Utility-based Resource Management for Future Mobile Communications Considering QoE

Vom Promotionsausschuss der
Technischen Universität Hamburg-Harburg
zur Erlangung des akademischen Grades
Doktor-Ingenieur (Dr.-Ing.)

genehmigte Dissertation

von
Ming Li

aus
Shandong, China

2016

Vorsitzender: Prof. Dr.-Ing. Gerhard Bauch
1. Gutachter: Prof. Dr.-Ing. Andreas Timm-Giel
2. Gutachter: Prof. Dr.-Ing. Christian Wietfeld

Tag der mündlichen Prüfung: 22.08.2016

Acknowledgement

This thesis was written during my research as a doctorate candidate at the Communication Networks Institute (ComNets) at the Hamburg University of Technology. It has been an incredibly rewarding experience that I got this opportunity to work alongside many great colleagues.

My deepest gratitude goes first and foremost to my supervisor Prof. Dr. Andreas Timm-Giel for his consistent and illuminating instruction. His guidance helped me in all the time of research and writing of this thesis. Without his constant encouragement and insightful comments, this work could not have reached its present form.

I would like to extend my sincerest thanks to Prof. Dr. Ulrich Killat for his constructive and inspiring ideas. I am deeply moved by his rigorous scientific attitude and dedicated spirit. Dr. Phuong Nga Tran had numerous discussions with me on convex optimization and gave me many critical ideas, for which I am extremely grateful. I am also very much appreciative of Dr. Lothar Kreft for his help in proof reading of the thesis, and for his encouragement and unwavering support in both my research work and personal life. I am deeply grateful for the insightful suggestions by Prof. Dr. Christian Wietfeld, who is my second examiner. His advice was a great help for me.

I would like to acknowledge all of my colleagues within the ComNets groups, of the Hamburg University of Technology and of the University of Bremen, for their support and help through these tough years. First, I would like to express my sincere gratitude to my project partners Dr. Xi Li, Dr. Umar Toseef, Dr. Thushara Weerawardane and Dr. Yasir Zaki who helped me overcome many challenges and difficulties side by side. Besides, I would also like to thank Dr. Maciej

Mühleisen for his support at both academic and personal level. Furthermore, I would extend the warmest thanks to Ilona Düring, Dr. Koojana Kuladinithi, Chunlei An, Dr. Yunqi Luo, Jonas Eymann, Christoph Petersen, Leo Krüger, Raphael Elsner, René Steinrücken, Frank Laue, Thomas Müller and Manuel Ponce-Ibarra for their great help. In addition, I would also like to thank my students Dimin Wang and Hüseyin K Tütüncüoglu for their good work.

Finally, I would like to thank my parents for their love and support throughout my life.

München, November 2016 *Ming Li*

Abstract

This work addresses the mobile networks, such as LTE (Long Term Evolution). The key objective is to improve the system performance in means of user satisfaction. Two main potential bottleneck links are identified, i.e. the Radio Interface and the transport network link from base station to core network.

Therefore, first a utility based radio scheduling algorithm is proposed. The scheduling algorithm, which maximizes the aggregated QoE, is proven to be optimal analytically. In a second step the resource limitation on the second bottleneck link, the S1 interface in the transport network is considered in addition. The mathematical problem is formulated considering the resource on both bottleneck link and maximizing the aggregated user satisfaction. The formulated problem is proven to be a convex optimization problem, and then solved using the Lagrangian relaxation method. In addition, computational advantageous heuristics are developed and compared in simulations against legacy approaches and the optimal solution.

Contents

List of Figures

List of Tables

Abbreviations

3GPP	The 3^{rd} Generation Partnership Project
ACK	Acknowledgement
AF	Assured Forwarding
AM	Acknowledgement Mode
AMBR	Aggregated Maximum Bit Rate
APN	Access Point Name
ARP	Address Resolution Protocol
ARQ	Automatic Repeat Request
AuC	Authentication Center
AWGN	Additive White Gaussian Noise
BE	Best Effort
CA	Carrier Aggregation
CAPEX	Capital Expenditure
CBR	Constant Bit Rate
CI	Confidence Interval
CoMP	Coordinated Multi-Point
C-RAN	Cloud Radio Access Network
CRC	Cyclic Redundancy Check
CQI	Channel Quality Indicator

D2D	Device-to-device
DiffServ	Differentiated Services
DL	Downlink
DRS	Dynamic Rate Shaping
DSL	Digital Subscriber Line
DVB	Digital Video Broadcasting
EDF	Earliest Deadline First
EF	Expedited Forwarding
eICIC	enhanced Inter-cell Interference Coordination
eNB/eNodeB	E-UTRAN Node B
EPS	Evolved Packet System
EPC	Evolved Packet Core
E-UTRAN	Evolved UMTS Terrestrial Radio Access Network
FRS	Fixed Rate Shaping
FSM	Finite-State Machine
FTP	File Transfer Protocol
GBR	Guaranteed Bit Rate
GGSN	GPRS Support Node
GPRS	General Packet Radio Service
GSA	Global mobile Suppliers Association
GSM	Global System for Mobile Communications
GTP	GPRS Tunnelling Protocol
HARQ	Hybrid Automatic Repeat Request
HLR	Home Location Register
HSPA	High Speed Packet Access
HSS	Home Subscriber Server
IAT	Inter-arrival Time
IMT	International Mobile Telecommunications

IP	Internet Protocol
ITU	International Telecommunication Union
ITU-R	ITU Radio Communication Sector
LTE	Long Term Evolution
LWDF	Largest Weighted Delay First
LWF	Longest Wait First
MAC	Medium Access Control
MBR	Maximum Bit Rate
MCS	Modulation and Coding Scheme
MIMO	Multiple Input Multiple Output
MME	Mobility Management Entity
MOS	Mean Opinion Score
MSS	Maximum Segment Size
MT	Mersenne Twister (Random Number Generator)
MTC	Machine-Type Communications
NAK	Negative Acknowledgement
NGMN	Next Generation Mobile Network
nRT	non Real-Time
OFDM	Orthogonal Frequency Division Multiplexing
OFDMA	OFDM Access
OPEX	Operating Expenses
PAPR	Peak-to-average Ratio
PCRF	Policy and Charging Control Function
PDCP	Packet Data Convergence Protocol
PDN-GW	Packet Data Network Gateway
PDU	Packet Data Unit
PHB	Per Hop Behavior
PHY	Physical Layer

PRB	Physical Resource Block
PS	Packet Switch
QoS	Quality of Service
QoE	Quality of Experience
RAN	Radio Access Network
RAU	Radio Access Unit
RD	Random Direction
RLC	Radio Link Control
RNC	Radio Network Controller
RNG	Random Number Generator
RRH	Remote Radio Head
RRM	Radio Resource Management
RRA	Radio Resource Allocation
RT	Real-Time
RTP	Real-Time Protocol
RWP	Random Way Point
SAE	System Architecture Evolution
SDU	Service Data Unit
SGSN	Serving GPRS Support Node
S-GW	Serving Gateway
SINR	Signal to Interference plus Noise Ratio
SON	Self-organizing Network
SC-FDMA	Single Carrier Frequency Division Multiplexing Access
SP	Strict Priority
TBS	Transport Block Size
TCP	Transport Protocol
TF	Transport Format
TTI	Transmission Time Interval

UDP	User Datagram Protocol
UE	User Equipment
UM	Un-acknowledgement Mode
UMTS	Universal Mobile Telecommunications System
UL	Uplink
VoIP	Voice over Internet Protocol
WFQ	Weighted Fair Queuing
WiMAX	Worldwide Interoperability for Microwave Access
WLAN	Wireless Local Area Network

Chapter 1

Introduction

Mobile communication continues to evolve at a lightning fast pace through endless innovations. In the last few decades, several generations of mobile communication have been developed that each generation brings significantly improved system performance, profoundly richer user experience, and dramatically reduced cost. In order to meet the ever growing mobile user demands and network loads, new revolutionary communication technologies and services are developed rapidly to keep up with this competition. The 3^{rd} Generation Partnership Project (3GPP) introduces Long Term Evolution (LTE) to ensure the competitiveness for the long term. In LTE (-Advanced), a new air interface and a new radio access network are introduced to provide significantly higher throughput and lower latency, greatly improved system capacity and coverage compared to those of the 3G systems. These improvements lead to increased users' satisfaction on the received services, known as *Quality of Experience* (QoE). However, the problem of spectrum scarcity will still exist in mobile networks. According to Cisco investigation published in 2015, the number of mobile-connected devices has already exceeded the number of people on earth, and by 2019 there will be nearly 1.5 mobile devices per capita [Cis15]. The

number of users and the traffic that they generate increase extremely fast, while the bandwidth resource stays very limited. Therefore, in order to guarantee a certain *Quality of Service* (QoS) and more importantly, to improve the users' perceived QoE, it is necessary to have an efficient resource management mechanisms.

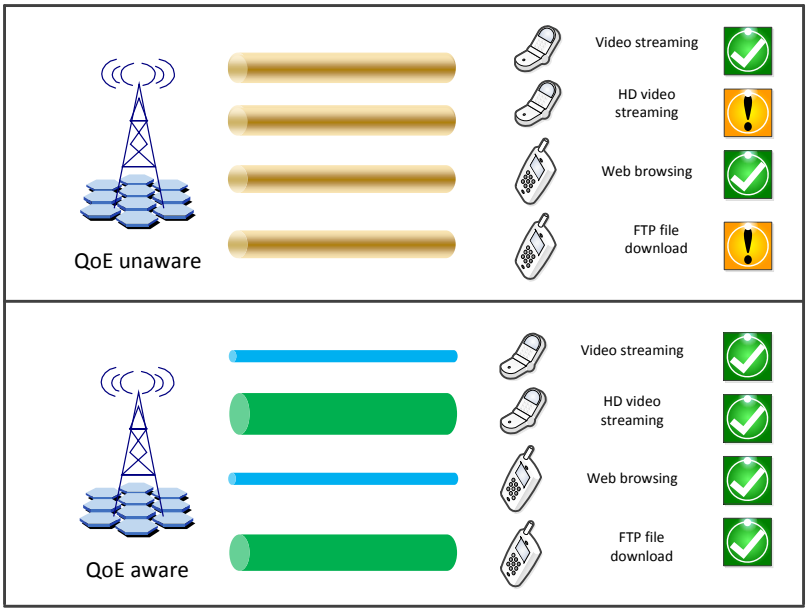


FIGURE 1.1: Example of the benefit by QoE based radio resource allocation

The key objective of radio resource allocation mechanisms is to make the best use of limited resources and to enhance the users' experience, under time varying channel conditions. The different scheduling disciplines regarding fairness, QoS provisioning and spectrum efficiency have been extensively researched. However, radio scheduling aims not only to improve the network performance, but also to increase the users' QoE. QoE is often measured by Mean Opinion Score (MOS) and can be

mathematically represented by a utility function. QoE, defined as "the overall acceptability of an application or service, as perceived subjectively by the end user" [IT07], is not a linear function of QoS (Quality of Service) parameters and very much depends on the application. For instance, for a user streaming high quality audio the difference in means of user satisfaction when served with 2 instead of 1 Mbit/s is low, if not negligible, whereas the user is obviously much more content when served with 100 instead of 50 Kbit/s.

One of the major innovations of this work is that a utility based radio resource scheduling framework with the focus of maximize the aggregated QoE in a cell is proposed. Figure 1.1 gives an example of the benefit by QoE based radio resource allocation. In this example, the video streaming and web browsing users need less amount of resources to get satisfied while the High-definition (HD) video and FTP (File Transfer Protocol) file downloading users need much higher resources. With the QoE aware scheduling, the radio resources are scheduled in a smarter way to the users who have the highest QoE gains. In other words, the resources are given to the users who are most demanding. For instance, if there are only few radio resources available, most of the resources will be given to the video streaming and web browsing users since they can achieve a high QoE with fewer resources compared to other users. On the other hand, if there are more resources available, more resources will be scheduled to the HD video and FTP users since the video streaming and web browsing users are already satisfied with less resources.

In this work, first some examples on how to build up the utility functions considering QoE are given. Then a utility based radio resource scheduling framework is proposed. The scheduling algorithm is proven to be optimal analytically. The performance of the scheduler is evaluated in several typical scenarios comparing against the well-known proportional fair scheduler.

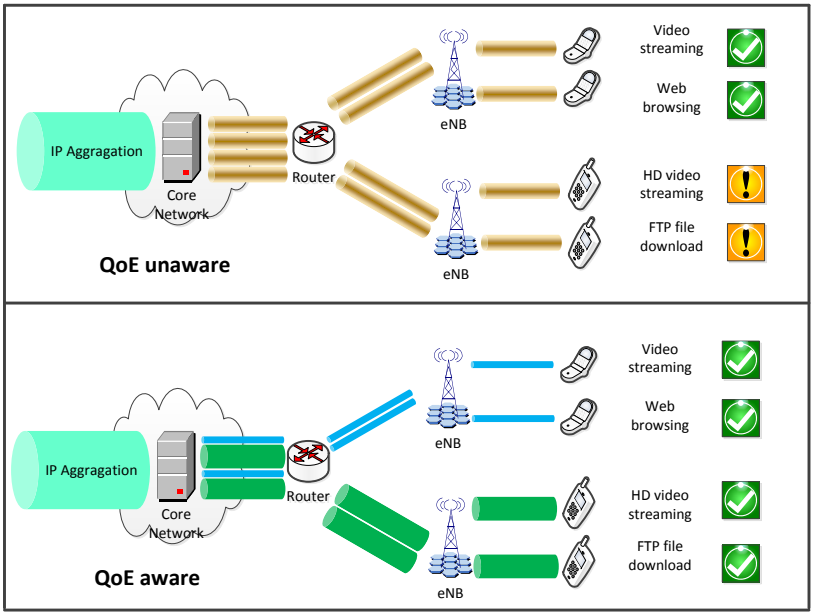


FIGURE 1.2: Example of the benefit by joint radio and transport optimized resource management

The radio scheduler can effectively utilize the radio resources. Nevertheless, LTE access network contains two main bottlenecks: air interface and transport backhaul. In legacy mobile networks, the transport backhaul is well dimensioned to cover the typical scenarios. However, the transport backhaul capacity is not been dimensioned to handle all the peak scenarios with the purpose of saving Capital expenditures (CAPEX) and operating expenses (OPEX). Therefore, the in case multiple cells experienced high loads at the same time, the transport backhaul could be overloaded as well. More importantly, the femtocell networks use cable or broadband xDSL as the last mile solution and usually its bandwidth is limited by the contract. Besides, Cloud-RAN (C-RAN) is a new cellular network architecture for the future mobile

network infrastructure [Ins11]. Since the baseband units (BBUs) are separated from the radio access units (RAUs, or referred as Remote Radio Heads, RRH) and moved to the cloud for centralized processing, the backhaul to the Core Networks is a challenge since it aggregates the traffic from over hundreds of cells.

A joint radio and transport optimized resource management scheme is designed considering both radio and transport network limitations which is the major focus of this work. The target is to optimize the resource allocation in terms of deciding optimum bearer rates for different users dynamically according to their application types as well as traffic variations and available network capacities of both radio and transport networks. The innovation of this work is an algorithm that dynamically optimizes the EPS (Evolved Packet Core) bearer rates for all users in the LTE networks considering the limitation of radio and transport resource with the objective of maximizing the accumulated users' QoE. Figure 1.2 gives an example of the benefit by the proposed joint radio and transport optimized resource management.

In this work, the EPS bearer shaping method as well as other existed resource management schemes at LTE transport networks are discussed first. Then, the typical scenarios which the proposed joint radio and transport optimized resource management schemes are applicable, e.g. femtocell cluster, base station with multiple cells and Cloud-RAN in LTE-advanced, are introduced. The resource allocation problem in multiple cells considering transport network limitations is formulated as a convex optimization problem, which maximizes the aggregated QoE, and it will be solved using the Lagrangian relaxation method. The EPS bearers are shaped according to their optimal rates for all users at the core network. The performance of the proposed algorithms is investigated and evaluated by simulations. Moreover, a discussion on how often to calculate and adjust the EPS bearer rates is given based on the complexity and performance investigations. At the end, radio scheduling heuristic methods with reduced complexity are proposed.

The performance is evaluated by simulations and compared with the joint radio and transport optimized resource management scheme.

Theoretically, the mathematical models proposed in this work are applicable to both LTE downlink and uplink. LTE downlink is focused in this work because most of the mobile traffic today is on downlink direction that data asymmetry between downlink and uplink on average can be more than 10:1 according to Nokia's investigation [Nok14].

1.1 Thesis Overview

In Chapter 2, LTE in general is introduced. The main motivation of LTE and its key features are explained. Then the overall system architecture, including both the radio access and core network of LTE is described in details according to the LTE standardization. The LTE femtocell, which is one of the typical scenarios focused in this work, is introduced. The LTE evolution roadmap to the future is discussed in the end.

The designed LTE simulator which is the first contribution of this work is described in Chapter 3. The importance of network simulation technique in communication networks is highlighted in the beginning. An introduction to OPNET modeler is given, which is a simulation tool used to build up the LTE simulator. The implemented network entities, channel model as well as the traffic models are explained stepwise in detail afterwards. This chapter concludes with the explanation of statistical evaluation methods used for performance evaluation.

In chapter 4, a utility-based radio scheduling framework in LTE considering QoE is proposed and evaluated, which is the second contribution. Firstly, the LTE scheduler and an overview of existing scheduling strategies are given. Besides, the state-of-art and the motivations of QoE based scheduling are introduced. The relationship between QoS and QoE is discussed afterwards, and how the QoE based utility

functions for various traffic are explained in details with several examples. Subsequently, a QoE based scheduler is proposed and analytically proven to be optimal. The scheduling procedures are illustrated step by step later. Finally, the performance gain of the proposed QoE based scheduler framework is compared against the conventional proportional fair scheduler by numerous simulations.

Chapter 5 proposes a joint radio and transport optimized resource management scheme, which is the most significant contribution of this work. This chapter first explains the legacy resource management schemes in core and transport networks. Various use cases which would be beneficial with a joint radio and transport resource management are introduced. Afterwards, an optimal QoE based dynamic rate shaping method is proposed. The joint resource management problem is formulated as a convex optimization problem and solved by the Langrangian relaxation method. Subsequently, the performance of this approach is evaluated by simulations. In the end, two heuristic methods with reduced complexity are proposed and compared against the optimal shaping method.

Chapter 6 gives the overall conclusion of the thesis with the highlights of all the main achievements. Finally, an outlook concerning future work concludes this thesis.

Chapter 2

Long Term Evolution

The roadmap of Next Generation Mobile Network (NGMN) is to provide mobile broadband services to end-users. To make this happen, 3GPP introduces Long Term Evolution (LTE) to ensure the competitiveness of the 3GPP technology family for the long term. In 2008, the ITU Radio communication Sector (ITU-R) of the International Telecommunication Union (ITU) defined the requirements of an International Mobile Telecommunications-Advanced (IMT-Advanced) compliant system which is marketed as 4G. LTE further evolved to LTE-Advanced to meet the ITU-R IMT-Advanced requirements. LTE introduces a new air interface and radio access network, which provide much higher throughput and low latency, greatly improved system capacity and coverage than the 3G systems. These improvements lead to increased expectations on the end-user quality of application services over LTE as compared to the existing 2G/3G systems. The improvements impose much higher requirements on the transport network, as well: (i) as the LTE system is designed for Packet Switched (PS) traffic, PS based data services are expected to dominate in LTE; (ii) the significantly improved throughput and capacity of the LTE air interface will put more load on the transport network; (iii) the higher expectations

on the responsiveness of interactive applications, and on voice quality increases the demand on lower packet delay in the transport network.

The Evolved Packet System (EPS) is composed by the Terrestrial Radio Access Network (E-UTRAN) which is the access part of the system, the Evolved Packet Core (EPC) which is the core part of the system, and mobile terminals. Nevertheless, LTE has become the colloquial name for the whole EPS system regularly used by 3GPP. LTE is pure IP based that means both real time and data services are served based on the IP protocol. LTE introduces a completely new radio interface and core network to its predecessor, providing substantially improved data performance. In order to support the new LTE radio interface, an EPC has been developed. The work on specifying the core network is known as System Architecture Evolution (SAE).

The main requirements of LTE included higher data rate, enhanced cell edge coverage, lower latency, lower complexity, improved system capacity, and spectrum flexibility. The initial LTE standard was finalized in December 2008 in 3GPP Release 8 [3GP09b], and the first publicly available LTE service was launched by TeliaSonera in Oslo and Stockholm on December 14, 2009. Within 5 years' time, 442 commercial LTE networks have been launched by operators in 147 countries by the end of 2015, according to data released by GSA (Global mobile Suppliers Association) which makes LTE the fastest developing mobile system ever [Ass15]. The LTE key features in Release 8 are summarized as follows [3GP09b]:

- Simplified flat all-IP based network architecture: the key benefits of flat IP architectures are lower cost, reduced system latency, and decoupled radio access and core network evolution.

- Significantly increased peak data rates: up to 100 Mbps in the downlink and 50 Mbps in the uplink in 20MHz spectrum. With the use of Multiple-input and Multiple-output (MIMO) and beam

forming techniques, even higher peak data rates can be achieved, e.g. 300 Mbps in the downlink and 75 Mbps in the uplink.

- Improved spectral efficiency and cell edge coverage: spectral efficiency in the downlink is targeted at 5 bps/Hz/cell and 2.5 bps/Hz/cell in the uplink, which is a four times improvement over its predecessor: High Speed Packet Access (HSPA).
- Reduced latency: both control plane and user plane latencies are reduced significantly, e.g. 5 ms user plane latency in upload condition which is up to 80% less than the delay in HSPA. These enhancements are especially beneficial to real-time services.
- Mobility: LTE provides high performance up to 120 km/h moving speed while maintaining services up to 350 km/h.
- Flexible radio planning: scalable usage of frequency spectrum from 1.25 MHz to 20 MHz.
- Efficient multiple access techniques: Orthogonal Frequency Division Multiplexing Access (OFDMA) and Single-carrier FDMA (SC-FDMA) as the multiple access techniques for downlink and uplink respectively.

2.1 LTE Multiple Access Schemes

OFDM has been adopted as the downlink transmission scheme for LTE downlink. OFDM is favored due to the fact that it is very robust against frequency selective fading channels, and therefore a low complexity of the equalization is needed to overcome frequency selectivity. It can significantly enhance the spectrum efficiency by adapting the modulation and coding scheme per sub-carrier according to the channel condition of each sub-carrier. In addition, OFDM is able to make use of different bandwidth by varying the number of OFDM sub-carriers used for transmission. OFDM is widely used in modern mobile and

wireless communication systems, e.g. Worldwide Interoperability for Microwave Access (WiMAX), Digital Video Broadcasting (DVB), and Wireless Local Area Network (WLAN).

2.1.1 OFDM and OFDMA

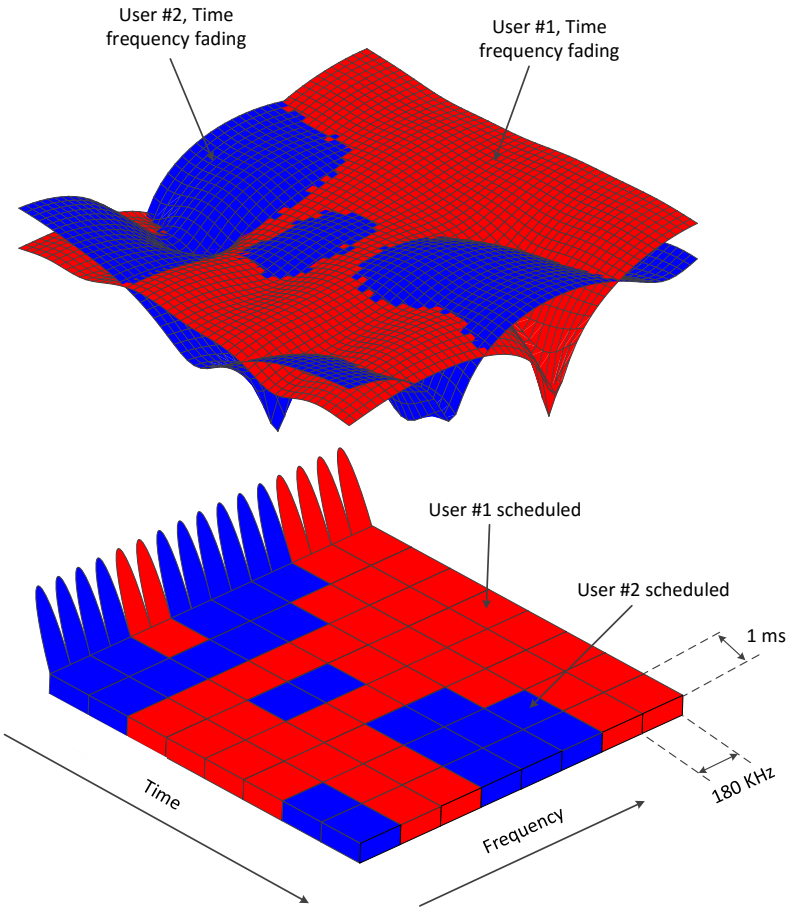


FIGURE 2.1: An example of channel dependent scheduling in LTE

Like FDM systems, in which signals from multiple transmitters are transmitted simultaneously over multiple sub-carriers, OFDM also uses multiple sub-carriers. In OFDM the sub-channels are orthogonal to each other. So they can be closely spaced to each other without causing interference. OFDM does not require guard intervals between the sub-channels and therefore can achieve a higher system spectral efficiency compared to conventional FDM systems. Besides, the OFDM sub-channels have narrow bandwidths and therefore OFDM is more resistant to frequency selective fading than single carrier systems. OFDM has a long symbol duration (low symbol/transmit rate) since a data stream is modulated into several sub-carriers before transmission. Therefore fading is slow enough for the channel to be considered as constant during one OFDM symbol interval [Dah07].

OFDM can be combined with multiple accesses with time, frequency separation of the users, allowing simultaneous transmissions of multiple mobile terminals. This scheme is referred as Orthogonal Frequency Division Multiple Access (OFDMA). In OFDMA, users are allocated a specific number of subcarriers for a predetermined amount of time. They are referred to as physical resource blocks (PRBs) in the LTE specifications which is the smallest granularity in LTE radio resource allocation. As shown in Figure 2.1, a PRB has a time duration of 1 ms and a frequency spacing of 180 KHz (12 subcarriers by 7 OFDM symbols). The LTE base station can make use of the different user channel conditions to allocate PRBs to the users periodically every 1 ms. Figure 2.1 gives an example of channel aware radio resource allocation in LTE downlink with 2 users. The upper part of the figure represents the user channel conditions. The lower part of the figure shows the resource along the time that the PRBs are allocated to the user with better channel condition along the time.

2.1.2 SC-FDMA

Single-carrier FDMA is a frequency-division multiple access scheme that is adopted for LTE uplink access and transmission. SC-FDMA has an additional DFT (Discrete Fourier Transform) procession step prior to the conventional OFDMA procession, and therefore can be seen as a special kind of OFDMA. Single-carrier is used in LTE uplink instead of multiple carriers OFDMA in LTE downlink. Comparing to OFDMA, SC-FDMA provides a lower peak-to-average ratio (PAPR) that can improve the transmission efficiency and reduce the cost of the amplifier and the power consumption at the same time, making it a very attractive characteristic for uplink transmission by mobile terminals.

2.2 LTE Overall System Architecture

Figure 2.2 shows the overall high-level network architecture of LTE, which is also referred as EPS. It is comprised of three main components: User Equipment (UE), Evolved UMTS Terrestrial Radio Access Network, and Evolved Packet Core. The E-UTRAN is the access part of the network which is responsible for radio-related functionalities including radio resource allocation and scheduling, coding and transmission. The EPC provides the core network functions such as mobility, policy, and security management. It is based on the Internet Protocol (IP) that enables communication to 3GPP radio access (LTE, 3G and 2G), non-3GPP radio access (WLAN and WiMAX), and fixed networks (Ethernet and fiber).

EPS is a connection-oriented transmission network. A bearer is defined to provide a "virtual" connection between two endpoints, e.g. a UE and a PDN-GW (Packet Data Network Gateway) with specific QoS attributes. Both EPC and E-UTRAN provide services according to the QoS parameters associated to the bearer. In LTE multiple bearers can be established for a user with multiple services, e.g., Voice over Internet

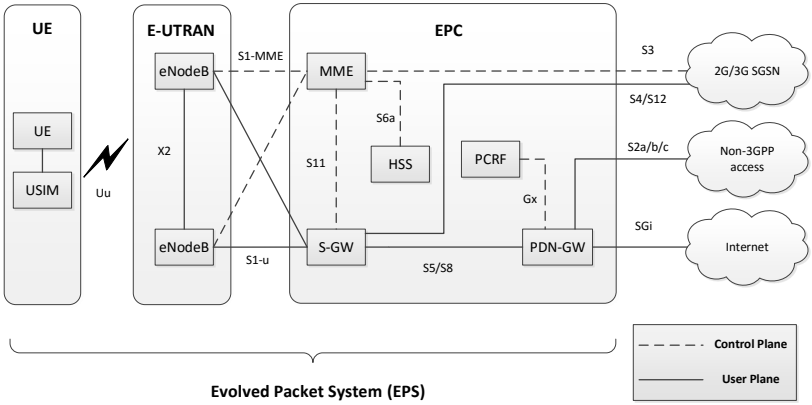


FIGURE 2.2: LTE system architecture

Protocol (VoIP), and at the same time downloading a file using File Transfer Protocol (FTP).

2.3 Core Network

The key component entities of EPC include Mobility Management Entity (MME), Serving Gateway (S-GW), and PDN Gateway (PDN-GW). Besides, there are some logical entities like Policy and Charging Control Function (PCRF), and Home Subscriber Server (HSS) [3GP15b].

- **MME**: It is a control plane node providing signaling and various controlling functions including: authentication, security, roaming, default/dedicated bearer establishment, tracking user mobility and handover.
- **S-GW**: It routes and forwards user data packets between the base station and the PDN gateway. It is the local mobility anchor point for inter-eNodeB handover, as well as, the mobility

anchoring for inter-3GPP mobility. In addition, it manages and stores UE contexts, e.g. parameters of the IP bearer service, network internal routing information.

- **PDN-GW**: Similar to GPRS support node (GGSN) and the serving GPRS support node (SGSN) with UMTS (Universal Mobile Telecommunications System) and GSM (Global System for Mobile communication), PDN-GW provides connectivity to external packet data networks. A UE may be served by more than one PDN-GWs simultaneously for accessing multiple PDNs that each packet data network is identified by an access point name (APN).

- **HSS**: It contains user related information supporting functionalities such as mobility management, call establishment and user authentication. The HSS is composed by the Home Location Register (HLR) and the Authentication Center (AuC).

- **PCRF**: It is responsible for enforcing the charging policy. In addition, it provides the QoS authorization that specifies how a certain data flow will be handled in accordance with the user's subscription.

2.4 Radio Access Network

The access network of LTE, named as E-UTRAN, contains only interconnected Evolved Node Bs (base station in LTE, abbreviated as eNodeB or eNB). Unlike the former UMTS/HSPA systems, there is no centralized Radio Network Controller (RNC) in E-UTRAN, and therefore LTE is been said to have a flat architecture. The flat architecture greatly reduces the handover latency and avoids the system vulnerable against RNC failures. The eNodeBs are connected with each other via X2 interface supporting functionalities such as the intra-LTE handover

and the inter-eNB coordination. An eNodeB is connected to the EPC via the S1 interface, and the UE via the Uu interface.

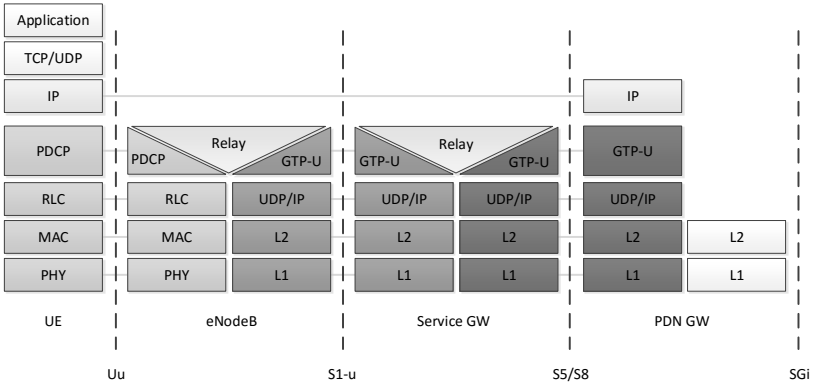


FIGURE 2.3: LTE user-plane protocol stack

The LTE protocols consist a user plane and a control plane. As the name suggests, the control plane protocols are responsible for establishing and controlling user connections, while the user plane protocols are used for user data transfer. Figure 2.3 shows the LTE user plane protocol stack over UE, E-UTRAN and EPC. The eNodeB acts as a bridge between the UE and the EPC that it can send and receive data both from and to the UE and the EPC. On one hand, it sends the user data securely to the EPC based on the GTP tunneling protocol, which is on top of the UDP/IP protocols, and vice versa. On the other hand, it manages the traffic flows between the UE and the eNodeB by scheduling the frequency spectrum resources in both downlink and uplink directions through a set of layer 1 and 2 protocols, which are explained in details below. Figure 2.4 further gives a detailed overview of the air interface protocol architecture and corresponding functions for the downlink [DS07].

- ***Packet Data Convergence Protocol*** **(PDCP)**: There is one PDCP entity per SAE bearer configured for a user. It performs

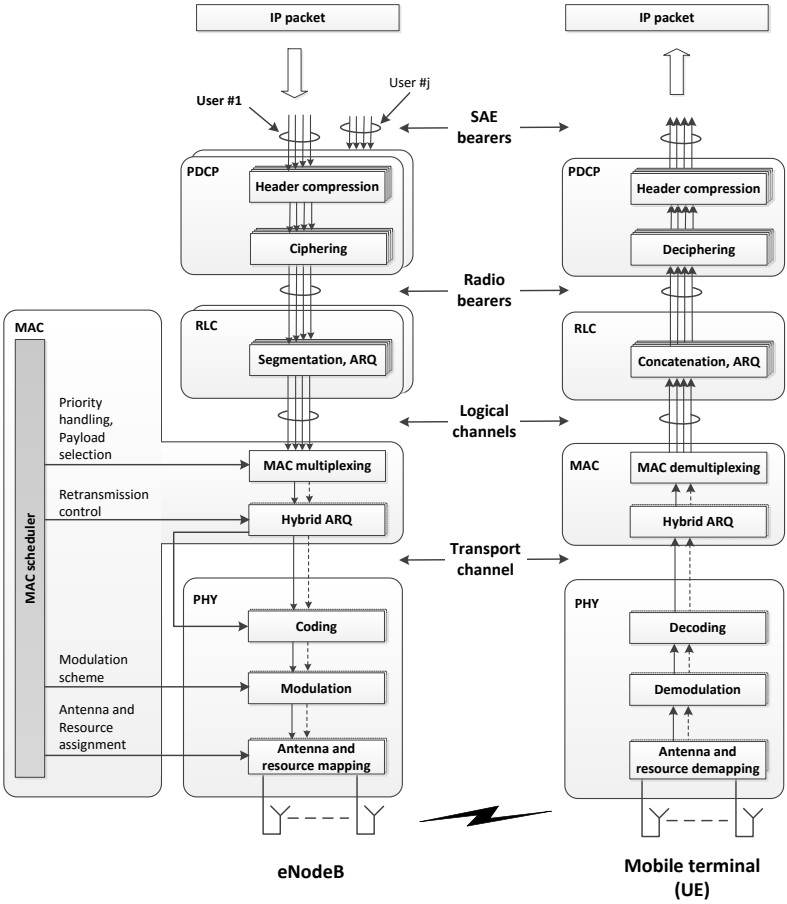


FIGURE 2.4: LTE protocol architecture in downlink, based on [DS07]

the IP header compression (and decompression) based on the ROCH (Robust Header Compression) standard in order to improve the efficiency of the radio interface. Besides, PDCP is responsible for ciphering and integrity protection. At the receiver side, the corresponding deciphering and decompression are performed. A detailed description of the PDCP functionality can be found in [3GP10b].

- ***Radio Link Control*** (**RLC**): It provides services to the PDCP layer in the form of radio bearer. There is one RLC entity per radio bearer configured for a user. RLC performs segmentation, concatenation and reassembling of the packets. It takes care of in-sequence packet delivery to the PDCP layer. Besides, it also performs error handling, based on the well-known Automatic Repeat Request (ARQ) methods.

  There are three operation modes supported in the RLC layer: Acknowledgement Mode (AM), Un-acknowledgement Mode (UM) and Transparent Mode (TM). AM is used to provide error-free and in-order transmission so that it is suitable for TCP based services, such as FTP and HTTP. UM provides segmentation and concatenation functionalities, but it does not support retransmission. Therefore UM is suitable for UDP based applications that can tolerate some losses, like VoIP. As the name suggested, Transparent Mode does not add any overhead to the data. It can be used for random access. More information can be found in [3GP10c].

- ***Medium Access Control*** (**MAC**): It provides services to the RLC in the form of logical channels. The major function of the MAC is radio resource allocation and scheduling of radio resources in both uplink and downlink considering users' QoS over the air interface. There is only one MAC entity per eNodeB, responsible for both downlink and uplink scheduling. In addition, it supports

Hybrid Automatic Repeat Request (HARQ) retransmissions for improved reliability[3GP10a].

- ***Physical Layer*** (**PHY**): It provides its services to the MAC in the form of transport channels. It handles physical layer functions including coding/decoding, modulation/demodulation, multiple antenna mapping etc.

Figure 2.5 gives an example of LTE data flow with three IP packets through all the protocol layers over the air interface in downlink direction [Dah07]. The header compression and ciphering is been performed at the PDCP layer. A PDCP header is added to each IP packet and sent to the RLC layer. The RLC layer then performs concatenation and/or segmentation and adds an RLC header for each RLC PDU (Packet Data Unit). The header information is used for retransmissions and in-order delivery in the UE. The MAC layer assembles a number of RLC PDUs into a MAC SDU (Service Data Unit), and adds a header to form a transport block. The size of a transport block depends on the user channel condition and the number of PRBs allocated to the UE. At the physical layer, a CRC (Cyclic Redundancy Check) is added to the transport block for error-detection. The PHY layer then transmits the modulated signal over the air interface.

2.4.1 Medium Access Control

The MAC layer of the eNodeB, which is the lowest sub-layer of layer 2, has the most important function in Radio Resource Management (RRM). It is responsible for radio resource allocation and packet scheduling for both downlink and uplink. It provides services to the RLC layer via logical channels, and connects to the PHY layer via transport channels. The MAC layer performs multiplexing and demultiplexing of the logical channels, and maps the logical channels to the appropriate transport channels. Besides, the MAC layer supports HARQ with soft combining, providing robustness against transmission errors.

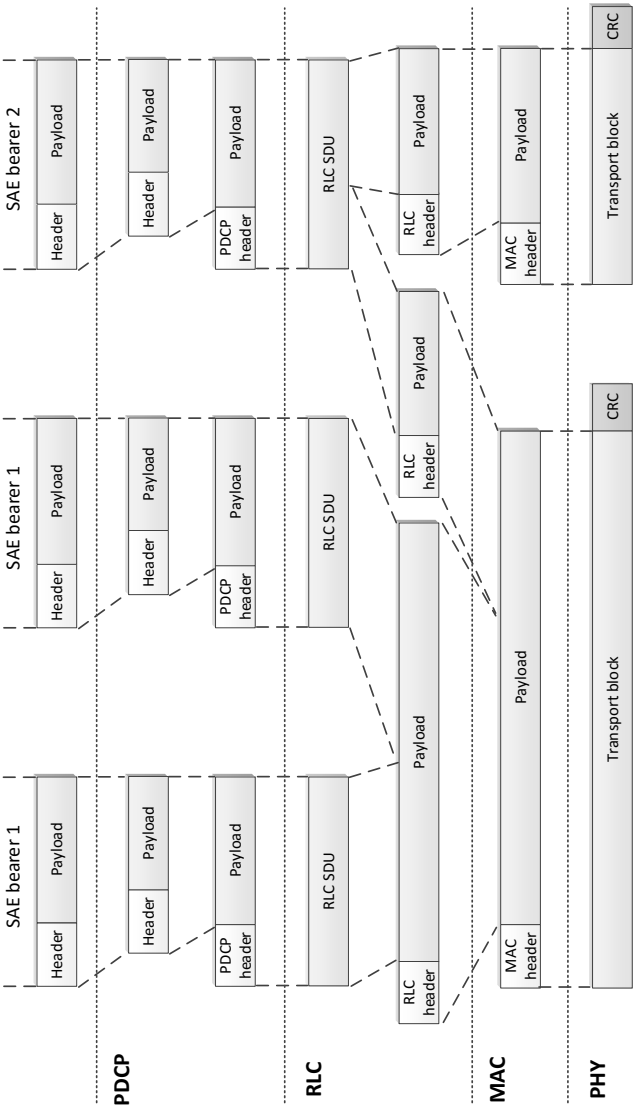


FIGURE 2.5: Example of LTE data flow, based on [DS07]

Logical Channels and Transport Channels

The MAC provides services to the RLC in the form of logical channels, which are generally classified into control channels, used for transmission of control and configuration information, and traffic channels, used for user data transmission.

TABLE 2.1: A list of logical channels and transport channels [3GP10a]

Logical channels	**Transport channels**
Broadcast Control Channel (BCCH)	Broadcast Channel (BCH)
Paging Control Channel (PCCH)	Paging Channel (PCH)
Dedicated Control Channel (DCCH)	Downlink Shared Channel (DL-SCH)
Multicast Control Channel (MCCH)	Multicast Channel (MCH)
Common Control Channel (CCCH)	Uplink Shared Channel (UL-SCH)
Multicast Traffic Channel (MTCH)	Random Access Channel (RACH)
Random Access Channel (RACH)	

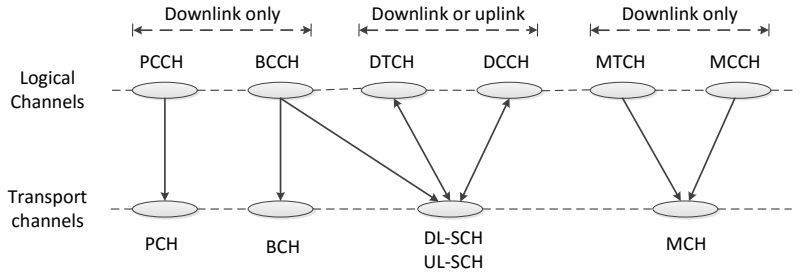


FIGURE 2.6: Example of mapping of logical channels to transport channels

In the meantime, the PHY offers services to MAC in the form of transport channels, which specifies how and with what characteristics the information is transmitted over the radio interface. There

is a Transport Format (TF) that contains the information about the transport block size, the modulation scheme, and the antenna mapping associated with a transport block.

The logical channels and control channels are listed in Table 2.1 according to 3GPP standards [3GP10a]. Figure 2.6 gives an example of mapping between the logical channels and the transport channels. For a detailed description of the logical and transport channels and the mapping between them, please refer to [Dah07].

Packet Scheduling

The MAC is responsible for the allocation of the time-frequency resources to one or multiple terminals for both uplink and downlink. As stated earlier, the minimum granularity of the time-frequency resource that the MAC can handle is a PRB, which has a time duration of 1 ms and a frequency spacing of 180 KHz. In every Transmission Time Interval (TTI), which is 1 ms, the MAC decides the allocation of PRBs to one or more UEs, as well as the corresponding transport block size, the modulation and coding scheme, and the antenna mapping (in case of multi-antenna transmission).

Because 3GPP does not specify any scheduling strategy, the radio resource allocation and packet scheduling has been widely researched with the key objective to make the best use of limited resources under time varying channel conditions. The user channel conditions are reported to the MAC by the Channel Quality Indicator (CQI), and therefore the MAC can decide how to make the best use of the radio resources. Figure 2.7 gives an example of the channel aware radio resource allocation in LTE downlink with 2 users. It shows that the radio resource allocation in LTE can exploit channel variations in both frequency and time domains. In addition to the channel quality, a well-designed scheduler should also take the QoS requirements as well as the buffer status and priorities into account. For instance, the real-time services such as VoIP, need a low delay while the best effort services are more sensitive to the data rate.

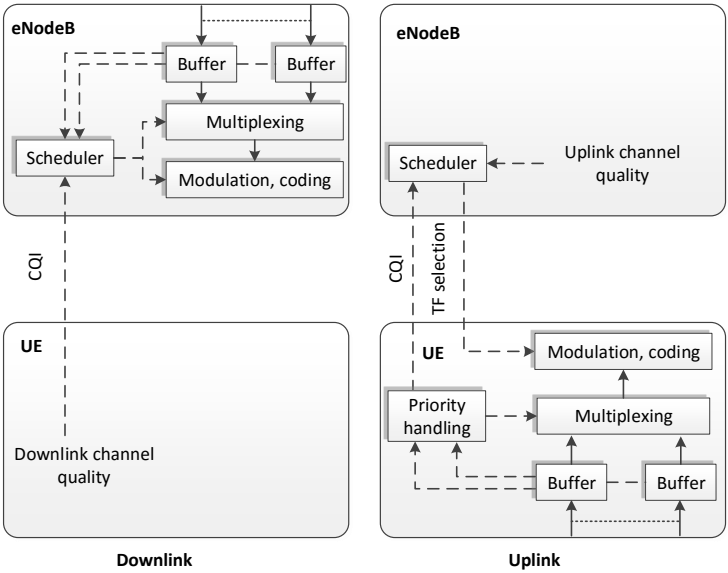


FIGURE 2.7: An overview of the downlink and uplink scheduling, based on [DS07]

The basic function of the uplink scheduler is similar to the downlink. Nevertheless, the uplink transmission can only use single carrier, which has an additional requirement that only continuous PRBs can be allocated to one UE. Besides, unlike the scheduling decision is per radio bearer based on downlink, the uplink scheduling decision is per UE based. The UE then decides which radio bearer(s) should be served. An overview of downlink and uplink scheduling mechanisms can be seen in Figure 2.7. Chapter 4 summarizes and compares the most widely used scheduling strategies.

Hybrid ARQ

Similar to the HSPA system, LTE uses the HARQ protocol with soft combining to provide robustness again transmission errors. Figure

2.8 gives an example of a HARQ process in LTE that multiple parallel stop-and-wait processes are used in the HARQ protocol [Dah07]. When a transport block is received, the receiver tries to decode it and informs the transmitter about the outcome of the decoding operation that whether a retransmission is needed through the ACK/NAK (Acknowledgement/Negative Acknowledgement) flag. With the use of multiple parallel stop-and-wait processes, the data can be delivered to the RLC out-of-sequence. For instance, the transport block 5 was successfully decoded before transport block 3, since it is being delayed by a retransmission. In LTE, the RLC is responsible for the packet reordering in the case of AM and UM transmission modes.

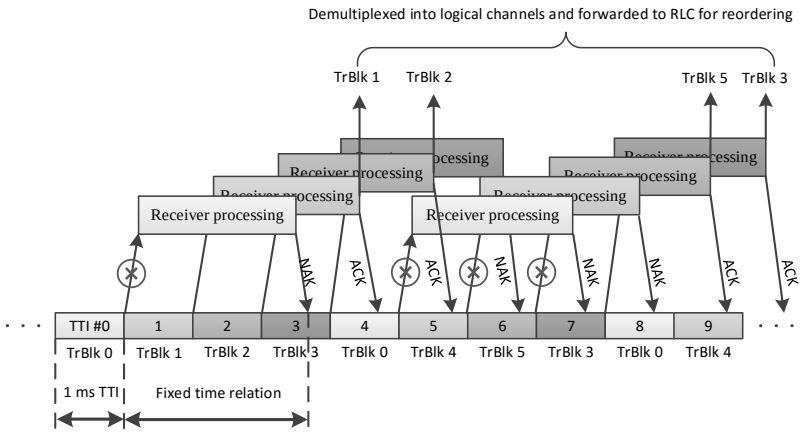


FIGURE 2.8: Example of a HARQ process in LTE, based on [DS07]

Although the HARQ can recover most of the transmission errors, some residual transmission errors may still exist and therefore it may fail to deliver error free data blocks to the RLC. For example, a NAK is incorrectly interpreted as an ACK by the transmitter in the order of 1% [Dah07]. The packet errors will trigger TCP congestion avoidance resulting a reduction of the TCP performance. In RLC AM, these

errors are further detected and recovered by the RLC retransmissions to deliver error-free data to higher layers.

2.5 LTE Femtocell

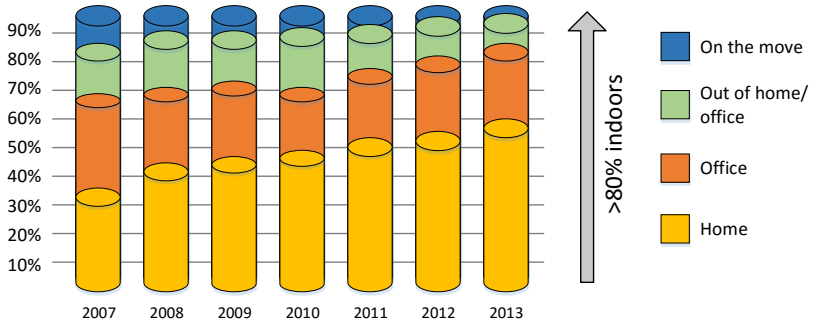


FIGURE 2.9: Mobile data traffic's phenomenal growth indoors [Nok11]

According to Cisco investigation [Cis15], global mobile data traffic will increase nearly tenfold between 2014 and 2019. Mobile data traffic will grow at a compound annual growth rate (CAGR) of 57% from 2014 to 2019, reaching 24.3 exabytes per month by 2019. It is an acknowledged fact that most of the mobile data is generated in the indoor environments such as homes, offices, shopping malls, hotels, and indoor venues. Figure 2.9 shows that the indoor traffic dominates today's mobile data traffic by more than 80%. However since the radio signals are seriously attenuated by the walls and other obstacles, the indoor performance is significantly poorer than outdoor. The latest trend to solve this problem is to introduce a large number of small cells, which can be connected to the operator's network by using existing user broadband (e.g. xDSL/Cable/Fiber) connection, for indoor coverage.

TABLE 2.2: Comparison of femto-, pico-, micro- and macrocell

	Femto	Pico	Micro/metro	Macro
Indoor/Outdoor	Indoor	Indoor or outdoor	Outdoor	Outdoor
Number of users	4 to 16	32 to 100	200	200 to 1000+
Max. tx power	20 to 100 mW	250 mW	2 to 10 W	40 to 100 W
Max. cell radius	10 to 50 m	200 m	2 km	10 to 40 km
Bandwidth	10 MHz	20 MHz	20, 40 MHz	100 MHz
Backhaul	DSL, cable, fiber	Microwave	Fiber, microwave	Fiber, microwave

The macrocells are traditional base stations with high transmission power and wide coverage up to tens of kilometers. Microcells, with a smaller size and reduced transmission power, are installed mainly in urban areas to offload the macrocells traffic. Even smaller picocells are deployed in areas with very dense phone usage, such as office buildings and shopping malls.

Table 2.2 gives a comparison of the different types of cells. As the name suggested, a femtocell is a very low-power cellular base station with small coverage, typically designed for use in a home or small business in the indoor environments only. Unlike the other three types of cells that are usually installed and maintained by the mobile network operators, the femtocells are usually installed by the end user in their homes or offices, and therefore the femtocells are designed to be autonomous. Cisco estimates that by 2019, 67% of mobile data traffic will be offloaded from the macro network [Cis14].

Figure 2.10 shows the femtocell system architecture along with the legacy LTE macrocell system. A femtocell base station is often been called Home eNodeB (HeNB). The femtocell stations have the same functions as macrocell stations and can provide such as radio resource management functions. Unlike the macrocell stations which directly

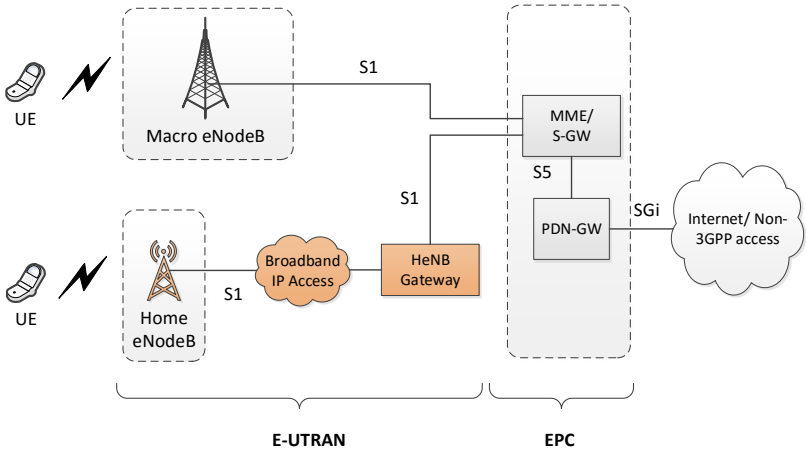


FIGURE 2.10: LTE femtocell system architecture [3GP09a]

connect to the EPC, the femtocell stations are connected using broadband IP, such as DSL or cable modems, to the EPC via a femtocell gateway. A femtocell gateway is composed by a security gateway that may connect to hundreds of femtocells, and a signaling gateway which aggregates, validates and authenticates traffic of the individual femtocell.

2.6 LTE Evolution Roadmap

3GPP initiated the investigation on LTE since 2004 and specified the first release of LTE (Rel-8) in March 2009. In 2008, the ITU Radio communication Sector (ITU-R) of the International Telecommunication Union (ITU) defined the requirements of an International Mobile Telecommunications-Advanced (IMT-Advanced) compliant system which is marketed as 4G. LTE further evolved to LTE-Advanced to meet the ITU-R IMT-Advanced requirements from the Rel-10. One of ultimate goals of the long term evolution is to provide one thousand

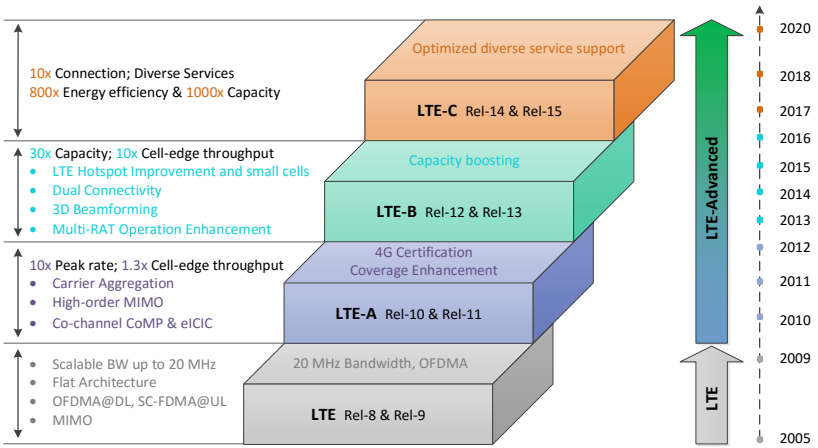


FIGURE 2.11: LTE evolution roadmap [Hua13]

times of capacity as its predecessor. LTE keeps evolving by introducing new features, supporting more services, by multiple market phases: LTE-A, LTE-B and LTC-C and more in the future, as shown in Figure 2.11 [Hua13].

In the initial release, Rel-8 defined the basis of LTE technologies, including a flat network architecture, exploiting OFDMA and SC-FDMA as donwlink (DL) and uplink (UL) multiple access technology, supporting a scalable bandwidth up to 20 MHz. Rel-9 introduced more features as dual layer beamforming, positioning, Self-organizing Network (SON).

From 3GPP Rel-10 onwards, LTE evolves to LTE-Advanced to meet and exceed the IMT-Advanced system requirements. LTE-A (Rel-10/Rel-11) supports wider bandwidths with carrier aggregation up to 100 MHz and higher-order spatial multiplexing with up to 8x8 MIMO in DL and 4x4 MIMO in UL. The cell peak rate has been boosted more than ten times up to 3 Gbps on downlink and 1.5 Gbps on uplink compared with Rel-8. Heterogeneous Networks (HetNets) containing

a variety of cell sizes, e.g. femtocell, picocell, as well as relay nodes are studied to improve the coverage. Coordinated Multi-Point (CoMP) and enhanced inter-cell interference coordination (eICIC) are the two key features in Rel-11 to improve performance especially at the edge of cells. eICIC is an interference control technology to prevent inter-cell interference by assigning different subframes (different time ranges) for different users in the cell egde. CoMP is a new technology based on network MIMO. By coordination of multiple sites, a user at the cell edge is able to receive and combine signals from multiple cell sites. And the user's transmission can be received and processed by multiple sites.

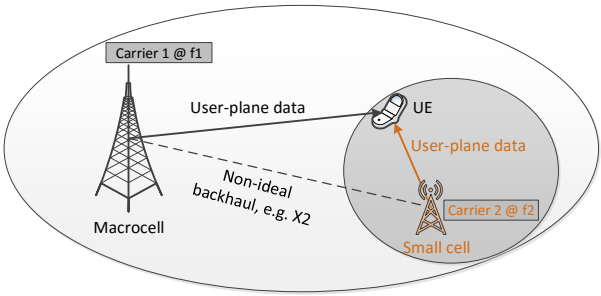


FIGURE 2.12: Example scenario of dual connectivity

LTE-B (Rel-12/Rel-13), which is currently under development, is the second phase of LTE-Advanced, targeting more than 30 fold increase of capacity and 10 fold increase of cell edge performance. LTE-B introduces LTE-Hi (LTE Hotspot improvement) to optimize the hotspot and indoor transmission of small cells. LTE-B improves the multi-antenna performance with 3D-beamforming which can further boost the performance of both SU-MIMO (Single-user MIMO) and MU-MIMO (Multi-user MIMO). Besides, LTE-B is designed to support better Multi-RAT (Multi-Radio Access Technology) interworking between LTE and WiFi, GSM and UMTS, providing better mobility

management, resource allocation and traffic offloading. In addition, dual connectivity, which is also known as the inter-site carrier aggregation, can split the data transmission over multiple base stations to improve the user throughput as seen in Figure 2.12.

LTE-C in the future is supposed to provide 1000 fold of capacity and 800 fold of energy efficiency, and support more users with diverse services. Some preliminary studies have been made, e.g. LTE in unlicensed spectrum, enhancements of CA, Machine-Type Communications (MTC), Device-to-device (D2D) communications, Full-Dimension MIMO, Indoor positioning, etc. With more spectrum, more small cells and higher transmission efficiency, the well-known 1000x capacity challenge can be conquered in near future [Hua13].

In addition, 3GPP has initiated preliminary studies on a new generation of 5G standards, which may be introduced in the early 2020s. Three distinct 5G network visions had emerged including a super-efficient mobile network, a super-fast mobile network and a converged fiber-wireless network [GSM15]. A 5G workshop (5G Workshop - The Start of Something) was held by 3GPP in September 2015. Three high level use cases are addressed in the workshop: enhanced mobile broadband, massive machine type communications and ultra-reliable and low Latency Communications [3GP15a].

Chapter 3

LTE Simulation Method

3.1 Introduction

In an era of information explosion, in order to meet the ever growing mobile user demands and network loads, new revolutionary communication technologies and services are developed rapidly to keep up with this competition. The increased complexity of communication protocols, large scale networks and coexistence of multiple communication standards put great challenges for the network researchers and developers. No matter deploying a new network, upgrading the current the network or testing new protocols, assessments on the network performance and reliability must be performed in order to reduce the investment risk of network construction and improve the network performance. For instance, the network engineers need to optimize the performance in protocol developments and the operators need to assure that the networks are designed properly to meet the demands.

Various methods, including mathematical modeling and analysis, prototype implementation, and network modeling and simulation, are

used by network engineers. Mathematical analysis has a lower complexity and is usually faster than the network simulation and the prototype method. Nevertheless, it is usually not applicable for end-to-end large scale networks since the network component behavior or performance is dependent on the remaining parts of the network. Therefore the Kleinrock's independence assumption is violated [Kle75], since the end-to-end network cannot be decomposed and analyzed hop-by-hop independently. The prototype implementation method can offer a straightforward insight on the system performance in the real world. However, it is very complicated and costly to set up the test-bed including the hardware assembling and protocol implementation, and so on. Owing to this fact, this method is only applicable to small scale networks. In addition, the prototype implementation is sometime not possible for the future technologies that not yet available.

Network simulation, which is a technique that the network behavior is modeled in a program by modeling the network entities and their interactions based on some mathematical models, or playing back the observations captured from a real network, can be used to study large scale networks with high accuracy. Unlike the prototype method that is very time consuming and costly setting up a test bed with multiple servers, routers and links, etc., network simulation allows engineers and researchers to test the network behavior in a more efficient way. Furthermore, network simulation can easily reproduce some networking phenomena, such as wireless radio interference.

Therefore, the network simulation is adopted for the investigations in this work due to the above mentioned merits. There are many popular network simulators available, such as OPNET Modeler (or Riverbed Modeler, OPNET was purchased by Riverbed) [Riv16], NS [NS16], openWNS [OPE15], OMNeT++ [OMN16], NetSim [Net16], etc. Among them, OPNET Modeler is one of the most successful simulators with high reputations, and widely used by many operators, vendors, infrastructure providers. Besides, it is widely used in military,

education, banking, insurance areas as well. It provides a rich model library, including many protocols and vendor nodes which are well documented. The OPNET Modeler is therefore adopted in this work to design, implement and evaluate the LTE simulation model.

The rest of this chapter is organized as follows: Section 3.2 gives a general introduction to OPNET Modeler. The simulation framework and implementation details are discussed in Section 3.3. The user traffic models are explained afterwards in Section 3.3.5, and the statistical analysis method on simulation results processing is illustrated in Section 3.4.

3.2 OPNET Modeler

The OPNET Modeler was initially created by two PhDs in MIT (Massachusetts Institute of Technology) in 1986 and became commercialized one year later. So far it has thousands of clients, covering a variety of companies, banks and governments in many areas. For example, the infrastructure companies like Cisco, Nokia, and operators as AT&T are using OPNET for researches, simulations and tests. OPNET Modeler was purchased by Riverbed Technology, which is a leading company in application performance infrastructure and has 26,000+ customers include 97% of the Fortune 100 and 98% of the Forbes Global 100 with around 1 billion dollars in annual revenue [Riv16]. OPNET Modeler was renamed as Riverbed Modeler since version 18.0. The simulator used in this work is based on the OPNET Modeler version 17.5, however the developed model is compatible with the newer versions.

An overview of the OPNET Modeler is shown in Figure 3.1. The OPNET Modeler offers many advanced features including a powerful and user friendly graphical interface, a hierarchical and objected oriented approach, comprehensive tools for result analysis and a build-in event tracing debugging tool (OPNET Debugger (ODB)). The OPNET

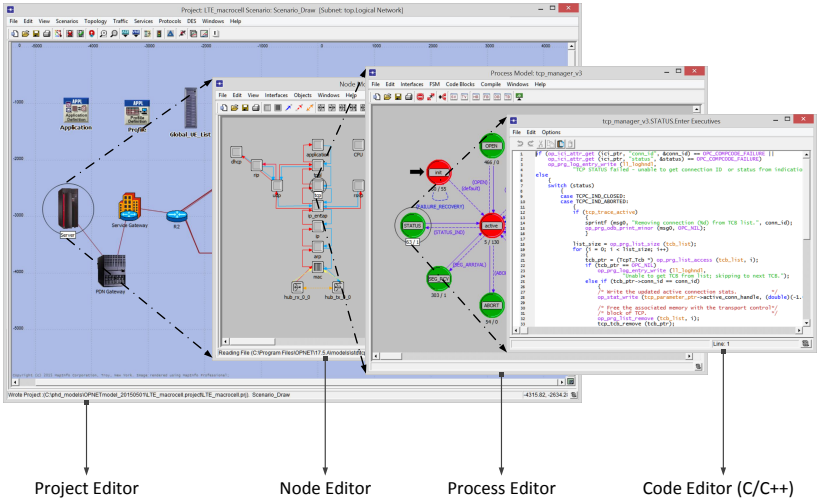


FIGURE 3.1: OPNET Modeler editor overview

Modeler is able to perform end to end network architecture design, system level simulation, protocol development and optimization, network application optimization and deployment analysis, etc.

The OPNET Modeler uses hierarchical network modeling. From the protocol point of view, it follows the Open Systems Interconnect (OSI) architecture, with a top to bottom protocol structure from application layer, TCP layer, IP layer, ARP layer, MAC layer to Physical layer. From the network topology point of view, it provides a 3-layer modeling mechanism: the top layer shows the whole network layout seen in the project editor; the middle layer is the node layer with a layer based model seen in the node editor, reflecting the device/router/server behavior; the bottom layer is the process layer, modeled by a Finite-State Machine (FSM) seen in the process editor. The detailed coding for each state of a FSM can be edited in the code editor. The 3-layer models correspond to the networks, devices, and protocols from top to bottom, which reflects the construction of a real network.

The object-oriented modeling is used in the OPNET Modeler. The same class of nodes can use a same node model with different configurations. For instance, deploying multiple eNBs in a scenario can use a same eNB model, but these eNBs can be configured with different bandwidth, location, transmission power and other parameters. The detailed modeling of protocols is based on the FSM which is driven by discrete events. The FSM does not have to check whether to change to another state from time to time, but only gives a response until a packet arrived or other events. Therefore, compared to time driven simulations, this improves the computation efficiency dramatically.

In addition, the OPNET Modeler has a rich library for creating communication links, packet formats, and interface control information and so on. And to be noticed, all the source codes and libraries in OPNET Modelers are well documented. For further information of OPNET Modeler, please refer to [Riv16].

3.3 Simulator Design

3.3.1 LTE Reference Model and Scenario

The objective of this work is to investigate the utility based resource management schemes in LTE considering QoE and evaluate the proposed framework by simulation. QoE is defined as "the overall acceptability of an application or service, as perceived subjectively by the end user" [IT07] and it depends on the end-to-end system performance. LTE is composed by two components: the access network and the core network. Both are responsible for resource management, and therefore both have important impact on the system performance. For example, in the LTE core network, the access gateways control the user data rate by traffic shaping. In the access network, the radio resource management is mainly done by radio scheduling which is one of

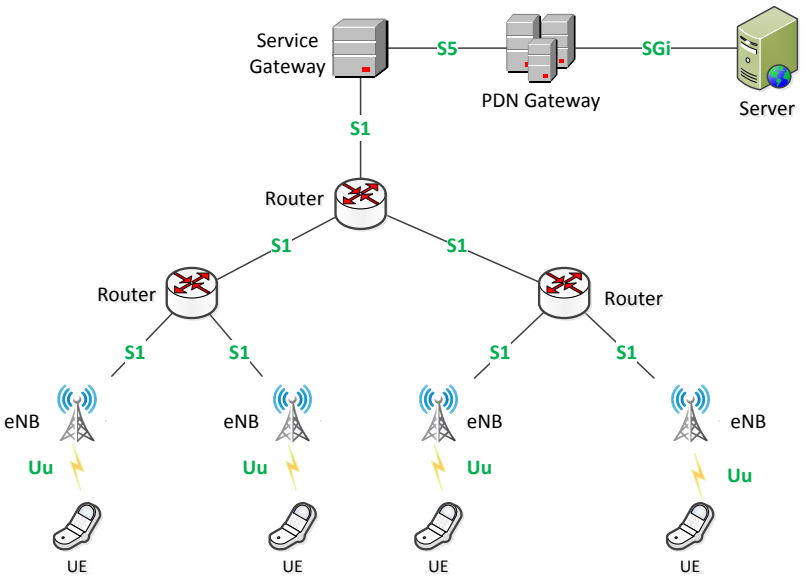


FIGURE 3.2: LTE reference model

the main functions of the eNB. In addition, the routers in the transport network, which connects the core and the access networks, have traffic management functions, e.g. service differentiation by transport scheduling. These important elements that have major impacts on the end-to-end user performance are considered in the LTE simulation reference model, shown in in Figure 3.2. The simulation reference model includes a server, PDN and service gateways, and a couple of routers, base stations and users.

The simulator is implemented according to the reference model and an example scenario is shown in Figure 3.3. The high layer protocols are relying on the build-in models provided by OPNET Modeler with some extensions, such as the application and TCP/IP layers. However, the OPNET LTE model has very complex implementations on the control plane protocols. The control plane protocols are mainly

responsible for establishing and controlling user connections, which are not main factors concerned in this work. Besides, the radio related protocols in the OPNET LTE model is not well structured following 3GPP specifications, that all the radio related protocols (PDCP, RLC and MAC) are implemented in a single layer. For these reasons, a completely new implementation is conducted, which is one contrition of this thesis work. The protocols for LTE can be separated into control plane protocols and user plane protocols. The control plane protocols are simplified modelled. The user plane protocols are used for resource management and data transfer are modelled in details with necessary details according to 3GPP specifications. The 1-tier network layout, PDN and service gateways, channel and mobility models, as well as the radio scheduler are implemented by the author. The simulator architecture and some other functionalities, e.g. GTP tunneling, PDCP layer and RLC layer protocols, were developed in collaboration with other colleagues [Zak12] [Tos13].

In the following sections, the major node entities including UE, eNB, PDN and service gateways, and their interconnections are further introduced. The network topology and layout, channel and mobility models are discussed afterwards.

3.3.2 LTE Node Models

3.3.2.1 User Equipment Node Model

The UE node model and its protocol stack are shown in Figure 3.4. The node is modified based on the OPNET LTE user node. It follows the OSI top-down protocol standard. The standard build-in protocols in OPNET Modeler, including application, TCP/UDP and IP are adopted as high layer protocols. The LTE Uu protocols, including PDCP, RLC, MAC and physical layer, are newly implemented according to 3GPP standards.

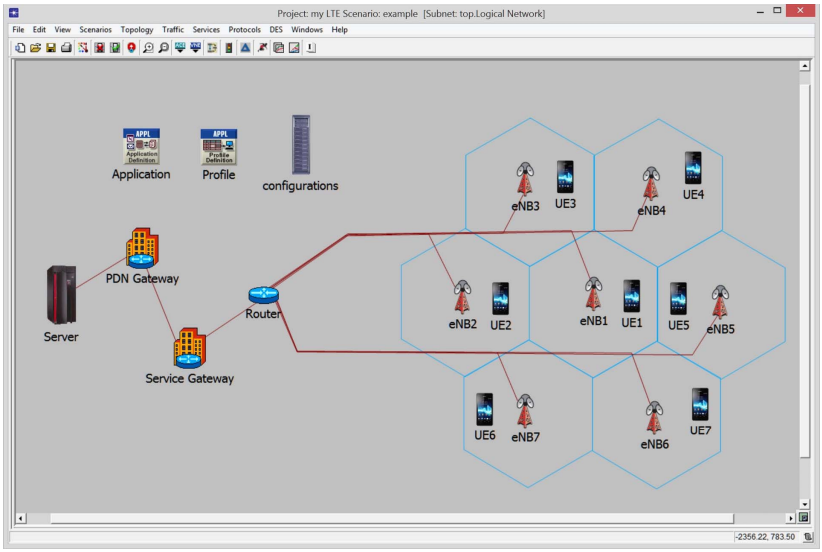


FIGURE 3.3: An example scenario in OPNET Modeler

The LTE Uu protocols were introduced in Section 2.4. The packets need to buffered at the PDCP layer before scheduling. Each bearer has its own buffer, therefore the packets need to be classified and put into its corresponding buffer. The main functionalities of the PDCP layer are modelled, including uplink packet classification based on QoS attributes, UL user and bearer identification, UL PDCP buffer management and DL PDCP de-capsulation. The concatenation/segmentation function of the PDCP PDU into RLC PDU for UL traffic and the reassemble function for DL traffic are supported. The UL HARQ is covered by the MAC layer. The physical channels are emulated at the MAC layer by a link-to-system mapping, which will be discussed in Chapter 4.

There are many parameters that can be configured separately for the individual users. For instance, some parameters, such as the serving eNB and cell, the application type and profile, the mobility speed, the

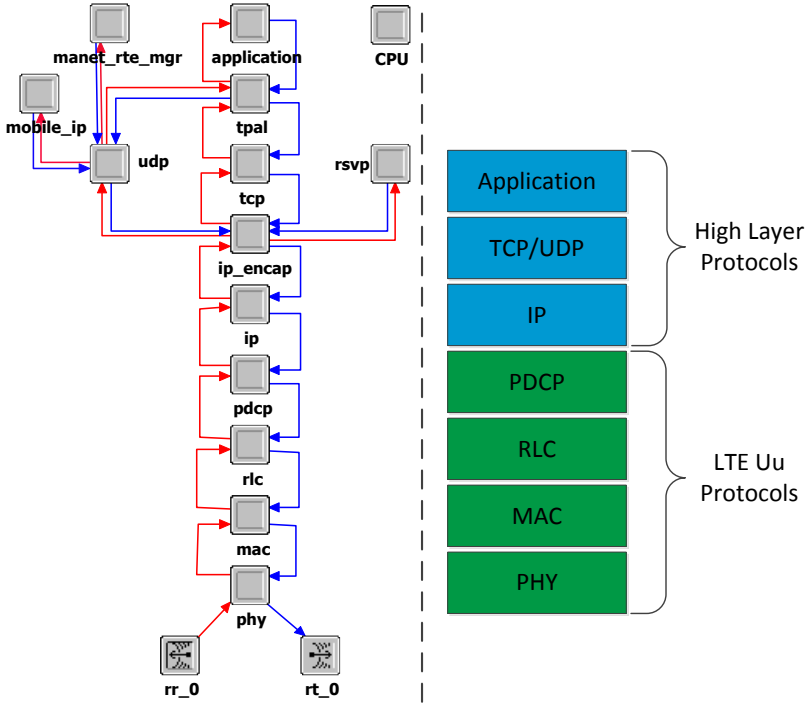


FIGURE 3.4: LTE UE node model

RLC operation mode, are associated with a user.

3.3.2.2 eNB Node Model

In LTE, the base station design is the most important part in the radio access network. In LTE, the eNB is responsible for the radio resource management and scheduling. Besides, it provides a tunnel connecting to the transport network. As it is shown Figure 3.5, there are two parts in the eNB: the transport protocols and the Uu protocols.

The LTE Uu protocol stack on the right provides the interface to the UE. The PDCP layer on the top has the following functions: DL

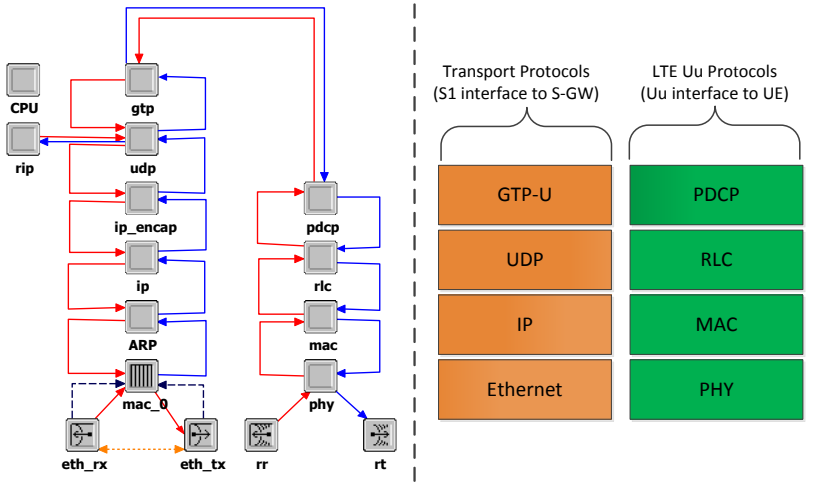


FIGURE 3.5: LTE eNodeB node model

packet classification based on QoS attributes, DL user and bearer identification, DL PDCP buffer management and UL PDCP de-capsulation. There are three operation modes supported in RLC layer: Transparent Mode, Acknowledgement Mode (AM) and Un-acknowledgement Mode (UM). Besides, the concatenation/segmentation function of the PDCP PDU into RLC PDU for DL traffic and the reassemble function for UL traffic are supported by the RLC. Since multiple users served by the same eNB are sharing the same radio channels, the Media Access Control (MAC) layer is responsible for resource allocation and scheduling. More details will be discussed in Chapter 4. The physical channels are emulated at the MAC layer by a link-to-system mapping, which will be discussed in Chapter 4.

The transport protocols include GTP-U, UDP/IP and Ethernet. The GTP-U terminates the transport network tunneling for DL traffic and tunnels the UL traffic towards the service gateway. There are

many configurable parameters are associated to each eNB, such as the scheduling method, the transmission power and the cell spectrum bandwidth.

3.3.2.3 Service Gateway Model

The service gateway routes and forwards user data packets to the PDN gateway or eNBs, while also acting as the mobility anchor for the user plane during inter-eNodeB handovers and as the anchor for mobility between LTE and other 3GPP technologies. It is responsible for creation, deletion, and modification of bearers for individual users connected to the EPS [3GP11b]. In this work, only the performance of LTE is focused, therefore the handover to other 3GPP technologies are not considered. So the service gateway is implemented in a simplified way. Figure 3.6 shows that the service gateway connects to the PDN gateway and the eNB.

The service gateway stores and manages the IP bearer service. A bearer shaping function is implemented based on the Token Bucket algorithm [Tan02]. The bearer shaping, which is a resource management in the LTE core network, is used to control the bearer rate. The bearer shaping is discussed and evaluated in Chapter 5. A bearer is a traffic separation element that enables differentiated treatment of traffic based on its QoS requirements, and provides a logical path between the UE and the gateway located in the EPC. Each bearer is assigned to a certain data rate, according to the QoS parameters (e.g. Guaranteed Bit Rate (GBR) and Maximum Bit Rate (MBR) for the GBR type bearers, and Aggregate Maximum Bit Rate-AMBR for the non-GBR type bearers). For more information, please refer to the traffic shaping in Chapter 5.1.1.

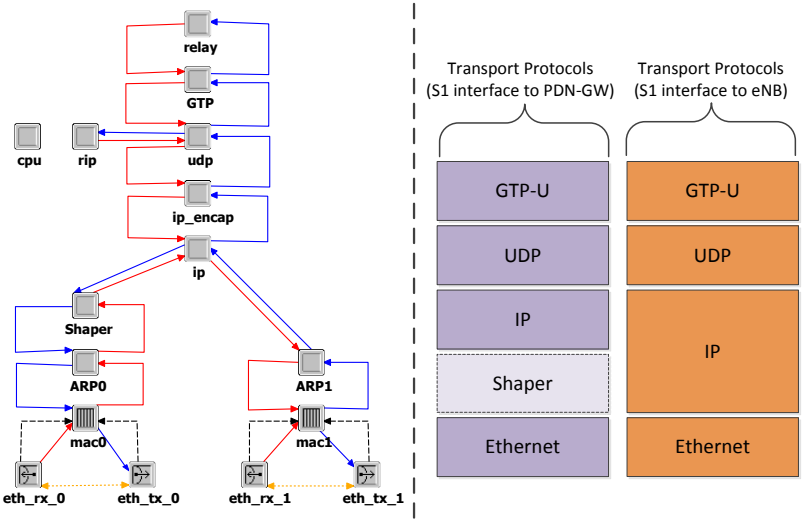


FIGURE 3.6: LTE S-GW node model

3.3.2.4 PDN Gateway Model

The PDN gateway shown in Figure 3.7 provides the connection to the external packet data networks. A UE may have simultaneous connectivities with more than one PDN gateway for accessing multiple PDNs. The traffic from different PDNs belongs to the same UE is forwarded to the UE's serving gateway. Another key role of the PDN gateway is to act as an anchor for mobility between 3GPP and non-3GPP technologies which is not been considered in this work.

3.3.3 Network Layout, Channel and Mobility Model

A seven eNBs/cells hexagonal network layout, which is well known as the 1-tier scenario, is implemented in this work. There is one cell per eNB, and each cell has a hexagonal coverage that each user is served by a station which is closest to the user. There is one eNB in the

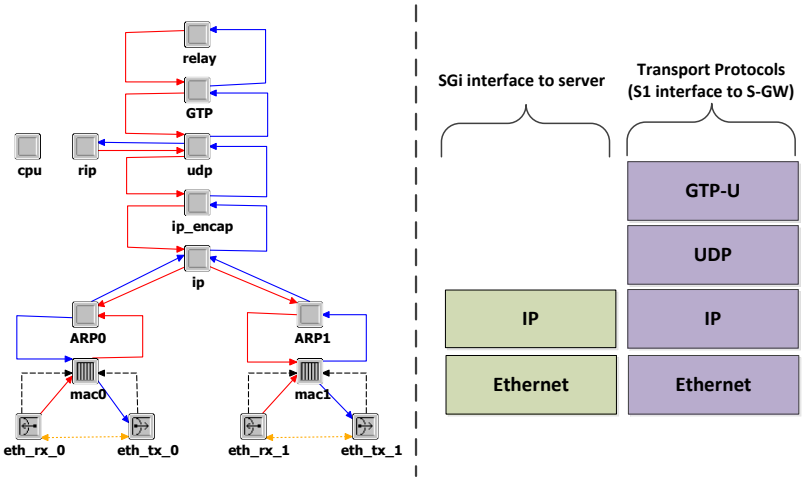


FIGURE 3.7: LTE PDN-GW node model

middle, while six other eNBs are around the centered eNB as the first tier with an inter-eNB distance of 500 meters. The model is designed to support femtocells that could be plugged into any places within the 1-tier scenario's coverage.

The users channel conditions are updated in each TTI, which is 1 ms. The radio channel is modelled considering the three most significant impact factors: path loss, slow fading and fast fading. Besides, the interferences among the eNBs are dynamically calculated. The channel condition in the term of Signal to Interference plus Noise Ratio (SINR) is updated for every PRB of all the users, considering the path loss, the slow and fast fading, and the interferences. The SINR has a critical impact on the radio performance. A higher SINR means a better channel condition. With higher SINR, higher modulation and coding schemes can be used, which results high throughput over the radio interface. Due to the frequency and time selectivity, the SINR values over different PRBs of a user are different. Each cell supports

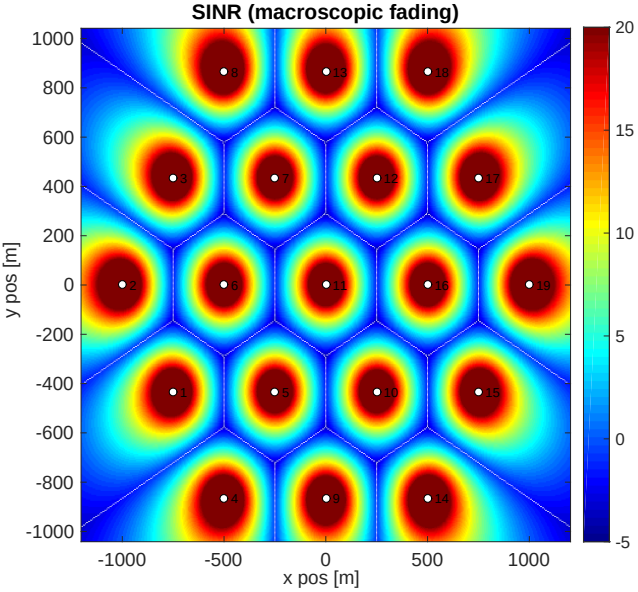


FIGURE 3.8: SINR map without fading

configurable bandwidth, scaling from 1.4 MHz with 6 PRBs, up to 20 MHz with 100 PRBs, according to 3GPP specifications [3GP06].

Path Loss: The urban area macro cell path loss model is applicable for scenarios in urban and suburban areas outside the high rise core where the buildings are of nearly uniform height [3GP11a].

$$L = 40 \cdot \left(1 - 4 \cdot 10^{-3} \cdot D_{hb}\right) \cdot \log_{10}(R) - 18 \cdot \log_{10}(D_{hb}) + 21 \cdot \log_{10}(f) + 80$$

where: L is the path loss in dB.

R is the base station-UE separation in kilometers.

f is the carrier frequency in MHz.

D_{hb} is the base station antenna height in meters, measured from the average rooftop level.

Considering a carrier frequency of 2000 MHz and a base station antenna height of 15 meters, the path loss is given by:

$$L = 128.1 + 37.6 \cdot \log_{10}(R)$$

Figure 3.8 shows a SINR map considering the macroscopic path loss generated by the LTE TU Vienna simulator [TBK+15a]. The users at the cell center have very high SINR values so that high modulation and coding schemes can be used. The users at the cell boarder have very low SINR values that nearly only the lowest modulation and coding scheme can be used.

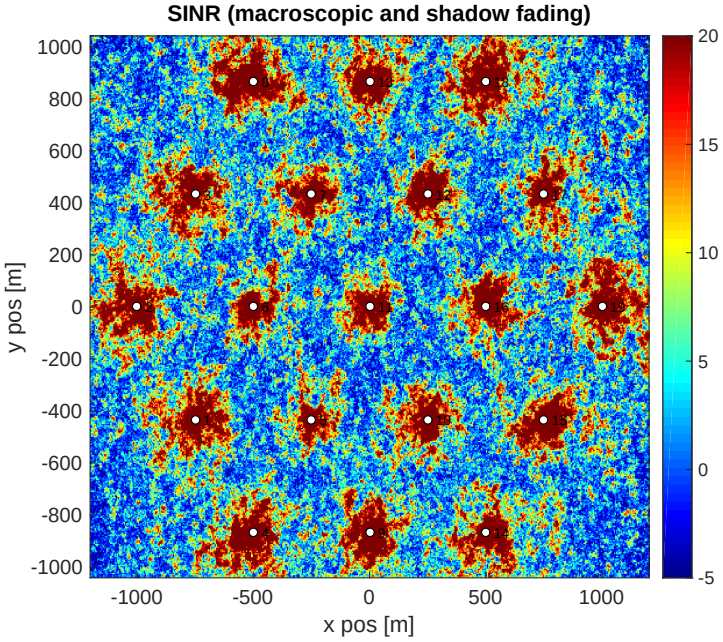


FIGURE 3.9: SINR map with slow fading

Slow Fading: It is caused by shadowing from obstacles in the propagation path between the UE and the eNB affecting the wave propagation. A two-dimensional Gaussian process with appropriate spatial correlation is desired in order to capture the macrocell diversity in a realistic way [CG03]. A low-complexity method capable of introducing space correlation into the Gaussian process while still preserving its statistical properties as well as inter-site correlation has been used by the TU Vienna LTE downlink simulator [Cla05].

Figure 3.9 shows a SINR map considering macroscopic path loss and slow fading generated by the LTE TU Vienna simulator. Each eNB has its own slow fading map, and the inter-site correlations are considered in the construction of the maps. Comparing to Figure 3.8, the SINR map shows higher randomness reflecting the irregularities of the geographical characteristics of the terrain. In this work, the slow fading maps used in the OPNET simulator are pre-generated.

Fast Fading: Due to the multipath propagation of the wireless signal, fast fading occurs when the coherence time of the channel (time selectivity) is small relative to the symbol duration. For different frequency with different wavelengths, a lightly different phase shift of superimposing components. So the channel is frequency selective due to the superimposition of multipath components. ITU defines the Pedestrian and Vehicular models for simulating the fast fading effects for low and high mobility users [3GP02]. Figure 3.10 gives an example of fading effects based on the ITU Pedestrian A model with 5 km/h speed within 1s duration over 20 MHz bandwidth at 2 GHz frequency. The channel is both time and frequency selective that the fast fading is different over different PRBs and changes over time. Temporary deep fading higher than 30 dB can be observed. Therefore the scheduler should consider the variations of channel conditions using time diversity to avoid a temporary deep fade.

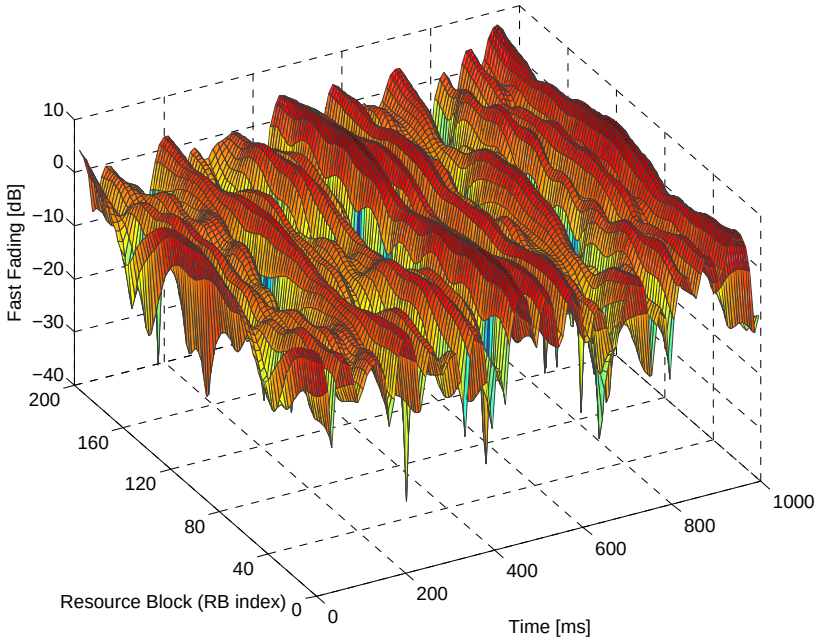


FIGURE 3.10: An example of fast fading effects

Inter-eNB Interference: Unlike the previous GSM and UMTS systems, LTE is designed to have a frequency reuse factor 1. The neighbor cells use the same spectrum band, therefore the inter-cell interference has a critical impact on the channel condition. The interfere needs to be considered in a multiple cell scenario. Suppose there are N_c cells/eNBs in total, the SINR value of a user in cell j over PRB i is calculated as:

$$SINR_i = \frac{P^{tx}_{i,j} \cdot G_{i,j}}{N_0 + \sum\limits_{k=1,k\neq j}^{N_c} \xi_{i,k} \cdot P^{tx}_{i,k} \cdot G_{i,k}}$$

Where $P^{tx}_{i,j}$ is the transmission power of the serving cell j over PRB i while $P^{tx}_{i,k}$ is the transmission power of the neighbor cell k over PRB

i. $G_{i,j}$ is the channel fading considering the antenna gain, the path loss, the slow and fast fading from the serving cell while $G_{i,k}$ is the channel fading from neighbor cell k. $\xi_{i,k}$ equals to one when the PRB i in cell k is been used, otherwise it is set to 0. N_0 is the Additive White Gaussian Noise (AWGN).

Mobility Model: Two well-known mobility models have been implemented in the simulator: Random Way Point (RWP) and Random Direction (RD) [CBD02].

RWP randomly selects an end point within the cell coverage and moves towards the point. Once the user reaches the point, a new destination is selected. In RWP, the user density is not uniformly distributed in a cell. The user density is high in the cell center and low in the cell edge. This property is used to model the case that base stations are favored to be deployed in the places with high user density in reality. RD overcomes the non-uniform user density problem by randomly selecting a moving direction towards to the boundaries. RD has a uniform user density within the cell coverage. Figure 3.11 shows the trajectories of user movements based on the RD model.

3.3.4 Link-to-System Mapping

Generally, the network simulations are divided into two levels: link level simulations and system level simulations. In link level simulations, as the name suggests, the performance of the radio links as well as the physical layer is evaluated on symbol or bit level, considering the modulation, coding schemes, antenna patterns, MIMO, equalizations, etc. On the other hand, the system level simulations, focusing on higher layers and large-scale networks, are able to evaluate the overall system performance. It is favorable to have a simulation model combining both link and system level simulations. However, in reality, it is not feasible to consider link level aspects in large-scale scenarios (e.g. with multiple users and base stations) due to complexity. However, in order to have

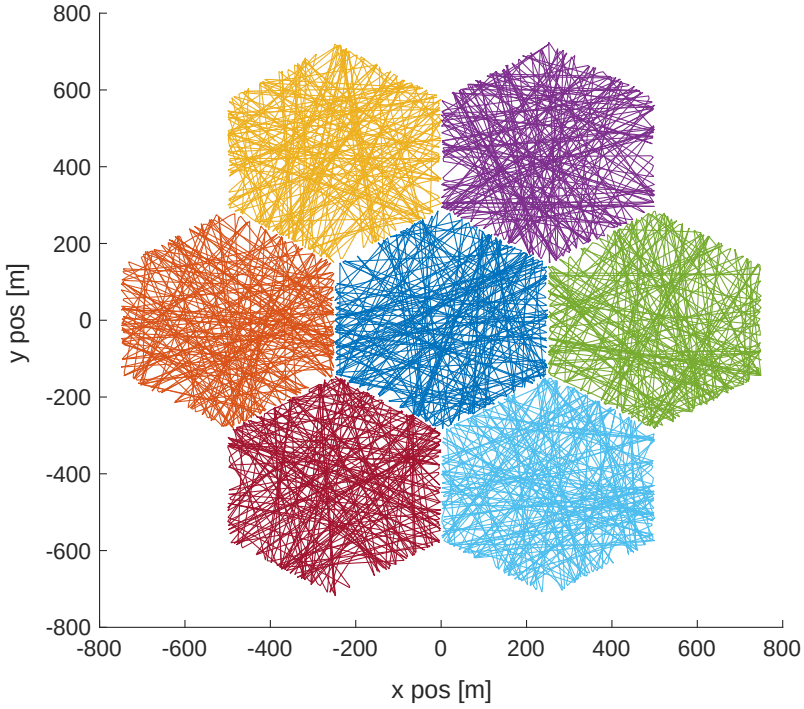


FIGURE 3.11: An example of user moving traces with RD mobility model

an accurate system level model for performance evaluation, the link level performance is emulated statistically and mapped to system level by a Link-to-System mapping method.

In this work, a statistical Block Error Ratio (BLER) model is adopted based on the link-to-system mapping. In the LTE scheduling, the modulation and coding scheme (MCS) is selected according to the channel quality, which is obtained from CQI reports. Figure 3.12 shows BLER curves from SISO (Single-input Single-output) AWGN simulations for 15 CQI values without using HARQ [TBK+15b]. Each curve is spaced approximately 2 dB from each other. The MCS is chosen as

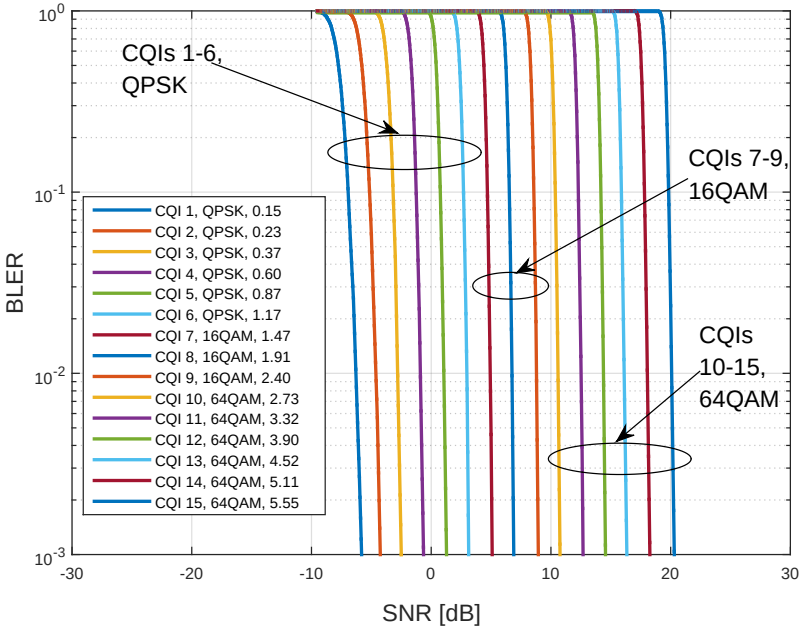


FIGURE 3.12: BLER curves from SISO AWGN simulations for 15 CQI values

high as possible on one hand. On the other hand, the expected BLER value should be smaller than 10% with the selected MCS. By plotting a 10% BLER line, the 10% BLER intersection points for 15 CQI values are obtained and they are drawn in Figure 3.13. Therefore in the system level simulation, a given effective SINR value can be mapped to a CQI value, and the corresponding modulation and coding schemes are decided. For example, for an effective SINR value of 10 dB, it is mapped to CQI 9 (by a floor function), and the corresponding MCS is known.

LTE uses OFDM technique and the signals can be mapped to different subchannels (i.e., PRBs) for transmission. The signal over different PRBs have different channel conditions due to the frequency

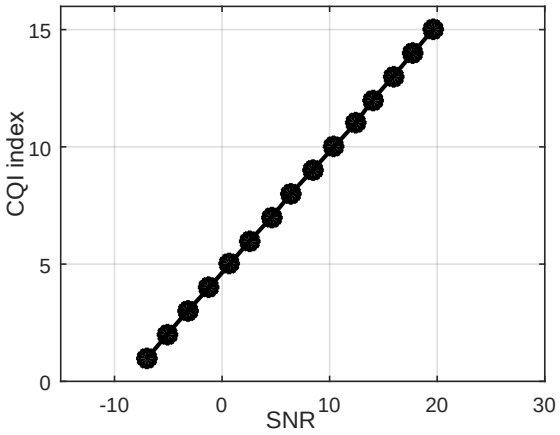


FIGURE 3.13: CQI mapping. BLER=10% points for 15 CQI values from Figure 3.12

selective fast fading, the interferences from neighboring cells and random Gaussian noises. An effective SINR per user must be determined considering the channel conditions over multiple PRBs. The Exponential Effective SINR Mapping (EESM) method is adopted in this work. The EESM method was widely used because of its accuracy and simplicity [BAS+05]. The effective SINR over multiple PRBs is calculated based on eq. (3.1).

$$SINR_{eff} = -\beta \cdot \ln\left[\frac{1}{N}\sum_{n=1}^{N}\exp\left(-\frac{SINR_n}{\beta}\right)\right] \tag{3.1}$$

where N is the total number of PRBs, $SINR_n$ is the SINR value of the n_{th} PRB and β is the MCS dependent scaling factor. The different β for different MCSs can be found in [KSW+08]. As long as the user's effective SINR is determined, the CQI can be determined and the appropriate MCS is chosen correspondingly. This information is taken into consideration for the radio resource allocation. Based on user's

MCS and the number of PRBs the user gets, the Transport Block Size (TBS) is determined based on the standard table from 3GPP [3GP10d] (see Appendix D.1). The TBS is the amount of data can be transmitted for the user in this TTI.

In LTE, the HARQ is used to handle the transmission error. It includes two functionalities to increase the retransmission success probabilities: chase combining and incremental redundancy. Chase combining retransmits identical copies of the original transmission. The maximum ratio combining technique is used at receiver to sum the packets. According to 3GPP specifications [3GP10a], every bearer supports up to 8 stop-and-wait simultaneous HARQ processes. In one TTI, a user can either perform a pending retransmission or a new transmission, but not at the same time. In case of a transmission error, a retransmission is performed 8 ms later. This behavior is modelled in the simulator. The transmission error rate is modelled using a statistical method. The error rate for new transmission is set to 10% that is the same as the target BLER rate for the effective SINR to CQI/MCS mapping. Due to chase combining and incremental redundancy, the retransmission error rate can be reduced effectively. In this work, the first retransmission error rate is set to 1% while the second retransmission is set to 0%.

3.3.5 User Traffic Models

In network simulations, it is critical to model the network services and traffic types in a proper way to get a meaningful evaluation of the network performance. Different from the previous UMTS networks, LTE has a full-IP packet based network. Voice over IP (VoIP) is supported in LTE to replace the conventional circuit-switch based voice service. Some other typical real time services include real-time gaming and streaming. These real-time services, which use User Datagram Protocol (UDP) as the transport layer protocol, have critical requirements on the delay. Besides, the buffered video streaming, web browsing and file downloading/uploading are the most widely used services nowadays

and create massive traffic on the network. These services rely on the Transmission Control Protocol (TCP).

3.3.5.1 UDP Based Application Models

Voice over IP (VoIP): VoIP services are based on the Real-Time Protocol (RTP) at the application layer. The VoIP services use UDP as the transport layer protocol. UDP uses a simple but fast connectionless model with no guarantee of successful delivery. UDP is favored for the real-time VoIP services to meet the critical delay requirements. The GSM Enhanced Full Rate (EFR) codec is one of the most widely used voice codecs. It belongs to the family of Adaptive Multi-Rate (AMR) codecs, with an application data rate of 12.2 kbps. Some other codecs are supported as well which are summarized in Table 3.1. The voice service is modeled based on an ON/OFF model representing the talk spurt and silence periods. In addition, the compression/decompression and de-jitter buffer delay are modelled. The default setting are summarized in Table 3.2.

TABLE 3.1: Codec properties

Codec Properties				
Codecs	GSM EFR	G.711	G.729a	G.722.2
Codec rate	12.2 Kbps	64 Kbps	8 Kbps	23.85 Kps
Frame delay size	20 ms	20 ms	10 ms	20 ms
Frames per packet	1	1	2	1

Real-time video streaming: Similar to the VoIP services, the real-time video streaming services are based on RTP. A Constant Bit Rate (CBR) video model is adopted with two important parameters:

TABLE 3.2: Codec settings

Codecs Settings	
Silence length	Exponentially distributed with mean of 3 seconds
Talk spurt length	Exponentially distributed with mean of 3 seconds
Compression delay	20 ms
Decompression delay	20 ms
De-jitter buffer delay	40 ms

the frame size and the inter-frame time interval (frame rate). The real-time video streaming service is studied in the Appendix A.

3.3.5.2 TCP Based Application Model

TCP provides reliable error-free transmission services to applications. Buffered video streaming, web browsing and file downloading/uploading are TCP based applications. The OPNET Modeler supports file transfer in both uplink and downlink directions. The GET command, which is sent from users to the server, triggers a file downloading and PUT command triggers a file uploading. By configuring the inter-arrival time (IAT) or the idle time, a series of the file transfers of a user is modeled. The TCP based traffic model can be seen in Figure 3.14 and some configurable parameters are shown as follows:

- **GET percentage**: The percentage of file downloading. For example, if it is set to 50%, 50% of file transfers are file downloading and 50% are file uploading.
- **File size**: The file size in bytes. It can be a fixed value or can follow different kinds of distributions, e.g. exponential, uniform, normal, etc.

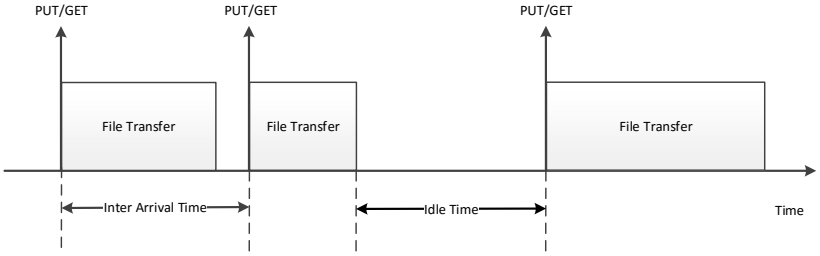


FIGURE 3.14: TCP based traffic model

- **IAT**: Time between the starting time (when the GET/PUT command starts) of two consecutive file transfers.

  Maximum simultaneous file transfers per UE: It might happen that the next file transfer request starts before the ending of previous file transfers. This parameter limits the number of maximum simultaneous file transfers. If the limit is reached, the new request will be buffered until a file transfer is finished and the number is below the limit.

- **Idle/Reading time**: The times between the ending of the previous file transfer and the beginning of the next file transfer. For instance, this can be used to model the human behavior of web browsing services. Humans need some time to read the current website before taking the next action. If the idle/reading time is used instead of IAT, there is maximum one file transfer per UE in parallel, since a file transfer always starts after the ending of the previous file transfer. In a simulation run, either the IAT or Idle/Reading time should be configured.

The OPNET modeler supports a variety of TCP flavors with many parameter settings. The default flavor used in this work is TCP New Reno which is one of the most widely used TCP flavors nowadays. Some typical settings are listed as follows:

- **TCP flavor**: New Reno with fast recovery and fast retransmit enabled.
- **Maximum Segment Size (MMS)**: 1400 Bytes. This is to avoid packet segmentation at the IP layer since more headers are added by GTP tunneling in LTE transport networks.
- **TCP receive window size**: 64 KBytes.

3.4 Statistical Evaluation

The primary focus of network simulation is to evaluate the designed system performance in steady-state. The estimation of the steady-state performance is done by analyzing and post-processing the collected statistical data from the simulation runs. In the beginning of a simulation run, the system needs some time to initialize the system, setting up the routing table, starting the application profiles, etc. For instance, many users may start file transfers in a burst. In addition, it needs some time until the TCP based application to change the initial slow start state into the congestion avoidance state. This initial phase is called "warm-up" period, and the results during this period should be omitted in result collection.

The system performance is evaluated based on the estimating of the steady-state mean. The batch means method and independent replications method are two of the most widely used methodologies. The batch means method simply splits a simulation run into a number of contiguous non-overlapping batches. The sample means of the batches are used for further statistical data processing. Nevertheless, the correlations among the batch means very likely exist in a simulation run. In the independent replications method, the means are calculated from the independent replications of simulation runs to avoid the possible correlations. The method is adopted in this work and confidence

intervals are used as an indication for the quality of the estimation of the mean.

3.4.1 Independent Replications

In this method, m independent simulations are run with the same initial states and same system settings. Each simulation run uses a different seed in a Random Number Generator (RNG) so that each run uses a different series of random numbers. In the OPNET modeler, the default RNG is Mersenne Twister (MT), which is currently the most widely used RNG in network simulation. Assuming that $x_{i1}, x_{i2}, x_{i3}, ..., x_{iN}$ are the samples obtained in i_{th} simulation run, the sample mean is calculated by:

$$y_i = \frac{1}{N} \sum_{j=1}^{N} x_{ij} \qquad i = 1, 2, 3, ..., m$$

In total, m sample means are obtained by m simulation runs. The mean and the variance of m sample means are calculated below and they can be used in confidence interval calculation.

$$y\,(m) = \frac{1}{m} \sum_{j=1}^{m} y_j$$

$$S^2\,(m) = \frac{1}{m-1} \sum_{j=1}^{m} \left(y_j - y\,(m)\right)^2$$

3.4.2 Confidence Interval

The Confidence Interval (CI) [And84] is often used to express the precision and uncertainty associated with a sampling method, e.g. batch means and independent replications. The CI is a range which contains the real mean μ with α ($0\% < \alpha < 100\%$) confidence level. Mathematically, let Y be a random sample and μ is the real mean ($E\,(Y) = \mu$),

with confidence level α. The CI is an interval $([l(Y), u(Y)])$ with the following property:

$$P\{l(Y) < \mu < u(Y)\} = 1 - \alpha$$

Assuming there are m samples $(y_1, y_2, y_3, ..., y_m)$ obtained by independent replication of simulations. $y(m)$ and $S^2(m)$ are the sample mean and the variance of samples. Suppose the real variance is unknown, the CI is calculated as:

$$\left[y(m) - \frac{t_{(\alpha/2,m-1)} \cdot S(m)}{\sqrt{m}}, y(m) + \frac{t_{(\alpha/2,m-1)} \cdot S(m)}{\sqrt{m}}\right]$$

where $t_{(\alpha/2,m-1)}$ represents the upper critical value of the Student-T distribution with *m-1* degrees of freedom. The Student-T distribution is in use due to unknown variance leads to a larger confidence interval comparing to known variance. In case the real variance σ is known, the Student-T distribution is replaced by the Normal distribution and the CI can be calculated as:

$$\left[y(m) - \frac{z_{\alpha/2} \cdot \sigma}{\sqrt{m}}, y(m) + \frac{z_{\alpha/2} \cdot \sigma}{\sqrt{m}}\right]$$

where $z_{\alpha/2}$ represents the upper critical value of the Normal distribution. To be noticed, if the number of samples is sufficiently large, e.g. bigger than 30 [SS15], the Student-T distribution can be approximated by the Normal distribution.

Chapter 4

Utiliy-based Radio Scheduling in LTE

As detailed in Chapter 1, the LTE access network contains two main bottlenecks: the air interface and the transport backhaul [3GP12]. In legacy mobile networks, the bandwidth of transport backhaul is well dimensioned by mobile operators considering the traffic mixture and delay requirements to cover the typical scenarios [LLT+12], and therefore, the transport backhaul is typically not the bottleneck. However, in case multiple cells have very high traffic load at the same time, the transport backhaul could be overloaded. Besides, LTE femtocell networks use cable or broadband xDSL as the last mile transport and usually its bandwidth is limited by the contract as well.

In this chapter, the transport backhaul limitation is not been considered. The transport backhaul limitation will be considered and studied in Chapter 5. Figure 4.1 shows an example scenario with an air interface bottleneck. In this scenario, there are three cells, each with an average radio capacity of 40 Mbps, sharing the same transport backhaul serving with 100 Mbps. Cell A has a high traffic load while cell B and C are nearly idle. Therefore, the transport backhaul is not a

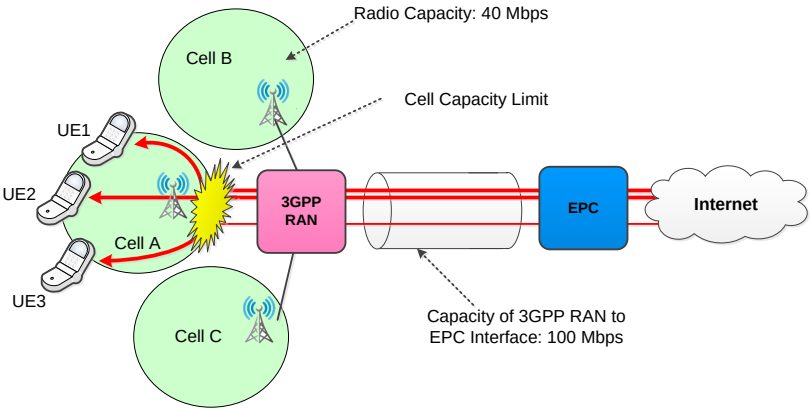


FIGURE 4.1: Performance limited by cell capacity (example capacities) [3GP12]

limiting factor in this scenario. Over the air interface, a key mechanism in the LTE traffic management is the packet scheduler, which decides the amount of radio resources allocated to each active users.

The structure of of this chapter is organized as follows: in Section 4.1, the LTE scheduler and an overview of existing scheduling strategies are discussed. Besides, the state-of-art and the motivations of QoE/utility-based scheduling are introduced. The relationship between QoS and QoE is discussed afterwards, and how the QoE based utility functions for various traffic are explained in details with several examples. Subsequently in Section 4.2, a QoE based scheduler is proposed and proven to be optimal analytically. The scheduling procedures are illustrated step by step. Afterwards in Section 4.3, the performance gain of the proposed QoE base scheduler framework is compared against the conventional proportional fair scheduler by numerous simulations. Section 4.4 gives a summary of this chapter.

4.1 State of the Art

4.1.1 LTE Scheduler in General

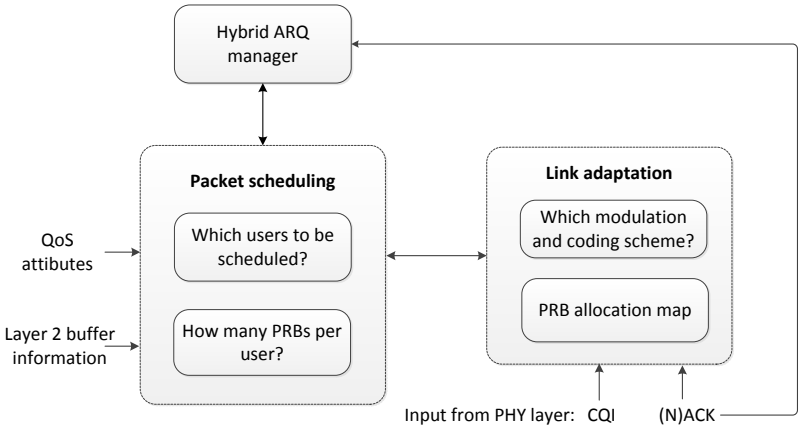


FIGURE 4.2: General packet scheduling framework

Figure 4.2 shows a general packet scheduling framework in LTE downlink, which mainly contains two parts: packet scheduling and link adaptation. In LTE, since the packet scheduling algorithm is not specified in the standard, eNodeB vendor can define their own scheduling algorithms. The radio resource allocation and packet scheduling is performed periodically every TTI (Transmission Time Interval, 1 ms in LTE). In every TTI, the packet scheduling selects which active users to be scheduled and decides the allocation of radio resources in terms of PRBs among the users. The scheduler is aware of the user channel conditions based on the CQI reports input from the physical layer, and it determines the corresponding MCS (Modulation and Coding scheme). Afterwards, based on the amount of PRBs allocated to the users and MCS, the TBS (Transmission Block Size), which is the amount of data that can be transmitted in the current TTI, is determined separately

for the users to be scheduled. The PRB allocation map as well as the corresponding MCS are broadcasted to the UEs over the control channel PDCCH. The scheduled UEs will receive their packets by accessing the corresponding PDSCH payload.

In general, the scheduling decision depends on various factors as seen from Figure 4.2, e.g. CQI reports, QoS attributes, buffer/queue status and so on, which are introduced as follows:

- Channel Quality: obtained by CQI reports from the users and can be used in channel aware scheduling. The users with good channel quality are preferred to be scheduled.
- HARQ: it needs to decide whether there is a pending retransmission. For complexly reasons, the scheduler cannot schedule a pending retransmission and a new transmission at the same time for the same user according to the 3GPP specification [3GP10a].
- QoS attributes: different bearer traffic associated with different QCI (QoS Class Identifier) have different QoS requirements. The scheduler needs to consider these QoS attributes to meet their requirements.
- Buffer/queue status: the buffer/queue status includes the queue length and queueing delay. The former one needs to be considered in order to avoid buffer overflow. The latter one is considered mainly for real-time delay sensitive traffic flows to meet the delay requirements.
- Scheduling history: this is used to improve the fairness among the users. The users in bad channel condition might starve or not be served for a long time causing a severe fairness problem.

TABLE 4.1: A summary of LTE scheduling strategies

Type	Scheduler	Simplicity	Spectrum efficiency	Fairness	QoS provisioning
Channel-unaware (Legacy wired system)	FIFO	++	-	-	-
	Round Robin	++	-	++	-
	Blind Equal Throughput	++	-	++	-
	Resource Preemption	+	-	-	+
	Guaranteed Delay	+	-	-	+
Channel-aware QoS-unaware (Non-real time services)	Maximum Throughput /Maximum C/I	++	++	-	-
	Proportional Fair	+	+	+	-
	Throughput to Average	+	+	+	-
	Joint Time and Frequency Domain Scheduler	++	+	+	+
Channel-aware QoS-aware (Real time services)	Scheduler for Guaranteed Data-Rate	+	-	+	++
	Scheduler for Guaranteed Delay Requirements	+	-	+	++
	Semi-persistent Scheduler for VoIP Support	-	-	+	++

4.1.2 Overview of the Scheduling Strategies for LTE Downlink

Resource allocation has been intensively researched in wireless and mobile communication networks. The radio resource allocation is always one of the most important research topics in LTE. The key objective of radio resource allocation mechanisms is to make the best use of limited resources and to enhance the users' experience, under time varying channel conditions. [CPG+13] gives a summary on downlink packet

scheduling in LTE cellular networks. As seen from Table 4.1, the most widely used schedulers in LTE are classified into three categories: i) channel-unaware; ii) channel-aware; QoS-unaware; iii) channel-aware; QoS-aware [CPG+13]. Besides, some researches focus on energy-aware scheduling. However, it is mainly for LTE uplink and not been considered in this work. The scheduling in LTE depends on many factors and has various targets which normally conflict with each other. The key design targets include simplicity, spectrum efficiency, fairness, QoS provisioning and so on. The compromise among different targets needs to be considered in designing scheduling schemes.

4.1.2.1 Channel-unaware

These schedulers are designed based on the assumption of time invariant and error free transmission media and mostly used in legacy wired networks [Tan07]. Since the user channel conditions are not considered, the spectrum efficiency is low. However, these schedulers are with low complexity and good fairness performance, and they can be used as a benchmark for designing LTE schedulers.

FIFO serves the users in First In First Out order, therefore the fairness is not considered at all. On the contrary, Round Robin serves the users with a fair share of available resources while the Blind Equal Throughput targets for throughput fairness. Resource Preemption gives absolute priority to high priority users, which provides better performance to the users with high priorities. However, the users with low priorities might get starvation. Guaranteed Delay considers the queuing time to meet the delay requirements of guaranteed delay services. Earliest Deadline First (EDF) and Largest Weighted Delay First (LWDF) are two disciplines [Tan07],[LL03].

4.1.2.2 Channel-aware, QoS-unaware

In mobile networks, the scheduling can take advantage of the CQI reports from the users to have multi-user diversity gain and achieve better spectrum efficiency, by allocating the resources to the users with relatively good channel conditions. The following schemes are mainly designed for best effort traffic, which has no crucial QoS requirements. The Maximum Throughput targets to allocate the resources only to the users with the best channel condition regarding less the fairness. It is an opposite extreme scheduler comparing to the Blind Equal Throughput. Proportional Fair [KLZ09] and Throughput to Average [KPK+08] make a compromise between Maximum Throughput and Blind Equal Throughput, thus a trade-off between fairness and spectrum efficiency. Joint Time and Frequency Domain Schedulers [PPM+07] have a two-step approach and offers a good trade-off between fairness and spectrum efficiency. The first step, a Time Domain Packet Scheduler (TDPS) selects a subset of active users as well as HARQ users to be scheduled, considering the channel condition, scheduling history and QoS attributes. In the second step, only selected users are scheduled by a Frequency Domain Packet Scheduler (FDPS), taking the buffer status information into consideration. Thanks to this two-step approach, the complexity of the packet scheduler is simplified since only few amounts of users are pre-selected for scheduling. Therefore, this scheduler is well researched and widely adopted by the vendors.

4.1.2.3 Channel-aware, QoS-aware

In real systems, the real time traffic is scheduled prior to the elastic traffic. The above-mentioned schemes do not consider the QoS requirements, and therefore not suitable for high priority traffic with strong QoS requirements. Scheduler for Guaranteed Data-Rate [MPKM08] and Scheduler for Guaranteed Delay Requirements (modified LWDF seen in [AKR+01]) are designed to guarantee the minimum requited

data rates or delay constraints. Semi-persistent scheduler for VoIP Support is mainly designed to serve VoIP traffic exploiting its periodic behavior (e.g. with GSM EFR codecs, a packet is generated every 20ms when the user is active). It reduces the use of control channels by pre-allocating resources to the active users, and therefore improves the cell capacity [FLKV08].

4.1.3 QoE/Utility-based Scheduling

The different scheduling disciplines regarding fairness, QoS provisioning, spectral efficiency and spectrum efficiency are discussed above. However in general, radio scheduling aims not only to improve the network performance, but also to increase users' QoE, which is often represented by a utility function. QoE, defined as "the overall acceptability of an application or service, as perceived subjectively by the end user" [IT07], is not a linear function of QoS (Quality of Service) parameters and depends on the application. For example, for a user streaming high quality audio the difference in means of user satisfaction when served with 2 instead of 1 Mbit/s is low, if not negligible, whereas the user is obviously much more content when served with 100 instead of 50 Kbit/s. Therefore QoE has drawn increasing attention in research in communication networks and mapping between QoS and QoE for different applications have been proposed based on experiments and different models [FHTG10].

Utility is an abstract concept from the field of Economics and is derived mainly from Von Neumann and Morgenstern [NM44]. The utility-based resource allocation targets to maximize the aggregated user satisfaction under resource constraints. This approach can be formulated as an optimization problem maximizing the aggregated utility in the system, subject to resource limitations. One of the main advantages of the utility-based optimization is that, for a new traffic type or user category, only a new corresponding utility function needs to

be added to the existing optimization formulations. Besides, the optimization formulation often allows analytical evaluation [SL05].

Utility-based radio resource allocation has been widely researched. A utility-based adaptive bandwidth sharing for elastic traffic is proposed in [Rza05] for wired networks. This is formulated as a concave maximization problem with linear constraints. It is shown by the Karush–Kuhn–Tucker (KKT) conditions [Rza05] that the optimal bandwidth allocation can be achieved. The utility-based optimal resource allocation is studied in wireless networks in [KL05], [KL08], and [CWC+11] considering the channel conditions. [KL05] proposes a heuristic to maximize the total utility in an iterative matter. In [KL08] the authors further prove that the performance gap between their proposed mechanism and the optimal solution is bounded. A unified utility function for QoS traffic and best effort traffic is proposed in [CWC+11]. In [LXZ+12] and [TCJ+11], the Mean Opinion Score (MOS) is used as a common utility metric for the quality of experience and the performance and feasibility are studied in LTE networks.

All the above-mentioned papers with utility based radio resource allocation are theoretical work for general mobile networks, for example in [KL05], [KL08], and [CWC+11]. However, in LTE, many practical factors must be taken into consideration. Firstly, only integer amount of radio resources in terms of PRBs can be allocated to users. Secondly, the buffer status information needs to be considered rather than a simple full buffer model. For instance, there is only one packet generated every 20ms for the VoIP users with GSM EFR codec. Besides, in case of TCP based traffic, there is a limited amount of packets in the buffer, especially during the slow-start phase and fast recovery phase in case of packet loss. Thirdly, HARQ is used for retransmission and it should be taken into account in scheduling. In the same TTI, either the retransmission or a transmission is allowed for one user. Fourthly, service differentiation has not been studied in the related works. Finally yet importantly, the performance of utility based radio resource

allocation is not intensively evaluated under a system level simulation implemented based on 3GPP specifications.

Therefore, in the following part of the work, how the utility functions should be constructed considering QoE will be discussed. The system model and assumptions are presented afterwards. An QoE based scheduling model is proposed and proven to be optimal theoretically. Then, a practical scheduler is designed based on the extension of the optimal model. The major focus of this work is for non-guaranteed bit rate (non-GBR) traffic. Because the GBR traffic is mapped to GBR bearers which have a higher priority over the non-GBR traffic, which is mapped to non-GBR bearers (see Session 4.2.1). Besides, there is admission control mechanisms in LTE to make sure that when the GBR user is accepted only when there are enough resources left to guarantee certain QoS requirements [PKF+09]. Nevertheless, the extension of the designed scheduler with GBR traffic support is studied in the Appendix A. In Section 4.3, the performance of the proposed scheduler framework is evaluated by simulation.

4.2 QoE-based Scheduler Design

4.2.1 LTE Quality of Service

In general, Quality of Service (QoS) is the overall performance of a telephony or computer network. In order to measure the QoS quantitatively, some key performance indicators are often used, such as transmission delay, data rate, error rate and jitter and so on. In LTE, 3GPP defined the LTE QoS framework using the Evolved Packet System bearer model [3GP15c]. The EPS bearer provides a logical path from the end user in E-UTRAN to the PDN gateway in EPC. All the packets mapping to the same EPS bearer have the same forwarding treatments, e.g. packet scheduling, queue management, rate shaping,

etc. Different EPS bearers have distinguished QoS requirements defined by 3GPP shown in Figure 4.3, allowing service differentiation based on their corresponding priority settings. Each bearer has a set of parameters describing the required network performance.

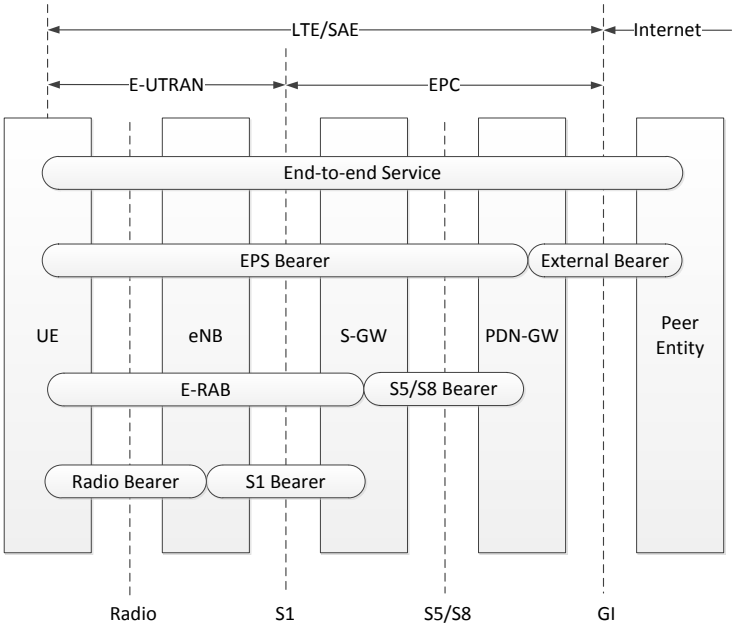


FIGURE 4.3: LTE bearer system overview

LTE supports multiple media services including voice call, live streaming, HTTP, FTP and so on. There are nine different types of bearers, identified by QoS Class Indicator (QCI). Each QCI is associated with a set of parameters, including resource/bearer type, Allocation and Retention Priority (ARP), packet delay budget and packet loss rate requirement (see Table 4.2). These bearers are been classified into two main categories: GBR and non-GBR traffic served by GBR and non-GBR bearers accordingly.

- Guaranteed Bit Rate bearers (QCI 1 to 4): these bearers provide services with guaranteed minimum data rates as the name suggested. They mainly serve high priority real time traffic, e.g. conversational voice or videos, real time gaming, etc. The GBR bearers have a higher allocation priority over non-GBR bearers and only second to IMS (IP Multimedia Subsystem) signaling. Besides, the GBR bearers with QCI 1 to 3, which provide real time services, have very tight delay budgets and rather loose packet loss rate requirements. This is coherent with the demands of the real time traffic since they are very sensitive to the delay. In reality, these services are usually based on UDP to have a fast end-to-end delivery without retransmission. To be noticed, the Maximum Bit Rate (MBR), which limits the maximum bearer rate, can be associated in a bearer.

- Non-Guaranteed Bit Rate bearers (QCI 5 to 9): on the contrary, the non-GBR bearers mainly serve non real time services with a lower priority compared to GBR bearers. They are been served only when there are resources left after serving the high priority GBR traffic. There is no guarantee on the service rate. These bearers are mainly used for best effort applications (e.g. QCI 6, 8-9), like web browsing (HTTP), file transfer (FTP) and so on. For such TCP based applications, delay is not as critical as for real time applications. Nevertheless TCP based applications have tight requirement on the packet loss, since retransmission is needed to ensure a reliable transmission in case of packet loss, and retransmission degrades the TCP performance. Although there is no Maximum Bit Rate (MBR) associated with non GBR bearers, the Aggregated Maximum Bit Rate (AMBR) is used to limit the aggregated maximum rate over all the non GBR bearers of a user.

TABLE 4.2: LTE standardized QCIs and their parameters

QCI	Resource Type	Priority	Packet Delay Budget	Packet Loss Rate	Example Services
1	GBR	2	100 ms	10-2	Conversational Voice
2		4	150 ms	10-3	Conversational Video (Live Streaming)
3		3	50 ms	10-3	Real Time Gaming
4		5	300 ms	10-6	Non-Conversational Video (Buffered Streaming)
5	Non-GBR	1	100 ms	10-6	IMS Signalling
6		6	300 ms	10-6	Video (Buffered Streaming) TCP-based (e.g., www, e-mail, chat, ftp, p2p file sharing, progressive video etc.)
7		7	100 ms	10-3	Voice, Video (Live Streaming) Interactive Gaming
8		8	300 ms	10-6	Video (Buffered Streaming) TCP-based (e.g., www, e-mail, chat, ftp, p2p file sharing, progressive video, etc.)
9		9			

4.2.2 Utility/QoE Functions

QoS describes the ability to provide services with an assured service level from the network perspective. Quality of Experience (QoE), on the other hand, reflects the end-to-end performance level from the user perspective. It indicates how well a system can satisfy the users from the user point of view. QoE is very much dependent on the QoS that the network provides, and a better network QoS will result in better QoE in general. However, this statement does not hold for all circumstances. For example, increasing bandwidth does not help so much for VoIP users in case of a large end-to-end network delay. Moreover, if the network performance is limited by some part of the network due to

congestion, improving the QoS performance of the other parts of non-congested networks can hardly contribute any QoE improvements.

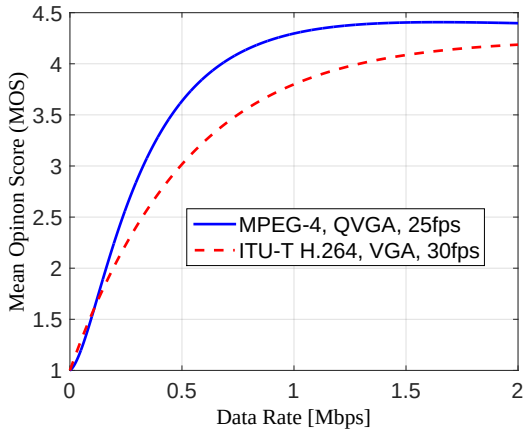
In order to provide high end user satisfaction, it is necessary to understand the QoS requirements to achieve a perceived level of QoE for different traffic and user categories. With the knowledge of the QoS requirements on the network, the proper resource management schemes can be designed accordingly at different network domains, such as radio scheduler at base station, traffic management schemes at core network, and so on. Consequently, there is a strong demand to study the relationship between QoE and QoS while designing a QoE-driven scheduler.

The Mean Opinion Score (MOS) is the most widely used subjective quality measure for QoE for decades. It was introduced in ITU-T Recommendation P.800 [IT12a] to measure the speech quality by subjective assessment methods. Its value ranges from 1 to 5, with 1 being the worst and 5 being the best experience. The MOS can be measured by subjective test. For instance, many testers in a "quiet" room listen to the audio and conduct the assessment of the voice speech quality. In addition, many mathematical models are developed for the MOS, e.g. the E-model [IT15] for VoIP with many different codecs. The extension to real time GBR traffic is discussed in Appendix A.

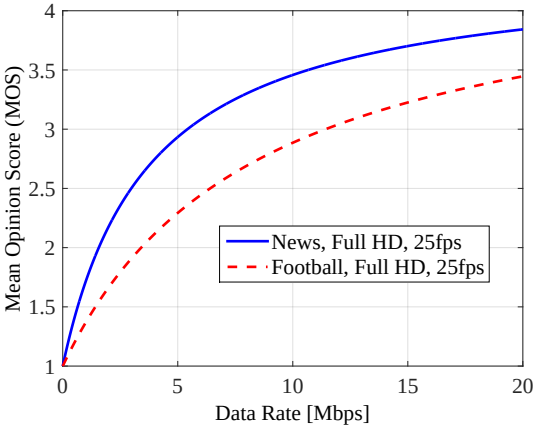
4.2.2.1 Video Streaming

The ITU-T recommendation G.1070 [IT12b] defined an assessment method to quantify the quality of video streaming applications, considering both voice speech and visual quality of video streams. The estimated video quality mainly takes video codec, frame rate, bit rate and packet loss rate into consideration. The mathematical formulation for video quality (V_q) assessment is given in the formula:

$$V_q = 1 + I_{coding} \cdot \exp\left(-\frac{P_{plV}}{D_{PplV}}\right)$$



(A) Curve fit of web browsing MOS model



(B) Curve fit of video streaming MOS model

FIGURE 4.4: Curve fitting of ITU MOS models with sigmoid function

where I_{coding} is the coding distortion for a given frame and bit rate while D_{PplV} is the degree of video quality robustness to the packet loss rate, marked as P_{plV}.

This recommendation provides the default parameters for calculating the MOS for some video codecs with variable resolutions, e.g. MPEG-4 with Quarter Video Graphics Array (QVGA) (320x240 pixels), and H.264 with VGA (640x480 pixels). Figure 4.4a shows the video quality in the term of MOS over the data rate. For both codecs, the MOS monotonically increases with the data rate. However, the marginal MOS, which is the first order derivative of the curve, is monotonically decreasing with the data rate, and this can be seen in the figure that the slope of the curves is getting progressively smaller. For example, the video with H.264 codec reaches almost 4 at 1 Mbps data rate, which means the user already has a very good user experience. There is little room for MOS improvement by getting additional 1 Mbps. Considering these two properties, the MOS is a concave function over the data rate.

Besides, many models extend the ITU-T G.1070 method to support videos with high qualities and high data rates, e.g. HD model [YH08], Temporal Complexity Model [HS12], Enhanced Temporal Complexity Model [WZD+11]. Figure 4.4b shows that the MOS is a concave function over the data rate for HD videos according to the HD model. [LAS+13] gives a summary and comparison of the models. Except this model from ITU-T, numerous previous researches for the video quality assessment are conducted from many perspectives. A new method on video quality assessment named Video Structural SIMilarity (VSSIM) is proposed in [WLB02]. The video functions for different video sequences obtained with transcoding based the VSSIM index method are shown in [TCJ+11]. The study shows a similar concave behavior to the ITU-T model.

4.2.2.2 Web Browsing (HTTP)

The web browsing service has been a key service in mobile networks as well as legacy fixes networks. The Hypertext Transfer Protocol (HTTP) protocol is adopted to support the web service at the application layer.

An experimental based model is proposed in the ITU-T recommendation G.1030 that the user satisfaction of web browsing depends on the user's expected session time which is context dependent. The browsing session time is the time duration between requesting a search page and requested data downloaded [IT14]. Based on many surveys, the recommendation constructs a model to predict the user satisfaction level in terms of MOS based on the session time. The formula between the user satisfaction and the session time is given as:

$$MOS = \frac{4}{\ln\left(\frac{Min}{Max}\right)} \cdot \left[\ln\left(SessionTime\right) - \ln\left(Min\right)\right] + 5$$

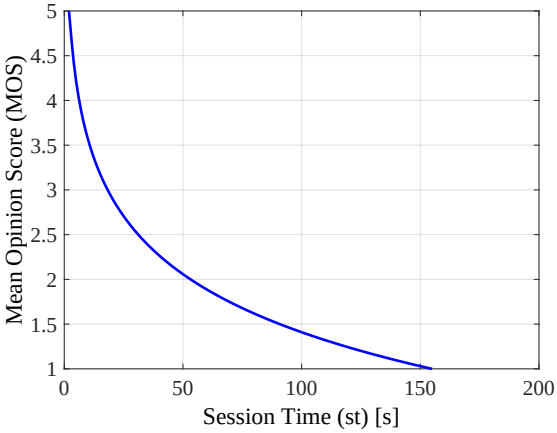


FIGURE 4.5: MOS over session time

The values Min and Max represent the minimum and maximum session time which depends on the experimental session context time. For example, the values are set to 2.16 seconds and 155 seconds when the user's expected session time is 60 seconds [IT14]. In addition, the MOS is ranged between 1 and 5. It is been observed that the MOS is a convex function of the session time from Figure 4.5. However,

the session time depends on the page size and average serving data rate. The session time has an inverse linear relationship with the data rate. Subsequently, the MOS is a concave function over the data rate for web browsing services. Some other QoE models for web traffic are proposed in [SFC10] using logarithmic and exponential methods. [KSM+13] conducts a subjective lab test and gives a summary of the existing assessment model for Web traffic. Figure 4.6 shows the curve fitting based on the lab test results provided in [KSM+13], showing the same concave property.

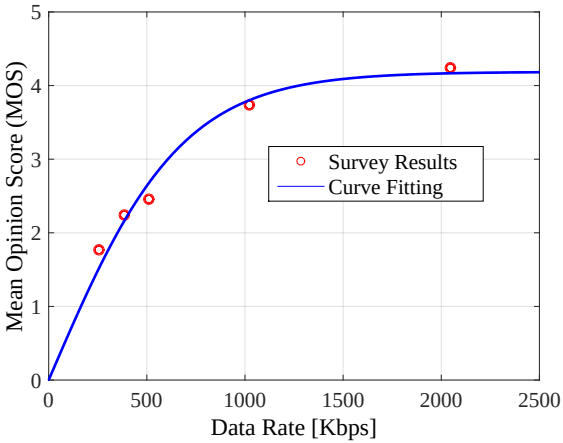


FIGURE 4.6: MOS over data rate

4.2.2.3 File Download (FTP)

File download, based on File Transfer Protocol (FTP) at the application layer, is typical elastic services. For the FTP services, the user's satisfaction purely depends on the average effective data rate, since the FTP file download time is proportional to the file size. In [LXZ+12], a QoE model of file download service is defined as:

$$MOS = \begin{cases} 1.0 & r < 8\text{kbps} \\ 2.5307 \cdot \log_{10}(0.3136 \cdot r) & 8\text{kbps} \leq r \leq 315\text{kbps} \\ 5.0 & r > 315\text{kbps} \end{cases}$$

The MOS has a logarithmical relationship with the data rate, truncated with the meaningful MOS range, which is between 1 and 5. Similar to video streaming and Web services, the concavity behavior of MOS over the data rate holds for the FTP traffic as well.

4.2.3 QoE-based Utility Functions Construction

From the QoE function analysis for video streaming, HTTP and FTP traffic, the QoE of these elastic services mainly depends on data rates and the MOS has a concave relationship with the data rates. In this part, the QoE-based utility functions are proposed to unify the mathematical formulation for different applications. Based on the proposed functions, the proper parameters can be easily found for various applications and user categories by the curving fitting method. Besides, owning a unified mathematical formulation allows us to transform the resource allocation problem to an optimization problem.

There are mainly two types of utility functions proposed in related works shown in Figure 4.7. For example, they are used in the study of utility-based resource allocation in wireless networks in [KL05], [KL08], and [CWC+11]. In both functions, the utility $u(r)$ is formulated as a function of the data rate r with a parameter α. The parameter α controls the shape of the curve. For instance, with a large α, the slopes of both functions are getting sharper.

Figure 4.7 shows the utility as well as their marginal utility curves. The function on the left is a typical concave function. With a larger α, the utility first increase faster, and then increase slower than with a smaller α. Correspondingly, it can be seen from the marginal utility

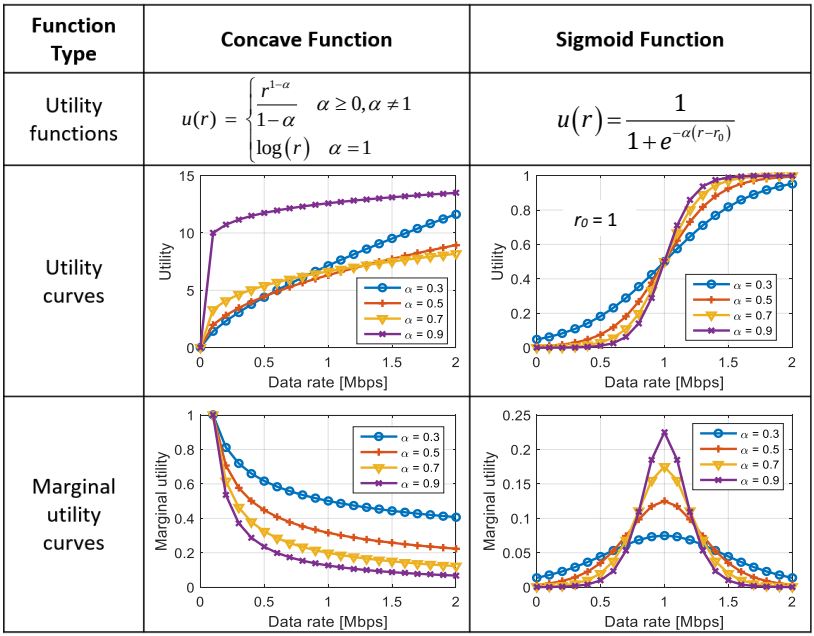


FIGURE 4.7: Summary of two utility functions

curves that, the marginal utility with a larger α is first higher and then smaller than that with a smaller α. When $\alpha < 1$, the utility is always positive, while on the opposite, the utility is always negative when $\alpha > 1$. However, the value of α cannot be set to 0 in the function $u(r) = \frac{r^{1-\alpha}}{1-\alpha}$ since the denominator cannot be 0. Nevertheless, when $\alpha \to 1$, the function is transformed into a logarithmic function by applying L'Hôpital's rule.

$$\lim_{\alpha \to 1} \frac{r^{1-\alpha}}{1-\alpha} = \lim_{\alpha \to 1} \frac{e^{\log(r)\cdot(1-\alpha)}}{1-\alpha} = \lim_{\alpha \to 1} \frac{\frac{d}{d\alpha}\left(e^{\log(r)\cdot(1-\alpha)}\right)}{\frac{d}{d\alpha}(1-\alpha)} = \lim_{\alpha \to 1} \frac{-\log(r)e^{\log(r)\cdot(1-\alpha)}}{-1} = \log(r)$$

The function on the right is a sigmoid function. The utility curves

are monotonically increasing. Nevertheless, there are two parts in the utility curves, which are separated by the critical point r_0. On the left side of the critical point, utility is a convex function while a concave function on the right side. Since a QoE-based concave utility function needs to be constructed for non GBR traffic, only the concave part of this sigmoid is used. The advantage of applying this sigmoid function is that the utility curve has a very good convergence property. Unlike the concave function on the left, the range of utility can go to infinity with increasing data rate; the sigmoid function converges to a center value with an infinitely large date rate. This property is critically important for building up a QoE based utility function since the MOS is limited to its maximum value 5. In the main part of this work, the concave part of the sigmoid function is adopted for further study. However the proposed methods and optimization models are feasible for any concave functions theoretically. Besides, the application with the left-hand concave function is studied in the Appendix B.

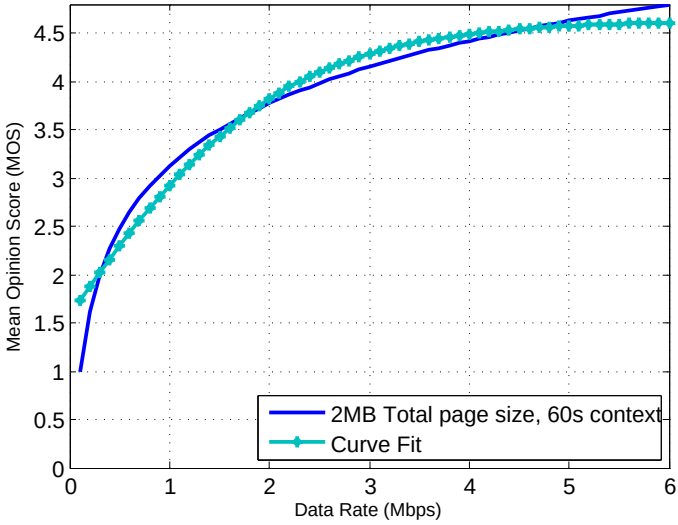
Since only the concave part of the sigmoid function is used, the critical point r_0 is set to 0. Besides, due to that the MOS is limited to the maximum value 5, two parameters A and D are introduced in the sigmoid function shown in eq. (4.1). The parameters α, A and D are determined by curve fitting according to the data obtained from the ITU-T documentations or experiments.

$$u(r) = \frac{A}{1 + e^{-\alpha \cdot r}} + D \tag{4.1}$$

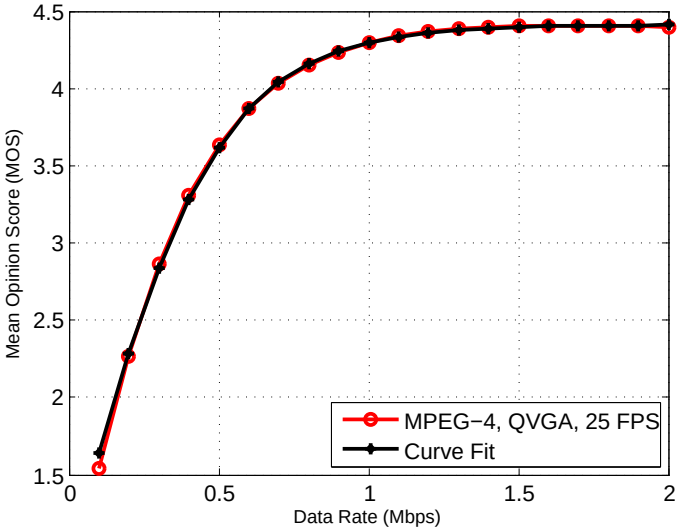
TABLE 4.3: Fitting parameter for Figure 4.8

	A	D	α	Norm of residuals
Web browsing	6.068	-1.435	0.9364	1.2022
Video Streaming	6.954	-2.542	4.104	0.1083

Figure 4.8 shows two examples of the curve fitting for web browsing and video streaming. It shows visually that with the sigmoid function,



(A) Curve fit of web browsing MOS model



(B) Curve fit of video streaming MOS model

FIGURE 4.8: Curve fitting for web and video streaming applications

by tuning the parameters A, D and α the curves are well fitted, especially for the video streaming curve. In addition, the goodness of curve fitting can quantified by the norm of residual which is expressed by:

$$\text{normofresiduals} = \sqrt{\sum_{i=1}^{n} d_i^2}$$

where n is the number of data samples and d is the difference between the original data and the value from the curve fit. The value of 0 means a perfect fit, so it is desired to have the norm of residuals as low as possible. The fitting parameters are shown in Table 4.3.

4.2.4 QoE-based Optimal Scheduling

In this section, a QoE-based scheduling framework for non GBR traffic is proposed. First, it is shown that the QoE-based radio resource allocation problem can be linearized without significant loss of accuracy. As a result, the problem is formulated as an optimization problem with the integer constant, since the LTE scheduler can only allocate an integer amount of PRBs to the individual users. It will be proven mathematically that the optimal resource allocation can be done in a simple iterative method, utilizing the concavity property of the QoE-based utility function. The method is optimal under two assumptions: no retransmission is considered; and the active users have enough data in the buffer to be transmitted in the current TTI. Then, a practical scheduler is designed based on the optimal allocation method that can be applied in a real system, taken the buffer information, the HARQ and other practical issues into consideration.

In the QoE function, the utility is a function of data rate. Nevertheless, the radio scheduler allocates the radio resource in terms of PRBs to its users every TTI. In LTE, the available bandwidth resources in terms of PRBs are scheduled to the active users periodically every TTI. A PRB, which has 180 KHz bandwidth, is the minimum resource

unit, which can be allocated to a user in LTE. The higher amount of the PRBs allocated to the user, the higher data rate the user can achieve. The user data rate depends on the MCS and the allocated number of PRBs. The allocated number of PRBs and MCS of a user are used to lookup the Transport Block Size (TBS). And the TBS is the amount of data that can be transmitted by the user in the current TTI. This gives the achievable throughput over the air interface, which is illustrated in Figure 4.9.

It can be seen from Figure 4.9, that the data rate has a near linear relationship with the number of PRBs, which identifies the allocated bandwidth to a user. Therefore, without significant loss of accuracy, the data rate r_i of user i is approximately proportional to the allocated bandwidth b_i, which can be described by the linear function $r_i = b_i \cdot \sigma_i$. The coefficient σ_i varies with the MCS. It is pre-computed for each MCS based on the data provided by 3GPP. Table C.1 in the appendix shows the curve fitting statistics including the R-square values, which measure how well the applied curves fit the input data.

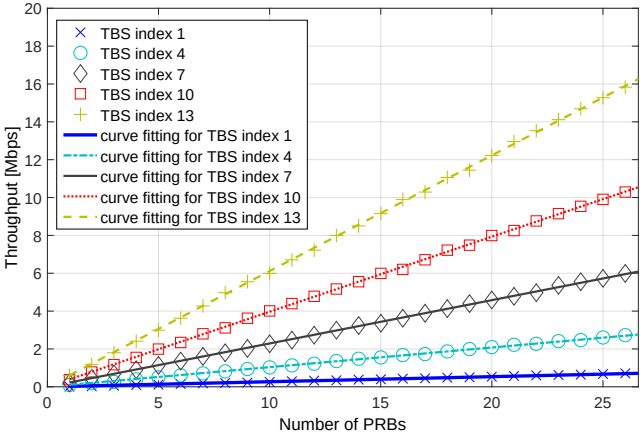


FIGURE 4.9: LTE throughput over number of PRBs

Therefore, the utility can be expressed as a function of bandwidth allocated to the user as $u(r_i) = u(b_i \cdot \sigma_i)$. The coefficient σ_i is a positive real number of user i in the current TTI, and therefore the achievable data rate r_i is directly proportional to the allocated bandwidth b_i. Consequently, the utility is a concave function of b_i given that the utility is a concave function of r_i. In LTE, the scheduler only allocates an integer amount of PRBs to the users, which means that the allocated bandwidth b_i can only be the non-negative integer multiple of 180 KHz. Considering this limitation, the utility is expressed as a function of the number of PRBs, and the problem is formulated as follows:

Problem I (4.2) is formulated as:

$$\max \left(\sum_{i=1}^{N} U_i(n_i) \right) \quad s.t. \quad \sum_{i=1}^{N} n_i = R \quad n_i \in Z^+ \tag{4.2}$$

with N is the number of users, R is the number of PRBs, n_i is a variable representing the number of resource blocks assigned to user i. U_i is the concave utility function of the user i depending on the number of PRBs the user gets.

Lemma: *The available PRBs are allocated to the user in an iterative manner. In each iteration, one available PRB is assigned to the user who has the maximum marginal utility, and the marginal utility of that user is updated for the next iteration. This method results in the optimal solution.*

Proof: Assume $\{N_1, N_2, ..., N_N\}$ is the optimal solution of the above problem, we have:

$$U_{\max} = \sum_{i=1}^{N} U_i(N_i)\Big|_{N_1+N_2+...+N_N=R} \geq \sum_{i=1}^{N} U_i\left(N_i{}'\right)\Big|_{N_1'+N_2'+...+N_N'=R} \quad \forall \left\{N_i{}'\right\} \tag{4.3}$$

Assume we have now $(R+1)$ PRBs, we have a new Problem II (4.4):

$$\max\left(\sum_{i=1}^{N} U_i\left(n_i\right)\right) \quad st. \quad \sum_{i=1}^{N} n_i = R+1 \quad \mathrm{r}_i \in Z^+ \tag{4.4}$$

Consider a feasible solution $\{N_1, N_2, ..., N_k+1, ..., N_N\}$ of Problem II with k is chosen such that:

$$U_k\left(N_k+1\right) - U_k\left(N_k\right) = \max_{i\in[1,N]}\left\{U_i\left(N_i+1\right) - U_i\left(N_i\right)\right\} \tag{4.5}$$

The resulting total utility is:

$$U = \sum_{i=1}^{N} U_i\left(N_i\right) + U_k\left(N_k+1\right) - U_k\left(N_k\right)$$

Consider another feasible solution of Problem II:
$\left\{N_1^{''}, N_2^{''}, ..., N_N^{''}\right\}\Big|_{\sum\limits_i^N N_i^{''}=R+1}$.

1. $N_k^{''} \geq N_k+1$ (user k gets equal or more than N_k+1 PRBs)
   $\left\{N_1^{''}, N_2^{''}, ..., N_k^{''}-1, N_N^{''}\right\}$ is a feasible solution of Problem I, because $N_1^{''}+N_2^{''}+...+N_k^{''}-1+...+N_N^{''} = R$. According to eq. (4.3),
   $\sum\limits_{i=1}^{N} U_i\left(N_i\right) \geq U_1\left(N_1^{''}\right) + U_2\left(N_2^{''}\right) + ... + U_k\left(N_k^{''}-1\right) + ... + U_N\left(N_n^{''}\right)$
   Furthermore, $U_k\left(N_k+1\right) - U_k\left(N_k\right) \geq U_k\left(N_k^{''}\right) - U_k\left(N_k^{''}-1\right)$ because U_k is a concave function, so its marginal utility decreases with increasing R. Therefore:
   $\sum\limits_{i=1}^{N} U_i\left(N_i\right) + U_k\left(N_k+1\right) - U_k\left(N_k\right) \geq U_1\left(N_1^{''}\right) + U_2\left(N_2^{''}\right) + ...$
   $... + U_k\left(N_k^{''}-1\right) + ... + U_N\left(N_n^{''}\right) + U_k\left(N_k^{''}\right) - U_k\left(N_k^{''}-1\right) \quad \forall\left\{N_i^{''}\right\}$

   $$\text{or} \sum_{i=1}^{N} U_i\left(N_i\right) + U_k\left(N_k+1\right) - U_k\left(N_k\right) \geq \sum_{i=1}^{N} U_i\left(N_i^{''}\right)$$

   $\{N_1, N_2, ..., N_k+1, ..., N_N\}$ is hence the optimal solution of the Problem II.

2. $N_k^{''} < N_k + 1$ (user k gets less than $N_k + 1$ PRBs)
   It exists q so that $N_q^{''} > N_q$ or $N_q^{''} \geq N_q + 1$. Since U_q is concave and has decreasing marginal utility, we have:

$$U_q\left(N_q+1\right)-U_q\left(N_q\right) \geq U_q\left(N_q^{''}\right)-U_q\left(N_q^{''}-1\right)$$

   According to eq. (4.5), we have:
   $U_k\left(N_k+1\right)-U_k\left(N_k\right) \geq U_q\left(N_q+1\right)-U_q\left(N_q\right) \geq U_q\left(N_q^{''}\right)-U_q\left(N_q^{''}-1\right)$
   And because $\left\{N_1^{''}, N_2^{''}, ..., N_q^{''}-1, N_N^{''}\right\}$ is a feasible solution of Problem I, following eg. (4.2), we have:
   $\sum_{i=1}^{N} U_i\left(N_i\right) \geq U_1\left(N_1^{''}\right)+U_2\left(N_2^{''}\right)+...+U_q\left(N_q^{''}-1\right)+...+U_N\left(N_n^{''}\right)$
   Therefore:
   $\sum_{i=1}^{N} U_i\left(N_i\right)+U_k\left(N_k+1\right)-U_k\left(N_k\right) \geq U_1\left(N_1^{''}\right)+U_2\left(N_2^{''}\right)+\ldots+$
   $U_k\left(N_k^{''}-1\right)+\ldots+U_N\left(N_n^{''}\right)+U_q\left(N_q^{''}\right)-U_q\left(N_q^{''}-1\right)$
   $\{N_1, N_2, ..., N_k+1, ..., N_N\}$ is hence the optimal solution of the Problem II.

This completes the proof of the lemma. To conclude, if there is one more available PRB, this PRB will be assigned to the user who has the maximum marginal utility given that the previous resource assignment is optimal. If the total number of resource blocks is 1 ($R = 1$), the optimal solution is trivial. The resource block is assigned to the user k such that: $U_k\left(1\right) = \max_i \left\{U_i\left(1\right)\right\}$. If $P = 2$, take the solution from $P = 1$, and assign the rest PRB to the user who achieves the maximum marginal utility. If $P = n$, take the optimal solution when $P = n - 1$, and assign the last PRB to the user who gets the maximum marginal utility.

4.2.5 QoE-based Scheduler Design

The QoE-based radio resource allocation method above is proven optimal under full buffer assumption. In this section, a practical scheduling

framework is proposed based on the optimal allocation method considering the retransmission and user buffer status. The scheduler is implemented in the simulator for performance evaluation.

4.2.5.1 Scheduling Procedures

Figure 4.10 shows the flow chart of the proposed QoE-based scheduler. The scheduling is performed periodically every 1 ms in each cell. The retransmission users are granted with highest priority. If any resources left, the resources are allocated to the active users in an iterative manner. This iterative method was proven optimal given full buffer assumption and without transmission error in Section 4.2.4. The detailed scheduling procedures are summarized as follows:

1. Initialize the candidates' lists. They are inserted into two lists:
   (a) HARQ list for the users with pending retransmission, which means the users experienced transmission errors 8 TTIs earlier;
   (b) new user list for the active users, which have packets/segments in PDCP/RLC buffers.
2. The retransmission users are granted with highest priority. Due to chase combining, the users will be allocated with the same amount of PRBs as they got for the last unsuccessful transmissions.
3. If there are any PRBs left, the new users in list b are taken into consideration for scheduling. The effective SINR for each user i is calculated based on the EESM SINR mapping method. Therefore, the MCS is determined correspondingly and the MCS dependent coefficient σ_i is obtained. The marginal utility for each user i by getting a PRB is calculated by:

$$u_i = u_i\left(n_i + 1\right) - u_i\left(n_i\right)$$

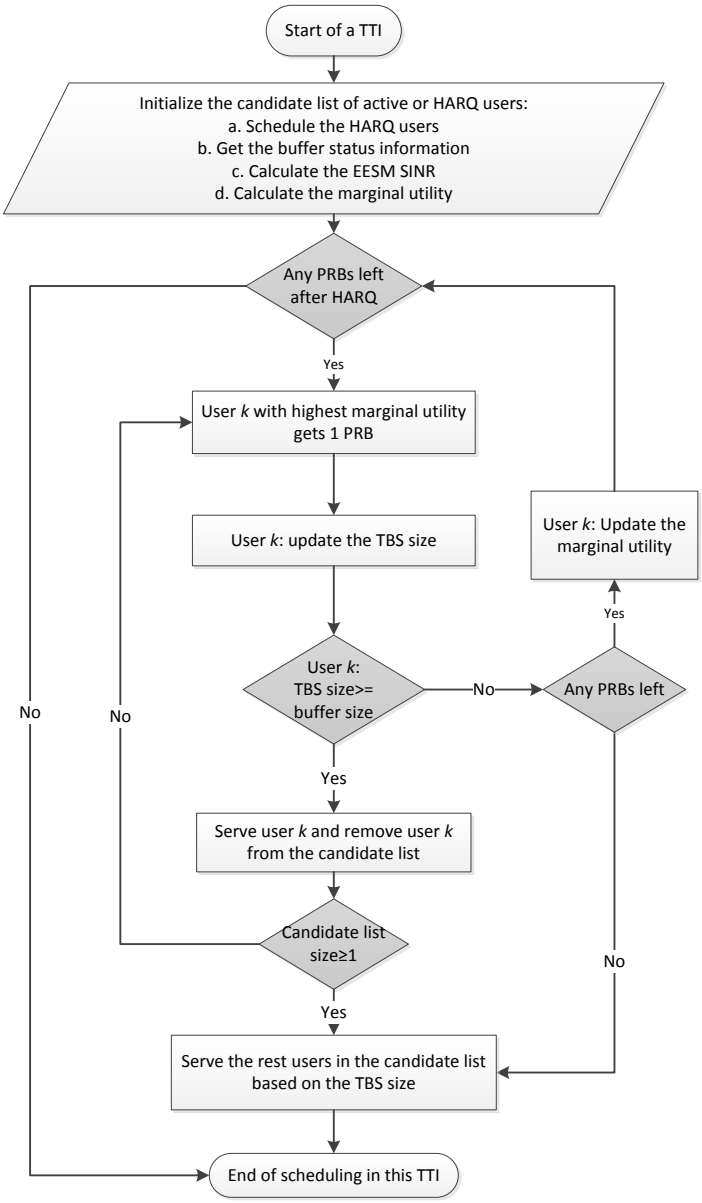


FIGURE 4.10: Flow chart of the proposed QoE-based scheduler

where n_i is initialize as 0 for all the users. In this work, the default utility function is $u_i(r_i) = \frac{A}{1+e^{-\alpha_i \cdot r_i}} + D = \frac{A}{1+e^{-\alpha_i \cdot b_i \cdot \sigma_i}} + D$. So it is initialized as:

$u_i = \left(\frac{A}{1+e^{-\alpha_i \cdot 180KHz \cdot \sigma_i}} + D\right) - \left(\frac{A}{1+e^{-\alpha_i \cdot 0 \cdot \sigma_i}} - D\right) = \frac{A}{1+e^{-\alpha_i \cdot 180KHz \cdot \sigma_i}} - \frac{A}{2}$

4. The user k with the highest marginal utility will get one PRB. The variable n_k indicating the number of PRBs allocated to user k is incremented by 1. The TBS of user k is updated and compared with the total buffer size (PDCP+RLC layer buffer).

   - If the current TBS can empty the buffer in this TTI, then the user k is served with the current TBS. Afterwards, the user k will be removed from the list b, which means that the radio resource allocation process for the user is finished. If there are any users still in the candidate list, then go to step 5; otherwise, go back to step 4;
   - In case that the current TBS cannot empty its buffer, if there are no available PRBs left, go to step 5; otherwise, the marginal utility of user k is updated and then go back to step 4.

5. Serve the rest users in the candidate list based on the TBS size. The scheduling process for the current TTI is finished.

In this work, the main target of the designed scheduler is to maximize the aggregated user QoE that all the active users with different services are treated equally. Nevertheless in reality, the network operators should support service differentiation over different user categories, e.g. platinum users, business users, economy users and so on. For instance, the business and platinum users are supposed to get better QoS and QoE than economy users since they are willing to pay more for getting premium services and therefore they can contribute a higher profit to the network operators. In addition, even with the same user category, the different services may be treated differently by the operators.

For example, the operators might be favorable to deliver a higher user experience for the video streaming and HTTP services other than FTP services on purpose or vise versa. For these reasons, a weighting factor is proposed to be added in the utility function as shown in the utility function eq. (4.6). Since the weighting factor is a constant value for individual user with specific services, the new utility function can be used directly in the proposed QoE based scheduler. Nevertheless, the target of the designed scheduler is changed to maximize the aggregated utility instead of aggregated QoE. The application of this utility function can be seen as an extension of this work, which is studied in the simulation scenario 4 in Section 4.3.3.

$$u_i = w_i \cdot \left(\frac{A}{1 + e^{-\alpha_i \cdot r_i}} + D \right) \tag{4.6}$$

4.3 Simulation Scenarios and Results

In this section, the performance of the proposed QoE based scheduler is studied by simulations. A legacy macro eNB/cell scenario which hexagonal coverage (see Section 3.3.3) is in use for evaluation. Both GBR and non-GBR traffic is taken into consideration. The VoIP application which belongs to GBR traffic is scheduled with absolute priority over non GBR traffic, including video streaming, HTTP and FTP. The proposed QoE based scheduler is supposed to achieve a better aggregated QoE by more intelligently allocate radio resources among the non GBR users considering their QoE requirements. A typical Proportional Fair scheduler [Zak12] is used for the comparison purpose. A random waypoint mobility model is applied to all the users to model the user movements. All scenarios are simulated with multiple seeds and considering the warm up period to generate confident simulation results. A summary of simulation settings is shown in Table 4.4.

Table 4.5 shows the detailed scenario settings. The study starts with a proof-of-concept scenario which only contains a video streaming,

TABLE 4.4: Simulation system settings

Parameter	Settings
Macro eNBs settings	Pathloss: 130.5 + 37.6log10(R), R in Km [IR09] Slow fading: Correlated Log normal, zero mean, 8db std. and 50 m correlation distance Small scale fading: 3GPP Pedestrian A Transmission power: 23dBm per PRB
TCP version	New Reno with 64Kbytes receiver buffer size
Traffic types	VoIP: GSM EFR, codec rate 12.2 kbps Video Streaming: TCP based full buffer streaming HTTP: 2MB page size FTP: 10MB file size
Mobility model	5Km/h,Random Direction
Number of PRBs	25 PRBs (5MHz spectrum at 2.6 GHz)
Simulation time	1000s (5 runs with different seeds) with warm up period of 300s

a HTTP and an FTP user with full buffer mode (idle time is set to 0). In the scenario 2 and 3, there are 10 VoIP users, 3 video streaming users, 3 HTTP users and 3 FTP users. The video streaming users are supposed to watch a long video and therefore are active all the times. In scenario 2, the HTTP and FTP users have no reading time. This means the next HTTP or FTP request starts immediately after the ending of last request. On the contrary, in the scenario 3, the HTTP and FTP users have pause/reading times between two consecutive file transfers (see Section 3.3.5.2). Because in reality, the users may spend some time on reading the website or processing the last download file. Comparing to the scenario 2, the traffic load of HTTP and FTP users is reduced. Since the users are not active all the times and introduce a higher randomness, it is meaningful to examine the performance of the proposed QoE based scheduling strategy under such circumstances. In

Table 4.5: Scenario settings

Scenario 1 Proof-of-concept	1 Video,1 HTTP, 1 FTP All active all the time
Scenario 2	10 VoIP; 3 Video, continuous streaming; 3 HTTP, 2MB page size, 0s reading time; 3 FTP, 10MB file size, 0s reading time;
Scenario 3	10 VoIP; 3 Video, continuous streaming; 3 HTTP, 2MB page size,50s reading time; 3 FTP, 10MB file size,50s reading time;
Scenario 4	10 VoIP; Bussiness users: 1 Video, 1 HTTP, 1 FTP; Economy users: 1 Video, 1 HTTP, 1 FTP; HTTP: 2MB page size, 0s reading time; FTP: 10MB page size, 0s reading time;

the scenario 4, the service differentiation over both different traffic types and user categories are evaluated. Two user categories are considered that business users are given a higher weighting factor than economy users. There are a video streaming user, a HTTP user and an FTP user in each of the user category.

4.3.1 Scenario 1 - Proof-of-concept Scenario

There is only a video streaming, a HTTP and an FTP user with the full buffer mode in this scenario. The performance of the proposed scheduling framework is compared against the legacy Proportional Fair scheduler. The purpose of this scenario is to prove the implemented scheduler is functional as expected. The QoE based utility curves and

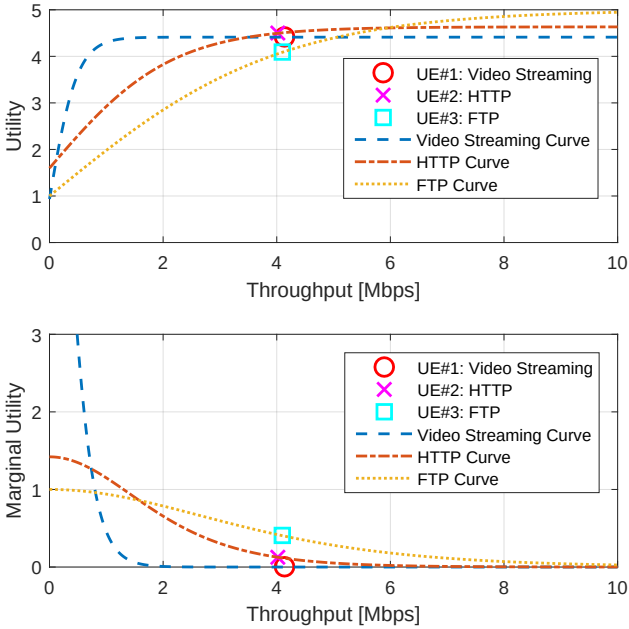


FIGURE 4.11: Performance of the proposed PF scheduler for scenario 1

their marginal utility curves for all three types of services are plotted in Figure 4.11 for proportional fair (PF) scheduler and Figure 4.12 for the proposed QoE based scheduler. The average downlink throughput of each user is marked in the corresponding curves.

If the simulation time is long enough, the average channel conditions for all the users tend to be the same due to random mobility. Therefore, in a legacy PF scheduler, all the users have similar throughput if there is no service differentiation considered. This behavior is observed in the simulation results shown in Figure 4.11. With similar throughput, the video streaming and HTTP users have a slightly higher QoE values than the FTP user. Nevertheless, the FTP user has a higher marginal utility. On the contrary, with the proposed QoE

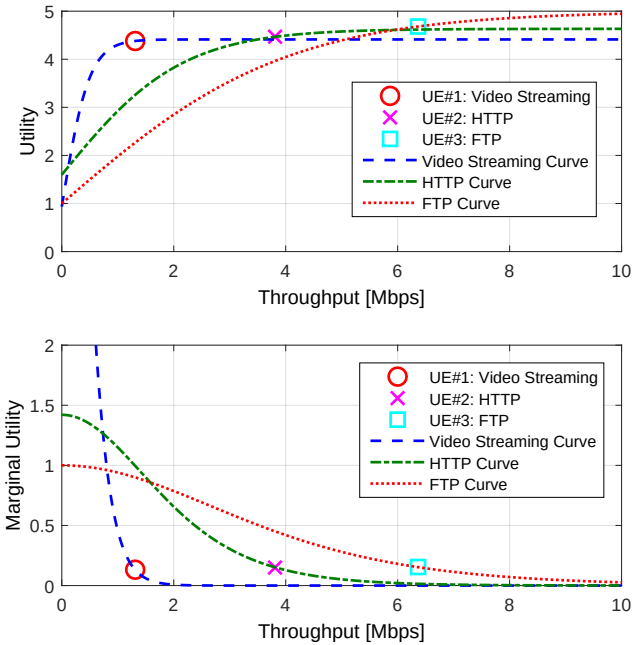


FIGURE 4.12: Performance of the proposed QoE based scheduler for scenario 1

based scheduler, all three users achieve very different throughput, but have a similar marginal utility value. Comparing to the PF scheduler, The HTTP user throughput nearly stays the same, but FTP user has a much higher throughput which is taken over from the video streaming user. However, the utility is nearly unchanged for the video streaming user because the user is already satisfied with a low throughput at around 1.5 to 2 Mbps. Allocating more resources to the video streaming user does not further improve its utility. This can be seen from its marginal utility figure that its marginal utility converges to 0 with a fastest speed. Therefore there is no margin to further improve the

utility when the marginal utility is very small. Among these three applications, the marginal utility of the FTP user converges to 0 with a slowest speed. The FTP user would benefit most with a throughput higher than 2 Mbps. Therefore, the proposed QoE based scheduler allocates most resources to the FTP user and least to the video streaming user. With a little utility degradation of the video streaming user, the utility of FTP user gets a higher gain. Consequently, the aggregated utility (QoE in terms of MOS) is improved by the proposed QoE based scheduler.

The principle of the QoE based scheduler is to allocate the resource to the user with highest marginal utility in an iterative matter in each TTI. Therefore theoretically, all the users would have the same marginal utility. This property can be used to calculate the needed amount of PRBs with certain amount of active users and known traffic types. In this scenario, if there are more PRBs available, e.g. with more cell bandwidth to 10 MHz, it is expected that the FTP user's throughput will increase most while the video streaming user the least. On the contrary, if there are less PRBs available, e.g. with cell bandwidth to 1.4 MHz, the utility of the video user will be least impacted due to its high marginal utility at a low throughput. Both throughput and utility of the HTTP and FTP users would be decreased substantially. As a result, when there are limited resources available, the video streaming user will get most of the resources since it can get higher utility with less amount of resources. The HTTP and FTP users will get more resources with higher amount of resources. With the proposed QoE based scheduler, it differentiates the users automatically with the purpose of maximizing the aggregated utility. With the legacy weighted PF scheduler, the service differentiation is fixed according to the weight settings without awareness of QoE.

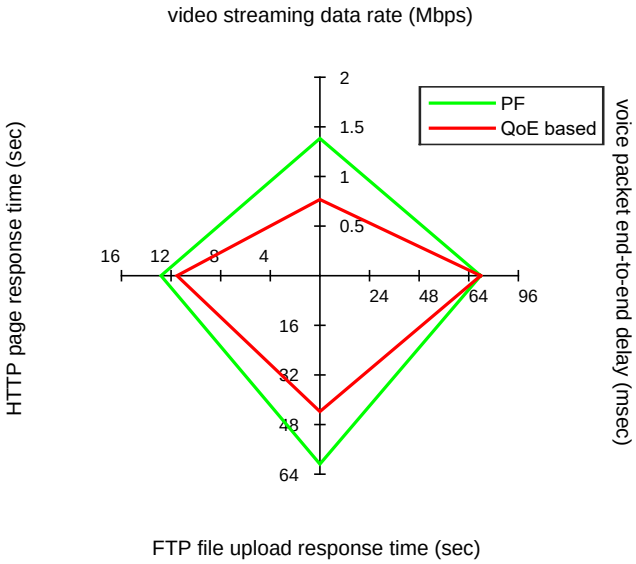


FIGURE 4.13: Application performance comparison for scenario 2

4.3.2 Scenario 2 and 3 - Varying Traffic Load

Figure 4.13 and 4.14 show the application performance and their corresponding MOS (the error bars show the 95% confidence interval) for scenario 2. Since the VoIP users only consume very little amount of resources (12.2 Kbps codec rate at the application layer) and they are prioritized over the whole network with absolute priority over non-GBR traffic, the voice packet end-to-end delays (including coding, decoding and de-jittering delay) are the same for both schedulers. With respect to the non GBR traffic, similar to scenario 1, the throughput of video streaming users is reduced dramatically from 1.38 Mbps down to 0.77 Mbps with the QoE based scheduler. More resources are given to the HTTP and FTP users. It can be seen from the figure that the average HTTP page response time is reduced from 12.8 seconds to 11.5 seconds. The average FTP download time is reduced greatly from 60.7 seconds

to 43.8 seconds so that the download time is reduced by reduced 30%. Correspondingly, the MOS of the FTP and HTTP users are improved significantly with the QoE based scheduler. Although the throughput of the video users is nearly halved, the MOS of the video streaming users nearly remain at the same level (reduced by 0.03). The reason is that the video MOS has a non-linear relation with the data rate. Thus the video users can remain satisfied with even less amount of resources. Therefore, the aggregated non-GBR user satisfaction in term of MOS is increased from 28.2 to 29.9 as shown in Figure 4.14.

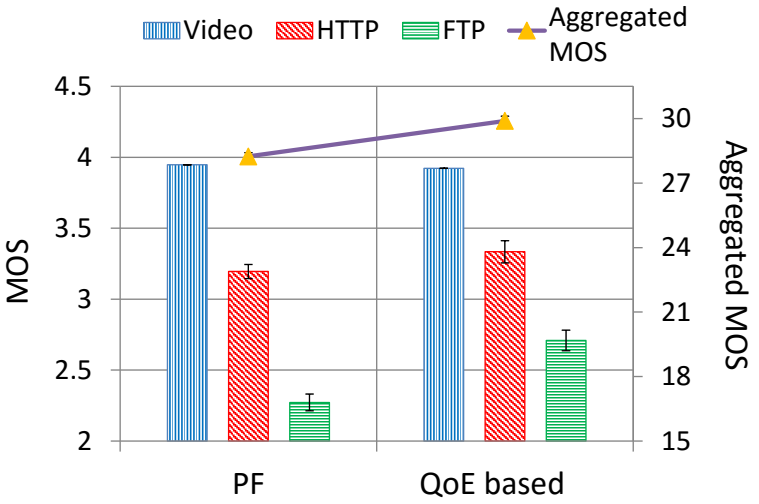


FIGURE 4.14: MOS comparison for scenario 2

Considering the minimum MOS value (scaling between 1 and 5) of each user is 1, the gain (G) is around 8.9% calculated based on eq. (4.7). MOS_{UT} and MOS_{PF} represent the aggregated MOS for all the non GRB users with the QoE based scheduler and PF scheduler, and N represents the total number of non GBR users.

$$G = \frac{(MOS_{UT} - N \cdot 1) - (MOS_{PF} - N \cdot 1)}{(MOS_{PF} - N \cdot 1)} \cdot 100\% = \frac{MOS_{UT} - MOS_{PF}}{(MOS_{UT} - N)} \cdot 100\% \tag{4.7}$$

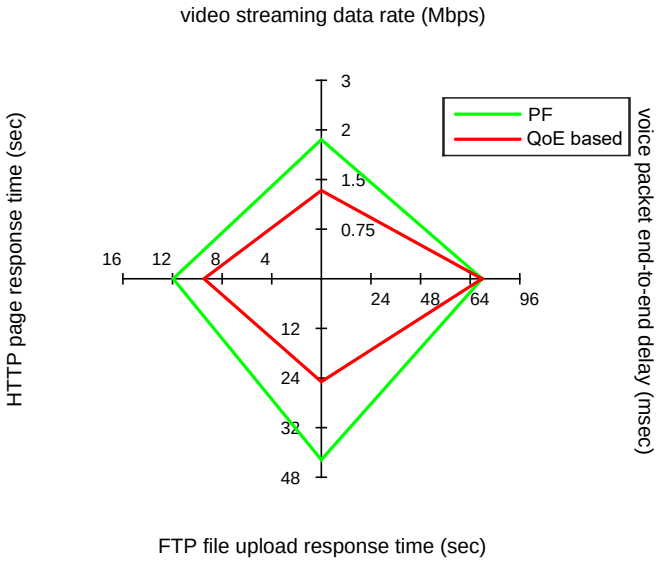


FIGURE 4.15: Application performance comparison for scenario 3

As explained, the HTTP and FTP users may spend some time on reading the website or processing the last downloaded file. The reading time is set to 50 seconds for both HTTP and FTP users in scenario 3 instead of 0 second in scenario 2. Therefore, the traffic load of HTTP and FTP users are reduced while the video streaming users remain the same. In another word, the traffic share of video streaming users is increased. Figure 4.15 and 4.16 show the application performance and their corresponding MOS for scenario 3. Same as in the previous scenarios, the gain comes from allocating more resources to the HTTP and FTP users while suppressing video streaming users' throughput. Since the traffic share of HTTP and FTP users is reduced, the gain

of the QoE based scheduler over PF scheduler is increased to 17.3% comparing to 8.9% in scenario 2. The HTTP page response time is reduced from 11.9 seconds to 9.4 seconds and the FTP file download time is nearly halved from 45.6 seconds to 25.9 seconds. Accordingly, the MOS value of FTP users is increased from 2.6 to 3.6 which means the user satisfaction level is increased from fair to good. Based on the comparison of the results between scenario 2 and 3, the conduction is proven that when there is higher traffic share of video streaming, the gain is higher. In principle, the more the marginal utility differs among the users by a PF scheduler, the higher room it is for the QoE based scheduler by more intelligently allocating the resources, and therefore a higher gain is expected.

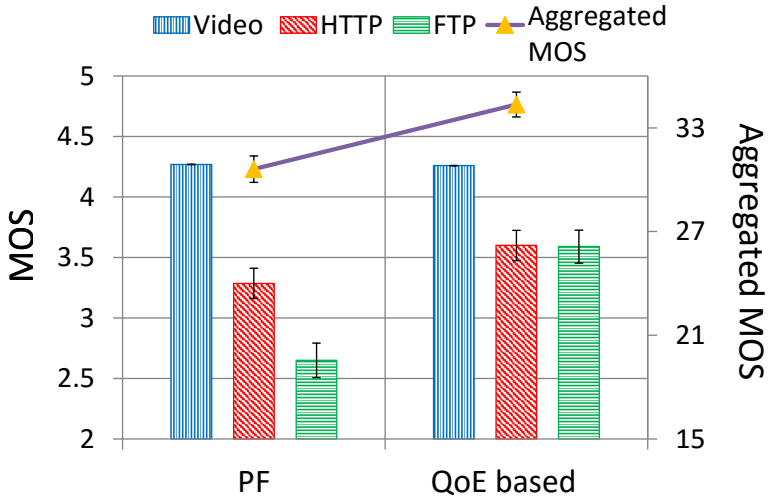


FIGURE 4.16: MOS comparison for scenario 3

4.3.3 Scenario 4 - Different User Category

In scenario 4, the service differentiation over both different traffic types and user categories are evaluated. There are a video streaming user, a

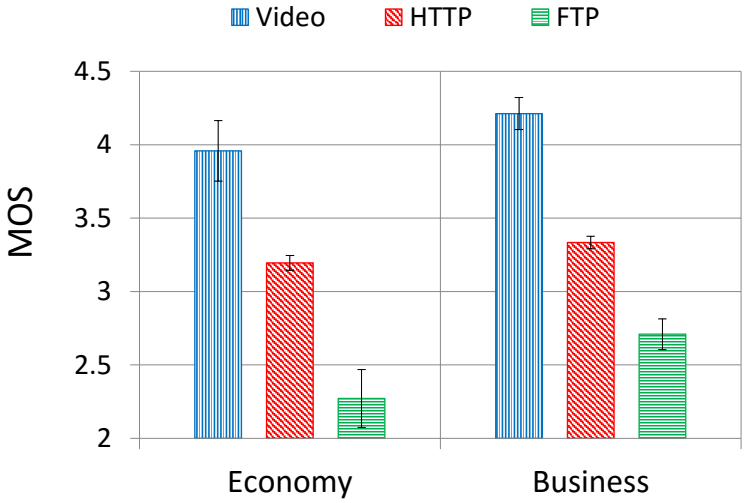


FIGURE 4.17: MOS comparison for scenario 4

HTTP user and an FTP user in each of the user category (3 business users and 3 economy users). The weighting factor w of business users are set to 2 while the economy users are set to 1. For the users with the same traffic type, both the utility and marginal utility are doubled for the business users over economy users with a same rate. Therefore, the business users are prioritized over the economy users. The HTTP page response time and FTP file download time is 10.9 and 38.7 seconds respectively for economy users while they are 7.4 and 22.6 seconds for business users. The video streaming throughput is 0.78 Mbps for economy user and 1.04 Mps for business user. The gains of business user over economy user range from 25% for the video streaming users to 41% for the FTP users, and 31% for the HTTP user. The service differentiation between the business and economy users is quite obvious. Nevertheless, the differences are not as high as 2 times, which is the weight difference. The reason is that the QoE based scheduler always try to maximize the aggregated utility and due to the concavity, the

double throughout does not give the doubled utility.

4.4 Summary

This chapter focuses on the design and implementation of an optimal utility-based radio scheduling considering QoE in LTE. The target of the scheduler is to maximize the aggregated QoE in a cell for elastic traffic whose QoE, which is measured by MOS, can be approximated by a concave relation over the data rate.

From the QoE function analysis for video streaming, HTTP and FTP traffic, the QoE of these elastic services mainly depends on data rates and the MOS has a concave relationship with the data rates. Therefore, concave utility functions are used to unify the mathematical formulation for different applications. Afterwards, a QoE based scheduler is proposed and proven to be optimal analytically. The simulation results show that the proposed QoE base scheduler framework significantly outperforms the conventional proportional fair scheduler.

Chapter 5

Joint Radio and Transport Optimized Resource Management

As discussed in Chapter 1, the transport backhaul is not optimal and can be a bottleneck for many typical scenarios. Usually, the operators do not own dedicated transport backhaul network, connecting the Radio Access Network (RAN) and the Evolved Packet Core (EPC). They need to lease bandwidth from 3^{rd} party IP service providers to get well managed transport network services for the LTE mobile backhaul (MBH). They normally lease the transport networks from 3^{rd} party Internet Service Providers (ISPs) with well-defined Service Level Agreements (SLAs). In order to fulfil the SLAs, the transport backhaul should be dimensioned by the mobile operators considering the traffic carried by the network and the QoS requirements [LLT+12]. In legacy mobile networks, the transport backhaul capacity is not been dimensioned to handle all the peak scenarios with the purpose of saving

CAPEX and OPEX. Therefore, in the case that multiple cells experienced high loads at the same time, the transport backhaul could be overloaded as well. More importantly, the femtocell networks use cable or broadband xDSL as the last mile solution and usually its bandwidth is limited by the contract.

As it is shown in Figure 5.1, the LTE access network contains two main bottlenecks: air interface and transport backhaul. On one hand, the transport backhaul could be limited as stated above. On the other hand, the radio capacity is limited by the available cell spectrum and user channel conditions at the air interface. The radio spectrum resources are scare and very expensive to use. The German government auctioned a total 358.8 MHz spectrum mainly used for LTE networks for 4.4 billion euro in 2010 [Ger10]. In reality, it might happen that some cells may be congested while the others not. For example, in this example scenario, both cell A and the aggregated transport backhaul are limited. If there is no mechanism to control the rate of users in congested cells, the cells may get more congested, while users in other cells do not get enough data rates. Furthermore, users may have poor QoE, because the data rate that the users get is not shaped according to their application types, but often heavily influenced by the transport protocol behavior like TCP and UDP.

It is motivated to optimize the resource allocation in terms of deciding optimum bearer rates for different users dynamically according to their application types as well as traffic variations and available network capacities of both radio and transport networks. As stated in Chapter 1, the innovation of this work is an algorithm that dynamically optimizes the EPS bearer rates for all users in the LTE networks considering the limitation of radio and transport resource with the objective of maximizing the accumulated users' QoE.

In this chapter, the EPS bearer shaping method as well as other existed resource management schemes at LTE transport networks are discussed first. Then, the typical scenarios are introduced, which the

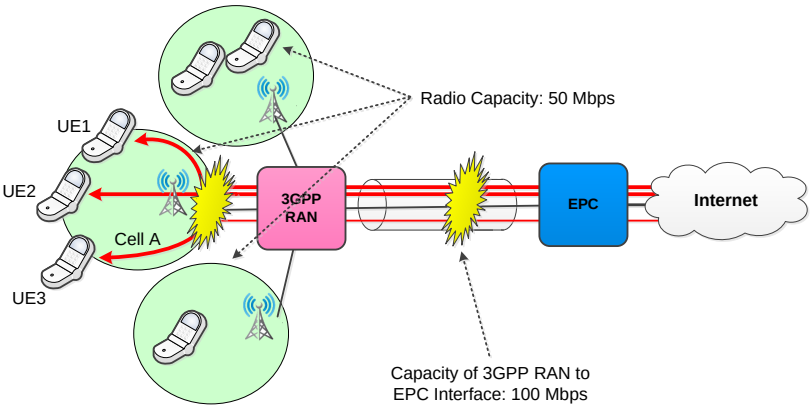


FIGURE 5.1: Performance limited by both cell and transport network capacity (example capacities) [3GP12]

proposed joint radio and transport optimized resource management schemes are applicable, e.g. femtocell cluster, base station with multiple cells and Cloud-RAN in LTE-advanced, are introduced. The resource allocation problem in multiple cells considering transport network limitations is formulated as a convex optimization problem, which maximizes the aggregated QoE, represented by the Mean Opinion Score (MOS), of all users using different applications, and then solve it using the Lagrangian relaxation method. The EPS bearers are shaped according to their optimal rates for all users at the core network. The performance of these proposed algorithms is investigated and evaluated by simulations. Moreover, a discussion on how often to adjust the shaping rates is given based on the complexity and performance investigations.

Although the optimal bearer shaping method can be generalized for LTE networks, the performance evaluation focuses only on the femtocell scenarios. However, a large scale network often serves hundreds of users and solving the optimization problem may not be feasible. In order to apply the proposed approach in large scale networks, efficient

heuristic algorithms are developed with reduced complexity and evaluated by simulations.

5.1 Resource Management Schemes at LTE Core and Transport Networks

5.1.1 IP per-bearer Traffic Shaping

In the transport backhaul network connecting the Radio Access Network (RAN) and Evolved Packet Core (EPC), as shown in Figure 5.1, traditional traffic management schemes (e.g. packet scheduling, queue management) are often applied on the packet level (i.e. link layer). In addition to that, the resource allocated to each user or an application flow in the transport network can also be controlled by a bearer shaping mechanism on the flow level (i.e. IP layer). In has been introduced in Chapter 2 that, 3GPP has defined a QoS framework based on the Evolved Packet System (EPS) bearer model to manage various traffic classes with different quality of service requirements. A bearer is a traffic separation element that enables differentiated treatment of traffic based on its QoS requirements, and provides a logical path between UE and the gateway located in EPC. Each bearer will be assigned to a certain data rate, according to the QoS parameters (e.g. Guaranteed Bit Rate - GBR and Maximum Bit Rate - MBR for the GBR type bearers, and Aggregate Maximum Bit Rate-AMBR for the non-GBR type bearers). Currently, the bearer rates are simply set by the mobile network operators or according to the subscriber contracts. Typically the maximum bearer rate is fixed and hardly achieved, which leads to none optimal resource utilization as well as poor users' perceived QoE. This problem becomes more severe in the LTE femtocell or picocell networks.

Figure 5.2 shows the overview of the downlink traffic shaping per bearer level at the access gateway (aGW) in EPC. The incoming packets

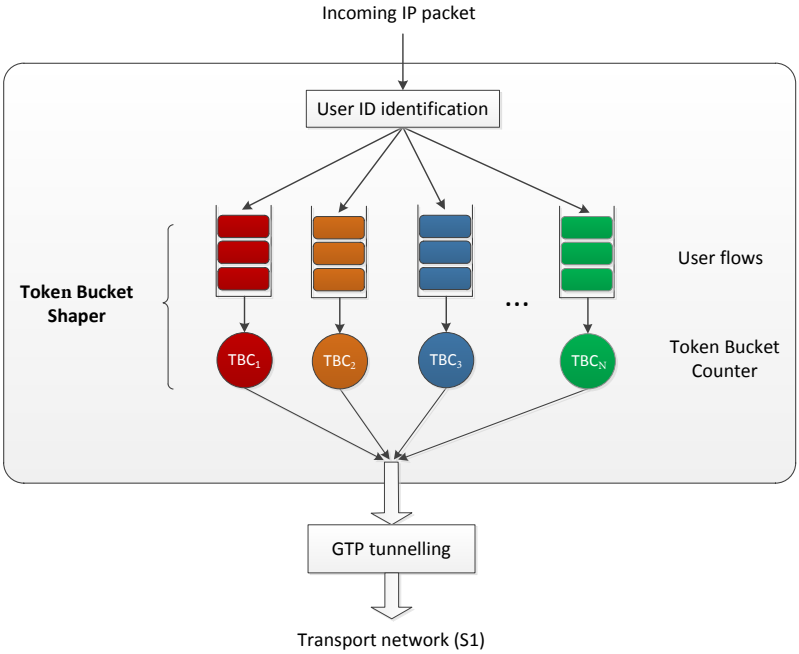


FIGURE 5.2: Token bucket traffic shaping

are sorted by identifying the user and bearer ID and then are inserted into corresponding token bucket shaper buffer. Each bearer has its separate buffer and shaper based on the token bucket algorithm [Tan02]. The bearer token shaping rates are set according to their corresponding optimal shaping rates which are calculated in Section 5.3. After shaping, the traffic of each bearer is aggregated and then carried through the transport network.

It is worth to mention that the proposed algorithm in this work can be used as an online function to dynamically change the optimal bearer shaping rate over time (periodically or on-demand). On one hand, the user channel conditions are changing over time due to fast fading and mobility. On the other hand, the users may become active or inactive

over time. Therefore, the optimization problem is being solved and the optimal shaping rates are updated over time. By default in this work, the update interval is 1ms which is 1 TTI (Transmission Time Interval) in LTE. However, other update intervals are studied in this work (discussed in Section 5.4.5).

5.1.2 Transport Scheduler and Shaper

As introduced in Section 4.2.1, there are nine different types of bearers, identified by QoS Class Indicator (QCI) in LTE. High priority real time traffic flows, e.g. voice/video telephony and real time gaming, are grouped into GBR bearers. Non-GBR bearers mainly serve non real time services with lower priority compared to GBR bearers. In order to provide required QoS over the transport networks, Differentiated Services (DiffServ) is enabled by mapping different bearers to corresponding PHBs (Per Hop Behavior) , which define the packet-forwarding properties associated with a class of traffic.

For instance, the conversational applications including voice, video conferencing and video games (QCI 1-4) are mapped to Expedited Forwarding (EF) PHBs with highest/stric priority since they requite low packet loss, delay and jitter. The applications with QCI 6-8 are mapped to Assured Forwarding (AF) PHBs, e.g. interactive gaming, buffered streaming, web browsing. AF PHBs have 4 priority classes (AF 1x to AF 4x), each class with 3 levels of dropping precedence: low, medium and high (AF x1 to AF x3). Different PHBs are used for different RAB types providing various QoS levels to end users in practice. The applications with the lowest priority are classified as Best Effort (BE) traffic and mapped to the BE PHB. To be noticed, the mapping of QoS classes to PHBs can be configured alternatively by operations depending on the application types and user categories.

Figure 5.3 shows an IP DiffServ architecture used in this work. The transport scheduler combines Strict Priority Queueing (SP) and Weighted Fair Queueing (WFQ) [FPR00], [TNC+01]. SP is used to

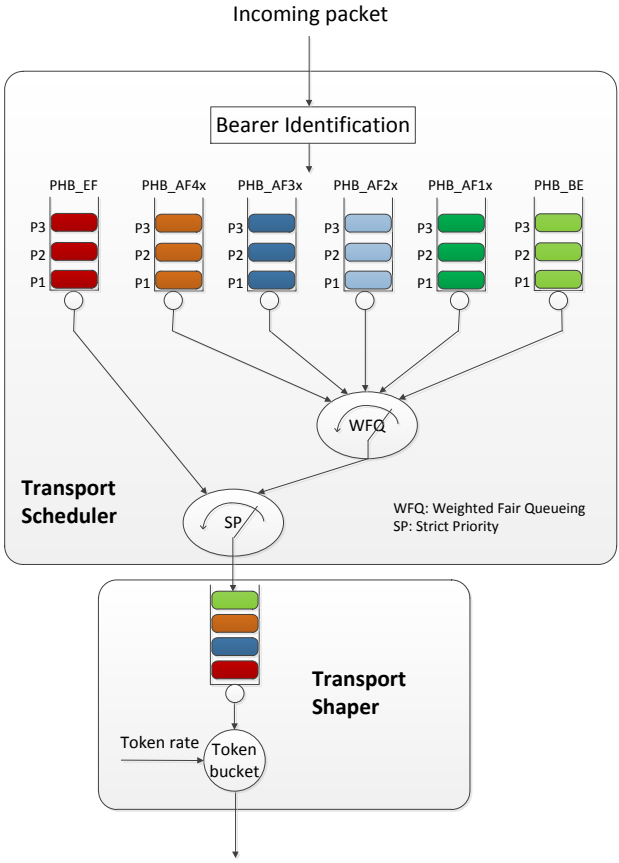


FIGURE 5.3: Weighted fair queueing scheduler and shaper

provide assured premium service to the EF PHB. When there is spare bandwidth left after serving the EF PHB, AF and BE PHBs are served in a weighted fair manner by setting different weights to different PHBs. At the egress port, the aggregated traffic after transport scheduler is shaped to a predefined data rate using Token Bucket Shaper. To be noticed, packet discarding mechanisms can be applied to PHB buffers

or the shaper buffer, e.g. WRED (Weighted Random Early Detection) [Li09].

5.2 Introduction and User Cases

5.2.1 Femtocell Clusters

In mobile communications, it is an acknowledged fact that most of the mobile data is generated indoors. More than 60% of voice calls and 90% of data traffic take place in indoor environments as pointed out by a recent survey [Eri15]. Usually the conventional solution for increasing network coverage is adding more network infrastructure such as tower cells or transmitters. However, the deployment of these macrocell elements can be very costly and increase the inter-cell interference in the network. With increasing number of indoor mobile users, operators have to find a better cost efficient solution to provide necessary Quality of Service (QoS) to the users. The latest trend to solve this problem is to introduce a large number of femtocells, which can be connected to the operator's network by using existing user broadband (e.g. xDSL/Cable/Fiber) connection, for indoor coverage. With this approach, operators can save costs for deploying more macrocells and offload some traffic from macrocell to femtocells. This can lead to a decrease in CAPEX/OPEX of operators and makes femtocell deployment a really lucrative option [ZHY11]. An increase in user satisfaction is also expected due to possible better indoor reception and longer battery life because of low transmission power levels [ZZWS13].

The number of users and the traffic that they generate increase extremely fast, while the bandwidth resource is very limited. Therefore, in order to guarantee a certain Quality of Service (QoS) and more importantly, to improve the users' perceived Quality of Experience (QoE), it is necessary to have an efficient resource management mechanisms.

In the LTE femtocell networks, a femtocell or multiple femtocells (a femtocell cluster), each serve multiple users, use broadband xDSL as the last mile transport and usually its bandwidth is limited by the contract, e.g. with a maximum of 16 Mbps data rate. Hence, the last mile is usually the bottleneck in the network.

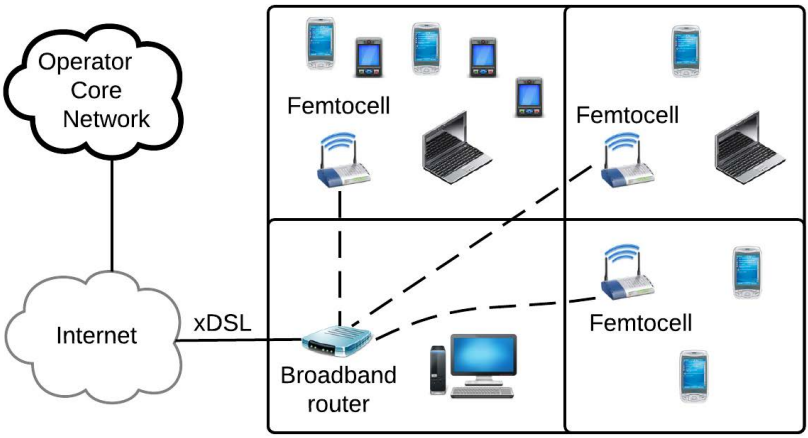


FIGURE 5.4: Example of a femtocell cluster

However, deploying femtocells raises many new challenges, one of which is the resource allocation for users serving by femtocells. Different from macrocells, whose transport network is usually appropriately dimensioned, the one of femtocells is bounded to the existing user's broadband (e.g. xDSL). Let's consider a scenario, in which multiple LTE femtocells share one transport network shown in Figure 5.4. Some cells are heavily loaded while the others are lightly loaded. Furthermore, the data rate of the transport network is lower than the total throughput of all cells. If the resource allocation is carried out independently at each cell without considering the transport network limitation, users in low loaded cells will tend to get a much higher end-to-end throughput than users in the highly loaded cells, resulting in

a low fairness and overall users' satisfaction. Therefore, the resource allocation should be done for all cells simultaneously in a coordinated manner, which considers not only the radio resources constraints but also the transport network limitations.

5.2.2 Multiple Cells in one eNB

The femtocell cluster is the major focus of this work due to its limited backhaul. However this work is not only limited to femtocell scenarios. Theoretically, the method can be applied to a cell or a group of cells, which share a limited transport network. As stated in the introduction of this chapter, in legacy macrocell networks, the transport network is carefully dimensioned according to desired level of SLAs. Nevertheless, the transport network can still be the major bottleneck in peak hours or certain situations, e.g. with many users under good channel conditions. Besides, with the exploding growth of mobile traffic, the transport network may not be able to fulfil the SLAs.

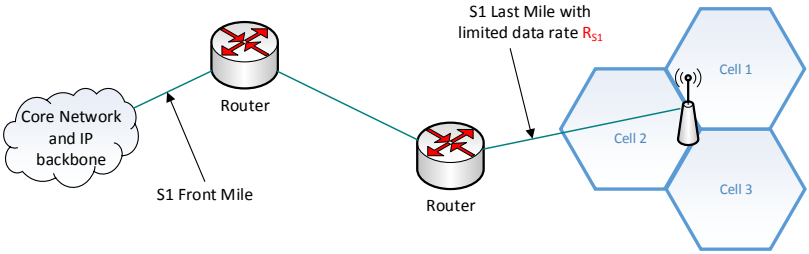


FIGURE 5.5: 3 cells in the same eNB

As a matter of fact, it is meaningful to consider the transport limitations in resource management. Figure 5.5 shows a typical legacy LTE network with 3 cells in one eNB. In this example, the last mile

access network to the eNB is limited, and it is shared by the 3 cells in the same eNB. Since all cells in the same eNB are in the same location, a centralized radio scheduler can be easily implemented inside the eNB without any signaling between the cells.

5.2.3 LTE Cloud Radio Access Network (C-RAN)

Global mobile data traffic will increase nearly tenfold between 2014 and 2019. Mobile data traffic will grow at a Compound Annual Growth Rate (CAGR) of 57 percent from 2014 to 2019, reaching 24.3 exabytes per month by 2019 [Cis15]. One solution is to install more base stations to handle the exploding growth of traffic volume. However, each legacy base station needs a complete system, including power supply, backup battery solution, cooling, a monitor system and so on. Deploying a large-scale amount of base stations needs expensive CAPEX/OPEX especially in the dense city area. Besides, the interference will become more critical with the increase of stations. This will put a challenge for interference management.

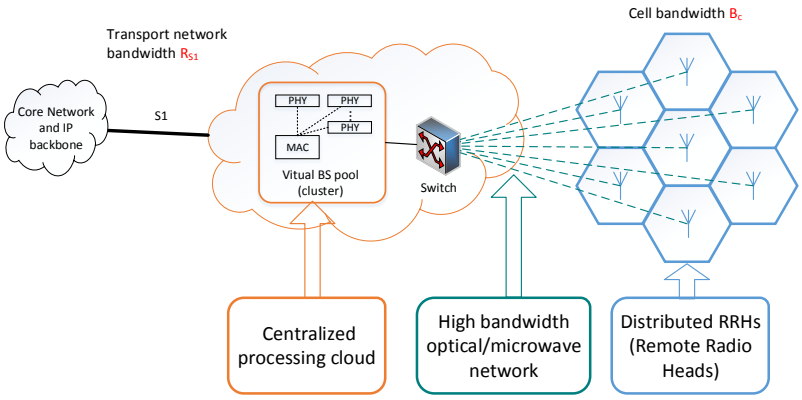


FIGURE 5.6: LTE C-RAN cluster

In order to overcome the above-mentioned problems, China Mobile Research Institute introduced the cloud Radio Access Network (known as Cloud RAN or C-RAN) for future mobile communication networks firstly in April 2010 in Beijing, China [Ins11]. The C-RAN is a centralized, cloud computing-based architecture. The C-RAN concept is widely accepted by operators and vendors, and it has been actively discussed by international standardization organizations. As shown in Figure 5.6, the Base Band Units (BBUs) for resource management and baseband processing are detached from the radio and antennas, and moved into the centralized processing cloud. Only the Remote Radio Heads (RRHs), including antennas, the radio and related amplification and filtering, are distributed among the geographical area for the signal coverage. They are connected to the centralized process cloud via high bandwidth optical/microwave networks. Nevertheless, the C-RAN puts additional challenge on the transport network since a LTE C-RAN cluster can support up to several hundreds of RRHs [Inc14].

5.3 QoE based Dynamic Rate Shaping (QoE-DRS)

In this section, a QoE based Dynamic Rate Shaping framework is introduced marked as QoE-DRS. The framework is valid for any traffic whose QoE can be modelled as strict concave functions. Figure 5.7 gives an overview of the proposed QoE-DRS. The proposed QoE-DRS is a method that decides the optimal rates of every active user in a cell cluster separately, which share a common transport network. The optimal rates are calculated by solving an optimization problem at the core network. The users are shaped according to their optimal rates based on traffic shaping explained in Section 5.1.1. The QoE-DRS framework is done in a periodic manner since the number of active users and the

users' channel conditions are changing over time. The default time interval in this work is 1 ms. A discussion on the shaping time interval is shown in Section 5.4.5.

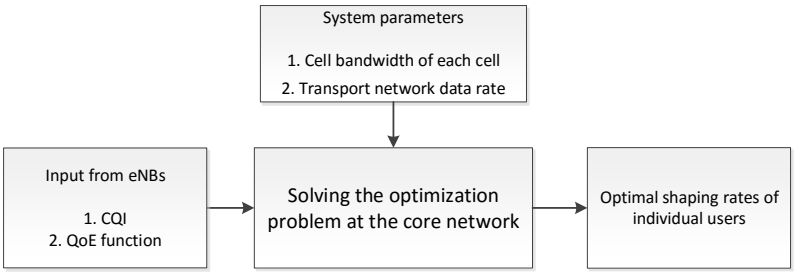


FIGURE 5.7: Overview of QoE-DRS

The resource management problem considering both the radio and transport limitations is formulated as a concave optimization problem in Section 5.3.1. The problem is proven to be convex and is solved by the Lagrangian relaxation method. The method is proven that it is valid for any kind of strict concave utility functions. The detailed solution for the QoE based utility function is shown in Section 5.3.3. In Section 5.3.4, a special case with the assumption that the transport network is the dominant limiting factor is studied, which is a typical use case in the femtocell scenario.

5.3.1 General Problem Formulation

In the above-mentioned scenarios, many cells share the same limited transport network. In order to design an optimized resource management scheme, both radio and transport limitations are obligated to be considered. In Chapter 4, the QoE based radio scheduling is proposed to maximize the aggregated QoE in a cell only considering the radio limitation. In this section, the work is extended to consider both radio

interface and transport backhaul limitations. The resource allocation problem is formulated as an optimization problem subject to the constraints from both radio and transport interfaces. The target of the optimization problem is to maximize the aggregated user QoE in a cell cluster. By solving the optimization problem, the optimal shaping rates for the users in a cell cluster are determined.

The resource management problem is formulated as follows:

$$\begin{aligned} &\max U, \quad U = \sum_c \sum_i u_{i,c}\left(r_{i,c}\right) \\ s.t. \quad &\forall c \quad \sum_i \frac{r_{i,c}}{\sigma_{i,c}} \leq B_c; \quad \sum_c \sum_i r_{i,c} \leq R_{S1} \\ &\forall r_{i,c} \geq 0 \end{aligned} \tag{5.1}$$

The objective is to maximize the aggregated QoE (U) which is the summation of the QoE of all the users. $b_{i,c}$ is the allocated bandwidth of user i in cell c. $r_{i,c}$ represents the data rate of user i in cell c, which is proportional to $b_{i,c}$ ($r_{i,c} = b_{i,c} \cdot \sigma_{i,c}$, see Section 4.2.4). The QoE of the user ($u_{i,c}$) is represented by a strictly concave QoE function of its data rate ($r_{i,c}$). Over the air interface, the total available cell bandwidth is limited by B_c in cell c, which is formulated as $\sum_i b_{i,c} = \sum_i \frac{r_{i,c}}{\sigma_{i,c}} \leq B_c$. Since all the users in a cell cluster share the same transport link (S1), the total rate over the transport network is limited by the date rate R_{S1}. Suppose there are N_c number of cells sharing the same transport network link, there will be $N_c + 1$ constraints in total.

$$H(U) = \begin{bmatrix} \frac{\partial^2 U}{\partial {r_1}^2} & \frac{\partial^2 U}{\partial r_1 \partial r_2} & \cdots & \frac{\partial^2 U}{\partial r_1 \partial r_n} \\ \frac{\partial^2 U}{\partial r_2 \partial r_1} & \frac{\partial^2 U}{\partial {r_2}^2} & \cdots & \frac{\partial^2 U}{\partial r_2 \partial r_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 U}{\partial r_n \partial r_1} & \frac{\partial^2 U}{\partial r_n \partial r_2} & \cdots & \frac{\partial^2 U}{\partial {r_n}^2} \end{bmatrix} \tag{5.2}$$

The Lagrangian relaxation method can be used to solve the problem if the problem is convex. A continuous, twice differentiable function

of several variables is convex on a convex set if and only if its Hessian matrix is positive semidefinite on the interior of the convex set [BNO03]. The Hessian matrix of the optimization target function (U) in eq. (5.1) is a square matrix of second-order partial derivatives of a scalar-valued function shown in eq. (5.2). Suppose there are n users in total in all the cells, the resulted Hessian matrix is a square matrix that has a size of $n \times n$. $\frac{\partial^2 U}{\partial r_j \partial r_k}$ represents the element at the $j's$ row and $k's$ column of the matrix. Since the QoE of an individual user only depends on his own data rate, $\frac{\partial \mathrm{u}_j}{\partial r_k} = 0$ when $j \neq k$. So when $j = k$,

$$\frac{\partial^2 U}{\partial {r_j}^2} = \frac{\partial}{\partial r_j}\left(\frac{\partial \mathrm{U}}{\partial r_j}\right) = \frac{\partial}{\partial r_j}\left(\frac{\partial \mathrm{u}_1 + \cdots + \partial \mathrm{u}_j + \cdots \partial u_n}{\partial r_j}\right) = \frac{\partial}{\partial r_j}\left(\frac{\partial \mathrm{u}_j}{\partial r_j}\right) = \frac{\partial^2 u_j}{\partial {r_j}^2} \tag{5.3}$$

and therefore the element becomes $\frac{\partial^2 r_j}{\partial {r_j}^2}$. Correspondingly, when $j \neq k$,

$$\frac{\partial^2 U}{\partial r_j \partial r_k} = \frac{\partial}{\partial r_j}\left(\frac{\partial \mathrm{U}}{\partial r_k}\right) = \frac{\partial}{\partial r_j}\left(\frac{\partial \mathrm{u}_1 + \cdots + \partial \mathrm{u}_k + \cdots \partial u_n}{\partial r_k}\right) = \frac{\partial}{\partial r_j}\left(\frac{\partial \mathrm{u}_k}{\partial r_k}\right) = 0 \tag{5.4}$$

As a result, the resulted Hessian matrix becomes a diagonal matrix. The element at $j's$ row and $j's$ column is $\frac{\partial^2 u_j}{\partial {r_j}^2}$, which is negative because u_j is a strictly concave function of r_j. Therefore, the Hessian matrix is a diagonal matrix, and all the elements on its diagonal are positive. The Hessian matrix ($H(U)$) is positive definite if the scalar $z^T \cdot H(U) \cdot z$ is positive for every non-zero column vector z of n real numbers [BNO03]. The prove for the Hessian matrix to be positive definite is shown in eq. (5.5).

$$
\begin{aligned}
z^T \cdot H(U) \cdot z &= \begin{bmatrix} a_1 & a_2 & \cdots & a_n \end{bmatrix} \begin{bmatrix} \frac{\partial^2 r_1}{\partial {r_1}^2} & 0 & \cdots & 0 \\ 0 & \frac{\partial^2 r_2}{\partial {r_2}^2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{\partial^2 r_n}{\partial {r_n}^2} \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix} \\
&= {a_1}^2 \cdot \frac{\partial^2 U}{\partial {r_1}^2} + {a_2}^2 \cdot \frac{\partial^2 U}{\partial {r_2}^2} + \cdots + {a_n}^2 \cdot \frac{\partial^2 U}{\partial {r_n}^2} \\
&< 0
\end{aligned}
\tag{5.5}
$$

The optimization problem is proven to be concave (convex upwards) since its Hessian matrix is negative definite, and therefore the Lagrangian relaxation method can be used to solve the problem. According to eq. (5.1), there is no equality constraint. Therefore, the Slater's condition is fulfilled which is a sufficient condition for strong duality to hold for a convex optimization problem [BNO03]. Due to the strong duality, there is no duality gap between the Lagrangian dual problem and the primal problem. The optimal solution of the Lagrangian dual problem is also the optimal solution for the primal problem. The Lagrangian dual problem is formulated and shown in eq. (5.6).

$$
\min_{\{\lambda \geq 0\}} \left\{ \overbrace{\max_{\{\mathbf{r} \geq 0\}} \left\{ \sum_c \sum_i u_{i,c}\left(r_{i,c}\right) - \sum_c \lambda_c \left(\sum_i \frac{r_{i,c}}{\sigma_{i,c}} - B_c \right) - \lambda_0 \left(\sum_c \sum_i r_{i,c} - R_{S1} \right) \right\}}^{f} \right\} \tag{5.6}
$$

Considering the problem f:

$$
\begin{aligned}
f &= \max_{\{\mathbf{r}\geq 0\}} \left\{ \sum_c \sum_i u_{i,c}\left(r_{i,c}\right) - \sum_c \lambda_c \left(\sum_i \tfrac{r_{i,c}}{\sigma_{i,c}} - B_c \right) - \lambda_0 \left(\sum_c \sum_i b_{i,c} - R_{S1} \right) \right\} \\
&= \max_{\{\mathbf{r}\geq 0\}} \left\{ \sum_c \sum_i \left(u_{i,c}\left(r_{i,c}\right) - \left(\tfrac{\lambda_c}{\sigma_{i,c}} + \lambda_0 \right) r_{i,c} \right) + \sum_c \lambda_c \cdot B_c + \lambda_0 \cdot R_{S1} \right\} \\
&= \sum_c \sum_i \max_{\{\mathbf{r}\geq 0\}} \left\{ \overbrace{u_{i,c}\left(r_{i,c}\right) - \left(\frac{\lambda_c}{\sigma_{i,c}} + \lambda_0 \right) r_{i,c}}^{L_{i,c}} \right\} + \sum_c \lambda_c \cdot B_c + \lambda_0 \cdot R_{S1}
\end{aligned}
\tag{5.7}
$$

Lemma: $\forall b_{i,c} \geq 0, L_{i,c}$ is a concave function and has one and only one maximum $L^*_{i,c}$. Then the dual problem becomes:

$$
\begin{gathered}
\min_{\lambda} f = \min_{\lambda} \left\{ \sum_c \sum_i L^*_{i,c} + \sum_c \lambda_c \cdot B_c + \lambda_0 \cdot B_{S1} \right\} \\
\forall \lambda \in R^+ \cup \{0\}
\end{gathered}
\tag{5.8}
$$

Proof: the function $L_{i,c}$ is twice differentiable and its second derivative given below is negative.

$$
\begin{aligned}
\frac{dL_{i,c}}{dr_{i,c}} &= \frac{du_{i,c}(r_{i,c})}{dr_{i,c}} - \left(\frac{\lambda_c}{\sigma_{i,c}} + \lambda_0 \right) \\
\frac{d^2 L_{i,c}}{dr^2_{i,c}} &= \frac{d^2 u_{i,c}(r_{i,c})}{dr^2_{i,c}} < 0
\end{aligned}
\tag{5.9}
$$

$L^*_{i,c}$ reaches its maximum at $r^*_{i,c}$ if $\frac{\partial L_{i,c}}{\partial r_{i,c}}\left(r^*_{i,c}\right) = 0$ and $r^*_{i,c} \geq 0$. Because $L^*_{i,c}$ is concave, if $r^*_{i,c}\Big|_{\frac{\partial L_{i,c}}{\partial r_{i,c}}\left(r^*_{i,c}\right)=0} < 0$, it reaches its maximum when $r_{i,c} = 0$. The formulation of $L^*_{i,c}$ is as follows:

$$
\max\left\{L_{i,c}\right\} = L^*_{i,c} = \begin{cases}
u_{i,c}\left(r^*_{i,c}\right) - \left(\frac{\lambda_c}{\sigma_{i,c}} + \lambda_0 \right) \cdot r^*_{i,c} \\
\qquad \text{if } \frac{du_{i,c}(r_{i,c})}{dr_{i,c}}\Big|_{r_{i,c}=0} \geq \left(\frac{\lambda_c}{\sigma_{i,c}} + \lambda_0 \right) \\
u_{i,c}\left(0\right) \\
\qquad \text{if } \frac{du_{i,c}(r_{i,c})}{dr_{i,c}}\Big|_{r_{i,c}=0} < \left(\frac{\lambda_c}{\sigma_{i,c}} + \lambda_0 \right)
\end{cases}
\tag{5.10}
$$

After calculating all the $L^*_{i,c}$, the problem shown in eq. (5.8) can be solved by a subgradient projection method to get the minimum of f which will be explained in the next section. Due to the strong duality, f has one and only one minimum which is also the solution of the primal problem.

5.3.2 Subgradient Projection Method

The subgradient projection method is an iteration that starts with an initial feasible vector $\lambda^0 \geq 0$, and generates the next iterate by taking a step along the negative subgradient direction -s_k of f at λ^k and then, by projecting on the set $\lambda \geq 0$ to maintain feasibility. Formally, a typical iteration of the subgradient projection method is given by $\lambda^{k+1} = \left[\lambda^k - t_k s_k\right]^+$ with t_k being a positive step size at the iteration k. In this work, the Modified Polyak's step size is applied [PE09]:

$$t_k = \gamma \frac{f\left(\lambda^k\right) - \hat{f}_k}{||s_k||^2} \quad \text{with} \quad \hat{f}_k = \min_{0 \leq j \leq k} f\left(\lambda^k\right) - \delta \tag{5.11}$$

$$t_k = \gamma \frac{f\left(\lambda^k\right) - \min_{0 \leq j \leq k} f\left(\lambda_k\right) + \delta}{||s_k||^2} \tag{5.12}$$

γ, δ are positive parameters. The minimum $f^* = \min_{0 \leq j \leq k} f\left(\lambda^k\right)$ is updated in each iteration, and the corresponding values of the dual variables are updated as well. The iterations are stopped when

$$f\left(\lambda^k\right) - \min_{0 \leq j \leq k} f\left(\lambda^k\right) \leq o$$

(o: a predefined value regarding to accuracy) for some iterations. The minimum f^* is the optimal solution which is also the optimal solution of the primal problem.

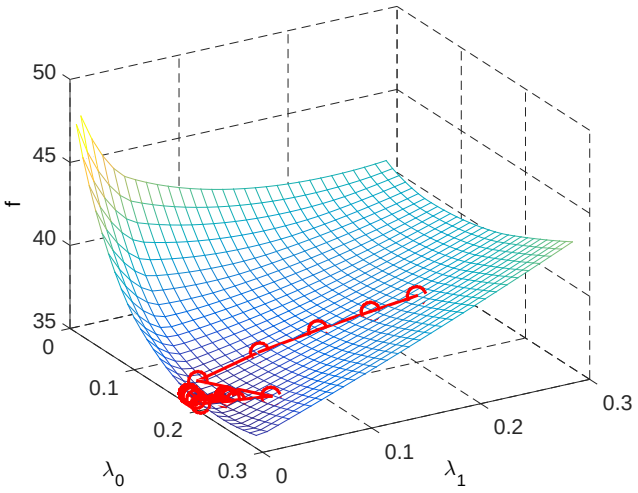


FIGURE 5.8: A visualized example of the subgradient method

Figure 5.8 gives a visualized example of the subgradient method. There is only one cell in this example, and therefore there are two constraints in the primal problem. The system performance is limited by the available cell bandwidth in the cell and the available data rate at the transport network. Correspondingly, there are two dual variables introduced in the Lagrangian dual problem that λ_0 is introduced to relax the data rate limitation at the transport network and λ_1 is introduced to relax the cell bandwidth limitation in the cell. The target is to find the minimum f over the two dual variables. The subgradient method starts with a feasible location. In this example, the initial point is set to $\lambda_0 = \lambda_1 = 0.2$. The subsequent point in each iteration is based on the current point, the negative subgradient direction and the step size based on the modified Polyak's step size [PE10]. It can be seen from the figure that the point is moving towards to the minimum point. The changing of the f over each iteration is shown correspondingly in

Figure 5.9. f is decreasing sharply in the first 5 iterations, but after 5 iterations, f is fluctuating. For this reason, after around 5 iterations, the minimum f^* over all the previous iterations nearly stay at the same value. If the minimum f^* is not changing for some iterations, the iterations are stopped. Knowing the values of the dual variables to get f^*, the corresponding values of the primal variables can be calculated. The detailed formulations are shown in Section 5.3.3 and 5.3.4.

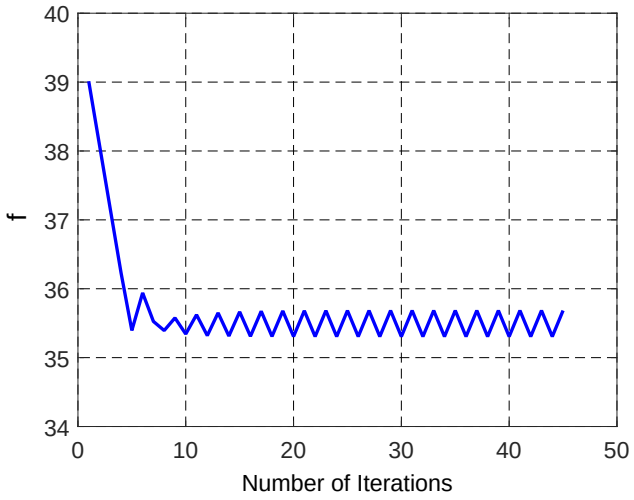


FIGURE 5.9: A visualized example of the subgradient method - how f varying over iterations

By solving the optimization problem, the optimal data rates of all the active users are obtained, e.g. the data rate of user i in cell c is ${r^*}_{i,c} = {b^*}_{i,c} \cdot \sigma_{i,c}$. Afterwards, the user token bucket shaping rates are set to their corresponding optimal date rates $r^*_{i,c}$. Since the user channel conditions are changing over time and the number of active users is variable, the optimal shaping rates are changing over time as well. So the optimization problem needs to be solved periodically and the optimal shaping rates are updated correspondingly. The performance

of the proposed method, the complexity of solving the problem and how often should the shaping rates be updated are investigated in Section 5.4.

5.3.3 Solution for the QoE Based Utility Function

In Section 5.3.1, the problem is formulated and solved in general for any kind of strict concave functions. A QoE based utility function which uses the concave part of a sigmoid function (see Figure 4.7) was studied and applied in Chapter 4. Since the concave part of the sigmoid function is a strict concave function, theoretically the proposed solution framework proposed in Section 5.3.1 can be used directly for the QoE based utility function. The detailed solution process is shown in this section.

The optimization problem is formulated as follows:

$$\begin{aligned} \max U, \quad & U = \sum_c \sum_i u_{i,c}\left(r_{i,c}\right) = \sum_c \sum_i \left(\frac{A}{1+e^{-\alpha_{i,c}\cdot r_{i,c}}} + D\right) \\ s.t. \quad & \forall c \quad \sum_i \frac{r_{i,c}}{\sigma_{i,c}} \leq B_c; \quad \sum_c \sum_i r_{i,c} \leq R_{S1} \\ & \forall i,c \quad r_{i,c} \geq 0 \end{aligned} \tag{5.13}$$

By introducing dual variables, the constraints are relaxed and absorbed into the optimization target function. Accordingly we get the Lagrangian dual problem and it is shown as:

$$\min_{\{\lambda \geq 0\}} \left\{ \overbrace{\max_{\{\mathbf{r} \geq 0\}} \left\{ \sum_c \sum_i \left(\frac{A}{1+e^{-\alpha_{i,c}\cdot r_{i,c}}} + D\right) - \sum_c \lambda_c \left(\sum_i \frac{r_{i,c}}{\sigma_{i,c}} - B_c\right) - \lambda_0 \left(\sum_c \sum_i r_{i,c} - R_{S1}\right) \right\}}^{f} \right\} \tag{5.14}$$

Consider the problem f:

$$f = \max_{\{\mathbf{r}\geq 0\}} \left\{ \sum_c \sum_i \left(\frac{A}{1+e^{-\alpha_{i,c}\cdot r_{i,c}}} + D \right) - \sum_c \lambda_c \left(\sum_i \frac{r_{i,c}}{\sigma_{i,c}} - B_c \right) - \lambda_0 \left(\sum_c \sum_i r_{i,c} - R_{S1} \right) \right\}$$
$$= \max_{\{\mathbf{r}\geq 0\}} \left\{ \sum_c \sum_i \left(\overbrace{\frac{A}{1+e^{-\alpha_{i,c}\cdot r_{i,c}}} + D - r_{i,c}\left(\frac{\lambda_c}{\sigma_{i,c}} + \lambda_0\right)}^{L_{i,c}} \right) + \sum_c \lambda_c \cdot B_c + \lambda_0 \cdot R_{S1} \right\} \tag{5.15}$$

The Hyperbolic tangent of x is defined as: $\tanh x = \frac{e^x - e^{-x}}{e^x + e^{-x}}$ and it can be deduced that $\frac{1}{2}\left(\tanh\frac{x}{2} + 1\right) = \frac{1}{1+e^{-x}}$. Consequently, the L_i can be represented by a hyperbolic function:

$$\begin{aligned} L_{i,c} &= \frac{A}{1+e^{-\alpha_{i,c}\cdot r_{i,c}}} + D - r_{i,c}\left(\frac{\lambda_c}{\sigma_{i,c}} + \lambda_0\right) \\ &= \frac{A}{2}\left[\tanh\left(\frac{\alpha_{i,c}\cdot r_{i,c}}{2}\right) + 1\right] + D - r_{i,c}\left(\frac{\lambda_c}{\sigma_{i,c}} + \lambda_0\right) \end{aligned} \tag{5.16}$$

The function $L^*_{i,c}$ is twice differentiable and its second derivative given below is negative. $\forall r_i \geqslant 0, L_{i,c}$ is a concave function and has one and only one maximum $L^*_{i,c}$.

$$\begin{aligned} \frac{dL_{i,c}}{dr_{i,c}} &= \frac{A\cdot\alpha_{i,c}}{4}\left[1 - \tanh^2\left(\frac{\alpha_{i,c}\cdot r_{i,c}}{2}\right)\right] - \left(\frac{\lambda_c}{\sigma_{i,c}} + \lambda_0\right) \\ \frac{d^2L_{i,c}}{dr^2_{i,c}} &= -\frac{A\cdot{\alpha_{i,c}}^2}{4}\tanh\left(\frac{\alpha_{i,c}\cdot r_{i,c}}{2}\right)\cdot\left[1 - \tanh^2\left(\frac{\alpha_{i,c}\cdot r_{i,c}}{2}\right)\right] < 0 \end{aligned} \tag{5.17}$$

$L_{i,c}$ reaches its maximum at $r^*_{i,c}$ if $\frac{\partial L_{i,c}}{\partial r_{i,c}}\left(r^*_{i,c}\right) = 0$ and ${r_{i,c}}^* \geq 0$. The formulation of L^*_i is shown in eq. (5.18) and the formulation of the corresponding $r^*_{i,c}$ is shown is eq. (5.19).

$$L^*_{i,c} = \max_{r_{i,c}\geq 0} L_{i,c} = \begin{cases} \frac{A}{2}\left[\tanh\left(\frac{\alpha_{i,c}\cdot {r_{i,c}}^*}{2}\right) + 1\right] + D - {r_{i,c}}^*\left(\frac{\lambda_c}{\sigma_{i,c}} + \lambda_0\right) \\ \qquad\qquad \frac{\lambda_c}{\sigma_{i,c}} + \lambda_0 \leq \frac{A\cdot\alpha_{i,c}}{4} \\ \frac{A}{2} + D \quad \frac{\lambda_c}{\sigma_{i,c}} + \lambda_0 > \frac{A\cdot\alpha_{i,c}}{4} \end{cases} \tag{5.18}$$

$$r^*_{i,c} = \begin{cases} \frac{2}{\alpha_{i,c}} \operatorname{arctanh} \sqrt{1 - \frac{4\left(\frac{\lambda_c}{\sigma_{i,c}} + \lambda_0\right)}{A \cdot \alpha_{i,c}}} & \frac{\lambda_c}{\sigma_{i,c}} + \lambda_0 \leq \frac{A\alpha_{i,c}}{4} \\ 0 & \frac{\lambda_c}{\sigma_{i,c}} + \lambda_0 > \frac{A\alpha_{i,c}}{4} \end{cases} \tag{5.19}$$

Within each iteration, the subgradient $\mathbf{s}$ needs to be updated which is calculated by eq. (5.20). The subgradient $\mathbf{s}$ is a column vector with $N_c + 1$ lines. The first element corresponds to the relaxed constraint of the transport network limitation. The remaining elements correspond to the relaxed constraints cell bandwidth limitations. Therefore, suppose there are N_c cells sharing the same transport network, there are $N_c + 1$ elements in the subgradient vector.

$$\mathbf{s} = \begin{bmatrix} \frac{\partial f}{\partial \lambda_0} \\ \frac{\partial f}{\partial \lambda_1} \\ \frac{\partial f}{\partial \lambda_2} \\ \bullet \\ \bullet \\ \bullet \\ \frac{\partial f}{\partial \lambda_{N_c}} \end{bmatrix} = \begin{bmatrix} R_{S1} + \sum_i \sum_c \left(\frac{\partial {L^*}_{i,c}}{\partial \lambda_0} \right) \\ B_1 + \sum_i \left(\frac{\partial {L^*}_{i,1}}{\partial \lambda_1} \right) \\ B_2 + \sum_i \left(\frac{\partial {L^*}_{i,2}}{\partial \lambda_2} \right) \\ \bullet \\ \bullet \\ \bullet \\ B_{N_c} + \sum_i \left(\frac{\partial {L^*}_{i,N_c}}{\partial \lambda_{N_c}} \right) \end{bmatrix} \tag{5.20}$$

The formulations of calculating on the partial deviation $\frac{\partial {L^*}_{i,c}}{\partial \lambda_0}$ and $\frac{\partial {L^*}_{i,c}}{\partial \lambda_c}$ are shown in eq. (5.21). With the subgradient s, the problem shown in eq. (5.8) is solved by a subgradient projection method to get the minimum of f which is explained in the Section 5.3.2.

$$\frac{\partial L^*{}_{i,c}}{\partial \lambda_0} = \begin{cases} -\frac{2}{\alpha_{i,c}} \operatorname{arctanh} \sqrt{1 - \frac{4\left(\frac{\lambda_c}{\sigma_{i,c}} + \lambda_0\right)}{A \cdot \alpha_{i,c}}} & \frac{\lambda_c}{\sigma_{i,c}} + \lambda_0 \leq \frac{A\alpha_{i,c}}{4} \\ 0 & \frac{\lambda_c}{\sigma_{i,c}} + \lambda_0 > \frac{A\alpha_{i,c}}{4} \end{cases}$$

$$\frac{\partial L^* i,c}{\partial^2 \lambda_c} = \begin{cases} -\frac{2}{\alpha_{i,c} \cdot \sigma_{i,c}} \operatorname{arctanh} \sqrt{1 - \frac{4\left(\frac{\lambda_c}{\sigma_{i,c}} + \lambda_0\right)}{A \cdot \alpha_{i,c}}} & \frac{\lambda_c}{\sigma_{i,c}} + \lambda_0 \leq \frac{A\alpha_{i,c}}{4} \\ 0 & \frac{\lambda_c}{\sigma_{i,c}} + \lambda_0 > \frac{A\alpha_{i,c}}{4} \end{cases} \tag{5.21}$$

5.3.4 A Special Case: Bottleneck only at the Transport Network

The resource management problem is formulated and solved in Section 5.3.1 in general. Section 5.3.3 gives the detailed solutions for a QoE based utility function. Both the limitations of the transport network and the cell bandwidth are taken into consideration as constraints in the optimization problem. Nevertheless, there are two special cases:

- the bottleneck is only at the radio side limited by the cell bandwidth;
- the bottleneck is only at the transport network side limited by the backhaul data rate.

The first case is covered by Chapter 4. The second case is valid when the backhaul is the major limiting factor which is a typical assumption for femtocells. In 3GPP Release 12 [3GP13], various small cell enhancements were studied. Owing to the fact that small cells have very limited coverage and the users are very close to the small cell stations, very high SINR is potentially available at the end users. Not surprisingly, the highest supported modulation was increased from 64 QAM to 256 QAM. Combining 256 QAM with advanced 8×8 MIMO and Carrier Aggregation up to 5 component carriers in the downlink direction, a user's peak downlink data rate can be up to 3.9 Gbps.

Nevertheless, the femtocells normally use broadband services as backhaul, e.g. xDSL. According to a recent report in Q3 of 2015 [Tec15], the global average internet access speed is only 5.1 Mbps. South Korea has the highest average internet connection speed of 20.5 Mbps. So the user's achievable data rate and the cell capacity can be substantially larger than the backhaul limit in the femtocell scenarios. Therefore, it is important to study this special case which is useful for the femtocell scenarios.

Consider the assumption that there is no cell bandwidth limitation, there will be only one constraint. The optimization problem is formulated in eq. (5.22).

$$\begin{aligned} \max U, \quad & U = \sum_i u_i\left(r_i\right) = \sum_i \left(\tfrac{A}{1+e^{-\alpha_i \cdot r_i}} + D\right) \\ s.t. \quad & \sum_i r_i \le R_{S1}; \quad \forall i \quad r_i \ge 0 \end{aligned} \tag{5.22}$$

Accordingly, there is only one dual variable λ introduced in the Lagrangian dual problem which is shown in eq. (5.23).

$$\min_{\{\lambda \ge 0\}} \left\{ \overbrace{\max_{\{\mathbf{r} \ge 0\}} \left\{ \sum_i \left(\frac{A}{1+e^{-\alpha_i \cdot r_i}} + D \right) - \lambda \left(\sum_i r_i - R_{S1} \right) \right\}}^{f} \right\} \tag{5.23}$$

Consider the problem f:

$$\begin{aligned} f &= \max_{\{\mathbf{r} \ge 0\}} \left\{ \sum_i \left(\tfrac{A}{1+e^{-\alpha_i \cdot r_i}} + D \right) - \lambda \left(\sum_i r_i - R_{S1} \right) \right\} \\ &= \max_{\{\mathbf{r} \ge 0\}} \left\{ \sum_i \left(\overbrace{\frac{A}{1+e^{-\alpha_i \cdot r_i}} + D - \lambda r_i}^{L_i} \right) + \lambda R_{S1} \right\} \end{aligned} \tag{5.24}$$

L_i reaches its maximum at r_i^* if $\frac{\partial L_i}{\partial r_i}\left(r_i^*\right) = 0$ and ${r_i}^* \ge 0$. The

formulation of L_i^* is shown in eq. (5.25) and the formulation of the corresponding $r_{i,c}^*$ is shown in eq. (5.26).

$$L_i^* = \max_{r_i \geq 0} L_i = \begin{cases} \frac{A}{2}\left\{\tanh\left[\frac{\alpha_i {r_i}^*}{2}\right]+1\right\} + D - \lambda {r_i}^* & \text{if} \lambda \leq \frac{A\alpha_i}{4} \\ \frac{A}{2} + D & \text{if} \lambda > \frac{A\alpha_i}{4} \end{cases} \tag{5.25}$$

$$r_i^* = \begin{cases} \frac{2}{\alpha_i}\text{arctanh}\left(\sqrt{1-\frac{4\lambda}{A\alpha_i}}\right) & \text{if} \lambda \leq \frac{A\alpha_i}{4} \\ 0 & \text{if} \lambda > \frac{A\alpha_i}{4} \end{cases} \tag{5.26}$$

Let $\lambda_i = \frac{A\alpha_i}{4}$. Without loss of generality, it is assumed that: $0 \leq \lambda_1 \leq \lambda_2 \leq ... \leq \lambda_N$. According to eq. (5.25) and (5.26).

$$f = \begin{cases} \sum_{i=1}^{N}\left\{\frac{1}{2\alpha_i}\left[-4\text{arctanh}\left(\sqrt{1-\frac{4\lambda}{A\alpha_i}}\right)\lambda + \sqrt{A\alpha_i\left(A\alpha_i-4\lambda\right)} + \left(A+2D\right)\alpha_i\right]\right\} \\ +\lambda R_{S1} \qquad \lambda \leq \lambda_1 \\ \sum_{i=k+1}^{N}\left\{\frac{1}{2\alpha_i}\left[-4\text{arctanh}\left(\sqrt{1-\frac{4\lambda}{A\alpha_i}}\right)\lambda + \sqrt{A\alpha_i\left(A\alpha_i-4\lambda\right)} + \left(A+2D\right)\alpha_i\right]\right\} \\ +k\left(\frac{A}{2}+D\right) + \lambda R_{S1} \qquad \lambda_k < \lambda \leq \lambda_{k+1} \qquad \forall k \in [1, N-1] \\ N\left(\frac{A}{2}+D\right) + \lambda R_{S1} \qquad \lambda > \lambda_N \end{cases} \tag{5.27}$$

The function $f = \sum_i L_i^* + \lambda R_{S1}$ is visualized in Figure 5.10:

The function f is continuous. Except at the points where $\lambda = \lambda_i$ - as shown in Figure 5.10, it is twice differentiable and its second derivative is non-negative as shown below:

$$\frac{d\left(dL_i^*\right)}{d\lambda_i^2} = \begin{cases} \frac{A}{\sqrt{A\alpha_i(A\alpha_i-4\lambda)\lambda}} \geq 0 & if \quad \lambda \leq \frac{A\alpha_i}{4} \\ 0 & if \quad \lambda > \frac{A\alpha_i}{4} \end{cases} \tag{5.28}$$

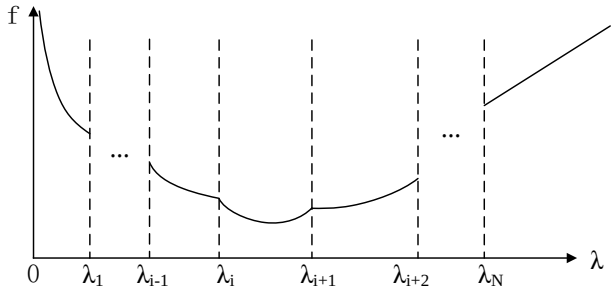


FIGURE 5.10: Visualization of the function f

$$\frac{d\,(df)}{d\lambda^2} = \sum_i \frac{d\,(dL_i^*)}{d\lambda^2} \geq 0 \qquad \forall\lambda \tag{5.29}$$

Therefore, f has one and only one minimum. Due to the strong duality, the solution of the dual problem is also the solution of the primal problem.

In order to find the minimum of f, the interval $(\lambda_i, \lambda_{i+1})$ containing it needs to be determined. This is done as follows: Let $\lambda_0 = 0$, at each λ_i, starting from $i = 0$, compute: $\left.\frac{\partial f}{\partial \lambda}\right|_{\lambda=\lambda_i+\delta}$ and $\left.\frac{\partial f}{\partial \lambda}\right|_{\lambda=\lambda_i-\delta}(\delta \to 0)$ as follows:

$$\begin{aligned}
\left.\frac{\partial f}{\partial \lambda}\right|_{\lambda=\lambda_i-\delta} &= \sum_i \left.\frac{\partial L_i^*}{\partial \lambda}\right|_{\lambda=\lambda_i-\delta} + R_{S1} = \sum_{k=i}^{N} \left[-\tfrac{2}{\alpha_k}\operatorname{arctanh}\left(\sqrt{1-\tfrac{4\lambda}{A\alpha_k}}\right)\right] + R_{S1} \\
\left.\frac{\partial f}{\partial \lambda}\right|_{\lambda=\lambda_i+\delta} &= \sum_i \left.\frac{\partial L_i^*}{\partial \lambda}\right|_{\lambda=\lambda_i+\delta} + R_{S1} = \sum_{k=i+1}^{N} \left[-\tfrac{2}{\alpha_k}\operatorname{arctanh}\left(\sqrt{1-\tfrac{4\lambda}{A\alpha_k}}\right)\right] + R_{S1}
\end{aligned} \tag{5.30}$$

At $\lambda_0 = 0$, we have: $\left.\frac{\partial f}{\partial \lambda}\right|_{\lambda\to 0} = \sum_i \left.\frac{\partial L_i^*}{\partial \lambda}\right|_{\lambda\to 0} = -\infty$

Therefore, if $\left.\frac{\partial f}{\partial \lambda}\right|_{\lambda=\lambda_1-\delta} > 0$, the minimum is in the interval $(0, \lambda_1)$. Otherwise,

- for $i > 0$, if $\left(\frac{\partial f}{\partial \lambda}\Big|_{\lambda=\lambda_i+\delta}\right)\left(\frac{\partial f}{\partial \lambda}\Big|_{\lambda=\lambda_i-\delta}\right) < 0$, the minimum is at λ_i
- if $\left(\frac{\partial f}{\partial \lambda}\Big|_{\lambda=\lambda_i+\delta}\right)\left(\frac{\partial f}{\partial \lambda}\Big|_{\lambda=\lambda_{i+1}-\delta}\right) < 0$, the minimum is between λ_i and λ_{i+1}

After determining the interval of the minimum, we solve the following equation to get the λ^* at which f reaches its minimum:

$$\frac{df}{d\lambda} = \sum_{k=i+1}^{N}\left[-\frac{2}{\alpha_k}\text{arctanh}\left(\sqrt{1-\frac{4\lambda}{A\alpha_k}}\right)\right] + R_{S1} = 0 \tag{5.31}$$

using the bisection method with initial points (λ^*,λ_{i+1}). The solution for the primal problem is:

$$r_k^* = \begin{cases} 0 & \forall k \leq i \\ \frac{2}{\alpha_k}\text{arctanh}\left(\sqrt{1-\frac{4\lambda^*}{A\alpha_k}}\right) & \forall i \leq k \leq N \end{cases} \tag{5.32}$$

5.4 Simulation Scenarios and Results

In this part, the performances of the QoE-DRS framework are evaluated by simulations and compared to the cases using the legacy algorithm of Fixed Rate Shaping (FRS) with a fixed maximum bearer rate (UE-AMBR), which is set to 7.2 Mbps by default since nowadays mobile operators provide up to a fixed 7.2 Mbps or even higher peak data rate [Mob15]. Over the air interface, the radio resource scheduler or MAC scheduler is responsible for scheduling the available air interface resource in each cell without coordination. GBR traffic (VoIP) is given to a strict priority over non GBR (video streaming, HTTP and FTP) traffic. A Longest Waiting-time First (LWF) scheduler is used for GRB

TABLE 5.1: General simulation settings

Parameters	**Settings**
Macro eNBs settings (fully loaded)	7 eNBs with hexagonal coverage, 500ms inter-eNB distance (center eNB located at the original point (0m,0m)) Pathloss: 130.5 + 37.6log10(R), R in Km [IR09] Slow fading: Correlated Log normal, zero mean, 8db std. and 50 m correlation distance Small scale fading: 3GPP Pedestrian A Transmission power: 23dBm per PRB
Femtocell cluster settings	Building size with 40mx40m, center coordinate:(200m,0m) 3 femtocell station coordinates: (210m, -10m), (190m, 10m), (210m,10m) Penetration loss (interference from macro eNBs) over the wall: 12dB mean with 8dB std. Pathloss: 41.1 + 16.9*log10(R), R in Km [IR09] Small scale fading: 3GPP Pedestrian A Transmission power: 0dBm per PRB
TCP version	New Reno with 64Kbytes receiver buffer size
Traffic types	VoIP: GSM EFR, codec rate 12.2 kbps Video Streaming: TCP based full buffer streaming HTTP: 2MB page size, Inter arrival time: exp. distributed with mean: 50s FTP: 10MB file size, Inter arrival time: exp. distributed with mean: 50s
Mobility model	5Km/h,Random waypoint
aGW shaper	Token Bucket algorithm, maximum token bucket size: 64KB
Transport limitation	16Mbps or 6 Mbps (aGW to femtocell cluster); 1Mbytes buffer size
Number of PRBs	25 PRBs (5MHz spectrum at 2.6 GHz)
Simulation time	1000s (5 runs with different seeds) with warm up period of 300s

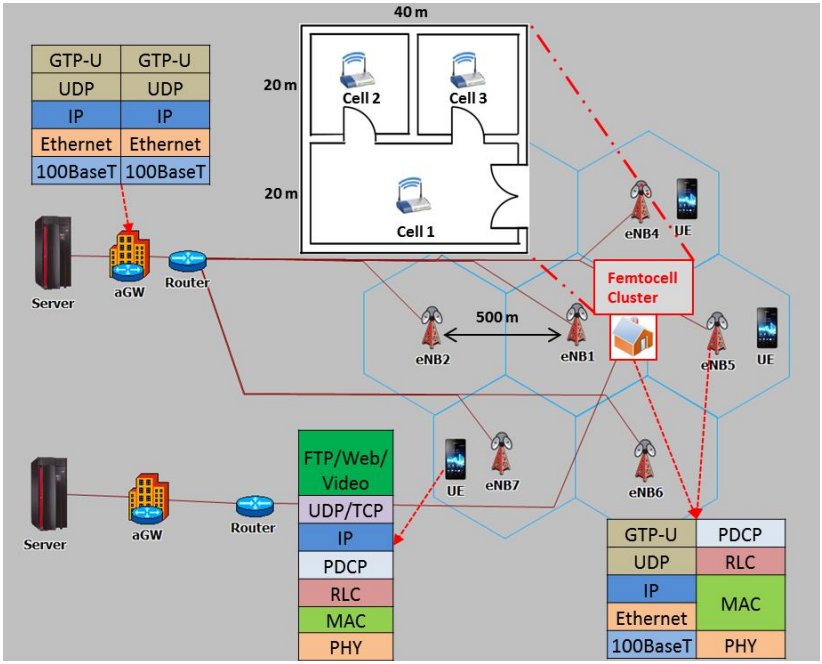


FIGURE 5.11: LTE femtocell cluster simulation model

traffic while a typical Proportional Fair (PF) scheduler is employed for the non-GBR traffic.

5.4.1 Lightly Loaded Scenario – Scenario 1

Figure 5.12 and 5.13 show the application performance and their corresponding MOS (the error bars show the 95% confidence interval) for the lightly loaded scenario. Since the VoIP users only consume very little amount of resources and they are prioritized throughout the network, the voice packet end-to-end delays (including coding, decoding and de-jittering delay) are the same with FRS and QoE-DRS. Regarding to the non-GBR traffic, with QoE-DRS the video streaming performance is reduced since its data rate is limited by shaping rate. Nevertheless,

TABLE 5.2: Scenario settings (scenario 1 to 3)

	Transport link limitation	Number of UEs
Scenario 1	16 Mbps	10 voips + 3 video in cell 1 3 HTTP in cell 2 3 FTP in cell 3
Scenario 2		10 voips + 10 video in cell 1 10 HTTP in ell 2
Scenario 3	6 Mbps	10 FTP in cell 3

HTTP and FTP users get more resources and the download response times are reduced dramatically. Correspondingly, the MOS of FTP and HTTP users are improved significantly with QoE-DRS. Though achieving less data rate with QoE-DRS, the MOS of video streaming users nearly remain the same. The reason is that the video MOS has non-linear relation with the data rates. Thus the video users can remain satisfied with even less amount of resources as seen from Figure 4.11. With little performance degradation of video users, the satisfactions of HTTP and FTP users are improved significantly. Therefore, the aggregated non-GBR user satisfaction in term of MOS is increased from 35.6 to 40.6 shown in Figure 5.13. Considering the minimum MOS value (scaling between 1 and 5) of each user is 1, the gain (G) is around 18.8% calculated based on eq. (5.33), which is already used in 4. MOS_2 and MOS_1 represents the aggregated MOS for all the non GRB users with QoE-DRS and FRS, and N represents the total number of non GBR users.

$$G = \frac{(MOS_2 - N \cdot 1) - (MOS_1 - N \cdot 1)}{(MOS_1 - N \cdot 1)} \cdot 100\% = \frac{MOS_2 - MOS_1}{(MOS_1 - N)} \cdot 100\% \tag{5.33}$$

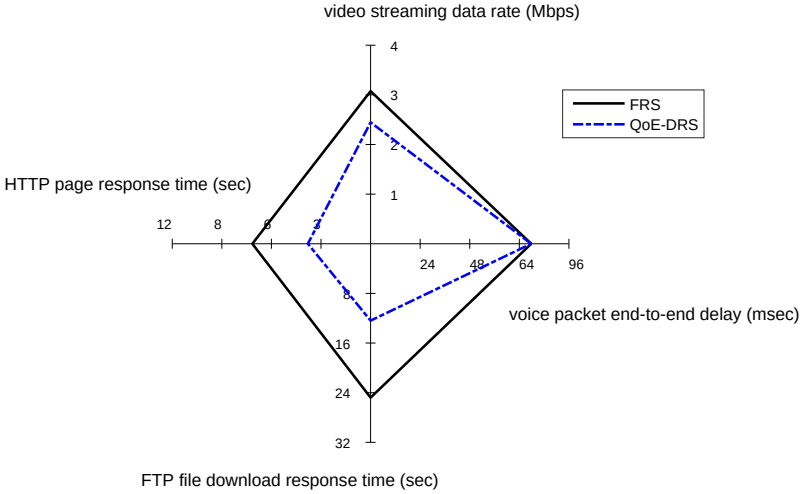


FIGURE 5.12: Application performance comparison for lightly loaded scenario

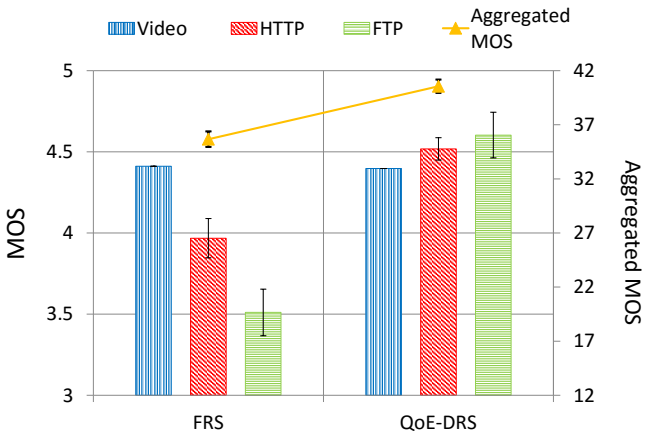


FIGURE 5.13: MOS comparison for lightly loaded scenario

5.4.2 Heavily Loaded Scenario – Scenario 2

Figure 5.14 and 5.15 show the system performance with increased load. Different from the lightly load scenario, the video users' QoE is improved with QoE-DRS. Video and HTTP users are getting more resources from FTP users. That is because with low data rate, video users have the highest marginal MOS, which is the first order derivative of the MOS curve over throughput, while the FTP user has the lowest. The FTP users need more resources to get the same satisfaction level as video and HTTP users. From Figure 5.15, it can be seen, both video and HTTP users have significantly better MOS with QoE-DRS. FTP users have slightly worse MOS with QoE-DRS, but the FTP users have very low MOS and they are not satisfied even with FRS. The aggregated MOS of non GBR users is increased from 77.8 to 84.3 which is 13.6%.

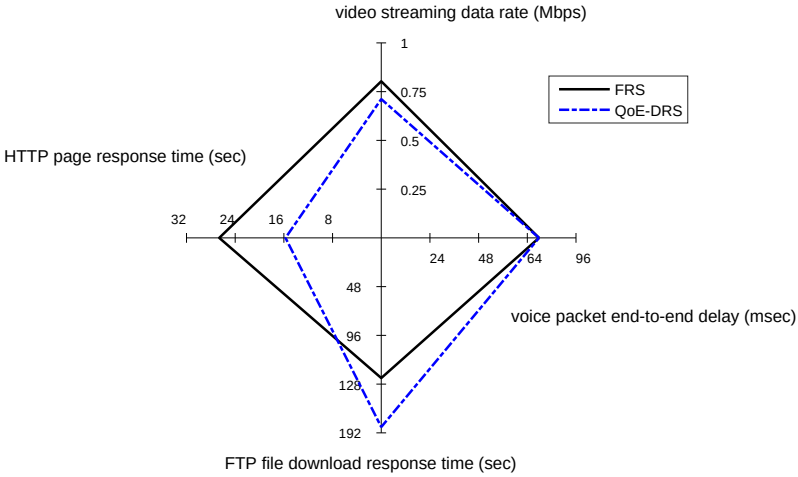


FIGURE 5.14: Application performance comparison for heavily loaded scenariol

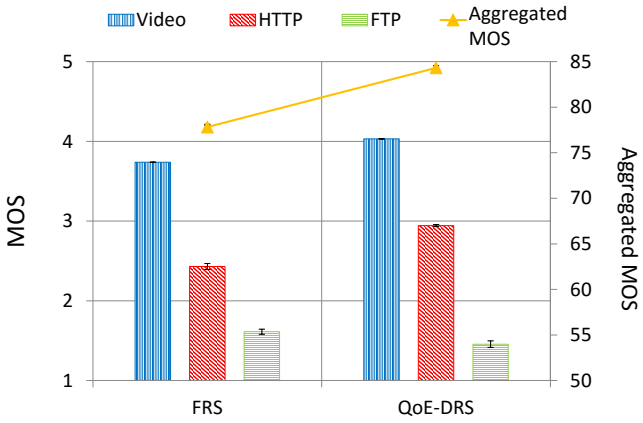


FIGURE 5.15: MOS comparison for heavily loaded scenario

5.4.3 Very Heavily Congested Scenario – Scenario 3

In scenario 3, the transport link is limited to 6 Mbps instead of 16 Mbps in scenario 2. As a result, the user plane is very congested. In this situation, the video users get most of the resources with QoE-DRS since they could be satisfied with little amount of resource. Figure 5.16 shows that HTTP and FTP users nearly cannot get any resources while video users still have very high MOS. With FRS, there is no service differentiation for the users with different applications since the AMBR is very high and it does not play a role. In addition, the TCP flows tend to have the equal share of the bottleneck link capacity. Therefore, the users with different applications have similar throughout, but their user satisfactions differ significantly. None of the users have a MOS over 3 which means that all users have poor user experiences. With QoE-DRS, most of resources are given to video streaming users and they achieve good user experiences since their MOS values are close to 4. The HTTP users and FTP users have poor user experiences with

both approaches. The aggregated MOS of non-GBR users is increased from 55.5 to 63.5 which has a gain of 31.4%.

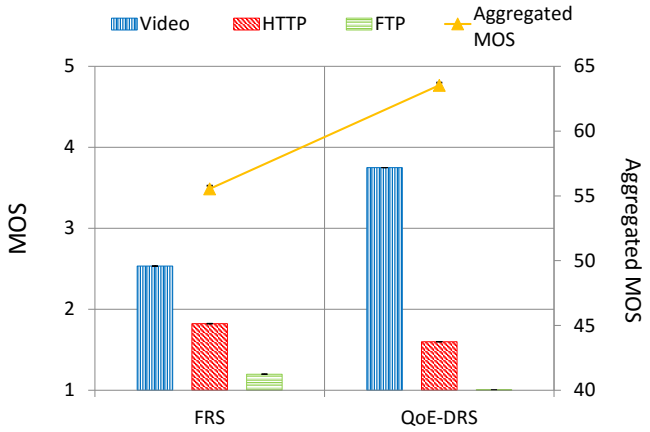


FIGURE 5.16: MOS comparison for very heavily congested loaded scenario

In all the investigated scenarios, the aggregated MOS with QoE-DRS is higher than FRS since the proposed QoE-based bearer shaping algorithm can dynamically and periodically adjust the shaping rate with the purpose of maximizing the aggregated user MOS for a cell cluster.

5.4.4 Discussion on the Complexity of the Proposed Algorithm

It is quite obvious that the QoE-DRS algorithm is able to make service differentiation, i.e. the resources are allocated to different services in a way to maximize the aggregated user satisfaction. However, in practice, it might be quite difficult to re-calculate the optimal shaping rates every 1ms. One reason is it might not be enough CPU power to solve the optimization problem in 1ms. Figure 5.17 shows that, under scenario

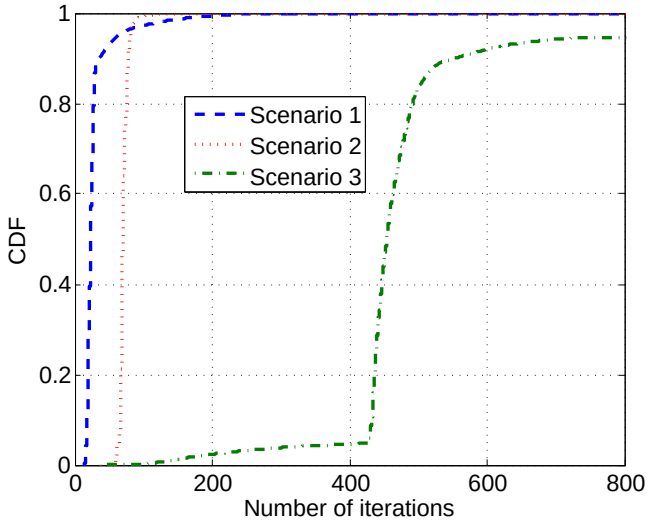


FIGURE 5.17: Number of iterations needed to get convergent

1 and 2, most of the cases (more than 97%), the optimal solution can be found with only 100 iterations which is solvable on a million second basis (less than 0.1 ms observed by simulation with Intel Core i7-3960X [Int16]). However, it might need a longer time to get convergent with sub-gradient method in rare cases. In scenario 3, when there are many active users and the transport link is very congested, it needs a lot more iterations to get the optimal results. Therefore the optimization problem might be not solvable on a millisecond basis. In the following section, a discussion on the impact of shaping rate time interval is presented (the time interval to calculate the shaping rates by solving an optimization problem). In addition, efficient heuristics with reduced complexity are proposed and evaluated against the QoE-DRS method in Section 5.5.

5.4.5 Discussion on the Shaping Rate Update Interval (Based on Scenario 1)

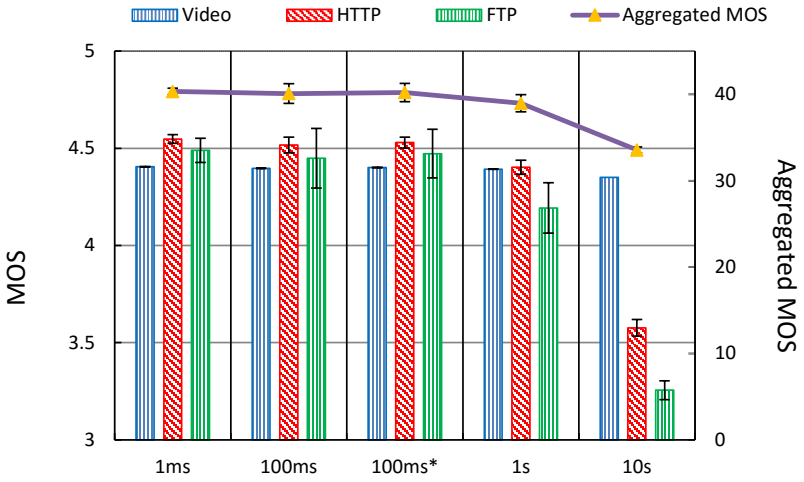


FIGURE 5.18: MOS comparison for variable shaping rate update interval

In reality, the optimization might not be done on 1ms basis, especially in large scale networks with a very large number of users under very congested scenarios. Besides, exhaustive signaling is needed on uplink to carry the user radio channel information when the shaping rate updates every 1ms. It is motivated to study the impact on the system performance when the optimal shaping rates are recalculated on longer time intervals. Figure 5.18 compares the performance with variable shaping rate update intervals based on scenario 1. With the 100ms update interval, the performance nearly remains the same as 1ms, and with 1s interval, slightly performance degradation can be seen. However, as expected, significant performance degradation is observed with 10s interval. Because the radio channel information is quite outdated after a long time. Besides, many users may become active or inactive

within this long period. As a result, the 100ms shaping update interval is favored from the investigation.

To be noticed, the signaling delay of the user channel information report on the uplink is not been considered in the investigation so far. One additional simulation with 5ms uplink signaling delay (from eNB to EPC, see 100ms* in Figure 5.18) is been studied with 100ms shaping update interval. The results show that the performance degradation is neglectable with 5ms uplink delay.

5.4.6 Coexistence with Transport Scheduler

As discussed in Section 5.1.2, the transport scheduler and shaper can provide service differentiation over the transport network. For example in general, the GBR traffic is mapped to the EF PHBs with highest priority. The best effort traffic is mapped to the BE PHB with lowest priority and the rest traffic is mapped to AF PHBs. In case of non-ideal transport network, the transport scheduler and shaper is responsible for deciding how different PHBs share the limited resources according to the QoS settings for different PHBs, e.g. weight settings and discarding mechanisms.

The transport scheduler is able to differentiate services per PHB class based according to the configurations. Nevertheless, the service differentiation is done in a fixed manner. For example, the weight settings are fixed in traditional transport scheduler, e.g. WFQ scheduler. Besides, unlike the proposed QoE-FRS which shapes and differentiates the traffic per bearer based, it cannot differentiate the traffic within the same PHB. For instance, many applications might be mapped to BE PHB and therefore they will receive the same treatment by the transport scheduler. Nevertheless, the proposed QoE-FRS is a function in the core network and can co-exist with the transport scheduler. On the downlink direction, the proposed QoE-FRS is supposed to shape the overall traffic, according to the transport and radio capacities prior of the transport and radio access network. As a result, the QoE-FRS

can leverage the resource management in LTE Radio Access Network (RAN) and transport network theoretically since they are not congested on the downlink direction.

TABLE 5.3: Scenario settings (coexistence with transport scheduler)

Method	Core Network	Transport Network	
	Shaping	PHB Settings	Weight
WFQ 1	FRS	Video: AF21 HTTP: AF11 FTP: BE	AF21: 100 AF11: 10 BE: 1
WFQ 2			AF21: 3 AF11: 2 BE: 1
SQ		All mapped to BE	-
QoE-DRS	QoE-DRS		

Table 5.3 shows the PHB mappings and WFQ transport scheduler weight settings used in the simulation. The single queue (SQ) has been used in the simulations of previous sections as well that all the traffic is mapped to the same PHB. In WFQ 1 and 2, video streaming, HTTP and FTP are mapped to AF21, AF11 and BE PHB correspondingly. Among the three PHBs, AF21 is given the highest weight and BE the lowest. In WFQ 1, the weights of three PHBs are set to 100:10:1. The weight differences are very high that the AF21 PHB's throughput is 100 times as high as BE PHB and 10 times as AF11 in case that there are enough packets in the buffer of all three PHBs. In WFQ 2, the weight differences are smaller comparing to WFQ 1. The legacy FRS is used to shape the bearers according to their maximum bearer rates. The performance of these methods is compared against the proposed QoE-DRS mechanism based on the scenario 1 to 3. In the QoE-DRS method, all the traffic is mapped to the BE PHB for demonstration.

Nevertheless, with the proposed QoE-DRS in the core network, the traffic is shaped according to the transport and radio access network capacities. The resource management in transport and access network can be leveraged.

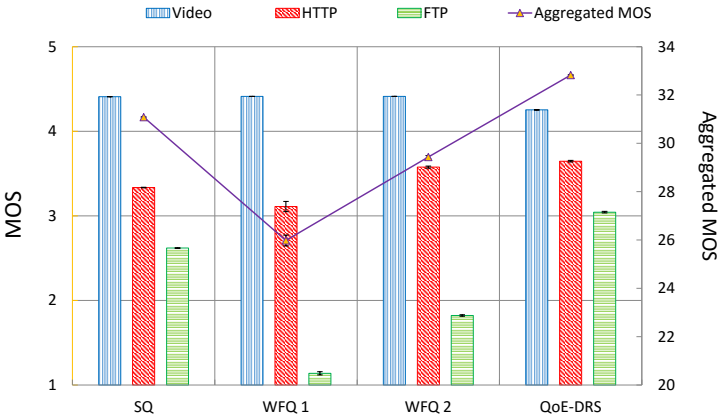


FIGURE 5.19: MOS comparison for lightly loaded scenario, with transport scheduler

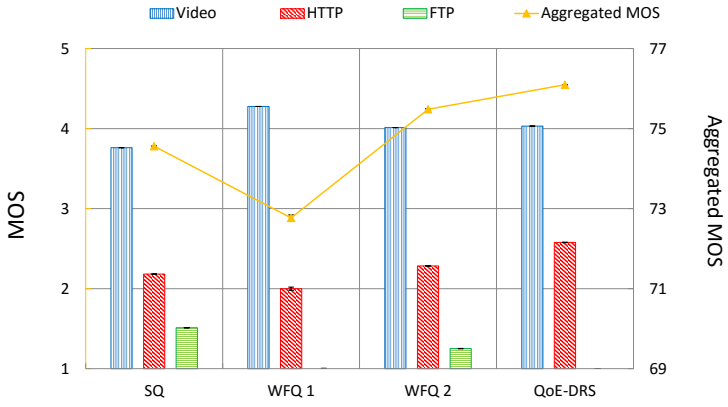


FIGURE 5.20: MOS comparison for heavily loaded scenario, with transport scheduler

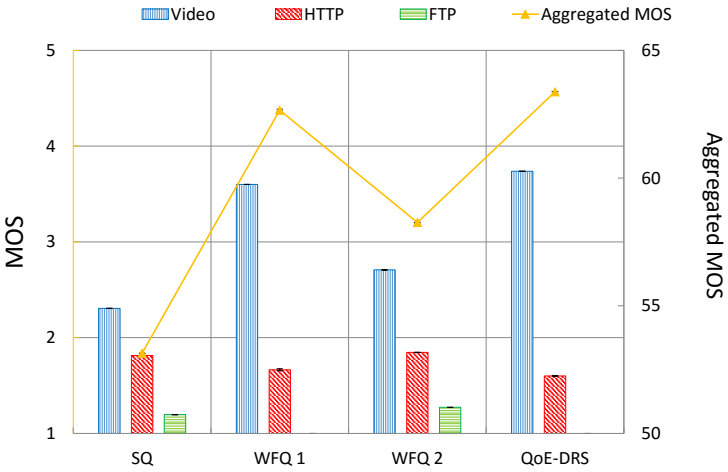


FIGURE 5.21: MOS comparison for very heavily congested scenario, with transport scheduler

Figure 5.19 - 5.21 compare the MOS with the four methods for all three scenarios. In this part, the full buffer mode is used for all the users that all the applications are set to active all the time by setting the idle time to zero. It can been seen from the three figures, the QoE-DRS method achieves the highest aggregated MOS for all the scenarios as expected. The proposed QoE-DRS can dynamically differentiate the traffic in a smarter way based on the optimal bearer rates calculated by solving the optimization problem. As referred to the legacy transport scheduler methods, the performance regarding to aggregated MOS differs with different scenarios. With SQ method, theoretically, all the users with different applications will have the same throughput. Due to no service differentiation, all the TCP flows tend to have a fair share of the network capacity. In the lighted loaded scenario, the SQ performs better than WFQ 1 and 2. Because the video streaming users are already very satisfied, allocating more resources to video streaming users will reduce the throughput of HTTP and FTP users, and therefore degrade the aggregated MOS by WFQ 1 and 2.

With increasing traffic load on the network, WFQ 2 becomes better than SP and WFQ 1. In this heavily loaded scenario, the FTP users can not be satisfied in any case. Therefore, allocate more resources to video streaming users and HTTP users will help to increase the aggregated MOS. Nevertheless, the WFQ 1 still allocate too much resource to video users comparing to WFQ 2. In scenario 3, the transport network is very congested, and WFQ 2 outperform WFQ 1 and SP methods. In this case, most of the resources should be allocated to the video streaming users since the video streaming users need less amount of resources to get a relatively high MOS.

5.5 Heuristic Design

Although the optimal bearer shaping method can be generalized for LTE networks, the performance evaluation focuses only on the femtocell scenarios. However, a large scale network often serves hundreds of users and solving the optimization problem may not be feasible. In order to apply the proposed approach in large scale networks, efficient radio scheduling heuristic algorithms are developed with reduced complexity.

Figure 5.22 shows the flow chart of the proposed QoE-based scheduler. The scheduling is performed periodically every 1 ms in each cell. The retransmission users are assured with highest priority. If any resources left, the resources are allocated to the active users in an iterative manner. This iterative method was proven optimal given full buffer assumption and without transmission error in Chapter 4 without transport limitations.

Two heuristics methods (different only on the priority factor for scheduling) are proposed in this section. The method 1 is based on the marginal QoE by getting one additional PRB which is used in Chapter 4. Besides, the method 2 is proposed by taking the marginal QoE and the channel quality coefficient (σ, introduced in Chapter 4.2.4) into consideration. The purpose is to allocate the PRB to the user with

highest marginal QoE to channel quality ratio. In case the transport network is a major bottleneck, the channel quality is becoming less critical in radio scheduling. The resource is preferred to allocate to the user with high marginal QoE but putting less traffic on the bottleneck link. The detailed scheduling procedures are summarized as follows:

1. For each cell, initialize the candidates' list. They are inserted into two lists:
   - HARQ list for the users with pending retransmission, which means the users experienced transmission errors 8 TTIs earlier;
   - new users list for the active users, which have packets/segments in PDCP/RLC buffers.
2. In all cells, the retransmission users are granted with highest priority. The users will be allocated with the same amount of PRBs as they got for the last unsuccessful transmissions.
3. If any PRBs left, the new users in list b are taken into consideration for scheduling. The effective SINR for each user i is calculated based on the EESM SINR mapping method. Therefore, the MCS is determined correspondingly and the MCS dependent indicator σ_i is obtained. The priority factor for each user i by getting a PRB is calculated by:

$$\text{Heuristic method 1} \quad m_{i,c} = u_{i,c}\left(n_{i,c}+1\right) - u_{i,c}\left(n_{i,c}\right)$$

$$\text{Heuristic method 2} \quad m_{i,c} = \frac{u_{i,c}(n_{i,c}+1) - u_{i,c}(n_{i,c})}{\sigma_{i,c}}$$

where $n_{i,c}$ is initialize as 0 for all the users. In this work, the default utility function is $u_{i,c}\left(r_{i,c}\right) = \frac{A}{1+e^{-\alpha_{i,c}\cdot r_{i,c}}} + B = \frac{A}{1+e^{-\alpha_{i,c}\cdot b_{i,c}\cdot \sigma_{i,c}}} + B$. So the $m_{i,c}$ is initialize as:

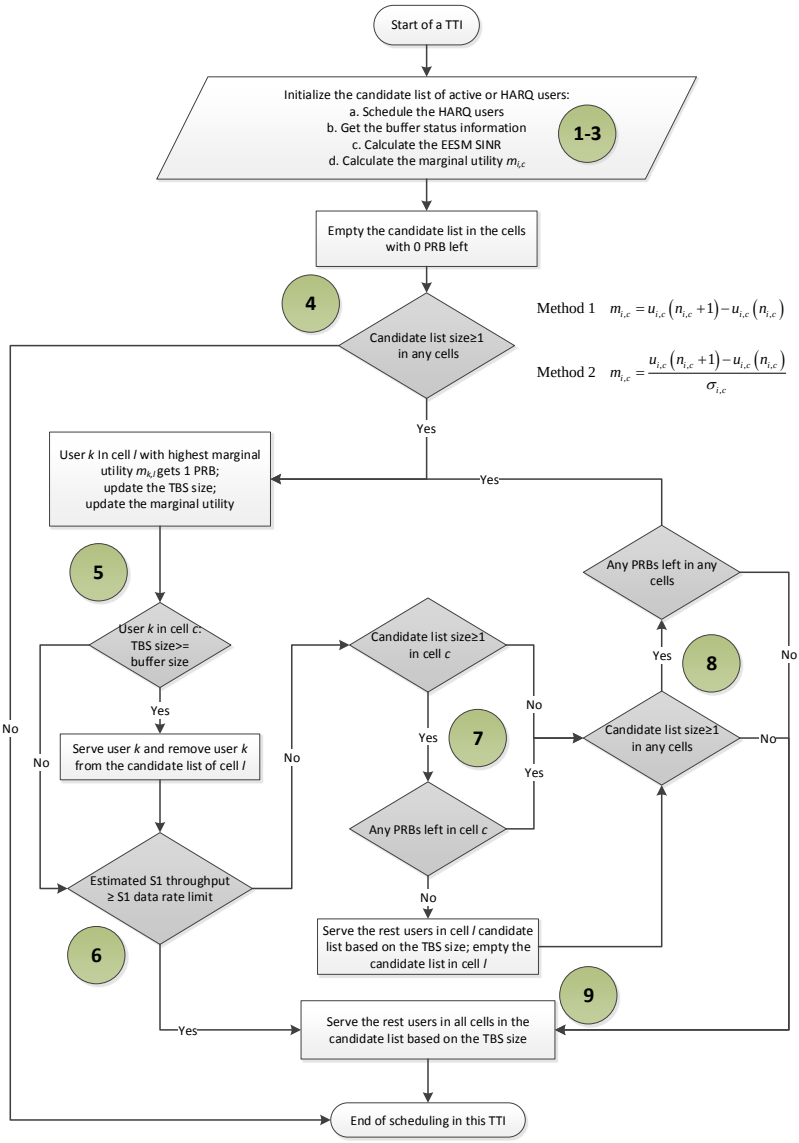


FIGURE 5.22: Flow chart of the proposed QoE-based scheduler

$$m_{i,c} = \frac{A}{1 + e^{-\alpha_{i,c} \cdot 0.18 \cdot \sigma_{i,c}}} - \frac{A}{2}$$

4. Empty the candidate list in the cells with 0 PRB. If the candidate list is not empty in any cells, then go to step 5; otherwise complete the scheduling in this TTI.

5. The user k in cell l that has the highest priority factor will get one PRB. The variable n_k indicating that the number of allocated to user k is updated to 1. The TBS of user k is updated and compared with the total buffer size (PDCP+RLC layer buffer). The priority factor of user k in cell l is updated by the formula below.

   Heuristic method 1 $\quad m_{k,l} = u_{_{k,l}}\left(n_{k,l} + 1\right) - u_{k,l}\left(n_{k,l}\right)$

   Heuristic method 2 $\quad m_{k,l} = \frac{u_{_{k,l}}(n_{k,l}+1) - u_{k,l}(n_{k,l})}{\sigma_{i,c}}$

   - If the current TBS can empty the buffer in this TTI, then the user k is served with the current TBS. Afterwards, the user k will be removed from the list b which means the radio resource allocation process for the user is finished. Then go to step 6;
   - Otherwise, go to step 6 directly.

6. Update the estimated achievable S1 throughput:

$$r_{est} = r_{est} + 0.18 \cdot \sigma_{k,l}$$

   r_{est} (initial value is set to 0 Mbps) is the estimated achievable S1 throughput with the unit of Mbps considering that the bandwidth of a PRB is 0.18 MHz. $\sigma_{k,l}$ is the channel quality related

coefficient for user k in cell l who gets a PRB in the current iteration. To be noticed, since the transmission is unpredictable, the throughput estimation is updated no matter the current transmission is successful or not. However, the retransmissions are not been taken into account. Therefore, all successful transmissions are counted for only one time.

- If $r_{est} \geq R_{S1}$, go to step 9;
- b) Otherwise, go to step 7.

7. Check whether there are any users and PRBs left in cell l.
   - If there is no remaining user in the candidate list, then go to step 8;
   - In case the candidate list is not empty, if there is no PRBs left in cell, then serve the remaining users in cell l and empty its candidate list.
8. Check whether there are any users and PRBs left in all cells.
   - If there are any users and any PRBs left, then go to step 5;
   - Otherwise, go to step 9.
9. Serve the remaining users in the candidate lists in all the cells based on the TBS sizes. The scheduling process for the current TTI is finished.

5.5.1 Simulation Results

5.5.1.1 Single Cell Scenario

In this part, the performance of the proposed heuristic is evaluated by simulations and compared to the solution of the optimization problem. Besides, the heuristic without considering the transport limitation (marked as w/o S1) is added for comparison. It is been proven that

TABLE 5.4: Scenario settings (single cell scenario)

Number of UEs in femtocell	4 Video users, 2 Web users, 2 FTP users
Traffic type	Video: full buffer, TCP based HTTP: 2 MB pagesize, reading time is 0s FTP: 10 MB pagesize, idle (reading) time is 0s
Transport limitation	16 Mbps, FIFO queue, 1 Mbytes buffer size
Number of PRBs	50 PRBs (10 MHz)

without transport limitations, the heuristic is optimal (see Chapter 4). As a result, it is meaningful to check the performance in case of transport limitation.

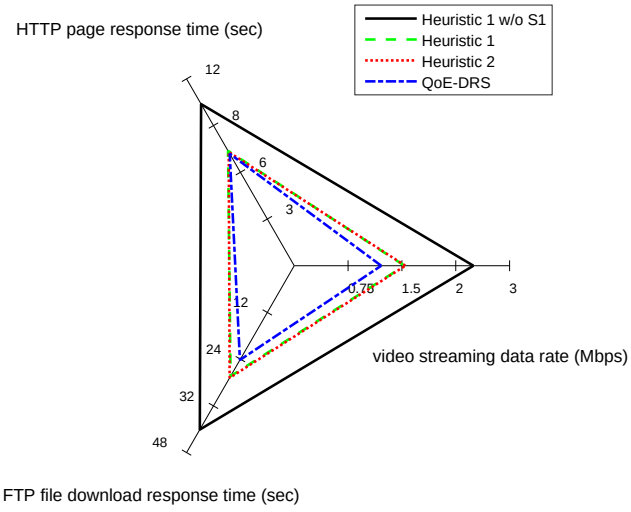


FIGURE 5.23: Application performance comparison

Figure 5.23 and Figure 5.24 show the application performance and their corresponding MOS (the error bars show the 95% confidence interval). Figure 5.24 shows the average user QoE calculated over multiple

samples. For each sample, the MOS is calculated based on a 5 seconds batch mean of the throughput since the human beings have the sense of user satisfaction in seconds instead of milliseconds [Shn84]. The error bars show the 95% confidence intervals, but they are very small and hardly visible. Considering the minimum MOS value (scaling between 1 and 5) of each user is 1, the gain is calculated based on (5.33).

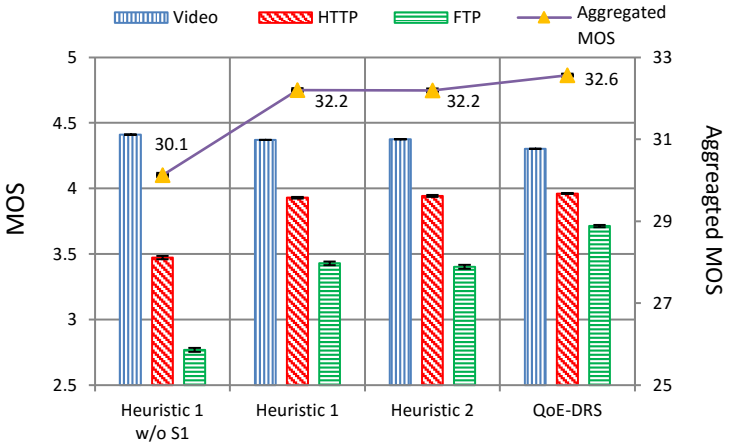


FIGURE 5.24: MOS comparison

For all three types of users, the application performance applying the heuristic approaches and the optimal approach is quite close (the three triangles shown in Figure 5.23). Looking at the video streaming users, using the Lagrangian method and the proposed heuristic, the video streaming data rates are reduced significantly compared to the heuristic without the awareness of the transport limitation. However, the MOS of video streaming users nearly remain the same as seen from Figure 5.24. The reason is that the video MOS has non-linear relation with the data rates. Thus, the video users can remain satisfied even with the less amount of resources. For example, the video streaming data rate is reduced from 2.08 Mbps with the heuristic without the awareness of the transport limitation to 1.01 Mbps with Lagrangian

method, while the corresponding's MOS is only reduced from 4.41 to 4.3. For the heuristic without the awareness of the transport limitation, it has been observed that it allocates too much air interface resources to the users which put excessive traffic to the transport network and cause the congestion. Due to the TCP congestion control, the TCP flows tend to have the equal share over the congested link. Consequently, video users have MOS around 4.4 while the FTP user has a MOS around 2.7. On the contrary, the HTTP and FTP users get more resources and the download response times are reduced dramatically for the Lagrangian method and proposed heuristic. With little performance degradation of video users, the satisfactions of HTTP and FTP users are improved significantly. For instance, the FTP file download time is reduced from 42.1 seconds with the heuristic without the awareness of the transport limitation to 28.8 seconds with proposed heuristic and 24.2 seconds with Lagrangian method, while the corresponding's MOS is increased from 2.77 to 3.43 and 3.71 separately. Regarding to the aggregated QoE, the proposed heuristic achieves 32.2 while the Lagrangian method gives 32.6 and the gain is around 1.5% based on formula 7. The gain is around 11.1% for the Lagrangian method against the proposed heuristic without the awareness of the transport limitation.

5.5.1.2 Multiple Cells Scenario - Asymmetric Cell Loads

In this part, the performances of the proposed heuristics are compared against the optimal bound obtained by the optimization model. Besides, the modified heuristic 1 without considering the transport limitation (marked as w/o S1) is added for comparison. The simulation results of heuristic 2 w/o S1 has similar performance as heuristic 1 w/o S1 and they are omitted in the work.

Figure 5.25 shows the average user MOS of all users in the femtocell cluster. For each user, the average MOS is a mean value over multiple samples. The 95% confidence intervals are shown with error bars, but they are very small and hardly visible. The figure shows that both

TABLE 5.5: Scenario settings
(multiple cells scenario - asymmetric cell loads)

Number of UEs in femtocell cluster	8 Video users in cell 1, 4 Web users in cell 2, 4 FTP users in cell 3
Traffic type	Video: full buffer, TCP based HTTP: 2 MB pagesize, reading time is 0s FTP: 10 MB pagesize, idle (reading) time is 0s
Transport limitation	32 Mbps, FIFO queue, 1 Mbytes buffer size
Number of PRBs	50 PRBs (10 MHz)

heuristic 1 and 2 achieve close to optimal bound results while heuristic 2 is slightly better than heuristic 1, and they all outperform the heuristic without the awareness of the transport limitations.

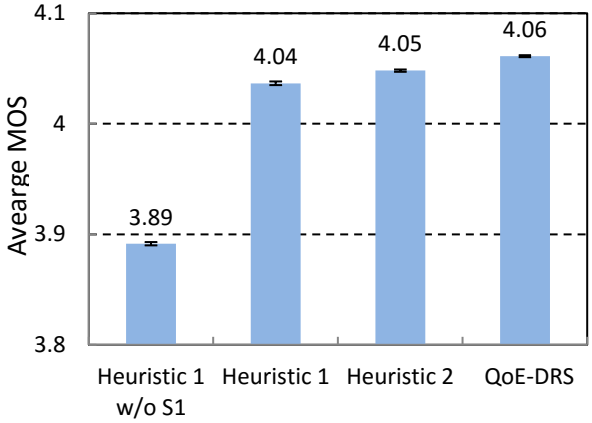


FIGURE 5.25: Average MOS of all users

Although the average MOS of all users do not differ so much among the heuristics, the performance of different user types is quite different which can be seen from the Complementary CDF (CCDF) curves in Figure 5.26. From the figure, with the heuristic 1 w/o S1, the video users get the best performance while the FTP users get worst. The

transport network is overloaded for the same reason explained the the single cell scenario. Due to the TCP congestion control, the TCP flows tend to have the equal amount of shares over the congested link. Thus the video users are very satisfied that their average MOS is over 4.3, but the FTP users only get an average MOS around 3. The optimal solution offloads some resources from video users to FTP users, and their average MOS increased to 3.9 from 3 while the video users' MOS only decreased by less than 0.1. With little performance degradation on the video users, the FTP user gets more satisfied.

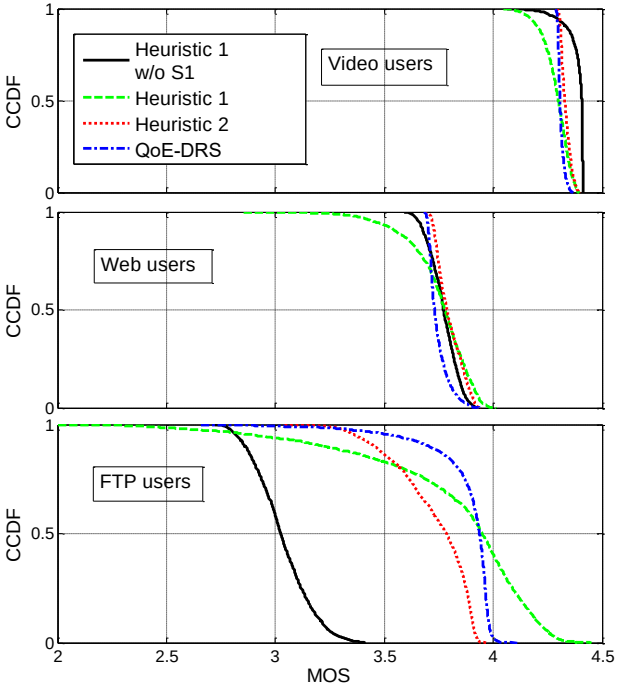


FIGURE 5.26: CCDF curves for different type of users

Both heuristic 1 and 2 show similar trends as the optimal solution. The detailed statistics of FTP users are summarized in Table 5.6. Heuristic 1 has closer average MOS to the optimal solution than

heuristic 2. However, the heuristic 1 has the worst performance on the 5^{th} percentile CDF of MOS (95^{th} percentile CCDF), which represents the worst 5% cases. Besides, the heuristic 1 has the largest standard deviation indicating the MOS fluctuates more. Additionally, heuristic 1 has the worst performance on the 5^{th} percentile CDF of MOS for video and web users as well. Heuristic 2 outperform heuristic 1 regarding to both the stability and the average MOS in the cluster. Therefore, heuristic 2 is recommended based on the analysis. More discussions will be given with the following scenarios.

TABLE 5.6: Statistics of FTP users

	5^{th} percentile CDF of MOS	Average MOS	std.
Heuristic 1 w/o S1	2.82	3.03	0.13
Heuristic 1	2.92	3.82	0.41
Heuristic 2	3.37	3.73	0.18
QoE-DRS	3.55	3.88	0.17

5.5.1.3 Multiple Cells Scenarios - Varying Traffic Load

In the last scenario, the performance of the proposed heuristics as well as the QoE-FRS mechanism are evaluated in an asymmetric scenario that there are different user types in different cells. In this part, the performance of the proposed heuristics is studied in several symmetric scenarios which were used to in Section 5.4.

Figure 5.27 and 5.28 show the application performance and their corresponding MOS for the lightly loaded scenario (scenario 1). The VoIP users consume very little amount of resources and they are prioritized throughout the network. As expected, the VoIP users are very satisfied for all of the schemes. As referred to the non-GBR traffic, with the heuristic 1 w/o S1, different users tend to get the same amount of data rate due to the congestion on the transport link. As stated in

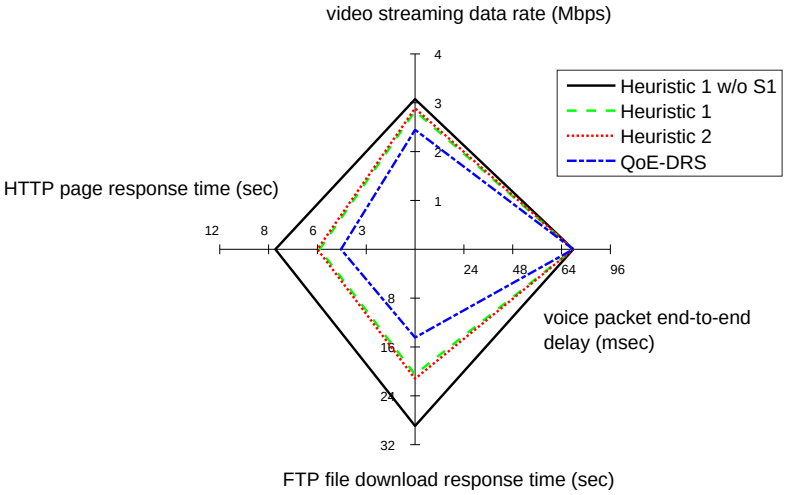


FIGURE 5.27: Application performance comparison for lightly loaded scenario

the last scenario, it allocates too much air interface resources to the users, which put too much traffic to the transport network and triggers TCP congestion control. On the contrary, with QoE-DRS, the video streaming data rate is reduced since its data rate is limited by shaping rate. Therefore, HTTP and FTP users could get more resources and the download response times are reduced substantially. The MOS of HTTP and FTP users are improved dramatically accordingly. With QoE-DRS, different users share the transport and air interface capacities in a dynamic and smart way. The satisfaction levels of HTTP and FTP users are improved significantly with only little performance degradation of the video users. Therefore, it managed to achieve the highest aggregated MOS over all the non-GBR users among the four schemes. The two proposed heuristics tend to allocate the resources in a similar way as the QoE-DRS that the HTTP and FTP users are given more resources by suppressing the data rates of the video users. The performance of the two heuristics are nearly the same. The aggregated

MOS is increased from 35.9 with heuristic 1 w/o S1 to 38.4 for both heuristics. Nevertheless, it is still worse than QoE-DRS which has an aggregated MOS of 40.7.

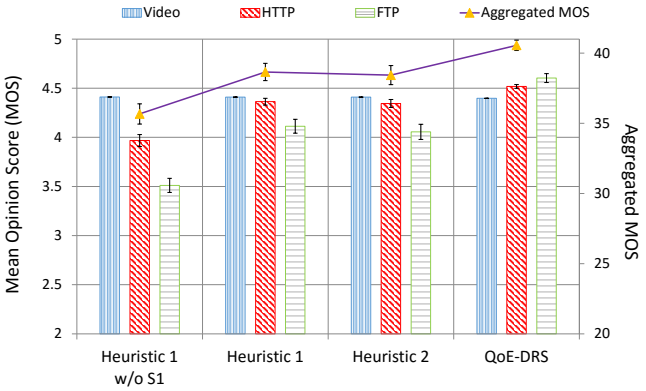


FIGURE 5.28: MOS comparison for lightly loaded scenario

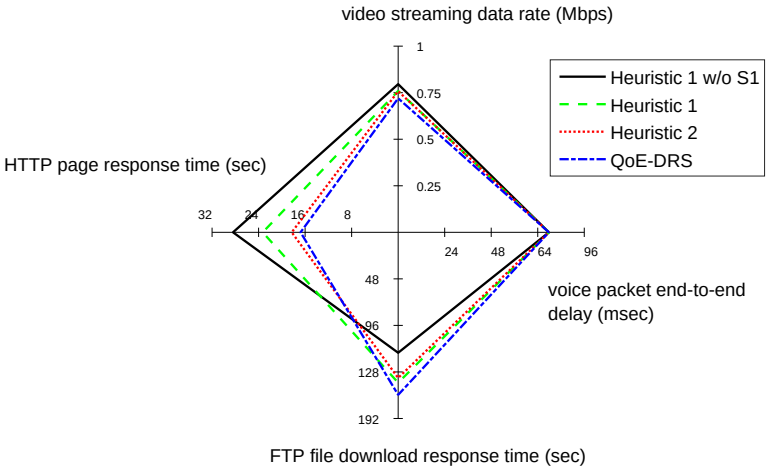


FIGURE 5.29: Application performance comparison for heavily loaded scenario

Figure 5.29 and 5.30 compare the system performance for the scenario 2 with increased load. Unlike the lightly loaded scenario, the FTP users' MOS is getting worse with QoE-DRS and the two heuristics comparing to the heuristic 1 w/o S1. This is because with increased number of users, it is not possible to satisfy all the users at the same time. The FTP users cannot be satisfied anyway, since they need more resources to achieve the same MOS comparing to video and HTTP users. Therefore, more resources are given to video and HTTP users since they need less resources to achieve a high MOS. Correspondingly, both the two heuristics and QoE-DRS obtained a higher aggregated MOS comparing to heuristic 1 w/o S1. QoE-DRS is still the best option achieving highest aggregated MOS. Between the two heuristics, heuristic 2 is better than heuristic 1, Both heuristics can fully utilize the transport link. But in heuristic 2, the user is prioritized based on the utility gain over the channel quality coefficient σ (σ is bigger with a better channel condition), while it is based only on the utility gain in heuristic 1. Therefore heuristic 2 gives higher priority to the user with a higher utility gain while putting less data on the congested transport link in each iteration when allocating PRBs to the users. The heuristic 1, however, only considers the marginal utility gain by getting an additional PRB. So heuristic 2 uses more air resources then heuristic 1. The aggregated MOS of non GBR users is increased from 77.8 to 83.3 with heuristic 2 and 84.3 with QoE-DRS.

In scenario 3, the transport link is limited to 6 Mbps instead of 16 Mbps in scenario 2 which means the transport link is very congested. Therefore in this scenario, the video users get most of the resources with QoE-DRS since the video users could be satisfied with little amount of resources. Figure 5.30 shows the HTTP and FTP users could hardly get any resources while the video users can still get very high MOS with QoE-DRS. Heuristic 2 has nearly the same performance as QoE-DRS. The aggregated MOS of non-GBR users is increased from 55.5 with heuristic 1 w/o S1 to 63.5 with heuristic 2 and QoE which has a

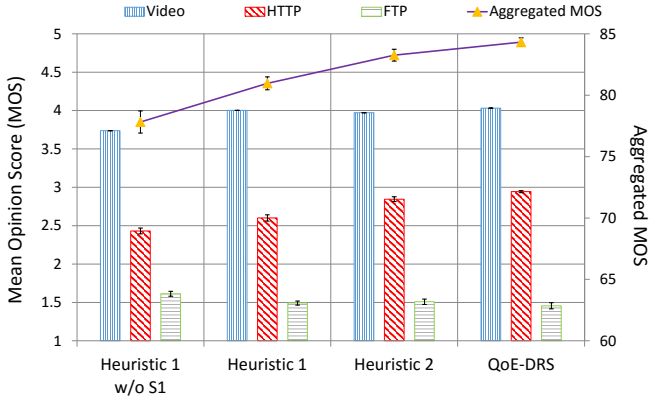


FIGURE 5.30: MOS comparison for heavily loaded scenario

gain of 31.4%. The heuristic 1 has an aggregated MOS of 61.9 which is slightly worse than heuristic 2 and QoE-DRS.

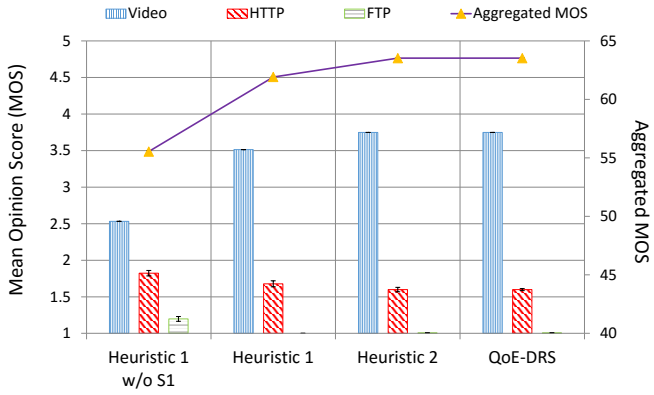


FIGURE 5.31: MOS comparison for very heavily congested loaded scenario

In all the investigated scenarios, QoE-DRS shows the best performance since it can dynamically and periodically adjust the shaping rate with the purpose of maximizing the aggregated user MOS for a cell

cluster. The proposed heuristic 1 w/o S1 behaves like the legacy PF scheduler that different users are getting the same amount of resources and therefore achieving the same data rate in long time average. Both heuristic 1 and 2 improve the system performance regarding to the aggregated MOS. Nevertheless, the heuristic 2 outperforms the heuristic 1 in the highly loaded and congested scenarios. Especially in the congested scenario, heuristic 2 has similar performance as the QoE-DRS mechanism. Therefore, heuristic 2 is highly recommended for the scenarios, providing the transport link is the major bottleneck.

5.6 Summary

In this chapter, a new resource management approach to dynamically set optimized EPS bearer shaping rates considering the limitations of both the air interface and the transport network is proposed, which maximizes the aggregated QoE in a cell cluster. This approach dynamically shares the transport link capacity among users according to their service types. Some typical scenarios, including femtocell clusters, multiple cells in an eNB and LTE C-RAN, are introduced. The problem is proven to be a concave optimization problem and is solved by Lagrangian relaxation. By numerous simulations, the proposed approach results in significant performance gain, with regard to the aggregated MOS, especially in case of heavy link congestion on the transport network in the investigated scenarios. Besides, the traffic shaping is done at the core network, and it is proven that the proposed shaping method can coexist with the existing resource management schemes at the transport network, e.g. transport network scheduler.

In order to apply this approach in real-time, the complexity of this approach is studied. The investigation suggests that an optimum periodic time interval shall be chosen as a tradeoff of signaling overhead, computation effort and achievable performance. In addition, two efficient heuristic algorithms (different only on the priority factor for

scheduling) with reduced complexity are developed for the large scale networks. Between the two heuristics, heuristic 2 is better than heuristic 1 and close to the optimal shaping method when the transport link is a major bottleneck. Both heuristics can fully utilize the transport link. But heuristic 2 uses more radio resources then heuristic 1. In heuristic 2, the user is prioritized based on the utility gain over the channel quality coefficient σ (σ is bigger with a better channel condition), while it is based only on the utility gain in heuristic 1. Therefore heuristic 2 is highly recommended.

Chapter 6

Conclusion

The thesis is targeted to enhance the resource management in LTE, with special focus on improving the end users' QoE. Firstly, the work focuses on enhancing the radio resource scheduling which is one of the most important research topics in LTE. A utility based optimal radio resource scheduling framework is proposed, which maximize the aggregated QoE in a cell. Secondly, the work is extended to consider both radio interface and transport backhaul limitations, which is specially useful for the scenarios with limited backhaul, e.g. the femtocells. A joint radio and transport optimized resource management scheme is proposed, which maximizes the aggregated QoE in a cell cluster which shares the same backhaul. Moreover, efficient heuristics with reduced complexity are proposed and evaluated.

The main work starts with an implementation of a comprehensive LTE simulation model with necessary details based on the discrete-event-based simulator OPNET Modeler in Chapter 3. The implementation done for this thesis includes all basic E-UTRAN and EPC network entities with fully implemented user-plane protocol stack according to the 3GPP specifications. The most important E-UTRAN and EPC nodes (UEs, eNodeBs, PDN-GW, S-GW) are implemented with the

respective protocols. A seven eNBs/cells hexagonal network layout, which is well known as 1-tier scenario, is implemented in this work. Besides, mobility models and a realistic channel model with considering interference among the eNBs are modeled to perform accurate simulations. Additionally, a femtocell cluster is modeled. Moreover, the existing and proposed resource management schemes in LTE, e.g. radio scheduling, transport network scheduling, traffic shaping, etc., have been developed in the simulator.

In Chapter 4, an optimized QoE based radio scheduler is proposed. At first, existing conventional scheduling methods as well as utility based scheduling strategies are summarized and discussed. It is shown by numerous examples that the QoE can be approximated by a concave relation over data rate for elastic traffic. Therefore, the QoE based radio scheduling problem becomes a convex optimization problem. The proposed QoE based radio scheduler is proven to be optimal analytically [LTTTG15b]. Besides, it allows application in real systems due to low complexity. By numerous simulations, the proposed scheduler significantly outperforms the conventional proportional fair scheduler.

In Chapter 5, both radio and transport network are considered. The most significant contribution of this thesis work is that a new resource management approach to dynamically set optimized EPS bearer shaping rates is proposed, which maximizes the aggregated QoE in a cell cluster scenario [LTTTG15a], [LTTG15], [LTTTG15b], [LTLTG15]. The algorithm is formulated as a convex optimization problem and solved by the Lagrangian relaxation method. This approach dynamically shares the transport link capacity among users according to their service types, which results in a significantly better aggregated QoE, as compared to the networks with the legacy fixed rate shaping. The simulation results show that, under the investigated scenarios the aggregated MOS value increases 18.8% in the case low traffic load, 13.6% in case of heavy load and 31.4% in case of very heavy link congestion.

In order to apply this approach in real-time, the complexity of this approach is studied. The investigation suggests that an optimum time interval (to recalculate the shaping rates by solving the optimization problem) shall be chosen as a tradeoff of signaling overhead, computation effort and achievable performance.

Although the proposed algorithm can be generalized for LTE networks, the performance evaluation focuses only on the femtocell scenarios. However, a large scale network in LTE often serves hundreds of users and solving the optimization problem may not be feasible. In order to apply the proposed approach in large scale networks, efficient heuristic algorithms are developed. The heuristics are based on the extension of the QoE based scheduling in Chapter 4 to support multiple cells and consider the transport limitation. The simulation results showed that both heuristic 1 and 2 (different only on the priority factor for scheduling) provide a better QoE for FTP users compared to the conventional resource allocation approach, in which the transport network is not taken into account. Furthermore, heuristic 2 has a better stability than 1 and results in a solution very close to the theoretical optimal bound. Therefore it is highly recommended.

The simulation model can be used for other research topics in LTE with necessary extensions. A restructured simulation model has been used by the author in the study of network virtualization and load balancing techniques in LTE networks [ZLZ+11], [LZL+12].

6.1 Outlook

In this work, the main target was to maximize the aggregated QoE in a cell or a cell cluster without considering the confliction between fairness and utilization. However, since the amount of physical channels in a cell is limited (e.g. 25 Physical Resource Blocks (PRBs) with 5MHz bandwidth in LTE), and only an integer amount of channels can be allocated to each user, users with bad channel conditions might almost

never get served, especially when the number of users is higher than the number of PRBs. Some users under bad channel conditions might starve or not be served for a long time causing a severe fairness problem. The problem can be solved through building the utility functions based on the Exponential Moving Average (EMA) rate instead of the instantaneous data rate. The advantage of this approach is, that it guarantees the users with very bad channel conditions still to be scheduled. This extension has been published in [LTWTG14], and can be found in Appendix B.

For future works, a more detailed contribution to real-time delay sensitive traffic can be studied. Some preliminary work on this extension has already been proposed in Appendix A. A utility function over average waiting times could be constructed for real-time traffic. A Max-Delay-Utility (MDU) scheduling method scheduling strategy has been evaluated in LTE networks.

In addition, the proposed resource management schemes are designed for LTE downlink. Nevertheless, the mathematical models in this work are applicable to both LTE downlink and uplink. In the future, practical modifications for LTE uplink can be studied.

Appendix A

QoE-based Scheduling for Real-time Services

In LTE, GBR bearers (QCI 1 to 4) provide services with guaranteed minimum data rates to high priority real time traffic, e.g. conversational voice or videos, real time gaming, etc. These services are referred as QoS traffic since they have tighter requirements on QoS performance, such as delay, packet loss rate, and jitter. The QoS traffic is scheduled prior to non QoS traffic based on channel-aware, QoS-aware scheduling strategies introduction in Chapter 4.1.2. For example, Scheduler for Guaranteed Data-Rate [MPKM08] and Scheduler for Guaranteed Delay Requirements (modified LWDF seen in [AKR+01], EXP seen in [CJS+07]) are designed to guarantee the minimum requited data rates or delay constraints performance.

There are two difficulties in designing joint channel- and queue aware scheduling algorithms [SL05]. First, there are many QoS requirements such as average delay, delay violation probability, jitter, etc., making it difficult to formulate the optimization goals. Second, the optimal solution algorithms are normally with exponential complexity. Unlike most joint channel and queue-aware scheduling policies,

[SL05] proposes a Max-Delay-Utility (MDU) scheduling method with an explicit optimization objective of maximizing the total utility with respect to the predicted average waiting times for OFDMA wireless networks. It further proves that the optimization objective turns out to be a linear function of the user data rate.

Nevertheless, the authors did not introduce how a utility function over average waiting times could be constructed. Besides, the MDU scheduling strategy has not yet been investigated and applied in LTE networks. These motivates us to study constructing utility functions considering QoE for real time traffic, apply the MDU scheduling method in LTE networks, and evaluate its performance by simulations.

A.1 Real Time Services

A.1.1 VoIP

VoIP services are based on the Real-Time Protocol (RTP) at the application layer. The voice service is modeled based on ON/OFF model representing the talk spurt and silence periods. The GSM Enhanced Full Rate (EFR) codecs is one the most widely used voice codecs. It belongs to the family of Adaptive Multi-Rate (AMR) codecs, with an application data rate of 12.2 kbps.

A.1.2 Real-time Video Conferencing

Similar to VoIP services, real-time video steaming services are based on RTP. A Constant Bit Rate (CBR) video model is adopted with two important parameters: frame size and inter-frame time interval (frame rate).

In this study, Microsoft's Lync Server 2013 online support documentation is taken as a reference for deciding the codec type and required bandwidth [Mic13]. There are many codec types for video applications and ITU's H.264 is one of the most common ones. It is

developed together with Moving Picture Experts Group (MPEG) and is also known as MPEG-4 Part 10 AVC (Advanced Video Coding). Another benefit of using H.264 is that it offers a scalable encoding rates, which can range from low rates such as 150 kbps and up to a few Mbps [Mic15]. For a peer-to-peer video conferencing application with H.264, Microsoft Lync Server reserves a bandwidth of 460 kbps for a typical stream.

A.2 E-model

ITU's E-model is one of the computational methods that predicts a user experience based on the equipment impairment factors [IT15]. Although the E-model is mainly used for assessing voice transmission qualities, there are some other ITU recommendations also mention its usage as a reference for other delay sensitive applications such as in [IT03]. The E-model estimates VoIP call quality on Mean Opinion Score (MOS) scale, which is considered as an important KPI (Key Performance Indicator) to characterize the end VoIP user Quality of Experience (QoE). Based on the empirical studies of an IP network a set of IP impairment profiles has been defined to emulate impairments (e.g. packet loss rate, packet delay and delay variation) in the transport network of LTE. MOS values are computed for VoIP calls made by users in the presence of IP impairments.

The primary output of the E-model is the "rating factor" R or R-factor which can be mapped to MOS to estimate the customer opinion on voice quality. The E-model assesses conversational voice quality by establishing a relationship between objectively measurable factors and subjective assessment of voice quality based on large scale of measurements. For a narrowband codec, the maximum value of R-factor computed by the E-model is 100 which corresponds the best possible achievable voice quality. Minimum value of R factor is 0, which is the

worst case. R-factor combines several transmission parameters considered relevant for an end-to-end transport connection as well as others which cover impairments due low bit rate codec, echo, background noise and electronic equipment, etc.

The R-factor is composed of five factors as stated below:

$$\mathrm{R} = \mathrm{R}_o - I_s - I_d - I_{e-eff} + A \tag{A.1}$$

where R_o represents the basic signal-to-noise ratio of a given environment of the talker. Factor I_s is the sum of all impairments which may occur more or less simultaneously with the voice transmission. I_d represents impairments caused by mouth-to-ear path delay. The advantage factor A provides compensation for impairment factor in return of other advantages enjoyed by the user, e.g. ease of access etc. I_{e-eff} is packet loss dependent Effective Equipment Impairment factor.

In OPNET Modeler, there are some pre-set values of E-model parameters for various scenes which is shown in Table 2.2. Under the assumption that the packet loss is random (independent), $\mathrm{I}_{\mathrm{e-eff}}$ is calculated using the equation with the packet-loss probability P_{pl}:

$$\mathrm{I}_{e-eff} = I_e + (95 - I_e)\,\frac{P_{pl}}{B_{pl} + P_{pl}} \tag{A.2}$$

TABLE A.1: The pre-set parameters of E-model in OPNET for various cases

Scene	**Communication System**	**Ro**	**Is**	**A**
Land Phone - Quiet Room	Conventional wirebound	94.77	1.43	0
Land Phone - Noisy Room	Conventional wirebound	90.74	5.67	0
Cell phone in building	Cellular Mobility in a building	85.91	2.32	5
Cell phone in SUV or sedan	Mobility across geographical area	80.73	3.24	10

The R-factor produced by the E-model can be mapped to MOS which ranges from 1 to 5, 1 being the worst and 5 the best perceived

quality. The expression used to map R-factor onto the MOS scale can be found in Appendix B of ITU-T G.107 [IT15].

$$MOS = \begin{cases} 1 & R \leq 0 \\ 1 + 0.035 \cdot R + R \cdot (R - 60) \cdot (100 - R) \cdot 7 \times 10^{-6} & 0 < R < 100 \\ 4.5 & R \geq 100 \end{cases} \tag{A.3}$$

In LTE scheduling, since the scheduler has no information on the packet loss, the construction of QoE based utility function only takes the packet delay into consideration in this preliminary work. Figure A.1a as R of E-model and in Figure A.1b as MOS value together with its curve fit for GSM EFR codec in the "Cell phone in SUV or sedan" scenario. The chosen curve fit function is:

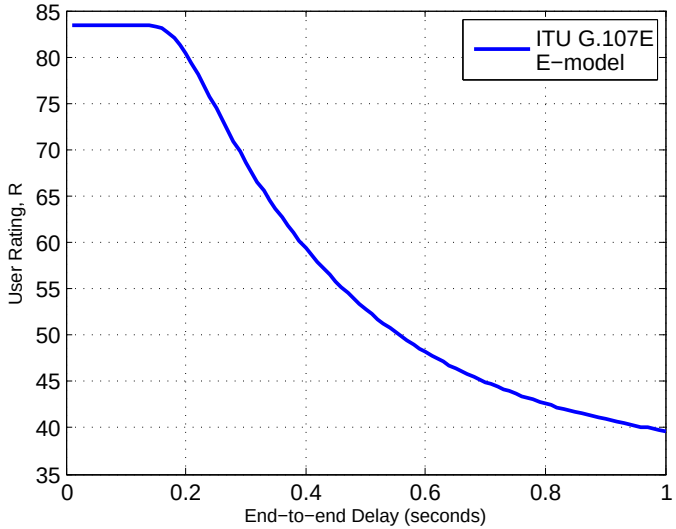
$$u(d) = \frac{A}{1 + e^{-\alpha \cdot (d - D)}} + B = \frac{A}{2}\left[\tanh\left[\frac{\alpha \cdot (d - D)}{2}\right] + 1\right] \tag{A.4}$$

The curve fit values of above parameters are: $A = 2.309$, $B = 2.043$, $\alpha = -8.105$ and $D = 0.3744$. The norm of residuals for this curve fitting is 0.7663.

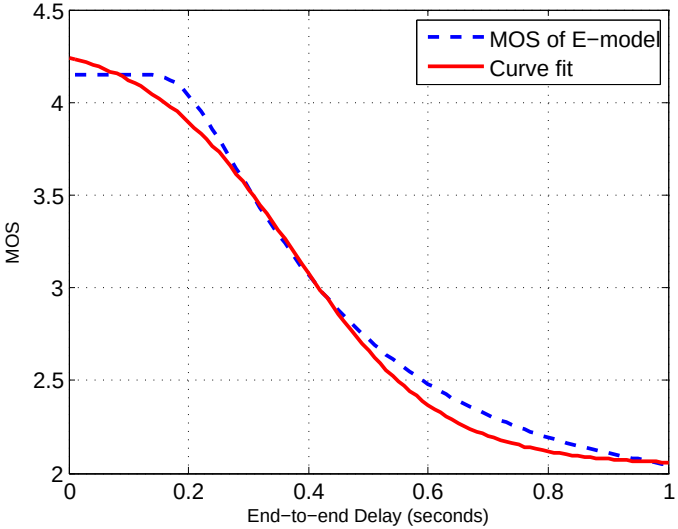
A.3 Max-Delay-Utility (MDU) Scheduling

MDU scheduling method is proposed by [SLC09] with an explicit optimization objective of maximizing the total utility with respect to the predicted average waiting times. It further proves that the optimization objective turns out to be a linear function of the user data rate in OFDMA wireless networks. The problem formulation is shown below:

$$\begin{aligned} \text{maxU}, U &= \sum_{i=1}^{N} \frac{|u'_i(d_i)|}{\rho_i} r_i(n_i) \\ s.t. \quad & \sum_{i=1}^{N} n_i = R, n_i \in Z^+ \\ & r_i(n_i) \leq \frac{Q_i}{T_S} \end{aligned} \tag{A.5}$$



(A) R value of delay sensitive applications



(B) MOS value of delay sensitive applications and its curve fit

FIGURE A.1: User satisfaction model of delay sensitive applications

Where $u_i'(d_i)$ is the first order derivative of $u_i(d_i)$, which is the utility of user i calculated based on the average delay time d, updated based on the exponential moving average (EMA) method. The delay time has two components: the delayed time over the network and the waiting time at the eNodeB to be scheduled. The delayed time over the network can be obtained from Deep Packet Inspection (DPI) of RTP header. ρ_i is the arrival data rate, which reflects the codec rate. n_i is a non-negative integer value that is the amount of PRBs allocated to the user. Q_i is the buffer size of user i and T_s represents one TTI slot. $r_i(n_i) \leq \frac{Q_i}{T_S}$ means that the scheduler should not allocate more number of PBRs which is required to empty its buffer. At each TTI, both $u_i(d_i)$ and ρ_i are constant values. $r_i(n_i)$ can be seen as a linear function of n_i ($r_i(n_i) = 180kHz \cdot \sigma_i \cdot n_i$, see chapter 4.2). Therefore, the objective function is reformulated as:

$$\begin{aligned} &\text{maxU}, U = 180kHz \cdot \sum_{i=1}^{N} m_i \cdot n_i \\ &s.t. \quad \sum_{i=1}^{N} n_i = R, n_i \in Z^+ \\ &r_i(n_i) \leq \frac{Q_i}{T_S} \end{aligned} \tag{A.6}$$

where $m_i = \frac{|u'_i(d_i)|}{\rho_i} \cdot \sigma_i$ and it is a constant value in the current TTI. $|u'_i(d_i)|$ can be calculated as:

$$|u'_i(d_i)| = \left| \frac{A \cdot \alpha}{4} \left\{ 1 - \tanh^2 \left[\frac{\alpha \cdot (d_i - D)}{2} \right] \right\} \right| \tag{A.7}$$

The optimization problem then is transformed into a very simple linear optimization problem. The optimal solution is just to allocate PRBs to the user with biggest m_i first until its buffer can be emptied. m_i depends on three factors: the arrival data rate ρ_i, the user channel condition indicator σ_i, and the absolute value of marginal utility $|u'_i(d_i)|$. The arrival rate ρ_i is related to the codec rate of the application. The users who introduce lower loads are preferred to get served

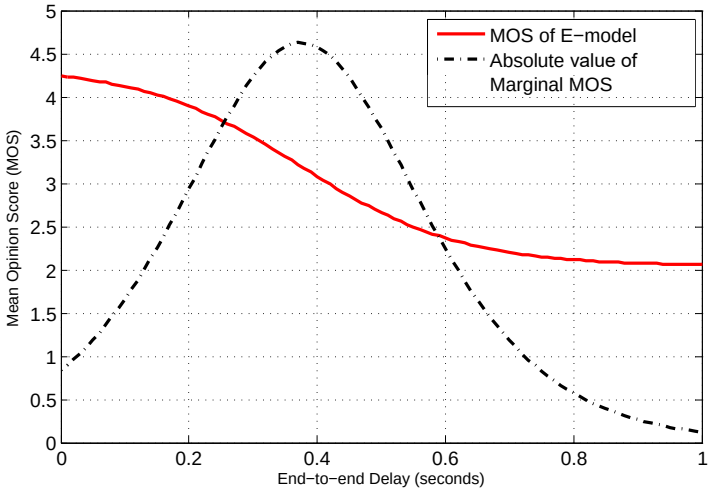


FIGURE A.2: MOS and absolute value of marginal MOS over delay

first. Under limited resource conditions, this is vital for achieving a high aggregated user satisfaction level.

Figure A.2 shows the MOS and the absolute value of marginal MOS over the delay. The absolute marginal utility has the highest value near when end-to-end delay is a little below 400 milliseconds. When end-to-end delay becomes very low or larger, the absolute marginal utility decreases. Therefore, the users who are around middle delay region are prioritized over users who have either already low delay values or who are beyond acceptable delay values. This approach can be seen as trying to improve the situation for users who have the potential to achieve lower delays and higher user satisfaction. Meanwhile, the users who are already suffering from very high end-to-end delays are less likely to reach acceptable delay regions. So prioritizing these users will not improve the situation, hence they have a lower priority than others.

Figure A.3 shows the flow chart of the scheduling procedure for real-time traffic. The scheduling is performed periodically every 1 ms

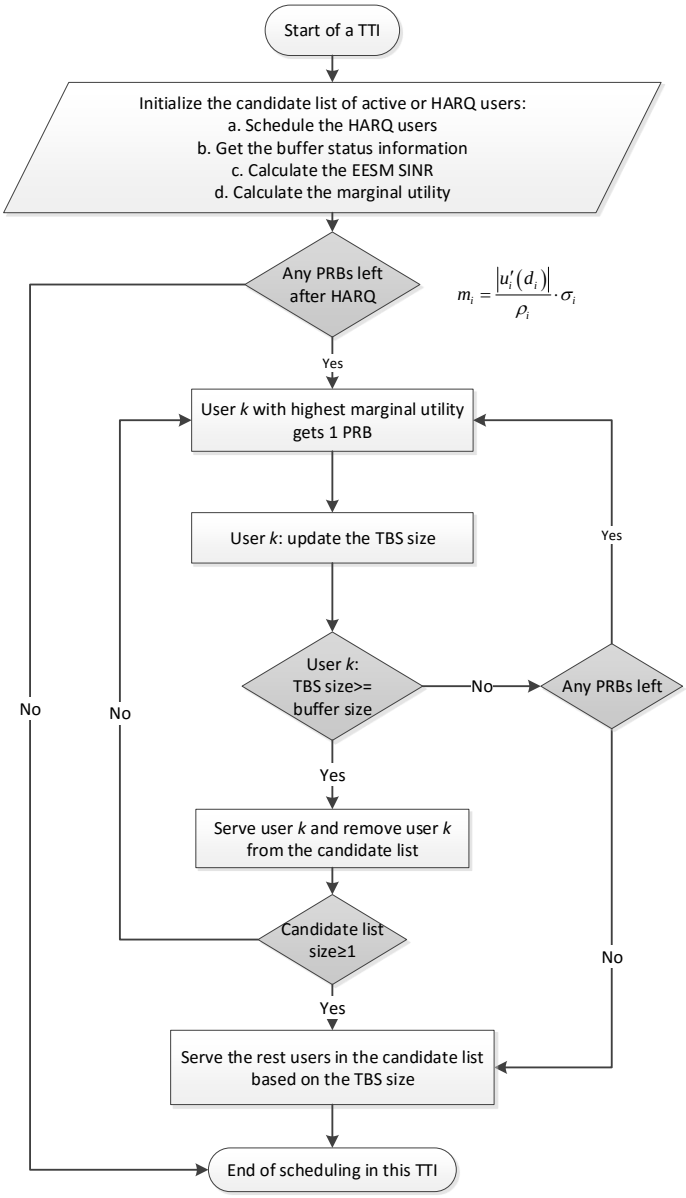


FIGURE A.3: Flow chart of the delay heuristic algorithm

in each cell. The retransmission users are assured with highest priority. If any resources left, the resources are allocated to the active users in an iterative manner. The detailed explanations can be found in Chapter 4.2.5.

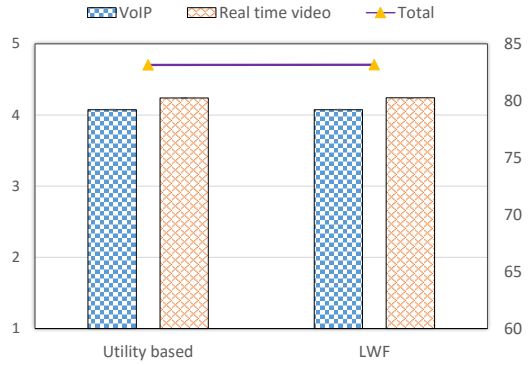
A.4 Simulation Scenario and Results

In this section, the proposed MDU scheduling framework is compared against with the LWF scheduler with real time traffic only scenarios. The MDU scheduler is designed with the purpose of maximizing the aggregated user satisfaction based on the proposed utility functions. As a comparison, the LWF scheduler simply priorities the user whose packets have the longest waiting time in the queues of eNodeB. The MDU scheduler prioritizes the candidate users based on their channel conditions, data rates and the estimated delay. Under limited resource condition, it prefers the users with lower data rates over high data rates in order to achieve a higher aggregated MOS. The scenario 1 is designed to prove this property by simulation. The other property of the MDU scheduler is that, if the users have already experienced very high delay over the backhaul or at the eNB, they are very unlikely to get satisfied and therefore they are scheduled with a low priority. Correspondingly, the scenario 2 is designed to prove this concept.

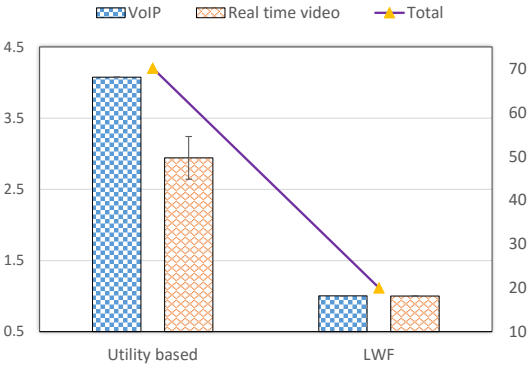
A.4.1 Scenario 1 with Varying Cell Bandwidth

There are 10 VoIP and 10 video conferencing users in one cell with 100 Mbps S1 capacity. The available number of PRBs is set to 25 and 6 PRBs in two different simulation runs, corresponding to a cell bandwidth of 5 MHz and 1.4 MHz respectively. The results for both cases are shown in Figure A.4. It can be seen from the figure that both schedulers have very similar performance when there are 25 PRBs available. Since the air interface has enough resources serving all the users, all the

users can achieve very high MOS with both schedulers. Nevertheless, when the number of PRBs is reduced to 6, the air interface does not have enough resources satisfying all the users. This can be observed from Figure A.4b that all the users have the lowest MOS of 1 with LWF scheduler.



(A) The average and aggregated MOS values with 25 available PRBs



(B) The average and aggregated MOS values with 6 available PRBs

FIGURE A.4: The average and aggregated MOS values for RT traffic only cases

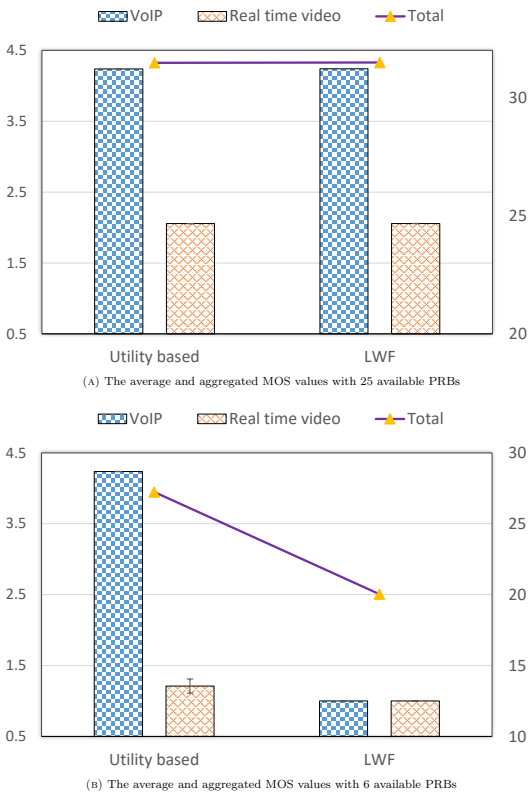
The MDU scheduler, on the other hand, prioritize the VoIP over video conferencing users since VoIP users need significantly less resources to get satisfied. Therefore, all the VoIP users are very satisfied

with the mean MOS over 4. The rest resources are given to the video conferencing users. Nevertheless, not all the conferencing users can be satisfied at the same time due to the limited resources. Some video conferencing users under bad channel conditions are scheduled with the lowest priority and experience high delay at eNB. If the delay becomes larger than 400 ms, the absolute marginal utility decreases, and therefore the scheduling priority will be further reduced. As a result, part of the video conferencing users are served achieving a higher MOS than 4, while the rest video conferencing users are hardly been served with a MOS of 1. In average, the mean MOS over the video conferencing users is around 3. By scarifying some users with bad channel conditions and high coding rates, the rest users can still be served and satisfied.

A.4.2 Scenario 2 with Varying Delays over the Backhaul

One advantage of the MDU scheduler is that it is not only aware the buffering delay of a packet at the eNB, but it also can take the total experienced delay of a packet since its creation at the RTP layer of a source node. If a packet has already experienced a very high delay over the backhaul, the packet is very likely to be outdated and useless at the application layer. On the other hand, if a packet has experienced a very low delay over the backhaul, the packet is still tolerable for some additional buffering delay at the eNB. Therefore, only the packet experienced with a middle range delay is preferred to be scheduled with the highest priority by the MDU scheduler.

In scenario 2, there are 10 active video conferencing users in the network and 5 of them are delayed by 0.7 second additionally at the backhaul. It can be seen from the Figure A.5, similarly to the scenario 1, the MDU scheduler has very close performance as the LWF scheduler with 25 available PRBs since there are enough resources to serve all the 10 users. However, when the number of available PRBs is reduced to



(A) The average and aggregated MOS values with 25 available PRBs

(B) The average and aggregated MOS values with 6 available PRBs

FIGURE A.5: The average and aggregated MOS values for video conferencing users with and without artificial delay

6, all the video conferencing users have the lowest MOS of 1 with LWF scheduler. This is because the LWF scheduler does not differentiate the users and try to serve all the users. Nevertheless, no user can be satisfied. The MDU scheduler, on the other hand, serves the 5 users without additional delay over the backhaul first, and these 5 users have a mean MOS value over 4. The rest resources are preferred allocating to the rest users with good channel conditions. The mean MOS for the 5 users with the additional backhaul delay is slightly better than

1. Generally speaking, the MDU scheduler gives up the users with the additional backhaul delay since they cannot achieve a high MOS anyway due to the high delay. The MDU scheduler is able to use the limited resources in a more intelligent way than the LWF scheduler and therefore can obtain a higher aggregated MOS in case there is not enough resources to serve all the real time users.

Appendix B

Radio Resource Allocation based on Moving Average Rates

In this work, a new algorithm is proposed to solve the downlink resource allocation problem in LTE networks taking the channel conditions into account. The problem is formulated as a convex optimization problem, which maximizes the aggregated utility of all users. To maximize the cell utility, the users with good channel conditions tend to get more resources because the utility increases with the data rate. However, since the amount of physical channels in a cell is limited (e.g. 25 Physical Resource Blocks (PRBs) with 5 MHz bandwidth in LTE), and only an integer amount of channels can be allocated to each user, users with bad channel conditions might almost never get served, especially when the number of users is higher than the number of PRBs. Since the existing algorithms do not consider the history of user data rates, some users might starve or not be served for a long time causing a severe fairness problem. This motivates the author to develop a new utility-based

approach for the resource allocation problem using the Exponential Moving Average (EMA) data rate. The advantage of this approach is that it guarantees the users with very bad channel conditions still to be scheduled, even if the number of Physical Resource Blocks (PRBs) is smaller than the number of users.

B.1 Problem Formulation and Solution

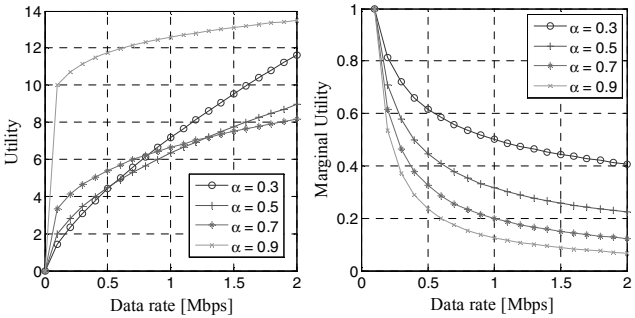


FIGURE B.1: Example of a utlity function

The utility as a function of data rate with the characteristics shown in Figure B.1 can be mathematically given by eq. (B.1):

$$u(r) = \begin{cases} \frac{r^{1-\alpha}}{1-\alpha}\alpha \geq 0, \alpha \neq 1 \\ \log(r)\alpha = 1 \end{cases} \tag{B.1}$$

where r is the user's data rate and α represents the traffic (or user) type. The influence of α on the utility can be seen in Figure B.1. Instead of using the instantaneous data rate as seen in many other papers [KL05], [KL08], and [CWC+11], we use the Exponential Moving Average (EMA) data rate . The use of EMA rate allows users with bad channel condition to be eventually scheduled. This is because

EMA rate takes the history of users' data rate into account and hence avoids a user getting starved or not being served for a long time.

The EMA data rate of user i at time t is given by:

$$\bar{r}_{i,t} = \overbrace{(1-\beta)\bar{r}_{i,t-1}}^{d_{i,t-1}} + \beta r_{i,t} = d_{i,t-1} + \beta r_{i,t} \tag{B.2}$$

$r_{i,t}$ is the estimation of user i's maximum instantaneous achievable date rate at time t, derived by the Shannon formula based on the bandwidth $b_{i,t}$ allocated to the user and the user's effective SINR. The effective SINR is calculated over all the available PRBs in its serving cell following the commonly used Exponential Effective SINR Mapping (EESM) method.

$$r_{i,t} = \sigma_{i,t} c_{i,t} \tag{B.3}$$

β is a constant smoothing factor between 0 and 1. A higher β discounts the older data rate faster. $\beta = 1$ results in an instantaneous rate.

In this work, we assume users only have positive utilities and correspondingly, $0 < \alpha < 1$. Since the radio resource allocation is done periodically every TTI (Transmission Time Interval) in LTE, we now focus on the solution algorithm for one TTI. The time index t is therefore discarded in the following part of this paper. The utility function of the user i can be reformulated as:

$$u_i\left(b_i, c_i, d_i, \alpha_i\right) = \frac{\left(d_i + \beta \cdot b_i \cdot c_i\right)^{1-\alpha_i}}{1-\alpha_i} \quad 0 < \alpha_i < 1 \tag{B.4}$$

Thus, the goal for the resource allocation is to maximize the aggregated utility, which can be expressed as:

$$\begin{aligned} \max U &= \sum_i u_i = \sum_i \frac{\left(d_i + \beta \cdot b_i \cdot c_i\right)^{1-\alpha_i}}{1-\alpha_i} \\ \text{s.t.} \quad & \sum_i b_i \leq B_c \quad \textit{and} \quad b_i \geq 0; \qquad b_i \in R \end{aligned} \tag{B.5}$$

where B_c is the total cell bandwidth.

The problem is convex and the Slater's condition is satisfied with $b_i = 0\,(\forall i)$, it has therefore a strong duality, which can be solved optimally using the Lagrangian decomposition method.

B.2 Lagrangian Dual Problem Formulation

The Lagrangian dual problem is:

$$\min_{\lambda}\left\{\max_{\{b_i\}}\left\{L=\sum_i \frac{(d_i+\beta_i c_i b_i)^{1-\alpha_i}}{1-\alpha_i}-\lambda\left(\sum_i b_i - B\right)\right\}\right\} \quad \text{s.t.} \quad \forall b_i \geq 0 \tag{B.6}$$

Consider the problem:

$$q=\max_{\{b_i\geq 0\}}\{L\}=\sum_i \max_{\{b_i\geq 0\}}\left\{\overbrace{\frac{(d_i+\beta_i c_i b_i)^{1-\alpha_i}}{1-\alpha_i}}^{L_i}-\lambda b_i\right\}+\lambda B \tag{B.7}$$

Lemma 1: $\forall b_i \geqslant 0, L_i$ is a concave function and has one and only one maximum L_i^*. Then the dual problem becomes $\min\limits_{\lambda} q = \min\limits_{\lambda}\left\{\sum\limits_i L_i^* + \lambda B\right\}$.

Proof: the function L_i is twice differentiable and its second derivative given below is negative.

$$\frac{\partial(\partial L_i)}{\partial b_i^2} = -\alpha_i b_i {c_i}^2 \beta_i^2 (d_i+\beta_i c_i b_i)^{-\alpha_i - 1} < 0 \quad \forall b_i \geq 0 \tag{B.8}$$

L_i reaches its maximum at b_i^* if $\frac{\partial L_i}{\partial b_i}(b_i^*) = 0$ and ${b_i}^* \cdot \sigma_i \geq b_{i0}$. Therefore, $b_i^* = \frac{1}{\sigma_i}\left[\frac{2}{\alpha_i}\text{arctanh}\left(\frac{\sqrt{A\alpha_i(A\alpha_i-4\lambda)}}{A\alpha_i}\right)+\mathrm{r_{i0}}\right]$ when $\lambda \leq \frac{A\alpha_i}{4}$. The formulation of L_i^* is as follows:

$$L_i^* = \max_{b_i \geq 0} L_i = \begin{cases} \frac{\alpha_i}{1-\alpha_i} \left(\frac{\lambda}{c_i \beta_i} \right)^{\frac{\alpha_i - 1}{\alpha_i}} + \frac{\lambda d_i}{c_i \beta_i} & \lambda \leq \frac{c_i \beta_i}{{d_i}^{\alpha_i}} \\ \frac{d_i^{1-\alpha_i}}{1-\alpha_i} & \lambda > \frac{c_i \beta_i}{{d_i}^{\alpha_i}} \end{cases} \tag{B.9}$$

Lemma2: f has one and only one minimum, which is also the optimal solution to the primal problem in (5).

Proof: Let $\lambda_i = \frac{c_i \beta_i}{{d_i}^{\alpha_i}}$

Without loss of generality, we assume: $0 \leq \lambda_1 \leq \lambda_2 \leq ... \leq \lambda_N$

According to Equation (9),

$$\sum_i {L_i}^* = \begin{cases} \sum\limits_{i=1}^{N} \left(\frac{\alpha_i}{1-\alpha_i} \left(\frac{\lambda}{c_i \beta_i} \right)^{\frac{\alpha_i - 1}{\alpha_i}} + \frac{\lambda d_i}{c_i \beta_i} \right) & \lambda \leq \lambda_1 \\ \sum\limits_{i=k+1}^{N} \left(\frac{\alpha_i}{1-\alpha_i} \left(\frac{\lambda}{c_i \beta_i} \right)^{\frac{\alpha_i - 1}{\alpha_i}} + \frac{\lambda d_i}{c_i \beta_i} \right) + \sum\limits_{i=1}^{k} \frac{d_i^{1-\alpha_i}}{1-\alpha^i} & \\ \qquad \lambda_k < \lambda \leq \lambda_{k+1} \forall k \in [1, N-1] & \\ \sum\limits_{i=1}^{N} \frac{d_i^{1-\alpha_i}}{1-\alpha^i} \quad \lambda > \lambda_N & \end{cases} \tag{B.10}$$

The function $q(\lambda) = \sum\limits_i L_i^* + \lambda B_c$ is visualized in figure.

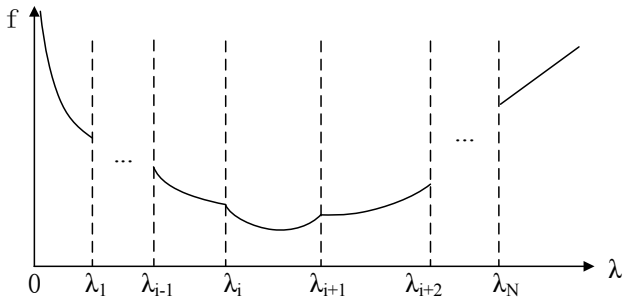


FIGURE B.2: Visualization of function $f(\lambda)$

The function q is continuous. Except at the points where $\lambda = \lambda_i$- as shown in Fig. 2, it is twice differentiable and its second derivative is non-negative as shown below:

$$\frac{\partial\left(\partial L_i^*\right)}{\partial \lambda^2} = \begin{cases} \frac{1}{\alpha_i {c_i}^2 \beta_i^2}\left(\frac{\lambda}{c_i\beta_i}\right)^{-\frac{1+\alpha_i}{\alpha_i}} & \text{if} \quad \lambda \leq \frac{c_i\beta_i}{{d_i}^{\alpha_i}} \\ 0 & \text{if} \quad \lambda > \frac{c_i\beta_i}{{d_i}^{\alpha_i}} \end{cases} \tag{B.11}$$

$$\frac{\partial(\partial q)}{\partial \lambda^2} = \sum_i \frac{\partial(\partial L_i^*)}{\partial \lambda^2} \geq 0 \quad \forall \lambda \tag{B.12}$$

Therefore, q has one and only one minimum. Due to the strong duality, the solution of the dual problem is also the solution of the primal problem.

B.2.1 Solution of the Dual Problem

In order to find the minimum of q, the interval $(\lambda_i, \lambda_{i+1})$ containing it needs to be determined. This is done as follows: Let $\lambda_0 = 0$, at each λ_i, starting from $i = 0$, compute: $\left.\frac{\partial q}{\partial \lambda}\right|_{\lambda=\lambda_i+\delta}$ and $\left.\frac{\partial q}{\partial \lambda}\right|_{\lambda=\lambda_i-\delta}(\delta \to 0)$ as follows:

$$\left.\frac{\partial q}{\partial \lambda}\right|_{\lambda=\lambda_i-\delta} = \sum_i \left.\frac{\partial L_i^*}{\partial \lambda}\right|_{\lambda=\lambda_i-\delta} = \sum_{k=i}^{N}\left(-\frac{1}{c_k\beta_k}\left(\frac{\lambda_i}{c_k\beta_k}\right)^{-\frac{1}{\alpha_k}} + \frac{d_k}{c_k\beta_k}\right) + B_c \tag{B.13}$$

$$\left.\frac{\partial q}{\partial \lambda}\right|_{\lambda=\lambda_i+\delta} = \sum_i \left.\frac{\partial L_i^*}{\partial \lambda}\right|_{\lambda=\lambda_i+\delta} = \sum_{k=i+1}^{N}\left(-\frac{1}{c_k\beta_k}\left(\frac{\lambda_i}{c_k\beta_k}\right)^{-\frac{1}{\alpha_k}} + \frac{d_k}{c_k\beta_k}\right) + B_c \tag{B.14}$$

At $\lambda_0 = 0$, we have: $\left.\frac{\partial q}{\partial \lambda}\right|_{\lambda\to 0} = \sum_i \left.\frac{\partial L_i^*}{\partial \lambda}\right|_{\lambda\to 0} = -\infty$

Therefore, if $\left.\frac{\partial q}{\partial \lambda}\right|_{\lambda=\lambda_1-\delta} > 0$, the minimum is in the interval $(0, \lambda_1)$. Otherwise,

- for $i > 0$, if $\left(\frac{\partial q}{\partial \lambda}\Big|_{\lambda=\lambda_i+\delta}\right)\left(\frac{\partial q}{\partial \lambda}\Big|_{\lambda=\lambda_i-\delta}\right) < 0$, the minimum is at λ_i
- if $\left(\frac{\partial q}{\partial \lambda}\Big|_{\lambda=\lambda_i+\delta}\right)\left(\frac{\partial q}{\partial \lambda}\Big|_{\lambda=\lambda_{i+1}-\delta}\right) < 0$, the minimum is between λ_i and λ_{i+1}

After determining the interval of the minimum, we solve the following equation to get the λ^* at which q reaches its minimum:

$$\frac{\partial q}{\partial \lambda} = \sum_{k=i+1}^{N}\left(-\frac{1}{c_k\beta_k}\left(\frac{\lambda}{c_k\beta_k}\right)^{-\frac{1}{\alpha_k}} + \frac{d_k}{c_k\beta_k}\right) + B_c = 0 \tag{B.15}$$

using the bisection method with initial points (λ^*,λ_{i+1}). The solution for the primal problem is:

$$b_k^* = \begin{cases} 0 & \forall k \le i \\ \frac{\left(\frac{\lambda^*}{c_k\beta_k}\right)^{-\frac{1}{\alpha_k}} - d_k}{c_k\beta_k} & \forall i \le k \le N \end{cases} \tag{B.16}$$

After obtaining the optimal solution on the allocated bandwidth to each user, we need to convert it to the number of PRBs. This is done as follows: Each user k is first assigned $\left\lfloor \frac{b_k}{180KHz} \right\rfloor$ PRBs (in LTE, each PRB has 180KHz bandwidth). The rest of the PRBs are distributed to users in the decreasing order of their utility gain if they are assigned more PRBs.

B.3 Simulation Scenarios and Results

This section presents the simulation results to analyze the proposed algorithm. We evaluate the difference in the system performance over different system parameters, i.e. β and α. We compare the EMA rate approach with the instantaneous rate approach by tuning β. Then, we demonstrate how α affects the system performance, e.g. cell throughput

and fairness among users. Finally, we show a scenario with mixed traffic types that gives some indication on choosing the parameters when using the EMA rate approach. All following investigations are based on the default simulation settings given in Table B.1.

TABLE B.1: System settings

Parameter	**Settings**
Number of eNBs/cells	1 eNBs serving 1 cell with 375m radius circular cell
Channel Model	Path loss: 128.1 + 37.6log10(R), R in Km Slow fading: Correlated Log normal, zero mean, 8db std. and 50 m correlation distance Fast fading: Jake's like model
Utility Function Related (default)	$\bar{r}$ has a unit of 100 Kbps in formula (1)
TCP version	New Reno
Traffic Type	TCP based file downloading with full buffer mode
Simulation Time	1000s (5 runs with different seeds)
Special settings for scenarios	
Scenario 1 and 2	15 PRBs (3MHz) per cell 25 Users without mobility (Stationary), 5 user groups (each 5 users) with different distances to eNB Scenario 1: $\alpha = 0.5$, different β in different runs Scenario 2: $\beta = 0.1$, different α in different runs
Scenario 3	25 PRBs (5MHz) per cell 10 Users with RWP mobility (120 Km/h) 5 groups (each 2 users) $\beta = 0.1$ different α for different users

B.3.1 Influence of the Smoothing Factor β, $\alpha = 0.5$

The proposed utility function is based on the user EMA rate which is updated every TTI according to Equation B.2. The coefficient β,

also known as the weighting multiplier, a constant smoothing factor between 0 and 1, represents the degree of weighting decrease. If $\beta = 1$, the EMA rate becomes the instantaneous data rate which has been seen from the previous papers [KL05], [KL08], and [CWC+11]. With a decreasing value of β, a longer and stronger history effect is considered. In the finance, β is often chosen as $\beta = \frac{2}{M+1}$, with M representing the time periods [DPP13].

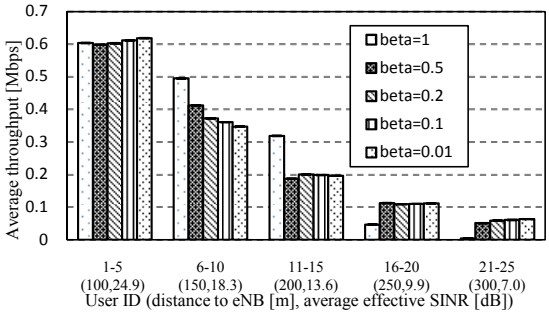


FIGURE B.3: Average user throughput over different β

In this scenario, we assume all the users have the same utility function representing the same traffic and user type, i.e. $\alpha = 0.5$ for all the users. This scenario is simulated with five different β (each β refers to one simulation run set with 5 seeds). Figure B.3 shows the average user throughput of the users with the same distance to the eNB. The confidence interval is also shown in the figure but it is too small to be visible. The x-axis shows the user id, user distance to eNB and the average effective SINR from the simulation. It is obvious that with increasing distance to eNB, the channel condition is getting worse, and the user average throughput decreases. When $\beta = 1$, corresponding to the instantaneous rate, the users 21-25 hardly get any throughput over a long time, which could cause a poor fairness. With a smaller β, which means the EMA rate is considered instead of the instantaneous data rate, the users 16-25 can get much higher throughputs by taking some

of the resources from users with good channel conditions, i.e. mainly from the users 6-10 in this specific artificial scenario. The throughput of users with bad channel conditions is improved by a smaller β at the cost of decreasing the total cell throughput as shown in Table B.2. The cell throughput decreases c.a. 7 8% when EMA is applied instead of instantaneous data rate. In reality, it is recommended that the operator should properly define the value of β in order to make a good trade-off between throughput and fairness, considering the number of PRBs and active number of users.

TABLE B.2: System behavior over β

β	1	0.5	0.2	0.1	0.01
Cell throughput [Mbps]	7.32	6.80	6.72	6.71	6.69

B.3.2 Influence of the Smoothing Factor α, $\beta = 1$

From the Equation B.1, it can be concluded that with the algorithm maximizes the cell throughput similar to the Max C/I scheduling algorithm [ZZL+12]. In Figure B.4, with , only users 1-10 with very good channels are scheduled, especially users 1-5 nearly get all the resources. Only 5.28 users in average are scheduled per TTI and a cell throughput around 10Mbps is achieved as shown in Table B.3. On the contrary, with a larger , the throughput of users 1-5 are reduced dramatically while the users 11-25 are scheduled, which leads to a better fairness regarding to user throughput. The cost is that more users are scheduled every TTI, which leads to a higher signalling overhead and the cell throughput is reduced (ca. 40% when increasing from 0.1 to 0.9).

The simulation results of the two scenarios above reveal that even though tuning α can improve the fairness, it is definitely not enough. Both parameters need to be tuned to guarantee certain fairness in the system. This confirms that our EMA rate approach can clearly improve

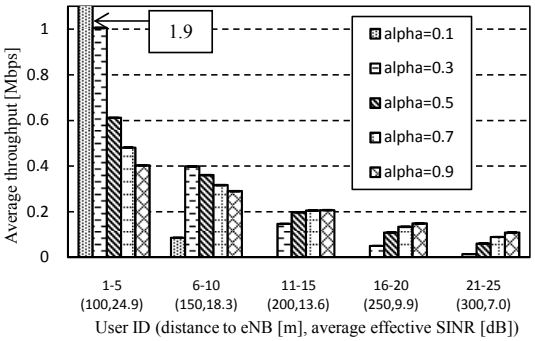


FIGURE B.4: Average user throughput over different α

TABLE B.3: System behavior over α

α	0.1	0.3	0.5	0.7	0.9
Cell throughput [Mbps]	9.96	8.09	6.71	6.15	5.81
Average number of scheduled users per TTI	5.28	11.29	13.58	13.63	13.64

the fairness compared to the instantaneous rate approach, especially when the number of users is larger than the number of PRBs.

B.3.3 Mixed Traffic Types

In this scenario, users with different α representing different traffic or user types coexist. Since user mobility is considered, all users are supposed to have the same average channel condition given a sufficient long simulation time. The parameter dominating the user throughput is α. The users with same should have the same performance which is proved by simulation results shown in Figure B.5.

The users 1 and 2 get the highest throughput while 9 and 10 lowest. Figure B.6 shows the achieved user marginal utility based on the long-term average throughput. It can be seen all users have a similar marginal utility (around 0.32). This is because users with a smaller α

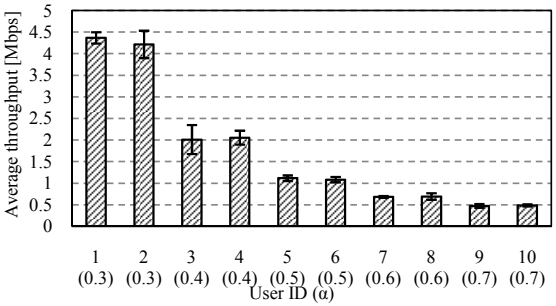


FIGURE B.5: Average user throughput for different user

demand a lower data rate to have the same marginal utility compared to ones with a larger α. This observation can be an indication for the operators to design proper utility functions for different traffic types and user categories.

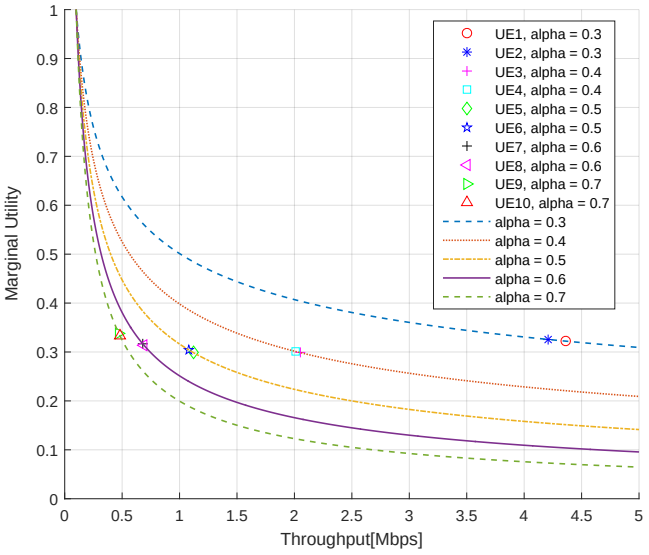


FIGURE B.6: User marginal utility based on average throughput

B.4 Conclusion

This work proposes a new approach to solve the utility-based resource allocation problem using the Exponential Moving Average rate. The problem is formulated as a convex optimization problem and analytically solved using Lagrangian decomposition method. The analytical solution is implemented in the LTE simulation model using OPNET. We have proved through the simulations that our approach overcomes the fairness drawback of the approach using instantaneous data rate. Our approach guarantees that all users including the ones with very bad channel conditions are still served. Furthermore, we also analyze the influence of the different parameters on the algorithm performance. By changing the system parameters α and β, the fairness can be tuned at the cost of cell throughput. This gives some indication on how to choose the right parameters for our proposed algorithm.

Appendix C

Curve Fitting Data

Table C.1: Curve fitting data for sigma values

TBS index	Sigma	R-square
0	26.75	0.9971
2	43.69	0.9989
4	70.78	0.999
6	103.8	0.9992
8	139.9	0.9995
10	175.7	0.9995
12	228.9	0.9995
14	286.8	0.9993
16	324.8	0.9994
18	396.6	0.9997
20	463.3	0.9996
22	538.6	0.9996
24	611.4	0.9997
26	738	0.9996

Table C.1 shows the curve fitting statistics of sigma values. Here the sigma represents the TBS size with respect to the MCS. In this study, MCS is an array of 14 values covering all TBS index range.

Here the R-square represents the reliability of the curve fitting by the cftool of MATLAB. By looking into the input and output, R-square returns a ratio value which tells how well the applied curve fit explains the variability in the input data. For R-square, ratio values which are close to 1 mean a good capture of the input data.

Appendix D

3GPP Transport Block Size

Table D.1 covers the full extend of mapping between the MCS and TBS according to the 3GPP standard [3GP10d]. The TBS value is telling how many bits can be send over the air interface in one TTI, so its size is the main factor for estimated throughput calculations.

TABLE D.1: Transport block size table according to 3GPP specifications

TBS index	**Number of PRB**									
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
0	16	32	56	88	120	152	176	208	224	256
1	24	56	88	144	176	208	224	256	328	344
2	32	72	144	176	208	256	296	328	376	424
3	40	104	176	208	256	328	392	440	504	568
4	56	120	208	256	328	408	488	552	632	696
5	72	144	224	328	424	504	600	680	776	872
6	328	176	256	392	504	600	712	808	936	1032
7	104	224	328	472	584	712	840	968	1096	1224
8	120	256	392	536	680	808	968	1096	1256	1384
9	136	296	456	616	776	936	1096	1256	1416	1544
10	144	328	504	680	872	1032	1224	1384	1544	1736
11	176	376	584	776	1000	1192	1384	1608	1800	2024
12	208	440	680	904	1128	1352	1608	1800	2024	2280
13	224	488	744	1000	1256	1544	1800	2024	2280	2536
14	256	552	840	1128	1416	1736	1992	2280	2600	2856
15	280	600	904	1224	1544	1800	2152	2472	2728	3112
16	328	632	968	1288	1608	1928	2280	2600	2984	3240
17	336	696	1064	1416	1800	2152	2536	2856	3240	3624
18	376	776	1160	1544	1992	2344	2792	3112	3624	4008
19	408	840	1288	1736	2152	2600	2984	3496	3880	4264
20	440	904	1384	1864	2344	2792	3240	3752	4136	4584
21	488	1000	1480	1992	2472	2984	3496	4008	4584	4968
22	520	1064	1608	2152	2664	3240	3752	4264	4776	5352
23	552	1128	1736	2280	2856	3496	4008	4584	5160	5736
24	584	1192	1800	2408	2984	3624	4264	4968	5544	5992
25	616	1256	1864	2536	3112	3752	4392	5160	5736	6200
26	712	1480	2216	2984	3752	4392	5160	5992	6712	7480

Bibliography

[3GP02] 3GPP. High Speed Downlink Packet Access (HSDPA); User Equipment (UE) radio transmission and reception (FDD). TR 25.890, 3rd Generation Partnership Project (3GPP), 03 2002.

[3GP06] 3GPP. Physical layer aspect for evolved Universal Terrestrial Radio Access (UTRA). TR 25.814, 3rd Generation Partnership Project (3GPP), 10 2006.

[3GP09a] 3GPP. Architecture aspects of Home Node B (HNB) / Home enhanced Node B (HeNB). TR 23.830, 3rd Generation Partnership Project (3GPP), 10 2009.

[3GP09b] 3GPP. Requirements for Evolved UTRA (E-UTRA) and Evolved UTRAN (E-UTRAN). TR 25.913, 3rd Generation Partnership Project (3GPP), 12 2009.

[3GP10a] 3GPP. Evolved Universal Terrestrial Radio Access (E-UTRA); Medium Access Control (MAC) protocol specification. TS 36.321, 3rd Generation Partnership Project (3GPP), 06 2010.

[3GP10b] 3GPP. Evolved Universal Terrestrial Radio Access (E-UTRA); Packet Data Convergence Protocol (PDCP) specification. TS 36.323, 3rd Generation Partnership Project (3GPP), 01 2010.

[3GP10c] 3GPP. Evolved Universal Terrestrial Radio Access (E-UTRA); Radio Link Control (RLC) protocol specification. TS 36.322, 3rd Generation Partnership Project (3GPP), 10 2010.

[3GP10d] 3GPP. Spreading and modulation (FDD). TS 25.213, 3rd Generation Partnership Project (3GPP), 10 2010.

[3GP11a] 3GPP. Evolved Universal Terrestrial Radio Access (E-UTRA); Radio Frequency (RF) system scenarios. TR 36.942, 3rd Generation Partnership Project (3GPP), 01 2011.

[3GP11b] 3GPP. Technical Specification Group Services and System Aspects; Network architecture (Release 10). TS 36.321, 3rd Generation Partnership Project (3GPP), 03 2011.

[3GP12] 3GPP. Feasibility study on user plane congestion management. TR 32.805, 3rd Generation Partnership Project (3GPP), 12 2012.

[3GP13] 3GPP. Scenarios and requirements for small cell enhancements for E-UTRA and E-UTRAN (Release 12). TR 36.932, 3rd Generation Partnership Project (3GPP), 3 2013.

[3GP15a] 3GPP. RAN 5G Workshop - The Start of Something. Technical report, 2015.

[3GP15b] 3GPP. Technical Specification Group Radio Access Network; Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Overall description; Stage 2 (Release 13). TS 36.300, 3rd Generation Partnership Project (3GPP), 12 2015.

[3GP15c] 3GPP. Technical Specification Group Services and System Aspects; Policy and charging control architecture (Release 13). TS 23.203, 3rd Generation Partnership Project (3GPP), 12 2015.

[AKR+01] M. Andrews, K. Kumaran, K. Ramanan, A. Stolyar, P. Whiting, and R. Vijayakumar. Providing quality of service over a shared wireless link. *Communications Magazine, IEEE*, 39(2):150–154, Feb 2001.

[And84] T. W. Anderson. *An Introduction to Multivariate Statistical Analysis.* Wiley, New York, NY, second edition, 1984.

[Ass15] GSA - The Global Mobile Suppliers Association. GSA confirms 360 LTE networks commercially launched by end 2014, strong growth in LTE-Advanced and VoLTE deployments. `http://www.gsacom.com/news/gsa_418.php`, 2015.

[BAS+05] K. Brueninghaus, D. Astely, T. Salzer, S. Visuri, A. Alexiou, S. Karger, and G.-A. Seraji. Link performance models for system level simulations of broadband radio access systems. In *Personal, Indoor and Mobile Radio Communications, 2005. PIMRC 2005. IEEE 16th International Symposium on*, volume 4, pages 2306–2311 Vol. 4, Sept 2005.

[BNO03] D. P. Bertsekas, A. Nedić, and A. E. Ozdaglar. *Convex Analysis and Optimization.* Athena Scientific, 2003.

[CBD02] Tracy Camp, Jeff Boleng, and Vanessa Davies. A survey of mobility models for ad hoc network research. *Wireless communciation and mobile computing (WCMC)*, 2:483–502, 2002.

[CG03] Xiaodong Cai and G.B. Giannakis. A two-dimensional channel simulation model for shadowing processes. *Vehicular Technology, IEEE Transactions on*, 52(6):1558–1567, Nov 2003.

[Cis14] Cisco. Cisco Visual Networking Index Predicts IP Traffic to Triple from 2014-2019; Growth Drivers Include Increasing Mobile Access, Demand for Video Services. Technical report, Cisco, 2014.

[Cis15] Cisco. Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2014–2019. Technical report, Cisco, 2015.

[CJS+07] Sunggu Choi, Kyungkoo Jun, Yeonseung Shin, Seokhoon Kang, and Byoungjo Choi. Mac scheduling scheme for voip traffic service in 3g lte. In *Vehicular Technology Conference, 2007. VTC-2007 Fall. 2007 IEEE 66th*, pages 1441–1445, Sept 2007.

[Cla05] H. Claussen. Efficient modelling of channel maps with correlated shadow fading in mobile radio systems. In *Personal, Indoor and Mobile Radio Communications, 2005. PIMRC 2005. IEEE 16th International Symposium on*, volume 1, pages 512–516, Sept 2005.

[CPG+13] 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. *Communications Surveys Tutorials, IEEE*, 15(2):678–700, Second 2013.

[CWC+11] Li Chen, Bin Wang, Li Chen, Xin Zhang, and Dacheng Yang. Utility-based resource allocation for mixed traffic in wireless networks. In *Computer Communications Workshops (INFOCOM WKSHPS), 2011 IEEE Conference on*, pages 91–96, April 2011.

[Dah07] E. Dahlman. *3G Evolution: HSPA and LTE for Mobile Broadband*. Electronics & Electrical. Elsevier Academic Press, 2007.

[DPP13] Jitendra Dangra Dileshwar Prasad Patel, Amit Vajpayee. Short Term Load Forecasting by Using Time Series Analysis through Smoothing Techniques. In *International Journal of Engineering Research and Technology (IJERT)*, Septermber 2013.

[DS07] Renaud Cuny David Soldani, Man Li. *QoS and QoE Management in UMTS Cellular Systems*. John Wiley & Sons, Ltd, 2007.

[Eri15] Ericsson. Ericsson Mobility Report, on the pulse of the networked society. Technical report, 2015.

[FHTG10] M. Fiedler, T. Hossfeld, and P. Tran-Gia. A generic quantitative relationship between quality of experience and quality of service. *Network, IEEE*, 24(2):36–41, March 2010.

[FLKV08] Yong Fan, P. Lunden, M. Kuusela, and M. Valkama. Efficient semi-persistent scheduling for voip on eutra downlink. In *Vehicular Technology Conference, 2008. VTC 2008-Fall. IEEE 68th*, pages 1–5, Sept 2008.

[FPR00] Tiziana Ferrari, Giovanni Pau, and Carla Raffaelli. Priority queueing applied to expedited forwarding: A measurement-based analysis. In Jon Crowcroft, James Roberts, and MikhailI. Smirnov, editors, *Quality of Future Internet Services*, volume 1922 of *Lecture Notes in Computer Science*, pages 167–181. Springer Berlin Heidelberg, 2000.

[Ger10] Germany Embraces 4G. `http://spectrum.ieee.org/tech-talk/telecom/wireless/germany-embraces-4g`, 2010. last accessed in February, 2016.

[GSM15] GSMHistory.com. What is 5g, Gg visions. Technical report, 2015.

[HS12] Jiarun Song Honglei Su, Fuzheng Yang. Packet-layer quality assessment for networked video. *International Journal of Computers Communications and Control (IJCCC)*, 7(3):565–573, September 2012.

[Hua13] Huawei. The second phase of LTE-Advanced (LTE-B 30-fold capacity Boosting to LTE). Technical report, Huawei, 2013.

[Inc14] Fujitsu Network Communications Inc. The Benefits of Cloud-RAN Architecture in Mobile Network Expansion. Technical report, 2014.

[Ins11] China Mobile Research Institute. C-RAN: The Road Towards Green RAN. Technical report, 2011.

[Int16] Intel Corporation. `http://www.intel.de/content/www/de/de/homepage.html`, 2016. last accessed in March, 2016.

[IR09] ITU-R. Guidelines for evaluation of radio interface technologies for IMT-Advanced. M Series M.2135.1, Recommendation Sector of International Telecommunication Union, 12 2009.

[IT03] ITU-T. One-way transmission time. Recommendation G.114, International Telecommunication Union, 5 2003.

[IT07] ITU-T. Definition of Quality of Experience (QoE). Technical report, Study Group 12, 1 2007.

[IT12a] ITU-T. Methods for subjective determination of transmission quality. Recommendation P.800, International Telecommunication Union, 7 2012.

[IT12b] ITU-T. Opinion model for video-telephony applications. Recommendation G.1070, International Telecommunication Union, 7 2012.

[IT14] ITU-T. Estimating end-to-end performance in IP networks for data applications. Recommendation G.1030, International Telecommunication Union, 2 2014.

[IT15] ITU-T. The E-model: a computational model for use in transmission planning. Recommendation G.107, International Telecommunication Union, 6 2015.

[KL05] Wen-Hsing Kuo and Wanjiun Liao. Utility-based optimal resource allocation in wireless networks. In *Global Telecommunications Conference, 2005. GLOBECOM '05. IEEE*, volume 6, pages 5 pp.–3512, Dec 2005.

[KL08] Wen-Hsing Kuo and Wanjiun Liao. Utility-based radio resource allocation for qos traffic in wireless networks. *Wireless Communications, IEEE Transactions on*, 7(7):2714–2722, July 2008.

[Kle75] Leonard Kleinrock. *Queueing Systems*, volume I: Theory. Wiley Interscience, 1975. (Published in Russian, 1979. Published in Japanese, 1979. Published in Hungarian, 1979. Published in Italian 1992.).

[KLZ09] R. Kwan, C. Leung, and Jie Zhang. Proportional fair multiuser scheduling in lte. *Signal Processing Letters, IEEE*, 16(6):461–464, June 2009.

[KPK+08] P. Kela, J. Puttonen, N. Kolehmainen, T. Ristaniemi, T. Henttonen, and Martti Moisio. Dynamic packet scheduling performance in UTRA Long Term Evolution downlink. In *Wireless Pervasive Computing, 2008. ISWPC 2008. 3rd International Symposium on*, pages 308–313, May 2008.

[KSM+13] A. Könsgen, A. Singh, A. Mahmoud, X. Li, C. Görg, M. Kus, M. Kayralci, and J. Grigutsch. Enhancing Quality of Experience (QoE) Assessment Models for Web Traffic. In *5th International Conference on Mobile Networks and Management (MONAMI)*, Cork, Ireland, September/September 2013.

[KSW+08] A. Köpke, M. Swigulski, K. Wessel, D. Willkomm, P. T. Klein Haneveld, T. E. V. Parker, O. W. Visser, H. S. Lichte, and S. Valentin. Simulating wireless and mobile networks in omnet++ the mixim vision. In *Proceedings of the 1st International Conference on Simulation Tools and Techniques for Communications, Networks and Systems & Workshops*, Simutools '08, pages 71:1–71:8, ICST, Brussels, Belgium, Belgium, 2008. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering).

[LAS+13] Xi Li, S. Aggarwal, A. Singh, A. Konsgen, C. Gorg, and M. Kus. Enhancing quality of experience (qoe) assessment models for video applications. In *Wireless and Mobile Networking Conference (WMNC), 2013 6th Joint IFIP*, pages 1–4, April 2013.

[Li09] Xi Li. *Radio Access Network Dimensioning for 3G UMTS*. PhD thesis, Bremen, 2009.

[LL03] D. Liu and Y.-H. Lee. An efficient scheduling discipline for packet switching networks using earliest deadline first round robin*. In *Computer Communications and Networks, 2003. ICCCN 2003. Proceedings. The 12th International Conference on*, pages 5–10, Oct 2003.

[LLT+12] X. Li, M. Li, U. Toseef, A. Timm-Giel, C. Goerg, D. Dulas, M. Nowacki, and R. Ruchala. Dimensioning of the shared transport network for collocated multiradio: LTE and HSDPA. In *Wireless and Mobile Computing, Networking and Communications (WiMob), 2012 IEEE 8th International Conference on*, pages 308–315, 2012.

[LTLTG15] Ming Li, Phuong Nga Tran, Xi Li, and Andreas Timm-Giel. Qoe-driven joint radio and transport optimized eps bearer rates of multi-services in lte. San Diego, USA, December 2015.

[LTTG15] Ming Li, Phuong Nga Tran, and Andreas Timm-Giel. Radio resource allocation in lte femtocell considering qoe. In *Leistungs-, Zuverlässigkeits- und Verlässlichkeitsbewertung von Kommunikationsnetzen und Verteilten Systemen 8. GI/ITG-Workshop MMBNet 2015*, number 302 in Berichte des Fachbereichs Informatik, pages 55–62. Universität Hamburg, Department Informatik, September 2015.

[LTTTG15a] Ming Li, Phuong Nga Tran, Hüseyin Tütüncüoglu, and Andreas Timm-Giel. Coordinated radio resource allocation in lte femtocell cluster considering transport limitations. *Communications (ICC), 2015 IEEE International Conference on*, pages 4716–4721, June 2015.

[LTTTG15b] Ming Li, Phuong Nga Tran, Hüseyin Tütüncüoglu, and Andreas Timm-Giel. Qoe-based radio resource allocation in lte femtocell considering transport limitations. *20th IEEE Symposium on Computers and Communication (ISCC)*, pages 656–661, May 2015.

[LTWTG14] Ming Li, Phuong Nga Tran, Dimin Wang, and Andreas Timm-Giel. Radio resource allocation in lte using utility functions based on moving average rates. *IEEE WCNC'14 Track 2 (MAC and Cross-Layer Design) (IEEE WCNC'14 Track 2 : MAC*, April 2014.

[LXZ+12] Fei Liu, Wei Xiang, Yueying Zhang, Kan Zheng, and Hui Zhao. A novel QoE-based carrier scheduling scheme in LTE-Advanced networks with multi-service. In *Vehicular Technology Conference (VTC Fall), 2012 IEEE*, pages 1–5, Sept 2012.

[LZL+12] Ming Li, Liang Zhao, Xi Li, Xiaona Li, Yasir Naseer Zaki, Andreas Timm-Giel, and Carmelita Goerg. Investigation of network virtualization and load balancing techniques in lte networks. *Vehicular Technology Conference (VTC Spring), 2012 IEEE 75th*, pages 1–5, May 2012.

[Mic13] Microsoft. Network bandwidth requirements for media traffic in Lync Server 2013. Technical report, 2013.

[Mic15] Microsoft. H264 Smooth Streaming 720p for 3G or 4G. Technical report, 2015.

[Mob15] Compare Mobile Broadband Deals. `http://www.broadbandchoice.co.uk/compare/mobile-broadband/`, 2015. last accessed in December, 2015.

[MPKM08] G. Monghal, K.I. Pedersen, I.Z. Kovacs, and P.E. Mogensen. QoS oriented time and frequency domain packet schedulers for the UTRAN Long Term Evolution. In *Vehicular Technology Conference, 2008. VTC Spring 2008. IEEE*, pages 2532–2536, May 2008.

[Net16] NetSim Simulator. `http://www.tetcos.com/netsim_gen.html`, 2016. last accessed in February, 2016.

[NM44] John Von Neumann and Oskar Morgenstern. *Theory of Games and Economic Behavior*. Princeton University Press, 1944.

[Nok11] Nokia. Improving 4G coverage - and capacity indoors and at hotspots with LTE femtocells. Technical report, Nokia, 2011.

[Nok14] Nokia. What is going on in Mobile Broadband Networks? Smartphone Traffic Analysis and Solutions. Technical report, 2014.

[NS16] NS-The Network Simulator. `https://www.nsnam.org`, 2016. last accessed in February, 2016.

[OMN16] OMNeT++ Discrete Event Simulator. `https://omnetpp.org`, 2016. last accessed in February, 2016.

[OPE15] openWNS Network Simulator. `http://docs.openwns.org/index.html`, 2015. last accessed in Septermber, 2015.

[PE09] Daniel P. Palomar and Yonina C. Eldar, editors. *Convex Optimization in Signal Processing and Communications*. Cambridge University Press, 2009. Cambridge Books Online.

[PE10] Daniel P. Palomar and Yonina C. Eldar, editors. *Convex optimization in signal processing and communications*. Cambridge University Press, Cambridge, UK, New York, 2010.

[PKF+09] K.I. Pedersen, T.E. Kolding, F. Frederiksen, I.Z. Kovacs, D. Laselva, and P.E. Mogensen. An overview of downlink radio resource management for utran long-term evolution. *Communications Magazine, IEEE*, 47(7):86–93, July 2009.

[PPM+07] A. Pokhariyal, K.I. Pedersen, G. Monghal, I.Z. Kovacs, C. Rosa, T.E. Kolding, and P.E. Mogensen. HARQ Aware Frequency Domain Packet Scheduler with Different Degrees of Fairness for the UTRAN Long Term Evolution. In *Vehicular Technology Conference, 2007. VTC2007-Spring. IEEE 65th*, pages 2761–2765, April 2007.

[Riv16] Riverbed Modeler. `http://de.riverbed.com`, 2016. last accessed in February, 2016.

[Rza05] J. Rzasa. Communication networking: An analytical approach [book review]. *Communications Magazine, IEEE*, 43(11):18–19, Nov 2005.

[SFC10] Junaid Shaikh, Markus Fiedler, and Denis Collange. Quality of experience from user and network perspectives. *Annales des Télécommunications*, 65(1-2):47–57, 2010.

[Shn84] Ben Shneiderman. Response time and display rate in human performance with computers. *ACM Comput. Surv.*, 16(3):265–285, September 1984.

[SL05] G. Song and Ye Li. Utility-based resource allocation and scheduling in ofdm-based wireless broadband networks. *Communications Magazine, IEEE*, 43(12):127–134, Dec 2005.

[SLC09] G. Song, Ye Li, and L.J. Cimini. Joint channel- and queue-aware scheduling for multiuser diversity in wireless ofdma networks. *Communications, IEEE Transactions on*, 57(7):2109–2121, July 2009.

[SS15] A. Shafiei and S.M. Saberali. A simple asymptotic bound on the error of the ordinary normal approximation to the student's t-distribution. *Communications Letters, IEEE*, 19(8):1295–1298, Aug 2015.

[Tan02] Andrew Tanenbaum. *Computer Networks*. Prentice Hall Professional Technical Reference, 4th edition, 2002.

[Tan07] Andrew S. Tanenbaum. *Modern Operating Systems*. Prentice Hall Press, Upper Saddle River, NJ, USA, 3rd edition, 2007.

[TBK+15a] M. Taranetz, T. Blazek, T. Kropfreiter, M.K. Muller, S. Schwarz, and M. Rupp. Runtime precoding: Enabling multipoint transmission in lte-advanced system-level simulations. *Access, IEEE*, 3:725–736, June 2015.

[TBK+15b] M. Taranetz, T. Blazek, T. Kropfreiter, M.K. Muller, S. Schwarz, and M. Rupp. Runtime precoding: Enabling multipoint transmission in lte-advanced system-level simulations. *IEEE Access*, 3:725–736, 2015.

[TCJ+11] S. Thakolsri, S. Cokbulan, D. Jurca, Z. Despotovic, and W. Kellerer. Qoe-driven cross-layer optimization in wireless networks addressing system efficiency and utility fairness. In *GLOBECOM Workshops (GC Wkshps), 2011 IEEE*, pages 12–17, Dec 2011.

[Tec15] Akamai Technologies. Akamai's [state of the internet], Q3 2015 report. Technical report, 2015.

[TNC+01] Fouad Tobagi, Waël Noureddine, Benjamin Chen, Athina Markopoulou, Chuck Fraleigh, Mansour Karam, Jose-Miguel Pulido, and Jun-ichi Kimura. Service differentiation in the internet to support multimedia traffic. In Sergio Palazzo, editor, *Evolutionary Trends of the Internet*, volume 2170 of *Lecture Notes in Computer Science*, pages 381–400. Springer Berlin Heidelberg, 2001.

[Tos13] Umar Toseef. *LTE Optimization and Resource Management in Wireless Heterogeneous Networks*. PhD thesis, Bremen, 2013.

[WLB02] Zhou Wang, Ligang Lu, and A.C. Bovik. Video quality assessment using structural distortion measurement. In *Image Processing. 2002. Proceedings. 2002 International Conference on*, volume 3, pages III–65–III–68 vol.3, 2002.

[WZD+11] Beibei Wang, Dekun Zou, Ran Ding, Tao Liu, S. Bhagavathy, N. Narvekar, and J. Bloom. Efficient frame complexity estimation and application to G.1070 vide quality monitoring. In *Quality of Multimedia Experience (QoMEX), 2011 Third International Workshop on*, pages 96–101, Sept 2011.

[YH08] K. Yamagishi and T. Hayashi. Parametric packet-layer model for monitoring video quality of IPTV services.

In *Communications, 2008. ICC '08. IEEE International Conference on*, pages 110–114, May 2008.

[Zak12] Y. Zaki. *Future Mobile Communications: LTE Optimization and Mobile Network Virtualization*. Advanced Studies Mobile Research Center Bremen. Springer Fachmedien Wiesbaden, 2012.

[ZHY11] Zhong Zheng, J. Hamalainen, and Ying Yang. Practical resource scheduling and power control optimization for lte femtocell networks. In *Multi-Carrier Systems Solutions (MC-SS), 2011 8th International Workshop on*, pages 1–5, May 2011.

[ZLZ+11] Liang Zhao, Ming Li, Yasir Naseer Zaki, Andreas Timm-Giel, and Carmelita Goerg. Lte virtualization: From theoretical gain to practical solution. *Teletraffic Congress (ITC), 2011 23rd International*, pages 71 –78, September 2011.

[ZZL+12] 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 Yevgeni Koucheryavy, Lefteris Mamatas, Ibrahim Matta, and Vassilis Tsaoussidis, editors, *Wired/Wireless Internet Communication*, volume 7277 of *Lecture Notes in Computer Science*, pages 323–331. Springer Berlin Heidelberg, 2012.

[ZZWS13] Qian Zhang, Xinning Zhu, Leijia Wu, and K. Sandrasegaran. A coloring-based resource allocation for ofdma femtocell networks. In *Wireless Communications and Networking Conference (WCNC), 2013 IEEE*, pages 673–678, April 2013.

Curriculum Vitae

Family Name	Li
Given Name	Ming
Birthday	20. Oct. 1984
Place of birth	Shandong, China

09.2000—06.2003	Laiwu No.1 Middle School, Laiwu, China
09.2003—02.2007	Shandong University, Jinan, China Bachelor of Electronic Information Science and Technology
03.2007—09.2009	Ulm University, Ulm Master of Communication Engineering
09.2007—09.2007	Part-time job at P3 Solutions, Ulm
10.2007—12.2007	Internship at Nokia Siemens Networks, Ulm
01.2008—09.2008	Master Thesis at Nokia Siemens Networks, Ulm
03.2010—03.2016	Research assistant and PhD candidate Working on projects funded by Nokia Siemens Networks Institute of Communication Networks Hamburg University of Technology, Hamburg, Germany
04.2016—now	Senior engineer at Intel, Munich