7	21	26 555	26 M	17324	26 M	9231	561 k	1397 k	30 k
20.2.0.2	37204	7.0.0.8	21	23 950	23 M	15645	23 M	8305	502 k	1261 k	26 k
20.2.0.2	33361	7.0.0.9	21	2 171	2266 k	1485	2225 k	686	40 k	119 k	2200
20.2.0.2	57808	7.0.0.10	21	22 344	23 M	15103	22 M	7241	431 k	1217 k	23 k
20.2.0.2	53807	7.0.0.11	21	42 050	42 M	27942	41 M	14108	842 k	2253 k	45 k

UEs IP@ — 208 M

Figure V.16 – Goodput measurement for FTP services (*wireshark print-screen*)

Figure V.16 provides a first measure of goodput in the analysed scenarios (Best Effort and Slo-Mo), which revealed an interesting fact concerning the radio resources utilization by the upper layers (i.e. TCP). In spite of the reduction of the cell throughput by Slo-Mo,

Figure V.16 shows an improvement in the amount of useful transmitted data (goodput) compared to Best-Effort scenario in around 4% (an increase of about 8 Mbytes).

In summary, Slo-Mo mechanism substantially improves the QoE for priority traffic (i.e. VoIP), which is compensated by a slight reduction in the cell throughput, but our first measure of goodput highlights a relevant improvement provided by Slo-Mo. We suspect that the goodput improvement is mainly due to the fact that Slo-Mo contributes to better adaptation of TCP protocol to the high variable radio conditions. Thus, Slo-Mo eliminates the TCP flow synchronization, which reduces the number of retransmissions of TCP data due to packet loss or TCP timer expiration [140]. TCP retransmissions are usually due to network congestion, which is the case of the Best-effort scenario in which the bottleneck is the eNB.

6 Conclusion and Future Work

The Slo-Mo mechanism described in this chapter allows for QoS improvement in mobile networks. It should be located on a unique point on the communication path, possibly in a P-GW. Slo-Mo aims to manage the QoS per aggregate based on a DSCP architecture and the transport protocols' congestion mechanisms. Contrary to similar mechanisms, Slo-Mo does not rely on protocol analysis, and may be applied to cyphered traffic. Slo-Mo does not rely on protocol analysis (e.g. ACK messages), nor any signaling protocol and may be applied to ciphered traffic. It does not make any assumption on the transport protocol (e.g. TCP, QUIC), provided it implements a congestion mechanism (e.g. cubic, reno).

Compared to other proposals (see Section 2.2), Slo-Mo implements a lightweight mechanism allowing mobile operators to regain control of QoS. It is easy to deploy and operate, since P-GW already implements most of the required features (multiple IP queues, at least one per UE [141]), except the closed-loop control.

Future work will focus on new increase/decrease functions (BAda block) as well as new QMon functions mainly inspired of last AQM algorithms (i.e. PIE, ARED), which have proved to be very effective. We also have started to test some increase/decrease functions (BAda block), which demonstrates be more adapted to rapidly changing radio conditions, since they try to imitate the TCP control flow algorithm.

This study has paved a promising way at Orange on cross-layer mechanisms towards enhanced customer experience and optimized network utilization. Finally, we work in collaboration with B-com in order to prepare a Slo-Mo Proof of Concept (PoC); we are also currently discussing with some telco vendors fur future integration of Slo-Mo in their products.

Conclusions and Perspectives

Conclusions

In this thesis, we have addressed QoS management issues in LTE/EPC mobile networks. Beside a cost-factor analysis and an evaluation of the standardised QoS architecture, we have also dedicated a significant part of the thesis to alternative QoS proposal and investigation. The envisaged QoS models aim at providing a cost-effective QoS scheme, better adapted to the current open and web-oriented mobile ecosystem than the current standards.

In chapter I and II, after reviewing the LTE/EPC architecture and protocols, we have thoroughly described typical QoS management schemes in packet networks. Through these analysis, it is clear these two worlds (fixed and mobile) have addressed these QoS issues with opposite approaches. On the one hand, the mobile world was a closed ecosystem in which mobile operators were providing and controlling services offered on their own networks. On the other hand, the fixed world is an open ecosystem mainly web oriented where the OTT are today the most important service providers. The meeting of these two worlds has had disruptive impacts on the mobile ecosystem, since standards of mobile networks have not evolved in the same direction, thereby giving way to loss of value for mobile operators. Moreover, third party players (i.e. OTTs) have started to propose and develop end-to-end congestion control mechanisms ever more agile and effective for managing customer experience in mobile networks. In a nutshell, access providers are in the process of losing their grip on customer experience.

In order to address the problem of the QoS in mobile networks, it was important to understand the most relevant weak points of 3GPP QoS model. In chapter III, we have investigated the cost-factors related to standardised mobile QoS which are mainly due to signaling procedures and sessions management, which have been inherited of legacy telco

standards. We have proposed an analytical model to evaluate the impact of the standardised QoS model in terms of *Context Load*, *Processing Load*, *Memory Access Rate* and *Radio Signaling Overhead* for LTE radio segment, which have been published in [53, 54]. Based on the findings and taking into account current mobile data traffic behaviour, we figured out that the 3GPP QoS model could have a negative impact on the performances of LTE/EPC elements. This performance depends on the degree of QoS granularity implemented in the network. Based on the conclusions of these three first chapters, we have defined the requirements and objectives of possible new QoS schemes for mobile networks.

In chapter IV we tackle the problems identified above, proposing a novel IP-centric QoS scheme called IP-aware, which have been published in [78, 90, 91, 92]. Our IP-aware scheme is mainly inspired on the DiffServ architecture, which has been widely studied and is currently one of the most used QoS schemes in the fixed Interned world. Our IP-aware scheme aims to be compatible with current and future mobile networks. In case of LTE networks, some modifications on the radio protocol stack are needed, since the current eNB does not implement the IP layer where we propose to manage the QoS. In order to evaluate its performance, we have attempted to recreate conditions close to the reality using the ns-3 simulator where we have implemented realistic traffic sources, which constitute the highest proportion of mobile data traffic (i.e. DASH, VoIP and web). Results showed that when multiple flows with different QoS requirements are simultaneously supported by a terminal our IP-aware scheme can achieve good performance and in some cases similar than the standardised QoS scheme. These studies have allowed to highlight remarkable benefits of IP-centric approach to manage QoS in mobile networks. Beside being cost-effective IP-Aware also provides good QoS performance, allowing for a unified solution in a fixed/mobile convergent world, which opens the door to access-agnostic applications. Clearly, our IP-aware approach will never be able to provide same QoS guarantees than current 3GPP solutions, because the IP-aware just aims to multiplex IP packets in an intelligent manner. This multiplexing stage has no impact on radio resource allocation, which is performed by a PF radio scheduler. We use a PF scheduler since it provides a trade-off between cell throughput optimization and fairness between the UEs.

In spite of remarkable proven benefits of the IP-centric approach, it seems that this kind of solution will be available in long term, since it requires modifications in eNB protocol stack. Those modifications implies standardisation evolution that is a long and difficult process in which Orange is involved and has already started to contribute on corresponding international bodies (i.e. 3GPP-[142], GSMA[2]). However, in the short term lightweight solutions are desired to overcome the limitations of current standards and allow operators to provide QoS guarantees adapted to the current mobile ecosystem.

It is also important to mention that third party actors (i.e. OTTs) have started to promote new end-to-end congestion control mechanisms, which will be possibly available in short term (e.g. CQIC, piStream). Therefore, access providers are in the process of losing their grip on customer experience to third party actors. Based on the foregoing, we have

[2]The GSM Association represents the interests of mobile operators worldwide

introduced Slo-Mo in chapter V, which is a lightweight implicit mechanism for managing QoS introduced in [136]. Slo-Mo must be located on a unique point on the communication path, possibly in a P-GW. With Slo-Mo, a self-adapted bottleneck is created in a node controlled by the operator when the point of spontaneous occurrence of congestion (e.g. eNB) is out of reach. Slo-Mo bottleneck is adjusted dynamically to the available resources taking advantage of TCP native flow control and the bufferbloat phenomenon, which has been widely studied in fixed networks ecosystem.

In order to evaluate Slo-Mo performances, we have implemented the Slo-Mo mechanism in the ns-3 simulator. As was demonstrated via simulation results, Slo-Mo enhances customer experience at a marginal cost and its deployment needs very few modifications in the involved node. Those modifications does not implies modifications of 3GPP standards, since in case of P-GW, it already implements a queue system per UE. Furthermore, as an implicit mechanism, Slo-Mo does not rely on protocol analysis, and may be applied to cyphered traffic. It does not make any assumption on the flow routing path as it can be applied independently to each direction (i.e. Uplink/Downlink). We believe that Slo-Mo is a "quick-win" allowing operators regain control on QoS and aims to stop its disintermediation.

Perspectives

Although we endeavored in this thesis to address various evaluation and design aspects of the QoS in mobile networks, this dissertation opens up many other avenues for future researches related to alternative QoS mechanisms. In this regard, there are several open topics that can be studied related to each contribution of the dissertation, which are proposed below.

- With regard to analytical models developed in chapter III, they can be valued via its integration in a mobile network dimensioning tools in order to enrich them by taking into consideration traffic pattern, LTE RRC state machine and providing new Key Performance Indicatorss (KPIs).

- With regard to our IP-Aware scheme, compare its performances considering IMS-based services as VoLTE and ViLTE are desired. This evaluation can be performed via simulations but the implementation of the VoLTE or/and ViLTE traffic model can be a real challenge, which could take a long time. It may be also interesting to compare the performances of our IP-aware scheme with other IP-centric schemes as those cited in chapter IV and also a new IP-aware scheme implementing Inter-bearer arrangements. Finally, we have mentioned that IP-centric approach is fully compatible with the new trends of mobile network architectures (i.e. SDN, NFV) and its integration could be also analyzed and evaluated.

- With regard to Slo-Mo mechanism, there are several open topics especially concerning Slo-Mo optimization and performance evaluation such as:
  - ◇ Compare Slo-Mo performances to equivalent mechanisms (e.g CQIC, piStream).
  - ◇ Enhance Slo-Mo simulator including other AQMs metrics and other increase/decrease functions as well as new metrics like the goodput and the retransmission rate at transport layer level (e.g. TCP). These metrics are necessary to well understand and evaluate the real benefits of Slo-Mo.
  - ◇ Moreover, we have observed that Slo-Mo provides a sort of stability to the bit-rate variation due to the fast evolution of radio conditions. This opens up the possibility to make use of Slo-Mo as a "TCP optimizer", in addition to its use in the QoS management. Therefore, this new use case become relevant taking into consideration the emergence of new transport protocols (e.g. QUIC, SPUD), which are becoming more closed to such a degree that the current TCP optimizers are unserviceable.
  - ◇ Slo-Mo principle can also be used in other communication networks having equivalent issues than mobile networks as WiFi networks, satellite networks, microwave links, legacy ADSL networks, etc.
  - ◇ Finally, in this dissertation we have proposed to use a DiffServ architecture to manage the QoS in our Slo-Mo implementation, but it is also possible to use other alternative mechanisms. In this regard, implicit QoS mechanisms seem to be a good option and among them the Shortest Queue First (SQF) mechanism [143] be the most promising.

In general, the QoS in mobile networks is starting to be redefined in the 3GPP [144], which aims to prepare future mobile networks to support new technologies as IoT, V2V. These new technologies will need a mobile network able to support a huge number of connected terminals. But as we showed in chapter III, the current 3GPP QoS scheme is not adapted for it. For that reason, new scalable and flexible QoS schemes are needed. We hope that our work concerning the IP-centric approach will contribute in the search and definition for future mobile QoS schemes.

Nevertheless, as previously mentioned, in short term the mobile networks urgently need an alternative solution to improve QoS since the standardised scheme is hardly ever used or only used for specific services as the VoLTE. For that reason, we have proposed Slo-Mo and we think that it is a promising solution. In this dissertation we lay the foundation and principles of Slo-Mo and we continue to explore ways of enhancing it.

List of Publications

International Publications:

[P1] William Diego. "A framework for Generating HTTP Adaptive Streaming Traffic in ns-3". In *ACM SIMUTools*, August 2016.
(Cited on page 88)

[P2] William Diego, Isabelle Hamchaoui, and Xavier Lagrange. "The Cost of QoS in LTE/EPC Mobile Networks Evaluation of Signalling Load". In *IEEE VTC Fall*, 2015.
(Cited on pages 47, 56, and 120)

[P3] William Diego, Isabelle Hamchaoui, and Xavier Lagrange. "Cost Factor Analysis of QoS in LTE/EPC Networks". In *IEEE CCNC*, 2016.
(Cited on pages 47 and 120)

[P4] William Diego, Hamchaoui Isabelle, and Xavier Lagrange. "Cross-layer Design and Performance Evaluation for IP-Centric QoS Model in LTE-EPC Networks". In *IFIP WMNC*, 2015.
(Cited on pages 73, 77, 120, and 144)

[P5] Isabelle Hamchaoui, William Diego, and Sebastien Jobert. "IP centric QoS model for mobile networks - Packet based QoS management for Intra-bearer arrangements". In *IEEE WCNC*, 2014.
(Cited on pages 77, 120, and 144)

Standardisation:

[S1] Isabelle Hamchaoui and William Diego. "Packet-oriented QoS management model for a wireless Access Point". In *Orange contribution to SG12, ITU-T C0255 Geneva, 05-14 May*, 2015.
(Cited on pages 77, 120, and 144)

[S2] Sebastien Jobert, Isabelle Hamchaoui, and William Diego. "Packet-oriented QoS management model for a wireless Access Point". In *Orange Internet-Draft to ICCRG Research Group, IETF draft-jobert-iccrg-IP-aware-AP-00.txt*, July 2013.
(Cited on pages 77, 120, and 144)

Patents:

[T1] William Diego, Isabelle Hamchaoui, and Fabrice Guillemin. "Procédés de traitement de paquets de données, dispositif, produit programme d'ordinateur, medium de stockage et neoud de réseau correspondants ". *Patent Application submitted to INPI N° 1561613* – France, November 2015.
(Cited on pages 103 and 121)

In preparation:

[H1] William Diego, Isabelle Hamchaoui, and Fabrice Guillemin. "SloMo: An Implicit Cross-Layer Mechanism for a Better Experience on Mobile Networks". In *to be submitted as a conference article*, –.
(Not cited)

Bibliography

[1] Cisco Systems Inc. White paper: "Cisco Visual Networking Index: Forecast and Methodology, 2013-2018", June 2014.
(Cited on pages 1, 142, and 147)

[2] Ericsson. White paper: "Mobility Report", June 2014.
(Cited on pages 1, 142, and 147)

[3] M. Vakulenko and VisionMobile in association with Ericsson. Report: "Telco Innovation Toolbox", December 2012.
(Cited on pages 1 and 142)

[4] 3GPP. "E-UTRA and E-UTRAN Overall description". TS 36.300 version 8.12.0 Release 8, 2010.
(Cited on pages 1 and 20)

[5] 3GPP. "QoS Concept and Architecture". TS 23.107 version 10.2.0 Release 10, 2011.
(Cited on pages 1 and 143)

[6] S. Ahmadi. *"LTE-Advanced: A Practical Systems Approach to Understanding 3GPP LTE Releases 10 and 11 Radio Access Technologies"*. ITPro collection. Elsevier Science, 2013.
(Cited on page 5)

[7] T. Ali-Yahiya. *"Understanding LTE and its Performance"*. SpringerLink : Bücher. Springer, 2011.
(Cited on page 5)

[8] 3GPP. "GPRS enhancements for E-UTRAN access". TS 23.401 version 8.18.0 Release 8, 2013.
(Cited on pages 6, 16, 17, 38, 48, and 56)

[9] 3GPP. "Non-Access-Stratum (NAS) protocol for Evolved Packet System (EPS)". TS 24.301 version 10.15.0, 2014.
(Cited on pages 7, 16, and 56)

[10] 3GPP. "E-UTRA and E-UTRAN; overall description". TS 36.300 version 8.12.0 Release 8, 2010.
(Cited on pages 8, 12, 37, and 161)

[11] Magnus Olsson and Catherine Mulligan. *"EPC and 4G packet networks: driving the mobile broadband revolution"*. Academic Press, 2012.
(Cited on page 8)

[12] Erik Dahlman, Stefan Parkvall, and Johan Skold. *"4G: LTE/LTE-advanced for mobile broadband"*. Academic press, 2013.
(Cited on pages 11 and 161)

[13] 3GPP. "Radio Link Control (RLC) protocol specification". TS 25.322 version 8.9.0 Release 8, 2010.
(Cited on page 12)

[14] 3GPP. "Universal Mobile Telecommunications System (UMTS); UTRA High Speed Downlink Packet Access (HSDPA); Overall description". TS 25.308 version 5.7.0 Release 5, 2004.
(Cited on page 13)

[15] Iffat Ahmed, Leonardo Badia, Nicola Baldo, and Marco Miozzo. "Design of a unified multimedia-aware framework for resource allocation in LTE femtocells". In *ACM MobiWac*, 2011.
(Cited on pages 15 and 45)

[16] Kian Chung Beh, Simon Armour, and Angela Doufexi. "Joint Time-Frequency domain proportional fair scheduler with HARQ for 3GPP LTE systems". In *IEEE VTC Fall*, 2008.
(Cited on pages 15 and 45)

[17] B. Bojovic and N. Baldo. "A new channel and QoS aware scheduler to enhance the capacity of voice over LTE systems". In *IEEE SSD*, 2014.
(Cited on pages 15, 45, 90, and 93)

[18] Thomas Bonald and James Roberts. "Scheduling network traffic". In *ACM SIGMETRICS*, 2007.
(Cited on pages 15 and 45)

[19] F. Capozzi, G. Piro, L.A Grieco, G. Boggia, and P. Camarda. "Downlink Packet Scheduling in LTE cellular networks: Key design issues and a survey". In *IEEE Communications Surveys & Tutorials*, 2013.
(Cited on pages 15, 39, 45, and 72)

[20] Patrick A. Hosein. "QoS control for WCDMA high speed packet data". In *IEEE WiMob Workshop*, 2002.
(Cited on pages 15 and 45)

[21] Troels Emil Kolding. "QoS-aware proportional fair packet scheduling with required activity detection". In *IEEE VTC Fall*, 2006.
(Cited on pages 15 and 45)

[22] Xin Liu, E. K P Chong, and N.B. Shroff. "Opportunistic transmission scheduling with resource-sharing constraints in wireless networks". In *IEEE Journal on Selected Areas in Communications*, 2001.
(Cited on pages 15 and 45)

[23] G. Monghal, K.I Pedersen, IZ. Kovacs, and P.E. Mogensen. "QoS oriented time and frequency domain packet schedulers for the UTRAN long term evolution". In *IEEE VTC Spring*, 2008.
(Cited on pages 15 and 45)

[24] M. Shariat, AU. Quddus, S.A Ghorashi, and R. Tafazolli. "Scheduling as an important cross-layer operation for emerging broadband wireless systems". In *IEEE Communications Surveys Tutorials*, 2009.
(Cited on pages 15 and 45)

[25] Sunil Suresh Kulkarni and Catherine Rosenberg. "Opportunistic Scheduling: Generalizations to include multiple constraints, multiple interfaces, and short term fairness". In *Springer Wireless Networks*, 2005.
(Cited on pages 15 and 45)

[26] Aditya Karnik and Catherine Rosenberg. "Optimal control of service rate and reliability in scheduling over wireless channels. In *IEEE COMSNETS*, 2012.
(Cited on pages 15 and 45)

[27] Frank Kelly. "charging and rate control for elastic traffic". In *European Transactions on Telecommunications*, 1997.
(Cited on page 15)

[28] 3GPP. "User Equipment (UE) procedures in idle mode". TS 36.304 version 8.10.0 Release 8, 2011.
(Cited on pages 16 and 56)

[29] William C Hardy and Luis Preface By-Cardoso. *"QoS: measurement and evaluation of telecommunications quality of service"*. John Wiley & Sons, Inc., 2001.
(Cited on page 24)

[30] ITU-T recommendation E.800. "Quality of Telecommunication Service: Concepts, Models, Objectives and Dependability Planning. Terms and Definitions Related to

the Quality of Telecommunication Services", September 2008.
(Cited on page 25)

[31] ETSI ETR003. "Network Aspects (NA); General Aspects of Quality of Service (QoS) and Network Performance (NP)". Technical report, 1994.
(Cited on page 25)

[32] Janusz Gozdecki, Andrzej Jajszczyk, and Rafal Stankiewicz. "Quality of Service Terminology in IP Networks". In *IEEE Communications Magazine*, 2003.
(Cited on pages 25 and 161)

[33] ITU-T..G.107. The E-model: a computational model for use in transmission planning. Technical report, 02, 2014.
(Cited on pages 26, 86, and 149)

[34] Conta A. Deering S. Rajahalme, J. and B Carpenter. "RFC 3697 IPv6 flow label specification" IETF. 2011.
(Cited on page 27)

[35] Bob Braden, David Clark, and Scott Shenker. "Integrated Services in the Internet Architecture: an overview". RFC 1633, RFC Editor, June 1994. http://www.rfc-editor.org/rfc/rfc1633.txt.
(Cited on pages 28 and 29)

[36] Steven Blake, David L. Black, Mark A. Carlson, Elwyn Davies, Zheng Wang, and Walter Weiss. "An Architecture for Differentiated Services". RFC 2475, RFC Editor, December 1998. http://www.rfc-editor.org/rfc/rfc2475.txt.
(Cited on pages 28, 29, and 32)

[37] R. Braden, Ed., L. Zhang, S. Berson, S. Herzog, and S. Jamin. "Resource ReSerVation Protocol (RSVP) – version 1 functional specification". In *RFC 2205*, September 1997.
(Cited on page 29)

[38] S. Shenker, C. Partridge, and R. Guerin. Specification of Guaranteed Quality of Service. Technical Report 2212, September 1997.
(Cited on page 29)

[39] 3GPP. "General Packet Radio Service (GPRS); Service description". TS 23.060 version 3.17.0 Release 99, 2006.
(Cited on pages 35, 36, and 37)

[40] David Soldani, Man Li, and Renaud Cuny. *"QoS and QoE management in UMTS cellular systems"*. John Wiley & Sons, 2007.
(Cited on page 37)

[41] Gerardo Gómez and Rafael Sánchez. *"End-to-end quality of service over cellular networks: data services performance optimization in 2G/3G"*. John Wiley & Sons, 2005.
(Cited on page 37)

[42] 3GPP. "QoS Concept and Architecture". TS 23.107 version 8.2.0 Release 8, 2011.
(Cited on page 37)

[43] 3GPP. "Policy and charging control architecture". TS 23.203 version 8.14.0 Release 8, 2012.
(Cited on pages 37, 39, and 42)

[44] Alcatel Lucent. "the LTE network architecture - a comprehensive tutorial". In *Strategic Whitepaper*, 2009.
(Cited on pages 39 and 162)

[45] Arash Asadi and Vincenzo Mancuso. "A survey on opportunistic scheduling in wireless communications". In *IEEE Communications Surveys & Tutorials*, 2013.
(Cited on pages 39, 45, and 72)

[46] F. Kelly. "Charging and rate control for elastic traffic". In *European Transactions on Telecommunications*, 1997.
(Cited on pages 39, 72, and 143)

[47] Telecoms.com. "Telecoms.com Intelligence Industry Survey 2015". In *Report*, 2015.
(Cited on page 42)

[48] J-JP Balbas, Stefan Rommer, and John Stenfelt. "Policy and charging control in the evolved packet system". In *IEEE Communications Magazine*, 2009.
(Cited on page 42)

[49] NETMANIAS TECH-BLOG. "LTE QoS: SDF and EPS bearer QoS". In *Technical Document*, 2013.
(Cited on pages 43 and 162)

[50] Jyrki TJ Penttinen. *"The LTE/SAE Deployment Handbook"*. John Wiley & Sons, 2011.
(Cited on page 44)

[51] Junxian Huang, Feng Qian, Yihua Guo, Yuanyuan Zhou, Qiang Xu, Z Morley Mao, Subhabrata Sen, and Oliver Spatscheck. "An in-depth study of LTE: effect of network protocol and application behavior on performance". In *ACM SIGCOMM*, 2013.
(Cited on pages 45 and 100)

[52] Yaxin Cao and Victor OK Li. "Scheduling algorithms in broadband wireless networks". In *Proceedings of the IEEE*, 2001.
(Cited on page 45)

[53] William Diego, Isabelle Hamchaoui, and Xavier Lagrange. "The Cost of QoS in LTE/EPC Mobile Networks Evaluation of Signalling Load". In *IEEE VTC Fall*, 2015.
(Cited on pages 47, 56, and 120)

[54] William Diego, Isabelle Hamchaoui, and Xavier Lagrange. "Cost Factor Analysis of QoS in LTE/EPC Networks". In *IEEE CCNC*, 2016.
(Cited on pages 47 and 120)

[55] Indra Widjaja, Peter Bosch, and Humberto La Roche. "Comparison of MME signaling loads for long-term-evolution architectures". In *IEEE VTC Fall*, 2009.
(Cited on pages 48 and 56)

[56] Chan-Kyu Han, Hyoung-Kee Choi, Jung Woo Baek, and Ho Woo Lee. "Evaluation of authentication signaling loads in 3GPP LTE/SAE networks". In *IEEE LCN*, 2009.
(Cited on page 48)

[57] Chan-Kyu Han and Hyoung-Kee Choi. "Security analysis of handover key management in 4G LTE/SAE networks". In *IEEE Transactions on Mobile Computing*, 2014.
(Cited on page 48)

[58] Pragya Kirti Gupta, RV Rajakumar, and C Senthil Kumar. "Analysis of impact of network activity on energy efficiency of 3GPP-LTE". In *IEEE INDICON*, 2012.
(Cited on page 48)

[59] Meng Wang, Michael Georgiades, and Rahim Tafazolli. "Signalling Cost evaluation of mobility management schemes for different core network architectural arrangements in 3GPP LTE/SAE". In *IEEE VTC Spring*, 2008.
(Cited on page 48)

[60] Dario S Tonesi, Luca Salgarelli, Yan Sun, and Thomas F La Porta. "Evaluation of Signaling Loads in 3GPP networks". In *IEEE Wireless Communications*, 2008.
(Cited on page 48)

[61] Giri Prassad Deivasigamani. "Dynamic configuration of inactivity timeouts for data radio bearers", 2013. US Patent App. 14/070,827.
(Cited on page 54)

[62] Jonathan Rodriguez. *"Fundamentals of 5G Mobile Networks"*. John Wiley & Sons, 2015.
(Cited on page 54)

[63] R Thomas, H Gilbert, and G Mazziotto. "Influence of the movement of the mobile station on the performance of a radio cellular network". In *Proc. 3rd Nordic Seminar*, pages 9–4, 1988.
(Cited on page 55)

[64] G Morales-Andres and M Villen-Altamirano. "An approach to modelling subscriber mobility in cellular radio networks". In *Proceedings of the Forum Telecom*, pages 185–189, 1987.
(Cited on page 55)

[65] Andreas Kunz, Tarik Taleb, and Stefan Schmid. "On minimizing Serving GW/MME relocations in LTE". In *ACM IWCMC*, 2010.
(Cited on page 55)

[66] 3GPP. "E-UTRAN; X2 Application Protocol (X2AP"). TS 23.107 version 8.9.0 Release 8, 2011.
(Cited on page 56)

[67] 3GPP. "LTE Radio Access Network (RAN) enhancements for diverse data applications". TR 36.822 version 11.0.0, 2012.
(Cited on pages 57, 59, 64, 66, and 165)

[68] Marc H Bornstein and Helen G Bornstein. "The pace of life". In *Nature 259*, 19 Feb. 1976.
(Cited on page 62)

[69] Ramzi Bassil, Ali Chehab, Imad Elhajj, and Ayman Kayssi. "Signaling oriented denial of service on LTE networks". In *ACM MSWiM*, 2012.
(Cited on page 64)

[70] 3GPP. "IP Multimedia Subsystem (IMS)". TS 23.228 version 10.9.0 Release 10, 2015.
(Cited on page 64)

[71] GSMA PRD N2020.01, "VoLTE Service Description and Implementation Guidelines", version 1.0, December 2014.
(Cited on page 64)

[72] 3GPP. "Signalling overhead of diverse data application". R2-116167 RAN WG2-76 San Francisco, USA, 14-18 November, 2011.
(Cited on page 68)

[73] 3GPP. "E-UTRA: RRC Protocol specification". TS 36.331 version 10.13.0 Release 10, 2014.
(Cited on page 69)

[74] Nokia. White paper: "Voice over LTE (VoLTE) Optimization", 2015.
(Cited on page 69)

[75] Gianluca Foddis, Rosario G Garroppo, Stefano Giordano, Gregorio Procissi, Simone Roma, and Simone Topazzi. "LTE traffic analysis for signalling load and energy consumption trade-off in mobile networks". In *IEEE ICC*, 2015.
(Cited on page 70)

[76] Nikola Zahariev, Yasir Zaki, Xi Li, Carmelita Goerg, Thushara Weerawardane, and Andreas Timm-Giel. "Optimized service aware lte mac scheduler with comparison against other well known schedulers". In *Springer Wired/Wireless Internet Communication*, 2012.
(Cited on page 72)

[77] Yasir Zaki, Thushara Weerawardane, C Gorg, and Andreas Timm-Giel. "Multi-QoS-aware fair scheduling for LTE". In *IEEE VTC spring*, 2011.
(Cited on page 72)

[78] William Diego, Hamchaoui Isabelle, and Xavier Lagrange. "Cross-layer Design and Performance Evaluation for IP-Centric QoS Model in LTE-EPC Networks". In *IFIP WMNC*, 2015.
(Cited on pages 73, 77, 120, and 144)

[79] ITU-T SG12. "Quality of Service Development Groups (QSDG)", May 2014.
(Cited on page 73)

[80] ETSI. "QoS of connections from current technologies to LTE". Technical report, ETSI STF 437, P. Pocta, 2012.
(Cited on page 73)

[81] GSMA. "Delivering an all-IP world", january 5, 2016. Accessed February 1, 2016.
(Cited on page 74)

[82] Aimin Sang, Xiaodong Wang, and Mohammad Madihian. "Differentiated TCP user perception over downlink packet data cellular systems". In *IEEE Transactions on Mobile Computing*, 2007.
(Cited on page 74)

[83] 3GPP. "Study on system enhancements for user plane congestion management". TR 23.705 version 0.11.0 Release 13, 2014.
(Cited on page 75)

[84] P. Szilgyi and C. Vulkn. "Application Aware Mechanisms in HSPA Systems". In *IARIA ICWMC*, 2012.
(Cited on pages 75, 76, 80, and 144)

[85] "Proposed high level principles for UPCON". In *Alcatel Lucent and AT&T contribution S2-130074 to 3GPP SA2, Prague, January 2013*.
(Cited on pages 75, 76, 82, and 144)

[86] D. Soldani, Hou Xiao Jun, and B. Luck. "Strategies for Mobile Broadband Growth: Traffic Segmentation for Better Customer Experience". In *IEEE VTC Spring*, 2011.
(Cited on page 75)

[87] A. Sang, X. Wang, and M. Madihian. "differentiated TCP user perception over downlink packet data cellular systems". In *IEEE TMC*, 2007.
(Cited on page 76)

[88] 3GPP. "Study on system enhancements for user plane congestion management". TR 23.705 version 13.0.0 Release 13, 2014.
(Cited on page 76)

[89] I. Hamchaoui, S. Jobert, and S. Boufelja. "IP aware radio scheduling – introducing IP QoS management in LTE networks". In *IEEE ICC Workshops*, 2013.
(Cited on page 77)

[90] Isabelle Hamchaoui, William Diego, and Sebastien Jobert. "IP centric QoS model for mobile networks - Packet based QoS management for Intra-bearer arrangements". In *IEEE WCNC*, 2014.
(Cited on pages 77, 120, and 144)

[91] Sebastien Jobert, Isabelle Hamchaoui, and William Diego. "Packet-oriented QoS management model for a wireless Access Point". In *Orange Internet-Draft to ICCRG Research Group, IETF draft-jobert-iccrg-IP-aware-AP-00.txt*, July 2013.
(Cited on pages 77, 120, and 144)

[92] Isabelle Hamchaoui and William Diego. "Packet-oriented QoS management model for a wireless Access Point". In *Orange contribution to SG12, ITU-T C0255 Geneva, 05-14 May*, 2015.
(Cited on pages 77, 120, and 144)

[93] B. Tirouvengadam and A. Nayak. "Hybrid Mode Radio Link Control for Efficient Video Transmission over 4GLTE Network". In *IEEE ICPADS*, 2012.
(Cited on page 84)

[94] 3GPP. "E-UTRA: MAC Protocol specification". TS 36.321 version 10.10.0 Release 10, 2014.
(Cited on page 84)

[95] Diego Kreutz, Fernando MV Ramos, P Esteves Verissimo, Christian Esteve Rothenberg, Siamak Azodolmolky, and Steve Uhlig. "Software-defined networking: A comprehensive survey". In *Proceedings of the IEEE*, 2015.
(Cited on page 84)

[96] Xin Jin, Li Erran Li, Laurent Vanbever, and Jennifer Rexford. "SoftCell: scalable and flexible cellular core network architecture". In *ACM CoNEXT*, 2013.
(Cited on pages 84 and 144)

[97] A. Gudipati, L. E. Li D. Perry, and S. Katti. "SoftRAN: Software defined radio access network". In *ACM SIGCOMM Workshop*, 2013.
(Cited on pages 84 and 144)

[98] L. E. Li, Z. M. Mao, and J. Rexford. "Toward software-defined cellular networks". In *IEEE EWSDN*, 2012.
(Cited on pages 84 and 144)

[99] K.-K. Yap, R. Sherwood, M. Kobayashi, T.-Y. Huang, N. Handigol M. Chan, N. McKeown, , and G. Parulkar. "Blueprint for introducing innovation into wireless mobile networks". In *ACM VISA Workshop*, 2010.
(Cited on pages 84 and 144)

[100] J. Kempf, B. Johansson, S. Pettersson, H. Luning, and T. Nilsson. "Moving the mobile evolved packet core to the cloud". In *IEEE WiMob*, 2012.
(Cited on pages 84 and 144)

[101] Li Erran Li, Z Morley Mao, and Jennifer Rexford. "cellSDN: Software-defined cellular networks". In *Technical Report, Princeton University (TR-922-12)*, 2012.
(Cited on page 84)

[102] "The LTE-EPC Network Simulator (LENA) project". In *http://iptechwiki.cttc.es/LTE-EPC Network Simulator (LENA).*
(Cited on page 85)

[103] Preben Elgaard Mogensen and J Wigard. "COST Action 231: Digital mobile radio towards future generation system, final report". Technical report, 1999.
(Cited on pages 85 and 113)

[104] 3GPP. "Proposal for LTE Channel Models". Technical Report R4-070572 TSG RAN WG4, meeting 43; Ericsson, Nokia, Motorola, Rohde & Schwarz; May, 2007.
(Cited on pages 85 and 113)

[105] A Pattavina and A Parini. "Modelling voice call inter-arrival and holding time distributions in mobile networks". In *ITC19*, 2005.
(Cited on page 86)

[106] Junqiang Guo, Fasheng Liu, and Zhiqiang Zhu. "Estimate the call duration distribution parameters in GSM system based on K-L divergence method". In *IEEE WiCom 2007.*
(Cited on page 86)

[107] Matteo Maria Andreozzi, Daniele Migliorini, Giovanni Stea, and Carlo Vallati. "Ns2Voip++, an enhanced module for VoIP simulations". In *IEEE ICST, https://gist.github.com/4391503*, 2010.
(Cited on pages 86, 149, and 166)

[108] Lingfen Sun and Emmanuel C Ifeachor. "Voice quality prediction models and their application in VoIP networks". In *IEEE Transactions*, 2006.
(Cited on pages 86 and 150)

[109] Rastin Pries, Zsolt Magyari, and Phuoc Tran-Gia. "An HTTP web traffic model based on the top one million visited web pages". In *IEEE NGI*, 2012.
(Cited on pages 86, 147, 148, and 165)

[110] Juan J Ramos-Muñoz, Jonathan Prados-Garzon, Pablo Ameigeiras, Jorge Navarro-Ortiz, and Juan M López-Soler. "Characteristics of mobile YouTube traffic". In *IEEE Wireless Communications*, 2014.
(Cited on pages 87, 88, 111, 150, 151, 152, 153, and 154)

[111] Ricky Yang and Harrison J. Son. "Youtube's Live TV Streaming in Mobile Devices - HLS and Adaptive". Technical report, NETMANIAS TECH-BLOG, 2013-10-30.
(Cited on pages 87, 152, and 154)

[112] William Diego. "A framework for Generating HTTP Adaptive Streaming Traffic in ns-3". In *ACM SIMUTools*, August 2016.
(Cited on page 88)

[113] 3GPP. "E-UTRA; Physical channels and modulation". TS 36.211 version 8.9.0 Release 8, 2009.
(Cited on page 89)

[114] D. Zhou, N. Baldo, and M. Miozzo. "Implementation and Validation of LTE Downlink Schedulers for ns-3". In *WNS3*, 2013.
(Cited on pages 89 and 90)

[115] 3GPP. "Policy and charging control architecture". TS 23.203 version 8.15.0 Release 8, 2014.
(Cited on page 92)

[116] Haiqing Jiang, Zeyu Liu, Yaogong Wang, Kyunghan Lee, and Injong Rhee. "Understanding bufferbloat in cellular networks". In *ACM SIGCOMM workshop*, 2012.
(Cited on page 100)

[117] Binh Nguyen, Arijit Banerjee, Vijay Gopalakrishnan, Sneha Kasera, Seungjoon Lee, Aman Shaikh, and Jacobus Van der Merwe. "Towards understanding TCP performance on LTE/EPC mobile networks". In *ACM SIGCOMM workshop*, 2014.
(Cited on page 100)

[118] Jeffrey Erman, Vijay Gopalakrishnan, Rittwik Jana, and Kadangode K Ramakrishnan. "Towards a SPDY'ier mobile web?". In *ACM CoNEXT*, 2013.
(Cited on page 100)

[119] Ankur Jain, Andreas Terzis, Hannu Flinck, Nurit Sprecher, Swaminathan, and Kevin Smith. "Mobile Throughput Guidance inband signaling protocol". Internet-Draft draft-flinck-mobile-throughput-guidance-03, IETF Secretariat, September 2015. http://www.ietf.org/internet-drafts/draft-flinck-mobile-throughput-guidance-03.txt.
(Cited on page 101)

[120] Feng Lu, Hao Du, Ankur Jain, Geoffrey M Voelker, Alex C Snoeren, and Andreas Terzis. "CQIC: Revisiting Cross-layer Congestion Control for Cellular Networks". In *ACM HotMobile*, 2015.
(Cited on page 101)

[121] Xiufeng Xie, Xinyu Zhang, Swarun Kumar, and Li Erran Li. "piStream: Physical Layer Informed Adaptive Video Streaming Over LTE". In *ACM MobiCom*, 2015.
(Cited on page 101)

[122] Keith Winstein, Anirudh Sivaraman, Hari Balakrishnan, et al. "Stochastic Forecasts Achieve High Throughput and Low Delay over Cellular Networks". In *NSDI*, pages 459–471, 2013.
(Cited on page 101)

[123] John Nagle. "On packet switches with infinite storage". In *IEEE Transactions communications*, 1987.
(Cited on page 102)

[124] Brough Turner. "Has AT&T Wireless data congestion been self-inflicted?". In *Brough Turner blog. Retrieved*, pages 02–28, 2012.
(Cited on page 102)

[125] Jim Gettys and Kathleen Nichols. "Bufferbloat: dark buffers in the internet". In *Communications of the ACM*, 2012.
(Cited on page 102)

[126] Nick Weaver and Jim Gettys. "Bufferbloat: what's wrong with the internet?". In *Communications of the ACM*, 2012.
(Cited on page 102)

[127] David Hayes, Ing-Jyh Tsang, David Ros, Andreas Petlund, and Bob Briscoe. "Internet latency: Causes, solutions and trade-offs". In *IEEE EuCNC*, 2015.
(Cited on page 102)

[128] Mark Allman, Sally Floyd, and Craig Partridge. "Increasing TCP's initial window". September 1998. RFC 2414, 1998.
(Cited on page 102)

[129] Van Jacobson. "Congestion avoidance and control". In *ACM SIGCOMM*, 1988.
(Cited on page 102)

[130] Peng Yang, Juan Shao, Wen Luo, Lisong Xu, Jitender Deogun, and Ying Lu. "TCP congestion avoidance algorithm identification". In *ACM ICDCS*, 2014.
(Cited on page 102)

[131] Kathleen Nichols and Van Jacobson. "Controlling queue delay". In *Communications of the ACM*, 2012.
(Cited on pages 102 and 103)

[132] Naeem Khademi, David Ros, and Michael Welzl. The new aqm kids on the block: an experimental evaluation of CoDel and PIE. In *IEEE INFOCOM Workshops*, 2014.
(Cited on page 103)

[133] Kathleen Nichols, Van Jacobson, Andrew McGregor, and Jana Iyengar. "Controlled Delay Active Queue Management". Internet-Draft draft-ietf-aqm-codel-02, IETF Secretariat, December 2015. http://www.ietf.org/internet-drafts/draft-ietf-aqm-codel-02.txt.
(Cited on pages 103 and 108)

[134] Rong Pan, Preethi Natarajan, Chiara Piglione, and Mythili Prabhu. "PIE: A lightweight control scheme to address the bufferbloat problem". Internet-Draft draft-pan-tsvwg-pie-00, IETF Secretariat, December 2012. http://www.ietf.org/internet-drafts/draft-pan-tsvwg-pie-00.txt.
(Cited on page 103)

[135] Sally Floyd, Ramakrishna Gummadi, Scott Shenker, et al. "Adaptive RED: An algorithm for increasing the robustness of red's active queue management", 2001.
(Cited on page 103)

[136] William Diego, Isabelle Hamchaoui, and Fabrice Guillemin. "Procédés de traitement de paquets de données, dispositif, produit programme d'ordinateur, medium de stockage et neoud de réseau correspondants ". *Patent Application submitted to INPI N° 1561613* – France, November 2015.
(Cited on pages 103 and 121)

[137] IETF Network Working Group. RFC 2475 "An Architecture for Differentiated Services". 1998.
(Cited on page 104)

[138] Hossein Falaki, Dimitrios Lymberopoulos, Ratul Mahajan, Srikanth Kandula, and Deborah Estrin. "A first look at traffic on smartphones". In *ACM SIGCOMM conference on Internet measurement*, 2010.
(Cited on page 111)

[139] Katsuhiko Ogata. *"Modern control engineering"*. Prentice Hall PTR, 2001.
(Cited on page 111)

[140] Mehmet Yavuz and Farid Khafizov. "TCP over wireless links with variable bandwidth". In *IEEE VTC Fall*, 2002.
(Cited on page 117)

[141] Cisco Systems Inc. "P-GW Administration Guide, StarOS Release 19", December 2015.
(Cited on page 117)

[142] 3GPP. "Solution to key issue on QoS framework". Technical Report S2-161280 SA WG2 Meeting #113 AH; Orange, Intel; February, 2016.
(Cited on page 120)

[143] Fabrice Guillemin and Alain Simonian. "Analysis of the Shortest Queue First service discipline with two classes". In *EAI ValueTools*, 2013.
(Cited on page 122)

[144] 3GPP. "Discussion on the QoS control in the NextGen RAN". R2-162622 RAN WG2-93 Dubrovnik, Croatia, 11-15 April, 2016.
(Cited on page 122)

[145] William Diego, Isabelle Hamchaoui, and Fabrice Guillemin. "SloMo: An Implicit Cross-Layer Mechanism for a Better Experience on Mobile Networks". In *to be submitted as a conference article*, –.
(Not cited)

[146] 3GPP2 C.R1002-0. "CDMA2000 Evaluation Methodology". Technical report, 2004.
(Cited on page 148)

[147] Yufei Cheng, Egemen K. Çetinkaya, and James P. G. Sterbenz. "Transactional traffic generator implementation in ns-3". In *EAI SIMUTools workshop*, 2013.
(Cited on page 148)

[148] Andrea Bacioccola, Claudio Cicconetti, and Giovanni Stea. "User-level performance evaluation of VoIP using ns-2". In *IEEE ICST*, 2007.
(Cited on page 149)

[149] Majed Haddad, Eitan Altman, Rachid El-Azouzi, Tania Jiménez, Salah Eddine Elayoubi, Sana Ben Jamaa, Arnaud Legout, and Ashwin Rao. "A survey on YouTube streaming service". In *In ICST 2011*.
(Cited on page 151)

[150] Fabio Sonnati. "Implementing a dual-threshold buffering strategy in Flash Media Server", 2006.
(Cited on page 151)

[151] Florian Wamser, Pedro Casas, Michael Seufert, Christian Moldovan, Phuoc Tran-Gia, and Tobias Hossfeld. Modeling the youtube stack: From packets to quality of experience. *Computer Networks*, 2016.
(Cited on pages 151, 152, and 154)

[152] Hyunwoo Nam, Bong Ho Kim, Doru Calin, and Henning G Schulzrinne. "Mobile video is inefficient: A traffic analysis". *Columbia University, Columbia University, 2013*.
(Cited on page 152)

[153] Pedro Casas, Pierdomenico Fiadino, Andreas Sackl, and Alessandro D'Alconzo. "YouTube in the move: Understanding the performance of Youtube in cellular networks". *In IEEE IFIP 2014*.
(Cited on page 152)

[154] Yao Liu, Fei Li, Lei Guo, Bo Shen, and Songqing Chen. "A comparative study of android and iOS for accessing internet streaming services". In *Passive and Active Measurement*. Springer Berlin Heidelberg, 2013.
(Cited on page 152)

Résumé en Français

Depuis quelques années le trafic de l'internet mobile ne cesse d'augmenter. Cette croissance soutenue est liée à plusieurs facteurs, parmi lesquels l'évolution des terminaux, la grande diversité des services et des applications disponibles et le déploiement des nouvelles technologies d'accès radio mobile (3G/4G). À cet égard, le standard 3GPP pour les réseaux LTE propose une architecture offrant une gestion fine de la QoS (par flux). Ce modèle, hérité des réseaux mobiles traditionnels orientés connexion, soulève des problèmes en termes de scalabilité, efficacité et performances.

Les travaux entrepris dans cette thèse ont pour objectif principal de proposer des solutions plus simples et moins coûteuses pour la gestion de la QoS dans les réseaux mobiles. À cette fin, à l'issue d'une étude et de l'évaluation de l'impact de la signalisation associée au modèle de QoS standard, deux modèles alternatifs ont été proposés. Nous proposons tout d'abord un modèle basée sur les mécanismes IP inspiré de l'approche DiffServ (par agrégat) largement étudié dans les réseaux IP fixes. Ce modèle fournit une gestion de la QoS simple, efficiente et rentable, tout en garantissant des performances équivalentes au modèle standard. Cependant, elle nécessite une remise à niveau de tous les eNB, et donc une longue phase de transition.

En conséquence, nous proposons SloMo qui vise à améliorer l'expérience des clients mobiles, mais avec un objectif de déploiement plus rapide. SloMo est une solution de gestion implicite de la QoS depuis un point unique situé sur le chemin des communications. SloMo exploite la dynamique instaurée par le mécanisme de contrôle de flux de TCP. Il vise à recréer un goulot d'étranglement dynamique dans un équipement contrôlé par l'opérateur lorsque les points de congestion réels ne sont pas accessibles. Une fois ce goulot d'étranglement déporté, il est alors aisé d'effectuer une gestion de la qualité IP classique dans l'équipement supportant Slo-Mo.

Mots-clés : *Réseau sans fil, Qualité de Service, LTE, DiffServ, Analyse de performance, IP centric, conception inter couches, TCP, convergence fixe et mobile*

1 Introduction

Ces dernières années, les différents acteurs issus du monde des réseaux mobiles ont constaté l'explosion du trafic data dans ce domaine. Ce processus a été accéléré récemment par la prolifération des forfait de data mobiles moins chers et par l'utilisation croissante des réseaux mobiles comme un substitut à la connectivité traditionnelle par ligne fixe.

Ce phénomène a été mis en évidence dans différents rapports comme [1, 2]. Il est due principalement à l'actuelle évolution du marché mobile en termes de Business Model et des utilisations. Poussé par l'augmentation de la puissance de calcul des terminaux mobiles (smartphones) et la grande diversité des services et des applications qui sont disponibles; cet écosystème est aussi impacté par un changement majeur dans l'équilibre du pouvoir entre ses différents acteurs.

En effet, les OTT sont devenu des acteurs clés dans le paysage mobile, apportant avec eux des interfaces orientées web, des services et des modèles issues du monde de l'Internet fixe au détriment des opérateurs mobiles [3]. Toutefois, les normes des réseaux mobiles n'ont pas évolué au même rythme et ne répondent plus aux exigences actuelles d'un modèle d'Internet plus ouvert et plus flexible.

2 Etat de l'art

Pour faire face à cet engorgement du trafic data mobile, les mécanismes de QoS permettent de privilégier les services sensibles tels que la vidéo en cas de congestion du réseau. À cet égard, l'organisme de standardisation issu du monde mobile (3GPP) propose un nouveau modèle capable d'offrir une gestion fine de la QoS dans la nouvelle évolution des systèmes mobile appelé LTE/EPC.

Contrairement aux anciennes normes mobiles (e.g. UMTS), qui se composent d'un domaine à commutation de circuits (CS) dédié à la voix et d'un domaine à commutation de paquets (PS) pour la data mobile, cette nouvelle norme est basé sur une architecture tout-IP avec un unique domaine à commutation de paquets pour la voix et la data mobile. Cependant, ce modèle a été hérité des anciens réseaux mobiles orientés connexion et soulève des problèmes en termes de scalabilité (nombre de bearers), efficacité (charge de signalisation) et performance (délai d'établissement des bearers).

A l'opposé, les mécanismes de QoS issus du monde de l'Internet, éprouvés depuis plusieurs décennies dans les réseaux de paquets, se caractérisent par leur robustesse et leur simplicité de mise en œuvre. Le transport des flux mobiles est aujourd'hui de plus en plus assuré par une couche IP, en particulier dans le cadre des réseaux LTE/EPC, l'adaptation de ces mécanismes issus du monde de l'Internet aux réseaux mobiles semble particulièrement prometteuse.

2.1 Gestion de la QoS du modèle 3GPP

Les systèmes de communication mobiles standardises par la 3GPP (2G/3G) se composent d'un domaine CS et un domaine PS, ce qui permet de séparer les services conversationnels des services data mobile. Dans le domaine PS, la quasi totalité des flux sont transportés en mode Best-Effort, sans aucune garantie de QoS. Grâce à cette séparation des domaines, les services conversationnels ne sont pas affectés par le trafic data mobile. En revanche, le nouveau système LTE/EPC est un réseau tout-IP, où le domaine CS n'existe plus. Cette nouvelle architecture présente des défis pour la gestion de la QoS, en particulier pour les services en temps réel. Par conséquent, des mécanismes pour gérer la QoS devraient être mises en œuvre, afin de permettre une coexistence harmonieuse entre les différents flux avec des exigences de QoS très variées.

Tel qu'elle est mentionnée ci-dessus, la gestion de la QoS dans les réseaux LTE/EPC est clairement orientée connexion [5]. Un tunnel virtuel, appelé bearer EPS, doit être mis en place entre les éléments finaux (i.e., UE, terminal et une Packet Data Network GateWay, P-GW) avant qu'un échange de trafic puisse être effectué entre eux. Il faut noter que ce bearer EPS est constituée de bearers locaux, établis entre des éléments voisins du réseau. Par exemple, un radio bearer est établi entre l'eNB et UE. Ce bearer EPS fournit un service de transport avec des attributs de QoS spécifiques. Lorsqu'un UE est attaché au réseau, un bearer par défaut avec une QoS par défaut est établie. D'autres bearer pouvent encore être mis en place, un par niveau requis de QoS. Les bearers sont exploités en mode connecté, c'est à dire qu'ils sont établis, modifiés ou libérés par moyen de protocoles de signalisation du plan de contrôle mobile.

Du côté de la P-GW, les flux de trafic dans le sens descendant sont envoyés aux bearer EPS pertinentes grâce à un filtre décrivant la correspondance entre les flux de trafic (i.e. identifiés par leurs en-têtes TCP/IP, etc) et les bearers. Côté eNB, les ressources radios sont allouées dynamiquement aux UEs actives selon l'algorithme d'ordonnancement radio, à chaque TTI (Transmission Time Interval). Différents exemples d'algorithmes d'ordonnancement peuvent être trouvés dans la littérature.

L'algorithme Proportional Fair (PF) [46] est l'un des plus connus et le plus implémenté par les équipementiers de réseaux mobiles. Il est largement utilisé pour l'exploitation des bearers par défaut. Cet algorithme propose un compromis entre l'optimisation de débit global de la cellule et l'équité entre les UEs. Les performances de cette solution sont excellentes en termes de QoS, mais elle présente l'inconvénient majeur de réduire considérablement la capacité globale de la cellule radio.

Malgré ses bonnes performances, on constate aujourd'hui que la solution recommandée par le 3GPP est peu déployée dans les réseaux mobiles tels qu'Orange. Actuellement, la solution typique consiste à utiliser uniquement le tunnel par défaut. Dans ce tunnel, la P-GW transmet tous les trafics sans aucune gestion de la QoS.

2.2 Coût de la gestion de la QoS 3GPP

Comme il a été mentionné précédemment, le modèle 3GPP utilise un circuit virtuel appelé "EPS bearer". Chaque terminal doit établir un EPS bearer par défaut avant qu'un échange de donnes ne soit effectué. Un terminal est capable d'établir plusieurs EPS bearers, un pour chaque niveau de QoS requis. Tous les bearers sont opérés en mode connecté, ceci implique son établissement, sa libération, ainsi que sa modification. Dans ce sens, il est très important d'évaluer le "coût" qu'implique l'utilisation du modèle de QoS normalisé.

Dans le chapitre III nous présentons un modèle analytique pour évaluer l'impact de la QoS standard. Ce modèle prend en compte principalement la dynamique instaurée par la machine à état lié à l'établissement et libération de bearers. Il prendsen compte également des mécanismes liés à la mobilité (i.e. Handover, Tracking Area Update) et la mixité de trafic que sont caractéristiques des réseaux mobiles d'aujourd'hui.

Les résultats obtenus par simulation montrent que la gestion de QoS standard pourrait avoir un impact négatif dans les réseaux mobiles, lié principalement à l'augmentation de la signalisation et du nombre de contextes que chaque équipement doit gérer. Finalement, lorsque la gestion de la QoS standard est déployée, un dimensionnement approprié est fortement conseillé en terme de mémoire et processeur.

3 Modele IP-centric

Inspirés du monde de l'Internet fixe, ce sujet de thèse a pour but l'étude et proposition d'une architecture de QoS IP-centric (mode sans connexion), moins minutieuse mais qui résout les problèmes mentionnés ci-dessus. Peu coûteuse et facile à gérer et à déployer, elle ouvre la porte à une convergence fixe-mobile simple et efficace. Ce modèle, inspiré des réseaux fixes et des architectures web, est en rupture avec les solutions actuellement normalisées par le 3GPP.

Dans ce sens, une architecture de QoS alternatif pour les réseaux mobiles peut être envisagée sur la base de l'expérience acquise sur l'Internet fixe: les réseaux IP fixes sont généralement utilisés en mode sans connexion, ce qui signifie que chaque paquet contient dans leur en-tête suffisamment d'informations pour être livré à la bonne destination avec la QoS requise. Cette façon de gérer la QoS pourrait simplifier le modèle de la QoS de la 3GPP et ouvrir la porte à une convergence fixe-mobile simple et efficace, grâce à une architecture de QoS IP-centric. Cette approche commence à être étudiée pour quelques acteurs du monde des réseaux mobile comme Nokia [84], Alcatel Lucent et AT&T [85].

Le modèle IP-centric [78, 90, 91, 92], pourrait être utilisés sur l'architecture classique 3GPP (mode orienté connexion), aussi bien que sur une architecture en mode sans connexion (e.g. sur la base de SDN - Software Defined Networking) qui actuellement commence a être proposé dans différents articles [96, 97, 98, 99, 100].

Dans le chapitre IV une architecture du modèle IP-centric est proposée sur la base de l'architecture actuel 3GPP. Par conséquent, la connectivité des UEs est encore exploité en mode connecté par le biais d'un bearer EPS, mais la QoS est gérée paquet par paquet; et les paquets sont transportés sur un unique bearer multi-QoS. Le marquage DSCP (DiffServ Code Point) devrait être potentiellement pris en compte par chaque noeud supportant la couche IP E2E (End-to-End) afin de prioriser correctement les paquets. Les paquets sont tout simplement priorisés en fonction de leur marquage DSCP à l'intérieur du bearer affecté à chaque UE.

Des autres fonctions bien connues de gestion de QoS IP pourraient être adaptées à l'écosystème mobile (e.g. AQM - Active Queue Management, CoDel - Controlled Delay, etc). En effet, ces mécanismes IP se sont révélées être flexibles, rentables, évolutives, facile à configurer et bien adaptés aux écosystèmes ouverts.

Le modèle IP-Centric propose donc une gestion de la QoS de niveau IP via un système de files d'attentes à priorité dans l'eNodeB. Présentée et discutée au sein de l'ITU-T [4] et de l'IETF [5], cette solution n'est cependant pas standardisée au 3GPP ni implémentée à la connaissance de l'auteur. Il est probable que cette gestion de QoS dans l'eNB inspirée des réseaux IP fixes finisse par s'imposer, mais elle nécessitera une remise à niveau de tous les eNode B, et donc une longue phase de transition.

4 Slo-Mo : un mécanisme implicite pour la gestion de la QoS

Les architectures IP-centric évoquées au chapitre IV impliquent une action en normalisation, puis un effort d'implémentation des constructeurs d'eNB, et enfin une mise à jour des eNB. Il s'agit donc d'un processus relativement long, que l'on ne peut espérer voir aboutir avant plusieurs années. Dans le chapitre V nous introduisons le mechanisme Slo-Mo qui vise à améliorer l'expérience des clients mobiles, mais avec un objectif de déploiement plus rapide.

Slo-Mo propose une solution légère et peu coûteuse permettant aux opérateurs d'accès de reprendre le contrôle de la gestion de la qualité et de stopper leur désintermédiation.

Slo-Mo est une solution de gestion implicite de la qualité de service depuis un point unique situé n'importe où sur le chemin des communications, comme par exemple une passerelle convergente fixe/mobile. Contrairement à d'autres solutions de gestion de la qualité à distance, Slo-Mo ne repose sur aucune analyse protocolaire du trafic - qui peut donc être chiffré - et ne fait pas d'hypothèse sur les protocoles utilisés (e.g. TCP, UDP). Slo-Mo ne requiert aucun routage particulier puisqu'il peut traiter chaque sens de communication indépendamment.

Slo-Mo vise à recréer un goulot d'étranglement dynamique dans un équipement con-

trôlé par l'opérateur lorsque les points de congestion réels ne sont pas accessibles. Le ralentissement réalisé par Slo-Mo s'ajuste en permanence aux ressources réelles dont dispose le terminal considéré. De ce fait, les files d'attente de Slo-Mo se remplissent légèrement, et celles des goulots d'étranglement réels (e.g. antennes 3G/4G) se vident. Il est alors aisé d'effectuer une gestion de la qualité IP classique dans l'équipement supportant Slo-Mo.

Slo-Mo améliore l'expérience client à un coût très marginal, permettant de l'inclure dans toutes les offres d'accès. Slo-Mo est un "quick-win" permettant de valoriser les réseaux du groupe en préservant la qualité d'expérience. Ce mécanisme est sans interaction avec l'OTT et à très faible impact sur le réseau.

Les performances de Slo-Mo ont été évaluées par simulation sous ns3 dans un scénario inspiré de conditions réelles (e.g. modèle de propagation, nombre d'utilisateurs, accès radio). Sur la base des résultats de simulations, des optimisations de Slo-Mo ont été proposées afin d'améliorer la QoS tout en limitant l'impact sur le débit global de la cellule. La configuration retenue permet d'améliorer la QoS des flux sensibles au prix d'une dégradation faible à modérée de la capacité globale de la cellule. Notons que notre évaluation de cette dégradation est pessimiste, car fondée sur le débit global incluant les pertes et retransmissions, pertes et retransmissions fortement réduites par Slo-Mo.

Appendix A

1 Traffic Modeling and Implementation

The landscape in mobile data traffic is changing rapidly, as reported by Cisco [1] and Ericsson [2]. Today, mobile devices as smartphones and tablets have better connectivity capabilities, which are close or some times better than fixed devices. Traffic patterns have also changed, Music streaming (e.g. Deezer, Spotify), Video on Demand (VoD) (e.g. Netflix, CanaPlay) and video streaming (e.g. YouTube, Dailymotion) are just examples of services that are available today, and that some years ago were almost nonexistent. This evolution of the usage of web-based services due to the unlimited data plans, added to the increased network capacity (i.e.) as well as the evolution of mobile services like VoIP (e.g. Skype, Viber) and instant messaging applications (e.g. WhatsApp, LINE), which are preferred instead of Telco conventional voice call and SMS services, brings new challenges for Telco operators. This leads them to innovate and search new solutions in order reduce the CAPEX and OPEX without impact the QoS.

1.1 Web Traffic Model

As the major contributor of transactional traffic, Hypertext Transfer Protocol (HTTP) is a pervasive application protocol and consumes a significant share of application flow in the Internet. Web traffic has transformed from plain-text Web pages to large size pages with embedded objects. An HTTP traffic model is needed to accurately represent and simulate Web traffic with the sustaining influence of HTTP over the Web. In order to have realistic traffic HTTP, we use HTTP traffic model proposed by [109], which was implemented on ns-3 simulator

In order to have realistic traffic HTTP, we use HTTP traffic models proposed by [109]:

The delay between two consecutive page requests is called "reading time". Nevertheless the *"reading time"* model proposed in [109] takes into account that users tends to move around web pages before one whole page finishes download its embedded objects, which

Parameter	Mean	Median	Max	Standard Deviation	Best fit
Main object size	31,561 Byte	19,471 Byte	8 MB	49,219 Byte	Weibull (28242.8,0.814944)
Compressed	22,468 Byte	11,535 Byte	8 MB	41,295 Byte	Weibull (19104.9,0.771807)
Number of main objects	2.19	1	212	2.63	Lognormal $\mu = 0.473844, \sigma = 0.688471$
Inline object size	23,915 Byte	10,284 Byte	8 MB	128,079 Byte	Lognormal $\mu = 9.17979, \sigma = 1.24646$
Compressed	21,208 Byte	7,338 Byte	8 MB	127,979 Byte	Lognormal $\mu = 8.91365, \sigma = 1.24816$
Number of inline objects	31.93	22	1920	37.65	Exponential $\mu = 31.9291$

Tableau A.1 – HTTP model parameters [109]

is not the case in smartphones. In this sense we use the "reading time" proposed by [146], which describes the necessary framework for simulating the performance of a cdma2000 system (reading time is modeled with an exponential distribution with mean = 30s). Figure A.2 shows the model behavior. The communication is composed of page requests of fixed size, each one followed by one main object plus zero or more embedded objects. The time between two consecutives object downloads is called server response time (in simulation is considered 1s for simplicity). The number of objects per page and their respective size could be modeled using the table A.1 models. We also established a maximum timeout for web page request in order to avoid errors during simulation, we fixed this parameter at $30s$. fter exceeded the timeout, session is closed and a new request is sends. LTE

Our ns-3 implementation was inspired of [147] HTTP traffic generator, which is able to generate HTTP 1.0 traffic as well as HTTP 1.1 traffic with persistent connection and pipelining. The proposed model is capable of generating both *Internet-like* traffic and *user-defined* traffic, which are the two working modes of this generator. Nevertheless, this simulator is based on old internet traffic traces and is not real time generator. It generates all traffic and schedule its execution time at the beginning of simulation and it is not adapted for channel varying environments as LTE or wifi.

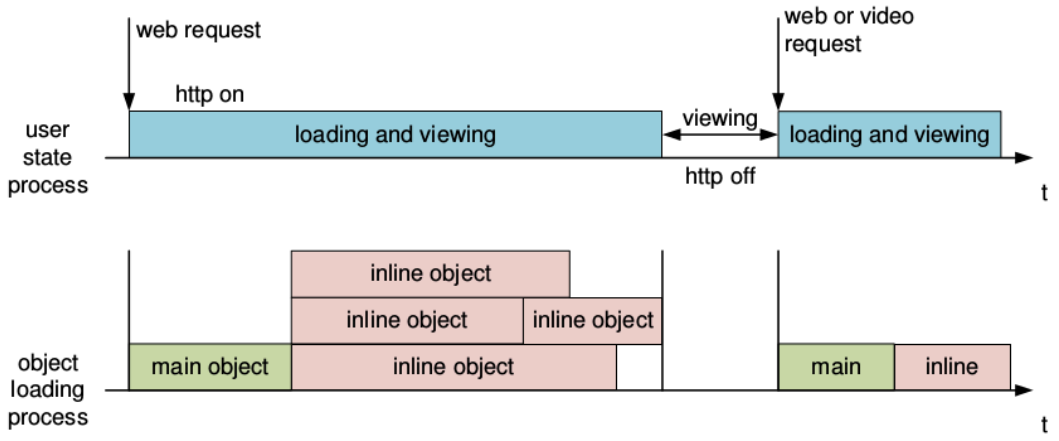


Tableau A.2 – Behavioral model

We use 5 top french web sites based on Alexa's ranks in order to simplify and harmonize the KPI measurements. The model traffic was implemented as is described above, where the number of objects and their size match with www.websiteoptimization.com measurements, see Table A.3. The KPI is the page delay (i.e. the time needed to receive a full page, including all the embedded objects, starting from the time the request is issued).

Web Site	**Avg. Size (Bytes)**	**Total Objs**
Orange.fr	2366380	129
pole-emploi.fr	1205991	320
Lefigaro.fr	1054458	50 LTE
Facebook.com (profile Facebook France)	1047690	185
Leboncoin.fr	728592	78

Tableau A.3 – Traffic distribution

1.2 Voice Traffic Model (AMR-NB)

Voice over IP (VoIP) is modeled according to [148]. The employed codec is the Adaptive Multirate Narrow Band (AMR-NB) (12.2 kbit/s) with VAD (no packets are sent during silences) and without header compression.

The model states that voice traffic at the source is characterized by two periods; an active or ON period and an inactive or OFF period. During the ON period, the source sends packets at regular intervals of length 110 Bytes (Packetization time), according to table A.4.LTE

Talkspurt duration - ON Time	Weibull distribution: Shape = 1.423 / Scale = 0.824
Silence duration - OFF Time	Weibull distribution: Shape = 0.899 / Scale = 1.089
Codec type	AMR-NB (12.2 kbps)
VAD model	One-to-one conversation
Packet length	32 bytes/frame + 78 bytes headers = 110 bytes

Tableau A.4 – VoIP model parameters (source: [107])

In order to measure the QoE of VoIP, we use the R-factor specified in ITU G.107 [33], which is expressed as:

$$R = R_o - I_s - I_d - I_{e-eff} + A$$

- R_o represents in principle the basic signal-to-noise ratio;

- I_s is a combination of all impairments which occur more or less simultaneously with the voice signal;
- I_d represents the impairments caused by delay and the effective equipment impairment;
- I_{e-eff} represents impairments caused by low bit-rate codecs;
- A allows for compensation of impairment factors when the user benefits from other types of access to the user.

Based on ITU G.107 recommendation, the R-factor equation can be simplified as:

$$R = 93.2 - I_d - I_e - A$$

Where A is 0 for wireline and 5 for wireless networks. The value of I_e, is calculated as:

$$I_e = a + b \times ln(1 + c \times \frac{P}{100})$$

Where, P is the Packet Loss Ratio (PLR) and a, b and c are codec fitting parameters, which are specified in [108] for AMR-NB (12.2 kbps - $a = 14.96$, $b = 16.68$ and $c = 30.11$). The value of I_d, which is impairment due to delay is calculated as:

$$I_d = 0.024 \times d + 0.11 \times (d - 177.3) \times H_{(d-177.3)}$$

Where: $H_{(x)} = 0$ if $x < 0$ and $H_{(x)} = 1$ otherwise.

1.3 YouTube Traffic Model

YouTube Characteristics

Most popular video streaming services (i.e. YouTube, Netflix) use HAS. HAS, split up media (i.e. video) into a series of small files called chuncks, which are then encoded using different video qualities [110]. Each chunk is transmitted individually as a single web object via plain HTTP. In the course of playout of the video, the client continuously assess available bandwidth and requests successive chunks for the data rate that can be supported. Typically, the client keeps a buffer of chunks to deal with eventual network issues (e.g. latency, packet loss, connection loss).

In order to provide compatibility with all browsers, devices, bandwidth and quality requirements, a wide range of encoding for every video file is available for clients, which can be selected according needed. Besides, a numerical identifier named "$itag$" is used in order to identify different encoding schemes of a video. The $itag$ " information is included in the HTTP requests. Moreover each chunk is transmitted individually as a single web object via plain HTTP. During the playout of the video, the client continuously estimates available bandwidth and requests chunks for the data rate that can be supported. The client try to keep a buffer of video data to deal with eventual network issues (e.g. latency, packet loss, connection loss), in order to perform it, a buffering strategy is necessary.

YouTube Buffer Management

The video buffering strategy is a key element on video streaming because it could permit the optimization of network resources (i.e. bandwidth, radio resources) or on the worst case the waste of it. Below is a brief description of most used video buffering strategy, which are described in [110, 149, 150, 151]:

- **Standard buffering:** In standard buffering strategy the device receives all video traffic (i.e. chunks) and keep data in the buffer until it is full. At this point, playback of the video starts and the video application tries to keep the buffer full along the video playback, in this regard the device will request to server the needed video data. In case the network condition are disturbed and the instant bandwidth goes below the value required for the video, the buffer data is used to fill the gap. In worst case when the buffer is empty, the video is interrupted until new video data are buffered.

- **Dual-threshold buffering:** In case of mobile networks, the standard buffering strategy is vulnerable to bandwidth drops, as well as being unable to exploit a increase of bandwidth. Dual-threshold buffering strategy is more flexible and try to fill the gaps of standard buffering strategy. It provides a resilience to data rate fluctuations or other adverse conditions associates to nature of wireless media.

  In the dual-threshold buffering strategy a initial buffering is performed before the playback of the video, which consists of the fill a first threshold B_{min} (*lower threshold*) in the buffer. In this strategy, instead of trying to keep the buffer full to this min level B_{min}, it tries to fill the buffer to a second higher level B_{max} (*upper threshold*). These additional video data will be useful if the network connection encounters temporary impairments. In worst case when the buffer is empty, the video playout can start after a short pre-loading time. Also, in case of an increase of data rate, a bigger buffering of video chunks to counteract the possibility of network malfunctions can be performed. Details of dual-threshold buffering are shown in Figure A.1

YouTube Traffic Models

We can found several YouTube measurement studies in literature. Much of them were focused on characterizing various aspects of YouTube videos, as well as its usage patterns and its strategy depending on supported hardware and software of terminals. But only few of these studies are focused in mobile YouTube traffic characterization.

In [152] authors analyze YouTube and Netflix video streaming using iOS and Android devices over wireless networks (WiFi, 3G and LTE). They found that when a client requests a video, the resolution is selected based on the device types (screen size), regardless of OSs on the devices or access networks. They also found that video players frequently terminated the TCP connection and open a new TCP connection to continue receiving the video content. The number of TCP connections varies depending on the playback buffer management policies of the video players running on different OSs. In [**?**] the above authors propose dynamic QoS-aware rules for LTE networks to select an appropriate video resolution under a fluctuating channel condition, in order to reduce the waste of the video content and enhance the QoE for end-users. They also found that YouTube uses a single TCP connection, which is opened and closed along the video streaming.

In [153] authors present an empirical study of the performance of YouTube in cellular networks. Is showed that the complex and dynamic CDN (Content Delivery Network) architecture of YouTube has better service performance in video over cellular networks in terms of improved QoE (delay and throughput) and user engagement compared to other CDNs providing HTTP video streaming.

In [154] authors analyze and compare the performance when Android and iOS devices are accessing Internet streaming services.

Is very hard to identify a valid reference of YouTube traffic characterization, because it is constantly evolving. In this paper we try to present an state-of-art of YouTube and highlight the most important evolutions that can impact its traffic characteristics. Moreover, we study the mobile ecosystem because is relative few explored field and additionally all our studies are focused on it. In this paper, we propose and evaluate a YouTube traffic generator based on models presented in [110, 151] and [111], which propose a YouTube traffic model described below.

Each application instance is composed by a video server that streams data via a TCP connection to a video client. The chunk duration and codec is 5 seconds. The client device uses a playlist information which provides several different profiles of video quality levels and is identified by the *itag* values, for example those presented in Table A.5. These profiles is used in order to choose the appropriate quality, according to the network and the devise capabilities.

At the beginning of the communication, the device requests a chunk with the lowest video quality (*itag* = 132, 266 kbps, 426 x 240), after that the device estimates its data rate based on the received chunk, it automatically selects the highest playable video quality

itag	Resolution	Encoding rate
132	426 x 240	266 kbps
92	426 x 240	395 kbps
93	640 x 360	758 kbps

Tableau A.5 – YouTube video quality information

and sends the YouTube server a request message. The ratio between the *Throttling Phase* throughput and the encoding rate is referred as the *throttling factor*. In [110] was found that mobiles use a throttling factor of 2.0, for encoding rates higher than 200 kbps. It is also showed that some terminals use a dual-threshold buffering strategy and others a standard buffering strategy, we implement the first one because the second one can be easily emulated using a high B_{max} and fixing $B_{min} = B_{max}$.

The server first sends an *Initial Burst*, corresponding to $35s$ seconds of video data, before to start the playing of the video. When the amount of data in the player buffer exceeds approximately $100s$ of video, the client aborts the TCP connection. The download is interrupted for approximately $60 - 70s$. When the amount of data in the player buffer falls below approximately $30s$, the terminal opens a new TCP connection to request to the server the next video segment. This behavior is repeated until the full video is downloaded. This threshold strategy avoids wasting data if the user aborts the video playback. The upper and lower threshold that we use are $100s$ and $30s$, respectively, but these values can be modified. We also established a maximum timeout for chunk request in order to avoid errors during simulation, we fixed this parameter at $30s$. After exceeded the timeout, the video session is stopped and it starts a new video session.

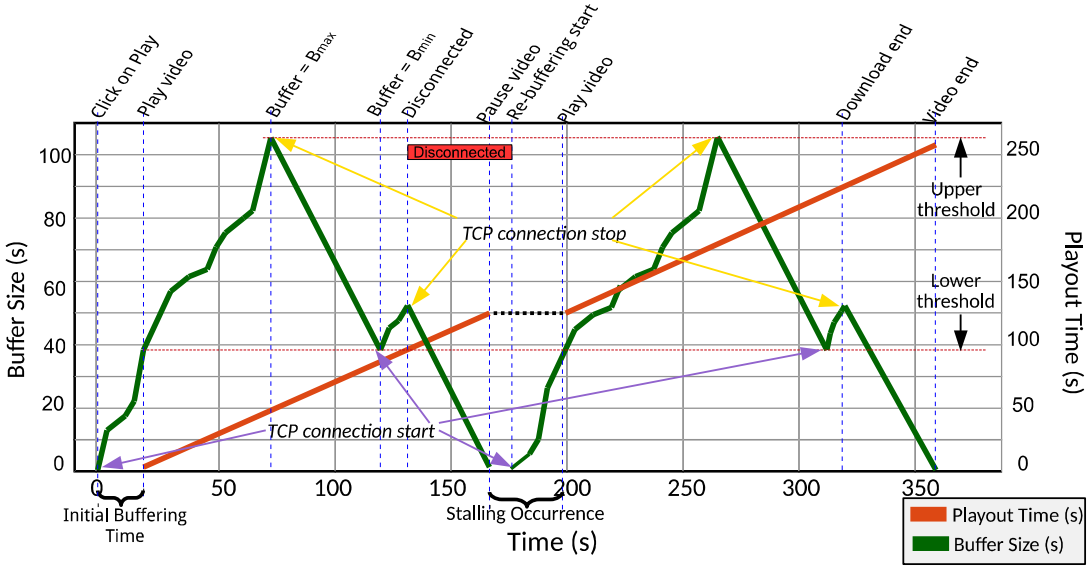


Figure A.1 – Dual-threshold buffering strategy Playout time and Buffer state illustration

YouTube Traffic model in ns-3

This section describes the design of our YouTube traffic generator for ns-3. We have implemented a specific module in ns-3 in order to emulate the delivery of mobile YouTube traffic. For this purpose, we have taken advantage of the traffic model described on [110, 151] and [111].

Basic Structure As shown in Figure A.2, the `YoutubeClient` and `YoutubeServer` applications are responsible for the major functionalities, such as generating the adaptive traffic, handling HAS processes, as well as recording and process statistics. When the mobile YouTube model starts, the `YoutubeClient` and `YoutubeServer` applications are installed in client and server nodes, respectively. Both applications start a new TCP connection; then the evolution of the communication between the `YoutubeClient` and `YoutubeServer` is described in Figure A.3.

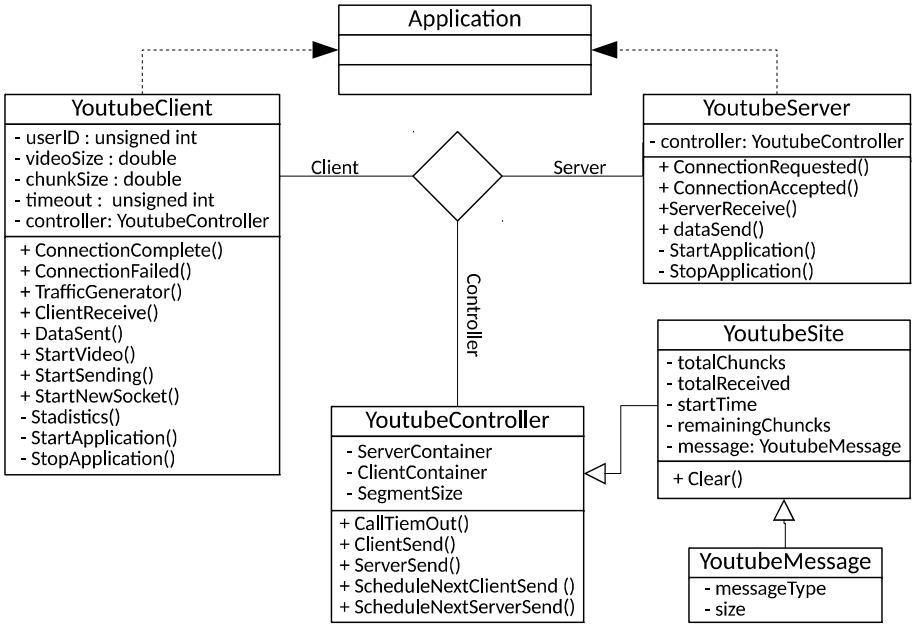


Figure A.2 – YouTube class diagram

We also define two main attributes, `VideoSize` and `Timeout`, which define the video duration and the maximum time to wait a video chunk before to restart the video session respectably.

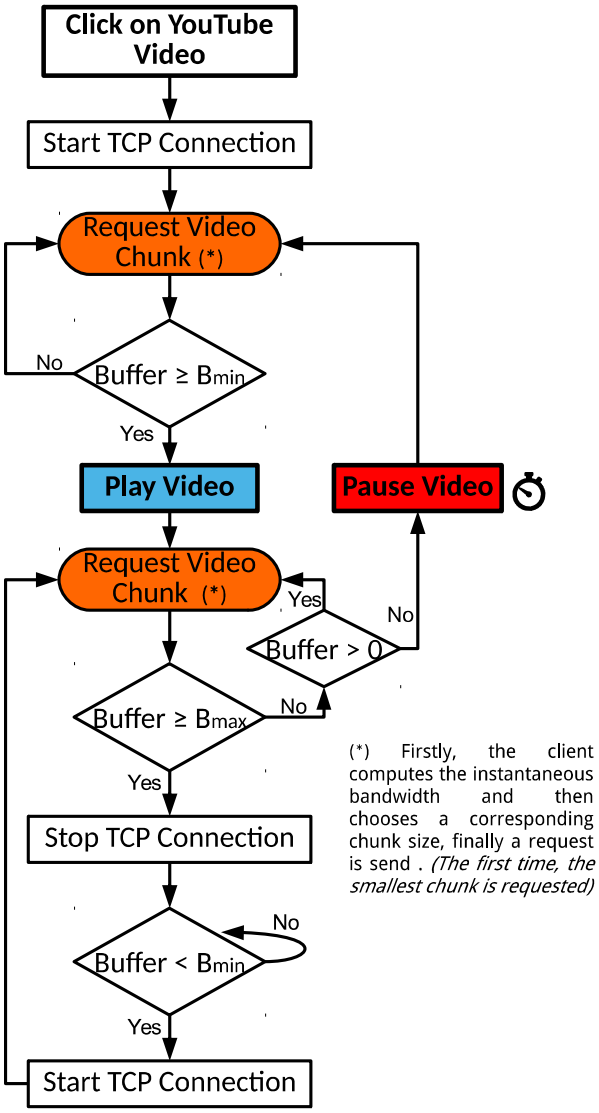


Figure A.3 – Flow description of our YouTube traffic generator

Appendix B

Multiple ON/OFF processes

Since each ON-OFF process is considered as a random ergodic process, the studied multiple ON/OFF processes can be modeled as an $M/G/n/n/n$ queueing system. Where the stationary probability π_0 that all applications are inactive (equation III.6) is the probability that the queueing system is empty. However, stationary state equations describing the $M/G/n/n/n$ are not able in the classic literature of queuing systems, thus, we evaluate by simulations equations (III.6) and (III.7).

Let f^i_{off} and f^i_{on} be the overall duration of all inactivity and activity periods of a process i measured during the simulation respectively. Hence, the probability that a process-i is inactive is computed as:

$$\pi'_{\text{off},i} = \frac{f^i_{\text{off}}}{f^i_{\text{off}} + f^i_{\text{on}}} \tag{B.1}$$

Taking into consideration a set of ON/OFF processes, let g_{off} be the overall duration of all periods of total inactivity (no active process) and g_{on} be the overall duration of all periods of total activity (at least one process is active) measured during the simulation. Thus, the probability that all applications are inactive can be computed as follows:

$$\pi'_0 = \frac{g_{\text{off}}}{g_{\text{off}} + g_{\text{on}}} \tag{B.2}$$

From equation (B.1), we can also compute the equation (III.6) as follows:

$$\pi''_0 = \prod_{i=1}^{n} \frac{f^i_{\text{off}}}{f^i_{\text{off}} + f^i_{\text{on}}} \tag{B.3}$$

Finally, we compute the stationary probability π_0 that all processes are inactive as

expressed in equation (III.7):

$$\pi_0 = \prod_{i=1}^{n} \frac{\mu_i}{\lambda_i + \mu_i}$$

We simulate the duration of the OFF states of a process i as an exponential random variable and the duration of its the ON state using the following random distributions:

- Exponential
- Lognomal
- Uniform
- Normal

Table B.1 shows simulation results taking into consideration 4 ON/OFF processes. We set as simulation input μ_i and λ_i values representing different applications behaviour. We computed the probability that all applications are inactive using equations (III.6), (B.2) and (B.3), which corroborates the hypotheses of the equations (III.6) and (III.7).

μ_i^{-1}	λ_i^{-1}	π_0	**Exponential**		**Lognormal**		**Uniform**		**Normal**	
			π_0'	π_0''	π_0'	π_0''	π_0'	π_0''	π_0'	π_0''
10, 5, 30, 2	100, 6, 200, 80	0.4766	0.4774	0.4772	0.4767	0.4765	0.4755	0.4753	0.4761	0.4762
	12, 30, 40, 5	0.1908	0.1914	0.1906	0.1928	0.194	0.1925	0.1925	0.191	0.1908
	11, 6, 4, 3	0.098	0.0983	0.099	0.0985	0.0989	0.0981	0.098	0.0977	0.098
3, 5, 2, 2	100, 6, 200, 80	0.5545	0.555	0.5549	0.5544	0.5541	0.5542	0.5543	0.5538	0.5539
	12, 30, 40, 5	0.4482	0.4473	0.4474	0.4513	0.4511	0.4482	0.4486	0.4494	0.449
	11, 6, 4, 3	0.1257	0.1252	0.1251	0.1253	0.1254	0.1252	0.1252	0.1255	0.1252

Tableau B.1 – Simulation result of multiple ON/OFF processes

List of Figures

I.1 3GPP LTE Architecture 6
I.2 Functional Split between E-UTRAN and EPC [10] 8
I.3 EPS bearer architecture 9
I.4 User Plane and Control Plane protocol stack 10
I.5 eNB: Uu interface Downlink protocol stack (simplified - *source: [12]*) . . 11
I.6 Physical Resource Block (PRB) structure 13
I.7 EMM-ECM-RRC states 16
I.8 Switch from CONNECTED state to IDLE state 17
I.9 Bearer states before/after Service Request procedure 17
I.10 LTE Attach and Default Bearer Setup Messaging 18
I.11 UE triggered Service Request procedure 19
I.12 Network triggered Service Request procedure 20
I.13 Handover procedure 20
I.14 X2 Handover procedure 21
I.15 Tracking Area Update procedure 22

II.1 QoS in Heterogeneous Network 24
II.2 ITU/ETSI and IETF approaches and general QoS model [32] 25
II.3 IntServ model 30
II.4 DiffServ domain 31
II.5 Packets Classification and Marking (Ingress Node) 32
II.6 Per-Hop Behavior (PHB) (Interior Node) 33

II.7 Queuing and Scheduling strategy . . . 33
II.8 Examples of main queuing algorithms . . . 34
II.9 IntServ vs DiffServ . . . 35
II.10 Packet Data Protocol (PDP) context . . . 36
II.11 Two EPS bearers across the different interfaces *(Source: [44])* . . . 39
II.12 Dedicated Bearer Setup Messaging . . . 40
II.13 P-GW: Dedicated Bearer Deactivation Call Flow . . . 41
II.14 UE: Dedicated Bearer Deactivation Call Flow . . . 41
II.15 Policy and Charging Control (PCC) logical architecture R8 . . . 42
II.16 QoS Parameters for SDF and EPS Bearer *(Source: [49])* . . . 43

III.1 Modeled RRC States . . . 52
III.2 Modeled Dedicated Bearer Activation/Deactivation . . . 54
III.3 Simulated LTE/EPC mobile network . . . 62
III.4 Impact of the inactivity timer (τ) on eNB, S-GW and MME *Processing Load* 64
III.5 *Processing Load* Increase compared to BE scenario . . . 65
III.6 Impact of the inactivity timer (τ) on eNB, S-GW and MME *Context Load* 65
III.7 Impact of the inactivity timer (τ) on eNB, S-GW and MME *Memory Access Rate* . . . 66
III.8 Memory Access Rate compared to BE scenario . . . 66
III.9 Impact of the VoLTE sessions arrival rate (α) on eNB, S-GW and MME *Processing Load* . . . 67
III.10 Impact of the VoLTE sessions arrival rate (α) on eNB, S-GW and MME *Memory Access Rate* . . . 68
III.11 Impact of the VoLTE sessions arrival rate (α) on eNB, S-GW and MME *Context Load* . . . 68
III.12 Ratio of the *Radio Signaling Overhead* to the mean Applications Data Rate 70

IV.1 Today's mobile networks QoS policies . . . 74
IV.2 IP-centric Architecture . . . 78
IV.3 Intra-bearer arrangements . . . 79
IV.4 Inter-bearer arrangements . . . 79
IV.5 Possible business model with IP-centric QoS approach . . . 80
IV.6 Virtual radio bearer design . . . 81
IV.7 Mono radio bearer design . . . 82
IV.8 eNB-Uu interface - IP-aware downstream protocol stack (simplified) . . . 83
IV.9 YouTube video service . . . 87

IV.10 YouTube video streaming strategy 88
IV.11 Two levels QoS aware schedulers 90
IV.12 Cell Throughput performance 93
IV.13 VoIP QoE in terms of Mean Score Opinion 94
IV.14 YouTube chunks distribution 94
IV.15 YouTube First Buffering Time 94
IV.16 YouTube's Flows Throughput 95
IV.17 Web Page Load Time 96

V.1 Traffic bottlenecks in LTE/EPC networks 99
V.2 TCP congestion control strategies 102
V.3 Bufferbloat illustration 103
V.4 Slo-Mo Mechanism: initial bottleneck location 104
V.5 Slo-Mo Mechanism: bottleneck de-location 105
V.6 Slo-Mo: bit-rate decrease *(High priority queue not depicted)* 107
V.7 Slo-Mo: bit-rate increase 107
V.8 Slo-Mo System Architecture (n Slo-Mo entities for n UEs) 108
V.9 CoDel example 109
V.10 QMon simplified block diagram based on CoDel criteria 110
V.11 Slo-Mo example 110
V.12 Slo-Mo rate tracking simplified block diagram 111
V.13 Simulated scenario 113
V.14 Cell Throughput performance 115
V.15 VoIP QoE in terms of Mean Score Opinion 116
V.16 Goodput measurement for FTP services (*wireshark print-screen*) 116

A.1 Dual-threshold buffering strategy Playout time and Buffer state illustration 153
A.2 YouTube class diagram 154
A.3 Flow description of our YouTube traffic generator 155

List of Tables

II.1 Standardized QCI characteristics 38

III.1 List of Symbols 50
III.2 Abbreviations of LTE/EPC procedures 57
III.3 Context and signaling events details of relevant LTE/EPC procedures . . . 58
III.4 RRC Signaling Overhead *(Source: [67])* 59
III.5 Scenario Parameters 62
III.6 Traffic Parameters 63
III.7 Scenarios 63
III.8 Mean Applications Data Rate 69

IV.1 Key Characteristics of 3GPP and IP-centric Approaches 82
IV.2 Simulation parameters 85
IV.3 Traffic distribution 86
IV.4 Web sites details 87
IV.5 YouTube video quality information 88
IV.6 Scenario parameters 92
IV.7 Performance analysis summary 96

V.1 Simulation parameters 113
V.2 Simulated Slo-Mo parameters 115

A.1 HTTP model parameters [109] 148
A.2 Behavioral model 148

A.3 Traffic distribution 149
A.4 VoIP model parameters (source: [107]) 149
A.5 YouTube video quality information 153

B.1 Simulation result of multiple ON/OFF processes 159

Résumé

Depuis quelques années le trafic de l'internet mobile ne cesse d'augmenter. Cette croissance soutenue est liée à plusieurs facteurs, parmi lesquels l'évolution des terminaux, la grande diversité des services et des applications disponibles et le déploiement des nouvelles technologies d'accès radio mobile (3G/4G). À cet égard, le standard 3GPP pour les réseaux LTE propose une architecture offrant une gestion fine de la QoS (par flux). Ce modèle, hérité des réseaux mobiles traditionnels orientés connexion, soulève des problèmes en termes de scalabilité, efficacité et performances.

Les travaux entrepris dans cette thèse ont pour objectif principal de proposer des solutions plus simples et moins coûteuses pour la gestion de la QoS dans les réseaux mobiles. À cette fin, à l'issue d'une étude et de l'évaluation de l'impact de la signalisation associée au modèle de QoS standard, deux modèles alternatifs ont été proposés. Nous proposons tout d'abord un modèle basée sur les mécanismes IP inspiré de l'approche DiffServ (par agrégat) largement étudié dans les réseaux IP fixes. Ce modèle fournit une gestion de la QoS simple, efficiente et rentable, tout en garantissant des performances équivalentes au modèle standard. Cependant, elle nécessite une remise à niveau de tous les eNB, et donc une longue phase de transition.

En conséquence, nous proposons SloMo qui vise à améliorer l'expérience des clients mobiles, mais avec un objectif de déploiement plus rapide. SloMo est une solution de gestion implicite de la QoS depuis un point unique situé sur le chemin des communications. SloMo exploite la dynamique instaurée par le mécanisme de contrôle de flux de TCP. Il vise à recréer un goulot d'étranglement dynamique dans un équipement contrôlé par l'opérateur lorsque les points de congestion réels ne sont pas accessibles. Une fois ce goulot d'étranglement déporté, il est alors aisé d'effectuer une gestion de la qualité IP classique dans l'équipement supportant Slo-Mo.

Mots-clés : Qualité de Service, Réseaux Mobiles, Conception de protocole, Analyse de performance, DiffServ, Active Queue Management, Protocole de transport TCP

Abstract

The mobile data landscape is changing rapidly and mobile operators are today facing the daunting challenge of providing cheap and valuable services to ever more demanding customers. As a consequence, cost reduction is actively sought by operators as well as Quality of Service (QoS) preservation.

Current 3GPP standards for LTE/EPC networks offer a fine tuning QoS (per-flow level), which inherits many characteristics of legacy telco networks. In spite of its good performance, such a QoS model reveals costly and cumbersome and finally, it remains very rarely deployed, thereby giving way to basic best-effort hegemony.

This thesis aims at improving QoS in mobile networks through cost-effective solutions; To this end, after an evaluation of the impact and cost of signaling associated with the standard QoS model, alternative schemes are proposed, such as the IP-centric QoS model (per aggregate) inspired from the DiffServ approach widely used in fixed IP networks. This model provides a simple, efficient and cost-effective IP level QoS management with a performance level similar to standardized solutions. However, as it requires enhancements in the eNB, this scheme cannot be expected in mobile networks before a rather long time.

Thus, we introduce Slo-Mo, which is a lightweight implicit mechanism for managing QoS from a distant point when the congestion point (e.g. eNB) is not able to do it. Slo-Mo creates a self-adaptive bottleneck which adjusts dynamically to the available resources taking advantage of TCP native flow control. Straightforward QoS management at IP level is then performed in the Slo-Mo node, leading to enhanced customer experience at a marginal cost and short term.

Keywords : Quality of Service, Mobile Networks, Protocol Design, Performance Analysis, DiffServ, Active Queue Management, TCP protocol

TELECOM Bretagne

n° d'ordre : 2016telb0406
Télécom Bretagne
Technopôle Brest-Iroise - CS 83818 - 29238 Brest Cedex 3
Tél : + 33(0) 29 00 11 11 - Fax : + 33(0) 29 00 10 00