Aston University
Birmingham

Quality-Driven Resource Utilization Methods for Video Streaming in Wireless Communication Networks

Mirghiasaldin Seyedebrahimi

Doctor of Philosophy

ASTON UNIVERSITY

March 2015

Aston University

Quality-Driven Resource Utilization Methods for Video Streaming in Wireless Communication Networks

Mirghiasaldin Seyedebrahimi

Doctor of Philosophy
2015

Summary

This research is focused on the optimisation of resource utilisation in wireless mobile networks with the consideration of the users' experienced quality of video streaming services. The study specifically considers the new generation of mobile communication networks, i.e. 4G-LTE, as the main research context. The background study provides an overview of the main properties of the relevant technologies investigated. These include video streaming protocols and networks, video service quality assessment methods, the infrastructure and related functionalities of LTE, and resource allocation algorithms in mobile communication systems.

A mathematical model based on an objective and no-reference quality assessment metric for video streaming, namely Pause Intensity, is developed in this work for the evaluation of the continuity of streaming services. The analytical model is verified by extensive simulation and subjective testing on the joint impairment effects of the pause duration and pause frequency. Various types of the video contents and different levels of the impairments have been used in the process of validation tests. It has been shown that Pause Intensity is closely correlated with the subjective quality measurement in terms of the Mean Opinion Score and this correlation property is content independent.

Based on the Pause Intensity metric, an optimised resource allocation approach is proposed for the given user requirements, communication system specifications and network performances. This approach concerns both system efficiency and fairness when establishing appropriate resource allocation algorithms, together with the consideration of the correlation between the required and allocated data rates per user. Pause Intensity plays a key role here, representing the required level of Quality of Experience (QoE) to ensure the best balance between system efficiency and fairness. The 3GPP Long Term Evolution (LTE) system is used as the main application environment where the proposed research framework is examined and the results are compared with existing scheduling methods on the achievable fairness, efficiency and correlation.

Adaptive video streaming technologies are also investigated and combined with our initiatives on determining the distribution of QoE performance across the network. The resulting scheduling process is controlled through the prioritization of users by considering their perceived quality for the services received. Meanwhile, a trade-off between fairness and efficiency is maintained through an online adjustment of the scheduler's parameters. Furthermore, Pause Intensity is applied to act as a regulator to realise the rate adaptation function during the end user's playback of the adaptive streaming service. The adaptive rates under various channel conditions and the shape of the QoE distribution amongst the users for different scheduling policies have been demonstrated in the context of LTE.

Finally, the work for interworking between mobile communication system at the macro-cell level and the different deployments of WiFi technologies throughout the macro-cell is presented. A QoE-driven approach is proposed to analyse the offloading mechanism of the user's data (e.g. video traffic) while the new rate distribution algorithm reshapes the network capacity across the macro-cell. The scheduling policy derived is used to regulate the performance of the resource allocation across the fair-efficient spectrum. The associated offloading mechanism can properly control the number of the users within the coverages of the macro-cell base station and each of the WiFi access points involved. The performance of the non-seamless and user-controlled mobile traffic offloading (through the mobile WiFi devices) has been evaluated and compared with that of the standard operator-controlled WiFi hotspots.

Keywords: Quality of Experience; Pause Intensity; Scheduling; Adaptive streaming; LTE;

This thesis is dedicated to:

Holly and Hannah who are the meaning of my life

and

to the memory of my father and my mother who are always alive in my heart

Acknowledgement

I would like to thank Dr Xiao-Hong Peng for his guidance, advice and support throughout the duration of my research. His constant enthusiasm and ability to generate ideas and directions of work hugely influenced the content of my research and without his input I would never have reached this point. I would also like to thank Dr Jort Van Mourik for his kind support and valuable guidance as my co-supervisor.

I must also thank Blackberry for their contribution in the sponsorship of the project and specially Dr Rob Harrison for his constant support and guidance and Dr Chris Jones for his valuable support.

My colleagues in the Adaptive Communication Research Group, in particular, Dr Colin Bailey have been a vital part of this journey. I have no doubt that our team work have greatly assisted my research work. Special thanks to Mr. Philip Travis for his kindness and help in providing and assisting with laboratory equipment and project requirements.

To My wife Holly, for her constant support and encouragement without which I doubt that I would have started this journey and reached this point.

Table of Contents

List of Abbreviations

3G	Third Generation of Mobile Telecommunications Technology
4G	Fourth Generation of Mobile Telecommunications Technology
3GP	3GPP File Format
3GPP	3rd Generation Partnership Project
ACK	Acknowledgement (positive)
ACR	Absolute Category Rating
AHS	Adaptive HTTP Streaming
AMC	Adaptive Modulation and Coding
ARQ	Automatic Repeat Request
CC	Chase Combining
CDF	Cumulative Distribution Function
CDN	Content Delivery Network
CQI	Channel Quality Indicator
CRC	Cyclic Redundancy Check
DASH	Dynamic Adaptive Streaming Over HTTP
DCR	Degradation Category Rating
DRM	Digital Restrictions Management
DVQ	Digital Video Quality
EPC	Evolved Packet Core
EUTRAN	Evolved Universal Terrestrial Radio Access Network
FDD	Frequency-Division Duplexing
FEC	Forward Error Correction
FIFO	First In First Out
FR	Full Reference
HARQ	Hybrid-Automatic Repeat Request
HD	High Definition
HDS	HTTP Dynamic Streaming
HLS	HTTP Live Streaming
HSPA	High Speed Packet Access
HTML	HyperText Markup Language
HTTP	Hyper Text Transfer Protocol
IETF	Internet Engineering Task Force
IIS	Internet Information Services
IP	Internet Protocol
IR	Incremental Redundancy
ISO	International Organization for Standardization
ITU	International Telecommunications Union
ITU-R	ITU Radio Communication Sector
ITU-T	ITU Telecommunication Standardization Sector
JND	Just Noticeable Difference
JVT	Joint Video Team
LTE	Long-Term Evolution
MAC	Medium Access Control

MCS	Modulation and Coding Scheme
MiFi	Mobile WiFi
MMT	MPEG Media Transport
MOS	Mean Opinion Score
MPD	Media Presentation Description
MPEG	Motion Pictures Expert Group
MSE	Mean Square Error
NACK	Negative Acknowledgement
NR	No Reference
NS-2	Network Simulator
OFDMA	Orthogonal Frequency-Division Multiple Access
OTT	Over-The-Top contents
PDF	Probability Density Function
PedA	ITU Pedestrian Wireless Channel Model
PHICH	Physical Hybrid-ARQ Indicator Channel
PI	Pause Intensity
PSNR	Peak Signal-Noise Ratio
PSS	Packet Switched Streaming Service
PVQM	Perceptual Video Quality Measure
QAM	Quadrature Amplitude Modulation
QoE	Quality Of Experience
QoS	Quality Of Service
QPSK	Quadrature Phase Shift Keying
RB	Resource Block
RE	Resource Element
RFC	Request For Comments
RLC	Radio Link Control
RR	Reduced Reference
RSN	Retransmission Sequence Number
RTMP	Real Time Messaging Protocol
RTP	Real Time Protocol
RTSP	Real Time Streaming Protocol
RTT	Round Trip Time
RV	Redundancy Version
SC-FDMA	Single Carrier-Frequency Division Multiple Access
SINR	Signal to Interference plus Noise Ratio
SNR	Signal to Noise Ratio
SON	Self-Organizing Network
SSIM	Structural Similarity (Index)
SVOD	Scalable Video Coding
TCP	Transfer Control Protocol
TDD	Time-Division Duplexing
TTI	Transmission Time Interval
UDP	User Datagram Protocol
UE	User Equipment
URL	Uniform Resource Location
UTRAN	Universal Terrestrial Radio Access Network
VehA	ITU Vehicular Wireless Channel Model

VoD	Video on Demand
VQEG	Video Quality Experts Group
VQM	Video Quality Metric
WAG	Wireless Access Gateway
WiFi	Wireless Fidelity
WLAN	Wireless Local Area Network
XML	Extensible Mark-up Language

Figures and Tables

Chapter 1
Introduction

The improved capacity of mobile communication systems, the growing number of mobile users, and the dominance of video related applications are among the major changes that the contemporary communication technologies have to face. These coincide with the dramatically improved capabilities of end users' devices in the form of Smartphones and Tablets. Ironically, these simultaneous changes at both the network and user sides, in terms of the increased capacity versus the growing demand, will continue to challenge network operators and service providers in maintaining a high level of users' satisfaction. This also pushes the network to approach its capacity limit. Consequently, in spite of the introduced enabling technologies to achieve a nearly full capacity of the wireless channel in both 'time' and 'frequency' domains (such as OFDMA), finding the optimum resource utilization solution still has a high priority in the research and development agenda.

The exact conditions of the problem and the applicability of a solution proposed for the resource utilization problem depend on the chosen context for investigation. Traditionally, the optimization problems for resource utilization in the context of wireless and mobile communications are aimed to achieve a certain level of efficiency and/or fairness under the constraints in frequency/time domains and in the system architecture. Quantitative evaluation parameters such as throughput, delay and loss rate are the most common objectives of this type of optimisation problem. However, as it will be shown in this work, the satisfaction of the user as a subjective and qualitative parameter may also be included as a target of the optimisation problem through a correlated quantitative parameter. In fact, the study of the resource utilization problem from a Quality of Experience's (QoE) point of view is in line with the main trend of the strategies for enhancing performance of the contemporary communication technologies. Meanwhile, the focus on a specific service such as 'video streaming' further elaborates the constraints of the proposed solution. This will be reflected in the formulation of our proposed optimisation problems.

In the context of the new generations of mobile communication systems (e.g. 4G-LTE), the following aspects are chosen to be examined in our work, which lead to optimised QoE-driven solutions for the resource utilization problem:

- the scheduler and resource allocation functions in the mobile base station,
- an online algorithm for maintaining a desired balance between efficiency and fairness in the scheduler
- the extension of the resource allocation function to a new performance evaluation that characterizes the correlation between the required and allocated data rates
- the rate adaptation algorithm and network resource management in an adaptive video streaming paradigm, and
- the distribution of resources across the cell under the coverage of a base station and through interworking with smaller cells such as WiFi hotspots

1.1 Novelty

There are mainly three areas in which the novelty of this work can be highlighted:

a) The main concept of the Pause Intensity as a quality assessment metric for video streaming has been introduced previously in [1]. However, the proposed analytical approach and the mathematical model achieved in my work provide a new closed form formula for Pause Intensity, which establishes the relationship between this metric and the network performance (throughput) and service level (video bitrate). This relationship reveals a comprehensive perspective of the statistical properties of the network and makes the quality assessment easily accessible by the network as well as the end user. This model has been proved to be an efficient and appropriate tool for assessing the perceived quality of video streaming services by users.

b) A new approach driven by the QoE requirement is proposed to optimize the resource allocation that can achieve desired trade-offs between efficiency and fairness. In addition, a measurement evaluating the correlation between the required and allocated data rates per user is also introduced and applied in the optimization process. Corresponding algorithms for the implementation of the proposed approach have been derived and examined thoroughly in the context of LTE.

c) The proposed metric and related optimised resource allocation solutions have been applied in the resource management functions, including the rate adaptation streaming service at the user side, and capacity distribution and interworking at the network side. This creates a wide range of applications for the assessment of the network and user's capabilities using our proposed framework.

1.2 Objectives

This work aims to provide a framework which can be used to optimise the resource utilisation in the new generations of wireless mobile systems and networks for video streaming services. Pause Intensity plays the key role of assessing QoE in both the user and network sides of the system. This work is focused on the quality issue that concerns the continuity of service delivery while the fidelity of the image is also involved indirectly in the assessment. The proposed analytical model is intended to be readily evaluated at the receiver and before the decoding process, which means that the framework can provide a fast packet based evaluation mechanism for video streams without decoding.

This work also targets a range of applications for the proposed framework to show the capability of a no-reference QoE metric and related optimisation problems to be used in various aspects of the network. These include scheduling, resource allocation and link adaptation in the last-mile network interface, rate adaptation at the user side and interworking between different types of wireless networks (e.g. for traffic offloading).

1.3 Structure

The thesis comprises the background studies and the contributions of the author in eight chapters. The background studies in Chapters 2 and 3 consist of the main attributes and related functionality of 3GPP Long Term Evolution technology (3GPP-LTE) as the main context of the research carried out. Furthermore, the related aspects of the video streaming techniques which have been used to examine the applicability of the proposed framework in relation with 3GPP-LTE are also explained in the background study. Background covers the general explanations of the above mentioned aspects of the work, the network infrastructure, employed protocols and related properties. The specifications and the capabilities of the developed MATLAB-based simulator for this work are also explained as a part of the background. This simulator provides the 4G/LTE-based context of the investigation for the rest of the chapters.

In Chapter 4 a mathematical model for Pause Intensity, as an objective, no-reference and packet based video streaming quality assessment metric is developed and supported by extensive simulation and subjective testing. Pause Intensity provides a base for the evaluation of the continuity of a video streaming service given the joint effects of the pause length and pause frequency (or underrun frequency). Subjective testing is carried out using various types of video clips and a wide range of Pause Intensity values contributed by various pause length and pause frequency. The investigation of the correlation between the perceived video quality and the pause length, underrun frequency and Pause Intensity reveals the advantage of Pause Intensity regardless of the video content (e.g. sport, news ...).

The performance of the standard resource allocation policies in the wireless communication networks vary based on their efficiency and fairness achievement. An optimised resource allocation approach is proposed in Chapter 5, which concerns not only the efficiency and the fairness of the allocation but also the quality of the delivered video and the correlation between the required and allocated data rates per user. Pause Intensity is adopted to play the key role in evaluating QoE in the scheduler. The framework is examined in the context of 3GPP Long Term Evolution (LTE) systems. The requirements and structure of the proposed PI-based framework are discussed, and the results are compared with existing scheduling methods on the performances of fairness, efficiency and correlation). It is also shown that the proposed framework can produce a balanced trade-off between the three parameters through the QoE-aware resource allocation process.

The online resource allocation algorithms and the distribution of the end users' QoE in an adaptive video streaming service are addressed in Chapter 6. Pause Intensity has been used to control the priority of the users during the "last mile" scheduling process and an online adjustment has been introduced to adaptively set the scheduler's parameter and provide and maintain a desired trade-off between fairness and efficiency. Furthermore, the same PI metric is examined as a regulator for rate adaptation during the end user's playback of an adaptive video stream service. The adaptive rates under various channel status and the shape of the QoE distribution amongst the users for different scheduling policies have been discussed. The feasibility of the framework implementation is examined within 3GPP-LTE and the results are compared with the most commonly existing scheduling methods.

In Chapter 7, the interworking between the mobile communication system at the level of macro-cell and short-range wireless LANs across the macro-cell is addressed. A PI-based and QoE driven method for offloading a part of the video traffic through WiFi access points and amending the distribution of the system capacity across the macro-cell is proposed. Pause Intensity leverages the scheduling policy to shape the performance of the resource allocation in a highly fair to highly efficient spectrum. The offloading mechanism influences the number of users under the direct coverage of the main macro-cell base station and those connected through the WiFi access entities.

Chapter 8 concludes the thesis and contains a summary of the contributions this work has made, followed by an outline of the opportunities for future work.

Chapter 2
Video Delivery Technologies and Quality Assessment Methods

2.1 Introduction

In line with the growing demands for video based services over the internet, the methods for storing and delivering these contents have been improved dramatically. Currently dominant video coding techniques, such as H.264/MPEG-4 [2, 3] as well as their successors such as H.265 [4], aim to provide a high level of compression and scalability required for the forthcoming ubiquitous video services. At the same time, new streaming methods such as the HTTP-based rate adaptive video streaming techniques have emerged, and existing methods have been improved in accordance with the scale of the existing networks and the upcoming demands of the users [5]. These developments influence the services which are under the control of the network operators as well as the internet-based Over-The-Top content (OTT).

The above mentioned evolutions of the video-related techniques alongside the growing expectation for high-level service quality have already created new challenges for the network operators and the carriers of the service [6]. Subsequently, existing strategies for a service such as multimedia streaming need to consider a wide range of heterogeneous specifications based on the network infrastructures, transmission protocols, communication systems and various types of user interfaces. This in particular is vital in time-variant and increasingly congested infrastructures such as wireless mobile communications. Furthermore, the specifications of the content of the service, transmission network infrastructure as well as the employed protocols all influence the effectiveness of the resource utilisation strategy and should be taken into account in the optimisation process.

This chapter is dedicated to the study of the above mentioned developments with regards to the video-related services in the context of the imminent mobile communication technologies. Our main focus will be on the latest delivery infrastructures, employed protocols and quality assessment

techniques for video streaming in the new generations of mobile communication systems (e.g. 3GPP 3G and 4G-LTE). The rest of this chapter is organized as follows: the latest delivery technique development including the delivery network infrastructure and various available streaming services will be explained first. Then the main streaming protocols which have already been used or are under development for the future standardization will be discussed. Finally, the approaches to service quality assessment and the support provided for these methods in mobile communication standards will be explained. The discussions comprise some examples from the publicly available protocols to proprietary protocols which are related to the recent mobile communication standards.

2.2 Content Delivery Infrastructure

Nowadays different types of contents are offered to the consumers through the internet-based services and other private or public networks. These include data in the form of text, still image, audio and video files. Video-related services are either based on the previously recorded contents or real-time broadcasting. The required functionalities to support these services on the user side and network side will obviously be different. Nevertheless, all of them comply with the general model of a communication service which consists of a source, a destination and a communication channel as depicted in Figure 2.1(a). Figure 2.1(b) shows the basic entities required for a real-time or on-demand video streaming service with a wireless access interface. This model consists of a client entity on the user side and a server entity as the content/service provider. The channel in Figure 2.1(b) comprises all of the elements involved in the communication facilitation for this service which may include some elements from the public network, operator's infrastructure and the user's last mile access interface.

Various aspects of this model such as the provider's delivery options, employed transport protocols as well as the application protocols will be discussed, together with the approaches to handling the quality assessment of the service, later in this chapter. As a prerequisite, however, the performance of this service mainly relies on the capability of the network infrastructure for providing the satisfactory streaming quality. This includes the latency and the throughput criteria required for a minimum service quality achievement based on the available resources. As an example, Table 2.1

Table 2.1 QoS thresholds for video services

	Interactive video conferencing	Video streaming
Loss rate	< 1%	< 5%
Delay	< 150 ms	< 5 s
Jitter	< 30 ms	N/A(*)

(*) due to the larger received buffer and less delay sensitivity, no significant jitter requirement needed to be defined

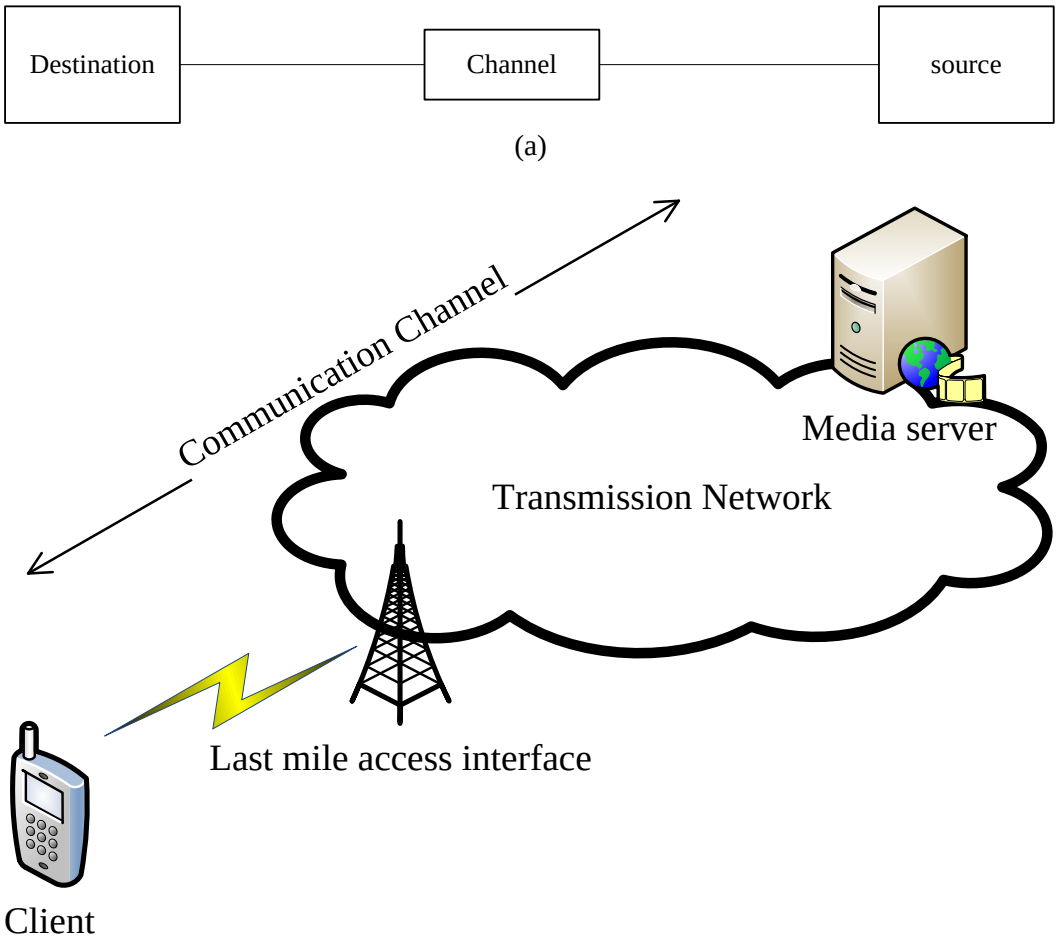


Figure 2.1 Simplified model of a video streaming service

depicts the ranges of tolerable loss, delay and jitter required to be met for the achievement of a satisfactory video service in such a network [7].

The service performance and its achievable quality are basically related to the structure of the delivery network and its efficiency. For example, in case of a globally accessible TCP-based mobile application the range of the values for Round Trip Time (RTT) and the effect of the retransmission mechanism over the latency of the service are two infrastructure-related parameters to determine the performance of this application. The latest state-of-the-art solutions for this type of challenge in a streaming scenario will be explained in the rest of this section. This includes the explanation of a general network solution for content delivery, Content Deliver Networks (CDN) [8], and the proposed infrastructure in wireless mobile communications for streaming services, i.e. Packet-switched Streaming Service (PSS) [9].

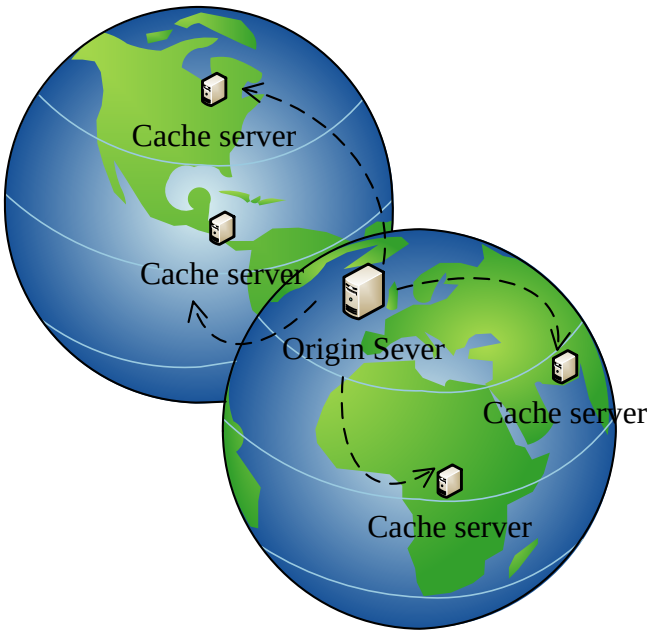


Figure 2.2 The main constituents of a globally distributed network for content delivery

2.2.1 Content Delivery Networks (CDN)

Service providers hold the original copy of their contents (audio, video and text, etc.) in a data centre where the servers are usually hosted by the delivery providers. However, the users of a worldwide service are globally distributed and their access pattern is not uniform over time. Furthermore, these servers deal with bursty traffic behaviour with probable high loads in short time intervals. Consequently, the required size of the data centre (including the number of servers) needs to be high enough to guarantee a satisfactory service quality for global users. Nevertheless, even with enough server equipment for handling the peak loads, the impact of a longer RTT over the performance of a transport protocol such as TCP is unavoidable for non-local consumers of the service.

The CDN with distributed architecture is one of the main existing solutions for the above mentioned problem. As shown in Figure 2.2, the geographically distributed locations for the CDN host servers all over the world (or all over the areas which are supposed to be covered by the service), aim to provide a server with a balanced load and optimal distance from any user across the world. These servers are potentially less likely to encounter traffic surge. Furthermore, the experienced average network latency and RTT will be much lower as compared to a unique and centralized data centre.

As presented in Figure 2.2 the ‘Origin Servers’ and the ‘Cache Servers’ are the main constituents of a Content Delivery Network. The contents which are going to be accessible to the end users are initially stored in the ‘Origin Server’. Since the ‘Cache Servers’ are updated with the duplicate of

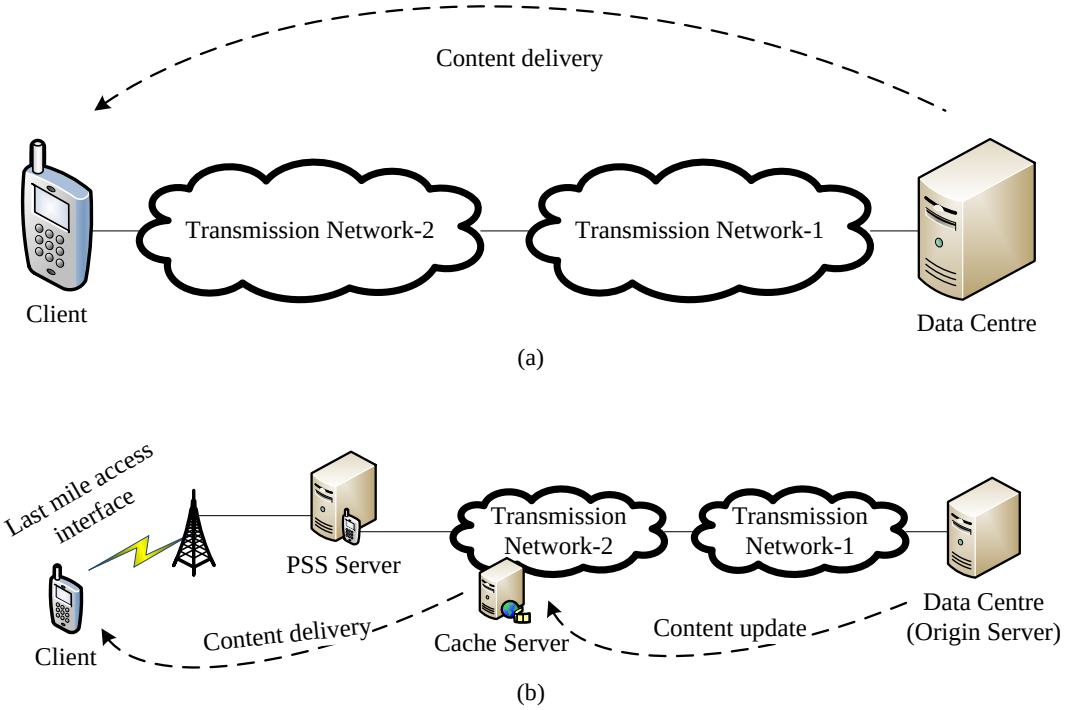


Figure 2.3 Comparison between: direct delivery from data centre (a) and distributed server in CDN (b)

the origin data, users who are asking for those contents will be redirected to the duplicate of the origin content in the nearest 'Cache Server'. It must be noted that some dynamic contents will still be fetched from the 'Origin Server' though this is done through the 'Cache Server'. 'Cache Server' is serving for the static contents in this case.

The performance of the content distribution networks can be improved by combining some other aspects of the service requirements in the content search and management algorithms. There are some examples in this area, such as using the cooperative approaches for tackling the effect of the roaming and mobility over the balance of the traffic [10], employing the quality-oriented algorithms for server selection optimisation [11], and using Peer to Peer technology to deliver the content from the nearby users instead of the Cache servers [12].

Based on the service requirement and the user's access interface specifications, some other network elements and functionalities may intervene in the delivery procedure. Figure 2.3 shows a specific streaming service, i.e. Packet-switched Streaming Service (3GPP-PSS), which is proposed as a delivery infrastructure mediator for streaming over mobile communication (e.g. 3G and 4G) [9]. This will be explained in more detail in the next subsection.

2.2.2 Content Delivery in Mobile Communications

The 3rd Generation Partnership Project (3GPP) is known as the main player in the area of the mobile communication standardization. The proposed 3GPP Long Term Evolution standard, known

as LTE/LTE-A [13], is widely accepted as the main candidate for the 4th Generation of mobile communication systems, or 4G defined by ITU [14]. This comes with a wide-ranging list of services, protocols and requirements defined under the same standardization body for the emerging communication networks including 3G, 4G and beyond.

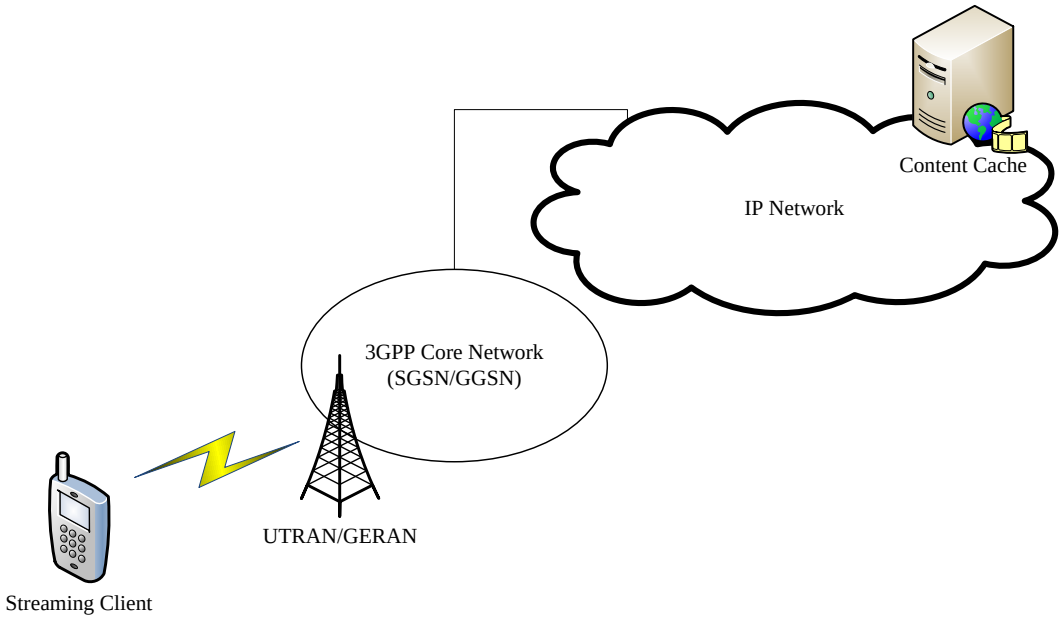
A packet-switched streaming service (PSS) has been proposed by 3GPP as a 'Transparent end-to-end packet switched streaming service' applicable to LTE/LTE-A [9]. PSS complies with the packet-switched structure of the network in 3GPP LTE/LTE-A. The PSS protocol is supposed to provide a framework for streaming services in mobile communications which bridge the gap between the existing trivial downloading protocols and the more complex conversational services. In PSS as a mobile streaming application both the protocol and the terminal specification complexity are taken into account.

PSS-related standards provide the bases required for the development of its protocols and codecs [15], session establishment, session control and data transport [16]. Furthermore, a specific file format suitable for video streaming known as 3GP file format has been defined as well [17]. IP Multimedia Subsystem (IMS) is the enabler of the IP multimedia applications in 3GPP [18]. The use of IMS standard alongside the multicast and broadcast protocols is another aspect of the PSS standards [19]. More importantly, the latest development of the rate adaptive streaming services is part of the 3GPP-PSS under the title of 3GPP-DASH [20]. In the rest of this section the architecture and protocols proposed for a packet-based streaming service in 3GPP will be explained. This will be followed by more details of the feasible delivery scenario in 3GPP, its employed data transport protocols, the structure of its quality assessment support and transactions which will give a more comprehensive insight into the streaming services in 3GPP. This information provides the main specifications of the context of the research toward the optimisation of the resource utilisation in this system.

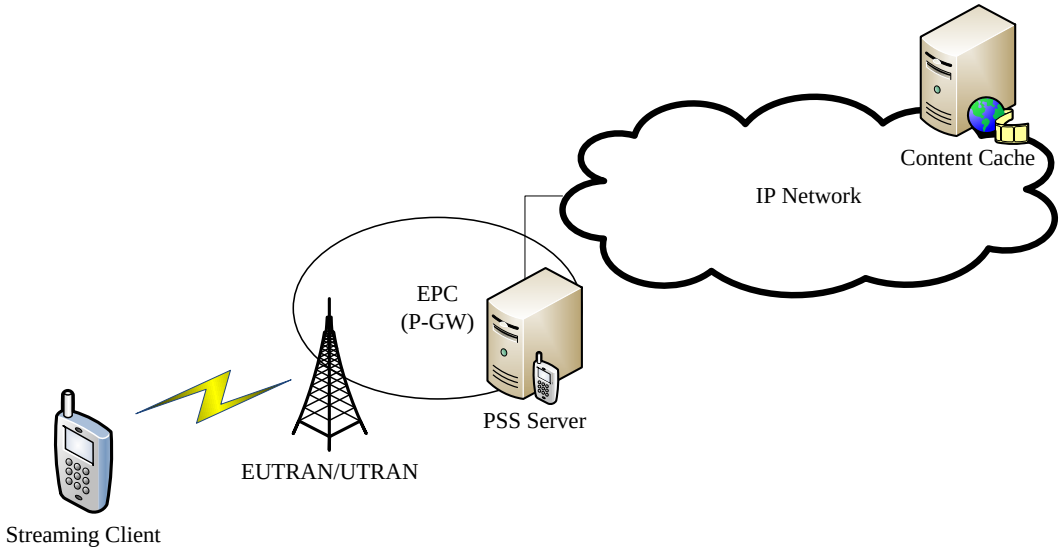
Streaming Service Architecture

A transmission network consists of different elements which provide the support for interworking between various coexisting network protocols. In 3GPP standardization for 3G and its predecessor mobile systems some functionalities are defined for adaptation to the public IP network. This comprises some extra gateway functionalities to adapt the packet-based IP network with its circuit switched compatible interface. However, in 4G, with the introduction of Evolved Packet Core (EPC) in LTE the structure of these connections will be slightly different. The EPC core of the mobile system in 4G is already packet-based and adapted with the existing IP network.

As shown in Figure 2.4, in the case of streaming service and for the mediation between the external IP network and the mobile system, 3GPP has also defined a Packet-based Streaming Server (PSS). The client, server and transmission network are the three main entities involved in a standard



(a)



(b)

Figure 2.4 a) 3GPP 3G/4G network interconnection b) infrastructure for streaming services

streaming service. The client is the initiator of the request for the streaming service. A content server provides the requested content in a streaming format. As it will be explained in sub-section 2.2.3, the streaming format could be progressive download, on-demand or live streaming. UTRAN/EUTRAN radio interface provides the connection to the network for an end-user while the established connection is going through a 3GPP defined core network toward the general IP network.

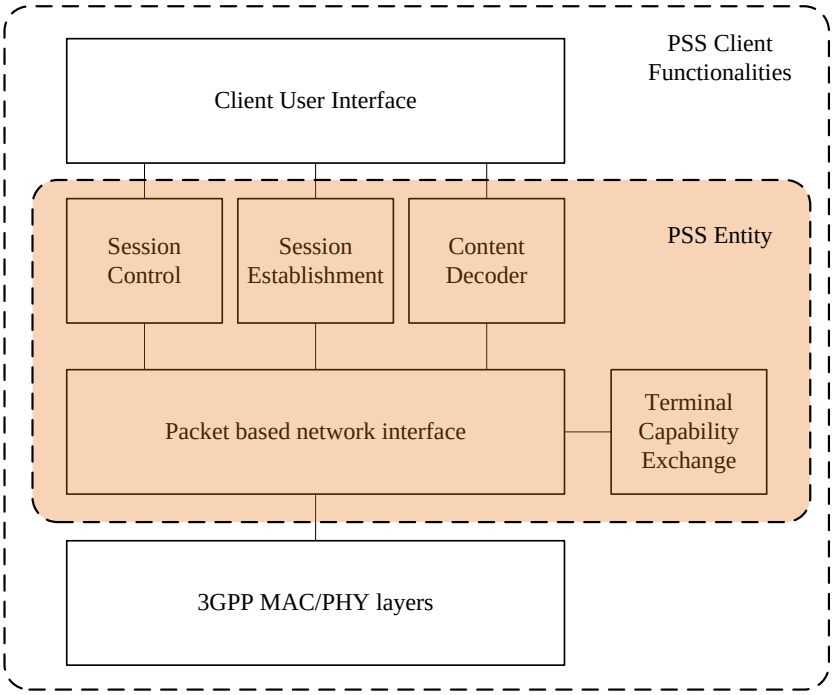


Figure 2.5 The functionalities included in a 3GPP streaming compatible client

Streaming Service Protocols

Figure 2.5 depicts the functionalities included in a 3GPP-PSS-compatible client. These functions will be supported by PSS server in the network side. PSS provides the connectivity based on which a client is receiving the requested stream from the content provider while its capabilities and preferences are taken into account by the PSS server. These functions include the establishment (e.g. invoking a PSS session from a browser) and control (e.g. setup, start and pause of a media stream) of the session as well as the messaging required for the user's capability exchange (e.g. adaptation based on the terminal capability). With the presence of the PSS in the streaming service, the content decoding will also be a part of the PSS functionalities. Depending on the type of the message and the network architecture the protocols may vary in both application and transport layers. However, in all cases the network layer protocol will be IP.

The standardized combination of protocols for application, transport and network layer in 3GPP-PSS, shown in Figure 2.6, the main audio and video contents can be transported over UDP using the real-time protocol, RTP. This is usually for the traditional streaming services. However, more recently the rate adaptive steaming technologies are mostly HTTP-based protocols over the TCP/IP transmission network. For example, the extra PSS features such as the end-users' capability is exchanged over RTSP. Transport protocol for RTSP can be chosen based on the main content's chosen protocol (i.e. UDP or TCP). As an option for the rate adaptive streaming over HTTP, the format of the file in application layer can also be a 3GP format.

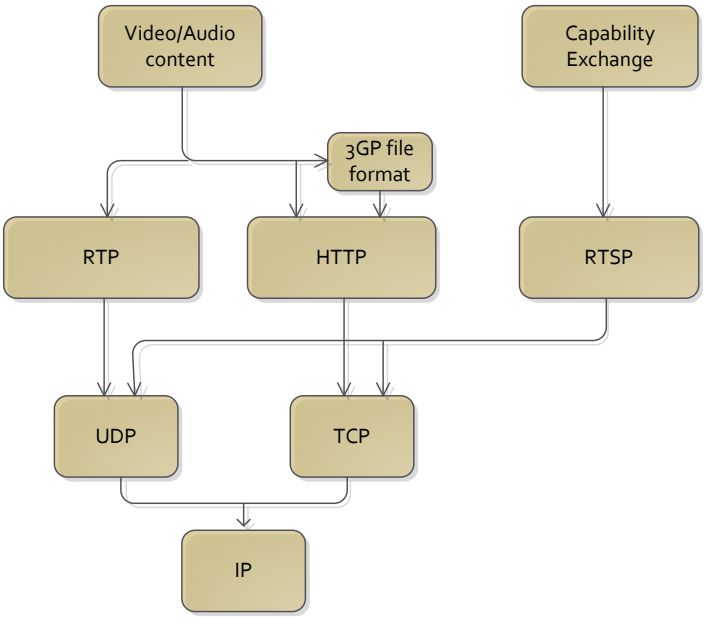


Figure 2.6 Protocol stack in 3GPP mobile-based streaming service

In the rest of this chapter a more detailed view of the main characteristics of streaming services and the formats and protocols which appear in a PSS-based interface will be discussed.

2.2.3 Video Delivery Methods

Video delivery as a streaming service is different from a normal file transfer in that the playback process of the delivered audio or video data can begin before the whole video data is received. The relation between video playback time at the user side and the actual time of the delivered content defines the exact delivery method to be employed. The phrase ‘actual time’ may refer to the video recording time (e.g. the event capturing time) in a live streaming or it may indicate the encoding and delivering time of a recorded raw video. It may even be a time reference which has been enforced through which the publishers’ right over the content accessibility can be protected. Furthermore, the constraint over the playback time and the approach which is used for caching the received data locally in the user-side can be used for delivery methods’ classification. The most common categories for video delivery which should generally be supported by a video service infrastructure will be explained further in this subsection.

Download and Progressive Download

Transferring a copy of a desired content as a file to the user-side and using the downloaded data when the whole file is transferred, is a simple and common approach to access the contents especially over the internet. This includes the video content in the form of a file available for download and the playback of the video later when download is completed. Although some video contents are still available for download and they do not require a specific video-related protocol, the main video-

sharing resources do not anymore support the access to their video contents through downloading process. This is mostly for protecting the publisher's right over their products.

For a user who is interested in just a part of a video or audio file, it will be a waste of resources to download the whole file. So, as an alternative to the 'download then playback' scenario, it is possible to start downloading the content and playing back the video while the download is in progress, hence the name 'progressive download'. No special protocol will be required in this case though the format of the file must be capable of being processed with its partially available content. Interleaved images, with half resolution at the beginning and followed by the full resolution, are examples of such a capability. It must be noted that due to the dependency of the decoding of each part of the content over the previous parts, the progressive streaming mechanism cannot adapt the data rate during the streaming procedure. Furthermore, the delivered data cannot be separated in multiple streams.

Progressive video delivery is usually carried out through a HTTP web server rather than a streaming server. There are different policies with regards to the availability of the content and the user's access to the downloaded data after playback. Downloaded data can be allowed to be stored in the user side's device, or similar to the streaming video, it can be unavailable at the end of the session.

On-Demand Streaming

The contemporary television broadcasters and popular audio and video content providers offer on-line access to their products through streaming services over the internet. These on-line video resources are usually accessible for some time after their first broadcast, hence often called 'catch-up' services. Furthermore, since the user chooses to watch the desired video in their convenient time, these services are sometimes called on-demand streaming or Video on Demand (VoD).

Personal computer interfaces and television users have initially been considered as the typical users of a video-on-demand service. This type of user is usually connected to the network through broadband access, while a standard 'set-top box' device is processing the delivered data. Subsequently, broadband access networks are the main context of the technological development for video on-demand streaming [21]. More importantly, heterogeneity was not the main concern of the initial variants of the service.

Due to the recent improvements of the access capacity provided by the new generations of mobile communication systems such as 4G-LTE, this trend has already changed [22]. This includes the consideration of a heterogeneous environment as well as the compatibility with mobile users and the developments of the network [23, 24].

Real-time Streaming

In contrast to a Video on Demand streaming service where all video frames are available during the delivery, in a real-time streaming (or live streaming) only the current frames are available for delivery from sender and playback on the user-side. Frames are not supposed to be processed after a certain time delay and any new session will start from its current frames. Due to the time constraints of the live streaming services, the required capability of the user to provide an efficient on-the-fly encoding process for the received video data is a challenge. Furthermore, the range of the tolerable delay for maintaining the service as a live service will be much lower than for an on-demand or progressive streaming.

Using a low delay transport protocol and a computationally low demand video coding technique is essential for the achievement of a satisfactory quality in real-time streaming. Later in this chapter the specifications of some protocols which have been proposed for real-time streaming will be discussed in more detail.

2.3 Streaming Protocols

As explained in the previous sections, a wide range of protocols are employed for data transmission and control of a streaming service including the standardized PSS in mobile communications. These protocols can be classified based on their server-client infrastructure, coding and compression properties or proprietary rights. Streaming protocols can also be categorized based on the role of the server and client in adaptation policy, the interaction with application layer or collaboration with transport layer. For example, the adaptive streaming development which is the main focus of this study can be considered as:

- A protocol which connects the client to the streaming server with the capability to follow the state of the clients during the session (hence it is called ‘Stateful’) e.g. Real Time Streaming Protocol, RTSP.
- A HTTP based protocol with no persistent connection between the media server and the client (hence ‘stateless’), e.g. most of the current HTTP-based and adaptive streaming services where the web infrastructure and HTTP-cashing have been enabled to mediate between client and the ‘origin’ media server.

The state of the art HTTP-based adaptive streaming technologies are provided by Apple with its HTTP Live Streaming (HLS) [25], Adobe with its Flash-based dynamic streaming [26], Microsoft with smooth streaming over Silverlight [27] and recent ISO and 3GPP standards with their HTTP-based DASH [20]. In addition, there are some service providers who provide the content delivery infrastructure for these technologies such as Akamai [28] and Limelight [29] content delivery

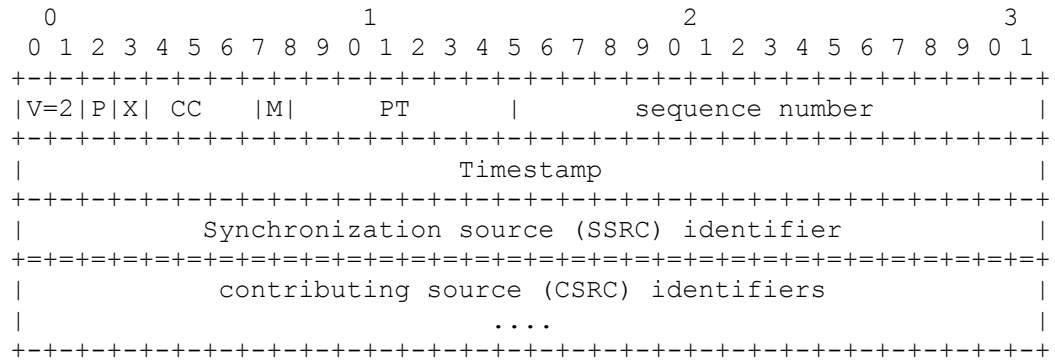


Figure 2.7 RTP header format [30]

networks (CDN). This section is dedicated to the explanation of the most common streaming protocols categorized based on the roles of the server and the client in the process of rate adaptation.

2.3.1 Server Driven Streaming

In a conventional streaming standard, the majority of the functions needed to support multimedia content delivery are handled in the server-side. These include the approach to access the multimedia components and handling the conversion between storage and delivery formats. Client is supposed to be the recipient of the service while it may process its local status to provide the required information for delivery format management (e.g. based on the channel status or buffer occupancy). Server actually remains the decision maker for delivery format adaptation and control during the session of the service. In this subsection, RTP/RTSP as an example of the protocols which can be used in a mobile communication paradigm and MMT as an emerging alternative will be discussed in more detail.

Real Time Protocols (RTP/RTSP)

Real Time Protocol (RTP) was introduced in the mid-1990s as an application layer protocol for delivery over IP networks [30]. This was a response to the shortcoming of the existing transport layer standards such as TCP and UDP in delivering real-time video and audio services. RTP compensates the simplicity of UDP protocol which is used in its Transport layer, while avoiding the inadequacy of the retransmission mechanism of TCP protocol for real-time streaming. Figure 2.7 depicts the customized structure of the overheads in RTP for multimedia streaming. As an example, the required synchronization for a real-time video data is provided by RTP timestamp. Furthermore, the format of the application layer data is defined by the marker and payload type overheads.

Real Time Streaming Protocol (RTSP), as a control protocol over IP networks, is suggested by the 3GPP standards for packet streaming in mobile communication alongside RTP for the delivery of the main data stream [16]. Both TCP and UDP can be used as transport protocol of RTSP messages. However, since latency is more tolerable for RTSP compared to RTP, the reliability of

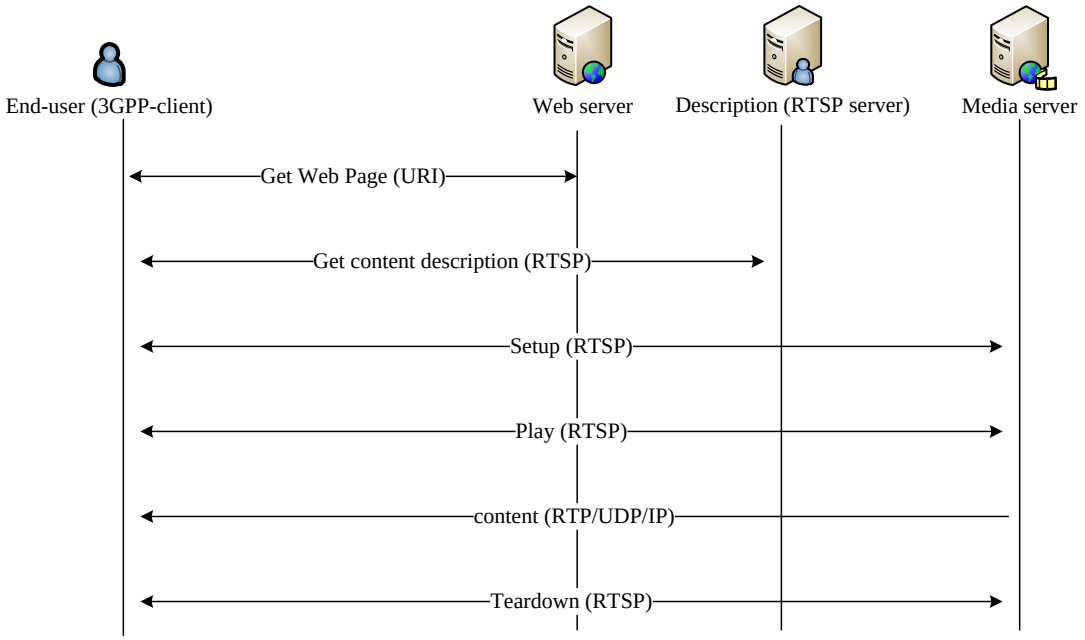


Figure 2.8 Schematic view of a basic streaming session using RTP/UDP for main content and RTSP for streaming control [9]

TCP is usually more preferable in the case of RTSP. With the help of RTSP the main control commands (such as setup and play) as well as the information of delivery mechanism (such as application type and delivery method) can be exchanged.

Figure 2.8 shows an example of the schematic view of a streaming session establishment and control in PSS. Service starts by invoking the required content from web server. Then RTSP is used to fetch the content description file as well as to run the initial setup and play procedures. Content is delivered toward the end-user through RTP over UDP. Finally, RTSP is used to shut down the service.

MPEG Media Transport (MMT)

MPEG Media Transport (MMT) is an example of contemporary server-driven streaming protocols which are under development as alternatives to the conventional real-time protocols such as RTP. MMT integrates a wide-ranging set of functionality and usages in one all-IP standard. This includes storage, delivery, stream multiplexing and other required functionalities for a comprehensive media transport protocol. This is a very robust design for a protocol compared to the conventional real-time protocols such as RTP. It is also intended to be used as a high efficiency media delivery in heterogeneous environments. MPEG has initiated the development of MMT considering the new challenges of the digital broadcasting and streaming over IP [31]. A client-driven adaptive streaming standard (i.e. MPEG-DASH) is also under development by MPEG and alongside MMT, which will be explained later in this chapter [32].

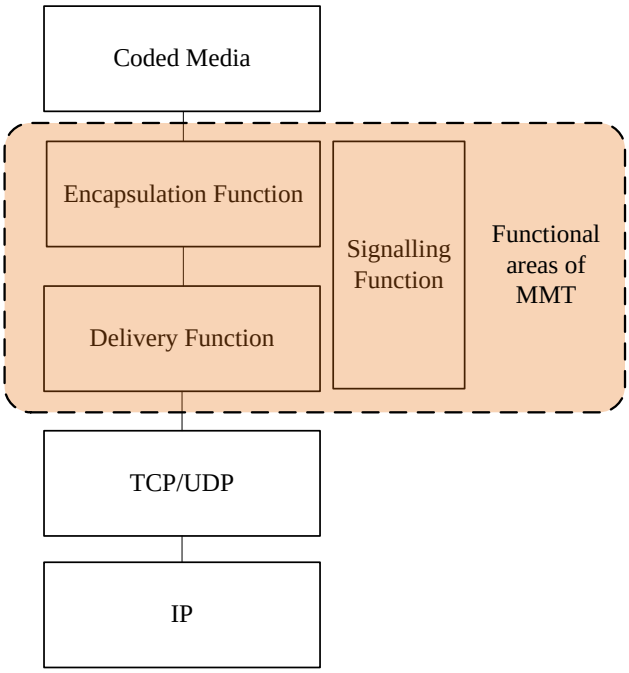


Figure 2.9 Functional areas of MMT

MPEG-DASH and MMT are designed for two different paradigms: in DASH the client manages the session and it should work with the existing infrastructure (e.g. HTTP). MMT is designed to be a replacement of RTP where the server manages the session and a specific server (similar to the RTP server) is needed to handle the procedure. It must be noted that the existing delivery network (i.e. CDN) doesn't support MMT at the moment.

As it has been depicted in Figure 2.9, MMT covers three major functionalities required for the delivery of multimedia content. This architecturally can be explained as encapsulation, delivery and signalling functions.

Encapsulation function: Using the ISO Base media file format, the encapsulation function defines the logical and physical structure of the content (i.e. mpeg4). This will be either cached at a network node for delivery preparation or will be packetized to be delivered over the network.

Delivery function: As an application layer protocol, the packetized data required to support the multimedia streaming through a heterogeneous network environment as well as the structure of the payload for encapsulated data will be defined by delivery function.

Signalling function: The configuration and format of the above mentioned protocol and payload structures will be managed by the signalling function.

2.3.2 Client-Driven Adaptive Streaming

In contrast to the above mentioned server-driven streaming standards, client has a high level of contribution in the adaptation process in a client-driven streaming standard. Furthermore, the current client-driven adaptive streaming protocols are mostly implemented as HTTP-based adaptive streaming. Subsequently in client-driven adaptive streaming services, the responsibility of the media server can be handed over to the edge servers in the delivery network. DASH (Dynamic Adaptive Streaming over HTTP) is one of the latest streaming protocols developed by 3GPP (i.e. 3GP-DASH) and adopted by MPEG (i.e. MPEG-DASH) as a client-driven protocol. This protocol alongside some other proprietary alternatives will be explained in more detail in this section.

3GP/MPEG-DASH

3GP-DASH, as a part of PSS, specifies the 3GPP approach for streaming service in a way that all service resources (the main media content and its metadata) are accessible through HTTP-related protocols. 3GP-DASH content consists of the main media content (e.g. video and audio data) and its description file (Media Presentation Description (MPD)) specified in binary and XML formats, respectively. The following formats have been defined by 3GPP as a requirement for 3GPP-DASH:

- ***Media Presentation Description (MPD):*** Describes the formats for announcing the HTTP-URLs for accessing the Segments of the main media data. It also provides sufficient information for demultiplexing, decoding and rendering the main media content in the client side.
- ***The Segment format:*** Specifies the formats of the response to an HTTP GET request based on the resource identified in the MPD. Segments comprise the coded media data and metadata.

3GP-DASH is the latest developed version of a previously available 3GPP Adaptive Streaming method known as Adaptive HTTP Streaming (AHS) in 3GPP-PSS (up to release 9 of [15]). 3GP-DASH is also different from former sever-driven and RTP-based streaming methods in 3GPP and can be deployed independently by using an HTTP-server. The adaptation for wireless in 3GP-DASH is actually achieved through the file format (i.e. 3GP file format), suggested codec and platform within the 3GPP-PSS specifications [15, 17]. Generally, the DASH client and server can be deployed independently or they may be sub-functions of the PSS client and server.

As it has been mentioned earlier, 3GPP adaptive streaming is the predecessor of MPEG-DASH of ISO. MPEG-DASH followed the 3GP-DASH to adopt it for its MPEG file format and in a wider scope as an international standard. The architecture of the elements and functionalities in MPEG-DASH and 3GP-DASH are almost the same. Main difference is the supported file format (3GP vs

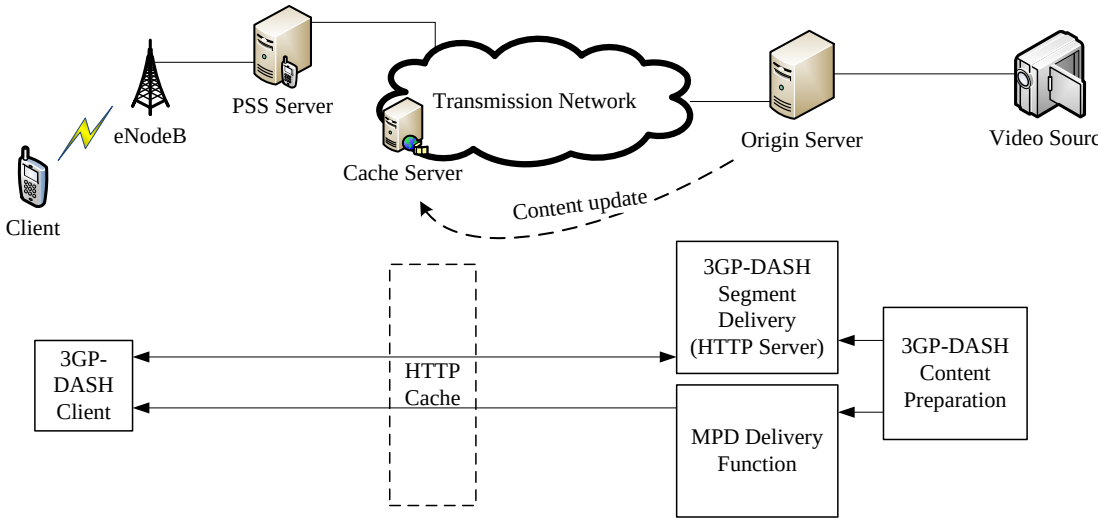


Figure 2.10 System Architecture for 3GP-DASH

MPEG) and the approach toward the representation and segmentation definitions in its media description [20, 32].

As it has been explained in the previous sections, in Content Distribution Networks (CDNs) the 'origin server' or its HTTP-caches can be used to deliver the main content of a streaming service (e.g. 3GP file data of a 3GP-DASH service). This has been depicted in Figure 2.10 where the metadata (MPD) and the main media contents (DASH segments) are delivered through an HTTP Cache. The DASH-client and the server of the DASH files (MPD and segments) must be able to handle the standard HTTP/1.1in this case [33]. Figure 2.11 shows the 3GP-DASH protocol stack comprising the media data and MPD over HTTP in the application layer and TCP/IP in lower layers. The main audio/video data is supposed to be transferred in a 3GP format in this case.

DASH Client

In HTTP-based streaming the intelligence and session control has moved from the network to the client. As an example, in 3GP-DASH the client fully controls the streaming session, i.e. it manages the on-time request and smooth playout of the sequence of the segments. It also adjusts the bitrates or other attributes as a reaction to the changes of the device status or the user's preferences [34]. So the main responsibility of the client's front-end will be: receiving the MPD, constructing and issuing the requests to the server-side, and receiving the Segments or part of the Segments.

DASH-related standards do not define any specific metric, algorithm or technique for the evaluation of available bandwidth and adaptive rate selection in the client side. The QoE parameters which have been defined to be provided by 3GP-DASH client (e.g. throughput, initial delay, buffer level ...) can be used for evaluation though these are options and can be changed based on the developers' adaptation algorithm.

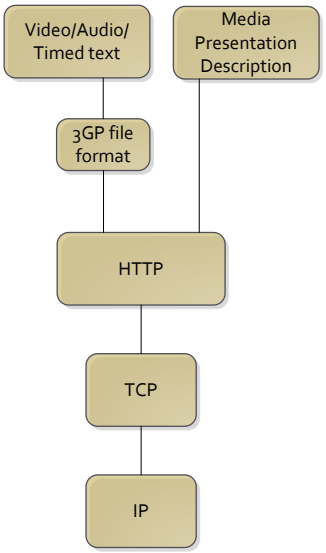


Figure 2.11 Protocol stack of 3GP-DASH

Media Presentation Detail

3GPP standardization was initially an RTP-based streaming which evolved as an adaptive http-based streaming (i.e. AHS) [15]. Then a separate standard was published for progressive as well as the Dynamic Adaptive Streaming over HTTP (3GP-DASH) [20]. 3GP-DASH can be deployed independently from its 3GPP predecessor using a standard HTTP/1.1 server as its host.

As it has been explained in the previous subsection, the 3GPP standard specifies the procedure for content delivery through the HTTP-based segments of the media content. This is done by using the provided descriptive manifest file called Media Presentation Description (MPD). MPD provides the information about the codec, language, Rights management (DRM), resolution etc. the procedure for determining the Segment size and updating the MPD may differ in the case of live streaming or on-demand streaming.

Figure 2.12 shows the hierarchical data model of MPD in 3GP-DASH which will be reflected in a XML description document with the elements exemplified in Figure 2.13. MPD data model comprises the following hierarchical sequences of data:

- ***Period:*** A period during which the available specifications of the encoded media content don't change. These include the set of available bitrates, languages, captions, subtitles etc.

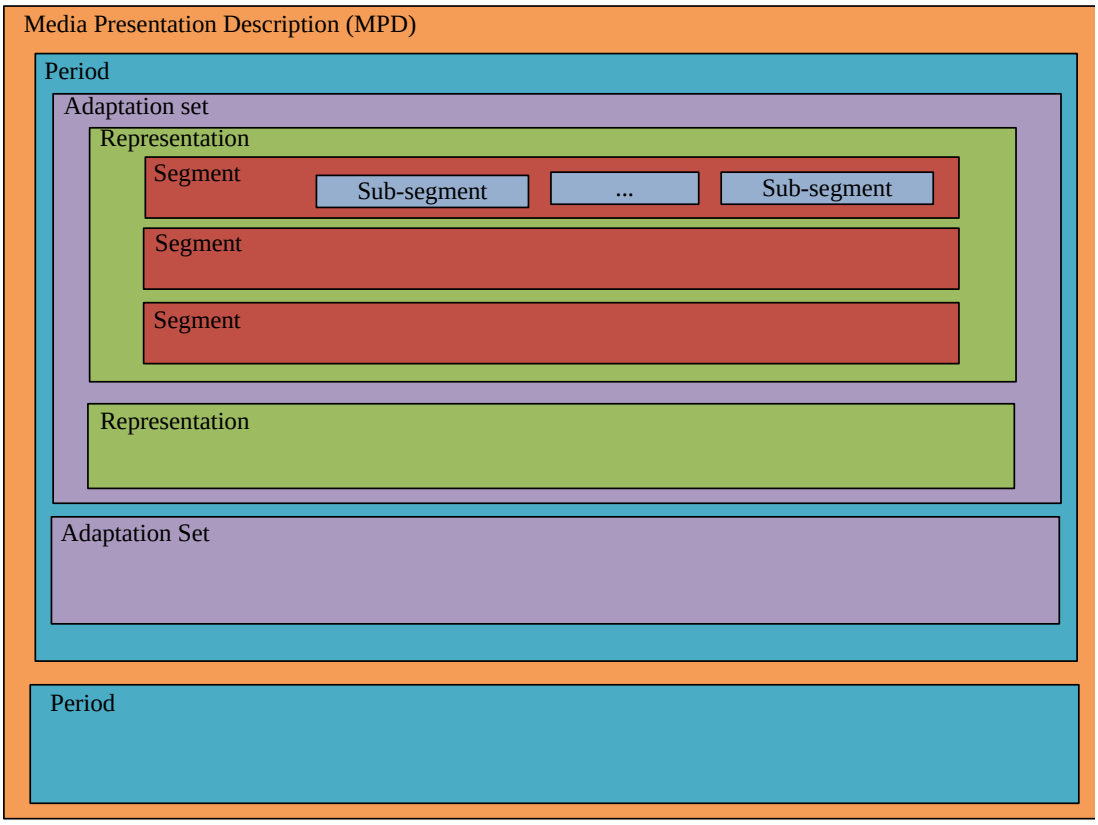


Figure 2.12 The structure of a Media Presentation Description in 3GPP-DASH

```
<?xml version="1.0"?>
<xs:schema targetNamespace="urn:mpeg:dash:schema:mpd:2011"
    attributeFormDefault="unqualified" elementFormDefault="qualified"
    xmlns:xs="http://www.w3.org/2001/XMLSchema"
    xmlns:xlink="http://www.w3.org/1999/xlink"
    xmlns:x3gpp="urn:3GPP:ns:DASH:MPD-ext:2011"
    xmlns="urn:mpeg:dash:schema:mpd:2011">
    <xs:annotation>
        <xs:appinfo>Media Presentation Description</xs:appinfo>
    </xs:annotation>

    <xs:import namespace="http://www.w3.org/1999/xlink"
schemaLocation="xlink.xsd"/>
    <xs:import namespace="urn:3GPP:ns:DASH:MPD-ext:2011" schemaLocation="3gpp-
2011.xsd"/>

    <!-- MPD: main element -->
    <xs:element name="MPD" type="MPDtype"/>
    ...

</xs:schema>
```

Figure 2.13 Overview of XML schema of the MPD

- ***Adaptation set:*** The ‘Adaptation Set’ represents the encoded versions the media content components. Usually there is, at least, an Adaptation Set for the main video component and a separate one for the main audio component. There may be more Adaptation Sets for other materials such as captions or audio descriptions.
- ***Representation:*** Representations provide the encoded versions of the components and the media stream. A single Representation is sufficient to render the contained component. Client actually can switch from Representation to Representation for the purpose of the adaptation.
- ***Segment:*** The media content is divided, in a timely manner, into Segments. A segment is the largest unit of data that can be retrieved with a single HTTP request within a Representation. A URL is provided to access each Segment. Segments are the largest portion of data that can be accessed through the provided URL. However, Segments may be further divided into ***Subsegments***, for example, based on the movie fragment.

MPD provides the address for accessing the desired content (i.e. segment index) with an appropriate specification (e.g. a certain video bitrate) which suits the current status of the client. This means that, for example, a video has been segmented in time (segments of 2-10 secs) and encoded with different code rates, each with a specific reference in the MPD. So the client can be referred to the most suitable video bitrate through its provided address in the MPD based on its current status.

The client may change the specification of the request whenever its status is changing. This is through referencing (indexing, URL address etc provided by MPD) to a different specification of the same content. The latest version of DASH also supports DRM for the most common video services including ‘subscription video on demand’ (SVOD) and ‘Over-the-top content’ (OTT) for various playback platforms.

State of the Art Alternatives for 3GP/MPEG-DASH

3GP/MPEG-DASH is a HTTP-based protocol which is capable of providing an adaptive rate streaming service. Currently a few other licence-free or proprietary standards are available which provide the similar functionalities and are widely used for streaming services. The most common protocols among them which can be considered as the predecessor or current alternatives to 3GP/MPEG-DASH will be explained briefly in this subsection.

HTTP Live Streaming (HLS): HLS is a streaming protocol based on MPEG2-TS, developed by Apple for their player (QuickTime) and any other devices which are based on Apple’s operating system, iOS. As it is suggested by its name, HLS is a HTTP-based protocol capable of being deployed over an HTTP server [25].

HTTP Dynamic Streaming (HDS): HDS is Adobe's standard for audio and video streaming with the capability of adaptive streaming over HTTP. Its user-side device must be compatible with Adobe Flash and its widely used player, Flash Player. HDS uses MPEG-4 (mp4 file format) for delivery. HDS is actually an alternative to the traditional RTMP protocol which was previously developed by Macromedia for adaptive bit rate streaming [26].

Microsoft Smooth Streaming: In Microsoft IIS smooth streaming, similarly, the server plays no part in the bit-rate switching process. The client-side decides when to request higher or lower bit rates from the server based on the monitored chunk download times, buffer fullness, rendered frame rates, and other factors [27].

Streaming in HTML5: HTML is the main protocol for presenting the contents of the World Wide Web to internet users. HTML5 is the latest revision of this standard. One of the major evolutions in HTML5 is the integration of video in the protocol as an alternative to the currently required external plug-in. Currently the video players and their plug-in capabilities are not consistent and they must be separately added to the browsers. HTML5 is an approach that aims to provide a consistent tool for a heterogeneous environment including the wide ranges of the devices which are available in wireless/mobile [35].

2.4 Video Service Quality Assessment

The quality of a communication service can be assessed technically from a system efficiency and operational point of view or based on the end-user's perception of the delivered service. The former aspect is usually referred to as the 'Quality of Service (QoS)' and the latter one is known as the 'Quality of Experience (QoE)'.

QoS is an objective concept meaning that it relies on one or more precisely defined quantitative scales. In contrast, QoE is a subjective parameter to assess the satisfaction of the end-user based on their observed service quality. Furthermore, QoE is largely related to the expectation of each user and varies from one person to another. The loss rate, throughput and latency are among the most common objective metrics for the evaluation of QoS in comparison with some predefined thresholds. Mean Opinion Score (MOS) is one of the metrics which reflects the subjective opinion of the user about the received service.

In the rest of this section QoE in video streaming services will be discussed in more detail. This includes the sources of the impairments, the approaches to assessing the experienced quality and the provided facilities for quality-related transactions among the network elements in the context of mobile communication standards.

2.4.1 The Sources of Streaming Impairments

The achievable quality of a streaming service is vastly related to the level of quality control provided by its application layer. This includes both the fidelity and the continuity of the service at the end-user's application layer. However, it is also essential to consider the impacts of the lower layers' protocols over the degradation of the quality of the service. This is especially more important with regards to the impacts of the transport layer protocol which is the carrier of the streaming packets. The impact of UDP as the transport layer protocol on RTP and the impact of TCP as the chosen transport protocol on HTTP-based streaming services are two examples of this challenge.

UDP is a connectionless transport protocol which is chosen for the streaming transport layer mainly due to its timely efficient performance. Without the congestion control and retransmission mechanisms, UDP is a protocol with optimum transmission time required for a time sensitive service such as video streaming. However, this is at the expense of no control of packet loss. UDP does not provide any recovery mechanism for lost packets which eventually leads to quality degradation in application layer. In most of the contemporary video compression formats, lost packets not only affect decoding the immediate involved frame but also it affects the other frames through the inter-frame prediction. These effects cause visual fidelity problems and can be seen as picture motion drift, blurriness, blockiness and loss of frames. It is possible to amend this shortage of UDP by employing error detection and correction techniques from other layers in a cross-layer enhancement paradigm (e.g. using CRC, FEC or hybrid algorithms in lower layers or in the application layer). However, the unreliability of UDP is still widely known as a source of fidelity impairments with regards to the streaming services [36].

TCP, in contrast to UDP, provides a high level of the accuracy for the received packets of data through its loss control mechanism. However, the streaming video packets may suffer a considerable delay due to the congestion control mechanism in TCP. This is especially more damaging in a live streaming scenario where the display time may elapse before the video part has arrived. Furthermore, the TCP congestion control reduces the throughput of the system as a result of congestion or random packet loss. This causes further destructive effect over the performance of the time-limited services such as live video streaming.

The type of the subsequent destructions to the service in the user-side may vary based on the incurred impairments. For example, discontinuity of the service during the video playback time (namely, pause, buffering or re-buffering) is a known result of the network performance inefficiency. This service interruption is mainly due to the limited available bandwidth, the latency of the TCP control and rate throttling mechanisms. The UDP protocol, however, is less likely to cause discontinuity of the service while its unreliability results in degraded image quality impairments. In the next subsection the methods for assessing the quality of the service will be discussed in more detail.

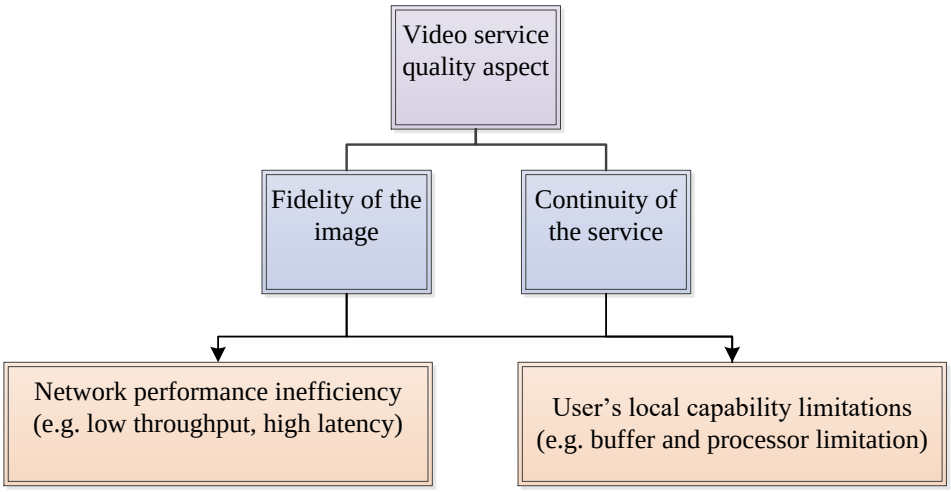


Figure 2.14 Different aspects of the video streaming quality assessments

2.4.2 Quality Assessment Metrics for Video Streaming

The assessment of the quality of a video streaming service can be defined from various points of view. As it has been depicted in Figure 2.14, quality assessment procedure can be a process to evaluate either the quality of the image or the continuity of the received video which is played-back at the receiver. In both cases the examined parameters which represent the quality of the service can be related to a local resource at user-side such as the pre-decoding buffer status or the process power allocated to the service. Quality representative parameters may also be related to the transport network performance e.g. the average throughput of the network or the incurred latency.

The performance evaluation methods employed for video streaming services can also be classified as objective or subjective methods (Figure 2.15). An objective approach employs a metric which provides a quantitative scale to assess and compare a certain characteristic of the service. The objective metric usually represents the exact status of the evaluated characteristic of the service in a precisely defined numerical form or based on a mathematical model. The level of the image

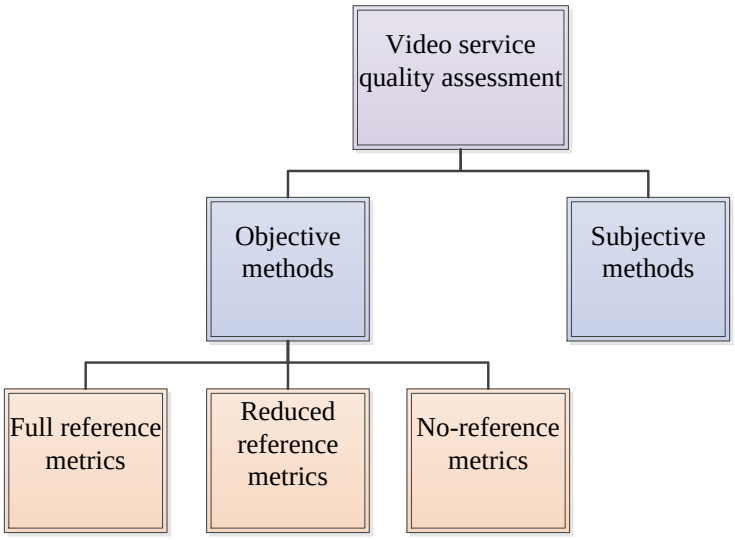


Figure 2.15 Different aspects of the video streaming quality assessments

distortion in comparison to the original transmitted image (known as the Peak Signal to Noise Ratio, PSNR) is an example of an objective metric.

A subjective approach, on the other hand, relies on a metric which represents different levels of the perceived service performance in a qualitative way. These levels can also be transferred to numerical scores if necessary. However, the employed transfer function is usually obtained from an experimental procedure per se. Mean Opinion Score (MOS) for user's satisfaction evaluation is an example of subjective metrics. For MOS rating, the voter (i.e. user of the service) gives a score of Excellent, Good, Fair, Poor or Bad to the experienced service quality (e.g. the quality of the video streaming playback). These values can be interpreted numerically as 5, 4, 3, 2, or 1, respectively.

Traditionally, the objective approaches to assessing the quality of the received video data in a video streaming service are classified further based on their dependency on the original video at the sender-side. The '***full reference (FR)'*** approaches are defined based on the comparison between the received video at the receiver and the original video at the sender-side. This means that, during the assessment time, the reference video needs to be sent to the receiver. Reference must not be exposed to any loss and must be at the receiver not later than the streamed video. Even with the help of the reliability of TCP, the extra required bandwidth and unavoidable delay to achieve high accuracy will make this approach impractical for on-line and on-the-fly assessments in a real application. However, in spite of this restriction, FR methods are still very accurate and suitable for offline mode assessment or evaluation of the applications during their development procedure and R&D tests.

In the '***Reduced Reference (RR)'*** method certain specifications of the transmitted video will be extracted and sent to the receiver for comparison. Some markers may also be inserted into the bit-stream to make some details of the reference available at the receiver. This is more practical in the case of online assessment though it still consumes extra bandwidth for sending additional information. However, the time constraint of the reference data which must arrive prior to the part of the video that it is being used to assess, remains as a challenge for RR methods.

The most practical method for video streaming assessment is the '***No Reference (NR)'*** method in which there is no need to access the reference signal. The assessment in this case depends on a pixel-based evaluation of the received image and/or a bit-stream valuation of the received stream of the data. Since the evaluation reference is supposed to be unavailable at the receiver, the outcome of the assessment must be interpreted based on some predefined assumptions. These assumptions are either compared with predefined thresholds or subjectively related to the user's satisfaction levels. The QoE metrics usually fall in this category of assessment methods.

Table 2.2 Examples of different quality assessment metrics

Metric	Reference dependency	System aspect	Example
PSNR	Full	Cumulative squared error (MSE)	[37], [38]
VQEG – ITU-T J.247	Full	Coding and packet loss degradation,	[54]
VQEG – ITU-T J.341	Full	HDTV, Freezing	[55]
PVQM	Full	Frame-based degradation comparison	[39]
MOSp	Full	MSE based on pattern recognition	[40], [41]
SSIM	Full	Structural (Blockwise) comparison	[42]-[45]
VQM	Full	Blurring, jerkiness, noise and colour distortion	[46]
DVQ, JND	Full	Perceptible frame distortion	[47], [48]
ITU-T J.247	Reduced	Frame by frame basis, original edge values is available	[49]
Bluriness metrics	No	High Spatial frequency attenuation	[50], [51]
Blockiness metrics	No	Luminance discontinuities	[52]
Buffer underrun	No	Buffer size and occupancy	[53]
Pause Intensity	No	Discontinuity of playback	[36]

Although FR metrics provide theoretically more accurate video quality evaluations, it has been shown that the RR and NR methods also produce a high level of correlation with perceived video quality. A satisfactory level of this correlation alongside the practicality of a NR method will justify its usage for on-line video service assessment [36].

Table 2.2 depicts some examples for each of the above mentioned assessment methods [37-55].

2.4.3 QoE Signalling in Mobile Video Streaming

Quality evaluation procedure is a task performed in the client-side of a video streaming service. The result of the evaluation procedure can be used locally as a service quality indicator or as a control parameter to improve the performance of the application. The assessment result can also be transferred to the server-side or any other intermediate network element to provide their required information for the service adaptation and improvement. Therefore, the quality-related frameworks comprise: 1) the requirements for quality evaluation process, locally and 2) the transactions between the user-side and other network elements with regards to the quality of the service. In this subsection

the QoE metrics defined in 3GPP mobile communication standards and the format of the QoE report will be discussed in more detail.

The QoE support in a mobile streaming service is actually an attribute defined as a part of the device capabilities. Device capability and user profile are the information which facilitate the server-side negotiation for matching process at the beginning of the streaming. The device capability description is exchanged between the profile server (provided by device manufacturer or mobile operator) and streaming server (i.e. PSS) as a part of the negotiation and matching mechanism [16].

Table 2.3 shows some of the attributes related to video streaming service included in the capability exchange signalling in 3GPP-PSS. These consist of the 'QoE support' attribute as well as the type of the carrier protocol (i.e. RTP/RTSP or HTTP) and the available options for rate adaptation mechanism. Figure 2.16 depicts the functional components of the capability exchange procedure given the protocol of transaction exemplified in Table 2.3. The mobile device sends (through the RTSP/HTTP data units) the URL of the location where the PSS can retrieve the user profile. PSS communicate with profile server (through HTTP Request/Response) to access the device description information. User can override or add an attribute during this process. The final negotiated attributes will be used by PSS to control the presentation of streamed media content to a mobile user.

In the case of a client with supported QoE attribute, the main video QoE-related metrics which are defined by 3GPP to be included in the report are:

- Average Throughput
- Initial Playout Delay
- Buffer Level
- MPD Information

As an example, Table 2.4 shows the main specifications defined for 'Average Throughput' and 'Buffer Level' which must be locally monitored and reported if necessary. 'Average Throughput' is one of the network-side monitored performances which is observed by the client-side during the measurement interval. 'Buffer Level' is a client–side examined performance and shows the buffer occupancy level measured during the playout at a certain rate.

The XML syntax (using the HTTP request with XML in its body) will be used for the transaction of the QoE metrics and reporting protocols. An example of this syntax has been depicted in Figure 2.17 and Figure 2.18 where the QoE reporting protocol is based on the HTTP request signalling.

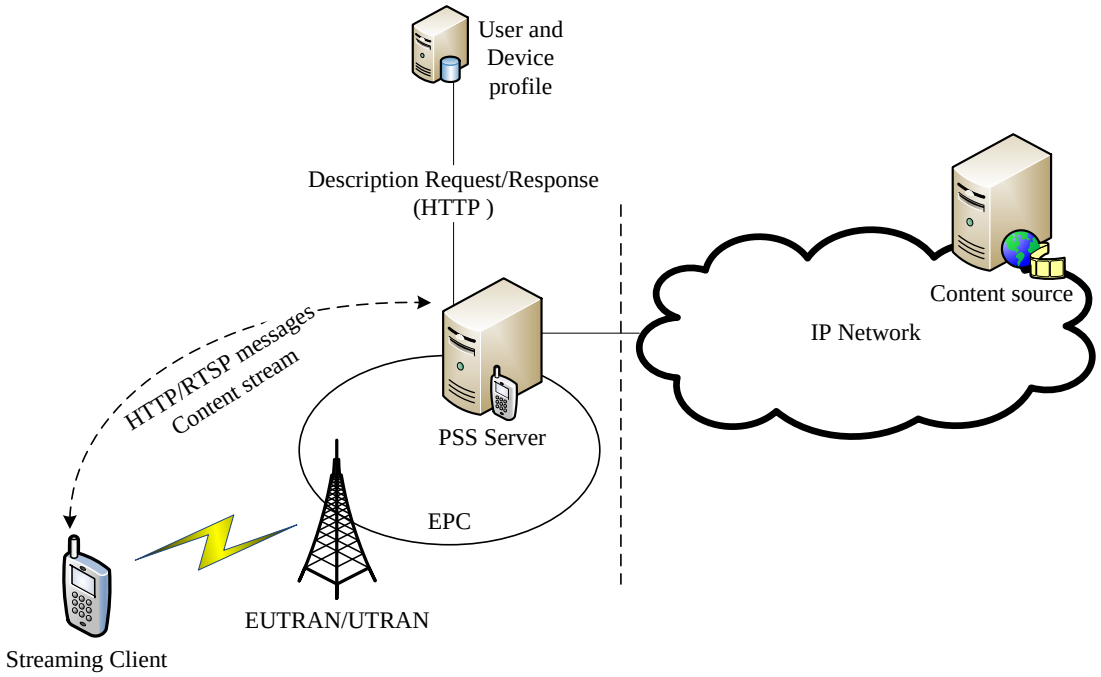


Figure 2.16 Logical system architecture of the capability negotiation mechanism applied in PSS (Functional components in PSS capability exchange)

Table 2.3 Examples of the user's attributes and capabilities including QoE support [15]

Attribute name	Definition
StreamingMethod	Support for RTP streaming, HTTP streaming, Progressive Download, or all.
AdaptationSupport	whether the device supports client buffer feedback signalling
QoESupport	supports for QoE signalling: RTSP/RTP-based or HTTP-based streaming
VideoDecodingByteRate*	defines the peak decoding byte rate the PSS client is able to support
VideoInitialPostDecoderBufferingPeriod*	defines the maximum initial post-decoder buffering period of video
VideoPreDecoderBufferSize*	defines the size of the hypothetical pre-decoder buffer

* These features are defined for the PSS video buffer model in the case of H.263 video type. Otherwise the file format will provide these parameters e.g. 3GP

Table 2.4 'Average Throughput' and 'Buffer Level' specifications as QoE metrics

Metric name	Metric parameter	description
Average Throughput	numbytes	The total number of the content bytes, i.e. the total number of bytes in the body of the HTTP responses, received during the measurement interval.
	activitytime	The activity time during the measurement interval in milliseconds. The activity time during the measurement interval is the time during which at least one GET request is still not completed (i.e. excluding inactivity time during the measurement interval).
	t	The real time of the start of the measurement interval
	duration	The time in milliseconds of the measurement interval
	accessbearer	Access bearer for the TCP connection for which the average throughput is reported
	inactivitytype	Type of the inactivity, if known and consistent throughout the reporting period: • User request (e.g. pause) • Client measure to control the buffer • Error case
Buffer Level	t	Time of the measurement of the buffer level.
	level	Level of the buffer in milliseconds. Indicates the playout duration for which media data of all active media components is available starting from the current playout time.

```
<?xml version="1.0"?>
<xs:schema targetNamespace="urn:3GPP:ns:PSS:AdaptiveHTTPStreaming:2009:qm"
    attributeFormDefault="unqualified"
    elementFormDefault="qualified"
    xmlns:xs="http://www.w3.org/2001/XMLSchema"
    xmlns:xlink="http://www.w3.org/1999/xlink"
    xmlns="urn:3GPP:ns:PSS:AdaptiveHTTPStreaming:2009:qm">

    <xs:annotation>
        <xs:appinfo>3GPP DASH Quality Reporting</xs:appinfo>
        <xs:documentation xml:lang="en">
            This Schema defines the quality reporting scheme information for 3GPP
DASH.
        </xs:documentation>
    </xs:annotation>

    <xs:element name="ThreeGPQualityReporting" type="SimpleQualityReportingType"/>

    <xs:complexType name="SimpleQualityReportingType">
        <xs:attribute name="apn" type="xs:string" use="optional"/>
        <xs:attribute name="format" type="FormatType" use="optional"/>
        <xs:attribute name="samplePercentage" type="xs:double" use="optional"/>
        <xs:attribute name="reportingServer" type="xs:anyURI" use="required"/>
        <xs:attribute name="reportingInterval" type="xs:unsignedInt"
use="optional"/>
    </xs:complexType>

    <xs:simpleType name="FormatType">
        <xs:restriction base="xs:string">
            <xs:enumeration value="uncompressed" />
            <xs:enumeration value="gzip" />
        </xs:restriction>
    </xs:simpleType>

</xs:schema>
```

Figure 2.17 Syntax of Quality Reporting Scheme Information [20]

```
POST http://www.exampleserver.com HTTP/1.1
Host: 192.68.1.1
User-Agent: Mozilla/4.0 (compatible; MSIE 8.0; Windows NT 6.1; Trident/4.0)
Content-Type: text/xml; charset=utf-8
Content-Length: 4408
<?xml version="1.0"?>
<ReceptionReport contentURI="http://www.example.com/content/content.mpd"
clientID="35848574673" xmlns="urn:3gpp:metadata:2011:HSD:receptionreport"
    xsi:schemaLocation="urn:3gpp:metadata:2011:HSD:receptionreport DASH-QoE-
Report.xsd"
    xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance">
    <QoeReport periodID="Period1" reportTime="2011-02-16T09:00:00"
reportPeriod="500">
        <QoeMetric>
            <HttpList>
                <HttpListEntry type="MPD"
url="http://www.example.com/content/content.mpd" trequest="2011-02-16T08:59:30"
tresponse="2011-02-16T08:59:31" interval="50">
                    <Trace s="2011-02-16T08:59:30Z" d="171" b="2367 1990 2463
1254"/>
                </HttpListEntry>
                <HttpListEntry type="InitializationSegment"
url="http://www.example.com/content/initRep1.3gp" trequest="2011-02-16T08:59:40"
tresponse="2011-02-16T08:59:41" interval="200">
                    <Trace s="2011-02-16T08:59:40.5Z" d="159" b="9345"/>
                </HttpListEntry>
                <HttpListEntry type="InitializationSegment"
url="http://www.example.com/content/initRep2.3gp" trequest="2011-02-16T08:59:41"
tresponse="2011-02-16T08:59:42" interval="200">
                    <Trace s="2011-02-16T08:59:41.5Z" d="123" b="6723"/>
                </HttpListEntry>
                <HttpListEntry type="InitializationSegment"
url="http://www.example.com/content/initRep3.3gp" trequest="2011-02-16T08:59:42"
tresponse="2011-02-16T08:59:43" interval="200">
                    <Trace s="2011-02-16T08:59:42.5Z" d="195" b="9786"/>
                </HttpListEntry>
            </HttpList>
        </QoeMetric>
        <QoeMetric>
            <InitialPlayoutDelay>10000</InitialPlayoutDelay>
        </QoeMetric>
    </QoeReport>
    <QoeReport periodID="Period1" reportTime="2011-02-16T09:08:20"
reportPeriod="500">
        <QoeMetric>
            <BufferLevel>
                <BufferLevelEntry t="2011-02-16T09:08:19" level="84673"/>
                <BufferLevelEntry t="2011-02-16T09:08:20" level="93874"/>
            </BufferLevel>
        </QoeMetric>
        <QoeMetric>
            <RepSwitchList>
                <RepSwitchEvent to="Rep2"/>
                <RepSwitchEvent to="Rep3"/>
            </RepSwitchList>
        </QoeMetric>
    </QoeReport>
</ReceptionReport>
```

Figure 2.18 An example QoE reporting protocol (based on HTTP POST request signalling)

2.5 Summary

The technical background, especially the latest developments of the video-related services in connection with mobile communication technologies has been discussed in this chapter. Discussions include the latest delivery network infrastructures, employed streaming protocols, various available streaming services and quality assessment techniques for video streaming in the new generations of mobile communication systems (e.g. 3GPP 3G and 4G-LTE). In addition, the quality assessment methods and the support provided for realizing these methods within mobile communication standards have been explained. Examples are provided from the publicly available protocols as well as proprietary protocols which are related to the recent mobile communication standards.

Chapter 3
Long Term Evolution (LTE): Functionality and Protocols

3.1 Introduction

The latest generation of mobile communication systems, 4G, has provided a vast amount of time-frequency resources with a considerably higher efficiency compared to its predecessors. The improved multiple access techniques in the physical layer and the approaches for resource utilization in the higher layers play important roles in this evolutional change. The OFDMA technique which has been employed in the 4G-LTE's physical layer and the robust link adaptation mechanism in the resource allocation process are two examples of the improvements that have been demonstrated. However, due to the growing number of the users and their rising demand for higher data rates, the improvement of the users' experienced quality of the delivered service remains a challenge. Especially, the popularity of the newly introduced video-based applications on Smartphones and Tablets, push the network to work at its capacity margin in spite of the higher system capacity provided. The achievable level of the service delivery quality is still limited and mainly relies on the resource management policies and proper settings of the related parameters.

An optimised resource utilisation solution which adequately takes into account the type of the service (e.g. the content of the service) is a typical approach for enhancing the level of the achievable quality. As a matter of fact, the effectiveness of a resource utilisation strategy in a changing environment such as wireless mobile communication not only relies on the amount of the available resources but also depends on the intelligence and adaptability of the allocation algorithms used. Subsequently, the algorithms which for example support the scheduling and link adaptation mechanisms play a central role in improving the performance of the system as they have a direct impact on the efficiency of resource allocation.

As it will be discussed later in the rest of this thesis, the proposed solutions for resource utilisation in our study are examined in the context of 4G-LTE. For this purpose, different aspects of data transport and resource utilisation procedures in LTE are discussed in this chapter. The rest of this chapter is organised as follows. The scope and the role of main network elements in LTE are explained first. The details of the employed functions and procedures such as retransmission, channel coding and frame timing are discussed next, followed by a functional block diagram based on the related 3GPP recommendations/specifications and some practical considerations. The similarities and the differences between the implementation in LTE and HSPA (as an example of the previous generation of the system) are also discussed, including the comparison among the initial transmission and retransmission timings, acknowledgment mechanisms, synchronization and adaptability of the operational processes. These comparisons clarify the benefit of the proposed solutions presented in this thesis.

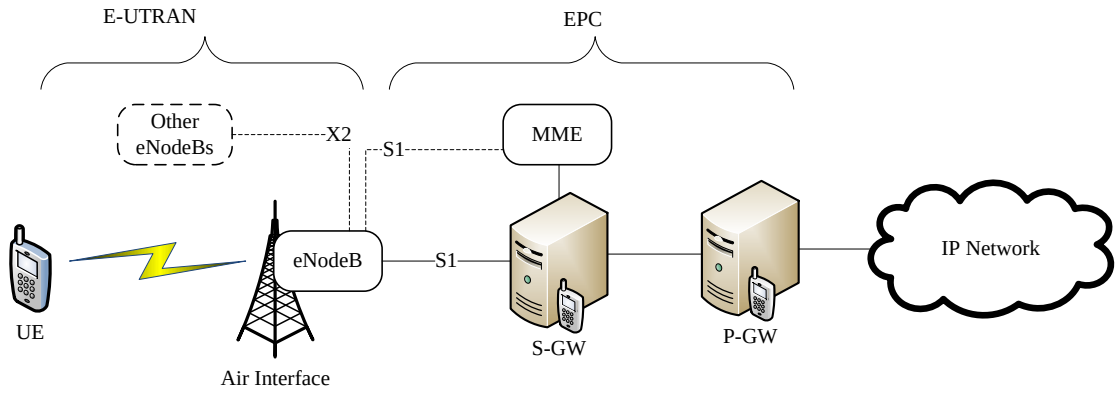
The structure of our MATLAB-based simulator which has been developed for investigating the behaviour of the scheduler in LTE and verifying the analytical models proposed is introduced in the final section, accompanied by some preliminary simulation results showing the characteristics of the simulated environment as well as the simulator performance.

3.2 Related Aspects to This Work

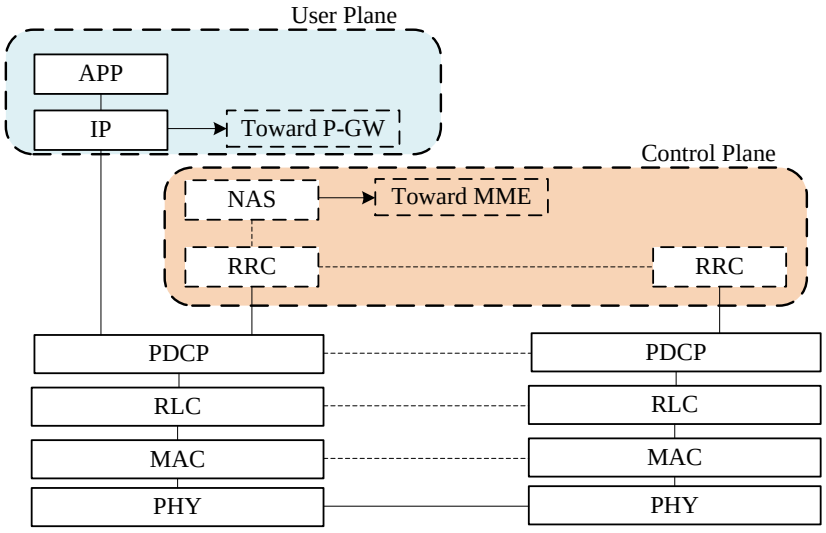
3.2.1 Overview

The International Telecommunication Union (ITU) has defined the specifications which are expected to be met in the 4th generation of the mobile communication systems best known as 4G [14]. The 4G standard aims to improve the capacity and the capability of the previous generations of the mobile communication systems through an evolved packet switched infrastructure. In the late-2000s, the 3rd Generation Partnership Project (3GPP) introduced the first version of a new generation of mobile communication systems as the main candidate for 4G [13]. This is called E-UTRAN (Evolved Universal Terrestrial Access Network) or Long Term Evolution (LTE). LTE aims to satisfy the users' demand for higher data rates, a better quality of the service as well as lower cost and less complexity of the system. The core and radio interface are fully packet-based and IP enabled. LTE is actually the access part of a system introduced by the same organisation and is known as the Evolved Packet System (EPS) [56, 57]. Evolved packet Core (EPC) is the core part of the EPS. LTE and its revised versions which later have been introduced as LTE-Advanced, provide a flexible range of bandwidth with high spectral efficiency and a high data rate achievement with short round trip time.

Figure 3.1(a) depicts the main network elements involved in a LTE network solution. The mobile base station or eNodeB (or eNB) [58], is the main network element in LTE alongside the LTE user equipment, UE [59], and the core network, EPC [57]. The structure is purely IP based and can carry



(a) E-UTRAN and EPC network elements



(b) user and control plane protocol stack

Figure 3.1 Overall EPS architecture and protocol stack

both real-time and non-real-time services. A combination of OFDMA (Orthogonal Frequency Division Multiple Access), high order modulation, large bandwidths and spatial multiplexing (up to 4x4, four antennas at both transmitter and receiver) have been used in LTE access solution. Subsequently, the data rate of 75 Mbps in the uplink and 300 Mbps in the downlink (using spatial multiplexing) are achievable in LTE. The mobile base stations in LTE are able to communicate with each other, facilitate their data and control exchanges through a dedicated interface, i.e. X2 [60]. This is important for the enhancement of the mobility services such as handover. The mobile base stations are also connected to the backbone through its specific interface S1 [61]. For Each user (i.e. UE) a serving gateway (S-GW) is responsible to terminate the traffics associated with the EPS. UEs' connectivity toward the external Packet Data Network (PDN) is provided through PDN gateway (P-GW).

The network intelligence is distributed amongst the eNodeB base stations to speed up the connection set-up and reduce the time required for a handover. So, no centralized intelligent controller is needed in LTE and eNodeBs are actually interconnected for this purpose. The protocol stack shown in Figure 3.1(b), including the MAC layer and its scheduling functionality are also fully represented in the eNodeB and UE. This enables a faster and more efficient radio resource utilisation. OFDMA in downlink and SC-FDMA (Single Carrier-Frequency Division Multiple Access, also known as DFT (Discrete Fourier Transform)) in uplink facilitate a high radio spectral efficiency as well as an efficient scheduling in both time and frequency domains [62]. Available bandwidths starting from 1.4 MHz and can reach up to 20 MHz. The frequency of the operating bands range from 700 MHz up to 2.7GHz with different arrangement for Frequency Division Duplex (FDD) or Time Division Duplex (TDD) in the case of uplink or downlink [58].

The resource allocation method and the solution for the allocated resources to be consumed by each user constitute the core of the proposed resource utilisation framework in our study. The associated functions like scheduling and link adaptation are implemented in the last mile's eNodeB in LTE. However, other network elements such as UE, transmission channels and the base stations' backbone connection devices also contribute in this process.

Obviously, the involved parameters from different layers of the system hold different importance, so it is impractical to consider all of them in an optimisation plan. Based on the above explained structure of LTE and by comparing it with the previous generations of the system, eNodeB is the main network element which embraces the majority of the LTE-related functions. UE replicates those functions in the user-side. The wireless communication channel is also a main contributor in this structure which mediates between eNodeB and UE. The main parameters from eNodeB, UE and communication channel are heavily considered in our research study and will be reviewed in this section.

3.2.2 Resource Utilization in LTE Base Station (eNodeB)

As it has been mentioned earlier, the base station or eNodeB in LTE plays a central role in the E-UTRAN infrastructure. Since resource allocation, link adaptation settings and synchronization for both the eNodeB and UE in LTE are managed by eNodeB, it is actually the task manager of the whole process. eNodeB acts as the decision maker for scheduling as well as other major control mechanisms such as Hybrid-ARQ and retransmission. Subsequently, our proposed analytical models as well as the structure of the developed simulator are mainly related to the eNodeB's functionality with the consideration of the required elements from the user side and the channel. Next subsections are dedicated to the explanation of two important aspects of the resource utilization in LTE required throughout the rest of this work, i.e. 'unit of resource' and 'resource allocation strategy' in LTE.

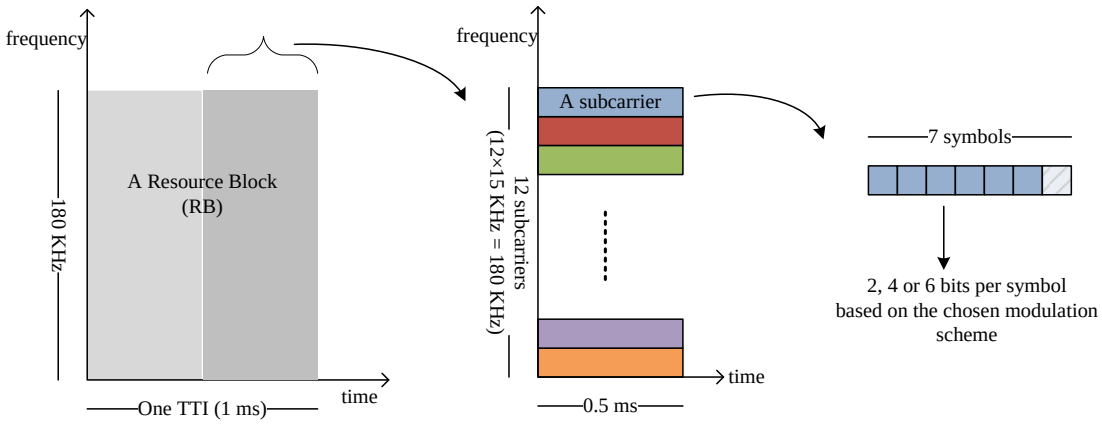


Figure 3.2 Resource Block as a unit of resource allocation

3.2.2.1 Unit of Resource in LTE

Scheduling in LTE is actually an allocation process of two dimensional resource in time and frequency domains. The resource allocation to the users is based on a unit of resource defined as a Resource Block (RB). Figure 3.2 shows the main elements of a Resource Block in LTE. Each RB includes 12 consecutive (or non-consecutive in special cases) subcarriers which represent the resource elements of the system (REs). Each RE represents a portion of the resource with 15 kHz width in frequency and the 0.5ms in time. It means that the unit of the resource in LTE will be of 180 KHz bandwidth for the length of 0.5ms. Since this allocation remains unchanged for two consecutive 0.5, the same resource allocation is actually used twice in each TTI (TTI-1ms=2×0.5ms) [62]. RB will be considered as the smallest unit of the resource allocation throughout this research and for the proposed optimisation problems. We further assume that the users do not share a unit of resource, which means each resource block is either allocated to a user or unused.

Each resource element (15 kHz, 0.5ms) consists of 7 symbols in a normal cyclic prefix. The length of each symbol depends on the type of the modulation which has been selected for that RB. The length of each symbol can be 2, 4 or 6 bits for the modulation types QPSK, 16QAM or 64QAM respectively. The overall number of the available RBs in LTE depends on the bandwidth which has been allocated to the system in each cell and ranges from 6 to 100. Table 3.1 shows the available bandwidth and RB combinations in LTE.

The control policy in scheduling and link adaptation is twofold. The first aspect of scheduling is to regulate the number of the RBs to be allocated to each user. The second aspect is the responsibility for defining the most suitable frequency slots (i.e. the set of subcarriers) which can be used to create RBs for allocation. Later we will discuss the channel status feedback (CQI) from UE to eNodeB, as a main parameter which affects the adequacy of each subcarrier to be allocated to each user.

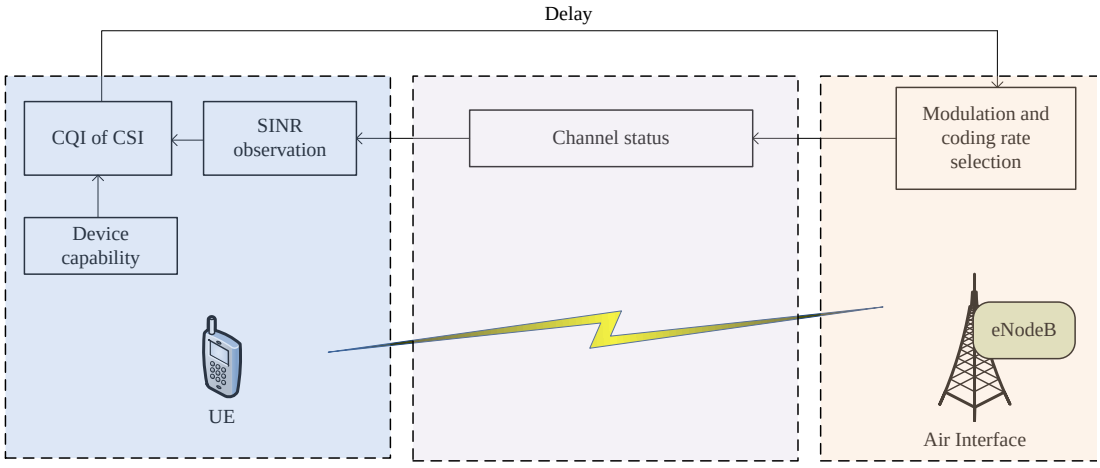


Figure 3.3 Link evaluation and CQI

As it has been explained, each resource element in a RB contains 7 symbols of the data and the length of these symbols depends on the selected modulation scheme. To accomplish an optimum process of modulation, the modulation type and code rate must be chosen properly based on the status of the channel for each user and each of the available subcarriers. This is known as Adaptive Modulation and Coding (AMC) [63, 64]. The link adaptation function in LTE is responsible for AMC and the adaptation of the most suitable modulation and coding rate (i.e. Modulation and Coding Scheme, MCS) for each resource block. Table 3.2 shows the supported combinations of modulation schemes in LTE. As it has been depicted in Figure 3.3, the adequacy of the adaptation is examined based on the latest available feedback of the status of the channel which has been reported from UE to eNodeB. This report is based on the latest observed SINR in UE and it varies due to the status of the channel. The amount of delay between the observation time in the UE and the utilization time of the report in eNodeB affect the efficiency of the adaptation strategy.

Table 3.1 Number of resource blocks in LTE

System Bandwidth (MHz)	Number of Resource Blocks
1.4	6
3	15
5	25
10	50
15	75
20	100

Table 3.2 Link adaptation options

Type of modulation	QPSK, 16QAM, 64 QAM
Modulation order (bits per symbol)	2, 4, 6
Coding rate	1/9, 1/6, 0.21 , ¼, 1/3, 0.42, ½, ...

3.2.2.2 Resource Allocation and Scheduling Strategy

eNodeB is actually a regulator of the resource allocation in LTE. It means that the algorithm based on which a specific time and frequency is allocated to a user is implemented in eNodeB but it applies to the UE elements as well. The scheduling functionality of eNodeB controls the transmission processes from base station toward the users (i.e. Downlink direction) as well as those from users toward the network base station (i.e. Uplink direction).

The strategy of the allocation and the scheduling algorithm is mainly designed to achieve a certain performance such as *efficiency* or *fairness*. The efficiency of the system can be represented by the overall created capacity through scheduling (e.g. the summation of the allocated data rates to the users) or the average information carried per unit of resource (e.g. bit/s/Hz). The latter definition actually takes into account the cost (i.e. bandwidth) of the created capacity. Fairness, in contrast, does not embody the absolute created capacity of the system. Fairness characterizes the way those allocated resources are distributed among the users.

In a heterogeneous environment and in the case of wireless communications, users with different demands are using devices with various capabilities and experiencing diverse channel status. Different received signal quality, various display resolution or the user's required data rate all contribute to the heterogeneity of the network. These differences may lead to a polarized distribution of the allocated resources and or an unbalanced Quality of Service (QoS) or Quality of Experience (QoE) among the users. Subsequently, the scheduling strategies have to define a specific target for their performance on both efficiency and fairness and a desired trade-off between these two aspects of the performance.

The Best-CQI scheduling algorithm (also known as max-C/I) is a well-known efficiency-oriented scheduling strategy in LTE. This method aims to maximize the total throughput of the system while prioritizing the users with the highest capability in their receiver to utilize the allocated resources. This capability is represented by the users' CQI feedback to the eNodeB's scheduler. Such a strategy obviously concentrates on the users with good reception or highly capable devices and neglects those users which experience a worse channel status. The fairness of this approach will be low especially in a network with various user statuses.

Round-Robin is an approach in which the scheduling algorithm handles the allocation in a circular way among the users without any specific priority. This approach does not neglect any user and has a fair performance but its efficiency depends on the average status of the users. This approach is simple for implementation but it is not ideal if a certain trade-off between efficiency and fairness has to be met.

MaxMin throughput is a fairness-oriented approach which aims to implement a fair allocation among the users regardless of the amount of the resources which need to be spent on each user. This means that more resources will be allocated to the users with low efficiency for making trade-offs with those with high efficiency. Allocating more resource to the users with poor channel status reduces the overall efficiency and capacity of the system but the fairness of the system will be very high.

A *proportional fair scheduling* policy makes an alternative to all the above algorithms by considering the history of the allocation to each user as well as its efficiency. A user with less allocated resources during the previous scheduling rounds will be prioritized in the current round and its efficiency for utilising the allocated resource is also taken into account. So, an efficient user with high capability for utilizing the allocated resource has priority as long as it is not consuming the resources excessively. An inefficient user, in contrast, does not have a priority unless too few resources have been allocated to it. This approach clearly makes a desirable balance between efficiency and fairness and has been adopted for different context including LTE network.

This subject will be discussed more analytically in the following chapters and a quality driven approach with more flexibility to create a desirable trade-off between efficiency and fairness will be proposed.

3.2.3 Channel Characteristics

The capacity of the system and the efficiency of the allocated bandwidth is a function of the status of the channel represented by the experienced SINR (Signal-to-Interference-and-Noise-Ratio) in the UE or eNodeB receiver. The Shannon channel capacity formula expresses the upper bound limit of this capacity as $C = B \times log_2\left(1 + \frac{S}{I+N}\right)$ in which *C*, *B*, *S*, *I* and *N* are the capacity, bandwidth, signal power, interference power and noise power, respectively. As it has been explained in the previous section, link adaptation is a function that matches the most suitable modulation scheme to each subcarrier based on the status of the channel at that frequency. This is actually an approach to achieve a closer result to the Shannon upper bound limit. A more optimal resource allocation achieves a result closer to this upper bound limit.

Table 3.3 Channel model as a power-delay profile for ITU-R-VehA scenario [65]

Tap No.	Relative delay (ns)	Average Power (dB)
1	0	0
2	310	-1
3	710	-9
4	1090	-10
5	1730	-15
6	2510	-20

There are different types of delay-power profiles which characterize the channel model based on the environment and users' specifications. These models are actually the channel impulse response modelled using a tapped delay line implementation based on the Doppler spectrum. Table 3.3 shows an example of delay-power profile for a vehicular user in a delay constrained and connection oriented environment (known as VehA), based on the ITU-R recommendation [65]. An optimal solution of resource allocation must reflect the adaptability to the status of the channel based on its profile. In the actual LTE device this is achieved indirectly through the feedback of the user about its channel status, i.e. mainly the SINR value. It is assumed that the experienced SINR which triggers a CQI index to be reported to eNodeB covers the effects of path loss, multipath fading and shadowing effects in the channel and mobility of the user. However in the LTE simulator the exact profile of the desired environment, similar to the one in Table 3.3, must be defined to make the channel modelling and estimation feasible.

The air interface carrier is defined based on the operating frequency band and the channel bandwidth. As an example, Table 3.4 shows the frequency bands defined for FDD. Figure 3.4 depicts the relation between a Resource Block in frequency domain (i.e. each RB as in Figure 3.2), channel

Table 3.4 E-UTRA operating bands for FDD [58]

E UTRA operating band	Uplink (UL) operating band			Downlink (DL) operating band		
	F_{UL_low} – F_{UL_high}			F_{DL_low} – F_{DL_high}		
1	1920 MHz	–	1980 MHz	2110 MHz	–	2170 MHz
2	1850 MHz	–	1910 MHz	1930 MHz	–	1990 MHz
3	1710 MHz	–	1785 MHz	1805 MHz	–	1880 MHz
4	1710 MHz	–	1755 MHz	2110 MHz	–	2155 MHz
5	824 MHz	–	849 MHz	869 MHz	–	894MHz
6 (NOTE 1)	830 MHz	–	840 MHz	875 MHz	–	885 MHz
7	2500 MHz	–	2570 MHz	2620 MHz	–	2690 MHz
8	880 MHz	–	915 MHz	925 MHz	–	960 MHz
9	1749.9 MHz	–	1784.9 MHz	1844.9 MHz	–	1879.9 MHz
10	1710 MHz	–	1770 MHz	2110 MHz	–	2170 MHz
11	1427.9 MHz	–	1447.9 MHz	1475.9 MHz	–	1495.9 MHz
12	698 MHz	–	716 MHz	728 MHz	–	746 MHz
13	777 MHz	–	787 MHz	746 MHz	–	756 MHz
14	788 MHz	–	798 MHz	758 MHz	–	768 MHz
15	Reserved			Reserved		
16	Reserved			Reserved		
17	704 MHz	–	716 MHz	734 MHz	–	746 MHz
18	815 MHz	–	830 MHz	860 MHz	–	875 MHz
19	830 MHz	–	845 MHz	875 MHz	–	890 MHz
20	832 MHz	–	862 MHz	791 MHz	–	821 MHz
21	1447.9 MHz	–	1462.9 MHz	1495.9 MHz	–	1510.9 MHz
22	3410 MHz	–	3490 MHz	3510 MHz	–	3590 MHz
23	2000 MHz	–	2020 MHz	2180 MHz	–	2200 MHz
24	1626.5 MHz	–	1660.5 MHz	1525 MHz	–	1559 MHz
25	1850 MHz	–	1915 MHz	1930 MHz	–	1995 MHz
26	814 MHz	–	849 MHz	859 MHz	–	894 MHz
27	807 MHz	–	824 MHz	852 MHz	–	869 MHz
28	703 MHz	–	748 MHz	758 MHz	–	803 MHz
29	N/A			717 MHz	–	728 MHz
30	2305 MHz	–	2315 MHz	2350 MHz	–	2360 MHz
31	452.5 MHz	–	457.5 MHz	462.5 MHz	–	467.5 MHz
32	N/A			1452 MHz	–	1496 MHz
Note 1: Band 6 is not applicable.						

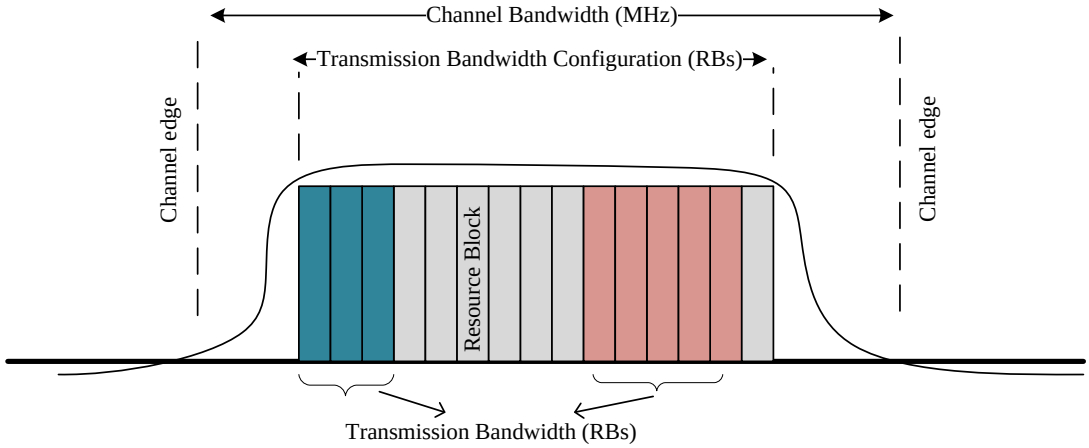


Figure 3.4 Definition of Channel Bandwidth and Transmission Bandwidth Configuration per carrier

bandwidth and the transmission bandwidth in LTE. Depicted definitions in Figure 3.4 are based on the following concepts which characterize the air interface of an LTE system in the frequency domain [58]:

- Carrier: The modulated waveform which is responsible for physical channels' transmission
- Operating band: A frequency range in which system operates and it is defined with a specific set of technical requirements. Table 3.4 provides an example extracted from LTE specifications.
- Transmission bandwidth: the bandwidth of each specific transmission between User Equipment (UE) and Base Station (eNodeB) which is measured in Resource Block units.
- Channel bandwidth: The radio frequency bandwidth to support a carrier with the transmission bandwidth configured in the uplink or downlink of a cell. The channel bandwidth is measured in MHz.
- Transmission bandwidth configuration: The highest transmission bandwidth allowed for uplink or downlink in a given channel bandwidth, measured in Resource Block units.
- Channel edge: The lowest and the highest frequency of the carrier, separated by the channel bandwidth.

3.2.4 User Equipment (UE)

In LTE, from a resource management's point of view, the user is under the control of eNodeB. However, the most influential contribution of the UE in the scheduling and link adaptation procedures is the Channel Quality Indicator (CQI) report. As it has been shown in Figure 3.3, CQI is the result of the channel evaluation in UE and contributes in the channel adaptation and resource allocation upon receiving at the eNodeB. The value of the CQI index is based on the experienced

SINR in UE and will be used to choose the suitable transmission scheme for transmission from eNodeB to UE (i.e. downlink). Table 3.5 depicts an example of the mapping between the experienced channel status at the user-side and the suggested CQI index and its corresponding modulation scheme. Code rate in column 4 of Table 3.5 represents the rate of the useful data in each symbol after taking out the parity bits. Based on this fact, the value of the CQI plays a central role in a function which expresses the capacity of the chosen transmission structure for the allocated resources in the optimisation plan.

3.3 Data Transport Procedure in LTE

The implementation of the MAC and Physical layers in the new generations of the mobile communication systems requires a combination of various complementary techniques. This combination provides the error correction capability with faster reaction to error and lower latency as well as the band width efficiency and a better performance of the receivers. These are accompanied by an efficient rate adaptation with regard to the changes in the channel status or the capability of the user's device. In the rest of this section the procedure which provides these capabilities at the transmitter and the receiver of LTE will be discussed. Since the functions which have been employed as the error control mechanism in LTE are at the core of this process, they need to be explained in advance.

3.3.1 The Hybrid Error Control Functionality

The rate of the erroneous received data is the most common quality benchmark for the evaluation of a digital communication service. Usually the error rate is calculated and compared for different system architectures versus the cost of the resources (energy or bandwidth) per unit of information

Table 3.5 Link adaptation and modulation scheme

SINR	CQI	Modulation Order	Code Rate
≤ -6.934	1(*)	2	0.1523
-5.147	2	2	0.2344
-3.180	3	2	0.3770
-1.254	4	2	0.6016
0.7610	5	2	0.8770
2.700	6	2	1.1758
4.697	7	4	1.4766
6.528	8	4	1.9141
8.576	9	4	2.4063
10.37	10	6	2.7305
12.30	11	6	3.3223
14.18	12	6	3.9023
15.89	13	6	4.5234
17.82	14	6	5.1152
≥ 19.83	15	6	5.5547

*** Users with SINR lower than a device related threshold will not be scheduled for using the resources**

delivered. During the last few decades of the development of the digital communication systems, the evaluation techniques and error control mechanisms have also been developed to guarantee a minimum required accuracy of the delivered information. This is more crucial in wireless and mobile communication systems in which the transmission signal experiences instantaneous changes in the condition of the communication medium (i.e. air interface).

The techniques for error control are normally classified into error detection and error correction methods. The type of the reaction to an error or loss in a digital communication system depends on the effect of that loss at the user side. This is based on the tolerance of the system and the minimum required quality of the service. The error could be left uncorrected as long as the quality is high enough to maintain the satisfactory communication session. It also can be recovered by dedicating a bandwidth to keep sending the lost data until the receiver confirms the successful reception of the lost data. More sophisticated methods may be used to detect and correct the erroneous data at the receiver. Automatic Repeat Request (ARQ) and Forward Error Correction (FEC) are two alternative solutions for the error control approaches in contemporary communication systems.

An error control mechanism known as Hybrid-ARQ (or HARQ), is the employed technique in the new generations of the mobile communication systems and based on the 3GPP standards. ARQ and FEC accompanied by an extra error detection method (i.e. CRC) constitute the HARQ technique in 4G-LTE [66-68]. The hybrid detection, correction and retransmission policy can exploit the advantages of each technique and complement their downside effects to achieve a more efficient and adaptive control. In the rest of this section this approach will be explained in more detail. The similarities and the differences between the HARQ implementations in LTE and its predecessors (e.g. HSPA) will also be discussed later in this chapter.

3.3.1.1 Automatic Repeat Request (ARQ)

Automatic Repeat Request (ARQ) is basically a combination of the detection of the error or loss at the receiving side and the retransmission of the erroneous data if required. In this error control method an error detection mechanism, such as CRC, helps the receiving side to realize the occurrence of the error. Receiver will inform the sender either the occurrence of error (namely NAK message or negative acknowledgment) or the correctness of received data (ACK message or positive acknowledgment) and ask for the proper reaction of sender in each case. Sender will consider the other aspects of its control policy and either retry the lost data or leave the system to deal with the event as a residual error [69].

There are a few features of the ARQ mechanism in which the synchronization of a packet stream and the reordering issue in case of retransmission events have been dealt with in different ways. Figure 3.5(a) depicts the basic idea of Stop-and-Wait-ARQ method which has been adopted in 3G and LTE as one of the components of HARQ [70]. In this method the sender will retry the

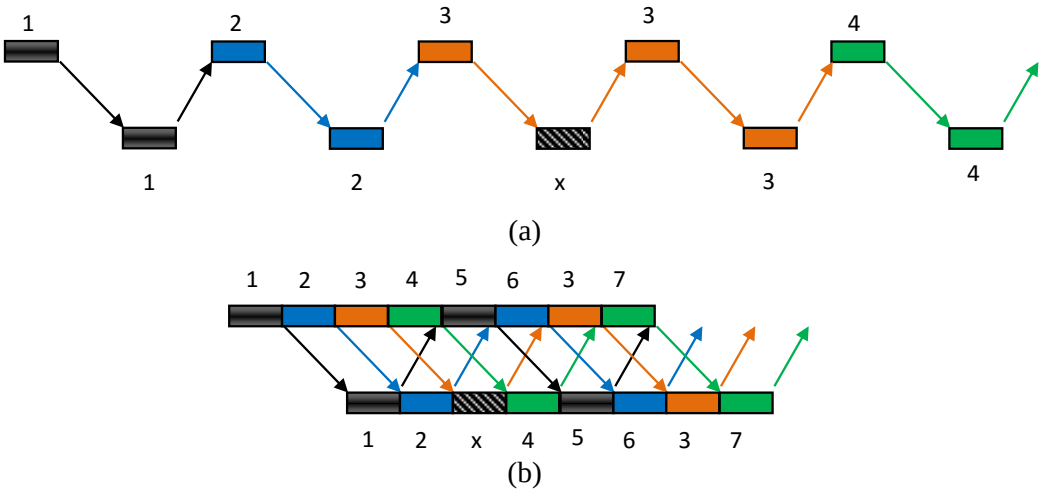


Figure 3.5 Single and multiple Stop-and-Wait-ARQ process

transmission of the lost data until either the receiver confirms the correctness of the received data or reaches the limitation on the number of continuous retransmissions. Then the sender attempts the transmission of new data.

Stop and Wait has a simple algorithm and is easy to implement. There is no reordering issue in this method and it needs a buffer just to hold the latest transmitted block of data (Transport Block) until it decides to send the next block. In spite of the simplicity and low buffer requirement, stop and wait ARQ impose a high delay to the service due to the idle round trip time (RTT) gap between sending the data and receiving the acknowledgment. For the ARQ which has been implemented as a part of HARQ function, it has been decided to manage more than one Stop and Wait process simultaneously in the same channel. Figure 3.5(b) shows this multiple process HARQ and re ordering issue in case of retransmission. The employed multiple process mechanism fills the waiting gap and reduces the overall delay of the system. Obviously the multiple processes in ARQ of HARQ violate the simplicity of the basic ARQ and add the complexity of timing and reordering whenever a retransmission is required.

3.3.1.2 Forward Error Correction (FEC)

Forward Error Correction (FEC) refers to an error detection and correction mechanism employed at the core of the Hybrid-ARQ in LTE. The FEC's correction entity is located at the receiver side. Sender transmits a combination of the data and some parity bits which have been derived from this data. Parity bits are generated through a specific time interleaving and convolutional process. The receiver uses these parity bits to detect and then to correct the error without asking for retransmission. Some specific state of the art FEC techniques are suggested for the implementations of HARQ in 3G and LTE as a part of their HARQ functionality. Table 3.6 shows an example of these FEC techniques

and their selected parameters for channel coding in the case of FDD implementation in 3GPP standards [71]. As it can be seen from the table the suggested coding for normal traffic (known as DCH in LTE and HSPA) is a Turbo code with 1/3 coding rate. This is a low coding rate, aiming to achieve robustness against error but using more than 66 percent of the bandwidth just for parities. To make this system more efficient while keeping the code rate low, a puncturing mechanism has been provided to select and send a smaller number of parities in the initial transmission attempt and send the rest of the parities in next retransmissions if necessary. In this way it will create a higher code rate if there is no need for retransmission and a lower code rate if retransmission is used. Consequently, this approach helps to improve the throughput of the system and achieve a lower latency.

3.3.1.3 Puncturing and Soft Combining

As it has been explained above, FEC adds a set of parity bits to the main data (also known as systematic bits) to make the error detection/correction feasible in the receiver. The amount of added parity bits varies based on the type of the encoder in the FEC and its code rate. However, a typical low coding rate FEC usually adds a considerable amount of parities to the systematic bits. The final rate of the transport block includes the systematic bits and the generated parity bits. This is not necessarily equal to the capacity provided by its available transport channel. Subsequently, it requires a regulation to match the output of the encoder with the available transport channel while keeping the coding rate of the initial transmission as high as possible. This is a kind of rate matching which can be achieved either through the setting of the code rate or by puncturing and prioritizing different sections of the output of the encoder to be sent subsequently during the retransmissions. The subsequent sets of the retransmitted data will be buffered and combined in the receiver to achieve a better accumulated signal to noise ratio and to increase the probability of the successful decoding in the FEC [70]. Figure 3.6 shows an example of buffer and combine mechanism in a ‘soft combining Hybrid-ARQ’ with punctured data. This example shows the first and the second transmission attempts failed and third one succeeded to decode the information without error. Each received retransmitted data will be combined by the buffered data from the previous failed transmission. This will add to the parity bits or accumulate the energy of the repeated bits and improve the performance of the decoder.

An encoder with a low code rate has an output with a high amount of parity bits compared to the systematic bits which create low bandwidth efficiency. However, as a self-decodable data it is more

Table 3.6 Channel encoders and their main parameter

Channel type	Coding type	Code rate
Broadcast (BCH)	Convolutional	1/2
Normal Traffic (DCH)	Turbo	1/3

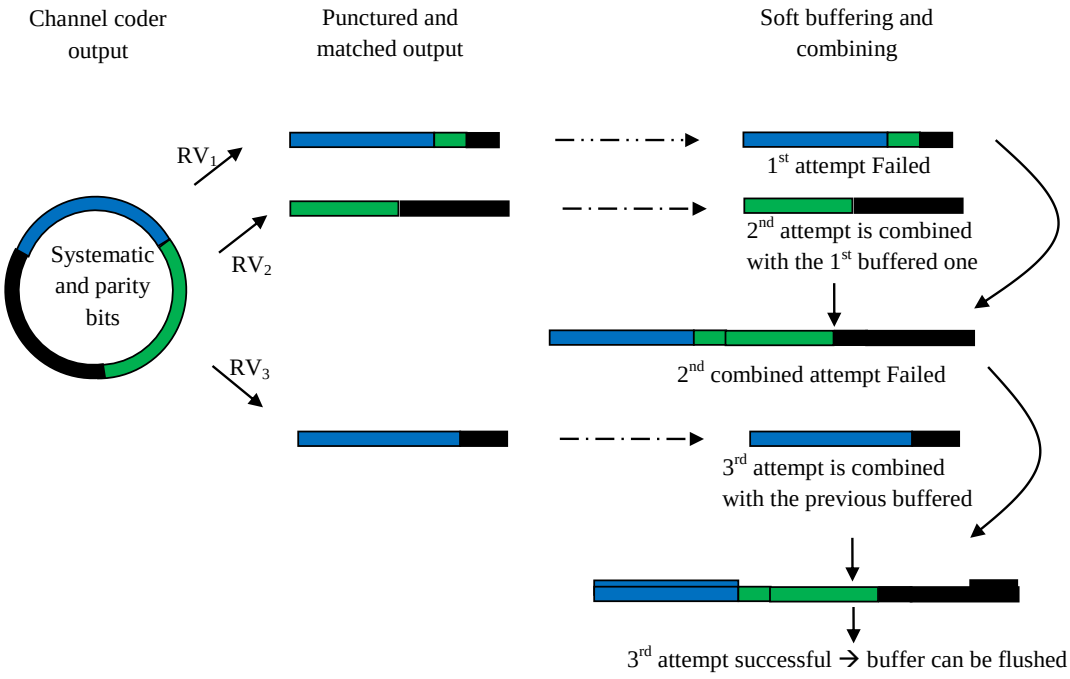


Figure 3.6 Different set of punctured data and their soft combing in the receiver

capable of the error correction without need for retransmission. The higher code rate of the first transmission process of the above explained rate matching will eliminate this robust error correction for the sake of the bandwidth efficiency. Furthermore it needs a kind of mechanism to be aware of the situation in which the main data (e.g. systematic bits in Turbo code) should be resent more than once. It also keeps the set of data self-decodable in the receiver. All these features in LTE and HSPA are planned through a Redundancy Version (RV) parameter. RV defines the combination of the data to be sent and synchronizes the subsequent transmission attempts based on the puncturing mechanism. The relation between the initial transmitted set of data and the subsequent retransmitted data defines the type of the employed soft combining HARQ. The following are the three implemented types of HARQ combining in 3G and LTE [70]:

Chase Combining (CC): It is used when the content of retransmissions are the same as the initial attempts. In this case there is no change in the coding rate and no code gain will be achieved. However, it results in the accumulation of the energy per bit and the improvement of the signal to noise ratio (i.e. the accumulated received E_b/N_0 for each retransmission).

Incremental Redundancy (IR): In this method the content of the initial transmission includes main data (e.g. the systematic bits), and retransmissions include different combinations of the systematic bits and parity bits. More systematic bits in the initial attempts results in a higher code rate when a lower coding gain is required for good receiving status. More retries in a bad receiving

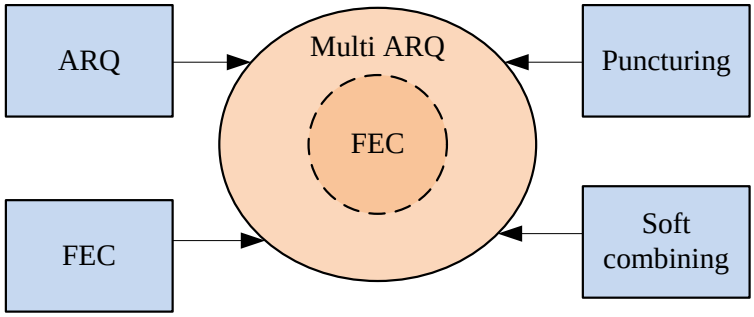


Figure 3.7 Combined properties of HARQ in 3G and LTE

condition adds more parity bits and decreases the code rate while increases the chance of the successful correction. Since the total number of data is limited, more retries provide a more chance for repeating each bit and boosting the energy of the repeated bit. This resembles the advantage of the Chase Combining (CC) method for repeated bits.

Partial IR: Occasionally, the redundancy control mechanism repeats the systematic bits in all retries while they are accompanied by different sets of the parity bits. Subsequently, for the systematic bits partial IR acts as a CC method and overall process performs similar to the IR method. This method can quickly provide a power gain for systematic bits and gradually reduce the code rate to improve the coding gain.

The current implementation of HARQ in 3G and LTE combines all above mentioned properties together as it has been shown in Figure 3.7. This combination simultaneously provides the error correction capability of a low code rate FEC, the band width efficiency of a high code rate FEC, and less delay with faster reaction to error.

3.3.2 Transmission-Side Procedure

Figure 3.8 shows a typical functional block diagram of an implemented transmission procedure in which all above mentioned components along with the implementation considerations and limitations in sender have been taken into account [67, 70]. In the sending side, first data will be coded with a low code rate (e.g. Turbo code, 1/3 code rate) then if there is a buffer constraint in the receiving side (User Equipment in downlink) data will be punctured to fit the combining buffer. Actually the capability of user equipment (UE) from buffer size point of view imposes a limitation on the maximum feasible rate of the corresponding sender in the network. The first rate matching function enforces this limitation and shape the rate of the output of the channel encoder. It must be noted that the first rate matching stage will be transparent if the receiving side is eNodeB or a UE with enough combining buffer. The second rate matching is responsible for puncturing the data to fit the available transport channel. It means that the available transport channel is the second limitation

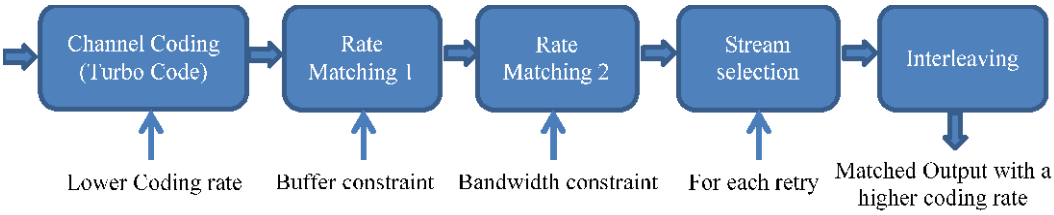


Figure 3.8 Functional block diagram (transmission side)

which reshapes the output of the channel encoder and probably reduces the final data rate of the service.

The output of the encoder includes a set of systematic bits which must be sent in the first transmission and two separate sets of parity bits which can be punctured and prioritized to be sent subsequently if retransmission is required. Stream selector will impose the policy and the order of the selection based on the retransmission policy. Actually the stream selector defines the type of the combiner by selecting the set of data which must follow the initial transmission in each retransmission. The degree of the similarity between the retransmitted content and the initial transmission defines the combining type. The combining type can be Chase Combining, Incremental Redundancy or partial IR as has been explained in the previous section. Redundancy Version (RV) is the parameter which set the selector switches and it will be explained in the next section.

Finally an interleaving stage (final bit collection) will map the bits of the matched data sets in a specific order into the transmission channel based on the types of the bits and transport format e.g. it defines the location of systematic and parity bits based on the symbol size of the modulation. Figure 3.9 shows an example of interleave for the systematic bits and two types of parity bits for 16QAM modulation in FDD [67]. Assume that the interleaver is rectangular with $N_{row} \times N_{col}$ cells. The number of rows for 16QAM in this example is $N_{row}=4$. The number of columns for N_{data} bits of information will be $\frac{N_{data}}{N_{row}}$. For systematic bits, N_{syst}, consider two intermediate values: $N_r = [\frac{N_{syst}}{N_{col}}]$ and $N_c = N_{syst} - N_r . N_{col}$. $N_c=0$ shows the divisibility of N_{syst} by N_{col} and $[.]$ is the round function toward the lower integer. Based on the value of N_c, cells in the rows 1 to N_r or N_{r+1} will be written by systematic bits and the rest of the cells will be filled by parities. Reading out and writing to the transport channel will be column by column.

It must be noted that there are some other internal interleaver which have their own specific mapping rules in the encoder. For example in LTE the mapping between input to the Turbo code internal interleaver, C_i, and its output, C_i', has been defined by the following relation: $C_i' = C_{p(i)}$ and $P(i) = (f_1 \times i + f_2 \times i^2)\ mod\ K$. The values of f_1 and f_2 for each i are given in a table

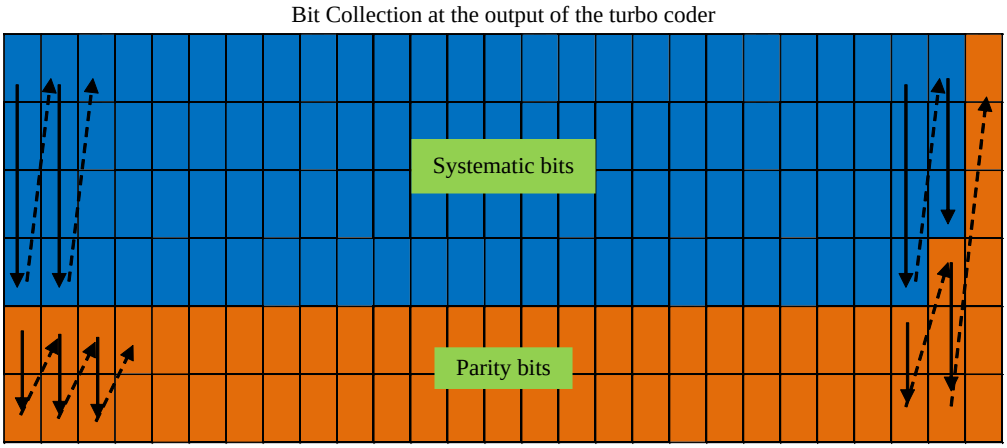


Figure 3.9 Functional block diagram (transmission side)

based on the value of *K* which is the transport block size [67]. So for each transport block size the mapping will be customized properly.

Although making a kind of diversity in time and frequency domains for each type of bits is the main reason of interleaving, different interleaving rules and the above mentioned relation between different types of bits is due to the inequality of the importance of the bits with regards to their carried information. For example, the systematic bits compared to the parities or the bits of each symbol which define the quarter in the constellation map, compared to the other bits of the symbol obviously do not have the same importance.

3.3.3 Receiving Side Procedure

As it has been shown in Figure 3.10, the receiving process begins with the process of the output of the physical layer's demodulator. At the beginning of each cycle of the process the transmitted data contains a new stream of bits. This self-decodable data will be passed directly to the decoder but there is no combining process. In a normal condition there is no error and the indication of CRC shows a successful decoding process. A positive acknowledgment will inform the sender and ask for

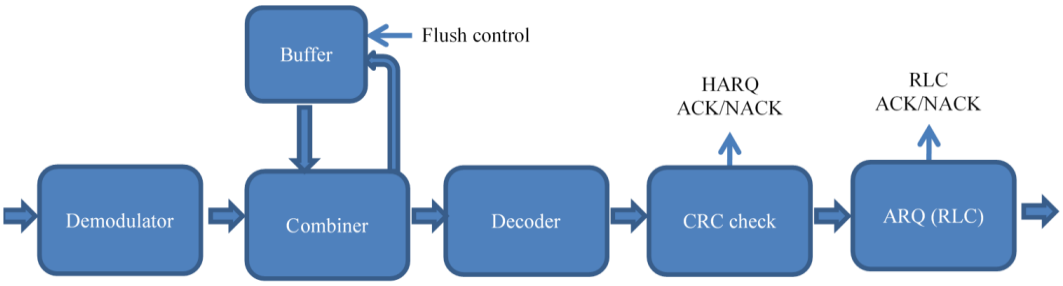


Figure 3.10 Functional block diagram (receiver side)

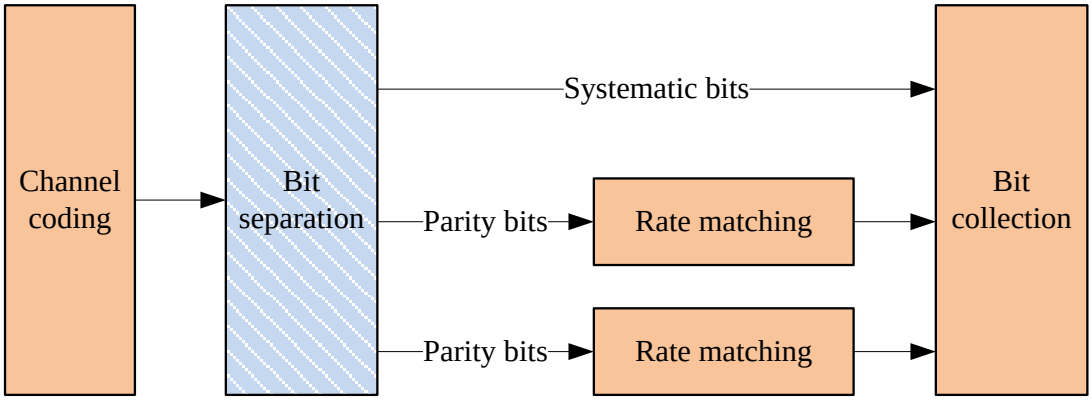


Figure 3.11 Turbo and convolutional channel coding process

a new data for that HARQ process. The next data have an indication of a new data and this indication will flush the buffer to prevent any data corruption in the combiner. In the case of error, CRC detects the error and a negative acknowledgment informs the sender and asks for retransmission. Current information remains in the buffer to be combined with next retransmission if any. The sender side retransmits the same set of data (Chase Combining) or a new set of the same data (Incremental Redundancy) and the receiver combines this with the previously stored information to improve the probability of correct decoding. This cycle will continue until the system reaches either a successful decoding or its retransmission limit. Any residual error will be dealt with in the ARQ of RLC which is a reliable correction method though it causes more latency. The effect and location of ARQ in RLC will be explained in more detail in next section.

It must be noted that the above explained functional block diagrams are applicable to downlink and uplink of LTE and its predecessor systems such as HSPA. There are some differences among the implementations of different types of systems and different architectures which customize this general diagram. For example, in the previous explanations it was assumed that the channel coding is a turbo code with three internal streams of data including one systematic and two parity bit streams. This has been shown in Figure 3.11. However, if a convolutional encoder has been used as the channel encoder (this is the case for paging and broadcasting traffic) there won't be a multiple internal stream. In this case, the redundant dashed block in Figure 3.11will be transparent and removed from the procedure. Table 3.7 shows some example of the traffics in LTE and HSPA with their choices of the Turbo code or convolutional code as the channel encoder and related code rate in each case.

The 1st stage of the rate matching in the downlink HARQ procedure is accompanied by a virtual buffer which is the mirror of the buffer capability of the User Equipment. As it is depicted in Figure 3.12, this combination enforces the first rate matching process. In contrast, no limitation needs to be

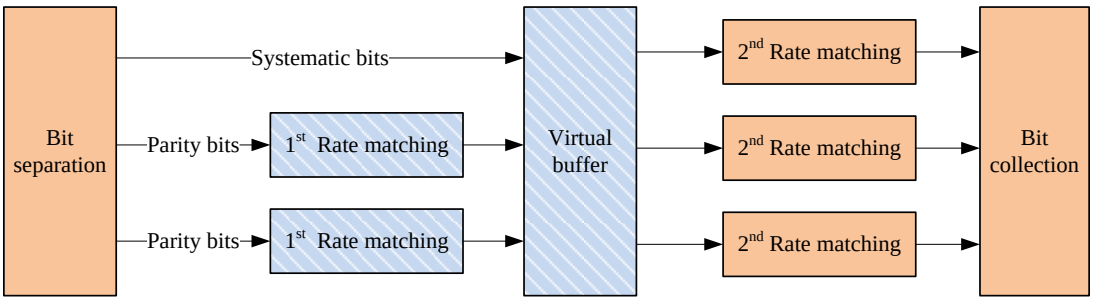


Figure 3.12 A virtual buffer represents the buffer capability of the UE in the downlink HARQ

imposed from user equipment toward the network as the first stage of the rate matching. Dashed blocks in Figure 3.12 depict the redundant and transparent processes in uplink direction in this case.

3.4 Implementation Considerations in LTE

The functionality and general principles which have been discussed in the previous section are the bases for the implementations of the data transport process in LTE. The sender attempts the first transmission with all necessary systematic information plus as many parity bits as required to fill the remaining dedicated channel space and makes a self-decodable set of data. Receiver decodes the received data and detects any error through the accompanied CRC code. In case of error it will inform the sender and ask for retransmission through the feedback signal (acknowledgment). When sender decides to attempt a retransmission, the type of HARQ will be chosen as CC (Chase Combining) or IR (Incremental Redundancy) by sending the same set of data as the initial transmission or a new combination of data, respectively. Redundancy Version (RV) in the sender will define this combination. Soft combining in CC provides a power gain through the accumulation of the retransmitted bits while the IR method provides a coding gain through the effort to send as few parity bits as possible.

However, there are some differences and specific considerations in the implementation of the above mentioned process for downlink compared to uplink and for FDD architecture compared to the TDD. LTE has also its own implementation characteristics which is different from the predecessor systems such as HSPA. This section is dedicated to the explanation of these specific

Table 3.7 Comparison between channel encoders' code rate in LTE and HSPA

System type	Traffic type	Channel coder	Code rate
HSPA	Broadcast traffic (BCH)	Convolutional	1/2
	Normal traffic (DCH)	Turbo	1/3
LTE	Broadcast traffic (BCH)	Tail biting Convolutional	1/3
	Normal traffic (SCH)	Turbo	1/3

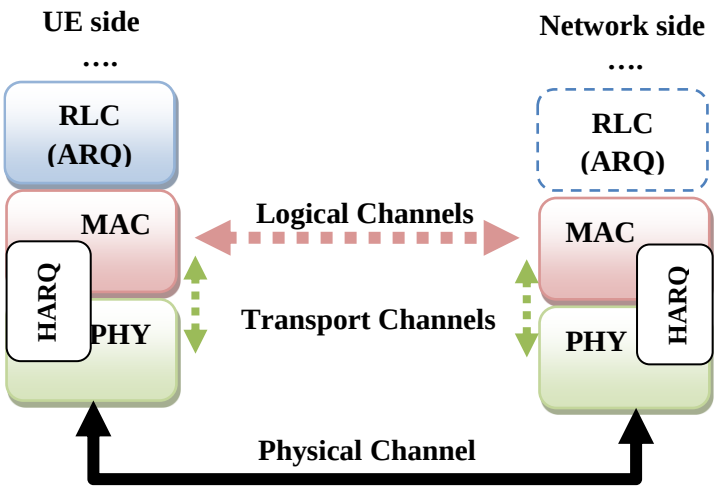


Figure 3.13 Layered channels of a data transport procedure

implementation considerations and the clarification of some important characteristics of the resource utilisation process in LTE.

3.4.1 Layered Protocol Stack and Signalling

The implementation of the user and control planes in LTE (i.e. the main traffic and the control communication between the user-side, UE, and the network interface, eNodeB) is based on a layered architecture which resembles the general protocol stack of the computer networking. Figure 3.13 shows the layers related to the data transport procedure in both user and network sides. Although the coordination and signalling of the channel occupancy originate from MAC layer, the actual channel coding, rate matching and combining which have been explained in the previous sections, are happening in the physical layer. Since there could be more than one bearer and process related to each physical channel, a channel mapping between logical channels in higher layers and their physical channel in lower layer is defined which is not necessarily one-to-one [67]. Figure 3.14 shows a typical relation between logical and transport channels which represent the idea of multiplexing and mapping the bearers between MAC and PHY layers. The signals related to any specific function such as HARQ goes through the physical channels while it has a dedicated location in the frame, a plan of allocation and mapping and its associated timing.

The overall structure of the layer stack for these functions is almost the same for Uplink and downlink in LTE or HSPA except for the role of RLC which will be explained later in this section. The differences are mainly related to the content of the information and the detail of their algorithms. For example, Figure 3.15 shows the information associated with HARQ in both LTE and HSPA and their related channels. Each HARQ process must be aware of the transmission set of data accompanied by its relevant control information including acknowledgment, processing number,

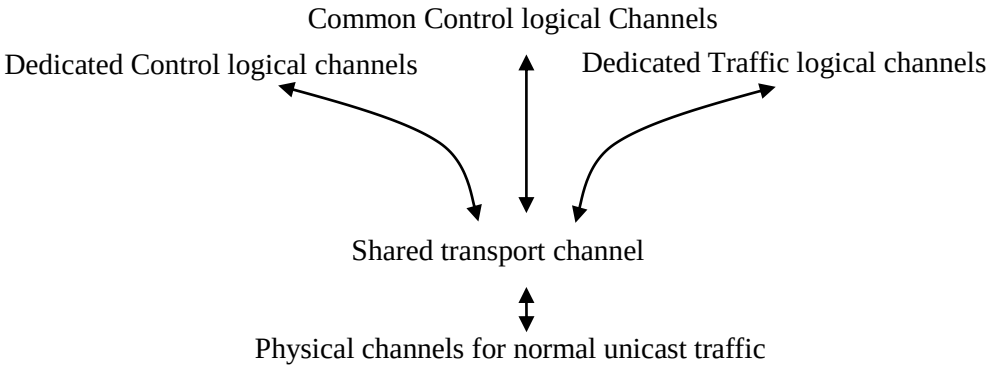


Figure 3.14 A typical relation between logical and transport channels

New Data Indicator, Redundancy Version (RV) and Retransmission Sequence Number (RSN). The channelization of this two-way communication varies based on the type of the system (HSPA or LTE) and for downlink and uplink. As it has been depicted in Figure 3.15, the main data stream and acknowledgment are the common signalling in all cases while the processing number has not been informed from UE toward the network interface in uplink. This is because of the synchronized structure of uplink process which provides implicit information about HARQ processing sequence in the base station.

Furthermore, the RSN which is a counter of number of consecutive retransmission has not been sent in downlink direction. RSN can be used by each of the multiple NodeBs connected to the same UE in soft handover situation. This helps them to recognize the provided correct data by other NodeBs when they have lost an intermediate block and its new data indicator. The value of RSN affects retransmission policy as well. Table 3.8 shows the effect of the value of RSN over the puncturing procedure and HARQ combining type. It shows that in case of RSN=0 (initial retransmission) the redundancy version enforces systematic bits into the transmission set of data. In

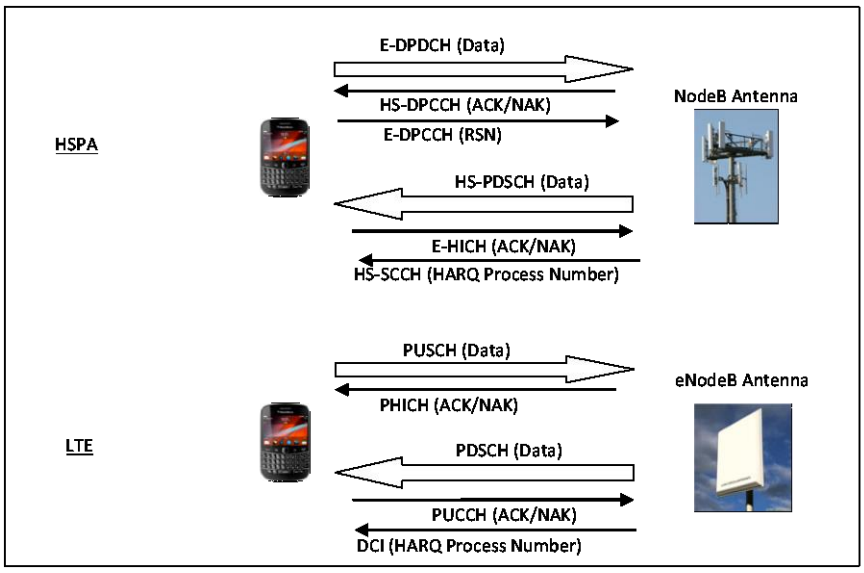


Figure 3.15 The information associated with HARQ in uplink and downlink

the case of more than one failed retransmission for the same data (RSN=2 in the table), based on the code rate either initial set of data or a new combination which includes both systematic and parity bits will be repeated [70]. Since the handover procedure has changed in LTE, as it has been shown in Figure 3.15 this overhead is no longer a part of the transmitted information in LTE.

3.4.2 Number of Parallel Process and Timing

In the basic 'Stop-and-Wait ARQ' mechanism and due to the waiting time to receive the acknowledgment, there will be a considerable delay in each processing cycle from first transmission attempt until next transmission/retransmission. To maintain the continuity of communication in a very high rate system (e.g. HSPA/LTE) and to make HARQ as fast as possible, the function of HARQ in LTE and HSPA is able to support more than one process in parallel and manage to react to their acknowledgment reports independently. This improves the overall throughput of the system while still has a low cost of overhead for acknowledgment and control.

However, a multiple process HARQ reduces the simplicity of ARQ timing. It needs to be customized for different situations and for different transmission schemes. Actually the optimum number of HARQ processes is a function of the required processing time for each HARQ process, the TTI duration and the propagation time between sender and the receiver. Consequently, the optimum number of HARQ processes will be different for LTE and HSPA. A smaller number of the parallel processes increase the delay while more parallel process increase the complexity of the system.

As a trade-off between latency and complexity, 6 processes in HSPA and 8 processes in LTE are typical configured values in FDD scheme. Since the allocation, timing and synchronization of TDD implementation is different from FDD, this number will be different in case of TDD and depends on the allocation plan in downlink and uplink.

Since each transport block will be transmitted in one TTI, initial transmission and retransmissions will not be in the same subframe and the number of subframes can be used as a base for HARQ retransmission timing. For example in FDD scheme, the acknowledgment of the HARQ process in subframe 'n' will be received at the time of subframe 'n+4' and can be retried, if necessary, in

Table 3.8 Relation between RSN and soft combining policy

RSN	Code rate (compare to 0.5)	RV
0 (first attempt)	0.3 (<0.5)	0 (mostly systematic bits)
	0.6 (>0.5)	0 (mostly systematic bits)
1 (second attempt)	0.3 (<0.5)	2 (systematic and parity bits)
	0.6 (>0.5)	3 (only parity bits)
2 (third attempt)	0.3 (<0.5)	0 (mostly systematic bits)
	0.6 (>0.5)	2 (systematic and parity bits)

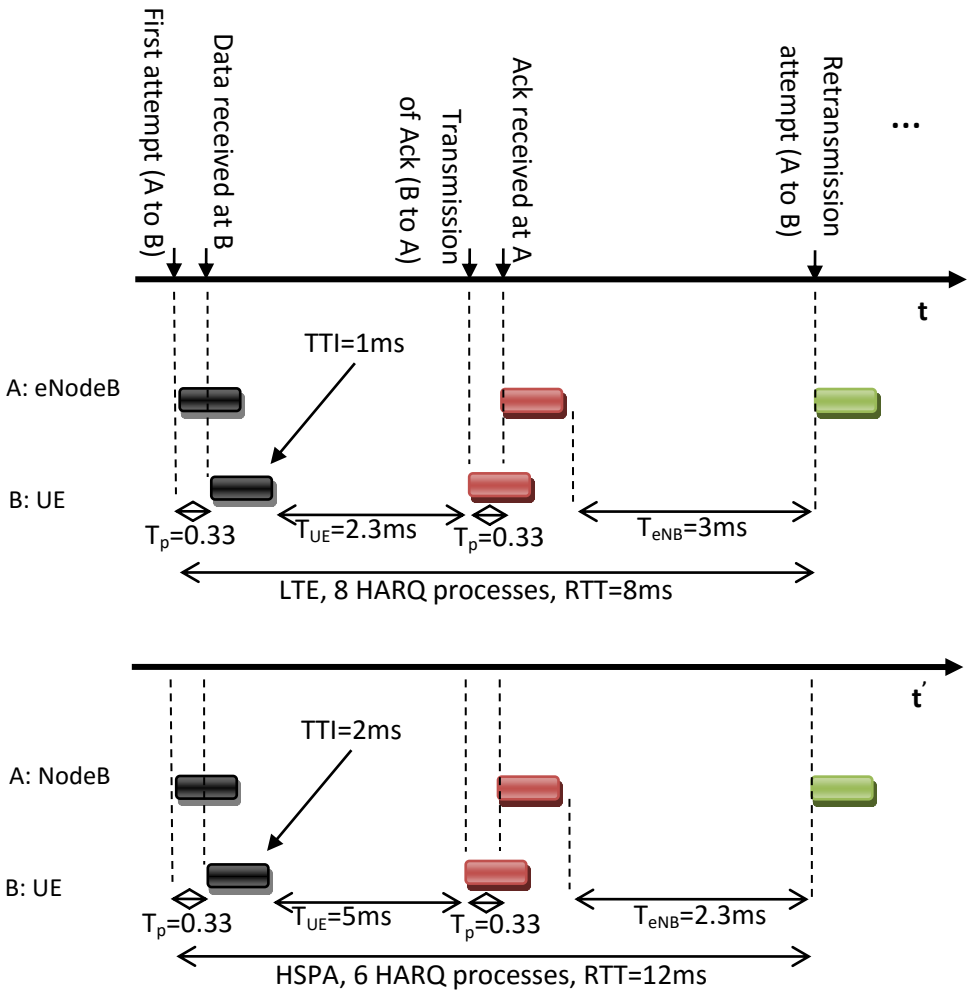


Figure 3.16 An example of the timing of HARQ transmission and retries

subframe 'n+8'. It means that the RTT of HARQ process in this example is almost fix and equals to 8 subframes which is 8ms. For HSPA with a longer TTI value and lower transmission rate, the RTT of HARQ will be proportionally higher unless a lower number for HARQ processes is required to be arranged.

The RTT of HARQ in TDD implementation follows a similar calculation but the exact number of subframe between initial attempt and probable retransmission (or next new data transmission) depends on the allocation plan and HARQ may experiences different RTT values.

Figure 3.16 shows an example of the timing of HARQ transmission attempts for HSPA and LTE (FDD implementation) and their typical HARQ Multi-process values between node A (e.g. eNodeB) and node B (i.e. the UE). The propagation time from A to B has been assumed to be 0.33ms (i.e. 100 km distance). For LTE the acknowledgement of subframe *i*, will be sent back in subframe *i+4*. The second attempt for the same data must be in the same HARQ process which means *i+8*. As it has been shown in Figure 3.16, the available time for processing the received subframe in the user-side, T_{UE}, and the network-side, T_{eNB}, are not the same. Considering the propagation time and the number of parallel processes in LTE as well as the transmission time interval, *TTI=1ms*, the

remaining time for process in the UE and eNodeB will be $T_{UE}=2.3ms$ and $T_{eNB}=3ms$, respectively. The received subframe at node A (B to A direction) must be synchronized with the beginning of the same number of subframe in A. Subsequently, the timing of the transmission from B to A is arranged to have a timing advance, T_{TA}, twice the propagation time ($T_{TA}=2\times0.33=0.66ms$ in this example) for the purpose of this synchronization.

For HSPA the number of HARQ processes, TTI and transmission times are different from LTE. So, the remaining processing time will be different too. An example of these values has been shown in Figure 3.16 for comparison. For 6 parallel HARQ process, *TTI=2ms* and the same 100Km distance between UE and NodeB the processing time for UE will be T_{UE}=5ms and for NodeB it will be $T_{NB}=2.3ms$. These values show the feasibility of using a device with lower computing specifications at the user-side. Furthermore, the base station in 3G-HSPA, i.e. NodeB, has less responsibility in the network compared to the eNodeB in LTE. This means that the NodeB's apparent lower processing time will be high enough for its tasks.

3.4.3 Synchronization and Adaptability of Retransmissions

Based on the architecture of the system, different strategies are applied to the timing and the format of the retransmission transport block of a HARQ process. Each transmission block is transmitted in a TTI with a specific transport format in a certain time. Transport format defines the transport block size, the number of the transport blocks (i.e. data rate), the frequency location (e.g. in FDD-LTE), the modulation and code rate schemes and so on. The timing means the relation between the time of the retransmitted subframe and the time of the received acknowledgement.

There are two options (adaptive and non-adaptive) for the transport format of the retransmission of a HARQ process compared to the initial transmission of the same process. The adaptive format allows the sender to change the format of the retransmission due to the change of the condition of the channel or the availability of the resources. The non-adaptive retransmission format does not allow any changes and enforce the same format as the initial transmission. Adaptive option provides a desirable flexibility in the frequency domain and it could be a kind of diversity for retransmitted data. However, it needs extra signalling and scheduling process. Downlink in both HSPA and LTE are adaptive so transport format can change in the case of retransmission. Uplink in HSPA and LTE are non-adaptive though uplink in LTE can be adaptive occasionally to avoid resource fragmentation or collision with random access.

The timing of the retransmission could be synchronous and scheduled in a predefined time after the initial transmission or asynchronous and flexible in time domain. Downlink in HSPA and LTE are asynchronous so retransmission can happen any time after the reception of the NACK. Uplink in HSPA and LTE are synchronous and the retransmission timing is scheduled. It must be noted that, unlike FDD, the synchronous timing in TDD doesn't mean a fix timing hop between the initial

Table 3.9 Adaptability and Synchronization of retransmission process in HSPA and LTE

System type	Direction	Adaptability	synchronization
HSPA	Uplink	Non-Adaptive	Synchronous
	Downlink	Adaptive	Asynchronous
LTE	Uplink	Non-Adaptive (can be adaptive occasionally)	Synchronous
	Downlink	Adaptive	Asynchronous

transmission and the corresponding retransmissions for all existing processes. The actual timing relation between initial transmission and the retransmissions in TDD depends on the specific applied uplink/downlink allocation in that system. Table 3.9 summarizes the above discussion.

3.4.4 Acknowledgment in Soft Handover and Buffer Flush Control

The functionality of HARQ in LTE and HSPA has been defined as a 'fast Hybrid ARQ with soft combining' mechanism which is a combination of HARQ protocol in MAC layer and redundancy control, buffering and soft combining in PHY layer. A few control messages have been defined to work together to achieve the sufficient control over this function. New Data Indicator, Acknowledgement, Redundancy Version (RV) and Retransmission Sequence Number (RSN) are the main control signalling related to the HARQ function.

Based on the mechanism of stop-and-wait ARQ, upon the receiving of the current transport block, receiver sends back a status report about the correctness of the received data. This is known as the positive acknowledgment (ACK) in the case of the correct data reception and the negative acknowledgment (NACK) otherwise. Whenever a retransmission is required, based on the acknowledgment report, the sender will use a different RV to make a new combination of the systematic and parity bits for retransmission. A New Data Indicator will be toggled to inform the receiver if the sender decides to send a new data, finishes the process of the previous transmission and releases the corresponding HARQ process. Receiver continues the combining of the received data with the previous buffered bits until it is informed about the new data reception.

Sending a one-to-one acknowledgement for each transported block of each parallel process provides a straightforward algorithm for the implementation of the system. However, due to the parallel processing of the subframes and the nonlinear timing relation between the uplink and downlink (I.e. the unbalanced duplex communications), this mechanism cannot be always one-to-one. For example, there is a one to one relation between each HARQ process and the acknowledgment report in the FDD scheme of the downlink of HSPA or LTE. In contrast, each acknowledgement message represents more than one HARQ processes in TDD (i.e. 'OR' of 'ACKs') and cannot be considered as an explicit one to one process.

Wrong interpretation of these control messages lead to a wrong action and a loss of data. For example ACK to NACK misinterpretation will enforce an unnecessary retransmission and degradation in throughput. NACK to ACK inversion can be a source of residual error and RLC involvement which leads to a higher latency. As it has been explained above, instead of the positive acknowledgment, actually the new data indicator controls the soft buffer flush while the negative acknowledgment directly controls the request for retransmission. Typical acceptable value for the probability of ACK to NACK is 10^{-2}. This value for NACK to ACK inversion which is more serious has been set between 10^{-3} and 10^{-4}. In LTE the bit of acknowledgment is repeated more than once and its carrier in physical layer (i.e. PHICH) is grouped and mapped into the multiple frequency allocations to control the rate of ACK/NACK error and make diversity in time and frequency.

The status of soft handover between different cells while the UE is at the edge of the cells is one of the challenging situations for the subframe process in uplink direction. In this case UE sends its transport block toward more than one network interface and receives multiple acknowledgments. When at least one received acknowledgment is positive (OR of the ACKs) the UE assumes a successful HARQ process and starts a new data. If one of the network interfaces which has sent a NACK to UE misses the new transport block message, it remains in the waiting state for previous block and this will corrupt its soft buffer combining in the next receiving process. For this reason in HSPA besides the ACK and the new data indicator, a two bits control message (RSN) will inform all connected NodeBs about the number of the continuous retransmission attempts for current block. This prevents them from missing the latest situation of successfully received packets and probable wrong soft combining even if they have missed intermediate new data indicator reports. As a part of the evolved features of LTE, a seamless/lossless handover through X2 interface provides the support for handover and RSN is not required in LTE.

3.4.5 Reordering and RLC Location

The fast reaction to the changes in the channel condition is one of the main advantages of using parallel and soft combined processes in LTE. This reaction is as fast as the duration of the HARQ RTT. The RTT is a function of the number of parallel processes, duration of the TTI, processing and propagation times. The result will be different in the case of HSPA and LTE given their different TTIs and data rate schemes. Furthermore, in all cases there will be a residual error which cannot be fixed through the provided fast mechanism. The residual errors propagate to the higher layers (e.g. TCP). Residual errors due to the misinterpretation of the acknowledgement and the wrong combining or wrong buffer flush timing are some examples which create residual errors. To increase the reliability of the system against residual errors, a complementary ARQ error correction has been provided in RLC. ARQ in RLC is happening less frequently compared to the HARQ and it is not as fast as HARQ process. Nevertheless, it resolves the majority of the residual errors and improves the reliability for higher layers. In addition, the result of the HARQ processes in the receiver is not

necessarily in the right order and must be reordered before going to the higher layers. This is a function between HARQ in the MAC and RLC protocol.

Since the location of the RLC layer of UE in HSPA is the same as LTE, their ARQ and reordering functions are similar in downlink direction. In contrast, unlike the eNodeB in LTE, the NodeB of HSPA doesn't embody the RLC layer at the network side for uplink reception. Actually, all of the received blocks, including the duplicated and the residual erroneous blocks must be transferred to the backbone of the NodeB to complete the process of the reordering and the ARQ. Consequently, ARQ and reordering in LTE is more efficient than HSPA and saves some of the capacity between eNodeB and core network.

3.5 LTE Simulator

A simulation tool which provides the required access to the scheduling and link adaptation functions alongside the channel model and a proper input data stream in the context of LTE is developed in this work. The simulator is capable of the modification of the related parameters from RLC, MAC and PHY layers as well as the characteristics of the video data stream in APP layer and channel fading and shadowing considerations in the air interface. These are the elements which are required for the verification of the proposed analytical models and the implementation algorithms. Furthermore, it provides access to the runtime data and the final performance parameters for analysing the behaviour of the system from the procedural and the final outcome point of views.

The required functionalities from eNodeB, User Equipment and the channel are developed using the object-oriented programming in MATLAB comply with the characteristics of the LTE system which is aimed to be investigated in this research [13]. The simulator is capable of working with our university's processing cluster and parallel processing platform to reduce the simulation time. The overall structure of the developed simulator in MATLAB and some aspects of the provided capabilities and performances will be explained in this section.

3.5.1 The Structure of the Simulator's Implementation

Figure 3.17 depicts the implemented LTE simulator structure. A traffic generator provides the input of the simulator based on the desired model of the data. For example, a specific data rate with a given packet size and inter-arrival time of the packets can define the required data slices. The resource allocation process (per Transmission Time Interval, TTI=1ms) is controlled through the scheduling policy, the bandwidth of the system (amount of the available resources blocks) as well as the status of the channel and the CQI feedback from UE. This will define the allocated rate to each user in each processing interval.

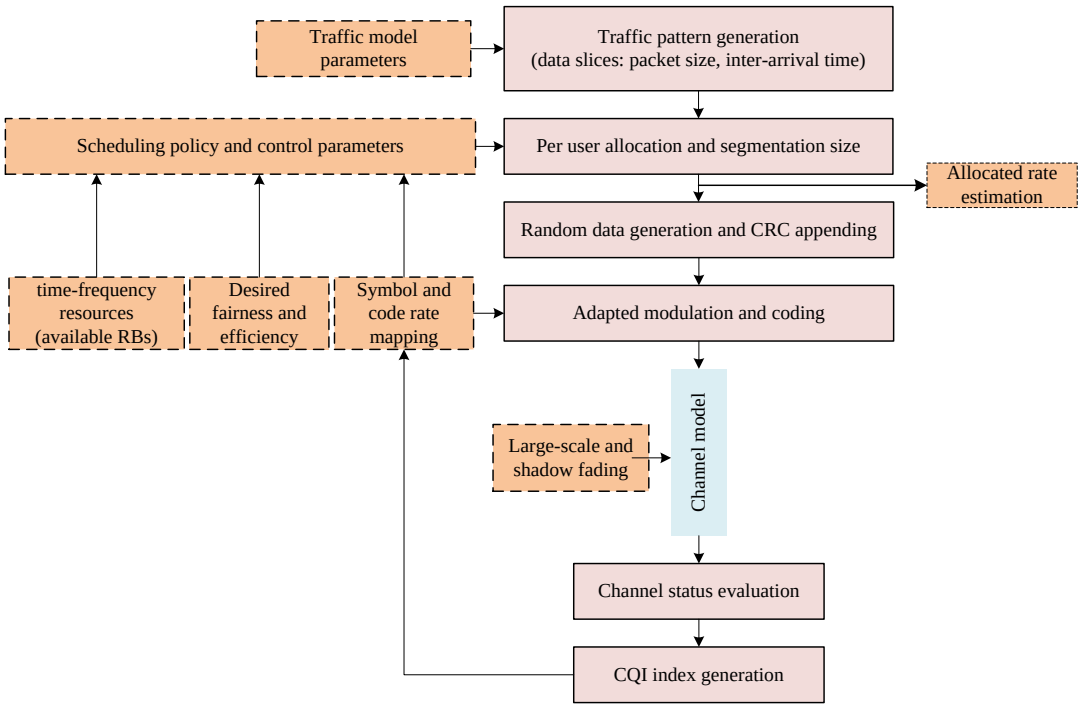


Figure 3.17 The structure of the implemented LTE simulator

Given the allocated data rate to each user and the size of the data which is supposed to be transmitted, a random data generator produces the random bit-stream which represents this data. The initial CRC overheads will be appended to each segment of the data at this stage. Modulation, channel coding and final transmission process will follow based on the adopted order of the modulation scheme and its symbol size (as it has been shown in Table 3.5). The channel model and its specifications such as the large scale and shadow fading are fully supported for one cell or more with or without the consideration of the inter-cell and intra-cell correlation. The decorrelation distance and the exact statistic of the shadow fading are definable as well.

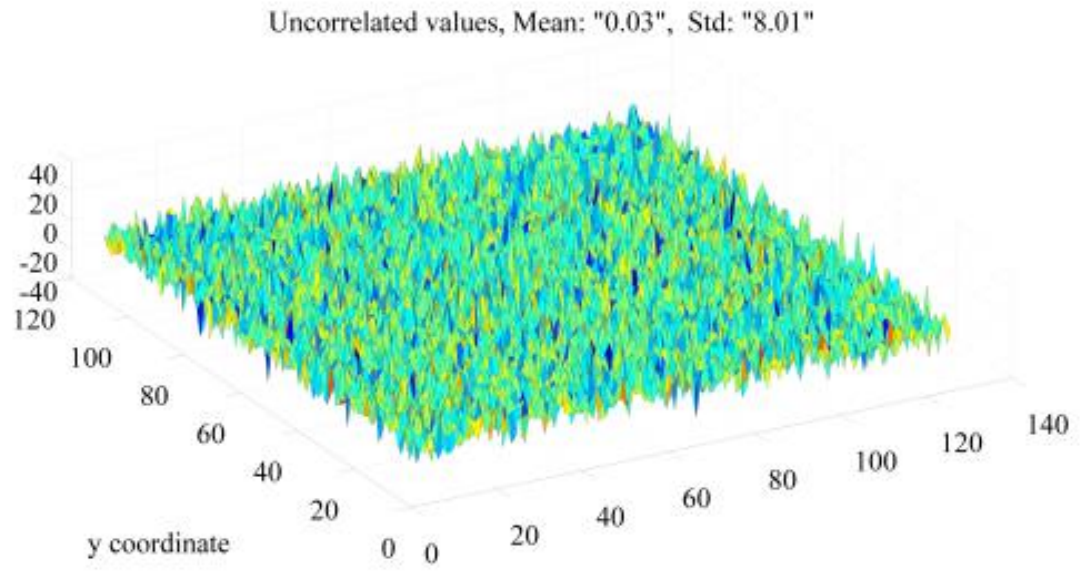
The signal strength attenuation experienced around a base station in a cell is defined based on the path loss and fading effects. These include the small scale fluctuation of the received power and the large scale shadowing effect (e.g. caused by obstacles with various types and distributions). This process is basically time-variant and location-dependant. The overall characteristics of the simulated channel model complies with the model explained in subsection 3.2.3 with the exact channel model parameters introduced in the rest of the Chapters whenever the simulator is used.

A separate program is developed to generate the spatially correlated values for shadow fading across each cell and a cluster of neighbouring cells. The implementation algorithm for this program is derived from a previously developed analytical model for spatial correlation consideration in a mobile radio system [72, 73]. Shadow fading is usually modelled as a zero mean log-normal distribution random variable with a non-zero deviation. However, the required spatial correlation

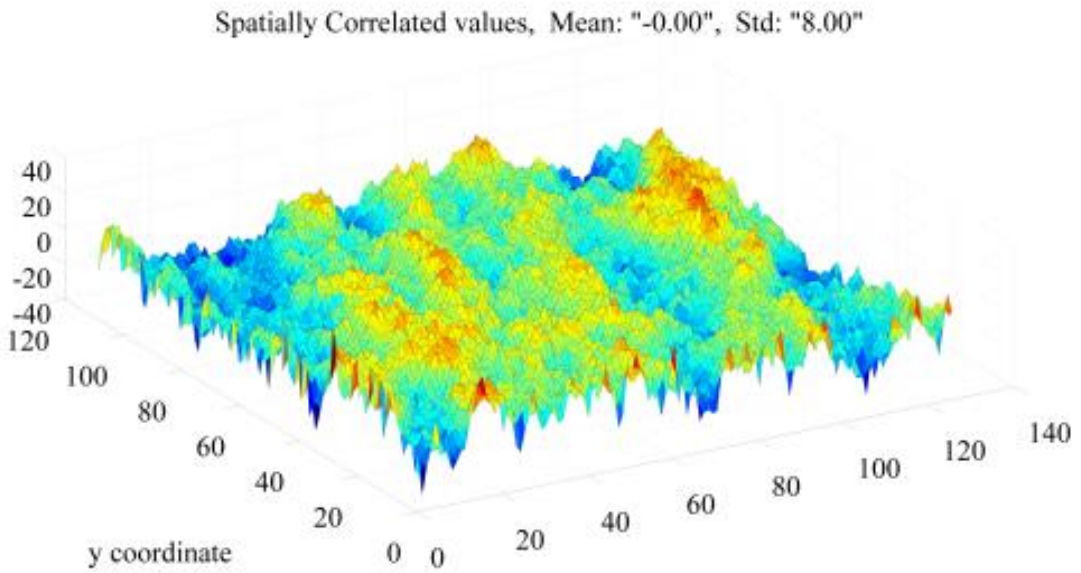
between neighbouring values in a cell need to be taken into account too. A simple exponential correlation between two points spaced by a distance x, can be defined as e^{-ax} where e^{-a} is the correlation coefficient between two points at distance $x=1$ to be empirically defined and be the base for finding the value of a in the corresponding environment [72].

The computational complexity and the required memory for the calculation of spatially-correlated shadow fading values in a cell with a reasonable resolution (e.g. 1 meter spacing between neighbouring points) also needs to be addressed. We have adopted the proposed low complexity approach in [72] where the points in the map of a cell are scanned one by one and just few neighbouring values around each point are taken into account for the spatial correlation calculation (e.g. two, four or eight points around each point determine the value of that point. More detail can be found in [72].

Figure 3.18 depicts an example for the uncorrelated and the spatially correlated shadow fading values across a cell. The spatially correlated values for an inner cell and first tier neighbouring cells are also depicted on Figure 3.19. In contrast to the completely random distribution of the uncorrelated values across the cell, the correlated values are changing smoothly from one point to any neighbouring points in the cell. Furthermore, the overall map of the cells across the cluster shows the same smooth distribution and the desired intra-cell correlation.

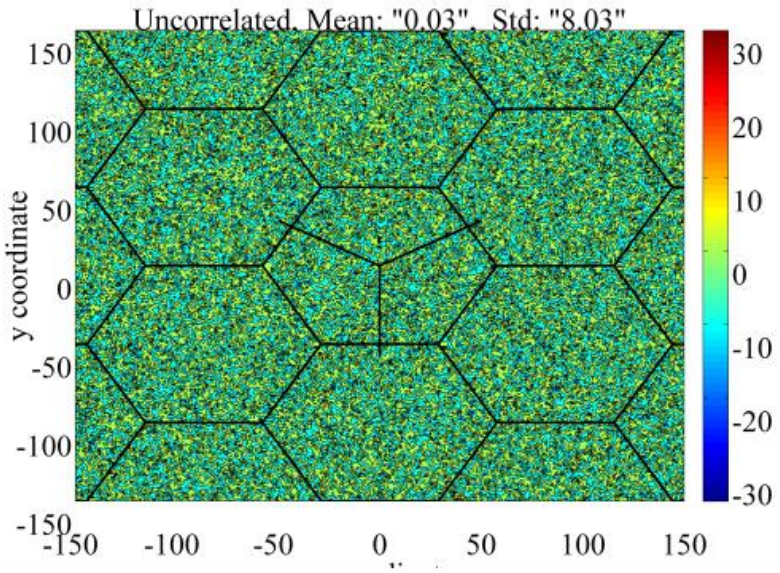


a) Seven sets of uncorrelated shadowing values (1st tier of a cluster)

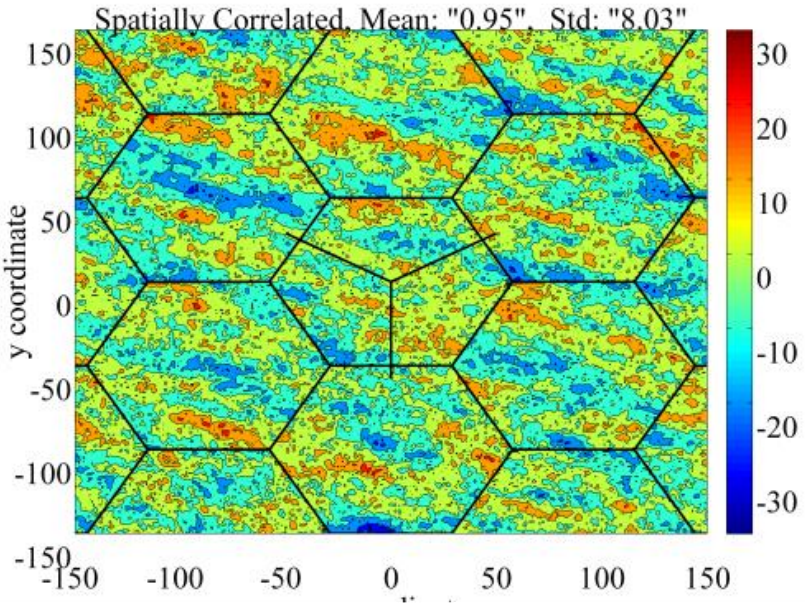


b) Values after inter/intra cell spatial correlation process

Figure 3.18 The shadow fading in one cell with and without the spatial correlation consideration (Shadowing effect: mean=0, deviation=8, decorrelation distance=25m, inter-site correlation=0.5). The cell radius is reduced for the sake of the presentation



a) Seven sets of uncorrelated shadowing values (1st tier of a cluster)



b) Values after inter/intra cell spatial correlation process

Figure 3.19 The shadow fading in a cluster with and without the inter/intra cell correlation consideration (Shadowing effect: mean=0, deviation=8, decorrelation distance=25m, inter-site correlation=0.5). Inter-site distance is reduced for the sake of the presentation

Our developed simulator provides the examining capability of the implementation algorithms for link adaptation and scheduling as well as the effect of the main operational parameters across the base station, UE or intermediate air interface. These include the time domain track of the behaviour of the system through the access to the runtime procedural data and the overall system performance through the evaluation of the final outcome of the system and its performance.

3.5.2 Simulator's Main Specifications

As it has been explained in the previous subsections, the simulator is developed using the object-oriented programming in MATLAB. The minimum system requirements for running the developed simulator are in the range of the normal specifications of contemporary laptops and personal computers. The minimum required computing power, memory and MATLAB software version for running this simulator smoothly (using the developed LTE-related functions, the employed functions from MATLAB library, Optimization and Communication toolboxes) are as follows:

- Processor: 1.6GHZ (per core) or higher
- Memory: 3Gig RAM or more (given windows 7 as OS)
- Software: MATLAB R11b or higher versions (default components + Optimization and Communication Toolboxes)

Although, the achievement of a satisfactory result is viable through the computing speed of a PC/Laptop device with multi core processor, it is also possible to use the capability of the computing cluster especially for some MATLAB functions which support parallel processing. This provides a high level of scalability for the software. The scalability of the simulator based on the computing time vs size of the network or number of the users is investigated in a quad-core laptop with the above minimum requirements. Figure 3.20 shows an example of the impact of the number of users over the simulation time for the following set of the simulation tasks with or without the bitwise operations of the PHY layer:

- the initial Network/UE settings
- eNodeB creation
- UE creation (including its channel status and traffic data scenario for whole simulation time).
- Allocation procedure (prioritization, link adaptation and scheduling algorithms) for 1000*TTI (i.e. 1000 ms).

Note that the simulation without the PHY layer operations means a limited estimation over the allocated data rate without the consideration of the line quality and the actual throughput of the system at the receiver. Figure 3.21 shows the boundary of the created system capacity for 5MHz bandwidth compared to its equivalent Shannon channel capacity. The gap depends on the actual capability of LTE structure and the CQI/SNR mapping and LTE-MCS (i.e. modulation and coding scheme)). The depicted steps of the achieved capacity in Figure 3.21 represents the switch between

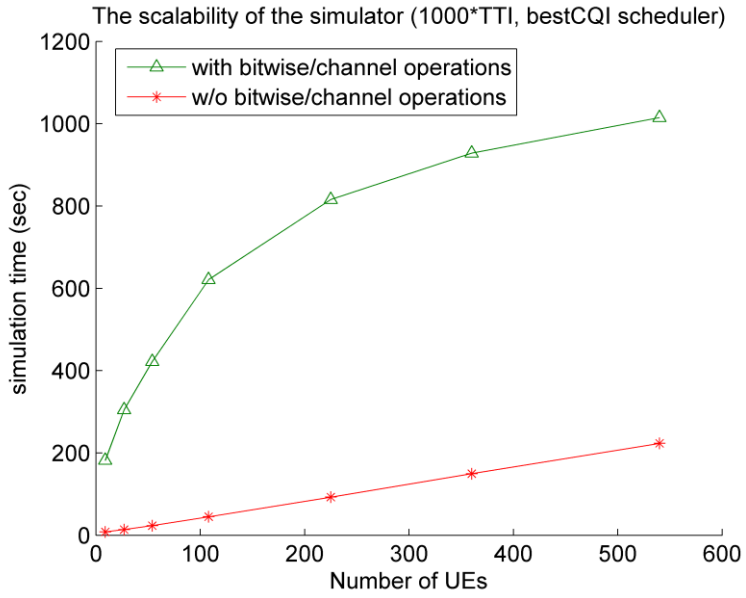


Figure 3.20 An example of the scalability of the simulator for 1000*TTI simulation time and a bestCQI scheduling algorithm

different modulation and code rates with the improvement of the channel status based on the link adaptation Functionality and mapping in Table 3.5.

This simulator also provides the ability to define fixed or moving users with or without a specific moving trace. It also provides the capability of the cross-layer algorithm implementation. As an example, the video bitrate can be adapted in an online fashion during the simulation process based on the channel status or other conditions from other layers or a parameter from MAC layer need to

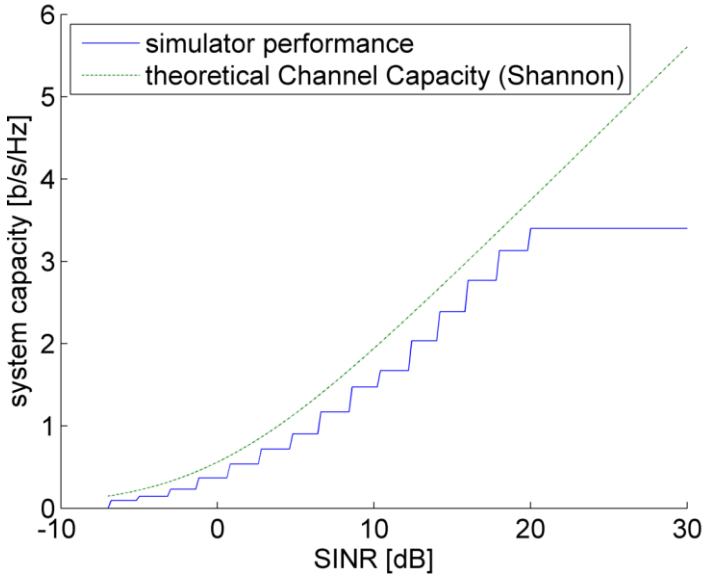


Figure 3.21 Created capacity vs channel status compared to the theoretical channel capacity (Shannon)

be adjusted online for maintaining a predefined value of fairness and efficiency when the instantaneous status of the users are changing.

The simulator has a modular structure flexible for the amendment, extension or substitution of each function or algorithm. It accepts formatted data from higher layers with desired segment sizes and inter-arrival times. At the moment a specific format to represent the video data has been employed while any other format related to other sources of data can be adopted in a similar way.

3.6 Summary

The roles of the main elements of LTE and selected parameters related to our research in the base station, communication channel and the user-side have been discussed in this chapter. These include the specifications of the employed Functionality and procedures such as retransmission, channel coding and frame timing as well as the transmitter and receiver's generic functional block diagram based on the 3GPP recommendations. Some important practical considerations have been reviewed afterward. The similarities and the differences between the implementation in LTE and the previous generation of the system (represented by HSPA) have also been studied. The timing of the initial transmission and retransmissions, acknowledgments mechanism, synchronization and adaptability of the processes have been used to highlight the differences which must be considered for applying the proposed solutions in the context of LTE-Downlink and LTE-uplink or HSPA downlink and uplink. The structure of our MATLAB-based simulator which has been developed to investigate the behaviour of the scheduler in LTE and verify the proposed analytical models has been described as well. Some preliminary simulation results have been used to depict the main characteristics of the simulator and the approaches to simulating the channel behaviour especially the large scale fading.

Chapter 4
Pause Intensity: A No-Reference Quality Assessment Metric for Video Streaming

4.1 Introduction

Video-on-demand services that utilize TCP, such as BBC's iPlayer, YouTube and Ustream, are rapidly growing with a prediction that Internet video traffic will account for 55 percent of all consumer Internet traffic in 2016 [5]. These services normally require a playout buffer to deal with the problems caused by TCP's congestion control mechanism. Due to network bandwidth scarcity and the demand for high definition quality video buffer underrun can occur, which results in a pause of a certain length before video playback can resume when sufficient data has been received in the buffer. Although TCP is designed to guarantee the reception of all packets to ensure the image quality, buffer underrun will cause impairments in video playout continuity which can affect the viewer's perceived quality.

Developing proper quality metrics is currently one of the research focuses on video technologies and services. However, the need for assessing the impairment in playout continuity has not been fully addressed. This quality issue is important for video streaming service providers to monitor the viewer's quality of experience (QoE), especially in TCP based streaming as the traditional metric: peak signal-to-noise-ratio (PSNR) is unsuitable for quality measurement in this scenario [74]. There are some attempts to characterize buffer behavior with specific metrics, such as the buffer underrun frequency or probability [75]. However, these metrics are unable to demonstrate their correlation with subjective results in quality assessment, which is discussed further in Section 4.2.

Our work builds upon a recently introduced new metric, namely *Pause Intensity* (PI) [1], and initial simulation and subjective testing results [76]. We show that viewer's QoE can be properly characterized by PI which is comprised of both pause frequency and pause duration, using the analytical model developed and extensive simulation and subjective testing results. We also establish

the connection between PI and network performance such as throughput and service levels such as the video playout rate. This unique feature makes PI a reference free metric which can be used to enable adaptive traffic control in streaming service delivery to meet the quality requirement defined by a certain PI value.

Discussions on the related work are given in Section 4.2. Section 4.3 describes the proposed model for Pause Intensity including the characteristics of buffer underrun, throughput and playback in the context of a streaming network and the derivation of a PI analytical model. In Section 4.4 the simulation and testing environments are set up, followed by the validation of the PI model by both simulation and subjective testing results in Section 4.5. The chapter is summarized in Section 4.6.

4.2 Related Work

As discussed in Subsection 2.4.2, metrics used for measuring video quality have typically been classified into three categories: full reference (FR), reduced reference (RR) and no reference (NR). FR metrics employ the original video as a reference point and the quality is determined by computing the difference between the original and distorted videos. The most common FR metric used is PSNR which is directly derived from the mean squared error (MSE). Due to the limitation of PSNR [77], [78] many other metrics have been proposed, such as structural similarity (SSIM) [79], video quality metric (VQM) [80], and other licensed tools such as SwissQual [81] and Kwill [82]. PSNR evaluation is especially not practical in the case of on-line streaming and on-the-fly assessment where some methods have been proposed to estimate the value of PSNR based on the video data characteristics [83].

NR metrics assess the content quality level without any knowledge of the original material. Commonly, this is done by identifying artifacts such as blurriness, blockiness, sharpness or a combination of related artifacts. Ferzli [84] and Liu [85] show that their metrics on blurriness and blockiness, respectively, are highly correlated with subjective data. The above metrics characterize the typical examples of artifacts that occur in best effort networks where data delivery is not guaranteed. In a TCP based streaming session, however, the artifacts such as blockiness, blurring and sharpness are not typically experienced as data delivery is fully ensured. The artifacts originated from the source of the delivery, for example due to the low rate encoding, are still happening and should be treated beyond the TCP layer.

The work presented in this chapter falls into the NR category and is concerned with a video evaluation tool for TCP based streaming. Other studies in this category have investigated varied aspects in the buffer underrun phenomenon. Kim [53] proposed a model to find a desirable buffer size determined by network characteristics and the underrun probability. However, no user satisfaction assessment is carried out apart from an assumption that a higher underrun probability can result in greater dissatisfaction.

The buffer starvation probability is a similar term to underrun probability, pause frequency, underflow probability or jitter frequency. It has been shown [86] that the starvation probability can be reduced by optimizing buffer settings and playout rate smoothing factors. The analysis of buffer delay has also been carried out, giving the conditions to avoid overflow and underflow in time varying channels [87] and the delay-underflow trade-off in a lossy network [88].

There are also schemes intended to control and reduce pause occurrence by using TCP-friendly stream rate adaptation to the changes in buffer occupancy [89] or optimizing TCP bandwidth allocation [90]. In [91] the buffer occupancy and network status have been used to evaluate and optimize the end user's perceived quality for video streaming, concerning the received video resolution levels but without considering continuity. The importance of mitigating buffer underrun in TCP based streaming has been highlighted in a recent survey paper [92].

Those metrics used to characterize buffer behavior including buffer-underrun or pause frequency and its related parameters fall short since they do not reflect viewer's QoE. This fact will be verified later in Section 4.5 using subjective assessment results. We will also prove that the new objective assessment metric, pause intensity, can precisely quantify the buffer underrun effect on the viewer's perception of playout quality, and that its correlation with the viewer's QoE is content independent.

4.3 Model of Pause Intensity

In this section, we show that the behavior of pause and play events during the playback follows the performance of throughput in a lossy channel. The characteristics of these events and throughput can be described using a unimodal probability density function (pdf) with their statistical moments (e.g. the mean value). Using the mean value we are able to develop a simple model for the playback buffer and derive the closed form formulation for pause duration, pause frequency and ultimately the pause intensity, as a function of throughput, video playout rate and receiver buffer settings.

4.3.1 The Characteristics of Buffer Underrun

In this section, the nature of buffer underrun and the resulting impairments are discussed. Buffer underrun is typically a result of inefficiency in the network due to bandwidth limitation and/or packet loss (e.g. TCP/IP over WiFi, congested network) which results in a throughput less than the required decoding rate of the video.

A pause is defined in this work as a temporary suspension of play followed by a period of playback which resumes from where the pause occurred. An overview of a typical streaming network is shown in Figure 4.1. The video player (decoder) provides the sink for the outgoing packets. The rate of the successfully received data from the TCP connection or termed throughput, η, is a function of the

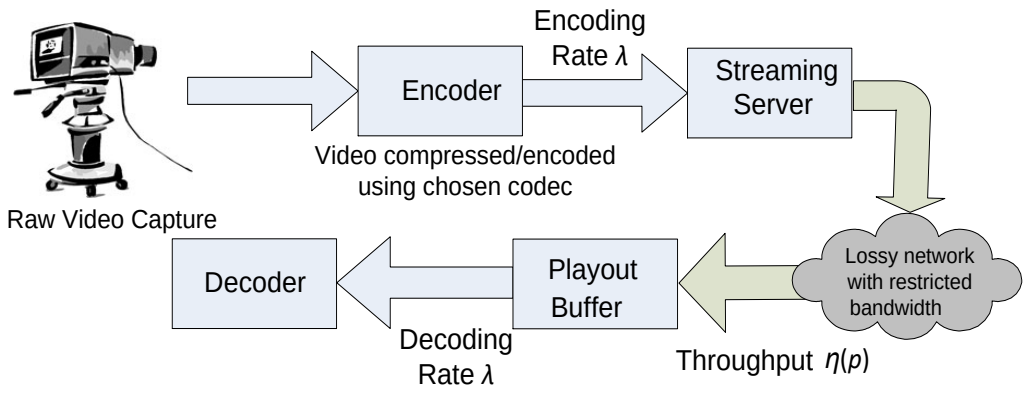


Figure 4.1 Video streaming architecture

packet loss probability, p, of the network. The required video playout rate, λ, is a characteristic of the video codec which is determined by the visual quality required of the video.

Figure 4.2 shows the structure of the playout buffer in connection with the network. The size of the playout buffer in the receiver is typically large enough to absorb the effect of small fluctuations in the network throughput and usually of sufficient size to store a few seconds worth of video data. The predefined threshold for video playback is referred to as q_{max}. Video playback will cease whenever the amount of stored packets in the buffer is less than the minimum threshold q_{min}. The packets are sorted using a FIFO algorithm. The instantaneous occupancy of the buffer is referred to as q.

When the video is played, an initial delay is experienced from the moment when the buffer receives data until the buffer occupancy reaches q_{max}. Following the initial delay, playback starts and one of three scenarios may take place as illustrated in Figure 4.3. Actually in the case where the average throughput is greater than or equal to the required playout rate λ, ($\eta_{average} \geq \lambda$), and provided that the receiver has enough memory to buffer the received data, there will be just an initial delay occurrence followed by a long period of video playback. When the available network throughput is less than λ, pause events will be experienced whenever the amount of data in the buffer falls below q_{min}. The analytical model presented in this section is based on the ($\eta_{average} < \lambda$) case.

4.3.2 Throughput Characteristics

Pause occurrence during the playback of a video stream is mainly caused by insufficient incoming data rate of the buffer or network throughput due to packet loss. Packet loss as a stochastic process has been studied extensively in the context of TCP/IP networks [93-95]. Since the detailed characteristics of the distribution of packet loss and throughput (e.g. average, skewness and kurtosis)

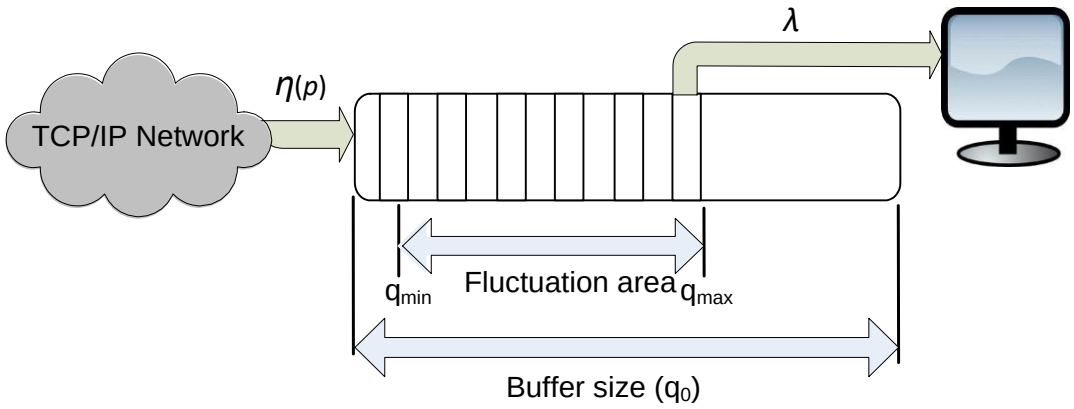


Figure 4.2 Buffer structure and related settings.

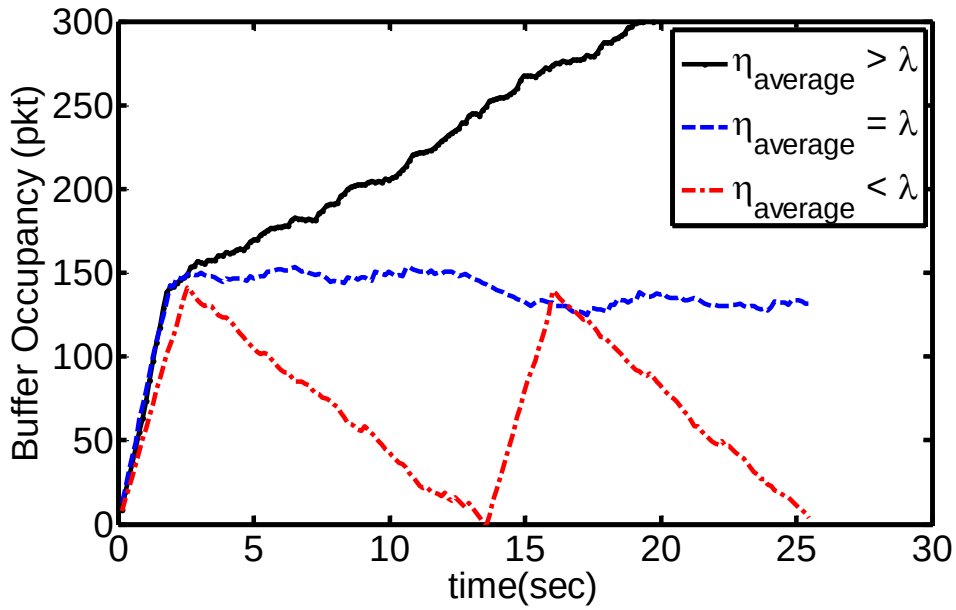


Figure 4.3 Buffer characteristics

depend on the network structure and traffic status, the following assumptions and considerations have been made for the subsequent discussions:

a) Bernoulli, Geometric and Gamma functions have been employed to fit the loss event distribution functions [93-95]. A continuous Gamma density function as the distribution function of 'probability of packet loss', p_l, is used in this work, so that any of the above mentioned functions could be utilized and compared. A Gamma distribution function is a continuous probability distribution function with two parameters, namely shape parameter k and scale parameter θ (both

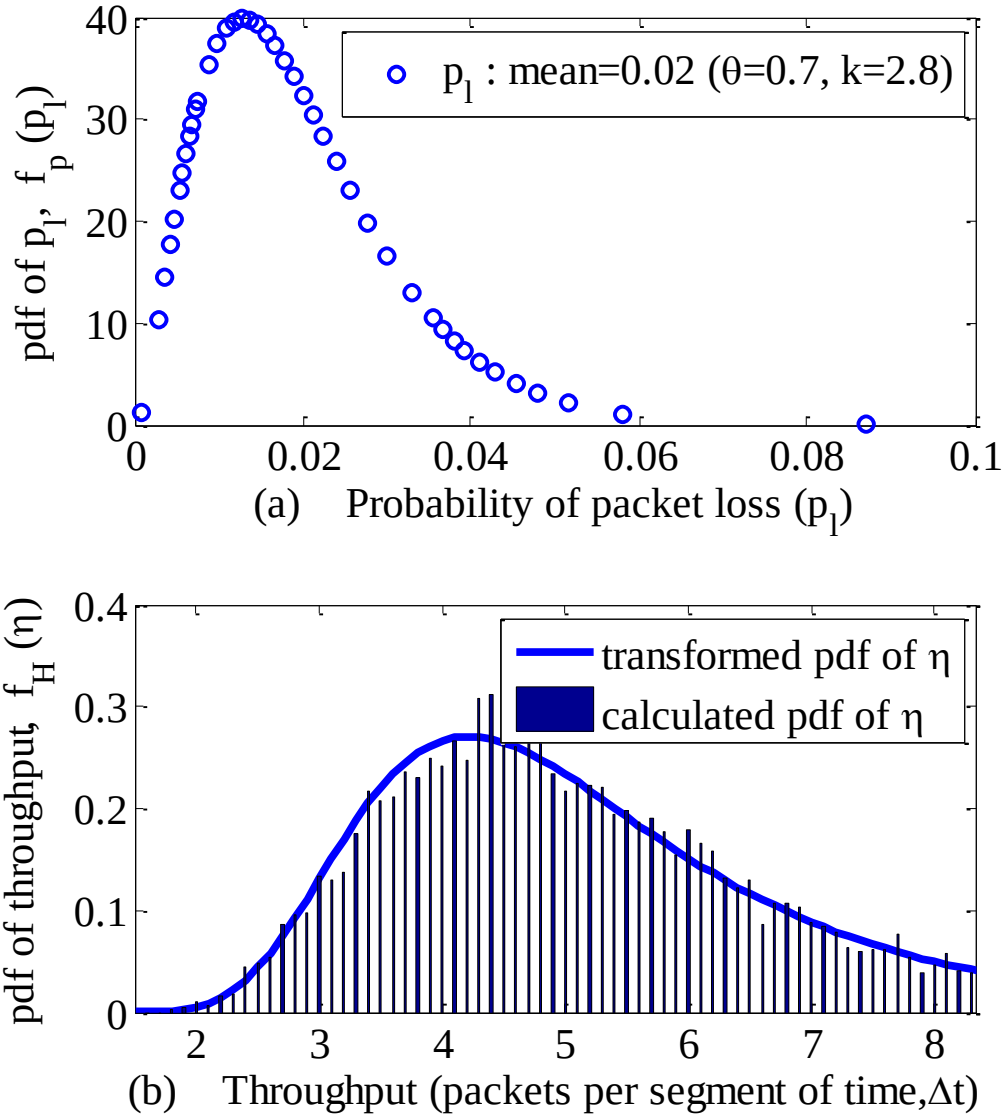


Figure 4.4 Examined pdf of (a) the probability of packet loss and (b) achieved throughput.

positive real numbers) through which the mean, variance and skewness of the distribution will be defined [100]. Figure 4.4(a) shows an example of the assumed pdf of p_l, $f_P(p_l)$, with an average packet loss rate of 0.02 and generated using *Gamma* ($k=2.8$, $\theta=0.7$) where k and θ are the shape and scale of a gamma distribution, respectively.

b) Although throughput monotonically decreases with respect to the packet loss rate in a TCP connection, the shape and pace of the change depend on the exact variant of TCP and the type of the network involved. In this work we adopt the widely accepted TCP-Reno model [96] and related specifications [97] for network throughput analysis. The throughput for a TCP-Reno connection is given as [96]:

$$\eta(p_l) = \begin{cases} \dfrac{1}{R\sqrt{\dfrac{2bp_l}{3}} + T_0 min\left(1,3\sqrt{\dfrac{3bp_l}{8}}\right)p_l(1+32{p_l}^2)} & with\ timeout \\ \dfrac{1}{R}\sqrt{\dfrac{3}{2bp_l}} & without\ timeout \end{cases} \quad (4\text{-}1)$$

where T_0 is the timeout in TCP-Reno and R is the round trip time. The throughput model without the timeout effect in (4-1) is invertible but not sufficient to analyze the pause duration behavior. Therefore, the non-invertible model of throughput with the timeout effect will be used in the rest of the chapter. The number of rounds for each increment in window size, b, in compliance with our simulation setting is assumed to be 2. Furthermore, the adequate range of p_l lower than 0.12 will satisfy the condition to assume $min\left(1,3\sqrt{\frac{3bp_l}{8}}\right) = 3\sqrt{\frac{3bp_l}{8}}$. Hence the throughput concerned as a function of p_l is:

$$\eta_{Reno}(p_l) = \frac{1}{\frac{2R}{\sqrt{3}}{p_l}^{\frac{1}{2}} + \frac{3\sqrt{3}T_0}{2}{p_l}^{\frac{3}{2}}(1+32{p_l}^2)} \quad (4\text{-}2)$$

The actual throughput of a TCP connection is also affected by the advertised window setting from the receiver side and bandwidth limitation due to a bottleneck in this connection. This will be considered later in the model formulation and evaluation.

c) To find out the transformed pdf of a random function (i.e. throughput, η) based on the known pdf of its random variable (i.e. probability of packet loss, p_l), the following relation will be used when the function is analytically invertible.

Given a function $y=g(x)$, if the distribution function of x is known as $f_x(x)$, then the distribution function of y is given by

$$f_Y(y) = \frac{f_X(g^{-1}(y))}{|g'(g^{-1}(y))|} \quad (4\text{-}3)$$

A variation of (4-3) in the case of y being a non-invertible function is

$$f_Y(y) = \sum_i \frac{f_X(x_i)}{|g'(x_i)|} \quad (4\text{-}4)$$

in which x_i is the *i-th* root of $y=g(x)$ for a given value of y. This relation can be applied to finding the pdf of throughput, $\eta(p_l)$, based on the known pdf of p_l without an explicitly inverted $p_l(\eta)$. Since η monotonically decreases with respect to p_l, it is possible to solve (4-2) numerically for a given value of η and find the corresponding pdf, f_H, using (4-4). Figure 4.4(b) shows the achieved pdf of throughput for the given pdf of the packet loss probability. To evaluate the accuracy of this approach, a histogram of 5000 calculated values of throughput is shown alongside. Throughput values are calculated based on random packet loss rates derived from the given Gamma pdf and show a good agreement with the achieved pdf.

4.3.3 Playback Characteristics

The achieved distribution of throughput will be used to analyze the pause behavior as follows. In Figure 4.5 the instantaneous growth of buffer occupancy, Δq_i, and the total occupancy, q, can be written as a function of throughput during the time interval Δt_i, i.e.:

$$\Delta q_i = \Delta t_i . \eta_i \quad \text{and} \quad q = \sum_i \Delta q_i = \sum_i \Delta t_i . \eta_i \tag{4-5}$$

where $\eta_i = \eta(p_i)$. it must be noted that during the pause time there is no packet consumption in the buffer, so λ has no influence during this period. If Δt_i is constant i.e. $\Delta t_i = \Delta t$, and small enough to assume more than one segment per each pause event, then the occupancy of buffer, q, after m segments will be a random variable equal to the summation of m random variable, i.e.:

$$q = \Delta t \sum_{i=1}^{m} \eta_i \tag{4-6}$$

Given the most common size of playback buffer at the receiver which is usually more than one second, a value of $\Delta t = 100ms$ will be sufficient in (4-6) to satisfy the above conditions. If q_0 is the difference between the minimum threshold (i.e. the trigger of pause) and the maximum threshold (i.e. the trigger of play resumption) of the buffer, finding the value $m=m_0$ that leads to $q=q_0$ is actually another way of presenting the pause duration, v, i.e.:

$$\begin{cases} v = m_0 \Delta t \\ where\ m_0\ satisfies: \ \ \Delta t \sum_{i=1}^{m_0} \eta(p_i) = q_0 \end{cases} \tag{4-7}$$

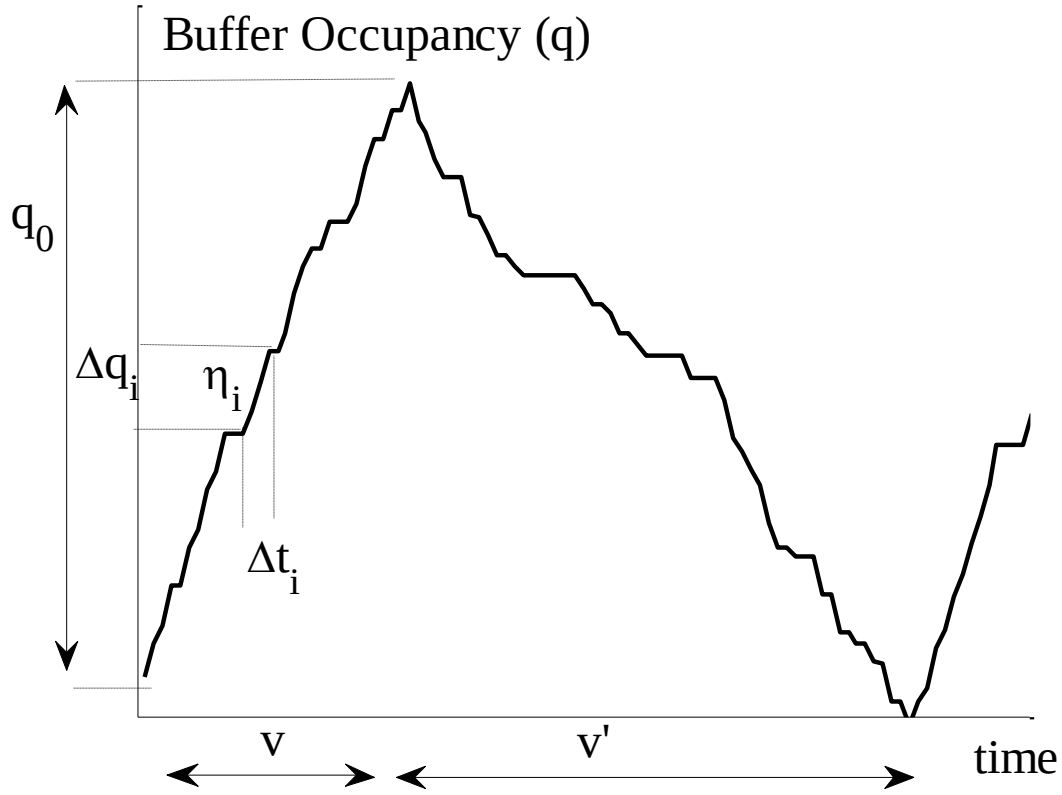


Figure 4.5 Buffer occupancy vs. time

The equation (4-6) is a summation of *m* values of a random variable with the known derived pdf described in the previous subsection. The *m* variables in (4-6) have an independent and identical distribution (i.i.d). As long as the assessment segment Δt is small enough, compared to the minimum possible length of a pause to guarantee a large value of *m* (e.g. *m>10*), the *Central Limit Theorem* can be used to find the distribution of *q* in (4-6), i.e. for η*:* $f_H(\mu_\eta, \sigma_\eta)$, hence we have

$$q: \mathcal{N}_Q\left(m\mu_\eta, \sqrt{m}\sigma_\eta\right) \tag{4-8}$$

where f_H is the pdf of throughput with mean and variance μ_η and σ_η, respectively (referring to Figure 4.4). In the proposed analytical model and the upcoming simulations and validation experiments, it is assumed that the fluctuation of the throughput is mainly due to the stochastic behavior of the channel during the observation period which lasts long enough to satisfy the i.i.d condition in (4-8). Any specific control mechanism which intervenes in the process and determines the network performance may affect the validity of this model.

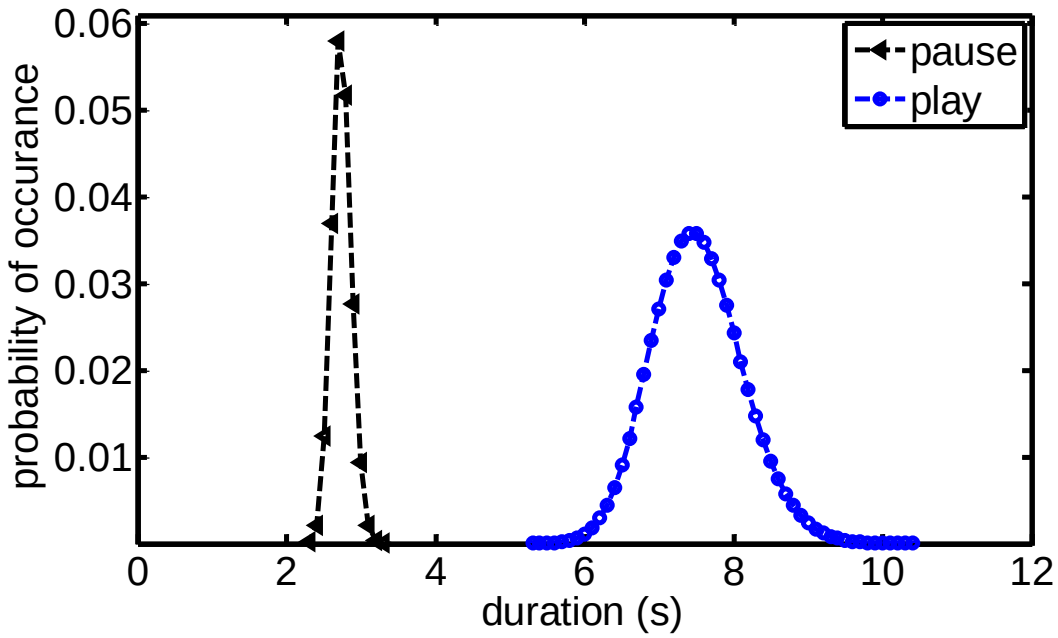


Figure 4.6 Distributions of pause and play durations

The pdf of buffer occupancy, N_Q, is a normal distribution with a mean equal to $m\mu_\eta$ and a variance equal to $m\sigma_\eta^2$. Obviously a subset of N_Q values, which are relevant to $q=q_0$ when $m=m_0$, will satisfy (4-7). Given the relation between m_0 and v in (4-7), these values will lead to the desired values for pause duration v, and their probabilities. Therefore, the probability of occurrences of the pause duration v can be expressed as:

$$\begin{cases} p(v) = p(q = q_0)|_{m=m_0} = \dfrac{1}{\sigma_\eta\sqrt{2\pi m_0}} e^{-\frac{(q_0 - m_0\mu_\eta)^2}{2m_0\sigma_\eta^2}} & \forall\, m_0 > 0 \\ given\ \ \eta: f_H(\mu_\eta, \sigma_\eta),\ \ q: \mathcal{N}_Q(m\mu_\eta, \sqrt{m}\sigma_\eta) \end{cases} \tag{4-9}$$

From (4-7), (4-8) and (4-9), we can conclude that for any given pdf of η with its μ_η and σ_η the probable pause durations can be described as a subset of normal distribution given in (4-8), regardless of the type of distributions for η. It is noted that (4-9) includes some samples of (4-8) which satisfy the condition given in (4-7). Figure 4.6 depicts (4-9) as the examples of p_l and η in Figure 4.4.

Using a similar method we can derive the distribution of play duration, v', shown in Figure 4.5, during which both throughput and the playout/code rate (λ) must be considered while λ is assumed to be constant, i.e.:

$$\begin{cases} v' = m_1 \Delta t \\ where\ m_1\ satisfies\ \ \Delta t \sum_{i=1}^{m_1} [\eta(p_i) - \lambda] = q_0 \end{cases} \tag{4-10}$$

Figure 4.6 shows the probability of occurrence of the play duration corresponding to the pause duration derived previously.

Although the above discussion gives an insight into the behavior of pause due to the impairment of transport connections, it does not give a direct closed form expression showing the relationship between the pause duration or pause frequency and network conditions such as throughput or the packet loss rate. Most importantly, it does not show how this pause or buffer behavior affect the quality perceived by the end users of video streaming services using the TCP protocol. In the next subsection, the analytical model of the quality metric, Pause Intensity, will be presented, which combines the statistical features of both pause duration and pause frequency.

4.3.4 Pause Intensity: Analytical Model Derivation

By applying the transformed pdf and the central limit theorem we have shown that throughput, pause duration and play duration have unimodal distributions. This allows us to use the statistical elements such as the mean or maximum probable value to achieve the closed form representations of buffer characteristics. Any other parameter such as pause frequency will be defined as a function of these representative values. In this subsection, we establish the model of pause intensity, a no-reference quality assessment metric, based on the buffer underrun properties for video streaming in TCP networks. These buffer underrun properties are characterized by the average pause duration $\bar{v}$, and average pause frequency $\overline{f_v}$, given network and buffer conditions and settings such as throughput, the code rate and receiver buffer fluctuation range.

Figure 4.7 shows a typical pause-play period with all related parameters, where the playout buffer is large enough to absorb the effect of small fluctuations. The trend of the accumulated data in the buffer between t_{v1} and t_{max} follows a line with the gradient equal to the average throughput in that period. The durations of pauses and plays are denoted by *v* and *v'*, respectively, and *w* represents the duration of a pause-play event. A pause event occurs when the number of buffered packets is reduced to q_{min} at time t_{v1}. Play will resume whenever the number of buffered packets reaches q_{max} at time t_{max}. If the average throughput is less than the required video playout rate λ, the next pause will occur at time t_{v2}. The buffered data during the pause-play event can be expressed as:

$$\begin{cases} q_{pause} - q_{min} = \dfrac{q_{max} - q_{min}}{t_{max} - t_{v1}}(t - t_{v1}), & t_{v1} < t \le t_{max} \\ q_{play} - q_{max} = \dfrac{q_{max} - q_{min}}{t_{max} - t_{v2}}(t - t_{max}), & t_{max} < t \le t_{v2} \end{cases} \tag{4-11}$$

Unlike the play time, there is no output from the buffer during the pause, which leads to:

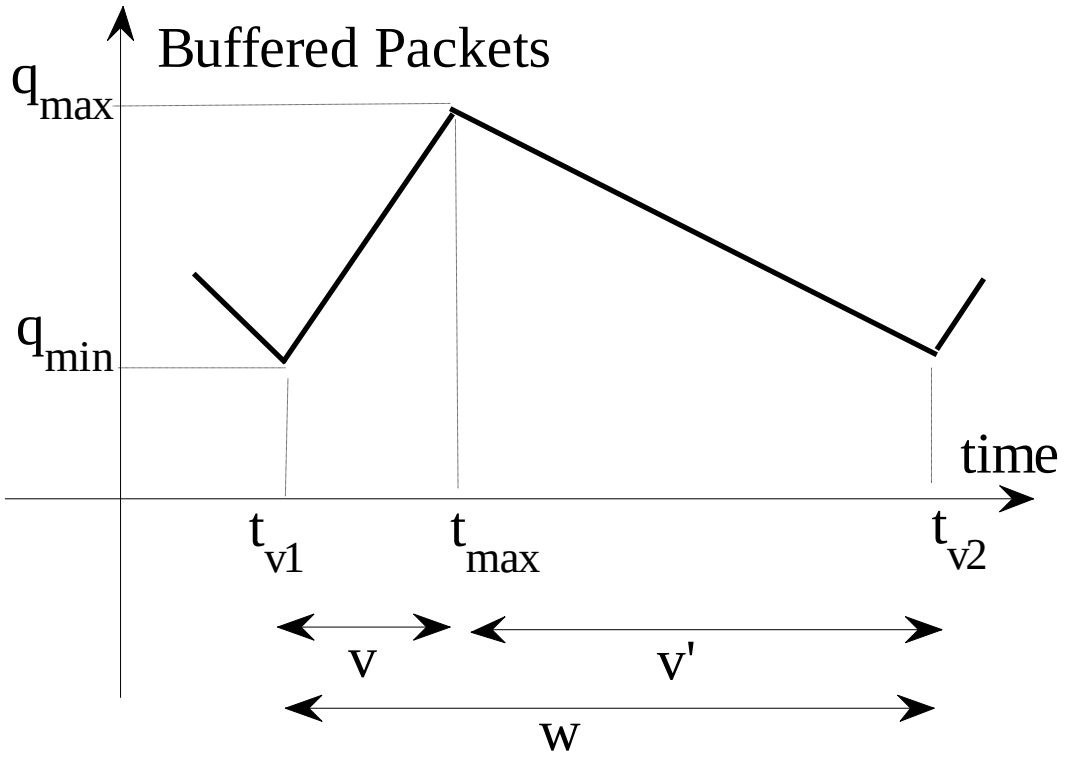


Figure 4.7 A typical pause-play period

$$\begin{cases} q_{pause} = q_{min} + \eta(t - t_v), & t_{v1} < t \le t_{max} \\ q_{play} = q_{max} + (\eta - \lambda).(t - t_{max}), & t_{max} < t \le t_{v2} \end{cases} \tag{4-12}$$

From (4-11) and (4-12), the parameters *v*, *v'* and *w* can be expressed as:

$$\begin{cases} v = t_{max} - t_{v1} = \dfrac{q_0}{\eta} \\ v' = t_{v2} - t_{max} = \dfrac{q_0}{\lambda - \eta} \\ w = \dfrac{q_0 \lambda}{\eta(\lambda - \eta)} \\ q_0 = q_{max} - q_{min} \end{cases} \tag{4-13}$$

where q_0 represents the fluctuation range in the buffer. Recalling the definitions of pause intensity, the average pause duration and pause frequency from [1] and discussions in Subsection 4.3.3, we use the average values to represent the parameters given in (4-13), and then we have.

$$\begin{cases} average\ pause\ duration = \bar{v} = \dfrac{q_0}{\eta} \\ pause\ frequency = \overline{f_v} = \dfrac{1}{\overline{w}} = \dfrac{\eta(\lambda - \eta)}{q_0 \lambda} \\ Pause\ Intensity = PI = \bar{v}.\overline{f_v} = 1 - \dfrac{\eta}{\lambda} \end{cases} \tag{4-14}$$

The pause duration $\bar{v}$ is a function of a dedicated buffer fluctuation range q_0 and network throughput $\eta(p)$. It is however not affected by the video playout rate λ as there is no output from the buffer during the pause. Pause intensity (PI) represents the relative effectiveness of throughput compared to the required playout rate λ and is not affected by q_0. The frequency of a pause-play sequence $\overline{f_v}$ is built upon the combination of the playout rate, network throughput and buffer settings (i.e. the buffer size). These features will be exploited further during the performance analysis in the following sections.

As explained in Subsection 4.3.2, we adopt TCP-Reno in our work and the throughput used in the PI model (4-14) is given in (4-2). Recalling the nature of a pause-play period *w*, intuitively it can be seen from Figure 4.7 that *w* has a finite value if $\eta_{average} < \lambda$. The change rate of *w* in (4-13) is given by:

$$dw = \frac{\partial w}{\partial \eta} d\eta + \frac{\partial w}{\partial \lambda} d\lambda = q_0 \left(\frac{-\lambda(\lambda - 2\eta)}{\eta^2 (\lambda - \eta)^2} d\eta + \frac{-1}{(\lambda - \eta)^2} d\lambda \right) \tag{4-15}$$

Furthermore, the change rate of *w* with respect to η, defined as β, can be determined by:

$$\begin{cases} \beta = \dfrac{dw}{d\eta} = q_0 \dfrac{-\lambda(\lambda - 2\eta)}{(\eta(\lambda - \eta)^2} \\ \lambda = \lambda_0\ (constant) \quad \therefore \quad \dfrac{d\lambda}{d\eta} = 0 \\ \eta = \min\left(\eta_{Reno}(p), BW_{min}, \dfrac{W_m}{RTT} \right) \end{cases} \tag{4-16}$$

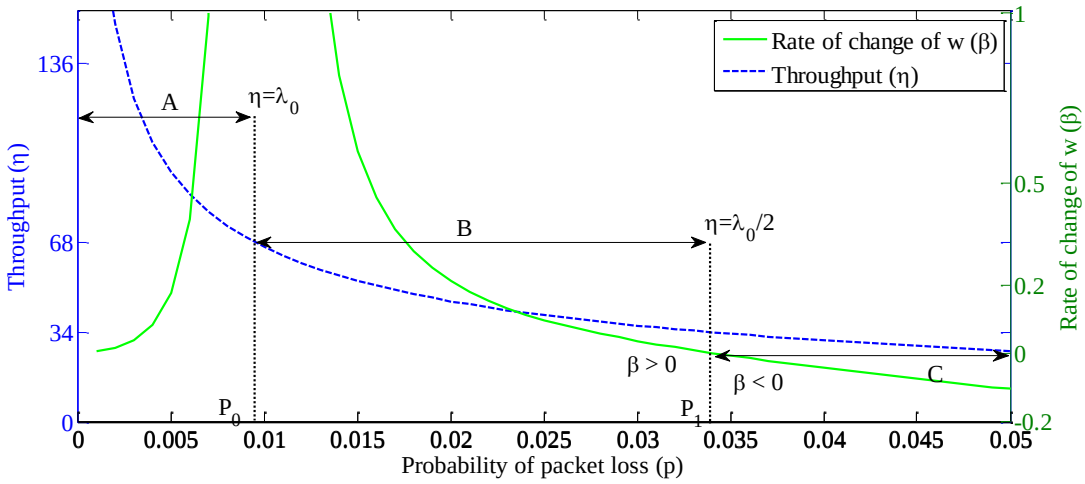


Figure 4.8 Critical points of pause-play sequence in relation to throughput and packet loss probability

in which the playout rate λ is considered to be a positive constant value without a continuous change. The actual throughput η is the minimum among the calculated throughput in (4-16), the bottleneck bandwidth BW_{min} and the advertised window size per round-trip time W_m/RTT.

Figure 4.8 depicts the features of the rate of change of w (i.e. $\beta = dw/d\eta$) given by (4-16) and the critical points against the probability of packet loss rate p. As it is shown, the vertical asymptote $p = P_0$ in which $\eta(p_0) = \lambda_0$ is the returning point of w from infinity, and consequently it is the point from which pauses begin to occur within the range denoted by B. From the point where $p=P_1$, which results in $\eta(p_1) = \lambda_0/2$, the pause duration starts exceeding the play duration and viewers satisfaction will be less likely within the range denoted by C. Later we will see that $p=P_1$ is the extremum point of pause frequency as well. Throughput in the range denoted by A is higher than what is required and pauses are therefore unlikely to happen.

The pause intensity model established in this section will be validated by simulation and subjective testing to show its effectiveness, accuracy and other features in the following sections.

4.4 Simulation and Subjective Testing

4.4.1 Simulation Setup

The simulation was carried out using NS-2 with the parameters specified in Table 4.1. A simple bottleneck was established with a single sender and receiver. Only one video stream was assumed and no background traffic was employed. Packet loss was set to vary from 0% to 12% which affects network throughput and consequently buffer behavior. An element of randomness was added into the timing of packet loss occurrence, i.e. for each value of the loss rate, 10 simulations with different

Table 4.1 Simulation Setup

Parameter	value	Parameter	Value
Connection	TCP(Reno)	**Video code rate λ_0**	100 kB/s
Bottleneck	1Mb/s	**Packet Loss Range**	0%-12%
Packet size	1500 B	q_{max}	200kB
$RTT_{average}$	128 ms	q_{min}	1.5kB
T_0	128 ms	**Window Size W_m**	20

timing of the packet loss event were executed. The mean and deviation of the simulation results for that loss rate can then be shown.

NS-2 was used to provide the output of the network and a separate module was written to simulate buffer behavior based on the discussion in Section 4.3. The buffer size (198.5kB or ~133 packets) was selected to be approximately double the video coding rate, which correlates to around 2 seconds of video. The playout rate and buffer size were also selected based on the requirements of live streaming services.

4.4.2 Subjective Testing Setup

In order to verify the success of the pause intensity metric, subjective testing was also carried out. Testers were instructed that ratings given for each video should represent their overall viewing experience and reflect their real viewing expectations. Due to the nature of the impairment, relatively long video sequences of 90 seconds were used. In accordance with ITU P.911 [98], a five scale overall quality rating, i.e. Mean Opinion Score (MOS) was employed with the terminology recommended by the ITU. In contrast to the previous work [76] which used the Degradation Category Rating (DCR), this work adopted the single stimulus method: Absolute Category Rating (ACR). DCR requires a reference video for each and every video, which would mean doubling the overall test time. Our clips were already relatively long, and anything longer would, I think, have caused the viewer to lose concentration and not rate the video clips properly. ACR doesn't require a reference, we can just show the clips. Furthermore, DCR focuses on the level of impairment, not the overall video quality (which is what we were trying to capture with PI, QoE). The ACR scale is defined: Excellent, good, fair, poor, bad. The DCR scale is defined: Imperceptible, Perceptible but not annoying, slightly annoying, annoying, very annoying. Our focus with PI is QoE, whereas DCR focuses on level of impairment. What we did for our tests was essentially a smaller version of crowdsourcing, so it's not an unknown concept in our area of research. There is however no universal standard that defines how crowdsourced based tests should be performed, and there probably needs to be one.

Table 4.2 shows the corresponding setting parameters for two groups of subjective testing that were carried out: Subjective Testing-1 and Subjective Testing-2.

Table 4.2 Subjective Testing Setup

	Subjective Testing 1				Subjective Testing 2
	MotoGP (M)	**Run (R1)**	**News (N)**	**Cartoon (C)**	**Rally (R2)**
Codec	H264	H264	H264	H264	H264
resolution	540x360	640x360	640x360	640x360	640x360
frame rate	30	25	25	25	30
encoding rate	840kbps	781kbps	781kbps	781kbps	788kbps
video length	90s	90s	90s	90s	90s
reviewers/ clip	19	17	16	17	20
no. of clips	16	10	10	10	12

In Subjective Testing-1, four different types of video sequences were used, namely MotoGP, Run, News and Cartoon, with similar parameters as shown in Table 4.2. 'MotoGP' have a large amount of fast paced motion in scenes and a large number of cuts from scene to scene. 'Run' contains a lot of motion in the image itself and, in addition to which there is a substantial amount of camera panning. 'News' on the other hand contains neither camera panning nor large amounts of motion in the clip and with few scene changes. 'Cartoon' contains a small amount of panning, a fair amount of movement within frames and a large number of scene changes. The choice of these clips was aimed to assess not only the correlation property of pause intensity (the new objective metric) with the subjective opinion scores, but also the content independency of PI.

In Subjective Testing-2, further investigation was carried out to assess the robustness of PI in terms of accommodating different buffer characteristics, especially to test if the metric still provides a good correlation with viewer experience in more extreme cases. This part of the work can therefore provide stress testing to evaluate the presence of either high pause frequency or long pause duration whilst maintaining a constant value of PI. The content named 'Rally' is chosen for the discontinuity to make an obvious contrast for testers. Characteristics of this video sequence are in the same range of Subjective Testing-1 given in Table 4.2.

The detailed compositions of each PI used for both groups of subjective testing are provided in Tables 4.3 and 4.5, respectively.

Figure 4.9 shows an example of the impairment characteristics in Subjective Testing-1 using the MotoGP video sequences. In this case, varying levels of pause intensity were used in a range of 0.05 – 0.75. As discussed previously, each pause intensity value is repeated because the pause duration and frequency is varied to make different compositions of the same pause intensity value. To reveal the buffer behavior for each pause intensity value, two different compositions for the same PI value were used. The PI values shown on the horizontal axis of Figure 4.9 represent the first and second scenarios alternatively. The first scenario has a lower pause frequency (represented by the black

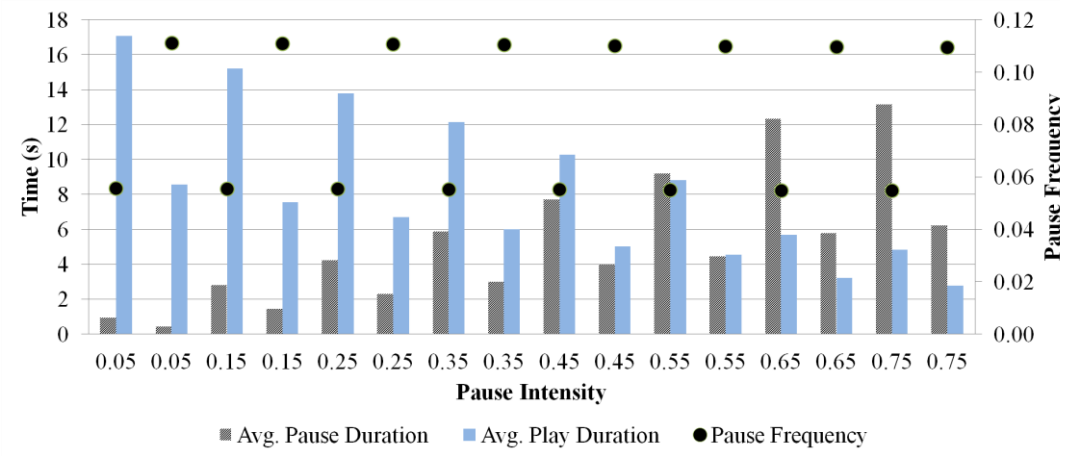


Figure 4.9 An example of the characteristics for subjective testing-1 (MotoGP video sequence scenarios)

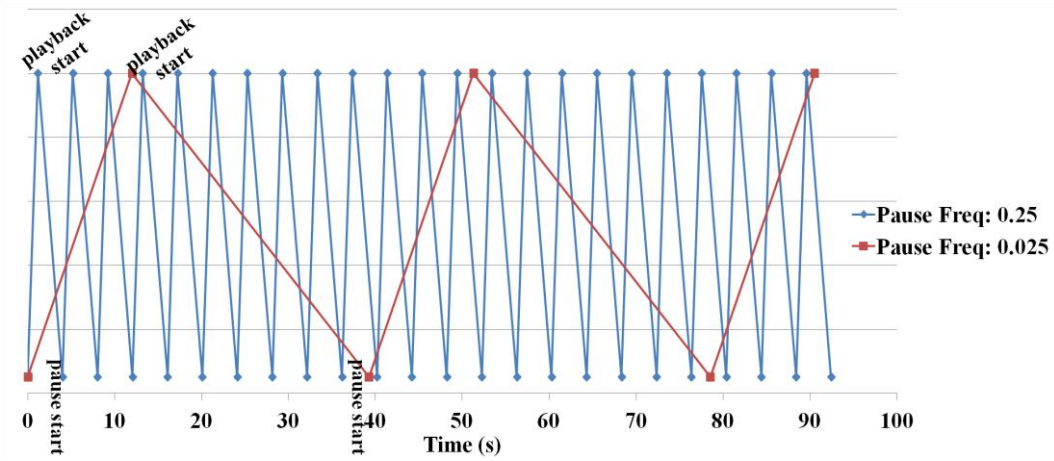


Figure 4.10 Pause characteristics of two videos with very different frequencies but the same PI

points in Figure 4.9) than the second scenario, meaning that higher value of PI results in higher pause and lower play duration regardless of the value of pause frequency. Furthermore, both the play duration and pause duration in first scenario are larger than those in scenario 2 which reflects the impact of pause frequency.

Figure 4.10 shows an example of videos used in Subjective Testing-2 (stress testing of PI with either extremely high pause frequency or long pause duration), which illustrates the contrast between pause-play scenarios with pause frequencies being 0.03 and 0.25, respectively, but with both having the same pause intensity. The upward slope is the time taken to fill up the buffer and the downward

slope shows the time when the video is played out. In addition, the significant difference in pause duration between the two videos is also clearly visible.

In all tests [99], video sequences with specified pause intensity were selected randomly from the pool of possible scenarios. Testers were allowed to vote for up to 6 clips in each run and each video was run between 14-20 times for the purpose of evaluation. Testers were invited through social media and local communities and largely did not possess any specific technical background. The testing environment also allowed the user to choose whether to continue to watch or not due to the level of impairment imposed.

4.5 Results and Analysis

4.5.1 Model Verification by Simulation

Figure 4.11 illustrates the results of the analytical models discussed in sub-sections 4.3.3 and 4.3.4 in comparison with the simulation results obtained in the environment with the parameters given in Table 4.1. As described earlier, each value produced in simulations represents the mean and deviation of 10 runs. Additionally, for each set of simulation test results, the best fitted curve to the model has been depicted. Clearly, both simulation and model results, as exhibited in Figure 4.11, are closely matched, which suggests that our model can successfully characterize the buffer underrun behavior with a high precision.

Figure 4.11(a) shows in pause duration against the probability of packet loss in the network. It may be recalled that pause duration is a function of throughput η and the buffer fluctuation range q_0, which is not dependent on the playout rate λ, as demonstrated in (4-14). It can be understood that as the packet loss probability increases, throughput decreases and therefore more time is required to fill the buffer to the playout threshold level q_{max}. It is also noticed that pause duration remains unchanged for the small values of p as playback buffer experiences a relatively stable network throughput under this condition.

Another important point we would like to make is based on the result of pause frequency or underrun probability, as shown in Figure 4.11(b). It reveals that pause frequency does not change monotonically as the packet loss probability increases. It increases with the loss probability when p is relatively small, and will adversely decrease when p increases up to a certain value (denoted by P_1 in Figure 4.11(b) and Figure 4.8). This is because the pause duration increases steadily with the loss probability as shown in Figure 4.11(a). As a result, the number of pauses will be reduced, so will the pause frequency. This phenomenon reveals that using pause frequency or the underrun probability alone simply cannot reflect the video playout quality as the viewer's QoE is directly affected by the change in loss conditions of the network.

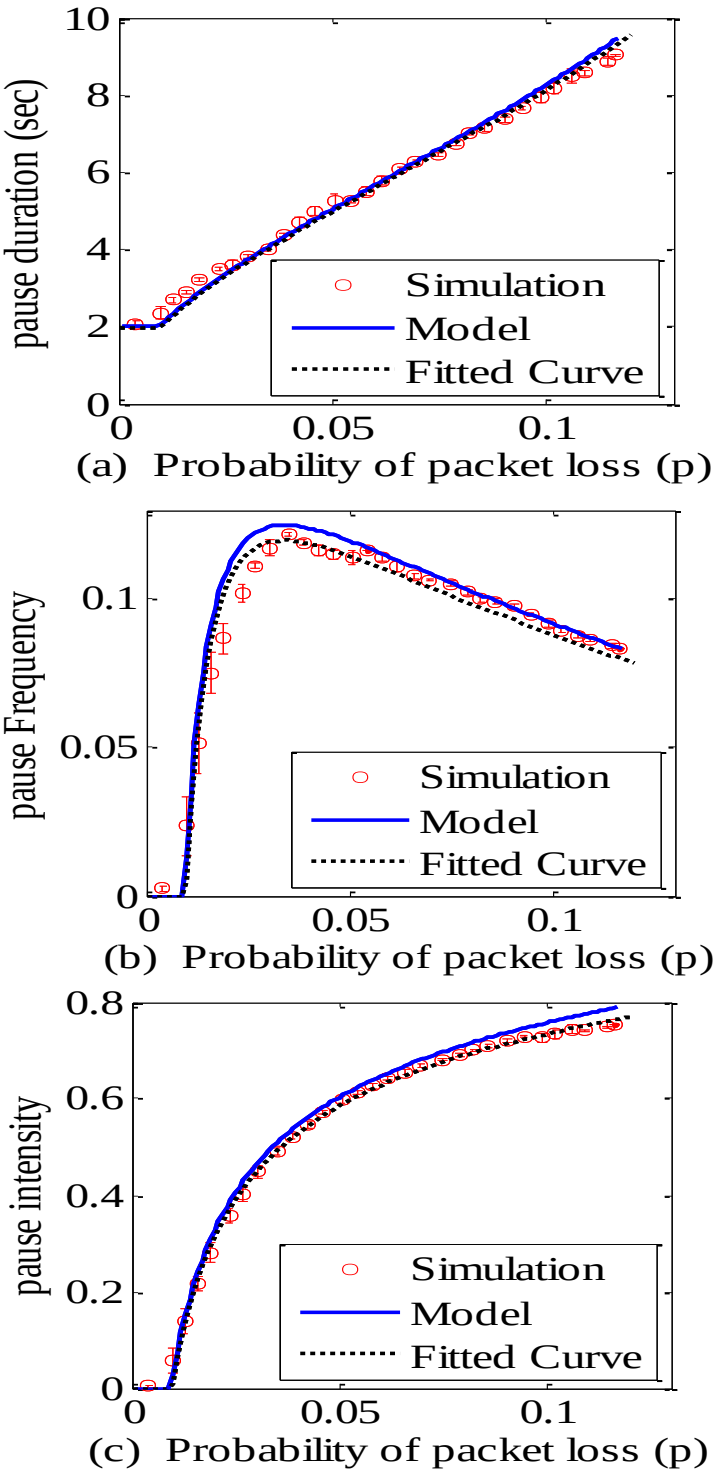


Figure 4.11 Model and simulation comparisons: (a) pause duration vs. loss probability; (b) pause frequency vs. loss probability; and (c) PI vs. loss probability.

Figure 4.11(c) shows the performance of pause intensity which changes monotonically in the whole range of the packet loss probability specified. The advantage of using PI is demonstrated as opposed to using a single parameter such as pause frequency as discussed above. PI is shown having a positive correlation with the packet loss probability, meaning that the metric is able to reflect network conditions, which is not the case for pause frequency and other parameters associated with buffer underrun such as average pause duration and average play duration. In summary, PI is a comprehensive metric that encompasses all the behavioral characteristics of the buffer. This conclusion will be further confirmed by subjective testing results discussed later.

4.5.2 Subjective Assessment-1 for Pause Intensity

As described in Section 4.4, 45 video sequences of four different types (MotoGP, Run, News and Cartoon) were used in the first group of subjective testing. The detailed setting parameters such as pause intensity, pause frequency and pause duration, and the testing results in terms of the MOS rating for each test are given in Table 4.3. Figure 4.12, constructed using the information provided in Table 4.3, demonstrates the advantages of using the pause intensity metric over pause frequency and pause duration.

First of all, Figure 4.12(a) shows two important features of this new metric: (1) pause intensity is closely correlated with the viewer's experienced quality for a wide range of PI values which encompass varying compositions of pause frequency and pause duration; and (2) this correlation is consistent with different types of video sequences tested, or in other words, pause intensity is highly content independent.

In Figures 4.12(b) and 4.12(c), the results show that both pause frequency and pause duration have poor correlation with MOS for all the video content used. For example from Table 4.3, the pause frequencies of Videos 1 and 15 are identical, but their corresponding mean opinion scores (MOS) are 4.35 and 1.43, respectively. A similar result can be found by evaluating the average pause duration in Videos 6 and 15. In this case, the pause duration shows a variation of just 0.34 seconds, but the MOS varies greatly (3.42 and 1.43, respectively). Clearly, it is confirmed that pause frequency or pause duration alone does not provide a good correlation with viewer opinion.

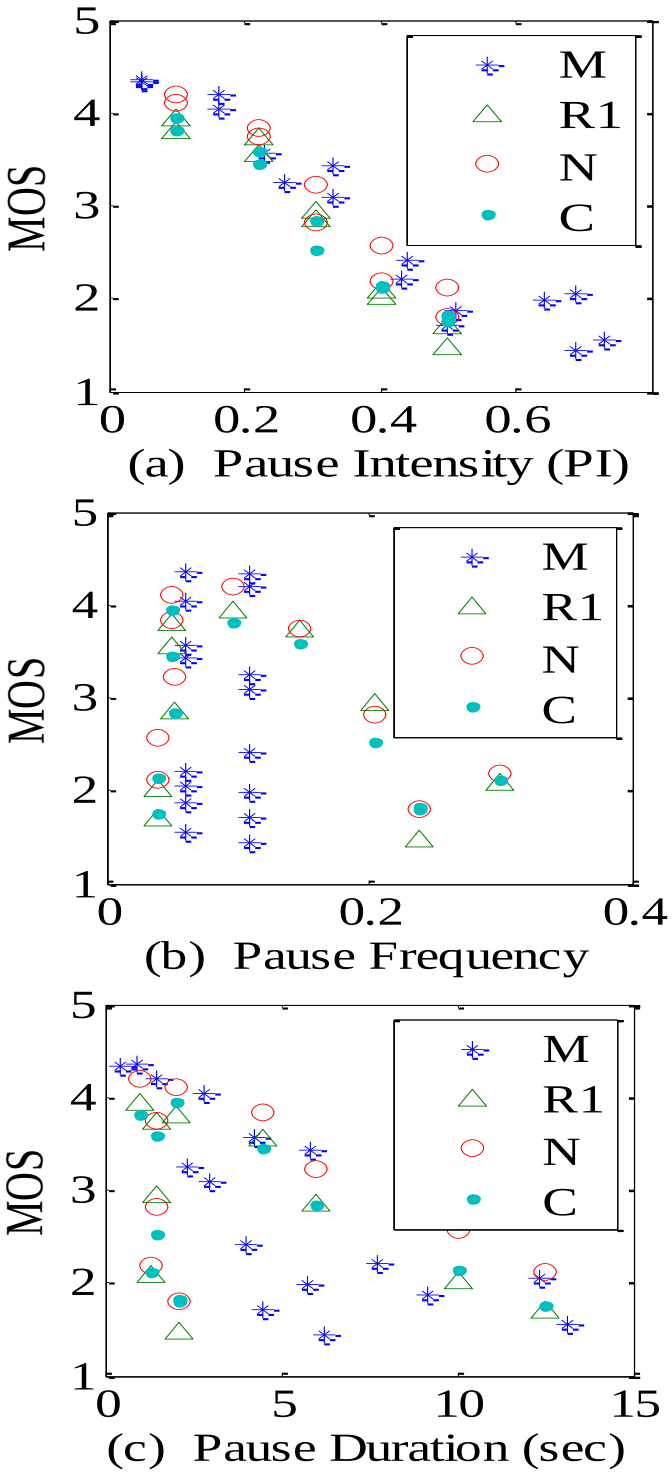


Figure 4.12 Results for Subjective Testing-1: MOS vs. (a) PI, (b) pause frequency, and (c) pause duration.

Table 4.3 Subjective Testing-1 Results

Video ID	Video content	PI	Pause Frequency	Average Pause Duration (sec)	MOS
0	M	0.05	0.06	0.93	4.35
1	M	0.05	0.11	0.43	4.34
2	M	0.16	0.06	2.8	4.03
3	M	0.16	0.11	1.46	4.19
4	M	0.23	0.06	4.22	3.57
5	M	0.26	0.11	2.31	3.24
6	M	0.33	0.06	5.87	3.42
7	M	0.33	0.11	3	3.08
8	M	0.43	0.06	7.72	2.21
9	M	0.44	0.11	3.97	2.41
10	M	0.51	0.06	9.2	1.88
11	M	0.5	0.11	4.47	1.72
12	M	0.69	0.06	12.33	2.05
13	M	0.64	0.11	5.79	1.99
14	M	0.73	0.06	13.16	1.55
15	M	0.69	0.11	6.21	1.43
16	R1	0.097561	0.04878	2	3.8
17	R1	0.097087	0.097087	1	3.933333
18	R1	0.219512	0.04878	4.5	3.55
19	R1	0.220588	0.147059	1.5	3.733333
20	R1	0.305344	0.050891	6	2.85
21	R1	0.306122	0.204082	1.5	2.933333
22	R1	0.401606	0.040161	10	2
23	R1	0.401198	0.299401	1.34	2.066667
24	R1	0.5	0.04	12.5	1.7
25	R1	0.5	0.238095	2.1	1.466667
26	N	0.097561	0.04878	2	4.11
27	N	0.097087	0.097087	1	4.1875
28	N	0.219512	0.04878	4.5	3.83
29	N	0.220588	0.147059	1.5	3.75
30	N	0.305344	0.050891	6	3.22
31	N	0.306122	0.204082	1.5	2.8125
32	N	0.401606	0.040161	10	2.56
33	N	0.401198	0.299401	1.34	2.1875
34	N	0.5	0.04	12.5	2.11
35	N	0.5	0.238095	2.1	1.8125
36	C	0.097561	0.04878	2	3.95
37	C	0.097087	0.097087	1	3.823529
38	C	0.219512	0.04878	4.5	3.45
39	C	0.220588	0.147059	1.5	3.588235
40	C	0.305344	0.050891	6	2.85
41	C	0.306122	0.204082	1.5	2.529412
42	C	0.401606	0.040161	10	2.15
43	C	0.401198	0.299401	1.34	2.117647
44	C	0.5	0.04	12.5	1.75
45	C	0.5	0.238095	2.1	1.823529

In addition, the Pearson Correlation Coefficient or simply correlation coefficient (*r*) is also adopted to evaluate the correlation performance of PI, pause frequency and pause duration with MOS, respectively, as shown in Table 4.4. The correlation coefficient, *r*, for all the types of video sequences is calculated based on the results in Table 4.3 and Table 4.5, and using the formula given below [100].

$$r = \frac{\sum(x - \bar{x})(y - \bar{y})}{\sqrt{\sum(x - \bar{x})^2 \ \sum(y - \bar{y})^2}} \tag{4-17}$$

where *x* and *y* are two sets of data, one representing PI, pause frequency, or pause duration and the other representing MOS; $\bar{x}$ and $\bar{y}$ are the mean of *x* and *y*, respectively. The correlation coefficient is used here to examine the linear relationship between PI, pause frequency or pause duration and MOS.

The correlation coefficients in Table 4.4 clearly show that PI has a consistent high correlation with the viewer's experienced quality while the performances of pause frequency and pause duration are inconsistent and they have low correlation in most cases, especially for pause frequency. We have

Table 4.4 Pearson Correlation Coefficient (r)

		Pause Frequency	Pause Duration	Pause Intensity
Subjective Testing -1	**Moto GP (M)**	-0.040	-0.760	-0.953
	Run (R1)	-0.316	-0.505	-0.972
	News (N)	-0.470	-0.381	-0.973
	Cartoon (C)	-0.355	-0.499	-0.979
Subjective Testing -2	**Rally (R2)**	-0.366	-0.254	-0.923

Table 4.5 Subjective Testing-2 Results

Video ID	PI	Pause Frequency	Average Pause Duration	MOS
0	0.10	0.01	11.76	3.76
1	0.10	0.09	**1.08**	**3.67**
2	0.22	0.02	9.29	3.93
3	0.22	0.19	**1.17**	**3.79**
4	0.29	0.25	**1.17**	**2.72**
5	0.31	0.03	12.00	3.00
6	0.31	0.03	12.52	3.09
7	0.33	0.31	**1.08**	**2.68**
8	0.40	0.30	**1.33**	**1.77**
9	0.42	0.02	18.32	1.93
10	0.47	0.02	25.98	1.59
11	0.50	0.33	**1.50**	**1.65**

also examined the Spearrman's rank correlation coefficient for this work and found that it gives very similar results to those produced by applying the Pearson correlation coefficient.

4.5.3 Subjective Assessment-2 for Pause Intensity

The second subjective testing session allows for stress testing of the pause intensity metric, where similar PI values are used with vastly different compositions. As shown in Table 4.5, the bolded values represent the sequences with a high pause frequency and shorter pause duration while the opposite scenarios are represented by the non bolded values. Given these extreme cases of pause frequency and pause duration combinations whilst maintaining similar PI values, the aim of this testing session is to explore the suitability of pause intensity in these scenarios. Based on Table 4.5 and the resulting Figure 4.13, we are able to assess the correlation performance of various buffer characteristics with the perceived viewer quality (MOS), in the same way as for subjective assessment-1.

Again, it is evident from Figure 4.13(a) and Table 4.5 that a clear correlation between MOS and PI can be obtained even with the huge difference in pause characteristics composition. For example, both Videos 2 and 3 in Table 4.5 have the same PI value of 0.22 and receive very similar MOS ratings, in spite of the very different pause frequency and duration compositions in these cases. The respective average pause frequency in Video 3 is 0.19, much higher than that in Video 2 which is just 0.02; while the average pause duration of Video 3 is around 1.2 seconds, much shorter than 9.3 seconds for Video 2. This property is also agreed by the Pearson Correlation Coefficient indicated.

Figure 4.13(b) shows, however, that the quality of experience of viewers as per MOS is inconsistent with the values of pause frequency. Videos 2 and 10, for example, have the same pause frequency (0.02), but their MOS values (3.93 and 1.50) do not match by a big margin, according to Table 4.5.

Figure 4.13(c) also indicates the shortcomings of using pause duration as a quality metric. Five videos (1, 3, 4, 7, 8) all have pause durations of around 1.1 seconds but the viewers' quality of experience varies greatly. It is also noticed that although the MOS ratings of Videos 0 and 1 are very close (3.76 and 3.67), their average pause durations are so different (11.76 and 1.08), showing no correlation between them.

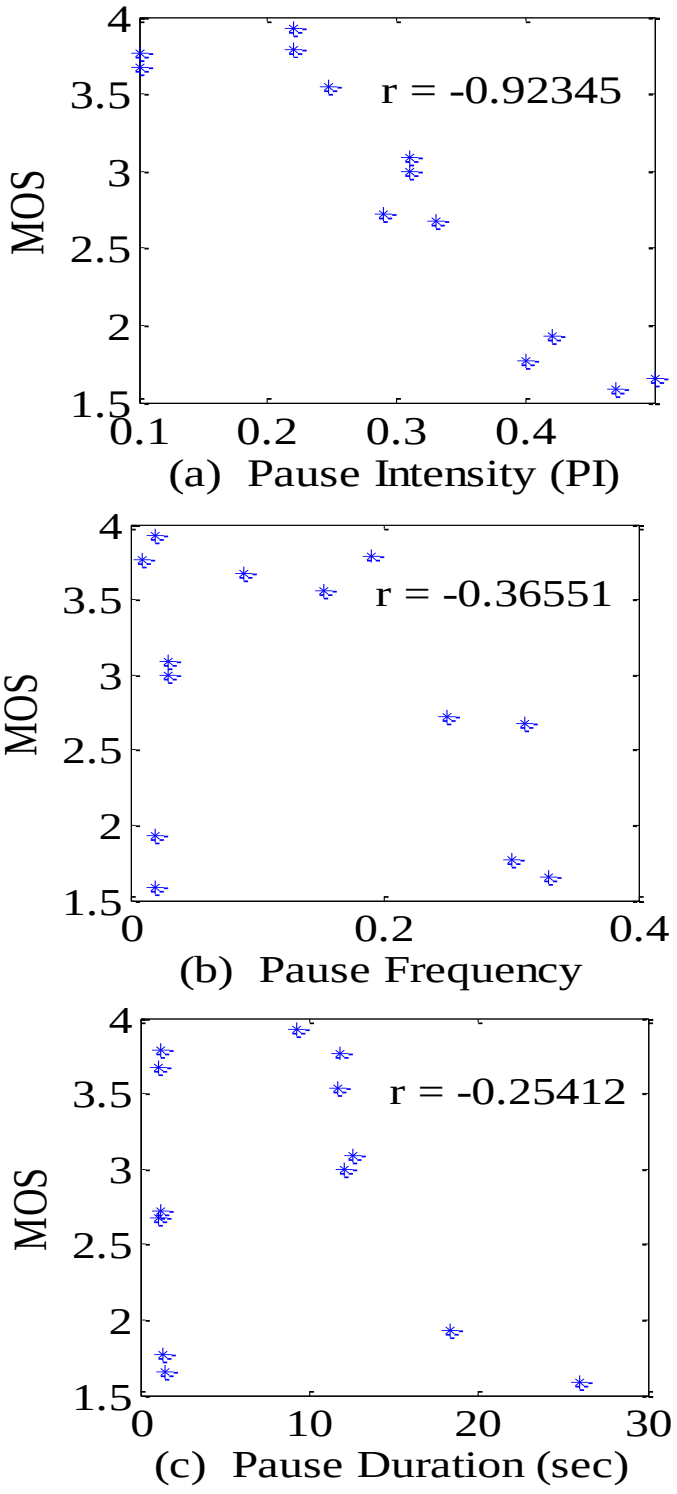


Figure 4.13 Results for Subjective Testing-2: MOS vs. (a) PI, (b) pause frequency, and (c) pause duration.

Looking back at Table 4.4, the correlation coefficient for PI in Subjective Testing-2 overwhelmingly outperforms the other two pause characteristics. In addition, although pause duration has shown some correlation with varied levels in Subjective Testing-1, this performance is inconsistent with testing scenarios such as in Subjective Testing-2 where a very low correlation is recorded for pause duration.

All the results have demonstrated that PI performs consistently and holds a linear relationship with MOS in varied testing scenarios and this relationship is content independent.

4.6 Summary

This chapter has explored the pause intensity metric in the context of video streaming in a TCP network, and provided a detailed analysis of the advantages of PI over other existing metrics that characterize buffer underrun. The analytical model developed reveals that the PI metric can be determined by both the video playout rate and network throughput and, more precisely, it is a ratio of the rate difference (λ-η) to the playout rate λ.

We have verified the model through extensive simulation using NS-2 and MATLAB, which demonstrates the high accuracy of the model established. We have also provided subjective testing results based on 57 video streaming clips, signifying that the PI metric has a very good correlation with the viewer's QoE in terms of the MOS ratings. In addition, we have shown that either pause frequency or pause duration is not sufficient on its own to reflect the perceived video playback quality by viewers. The results from subjective tests have confirmed the independency and consistency of our metric through testing different types of video clips and allowing vastly different compositions (constituted by both pause frequency and pause duration) for defining PI values.

Being a no-reference metric and related to network performance and system settings, PI values can be easily made available for network operators and services providers to predict user's QoE even before the video is actually played out at the receiver. As such, the conventional control mechanism such as rate adaptation can then be used more effectively and in a quality-guided manner. Further exploitation of this work will involve applying the higher-order factor analysis to the PI metric and realizing its benefit for real systems.

Chapter 5
Quality-Driven Scheduling for Video Streaming in LTE

5.1 Introduction

Although the upcoming generations of the mobile communication systems bring about a much greater network capacity, the user's demands for better and diverse applications also increase proportional to the capacity. This fact presents the challenge to the new era of mobile communications in terms of the improvement in user's satisfaction level or Quality of Experience (QoE) during the process of resource allocation to end users by the network. The subjective nature of QoE makes this task even more challenging.

In most cases, a proposed solution for improving QoE eventually ends up with a Quality of Service (QoS)-oriented solution that either lacks any relation with subjective quality measurement or needs an original reference (e.g. standard PSNR (peak signal-to-noise ratio) in video assessment). With regard to video streaming in the TCP/IP network, for instance, standard PSNR is an unsuitable metric for quality assessment due to fidelity related impairments being potentially recovered by the transport protocol; and as explained in Chapter 4, other metrics such as Pause Frequency are not properly correlated with subjective quality measurement [101].

In this chapter a quality-driven scheduling framework is proposed, aiming to address these problems and provide a sufficient access to the QoE criterion perceived by users in the implementation of the resource management functionality in the mobile communication system. Pause Intensity (PI) as an objective and no-reference quality assessment metric will be used to represent the subjective quality of the service in the network. The relation between PI and QoE in terms of the Mean Opinion Score (MOS) will be explained based on the fundamental models presented in Chapter 4. The framework is capable of working as a mediator between the user, network

(e.g. scheduler in Long Term Evolution (LTE)) and server to tailor the allocated resource or original offered data rate to achieve the required QoE.

PI is a metric for video stream quality without the need for decoding results at the receiver and can be worked out at various levels in the network provided that throughput and required average video bitrate are known. This is feasible on both the user side and last mile access point of the network (e.g. eNodeB in LTE) and provides a noticeable flexibility in the quality management.

The proposed method extends the scope of conventional scheduling schemes, taking into account both network performance and fairness as well as the user's demand. A PI-based scheduler not only addresses the unbalanced allocation to the users with poor channel status (i.e. lower SNR) but also is aware of the correlation between the required and the allocated data rates per user. As will be explained in Section 5.2, a PI-based scheduler can also reflect the need of the network efficiency.

The rest of the chapter is organized as follow. Related work and required background are explained in Section 5.1 and the proposed framework including the PI-based optimization system and an algorithm for its implementation are discussed in Section 5.2. In Section 5.3 the simulation results and performance of the framework will be discussed. Finally, Section 5.4 provides the conclusion.

5.2 Related Works and Background

5.2.1 QoE Evaluation

Service quality improvements through the optimization of throughput, delay and error rate (i.e. bit error rate, block error rate or packet loss) characterize mobile communication performance fairly well and have been addressed in the literature frequently [102]. They have been used in 3GPP LTE to standardize the thresholds for QoS at the level of bearers [103]. However, in spite of their importance and defined relation with the service quality through the QoS quantities, these parameters do not represent the actual quality experienced by the user. For this reason, QoE and an applicable metric with a defined relation with the perceived quality need also to be addressed.

In [104] an approach is proposed for resource allocation in LTE based on the user feedback, where PSNR is assumed to have a linear relation with MOS. Required metric is computed at the source and signalled along with the video if it is to be used in the network. A similar work uses PSNR for QoE representation directly [105]. The standard PSNR approach requires an original reference which is not accessible in a real-time and online evaluation procedure. In addition, it also needs decoding, recording and comparing the received video with the reference, making it irrelevant for online and reactive quality control.

In [106] the percentage of rebuffering has been defined as a metric for user satisfaction. This metric has been used to evaluate the capacity of LTE to support users with a certain level of QoE. In [107] QoE is represented by the number of pause interruption and a scheduling algorithm has been proposed to improve the QoE considering the buffer level at the receiver.

Actually, the concept of discontinuity of the service which is the main concern of this work, has been studied under various service and buffer related subjects such as buffer underrun, jitter, initial delay of service, buffer starvation, pause frequency etc. These parameters are mostly dependent on the user side's buffer setting. Furthermore, the exact correlation between the chosen QoE metric and user's satisfaction has not been investigated explicitly. As it will be explained in the following sections, we define the correlation as the relationship between the required and allocated data rates per user, which characterize how users' satisfaction (i.e. QoE) are addressed by the network's resource supply.

As introduced in Chapter 4, PI defines a quality assessment metric to characterize the playout buffer behavior and quantify the discontinuity impairment of video streaming services simplified in a closed form which resembles an M/M/1 process per user. Discontinuity is the main quality concern by the end user when the TCP transport protocol is used. PI takes both pause frequency and pause duration into account and can also be determined by network throughput, η, and required encoding rate (user demand), λ, according to (4-14):

$$\begin{cases} PI(\eta,\lambda) = 1 - \dfrac{\eta}{\lambda} \\ \eta \le \lambda, \qquad 0 \le PI \le 1 \end{cases} \tag{5-1}$$

Given a non-recorded streaming scenario, throughput is equal to or less than the required rate and the value of PI remain in the range of 0 and 1. A client adaptation algorithm which tries to increase the buffer fullness by requesting more data can reach λ as an upper bound.

PI does not rely on the buffer setting (e.g. playback buffer size), and can be assessed by users, given the network performance. Actually since PI is about the instantaneous data consumption and the fluctuation of the buffer occupancy, it gives a prior picture about the probable quality concerning the discontinuity. This attribute helps the proposed framework to calculate the metric wherever the information of current values of η and λ are accessible. The relation between PI and QoE in the form of MOS has been shown in Figure 5.1. This subjective testing result reveals the close correlation between PI and QoE with a correlation coefficient close to -1, as shown in Subsections 4.5.2 and 4.5.3. Moreover, it provides explicit quantitative measurements for exhibiting the discontinuity of streaming video replay, given the network performance (η) and a specific video bitrate (λ).

5.2.2 LTE Environment

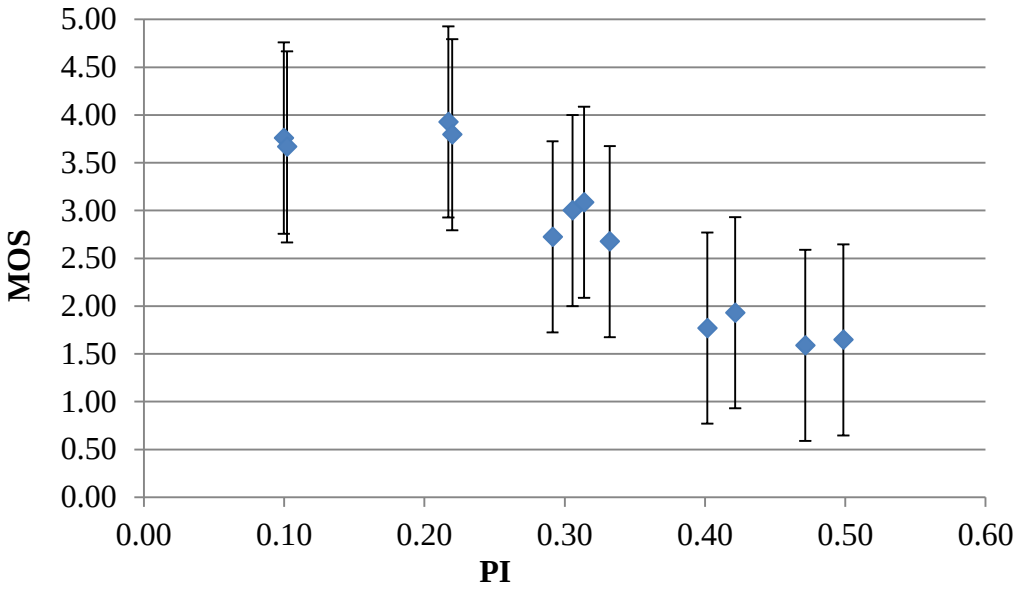


Figure 5.1 The relation between Pause Intensity and MOS

In this Subsection the proposed framework is explained in the context of LTE which has been standardized by 3GPP Rel.8 [13] and later extended to LTE-Advanced to fulfil the latest requirements of the next generation of mobile communication systems. Unlike previous generations of mobile systems, LTE has combined all of its radio access network functions into the base station known as eNodeB, which allows radio channel resources to be more efficiently managed.

Resource allocation and link adaptation in LTE are the two main functions which have been considered for QoE assurance in this work. The overall performance can be assessed comparing the resources allocated to each user with the total allocated resources in the system. As shown in Figure 5.2, PI is employed to balance the relation between resource allocation and link adaptation functions accompanied by scheduling algorithm and achieve a certain trade-off with regard to fairness, efficiency and the correlation between required and allocated data rates.

LTE is capable of providing an efficient resource management through the combination of flexible bandwidth from 1.4MHz to 20MHz, Orthogonal Frequency Division Multiple Access (OFDMA) and Adaptive Modulation and Coding (AMC) techniques. As it is illustrated in Figure 5.3, a Resource Block (RB) of (0.5ms, 180KHz) is defined as the resource allocation unit in LTE in a two dimensional time-frequency grid. The allocation in each 0.5ms (i.e. time slot) will remain the same for the next time slot and this makes a 1ms Transmission Time Interval (TTI) for each transmission process.

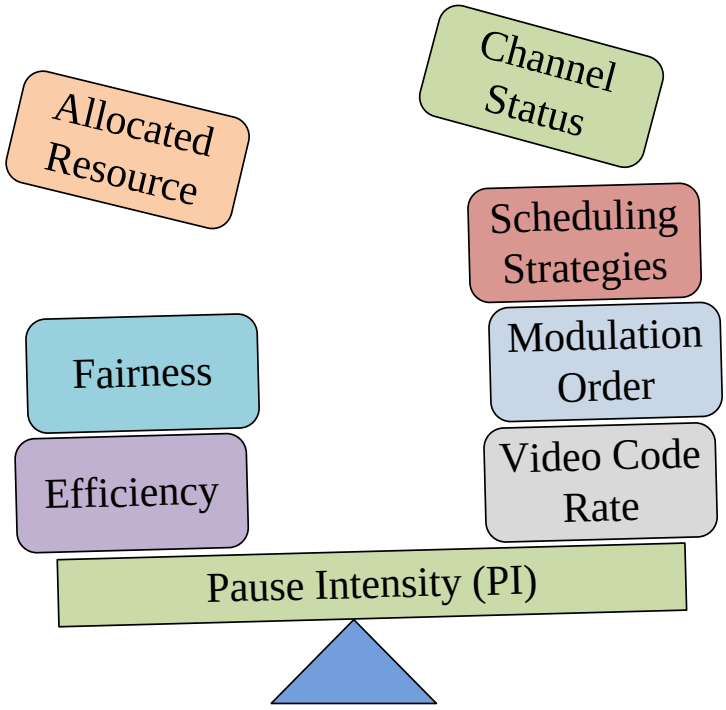


Figure 5.2 Contribution of PI in performance evaluation

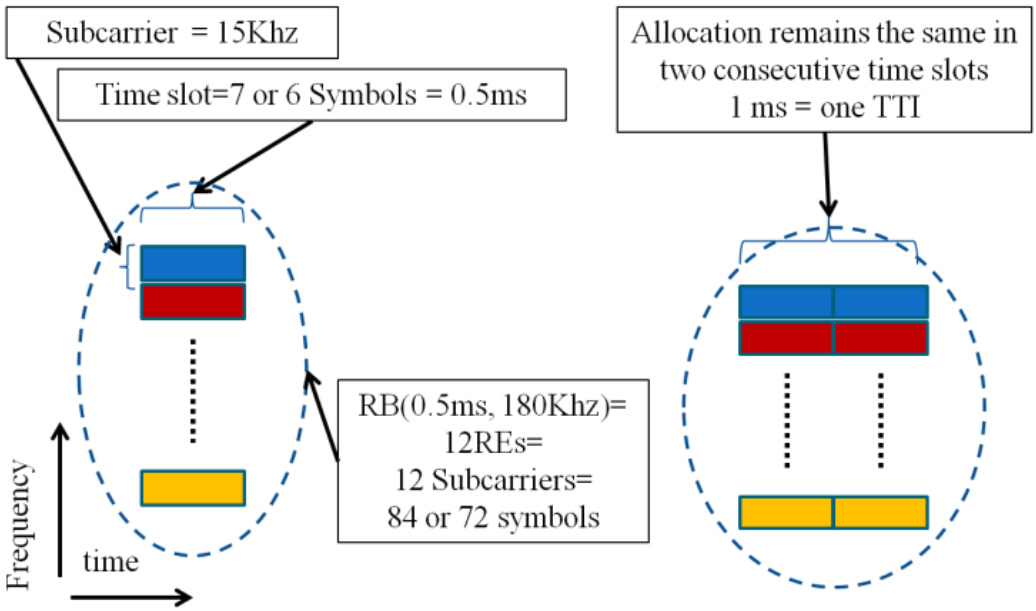


Figure 5.3 Resource Block as a unit of resource allocation in LTE

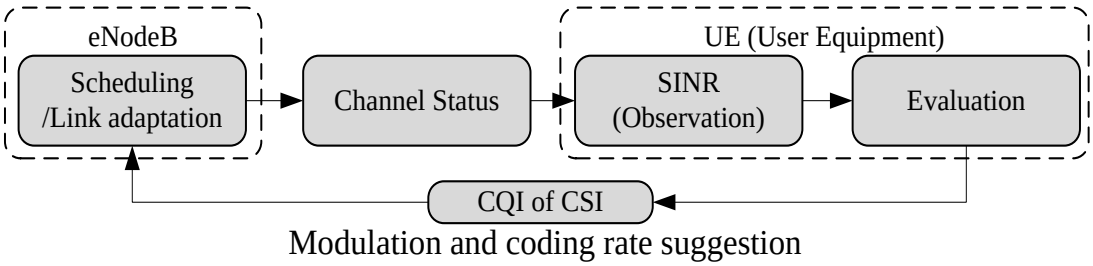


Figure 5.4 Channel quality feedback from user to eNodeB

The number of RBs depends on the employed bandwidth of the system and changes from 6 to 100 for 1.4MHz to 20MHz, respectively. A Resource Element (RE) of 6 or 7 symbols is the sub-unit inside each RB and occupies 15KHz of the allocated frequency resource as a subcarrier of the OFDMA technique for transmission. This RE conveys the user's data through applying the most suitable modulation and coding rate. The process is shown in Figure 5.4. The decision considering the modulation and coding scheme (MCS) is based on a channel quality report (CQI) from user to eNodeB which recommends the most suitable range of MCS for the next TTI to secure the minimum quality of 90% accuracy of the received blocks. In the CQI report, the capability of the users' device has to be taken into account too.

5.3 Proposed Framework

5.3.1 PI-based Optimization Problem

We have shown that PI is a function of video bitrate, λ, and throughput, η and can be interpreted as a resource utilization indicator based on the resource allocation parameters in LTE as follows [108]:

$$\mathrm{PI}(\eta,\lambda) = \mathrm{PI}(\eta(\mathrm{r_B},\overline{\mathrm{CQI}}),\lambda) \tag{5-2}$$

where throughput, η, is a function of the allocated resource blocks, r_B, and the selected modulation and code rate (through the average reported channel quality, $\overline{CQI}$).

The role of PI in an optimized resource allocation process can be described as:

$$\begin{cases} {r_B}^{opt} = \underset{r_B}{argmin}\, PI \\ PI = 1 - \dfrac{\eta}{\lambda} \\ 0 \le PI \le 1 \end{cases} \tag{5-3}$$

where r_B is the vector of the allocated resources (i.e. allocated RBs in LTE) and minimum values of Pause Intensity are sought for achieving the highest possible QoE for users (lower PI means higher QoE). Obviously due to the Pareto optimality of the problem any improvement in a user's QoE leads to lower QoE for other users [109]. This suggests recasting the problem in a way that guarantees a minimum quality for all users. More precisely, letting **PI** be a vector of PIs in a system with N users, we can rewrite (5-3) as:

$$\begin{cases} {r_B}^{opt} = \underset{r_B}{argmin} \max \boldsymbol{PI} \\ \boldsymbol{PI}_{1\times N}: (PI_1, PI_2, \dots, PI_N) \\ PI_k = 1 - \dfrac{\eta_k}{\lambda_k}, \qquad 0 \le PI_k \le 1 \quad \forall k \end{cases} \tag{5-4}$$

The problem arrangement in (5-4) minimises the worst case PI and considering the relation between PI and throughput and the fact that we discuss a situation in which throughput is less than the required data rate, λ, suggests a *MaxMin* interpretation of the problem as below:

$$\begin{cases} {r_B}^{opt} = \underset{r_B}{argmax} \min \boldsymbol{\xi} \\ \boldsymbol{\xi}_{1\times N}: (\xi_1, \xi_2, \dots, \xi_N) \\ \xi_k = \dfrac{\eta_k}{\lambda_k}, \qquad 0 \le \xi_k \le 1 \quad \forall k \end{cases} \tag{5-5}$$

where ξ represents the ratio of the throughput over the required data rate for each user. Throughput in (5-5) is a consequence of the allocated data rate in scheduler, given the user's channel status, modulation and coding rates. To evaluate and compare the PI values in each round in the scheduler (i.e. at eNodeB) the weighted average of the allocation will be considered as:

$$R_{i,k} = \left(1 - \frac{1}{t_w}\right) R_{i-1,\mathrm{k}} + \frac{1}{t_w} r_{\mathrm{i,k}} \tag{5-6}$$

where $R_{i,k}$ is the average allocated data rate to user k including allocation in current round (the *ith* round*)*, and $r_{i,k}$ is the latest allocation to user k (i.e. the allocated data rate if the available resources is given to user k). The average allocated data rate up to the current TTI and the average window size (i.e. degree of weighting) are represented by $R_{i-1,k}$ and t_w, respectively.

The main difference between (5-5) and a standard *MaxMin Throughput* optimization problem is the contribution of the user's required data rate, λ. This makes the fairness benchmark tending toward the ratio ξ instead of throughput η. Furthermore, the value of PI is calculated based on the average throughput derived from (5-6) which reveals a similarity between the proposed method and a

standard *Proportional Fair* scheduling method. In both methods the current allocation alongside the history of the previous allocations are taken into account to prevent excessive allocation to the user. Later in Section 5.4 the results of PI optimization problem in (5-3) - (5-5) will be compared with the standard *MaxMin Throughput* and *multicarrier proportional fair* schedulers [110].

5.3.2 An Efficient PI-based Scheduling Algorithm

Due to the properties mentioned above, the proposed method is expected to be proportionally fair with the resource allocation and correlated with the user's demand. However, the effort to reduce the PI value (i.e. improve the QoE) will reduce the efficiency of the system (i.e. reduces the total capacity of the system). This is similar to the well-known attribute of the standard *MaxMin Throughput* [111] and a result of the trade-off between the achieved fairness and efficiency. Therefore, the optimization system in (5-5) is expected to be fair with a high degree of correlation between the required and allocated data rates per user, although it is not aimed to optimize the system capacity or efficiency. We aim to address the issue of efficiency enhancement through the implementation algorithm of (5-3)-(5-5) while keeping both the fairness and correlation levels relatively high.

The algorithm, directly driven from the PI-based optimization system introduced in the previous subsection, may select the user with the highest PI in each round of the allocation, aiming to reduce its PI for the forthcoming time slots. A high PI value can be due to the high demand of the user (i.e. users with the higher video bitrates) or its bad channel status and low efficiency. In the case of low efficiency, users with PI close to 1 are likely to use the allocated resources in an inefficient way (by adopting low order modulation schemes and low channel coding rates). It is of course unlikely to achieve a high MOS score. In contrast, users with PI close to 0 are more likely to use their allocated resources efficiently through using high order modulation schemes and high coding rates even though fewer resources are allocated to them.

Therefore, in a PI-based scheduler, to guarantee an acceptable degree of system efficiency, users' efficiency must be considered alongside the latest PI values during the prioritization process in each allocation round. Subsequently a more efficient algorithm for the implementation of the proposed PI-based optimization system can be defined as:

$$\begin{cases} \mathrm{k}^* = \underset{\mathrm{k}}{argmax}\, \zeta_k \\ \zeta_k = r_{i,k} PI_{\mathrm{k}}, \quad \forall k \\ r_{i,k}\colon \text{ the } i^{th} \text{ allocated data rate} \\ \quad \text{if resource given to user } k \end{cases} \tag{5-7}$$

where both the efficiency of the user (given its current channel status), and its latest Pause Intensity value have been taken into account. For an inefficient user, it would have a better chance to be

Algorithm 5.1: PI-based scheduling algorithm (eNodeB)

1: ***for*** *all time-slots*
2: Provide a list of users with HOL packets
3: ***while*** *(RBs available)*
4: Choose the best MCS based on reported CQI
5: Calculate $\boldsymbol{r}$ *as in (5-7) given the channel status and resources*
6: Calculate ζ *as in (5-7) for all users in the list*
7: Rank the users based on their ζ *as in (5-7) and choose* $\boldsymbol{k^*}$
8: Calculate the produced capacity
9: Update all $\boldsymbol{R_{i,k}}$ *values as in (5-6) and estimate throughput*
10: Update all PI values
11: Update the list of non-allocated resources (available RBs)
12: Update the list of the users with HOL packets
13: ***end***
14: ***end***

selected for allocation due to its high PI value, while its low efficiency can deprioritize itself in the scheduling list. The implemented *PI-based scheduling algorithm* based on (5-7) has been summarized in Algorithm 5.1.

In the next section the simulation results of the PI-based optimization method and its modified algorithm are presented in comparison with other well-known scheduling methods.

5.4 Simulation Results and analysis

PI-based scheduling aims to provide a controlled QoE with consideration of both fairness and efficiency, together with the correlation between the required and allocated data rates per user being taken into account. In this section the proposed PI-based approach for achieving this goal is examined in the context of an LTE downlink scheduler with system setup, simulation results and their analysis.

5.4.1 Simulation Setup and Methodology

Table 5.1 shows the settings of a simulator which has been developed in MATLAB to examine the proposed algorithm for the optimization problem defined. The source of the users' data are the video stream data packets generated using a truncated Pareto model (for packets' inter-arrival-time and size) without any other background traffics. Data rates are chosen in a range suitable for achieving a standard quality of current video services (e.g. BBC-iPlayer). Users' Head-of-Line packets (HOL) are scheduled in a timely manner with no packet drop due to the delay in eNodeB. We choose the average window size, t_w, defined in (5-6) as 100ms (higher than LTE frame length, 10ms) to filter out the fluctuation of the average allocated resources, which is also long enough to capture the average video bitrate of user's data. The CQI mapping table and channel status generator presented in [111, 112] have been used in our simulator. To maintain the consistency of the results,

each video bitrate is corresponding to more than one user with different SNRs in the range of the defined CQI of LTE (for CQI=1~15). Users are distributed in one cell and the interference from the surrounding cells has been considered. The shadowing effect considering the inter/intra-cell spatial correlation has also been considered.

The proposed PI-based optimization system is compared with the standard *MaxMin Throughput*. In both cases the binary linear program with integer relaxation has been used to solve the problem, given the constraints driven from LTE's bandwidth and the number of Resource Blocks (Table 5.1). The performance of the suggested PI-based scheduling algorithm is compared with the standard *maxCI* (i.e. *best-CQI* in LTE) and multicarrier *Proportional Fair* schedulers [110] to demonstrate their differences in terms of resource allocation efficiency, fairness and the correlation between the required and the allocated data rates (per user).

The detailed performance analysis is provided in the next subsection, where the efficiency is represented by the total capacity of the system (the summation of the allocation of all of the users), the fairness by Jain's Index and correlation by Pearson's Linear Correlation Coefficient [100] with the input of the required and allocated data rates.

5.4.2 Results and Analysis

Figure 5.5 shows the overall comparison between the proposed PI-based optimization system in (5-3)-(5-5), its analogous *MaxMin Throughput* system, the modified PI-based optimization system given in (5-7), multicarrier proportional fair and best-CQI scheduling methods. To make a joint comparison, values are normalized to unify the scale of the diagrams. The proposed PI-based optimization system has the highest correlation between the required and allocated data rates per user

Table 5.1 Simulation Setup

Parameter	value
No. of Cells	1 (with the first tier interference)
Inter-site distance	2000 meters
Shadowing effect	mean=0, deviation=8 decorrelation distance=25meters, inter-site correlation=0.5
Channel model	PedA, speed=3km/h
Bandwidth	5MHz
No. of RBs (per TimeSlot)	25
Subcarrier	15KHz
Range of average SNR	-6 ~ 17 dB (CQI=1~15)
Average video code rate	200, 400, 600, 800, 1000 Kbps
No of Users	45
Each scheduling round	One TTI=1ms
Simulation time	5000*TTI
Video stream model	Truncated Pareto for packet size and inter-arrival time

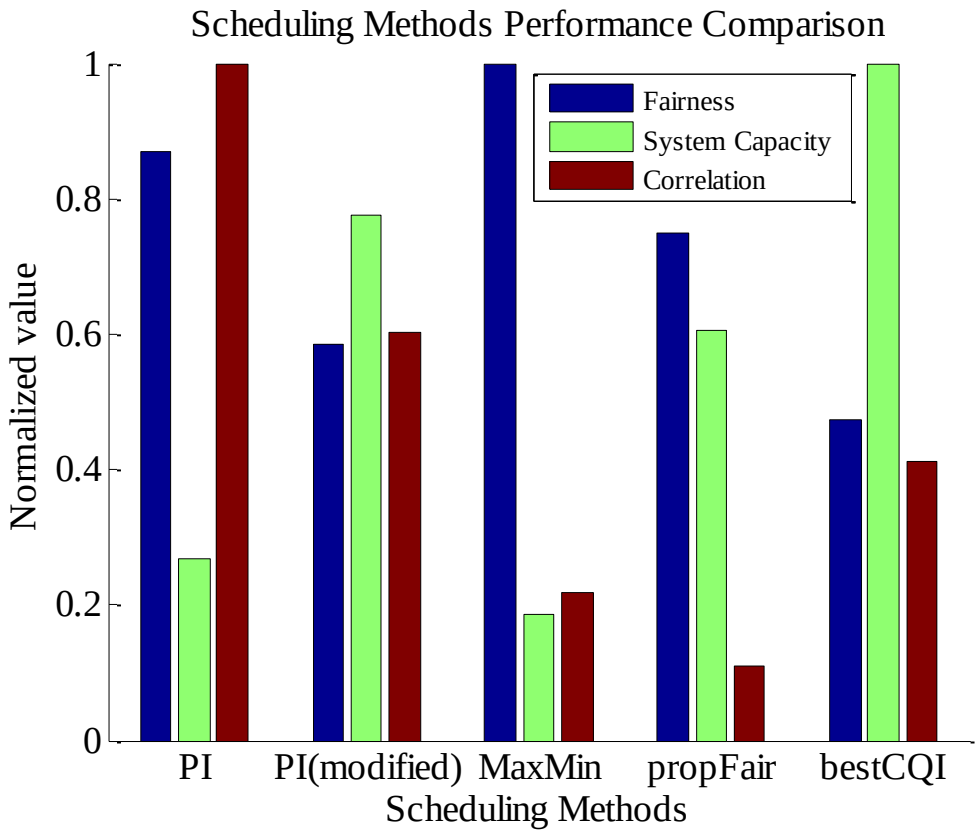


Figure 5.5 Proposed method's overall comparisons

and is capable of achieving a high degree of fairness comparable to the standard *MaxMin Throughput*. However the total capacity of the system is lower than the well-known scheduling algorithms such as Proportional fair and best-CQI.

The implementation of the modified PI-based scheduler in (5-7) provides a higher efficiency with the cost of poor fairness and low correlation between the required and allocated data rates though it still performs better than other methods. However it performs as a trade-off between the proportional fair method and best-CQI methods. The proposed algorithm in (5-7) achieves a higher fairness, compared to that of the best-CQI, and higher efficiency, compared to that of the proportional fair method, while maintaining the dominancy of the correlation between the required and allocated data rates.

The results of the allocated data rate as a function of the required data rate (video bitrate) in Figure 5.6 gives a better insight into the above discussed correlation coefficients. Results reveal the video bitrate dependency in the case of the proposed system in (5-5) and modified algorithm in (5-7). Although the best-CQI seems to be similar to a video bitrate dependent method, a closer look into the allocation data rate per user reveals the dominancy of the users with high SNR in each case regardless of their actual video bitrate. This difference has been shown in Figure 5.7 in which the

most beneficiary of the best-CQI allocation are the users with high value of SNR though among the privileged users there will be a video bitrate dependency which appears in Figure 5.6.

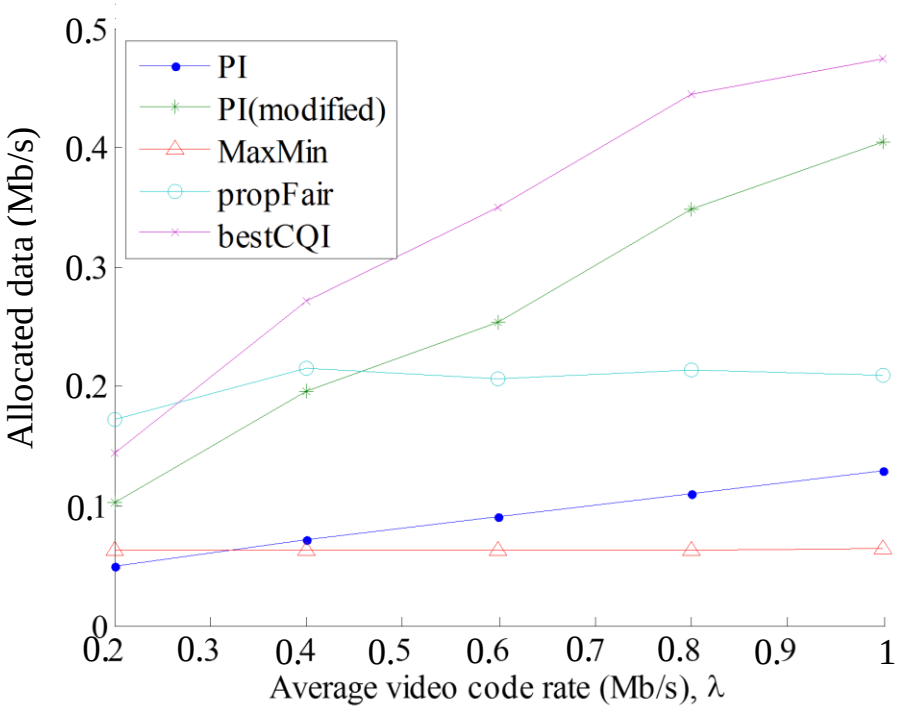


Figure 5.6 Dependency between the required and allocated data per user

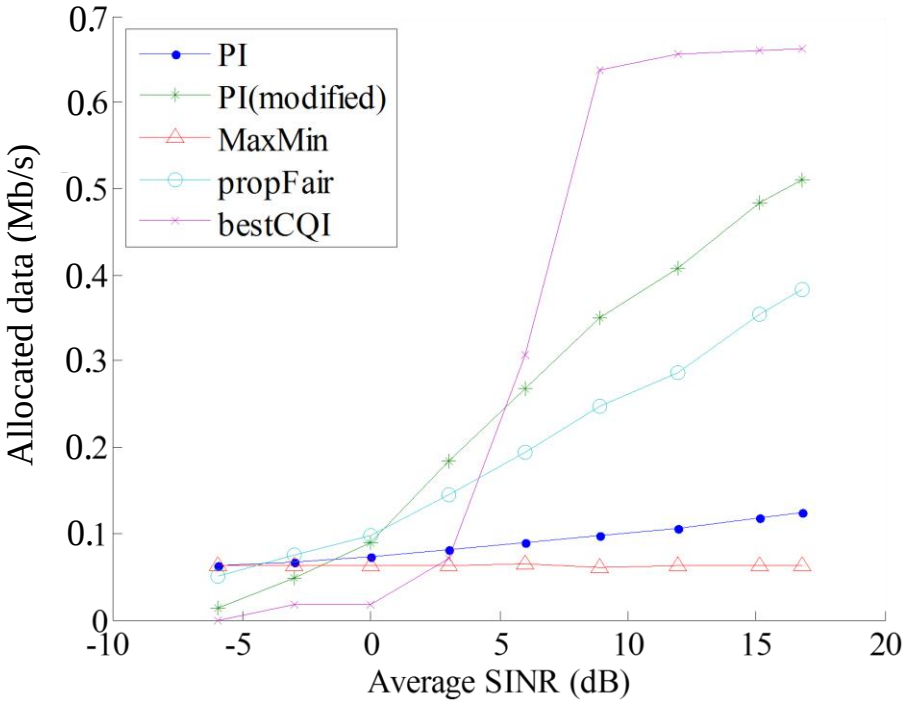


Figure 5.7 Dependency between the allocated data and user's channel status

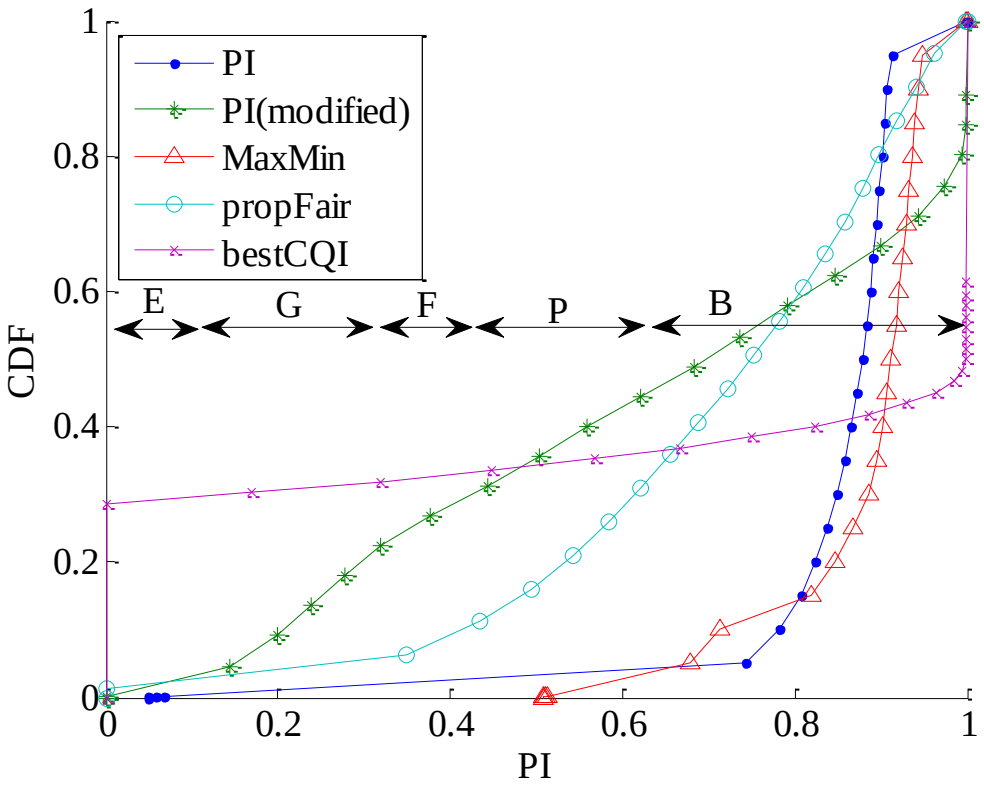


Figure 5.8 Achieved probability of Quality of Experience (QoE) for users with low SNR or high required data rate

The aim of the proposed method was to improve QoE through applying PI as it is shown that PI has a good correlation with MOS, demonstrated in Figure 5.1 and in Chapter 4. The achieved Cumulative Distribution Function (CDF) of QoE based on the distribution of achieved PI values for different scheduling methods is shown in Figure 5.8, especially featuring the users with low SNR (less than the 50th percentile or median of SNRs) or requiring high data rate (more than the 50th percentile or median of video bitrates). The performance of this group of users is usually hampered either by efficiency-oriented approaches (e.g. *best-CQI*) or fairness-oriented methods (e.g. *MaxMin Throughput*). In Figure 5.8, the areas labeled as E, G, F, P and B are related to the Excellent, Good, Fair, Poor and Bad scores of MOS, respectively. Obviously, if there is no concern about fairness, best-CQI can provide the highest probability of being in a satisfactory status of QoE represented by PI values in Figure 5.8. However, in the cases where fairness and/or correlation are also taken into account, the modified PI-based algorithm in (5-7) will provide the highest probability of being in a satisfactory level of QoE while maintaining the desired trade-off among fairness, efficiency and correlation.

Furthermore, best-CQI produces a flat CDF of PI for the above mentioned users. This means that some users receive a very good service and the rest of them experience a very poor service with a small likelihood to be in an intermediate situation. In the case of the proposed PI-based algorithm in (5-7) reasonably good services is achievable for most users in the network under various users' channel conditions.

5.5 Summary

In this chapter a framework has been proposed to adopt an objective, no-reference quality metric, Pause Intensity, in resource allocation in order to improve user's QoE in mobile wireless networks. To address this issue effectively, we have defined and investigated the relationship between the demand from users (QoE) and the provision by the network (resource allocation) in terms of the correlation of the two data rates: required and allocated. The proposed algorithm has been examined in the context of LTE scheduling and link adaptation schemes. The results have revealed the effectiveness of the proposed PI-based framework in terms of the improved QoE with tailored proportions of fairness, efficiency and correlation. The results have also been compared with other scheduling methods: multicarrier Proportional Fair and best-CQI, to exhibit the ability of the new scheme to trade-off among these three parameters.

As it is shown, the scheme is able to reach the aimed capacity, fairness and correlation targets by considering user's PI alongside the capability of the user in utilizing the allocated resources dynamically in each allocation round. This scheme can also predict user's QoE status based on the rate required by the user and the rate provided by the network, so that appropriate adaptive rate control or resource allocation methods can be applied. The latter option can be easily achieved from the network side as PI can be directly determined based on the network performance and required service level.

Chapter 6
Adaptive Resource Allocation for QoE-Aware Mobile Communication Networks

6.1 Introduction

In spite of the extended capability of the modern communication technologies that support a wide range of communication services, ensuring a high level of quality of service (QoS) or quality of experience (QoE) for end users remains to be a big challenge for network operators and service providers. This problem is further intensified by the growing demands for video streaming over mobile smartphones and tablets due to the limitations inherited in wireless and mobile communications environments.

Maintaining a good balance between the quality of video replay and the resource requirement is one of the main hindrances in video streaming services. However, it is generally possible to compromise on the quality issue for less resource dedication. This is especially desirable in the case of wireless communications with scarce spectrum but facing a high demand for mobile video services [113]. The current adaptive streaming service is an example of handling this trade-off, where multiple versions of the same video content with different video bitrates are made available for different user conditions and requirements.

In the 3GPP-DASH (Dynamic Adaptive Streaming over HTTP) standard [20], clients choose the code rates of the video content from the server (client-pull), without the intervention of the intermediate unit of the network, e.g. the base station in a mobile network. Collaboration between the base station and either side of an end-to-end video streaming system (server or client) can enhance the experience of clients being served in terms of their perceived quality. But this may entail extra information exchange among them and does not comply with the idea of the independent-client based adaptive service such as DASH. Furthermore, it may require additional processing overhead and standardization amendment which practically can be a limitation for the implementation of this idea.

To tackle this issue, a quality of experience (QoE) driven resource allocation scheme with scheduling algorithms for the last-mile scheduler is proposed in this work based on the video quality assessment metric, i.e. Pause Intensity (PI), which takes account of user's video bitrate (required data rate) and network performance (throughput) and can realistically characterize the demand-supply relationship of video streaming services, as discussed in Chapters 4 and 5. The proposed scheme provides the capability of online adjustment of system efficiency, fairness and correlation between the required and allocated data rates. The PI metric can be easily assessed by the scheduler on the network side without requiring extra information exchanged between users and the network. In addition, PI can also play a role in shaping the distribution of video bitrates for adaptive video streaming and reaching the required level of QoE for clients. The proposed algorithms are examined in the context of 3GPP-LTE (Long Term Evolution [13]) for both adaptive and non-adaptive video streaming scenarios, complied with the 3GPP and related standards for streaming services [9, 114].

It must be noted that, although the buffer fluctuation range is one of the elements involved in the PI derivation analysis, the final achieved formulation is merely based on the network performance (throughput) and required data rate (video bitrate). In the case of getting more robust to network performance fluctuation through buffer size control, the improvement will not necessarily result in less pause occurrence and will not be reflected in PI. However it will be captured by pause duration and pause frequency as explained in Chapter 4.

The rest of the chapter is organized as follows: The background and related works are explained in Section 6.2. The model description, proposed optimization system and its implementation algorithm are presented in Section 6.3. The simulation results and analysis are discussed in Section 6.4 and finally the summary is provided in Section 6.5.

6.2 Background and Related Works

This section discusses the current developments and related works including the overall structure of a content distributer and the nature of impairments in video transmission between the client-side and the network elements of a contemporary mobile communication system (i.e. under the 3GPP standards), in connection to the application of quality assessment.

6.2.1 Content Distribution Solutions

A general infrastructure model of the most common video streaming service including a client-side, a content source and its transmission channel is depicted in Figure 6.1. A global access to this service involves some elements from public networks, the operator's infrastructure and the user's last mile access interfaces. Subsequently, the performance improvement of the service requires a global consideration of these elements and protocols of intermediate networks [11].

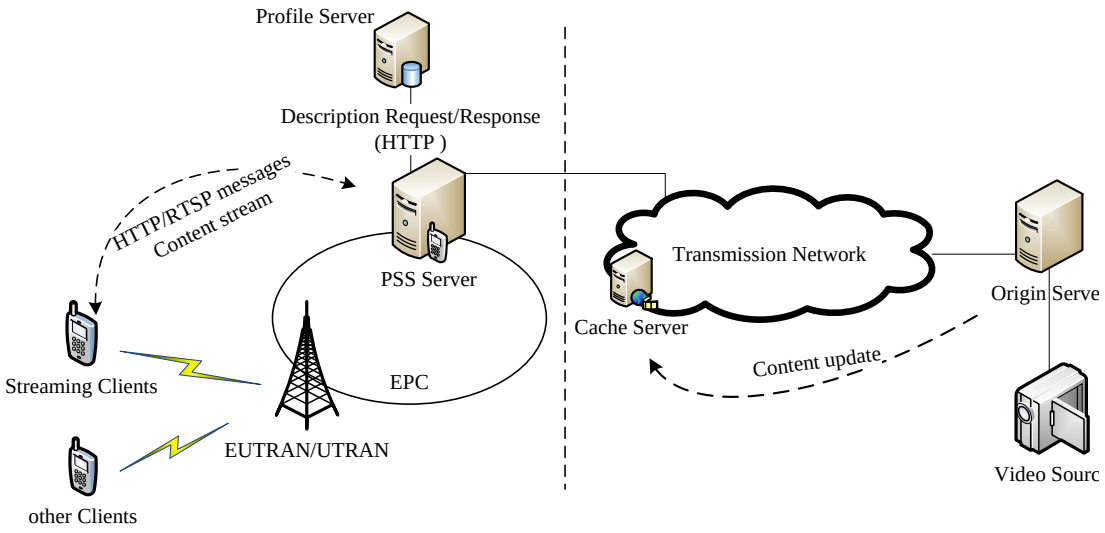


Figure 6.1 The network infrastructure of a globally accessible video streaming service over mobile communication and through a distributed network

The distributed system solutions, such as Content Delivery Networks (CDN), are widely used to reduce the complication in quality control for such a global service. A distributed system links up the locally manageable resources with the corresponding globally accessible services. For example, in Figure 6.1 the nearest ‘Cache Server' (edge server)’ to the user delivers the requested content on behalf of the ‘Origin Server’ to reduce the latency and the load of the traffic to/from the main server.

Some solutions are also introduced specifically for streaming over mobile communication systems. Packet-switched Streaming Service (PSS) is a package of solutions for different aspects of the video streaming over the new generations of the mobile communication systems, e.g. 3G and LTE [15]. PSS defines the functionalities required to support server-driven streaming solutions (e.g. RTP-based) or a client-driven adaptive streaming (e.g. 3GP-DASH) [20]. The PSS-related protocols aim to provide a packet-based framework which bridges the trivial downloading protocols and the more complex conversational services. However, with the introduction of the Evolved Packet Core (EPC) in LTE, some aspects of the PSS functionalities for packet and IP-based adaptation are already supported by the system without need for PSS.

The proposed solution in this work is applicable to the clients in Figure 6.1 with an HTTP-based and adaptive video streaming content. This content may have been distributed through CDN with or without the help of the PSS functions. A PSS function, such as the user description and profile server in Figure 6.1, provides a source of information for matching the content specifications with the user’s device capability and environment. Furthermore, the PSS analysis of the client-side buffer status

provides the elements required for the evaluation of the proposed quality metric at the network-side without extra transactions between the client and the network [15].

6.2.2 Sources of Impairments

There are two quality related aspects of a video service that can be compromised for less resource allocation during the communication process: the fidelity based quality of the image and the continuity of the service. Actually these two aspects are related to each other in terms of sharing the same amount of resource. For example, the discontinuity of playback is more likely during a video service with a higher level of visual quality (given a limited amount of bandwidth) [115]. Both fidelity and continuity based quality issues are related to QoE and can be generally assessed through subjective metrics, such as the mostly used MOS [98].

Two transport protocols, i.e. UDP for RTP (Real-time Transport Protocol) and TCP for the most recent HTTP-based streaming services also have impacts on the transmission impairment [116, 117]. For example, discontinuity during the video playback time (namely pause, buffering or under-run) is a result of the network performance inefficiency. This service interruption can be caused by the user's bandwidth limitation, the latency of TCP control mechanisms or its rate throttling. The UDP protocol, however, is less likely to cause discontinuity of the service but its unreliability normally results in degraded image quality.

Due to the subjective nature of QoE and the diversity of the related applications, a unified objective metric for QoE is still not available. Many variations of PSNR (Peak Signal to Noise Ratio) and SSIM (Structural SIMilarity) are used as video quality assessment tools to evaluate the performance of proposed solutions [118]. The occupancy of the playback buffer usually forms a base for the evaluation of continuity in video streaming. The occupancy level, probability of buffer underrun, initial delay and pause durations, pause frequency and jitter are some of the metrics which have been used to quantify the continuity of a video service [94, 119].

A communication model using the visual quality assessment metrics (e.g. PSNR) usually needs the output of the decoder and the original video reference. This type of metric is more suitable for performance analysis rather than an online quality assessment process [120, 121]. In contrast, the continuity based quality can be evaluated without the need of the original reference and decoder output. However, most of the continuity based metrics mentioned above do not have a good correlation with subjective QoE metrics such as MOS. Furthermore, a QoE-driven solution normally acquires extra information sent from the user to the network, which leads to additional control overhead or standard amendments.

As discussed in the previous chapters, Pause Intensity (PI) as a reference-less metric for continuity assessment is highly correlated with the subjective QoE metric, MOS, and this property is content

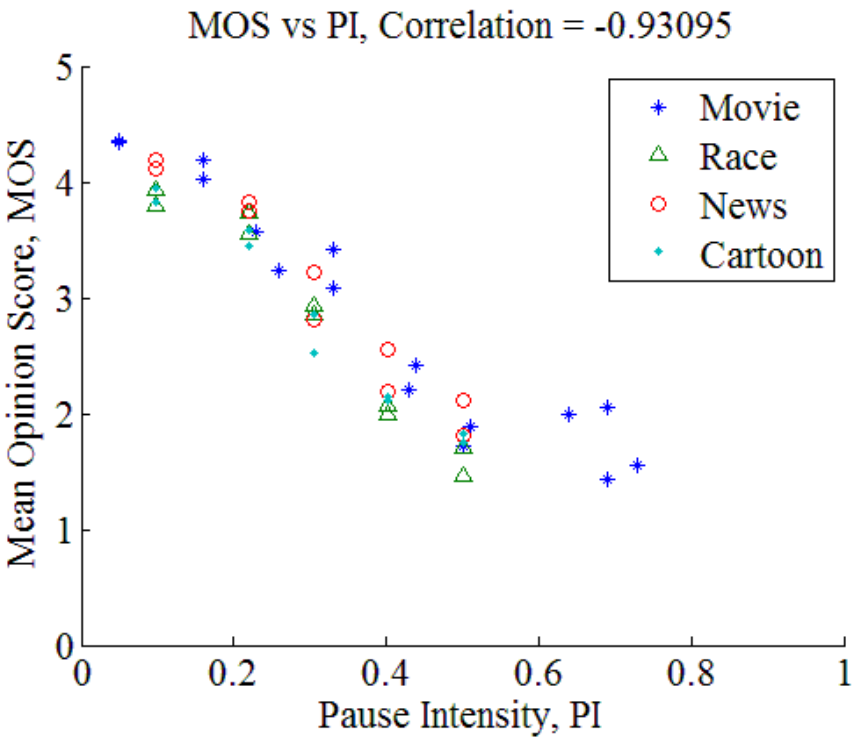


Figure 6.2 Correlation between MOS and PI produced from subjective testing using different video contents.

independent, as shown in Figure 6.2. Recalling formula (5-1), the PI value PI(η, λ), is determined by both network performance, i.e. throughput η, and the required data rate (video bitrate, λ) per user, i.e.:

$$\begin{cases} PI(\eta,\lambda) = 1 - \dfrac{\eta}{\lambda} \\ \eta \le \lambda, \qquad 0 \le PI \le 1 \end{cases} \tag{6-1}$$

The description of the PI model, buffer play-pause characterization and associated subjective tests are provided in Chapter 4 and the PI metric has been used for scheduling in 3GPP-LTE described in Chapter 5. PI is an objective representation of QoE which, as it will be shown in Section 6.3, can be evaluated locally both on the user side and on the network side without additional information exchange and hence no extra overhead for this purpose.

The proposed rate adaptation algorithm in this chapter is examined in the context of LTE/4G. The last mile resource allocation function in eNodeB (i.e. LTE's base station) plays the main role in the proposed idea alongside the link adaptation and channel status control, which will be detailed in the next section. Although the exact resource allocation policy in LTE has not been defined by 3GPP standards, the state of the art solutions are usually based on the general rationality of the resource allocation in mobile communication systems. Subsequently, the main parameters used to characterize

an adaptive streaming scheme are the efficiency of the system as a whole and the fairness to each user.

The most common resource allocation algorithms used to make an efficient or fair scheduler are *best-CQI*, *proportional fair* and *MaxMin throughput* schedulers [103, 122]. The CQI (Channel Quality Indication) in LTE is a feedback from the user to the base station to indicate the capability of the user for using the allocated resources, which is related to the modulation order and the channel coding (or code) rate. The *best-CQI* scheduler (also known as *maxC/I*) is focused on the efficiency of the system by targeting the users with the highest capability in each round of the allocation. In contrast, a scheduler such as *MaxMin throughput* scheduler achieves a high degree of fairness by allocating almost the equal resource to all users regardless of their CQIs.

A balanced allocation to each user can be achieved through the consideration of the efficiency of each user alongside the history of the allocation to that user. For example, in the *proportional fair* scheduler [123] a user with higher efficiency (i.e. better channel quality) will be served more than the users with poorer channel quality. Meanwhile, the comparison of the total allocation to all users will prevent the scheduler from excessive allocation to that user and force the scheduler to serve other users as well. Later in Section 6.4, the performance of our proposed algorithm will be compared with these common scheduling methods.

6.2.3 Quality assessment support in mobile communications

Generally, the assessment metrics for a service are presented in the form of Quality of Service (QoS) or Quality of Experience (QoE). QoS comprises system-oriented and quantitatively defined parameters such as latency, loss or throughput while QoE encompasses user-oriented and qualitatively defined scales such as ‘the opinion and judgment of the user about the service’. QoE may also be represented by a quantitative metric which has a proven correlation with the user’s experienced quality. Although several QoS and QoE-based approaches are available for video-related service assessment, very few of them suit an online and on-the-fly assessment of a streaming service over mobile communication. The employed Pause Intensity (PI) metric in this work provides the required quantitative presentation and the proven correlation with the subjective and QoE-based satisfaction of the user.

The required information for PI evaluation including video bitrate, λ, and average network throughput, η, is available at the client-side. This information can be estimated at the network-side through the functions related to the streaming service support in PSS. When the adaptive streaming service is provided independent from PSS (such as the independent 3GP-DASH servers in Figure

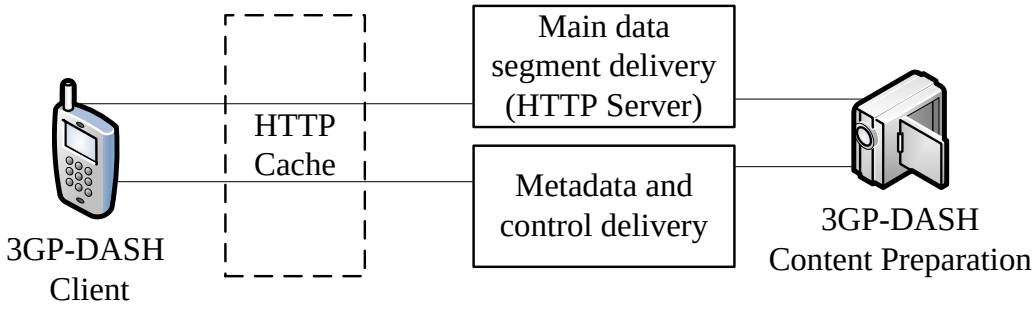


Figure 6.3 The infrastructure of an independent 3GP-DASH client-server

6.3), this information can still be transferred to the corresponding network element as a part of the provided quality-related transactions [15]. These capabilities are briefly explained below.

The QoE support in a contemporary mobile streaming service is actually an attribute defined as a part of the 'device capabilities'. 'Device capability' and 'user profile' are the information which facilitate the server-side negotiation as a matching process at the beginning of the streaming. As it has been shown in Figure 6.1, the device capability description is exchanged between the profile server (provided by device manufacturer or mobile operator) and streaming server (i.e. PSS) as a part of the negotiation and matching mechanisms. The attributes related to video streaming service in 3GPP-PSS consist of the 'QoE support enabler', the type of the carrier protocol (i.e. RTP/RTSP or HTTP) and the available options for rate adaptation mechanisms.

The mobile device sends the URL of the location where PSS can retrieve the user profile (through the RTSP/HTTP data units). PSS communicates with the profile server (through HTTP Request/Response) to access the device description information. User can override or add an attribute during this process. The final negotiated attributes will be used by PSS to control the presentation of the streamed media content to a mobile user. The average throughput, initial playout delay, buffer level and metadata Information are the main video-related QoE metrics which are defined by 3GPP and to be included in a quality report. The XML syntax (using the HTTP request with XML in its body) will be used for the transaction of the QoE metrics and reporting protocols.

In the next section the analytical model for the proposed QoE driven resource allocation model together with an implementation algorithm will be derived and explained.

6.3 QoE-Driven Optimisation and Adaptation Model

6.3.1 Resource Allocation Assumptions

LTE provides resources through a combination of Orthogonal Frequency Division Multiple Access (OFDMA) and Adaptive Modulation and Coding (AMC) techniques in a bandwidth range

from 1.4MHz to 20MHz. A resource allocation unit in LTE is defined as a 'Resource Block (RB)' in a two dimensional time-frequency grid. Each RB is a 180 KHz of bandwidth allocated for one time slot of 0.5ms. Allocation will remain the same in the next time slot which creates a 1ms Transmission Time Interval (TTI) for each transmission process.

Each client provides an evaluation of its channel status, i.e. signal-to-noise ratio (SNR) across the N_{RB} predefined resource blocks:

$$SNR \in \{SNR_{min}, \dots, SNR_{max}\}^{1\times N_{RB}} \tag{6-2}$$

A channel quality indicator (CQI) feedback will be generated based on this evaluation and the capability of the client's device:

$$CQI \in \{1,2, \dots, CQI_{max}\}^{1\times N_{RB}} \tag{6-3}$$

The value of CQI can be a result of a linear fitting of SNR value(s) or searching through a lookup table which reflects the capability of a user's device with regard to different modulation types and (channel) code rates (i.e. MCS) given the SNR values. CQI suggests a range of modulation types and code rates for which at least 90% accuracy will be achievable at the receiver. Given the selected modulation type and code rate (based on the CQI values) and the allocated resources, r_k, the total allocated data rate to user k ($k=1$ to N_{UE}) in the i^{th} round of the allocation, R_k^i, can be calculated as:

$$\begin{cases} R_k^i = C_k^T . r_k \\ C_k = C_k\big(CQI(SNR)\big) \in \mathbb{R}_{>0}{}^{1\times N_{RB}} \end{cases} \tag{6-4}$$

where C_k is the vector of the achievable capacities in the resource blocks for user k, given the corresponding CQI values. r_k is the vector of the allocation defined as follows:

$$\begin{cases} r_k = [r_{k,1}, r_{k,2}, .., r_{k,N_{RB}}]^T \in \{0,1\}^{N_{RB}}, \\ r_i . r_j^T = 0 \ \ \forall\, i \neq j\,, \\ \sum_{k=1}^{N_{UE}} \|r_k\|_1 \leq N_{RB} \end{cases} \tag{6-5}$$

$r_{k,l}=1$ indicates the allocation of the l^{th} resource block to user k and $r_{k,l}=0$ otherwise. Hence: 1) each resource block is supposed to be allocated just to one user; 2) all resource blocks can be allocated to

one user; and 3) allocated resources in each round can be less than the total number of available resources (i.e. some resources may remain unused in each round).

The weighted average of the allocated data rate after the i^{th} allocation round can be assessed as:

$$\overline{R_k^i} = \left(1 - \frac{1}{t_w}\right)\overline{R_k^{i-1}} + \frac{1}{t_w}R_k^i \tag{6-6}$$

where t_w is the average window size and must be large enough compared to the frame duration to filter out the fluctuation of the average allocated resources and capture the average video bitrate of the user (e.g. t_w*=100ms* will suffice for LTE with *10ms* frame duration). As it is depicted in Figure 6.4, the average incoming data rate of the video playback buffer at the receiver, η_k, can be expressed as:

$$\overline{\eta_k^i} = \beta_k \gamma_k \overline{R_k^i} \tag{6-7}$$

where β_k reflects the ratio of the pure video data rate, λ_k, to the whole incoming data rate at the scheduler related to that user, R'_k. This usually includes extra information such as voice, metadata etc. The channel quality, the robustness of the error detection/correction techniques (i.e. HARQ and ARQ) and suitability of the selected modulation type and code rate based on the received feedback are all reflected in γ_k.

6.3.2 Proposed QoE-Driven Optimization Method and Implementation Algorithm

A QoE driven allocation scheme aims to maximize the users' satisfaction level from the service continuity's point of view, which can be interpreted as a process of lowering pause intensity during the playback. This can be expressed as:

$$\begin{cases} r^* = \arg_r \ \min \max PI \\ \begin{cases} PI = \{PI_1, PI_2, \dots, PI_{N_{UE}}\}, \\ PI_i = 1 - \dfrac{\eta_i}{\lambda_i}, PI \in \{x | x \in \mathbb{R},\ \ 0 \le x \le 1\}^{1 \times N_{UE}} \\ \mathrm{H} \le \Lambda\,, \quad \mathrm{H} = \{\eta_1, \eta_2, \dots, \eta_{N_{UE}}\}, \\ \Lambda = \{\lambda_1, \lambda_2, \dots, \lambda_{N_{UE}}\} \end{cases} \end{cases} \tag{6-8}$$

As it has been examined in Chapter 5 and similar to the well-known attribute of a *MaxMin throughput*, the above optimization problem tends to be extremely fair and inefficient. To restore the

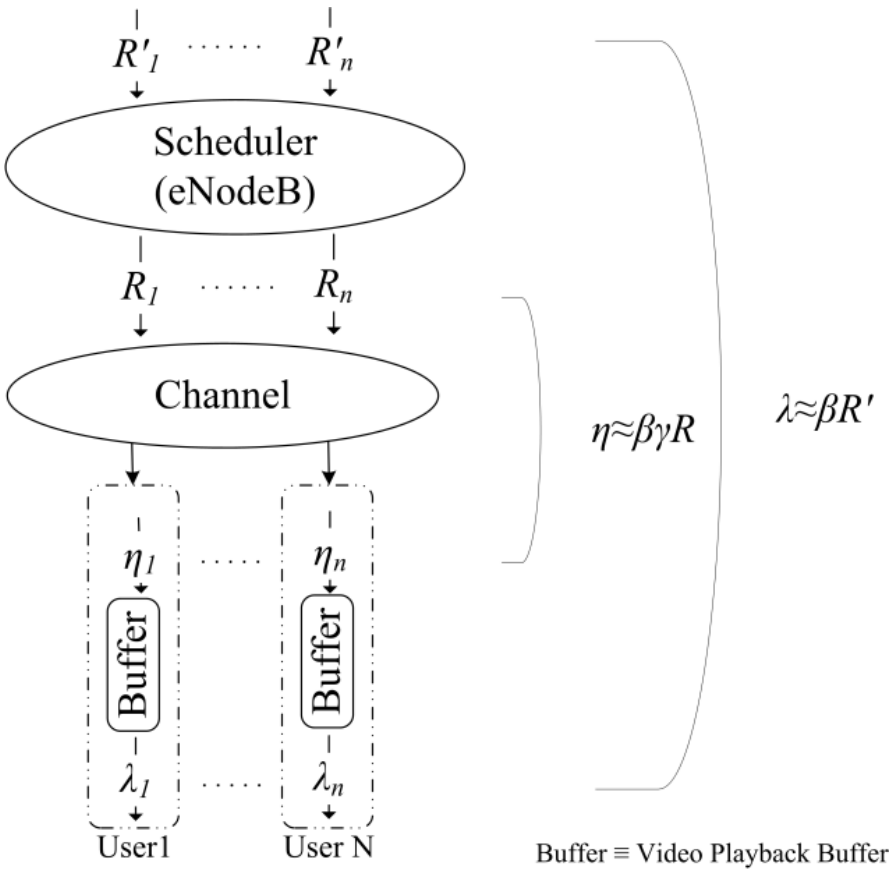


Figure 6.4 Model description and data rates' assumptions

efficiency of the system while maintaining the effect of the user's experienced quality, the problem in (6-8) can be rewritten as a weighted rate scheduling algorithm as follows:

$$\begin{cases} r^* = \arg_r max \sum_{k=0}^{N_{UE}} u_k \\ u_k = PI_k{}^{\alpha}.R_k\,, \quad u_k \in \mathbb{R}_{\geq 0} \end{cases} \tag{6-9}$$

u_k in (6-9) is the utility function where its first term (i.e. PI^{α}), reflects the effect of the clients satisfaction (i.e. QoE). The first term can also be viewed as the weight for the second term, R_k, which represents the user efficiency to consume the allocated resources. The value of α defines the trade-off between the efficiency and fairness, which will be discussed in Section 6.4.

Figure 6.5 shows the changes of the weight, $PI_k{}^{\alpha}$,of rate R_k in the proposed utility function for different values of α and versus a range of user channel status from poor to good (represented as the ratio of the achieved throughput, η, to the required data rate, λ). The depicted result justifies the trade-off between the efficiency and the fairness of the scheduler through the adjustment of α. The result

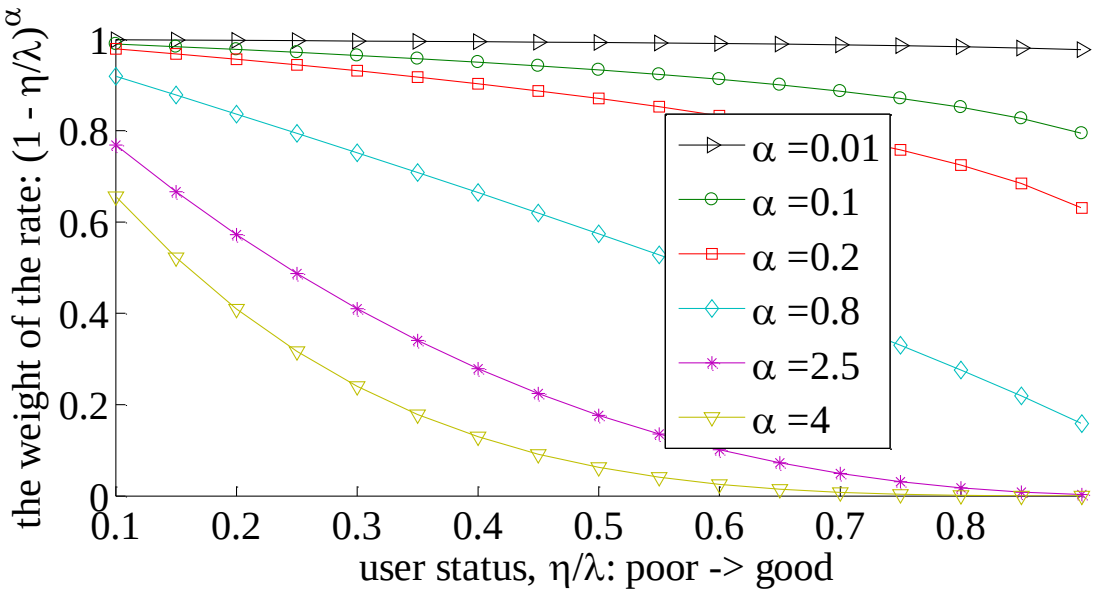


Figure 6.5 The weight of the rate in utility function as a function of parameter α.

shows almost the equal weight for all users in the case of smaller value of α and higher weight for users with poor channel status when α is greater. Therefore, with the value of α closer to zero, users with good channel status are expected to gain more benefit from their achievable rate and the allocation is more efficient. In contrast, users with poor channel status are expected to gain more benefit when the value of α increases, leading to a fairer allocation.

Usually in a wideband assessment of SNR at the receiver, a single average CQI will be generated to suggest the most suitable modulation scheme and code rate for the whole available allocation spectrum at the scheduler. Therefore all the elements of vector C_k in (6-4) will be equal to a certain value, $c_k(CQI)$, and in the i^{th} round of the allocation, (6-9) can be rewritten as a linear programing as follows:

$$\begin{cases} x^* = \arg_x \max f x^T \\ f \in \mathbb{R}^{1\times N_{UE}}, \quad f_k = PI_k{}^{\alpha}.c_k \\ x \in \mathbb{Z}_{\geq 0}{}^{1\times N_{UE}}, \quad x_k = \|r_k\|_1 \leq N_{RB} \end{cases} \tag{6-10}$$

where f_k combines the effect of the user's experienced quality, PI, with its achievable data rate in each resource block, c_k (i.e. user efficiency). The solution of (6-10), $x^*_k \geq 0$, represents the number of the allocated resources to user k and is an integer value while the original problem in (6-9) was a

binary integer programming problem. The evaluation of f_k in each round, given the throughput in (6-7) based on its corresponding average allocated data rate in (6-6), is as follows:

$$f_k = {PI_k}^{\alpha}.c_k = \left(1 - \frac{\eta_k}{\lambda_k}\right)^{\alpha}.c_k \tag{6-11}$$

PI as a QoE metric is by definition based on the client side information, while the actual allocation process is supposed to be on the network side. PI can also be treated as a pre-decoding QoE metric, so can be evaluated merely based on the network side information to avoid additional information exchanges between end users and the network. This implies that, given the network arrangement shown in Figure 6.4,

$$f_k = \left(1 - \frac{\gamma_k R_k}{R'_k}\right)^{\alpha}.c_k \tag{6-12}$$

Obviously, it is also possible to pull the users' evaluated PI from an element in the central network to do any specific load balancing and congestion control for those users that share the resources.

An algorithm for the implementation of the analytical models in (6-9) and (6-10) can be achieved by replacing the utility function in these models with a priority function. Based on the values of the priority function, the allocation process selects just one dominant user in each round and continues until all available resources are allocated. This algorithm can be expressed as:

$$\begin{cases} k^* = \arg_k \max u_k \\ \\ u_k = \left(1 - \frac{\gamma_k R_k}{R'_k}\right)^{\alpha}.c_k \end{cases} \tag{6-13}$$

where u_k appears as a priority function and a user with the highest value of u_k will be chosen in each round of the allocation. It will be shown in Section 6.4 that the result of the algorithm in (6-13) complies with those based on the models in (6-9) and (6-10), and (6-13) is for the practical implementation of the proposed model.

In the next section the result of the implementation algorithm in (6-13) will be compared with that of the analytical model in (6-10). The adjustment of α for achieving a certain trade-off between efficiency and fairness will be examined. In addition, a PI based adaptive video streaming scheme will be presented in the presence of this algorithm to show the effectiveness of PI as a QoE metric for both clients and the network.

6.3.3 PI-Based Rate Adaptive Video Streaming

In a client driven rate adaptive video streaming service (e.g. 3GPP-DASH), the client decides the suitable rate which has to be pulled from the server in each adaptation segment. As it is depicted in Figure 6.6(a) in a shared channel with limited available resources, the user has to decide the best trade-off between the desired fidelity of the image and the minimum acceptable continuity of the service. The user will ask for each segment of the video based on the adapted rate for that time segment. The required assessment in most of the existing technologies is based on the average incoming data rate of the playback buffer compared to a threshold (which is based on the video bitrate). A simplified decision making process for the adapted rate of the $(i+1)^{th}$ segment can be expressed as:

$$\lambda_{i+1} = \begin{cases} \eta_i^*, & \eta_i < \lambda_i \\ \lambda_i + S_v, & \eta_i > \lambda_i \end{cases} \quad (6\text{-}14)$$

where λ represents video bitrate to be adapted, η^* represents the rounded value of network throughput toward the nearest available video bitrate smaller than η. S_v represents the step granularity of the video bitrate. The rate adaptation condition in (6-14) can be reformed based on a minimum QoE threshold (a maximum acceptable discontinuity represented by $PI_{threshold}$) and expressed as:

$$PI_i > PI_{threshold} \quad (6\text{-}15)$$

where PI_i represents the assessed value of PI and $PI_{threshold}$ represents the maximum acceptable discontinuity of the service. Alternately (6-15) can be shown as:

$$\left(1 - \frac{\eta_i}{\lambda_i}\right) > PI_{threshold} \quad \rightarrow \quad \eta_i < (1 - PI_{threshold})\lambda_i \quad (6\text{-}16)$$

which resembles the initial form of the condition in (6-14) with the difference of the effect of the minimum desired QoE (i.e. $PI_{threshold}$). PI provides a quantitative and objective value and can be used to conduct a flexible and network oriented assessment for rate adaptation, instead of using the rigorous user oriented criteria in (6-14). A zero PI threshold produces the initial form in (6-14). As it will be shown later in Section 6.4, a predefined or broadcast non-zero PI threshold for users of a shared channel can produce a desired distribution of QoE from both continuity's and fidelity's points of view.

As depicted in Figure 6.6 (b), a PI driven rate adaptation mechanism (on the client side) actually has an interplay with the last mile's scheduler (e.g. in eNodeB) to shape the QoE distribution among

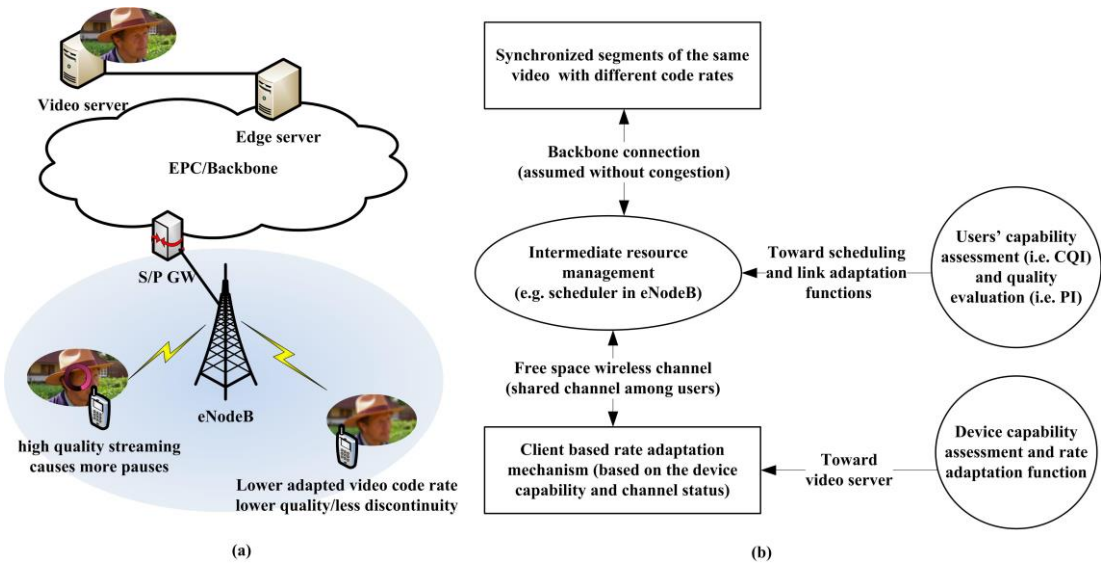


Figure 6.6 Different aspects of the QoE (fidelity and continuity) in an adaptive rate video streaming service and its implementation: (a) two users with similar channel status and the trade-off between fidelity of the image and the continuity of the service (b) the implemented model of an end-to-end rate adaptive video streaning service with the conisderation of the role of last mile scheduling policy.

the users of the shared resources. In Section 6.4 the results of these interactions will be discussed in more detail.

6.4 Simulation Results and Analysis

6.4.1 Simulation Setup

Table 6.1 shows the settings of the simulator developed in MATLAB, which is used to examine the proposed optimization method, its algorithm implementation and the QoE driven rate adaptive video streaming service. The source of the users' data is the video stream data packets generated using a truncated Pareto model (for packets' inter-arrival-time and size). No background traffic has been considered. Video bitrates are in the range of standard video quality of the state of the art technologies (e.g. for BBC-iPlayer 470 kbps -1500 kbps is the current range of non-HD video bitrates for desktop application). Users' Head-of-Line packets (HOL) are scheduled in a timely manner.

The CQI mapping table and the channel status generator presented in [112, 122] are used in our simulator. Each video bitrate is corresponding to more than one user with different SNRs in the range of the defined CQI (i.e. 1~15). This produces unbiased results with regards to the video bitrate or SNR distributions. Users are distributed in one cell with the consideration of the interference from

the first tier neighboring cells. Shadowing effect (inter/intra-cell spatial correlation) has been taken into account as well.

For the sake of comparison, *MaxMin Throughput* and *maxCI* (known as *best-CQI* in LTE) are taken as two extreme sides of the fairness/efficiency spectrum. The integer relaxation and rounding is employed to solve the integer linear programing problem in (6-10) with the constraints driven from LTE's available number of Resource Blocks for the given bandwidth in Table 6.1. The efficiency of the system is represented in b/s/Hz and is the ratio of the total created capacity (i.e. the summation of the total allocated data rates) to the system bandwidth. Fairness is evaluated among the users' allocated data rates using Jain's Index. Correlation between the users' required and allocated data rates is assessed by Pearson's Linear Correlation Coefficient.

6.4.2 Performance of the Proposed Algorithm

Figure 6.7 depicts the achievable efficiency, correlation and fairness of the proposed optimization method in (6-10) and its implementation algorithm in (6-13) for different values of α. The results show the achievable trade-off between fairness and efficiency based on the value of α. Increasing α improves the fairness (Figure 6.7(c)) and correlation (Figure 6.7(b)), but decreasing the efficiency of the scheduling process (Figure 6.7(a)). In contrast, a scheduler using smaller α will lower the levels of fairness and correlation, but increasing the efficiency. However, unlike the optimization problem in (6-10), the simplified algorithm in (6-13) allocates the resources in each round just to the dominant user (i.e. the most efficient user when α is close to zero). This leads to the over-performed efficiency

Table 6.1 Simulation Setup

Parameter	value
No. of Cells	1 (with the first tier interference)
Inter-site distance	2000 meters
Shadowing effect	mean=0, deviation=8 decorrelation distance=25m, inter-site correlation=0.5
Channel model	PedA, speed=3km/h
Bandwidth	5MHz, 20MHz
No. of RBs (per TimeSlot)	25, 100
Subcarrier	15KHz
Range of average SNR	-6 ~ 18 dB (CQI=1~15)
Average video code rate	156kbps ~ 1.5Mbps
No of Users	45
Each scheduling round	One TTI=1ms
Simulation time	10000*TTI (10 s)
Video stream model	Truncated Pareto for packet size and inter-arrival time

of (6-13) compared to (6-10) with higher correlation between the required and the allocated data rates when α approaches zero.

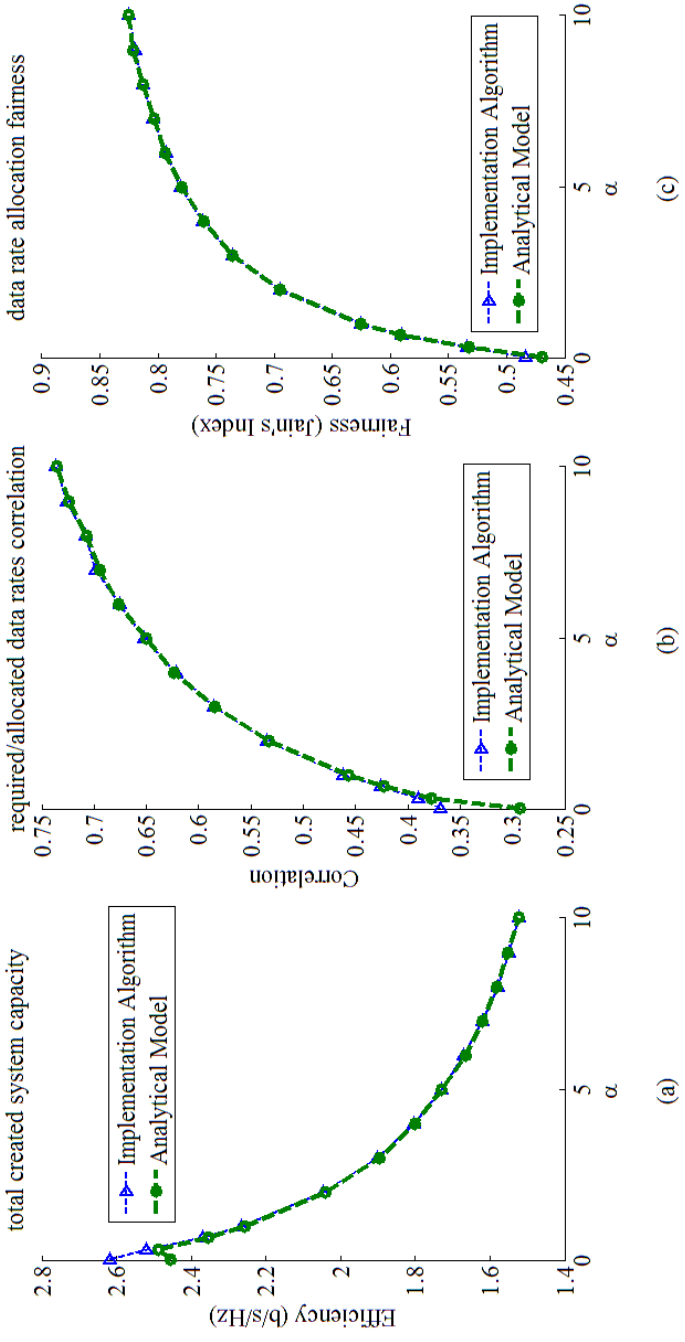


Figure 6.7 The achieved trade-off between efficiency, fairness and correlation for different value α in the proposed scheduling analytical model and implementation algorithm in sub-section 6.3.2: (a) Efficiency of the system vs. α (b) the correlation between required data rates and allocated data rates per user vs. α (c) fairness among allocated data rates to users vs. α .

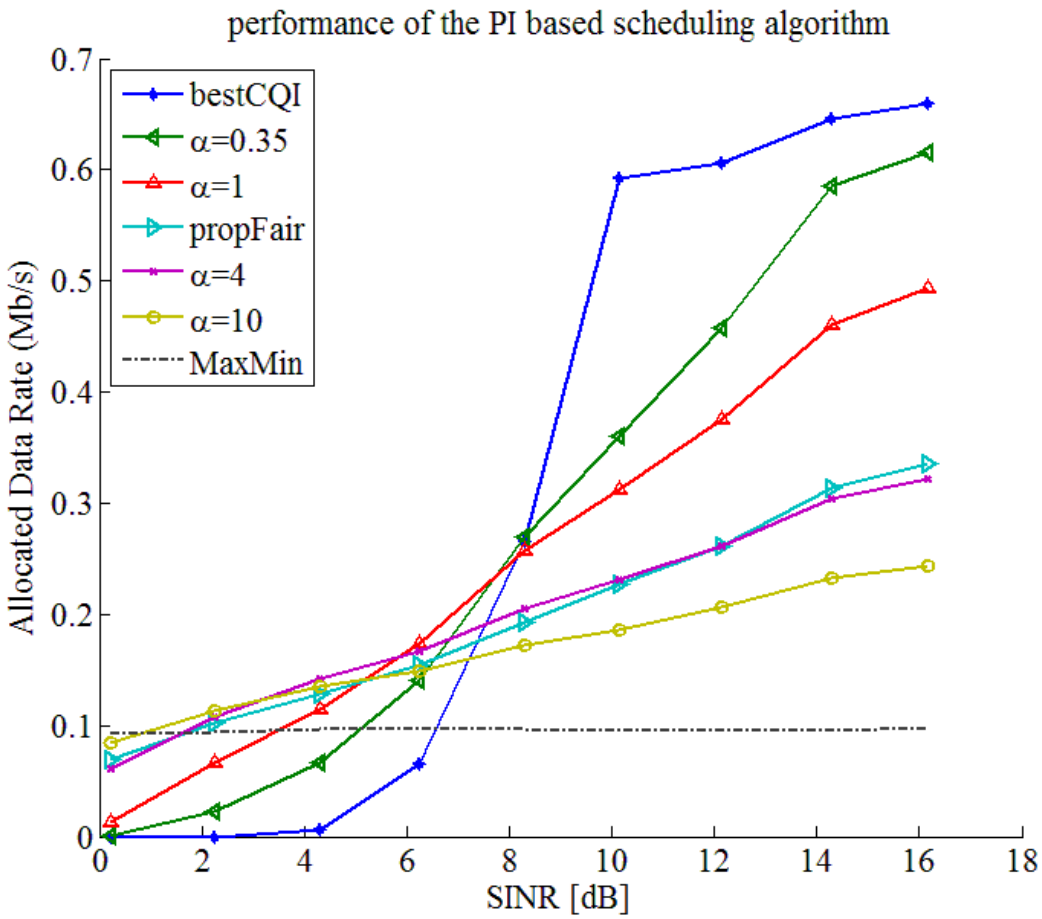


Figure 6.8 The performance of the proposed scheduling algorithm based on the allocated data rates vs the status of the users' channel represented by SINR.

Figure 6.8 shows the performance of the implementation algorithm in (6-13) as a function of the users' channel status. The performance of the scheduler, with different values of α, lay between extremely efficient (e.g. *best-CQI*) and very fair (e.g. *MaxMin throughput*).

6.4.3 Online Adjustment of α-Parameter

The suitable value of α can be chosen based on the desired trade-off between fairness and efficiency, illustrated in Figure 6.9. It can be a fixed predefined value based on the nominal characteristics of the system such as the system performance in Figure 6.7. Parameter α can also be adjusted online based on the assumption about its relationship with the desired fairness or efficiency. With the assumption that the change rate of α with respect to fairness, *f*, is a constant μ (i.e. $\frac{d\alpha}{df} = \mu$), the value of α can be set online as $\alpha_i=\alpha_{i-1}+\mu(f_{target} - f_{current})$ in each iteration where f_{target} is the desired fairness, $f_{current}$ is the achieved fairness via $\alpha=\alpha_{i-1}$ and α_i is the new α to be set.

As it is depicted in Figure 6.9 with two different values of μ and fairness target 0 .75, the value of αi approaches an adequate range after a transient time. It will be amended later, accordingly, with the changes in the situation (e.g. changes in the number of active stations, channel status, background traffic, etc.). The value of μ defines the step of the adjustment in each iteration and affects the speed of the convergence. Smaller μ produces smoother changes of α with less fluctuation in the produced fairness and efficiency (i.e. smoother change in the resource allocation) though this will extend the convergence time. The sufficiency of the achieved convergence time depends on the service demanded. Some alternative online adjustment methods, such as those suggested in [122], are available, which can be tailored for our purpose to achieve shorter transient time if necessary.

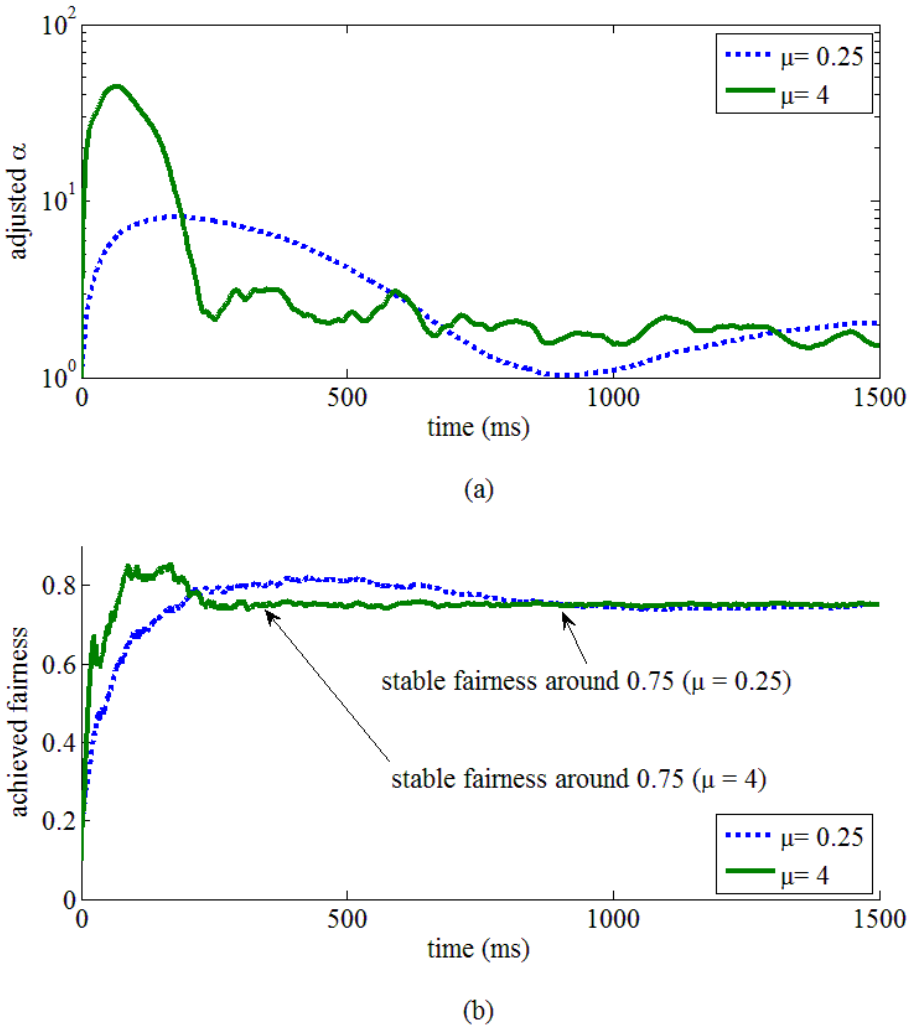


Figure 6.9 Online adjustment of α for fairness target 0.75 and the step size of the amendment μ=0.25 and 4: (a) adjusted values of α vs time (b) achieved fairness using the adjusted α in the scheduler

6.4.4 Client QoE-Driven Rate Adaptation

A PI based criteria for rate adaptation for video streaming has been introduced in Subsection 6.3.3. Figure 6.10 provides an insight into the performance of such an adaptive scheme compared to non-adaptive video streaming from both the user's and network's points of view. The initial values of the video bitrate for all of the users are a default value (780kbps in this example) and in the case of the adaptive streaming, video bitrates can vary (above or below the initial value, i.e. 156kbps~1.5Mbps).

Figure 6.10(a) depicts the adopted rates for two users with distinctive channel status, where the user with higher capability gradually acquires more image quality through the higher video bitrate. The user with poor channel status has to reduce the requested image quality to maintain an acceptable continuity for the service. The cost of good continuity for users with bad channel status will be lower levels of fidelity for their image. However, the users with higher capability and better channel status will be served with higher video bitrate. This has been shown in Figure 6.10(b) where the single choice of the video bitrate in non-adaptive service is expanded across a wide range of available rates higher or lower than the initial value. Figure 6.10(c) shows the achieved continuity of the service in each case. Since the adaptive streaming mechanism can reduce the requested quality of the video if necessary, it maintains the continuity of the service and achieves higher probabilities of being low PI instead.

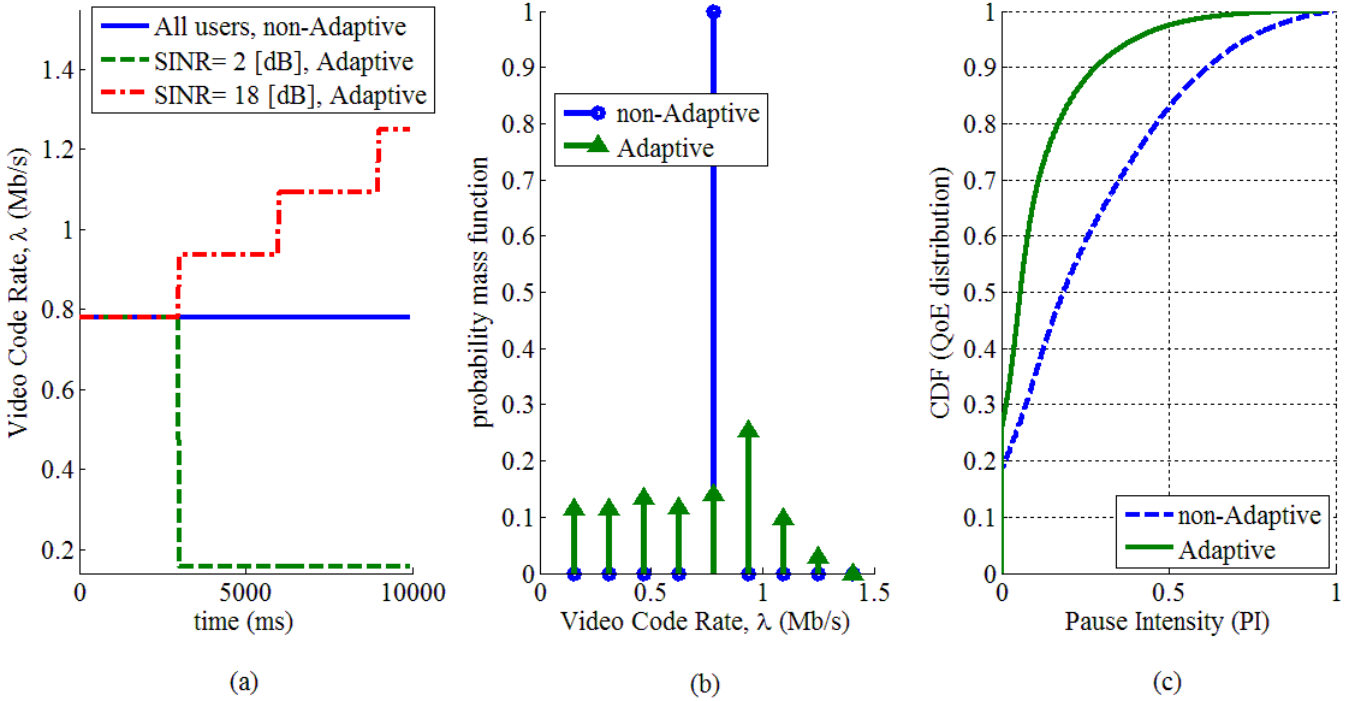


Figure 6.10 Adaptive video streaming performance compared to a non-adaptive service: (a) adapted rates for two users with distinctive channel status (b) the spectrum of the adapted rates of all users compared to the fixed rate of non-adaptive service (c) overall QoE performance of adaptive service compared to the non-adaptive service from continuity point of view

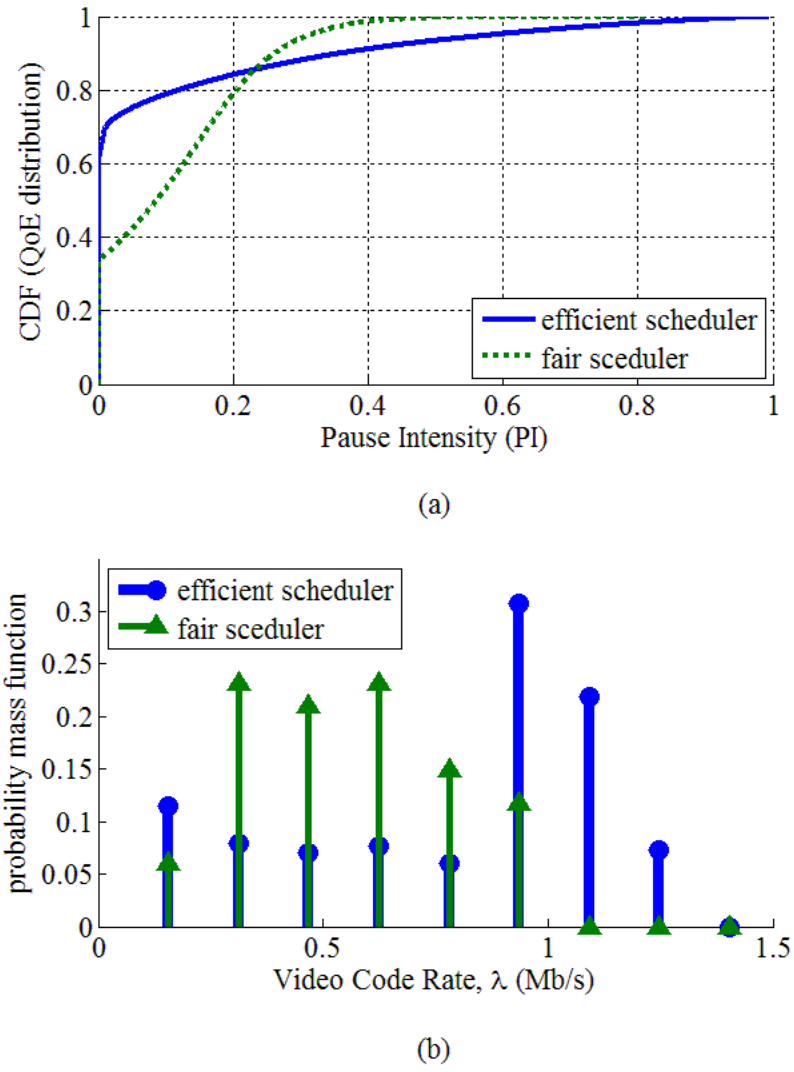


Figure 6.11 The effect of the last mile scheduler over the performance of the rate adaptation process: (a) achieved QoE of an efficient scheduler compared to a fair scheduler from continuity point of view (b) a comparison between the spectrum of the adopted rates of an efficient and a fair scheduler

Figure 6.11 depicts the interplay between the last mile scheduler (i.e. scheduling and rate adaptation functions of eNodeB in LTE) and the client side rate adaptation mechanism. The results of two distinctive efficient and fair schedulers with α=0.3 and α=3, respectively and as discussed in Section 6.3, are provided in Figure 6.11 for the purpose of comparison. On the client side, the rate adaptation mechanism chooses the desired rate of the video based on the performance of the network. As shown in Figure 6.6(b), the last mile wireless channel is supposed to be the main resource bottleneck. Therefore, the scheduling policy used in eNodeB that considers the capability of the user's device is the main factor affecting the network performance.

The user is expected to choose its video bitrate not only depending on its channel quality but also under the resource constraint which is related to the status of other users in the same cell (shared resources). The distribution of the QoE from the continuity's point of view (represented by Pause Intensity in Figure 6.11(a)) and from the fidelity's point of view (represented by the spectrum of the adapted video bitrates in Figure 6.11(b)) are highly polarized in the case of an efficient scheduler. It means that the rate of an adaptive video streaming service will be either very high or very low with a low possibility of being intermediate values.

An efficient scheduler provides more resources to users with better channel quality, hence higher video bitrates will be fetched by them. Consequently, users with poorer channel quality experience will limit network performance. Therefore they adopt lower video bitrates to maintain the minimum desired level of continuity of the service. A wider range of video bitrates will be chosen by the fair scheduler though the maximum video bitrate is restricted in this case. This fact has been reflected in the result where a smoother change in the distribution of QoE (for continuity) is exhibited in the range of $0 \leq PI < 1$.

As it has been explained in Section 6-3, the video bitrate adaptation process can be defined based on the acceptable continuity status (i.e. maximum acceptable Pause Intensity, PIthreshold). The result of the quality metric involvement in the user-side's rate adaption process is shown in Figure 6.12. The average adapted video bitrate throughout the simulation period is depicted as a function of the maximum acceptable Pause Intensity, PIthreshold. The implemented mechanism successfully maintains the adapted video bitrates which are low enough to achieve an almost pause-less playback (i.e. 500kbps average video bitrate for PIthreshold=0.01). In contrast, with a higher level of tolerable discontinuity for the service the implemented mechanism allows the users with good channel status and high device capability to increase their adapted video bitrates (i.e. 800kbps average video bitrate for PIthreshold=0.6).

The distribution of QoE for different values of PIthreshold, in terms of the continuity, is shown in Figure 6.13. Higher expectation for continuity when PIthreshold=0.01 has resulted in a distribution with more probable 'Excellent' quality. However, as it has been shown in Figure 6.12 the average visual quality will be lower in this case. In contrast, the service quality is less likely to be 'Excellent' or 'Good' in the case of the higher values of PIthreshold. The higher value of PIthreshold degrades the continuity of the service though it increases the average visual quality (as shown in Figure 6.12).

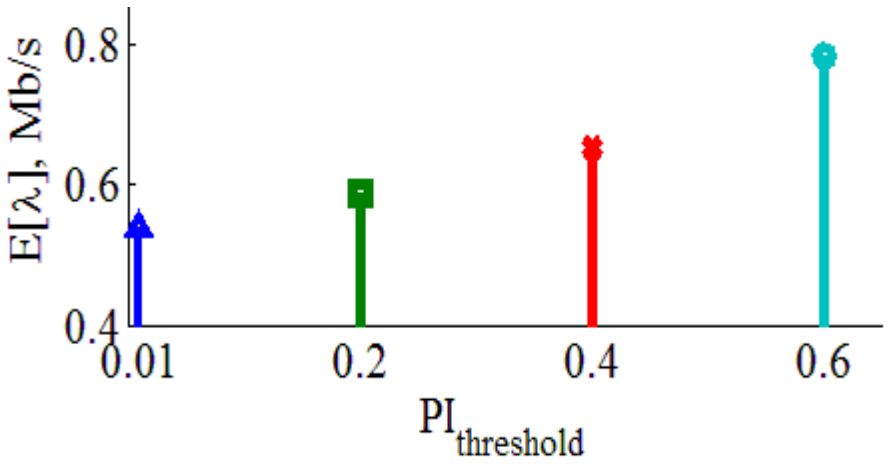


Figure 6.12 Average video code rate against the quality threshold. The Pause Intensity threshold (PIthreshold) represents the acceptable quality from continuity point of view while the average video code rate embodies the image quality/fidelity.

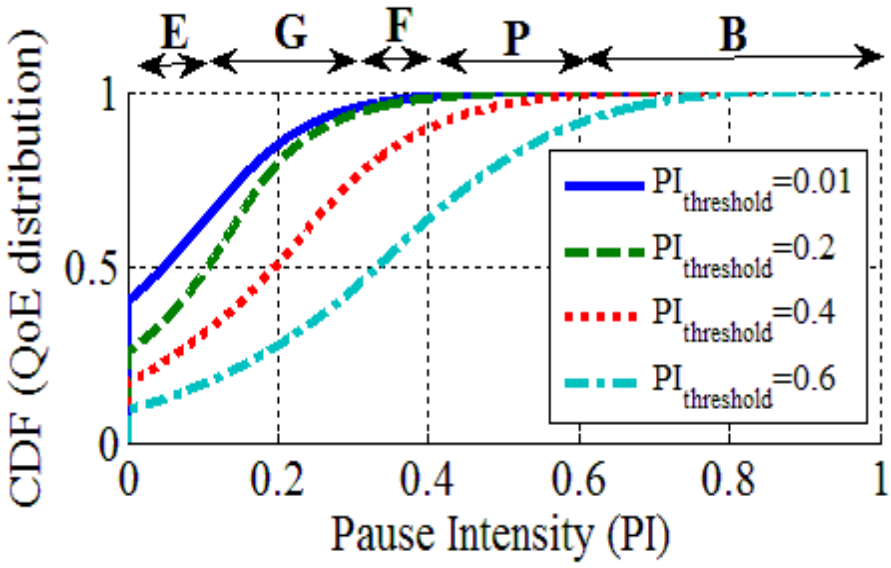


Figure 6.13 The effect of different rate adaptation condition (PIthreshold) over the achieved QoE distribution: PI distribution actually represents the continuity aspect of the QoE. The subjective aspect of the quality is also depicted through the corresponding ranges of MOS values from E (Excellent) to B (Bad).

6.5 Summary

In this chapter a QoE driven adaptive scheduler has been proposed and examined in the context of a wireless mobile communication system providing video streaming services with adaptive code rates. An algorithm for the implementation of the established analytical model (i.e. the LP problem) has been presented. Pause Intensity is adopted to quantify the continuity aspect of the service with the capability of being evaluated on both client and network sides. The proposed algorithm provides a flexible tool to achieve a desired trade-off between fairness and efficiency. Furthermore, the effectiveness of the online adjustment method for the scheduler parameter to maintain the desired level of fairness or efficiency has also been shown.

PI has been used to regulate user's video bitrate on the client side and to shape the distribution of the QoE related performance among users in collaboration with the scheduler on the network side. The performance trade-offs between efficient and fair schedulers in both adaptive and non-adaptive modes for streaming have been analyzed.

Chapter 7
Optimal Video Offloading Through LTE-WiFi Interworking

7.1 Introduction

Issues on network capacity enhancement, cell edge performance improvement and fast growing data traffic, especially the dominancy of video traffic are amongst the main concerns of the current mobile systems development [113]. This has been reflected in the latest standardization of the advanced Long Term Evolution (3GPP-LTE-A) [13]. The combined technological solutions such as Small Cells, Wi-Fi offloading, Relays and Self-Organizing Network (SON) have been considered within the mobile communication system to overcome its capacity limitation. These solutions would be especially attractive if they do not require a complex upgrade of the existing network architecture.

A proximity service, such as offloading through interworking with WiFi hotspots (based on the IEEE 802.11u wireless LAN (WLAN) [124], recently defined as Passpoint), acts as a congestion reliever in LTE networks [125]. It also improves the user's QoE, reshapes the distribution of network capacity, and controls the radio power consumption. Higher throughput and lower cost per bit are expected to be achieved through the deployment of small cells or interworking with low cost and previously established WLAN hotspots.

The current standardization on the data offloading mechanism is mainly dealing with the IP routing mechanism, breakout point functions and methods for redirecting user data from LTE to a different connection point. The proposed architecture of the IP Flow Mobility (IFOM) in relation with LTE's Evolved Packet Core (EPC) [126] is an example of these efforts. However, other aspects of the traffic offloading in LTE such as the shape of the redistributed capacity, the effects of the scheduling policy, fairness and efficiency factor trade-offs and the type of the hotspot's backbone

connection are still open for further study. In this work we examine these aspects in the context of data offloading in LTE for video streaming traffic.

We propose an offloading solution for video streaming services through the interworking between LTE's base station (i.e. eNodeB) and WLAN hotspots with the involvement of a QoE metric. Pause Intensity (PI), as a no-reference and packet based metric for QoE evaluation plays a key role in the resource allocation strategy in this solution. The effect of the resource allocation policy on the offloading performance will be discussed in detail later. We will also compare the performance between two different offloading methods, one through the hotspot which has an independent backbone connection and the other using a WiFi access point as a user of a macro-cell base station. The former type of the WiFi access points are defined in 3GPP standards as Interworking WLAN (I-WLAN) [125]. The latter type is commercially known as Mobile WiFi device (MiFi). MiFi resembles a small cell or femtocell with an unlicensed radio spectrum and a macro-cell dependent backbone connection.

MiFi is available as a standalone LTE-WiFi device with improved reception quality. Tablets and smartphones can also bridge their mobile 3G/4G connection with their WiFi port to serve as a MiFi for neighboring users. However, these personal devices have a lower reception quality compared to a standalone MiFi device. Given these two options for the user equipment (i.e. UE) to choose for connecting to the backbone network, as explained above, the resource allocation mechanism at the eNodeB needs additional intelligence to ensure the overall network performance to be optimal in terms of the best trade-off between efficiency and fairness. This issue is addressed by the scheme proposed in this work.

The rest of the chapter is organized as follows. The background and related works are explained in the second section. The proposed rate redistribution scheme for the LTE-WiFi interworking system is presented in Section 7.3. The simulation results and analysis are discussed in Section 7.4 and finally the conclusion is given in Section 7.5.

7.2 Interworking in LTE Networks

The network capacity or the total offered data rate supported by the new generation mobile communication technologies has not been fully utilized across the area covered by a macro-cell base station. Furthermore, any promised capacity is highly conditional and depends on the reception quality of users' device, the distribution of users in the cell and, most importantly, the resource allocation and scheduling policy. In fact, most of the users at the edge of the cell will not benefit from the improved capacity of the network if they merely rely on their direct connection to the macro-cell base station. 3GPP standardization has considered WLAN traffic offloading as a solution for maintaining the performance of the service at its required level [125]. This will be the case whenever users are better off in receiving their service through a WiFi hotspot than a direct macro-cell

connection. Similar concepts such as small cells and femtocells are also defined using the licensed radio spectrum of the macro-cell.

Although the interworking between WLAN and LTE is a fairly new subject, the main concept of offloading through interworking between different types of the networks (due to the heterogeneity of the network) has been connected to other technical issues such as load balancing [127, 128], resource management [129] and congestion control [130].

The different aspects of the WLAN offloading mechanism in the context of the new generations of the mobile communication systems (e.g. LTE-Advanced) is currently under development. An offloading process through WiFi hotspot can be seamless if without service interruption or non-seamless otherwise [131]. Offloading may require the initiation and signaling of users' equipment and can also happen even without the intervention of users' device (e.g. client and network based IP mobility in IFOM [126]). The performance of the standardized traffic offloading mechanism is also being investigated alongside small-cell and device to device communication technologies [132-134]. In collaboration with higher layer's protocols, the traffic offloading mechanism has been employed to improve the performance at the application layer in a cross-layer paradigm [135].

All the works mentioned above are based on the standard interworking between offloading hotspots and the mobile system while the new emerging standalone and user-controlled mobile WiFi devices are neglected in these scenarios. Traffic offloading through a mobile WiFi device reshapes the distribution of the available capacity (in terms of the allocated data rate to users) throughout the cell. However, it still relies on the capacity of the mobile base station. This important aspect regarding the current offloading mechanism and an algorithm for evaluating and improving its performance will be discussed in the following sections.

7.3 Model of Rate Allocation

Different scenarios involving the traffic offloading mechanism have been depicted in Figure 7.1. In this network, user equipment 1(UE1) has a normal connection to the server directly through the macro-cell base station (known as eNodeB in LTE). UE2, however, receives the service through offloading via a WiFi access point which has a backbone connection independent from macro-cell. In this case, a router (RA) and a wireless access gateway (WAG) are usually used to provide the backbone connection for enabling IP mobility and interworking with LTE-EPC, which is the standard form of the offloading mechanism and known as I-WLAN. For UE3, the connected WiFi access point is a portable/mobile WiFi device called Mobile WiFi or MiFi. MiFi is connected to the backbone through the LTE base station. UE3 will remain a user of the macro-cell base station. In contrast to the I-WLAN, a MiFi is not under the control of the network operators and can serve a limited number of mobile or fixed devices for partial or full offloading purposes.

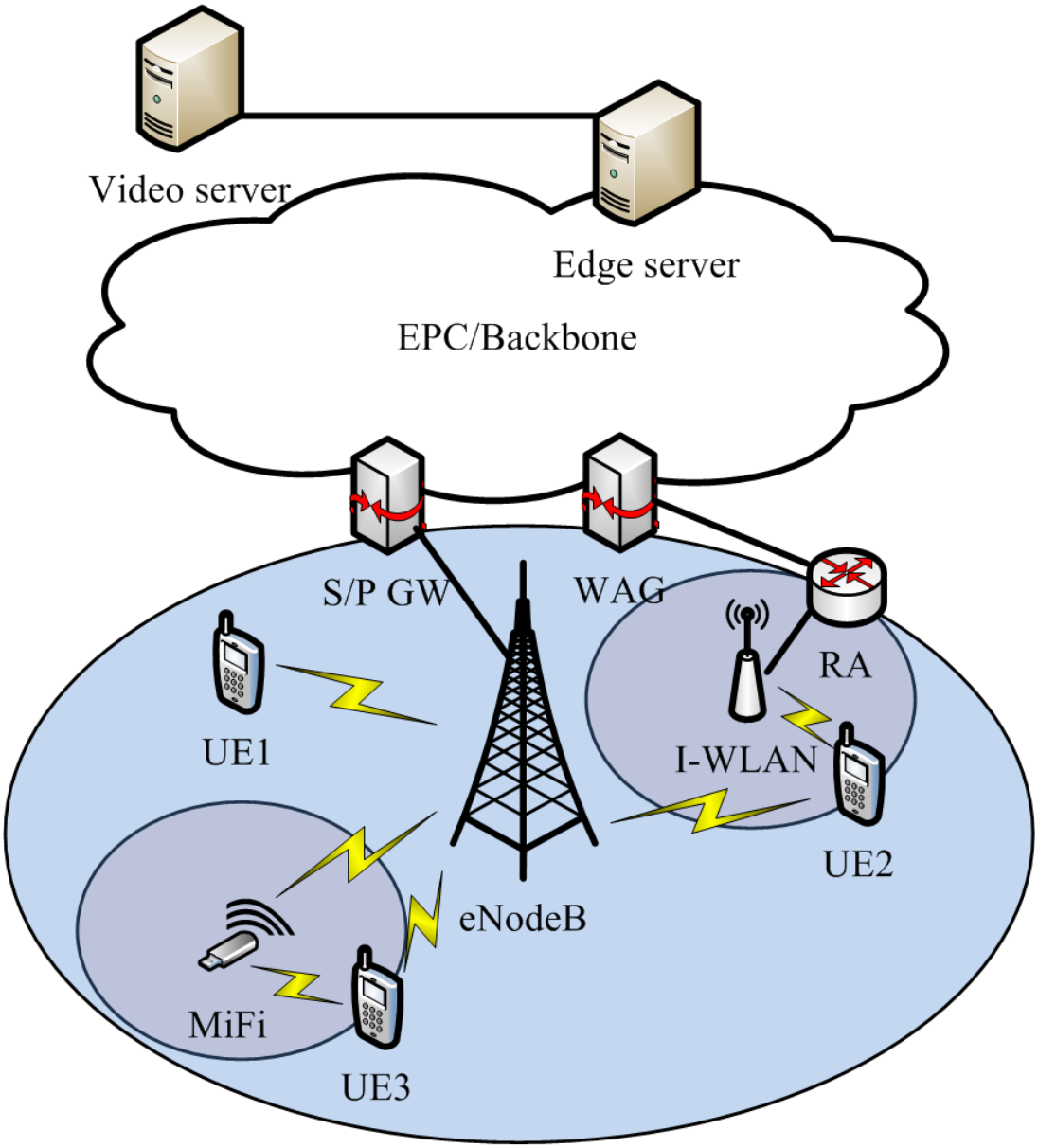


Figure 7.1 Different types of video streaming service provisioning through a direct connection to the base station (UE1), offloading through an independent WiFi access point (UE2 through I-WLAN) and offloading through a mobile WiFi (UE3 through MiFi)

In the rest of this section, based on the scenarios mentioned above, the quality assessment metric employed, the conditions for resource allocation in LTE and the rate redistribution algorithm will be discussed in the context of video streaming services. The performance of the proposed algorithm will be analyzed later in Section 7.4

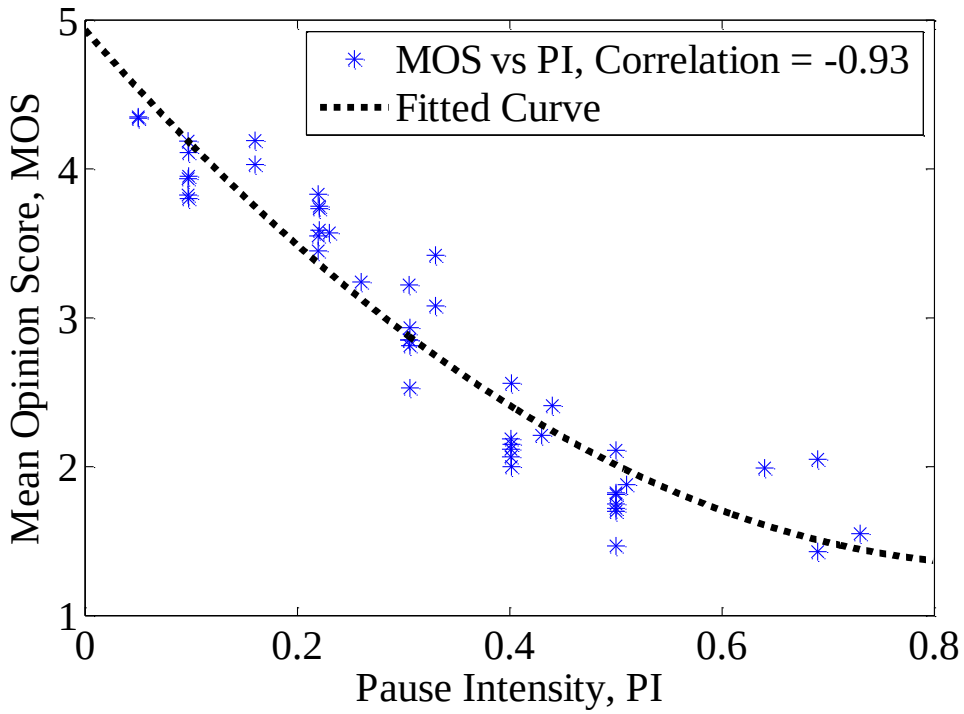


Figure 7.2 Correlation between MOS and PI achieved from a subjective testing with different video contents. MOS and PI are highly correlated (negatively)

7.3.1 QoE Metric

Pause Intensity, PI, is a metric for quality assessment which quantifies the video streaming playback discontinuity through playout buffer behavior characterization. As explained in Chapter 4, PI takes both pause duration and pause frequency into account and can also be represented by the network performance (throughput) and service level (code rate) parameters, i.e.:

$$PI = 1 - \frac{\eta}{\lambda} \tag{7-1}$$

where η is the network throughput and λ is the video bitrate (or required data rate) per user. In a streaming scenario η is normally less than or equal to λ and $0 \leq PI \leq 1$.

In Chapter 4 PI shown to be closely correlated with the subjective quality evaluation metric, *MOS* (Mean Opinion Score), as depicted in Figure 7.2. With this feature, *PI* has been employed in Chapter 5 and Chapter 6 to estimate the level of user's QoE for network operators and service providers to allocate proper resources to end users in 3GPP-LTE networks [136].

In a macro-cell serving a large number of mobile devices, it is essential to maintain a good balance between the efficiency and fairness in delivering data to users with different conditions. As an extension to the original PI metric, the high-order PI can be introduced to act as a weighting coefficient for rate allocation with QoE awareness. In Chapter 6, this coefficient is defined as PI^{α}, where $0 \leq \alpha$ [137]. The detailed descriptions with regard to how the high-order PI is applied to the rate redistribution algorithm and how the best tradeoff can be achieved with this algorithm are given in Subsection 7.3.3 and Section 7.4 respectively.

7.3.2 Resource Allocation for LTE

Each user of the LTE base station, eNodeB, provides an evaluation of its channel status across the available resource blocks, NRB, based on the signal-to-noise ratio, such as

$$SNR \in \{SNR_{min}, \dots, SNR_{max}\}^{1\times N_{RB}} \tag{7-2}$$

A CQI (Channel Quality Indication) feedback will be generated based on the above evaluation alongside the capability of the client's device, which is defined as

$$CQI \in \{1,2, \dots, CQI_{max}\}^{1\times N_{RB}} \tag{7-3}$$

The value of the CQI can be a result of a linear fitting of SINR value(s) or searching through a lookup table similar to the example shown in Table 7.1. These values reflect the capability of the user's device with regard to different modulation and channel code rates (i.e. MCS) to achieve a minimum acceptable error rate. CQI suggests a range of modulation orders and code rates for which at least a 90% successful rate will be achievable at the receiver. Given the selected modulation order and channel code rate (based on the CQI values) and the allocated resources, r_k, the total allocated data rate to user k ($k=1$ to N_{UE}) in the i^{th} round of the allocation, R_k^i, can be calculated as:

$$\begin{cases} R_k^i = C_k^T . r_k \\ C_k \in {\mathbb{R}_{>0}}^{N_{RB}\times 1}, r_k \in \{0,1\}^{N_{RB}\times 1} \end{cases} \tag{7-4}$$

where r_k is the vector of the allocation and C_k is the vector of the achievable capacities in the resource blocks for user k, given the corresponding CQI values (i.e. $C_k=f(CQI(SINR))$ as in Table 7.1) . This principle will be used to model the overall capacity of the system in the next subsection.

Table 7.1 Link adaptation and modulation scheme

SINR	CQI	Modulation Order	Code Rate
≤ -6.934	1(*)	2	0.1523
-5.147	2	2	0.2344
-3.180	3	2	0.3770
-1.254	4	2	0.6016
0.7610	5	2	0.8770
2.700	6	2	1.1758
4.697	7	4	1.4766
6.528	8	4	1.9141
8.576	9	4	2.4063
10.37	10	6	2.7305
12.30	11	6	3.3223
14.18	12	6	3.9023
15.89	13	6	4.5234
17.82	14	6	5.1152
≥ 19.83	15	6	5.5547

*** Users with SINR lower than a device related threshold will not be scheduled for using the resources**

7.3.3 Rate Redistribution Algorithm

Based on the structure of the last-mile wireless connection for a video streaming service (shown in Figure 7.1), user can either be served through the LTE macro-cell base station, i.e. eNodeB, or through a WLAN hotspot whenever an offloading mechanism is in place. Each hotspot can be connected to the backbone through its broadband connection which is independent from the mobile base station. It can also be connected to the backbone as a user of the mobile base station as it has been explained previously as a MiFi device.

An independent broadband connection to the backbone for WiFi hotspots will increase the overall capacity in terms of the allocated data rate in the cell. On the other hand, a hotspot which is connected to the base station doesn't increase the total available resources in the cell. However, it increases the capacity since it has better reception and can help improve the resource distribution with its own capacity. Wherever a MiFi device provides a better reception compared to eNodeB for mobile users, the link adaptation and resource utilization will be improved for those users. In all cases the locations of the offloading hotspots and the policy of the scheduler in the macro-cell base station are the main factors for reshaping the distribution of the data rates to the users.

Consider a macro-cell with an LTE base station with N_{UE} mobile users and N_{AP} WiFi hotspots. The total provided capacity in the cell will be a combination of the allocated data rates to the users by the main base station and the surrounding hotspots, which is expressed as:

$$R_t = R_{eNB} + \sum_{k=1}^{N_{AP}} R_{\mathrm{AP}_k} \tag{7-5}$$

where R_{eNB} and R_{AP} are the allocated data rates to the users covered by the base station and WiFi hotspots. Scheduling and link adaptation functions in the base station serve all users in the cell except for those who are under the coverage of the I-WLAN hotspots. Those users are assumed to be served fully through a sufficient WiFi broadband connection. As a result, the R_{APk} will be the summation of all the offloaded traffic of the k^{th} I-WLAN. However, in the case of MiFi, the Head-of-Line packets (HOL) related to the users under its coverage will be scheduled by eNodeB as a part of the MiFi's data.

An efficiency-oriented scheduler in eNodeB (e.g. bestCQI) will maximize the total allocated data rate, R_t, while a fairness-oriented scheduling policy will provide a flat data rate allocation among the users. The efficiency and fairness of the system will be evaluated in this chapter based on the total allocated data rates and the Jain's index of the allocated data rates, respectively.

By using the high-order PI metric defined in Subsection 7.3.1 and the user capacity in (7-4) we have:

$$\begin{cases} x^* = \arg_x \max f^T x \\ f \in \mathbb{R}^{N_{UE}\times 1}, \quad f_k = {PI_k}^{\alpha}.\bar{c_k} \\ x \in {\mathbb{Z}_{\geq 0}}^{N_{UE}\times 1}, \quad x_k \leq N_{RB} \end{cases} \tag{7-6}$$

where f_k (regarded as a utility function) combines the effect of the user's experienced quality, PI, with its average achievable data rate, $\bar{c_k}$ (i.e. user efficiency). The integer value $x^*_k \geq 0$ is the number of the allocated resources to user k (i.e. $x^*_k = ||r_k||_1$) during the scheduling process based on which the total allocated data rate can be expressed as:

$$R_{eNB} = \bar{C}.x^T, \qquad \bar{C} = \left[\bar{c_1}, \dots, \overline{c_{N_{UE}}}\right] \in \mathbb{R}^{1\times N_{UE}} \tag{7-7}$$

where C is the vector of the achievable data rates for each user if a single resource block be allocated to that user. c_k is a function of the SINR and CQI at the receiving side and has been assumed to be the same for all resource blocks (i.e. throughout the available bandwidth). Since a MiFi device mediates between its covered users and the mobile base station, the covered users will benefit from its higher SINR and subsequent higher c_k.

Parameter α $(\alpha \geq 0)$ defines the degree of the influence of QoE over the scheduling process. Since $0 \leq PI \leq 1$, the value of c_k dominates the utility function, f_k, for smaller value of α. Consequently, it enables a more efficient scheduling process. In contrast, the larger value of α reduces the efficiency but increases fairness among the users due to the consideration of user QoE. Further discussion on this with performance results will be presented later in Section 7.4. An implementation algorithm for the optimization model in (7-6) can be expressed as follows:

$$\begin{cases} k^* = \arg_{\ k} \ \max u_k \\ \\ u_k = PI^{\alpha}.c_k \end{cases} \tag{7-8}$$

where u_k as a priority function allows the scheduler to choose a user, k, with the dominant value of u_k in each round of the scheduling. This reduces the complexity of the implementation and the processing requirements compared to (7-6), and its performance closely resembles the model. In the next section, the performance of the proposed implementation algorithm in (7-8) will be compared with the optimisation model in (7-6) for different interworking scenarios.

7.4 Simulation Results and Analysis

7.4.1 Simulation Setup

Table 7.2 shows the settings of the simulator developed in Matlab to examine the proposed QoE driven rate redistribution algorithm for interworking between WiFi and LTE. The users' data is the video stream data packets generated using a truncated Pareto model (for packets' inter-arrival-time and size). No background traffic is considered. Video bitrate has been set to a standard video streaming quality of 790kbps (e.g. as in BBC-iPlayer). Users' Head-of-Line packets (HOL) are scheduled in a timely manner with no packet drop due to the delay in the scheduler (i.e. eNodeB). Table 7.1 defines the implemented mapping between SINR and CQI for lower layer's function settings such as the modulation order and channel code rate.

The same video bitrate corresponds to the users with different SINRs in the range of the defined CQI for LTE (i.e. 1~15). Users are distributed in one cell considering the interference from the first tier neighboring cells. A pedestrian user's model of fading and the shadowing effect (inter/intra-cell spatial correlation) have been taken into account as well. The subjective quality metric of the service, MOS, has been estimated through the fitting of the stable value of PI given the subjective test results in Figure 7.2. Figures 7.3, 7.4 and 7.5 show the received power, geometric SINR (i.e. the SINR calculated based on the Euclidian geometric model) and CQI distribution of the simulated network, respectively.

Table 7.2 Simulation Setup

Parameter	value
No. of Cells	1 (with the first tier interference)
Inter-site distance	2000 meters
Shadowing effect	mean=0, deviation=8 decorrelation distance=25m, inter-site correlation=0.5
Channel model	PedA, speed=3km/h
Bandwidth	20MHz
No. of RBs (per TimeSlot)	100
Subcarrier	15KHz
Range of average SINR	-6 ~ 18 dB (CQI=1~15)
Average video code rate	790 kbps
No of Users	130 (in each simulation run)
Each scheduling round	One TTI=1ms
Simulation time	10000*TTI (10 s)
Video stream model	Truncated Pareto for packet size and inter-arrival time
WiFi standard/coverage	802.11g / 100 m (max radius)
MiFi device antenna gain	15dB (relative to geometric SINR)

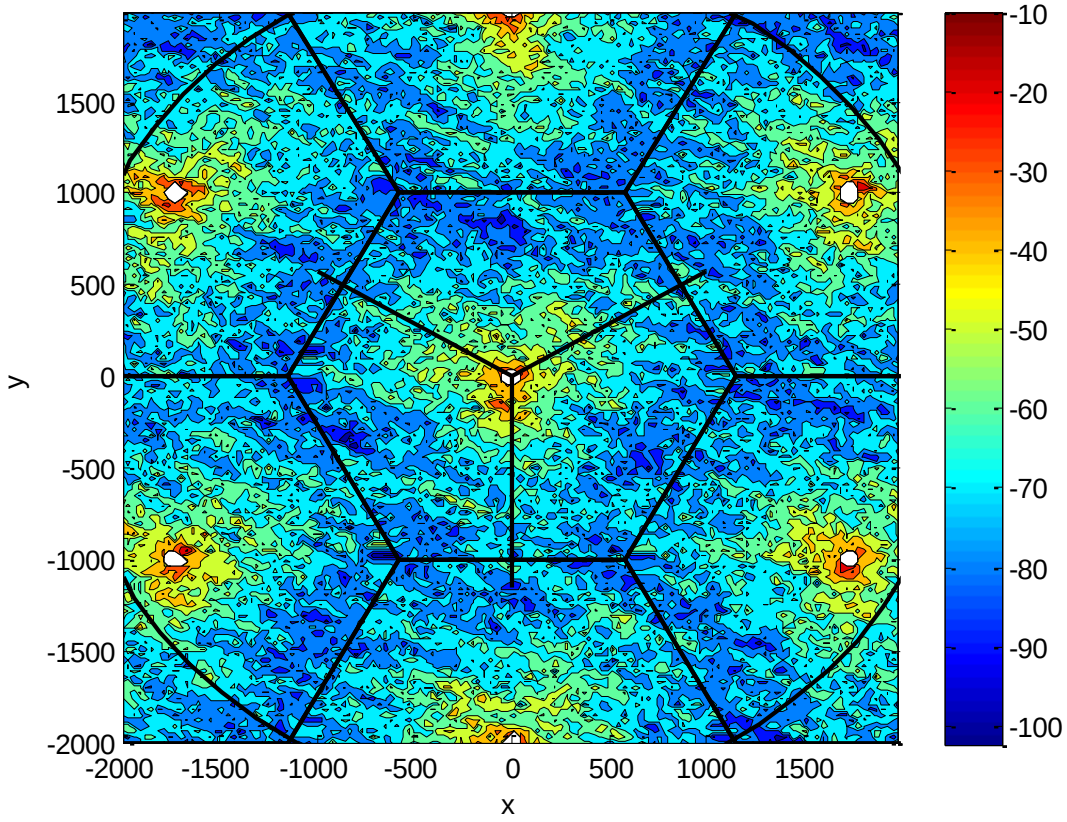


Figure 7.3 The geometric properties of the under study network: Received power [dBm]

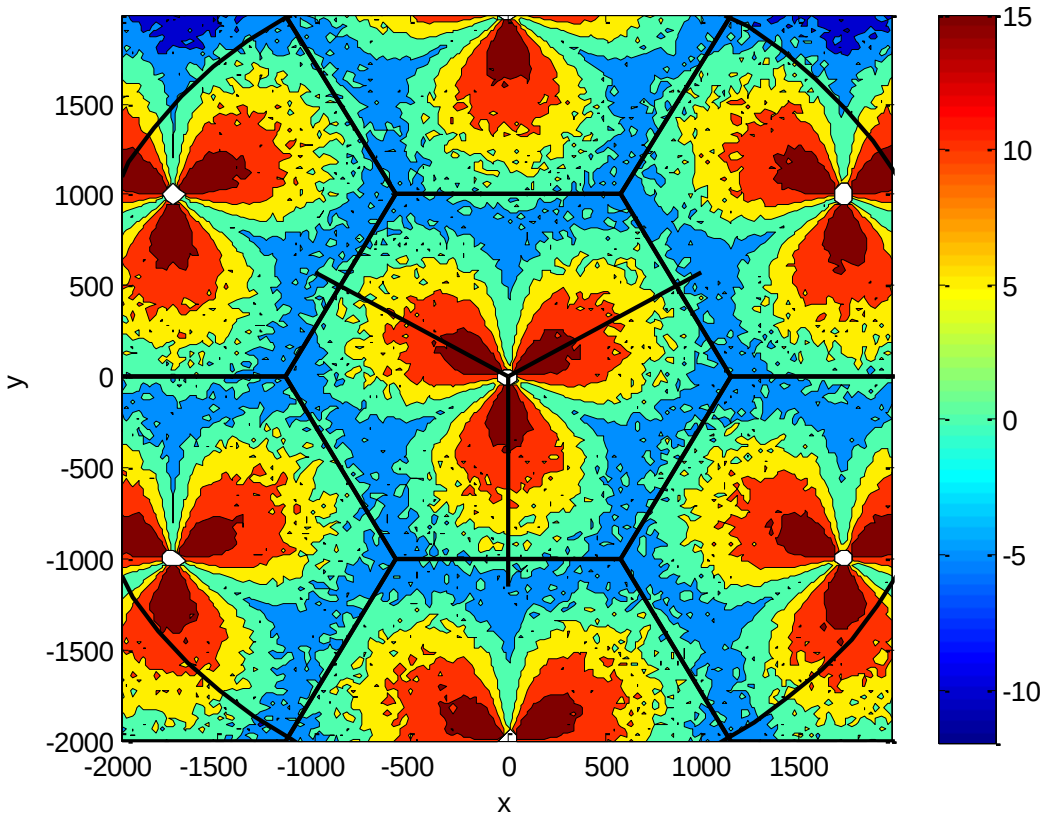


Figure 7.4 The geometric properties of the under study network: Geometric SINR [dB]

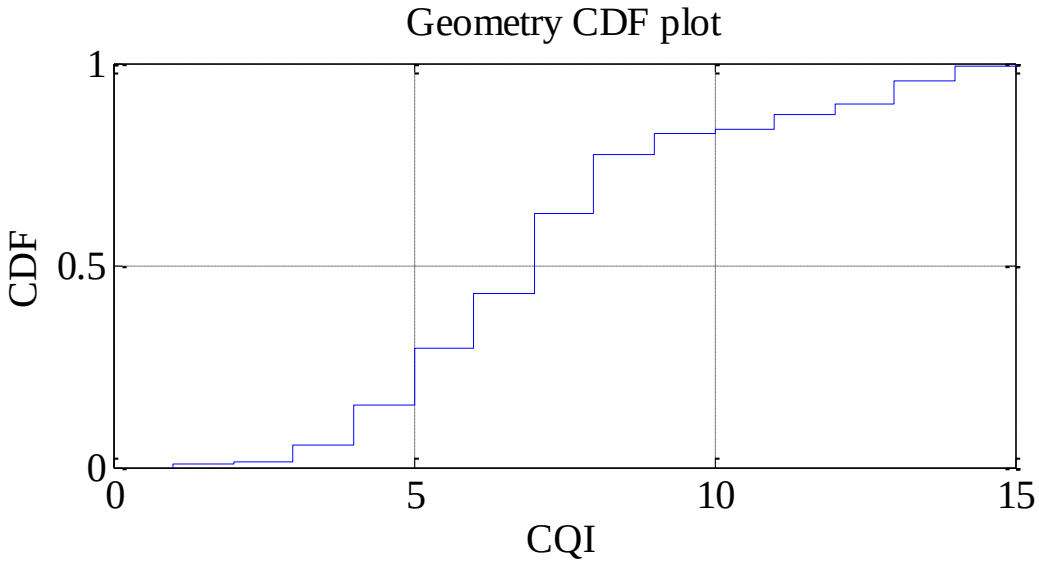


Figure 7.5 The geometric properties of the under study network: CQI values distribution

7.4.2 Results and Analysis

All the results presented in this section are obtained through five independent runs of the simulation for 100 randomly located users and 30 users with fixed locations. Users with fixed locations are capable of acting as a LTE mobile user and/or WiFi offloading device. Figure 7.6 shows the total 500 examined locations of the users and the coverage of the remaining 30 users when they act as WiFi access points. Figure 7.7 shows the total provided capacity, in terms of the allocated data rate, with and without WiFi interworking. And in Figure 7.8 the distribution of the achieved PI (i.e. QoE) is depicted. The comparison between the results of the implementation algorithm in (7-8) and those for the corresponding analytical model in (7-6) are shown to be closely matched in Figure 7.7 & Figure 7.8.

The WiFi hotspots with the independent backbone connection (i.e. I-WLAN) bring in extra capacity and increase the total capacity in the cell. In contrast, the standalone MiFi devices which rely on the capacity of the base station do not increase the available resources in the cell. However, due to their better reception quality compared to the neighboring mobile users, their link adaptation schemes provide a higher data rate for users under their coverage. The results in Figures 7.7 and 7.8 show that the macro-cell will benefit from interworking scenarios regardless of the type of the WiFi backbone connections. However, as it will be shown later in this section, this will not always be the case for any scheduling policy used by the base station.

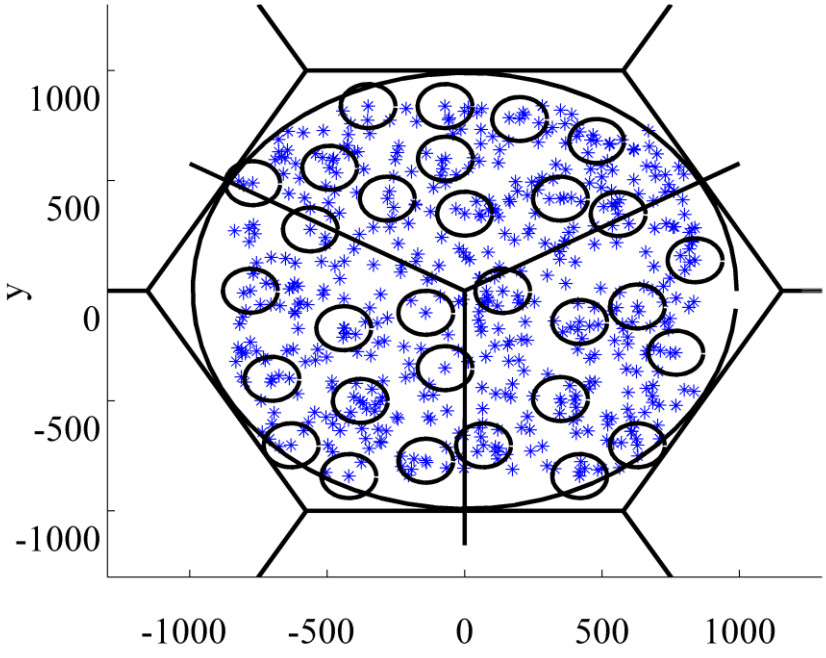


Figure 7.6 A snapshot of the users' locations and the WiFi access points' coverage

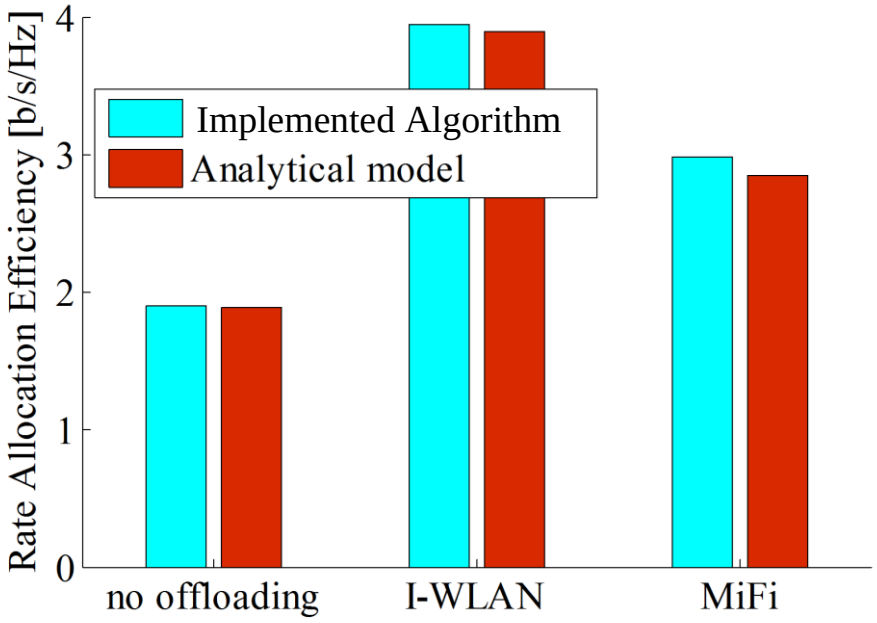


Figure 7.7 Comparisons between different LTE-WiFi interworking scenarios based on the proposed analytical model and the implemented algorithm in (7-6) and (7-8): Rate allocation efficiency in different scenarios

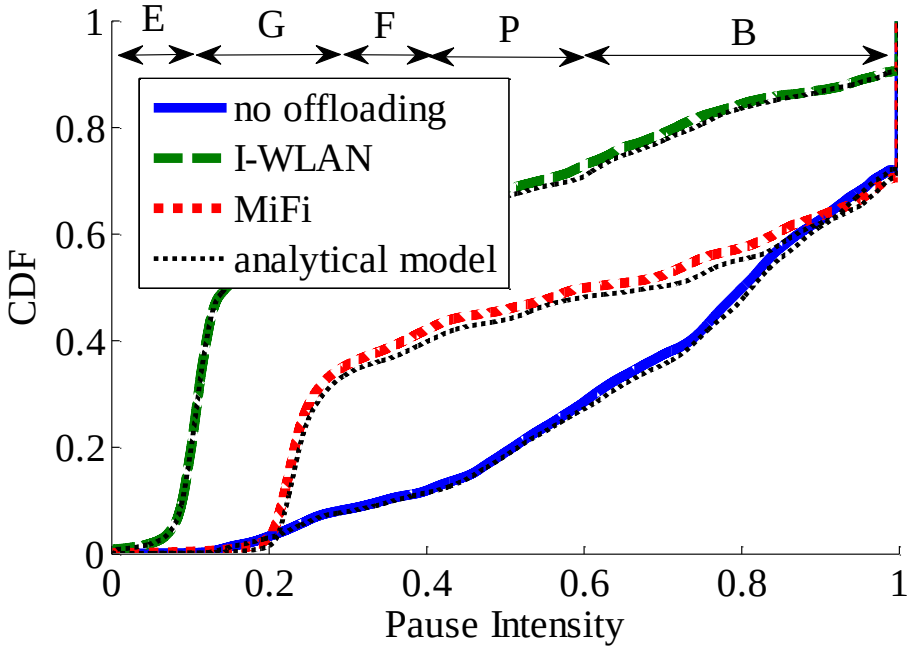


Figure 7.8 Comparisons between different LTE-WiFi interworking scenarios based on the proposed analytical model and the implemented algorithm in (7-6) and (7-8): Achieved QoE based on the PI values. The ranges of the values marked by E, G, F, P and B in (c) represent Excellent, Good, Fair, Poor and Bad qualities based on the corresponding MOS values in Figure 7.2

Figures 7.9 to 7.13 reveal how the interworking mechanisms discussed above reshape the distribution of the allocated data rate throughout the cell, in the form of two-dimensional geometric maps. The color bar on the right of Figure 7.9 shows the correspondence between the color used in the map and the allocated rate. The achieved distribution shown in Figure 7.9, where no offloading mechanism is applied, resembles the shape of the geometric SINR in the cell illustrated by Figure 7.10. In the case of I-WLAN offloading in Figure 7.10, the extra capacity provided by the independent backbone connection of the WiFi hotspots improves the capacity of the cell. Furthermore, Figure 7.10 demonstrates the effect of the WiFi hotspots' locations on the redistribution of the allocated data rate, compared to the case without offloading given in Figure 7.9.

Figures 7.11 to 7.13 show the redistributed allocated data rate for the MiFi interworking scenario. The scheduling algorithm used by the macro-cell base station is given in (7-8) with α=0.1, 1 and 10, respectively. As it can be seen, small α tends to enable a more efficient scheduling policy while large α will result in a more fair scheduling policy. These results reveal how the MiFi offloading performance will be affected by the scheduling policy. This is the consequence of the dependency of MiFi over the mobile base station. From Table 7.3, we can also see the trade-off between the fairness and efficiency of the system, i.e. the parameters representing both fairness and efficiency are proportional to the value of the scheduler parameter α but in the opposite trends. In the same table the achieved QoE for users are also exhibited through the comparison of the 50th percentile of the achieved MOS values in each case.

The contrast among the allocated rates, in terms of the difference in color displayed in the rate distribution maps in Figures 7.9 to 7.13, also indicates the levels of fairness and efficiency. The parameter α can be used to regulate the level of contrast to control the balance between fairness and performance. For example, increasing α will reduce the contrast of the rate redistribution map and make the system more fair in terms of resource allocation. However, a very fair scheduler like the one shown in Figure 7.13 would decrease the satisfaction level of users as m any premier users are allocated unacceptably low rates and the network efficiency measured by the throughput will be hampered as a result. The opposite is also true if a very efficient scheduler is used when α is small.

Table 7.3 MiFi Parameters

Scheduler parameter (α)	Achieved Efficiency (b/s/Hz)	Achieved fairness (J-index)	Achieved MOS Quality (50^{Th} percentile)
0.1	3.2	0.63	50% above 3.2 (Avg. Good)
1	2.9	0.68	50% above 2.1 (Avg. Fair)
10	2.16	0.84	50% above 1.6 (Avg. Poor)

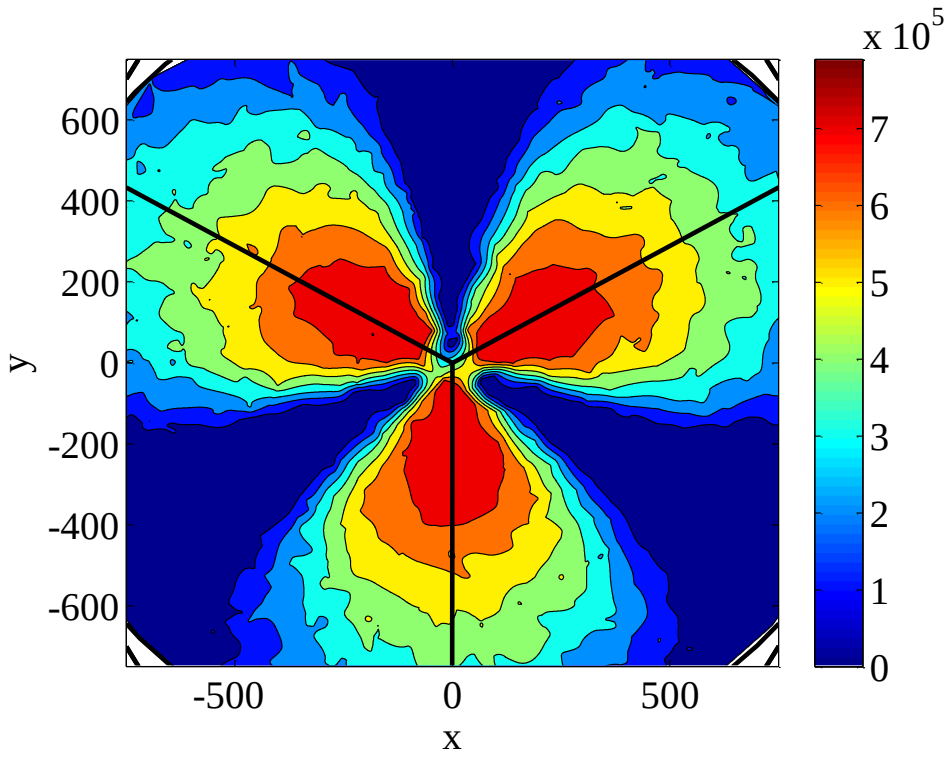


Figure 7.9 Allocated rate distribution throughout the cell: without traffic offloading

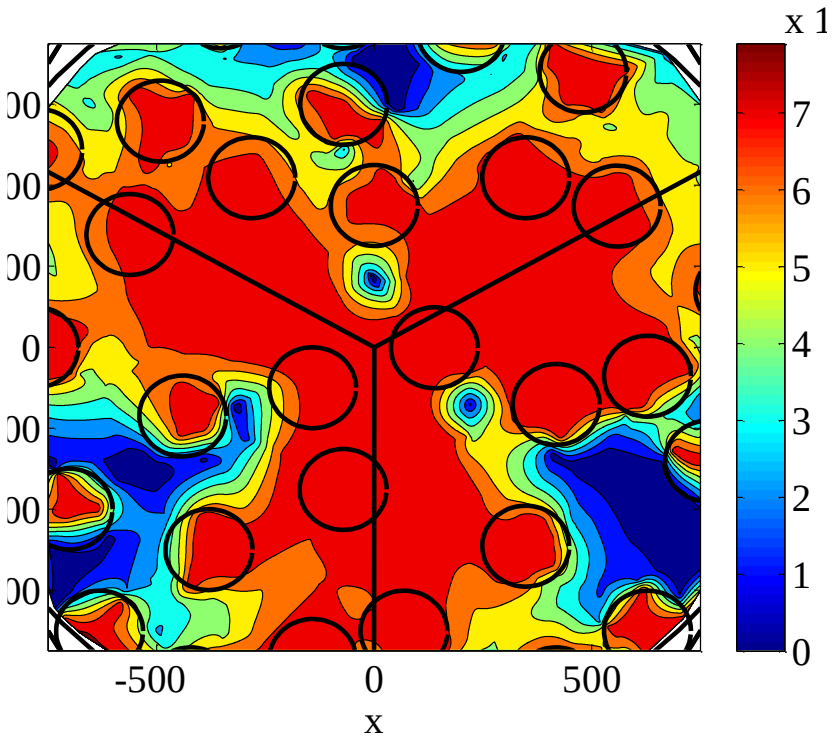


Figure 7.10 Allocated rate distribution throughout the cell: I-WLAN hotspots

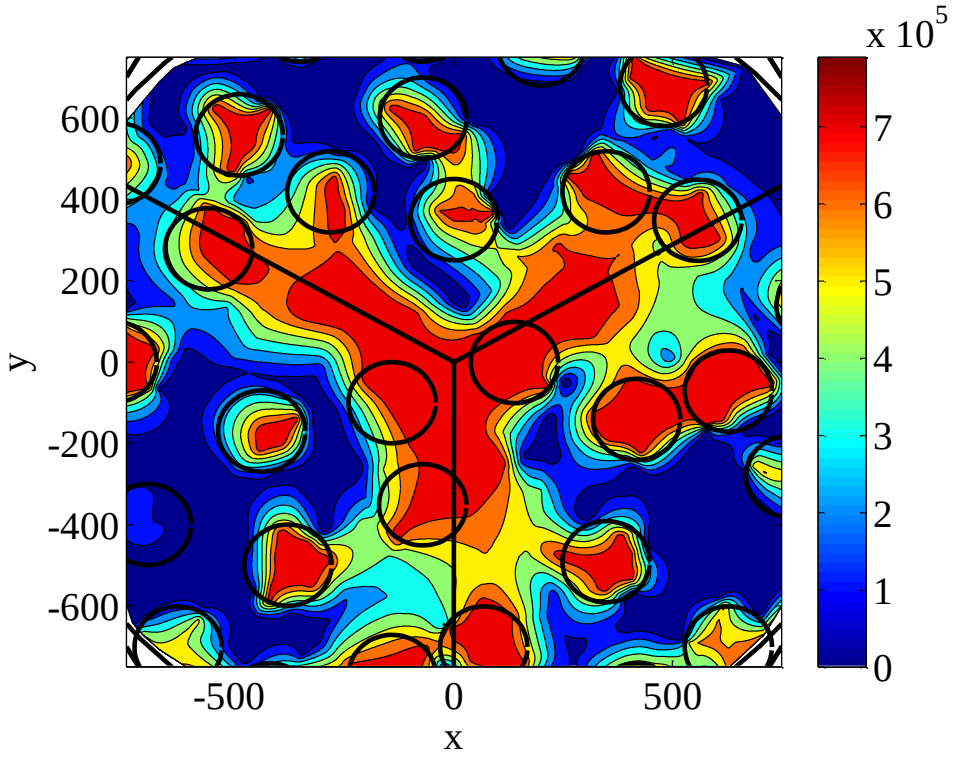


Figure 7.11 Allocated rate distribution throughout the cell: MiFi and an efficient scheduler (α=0.1)

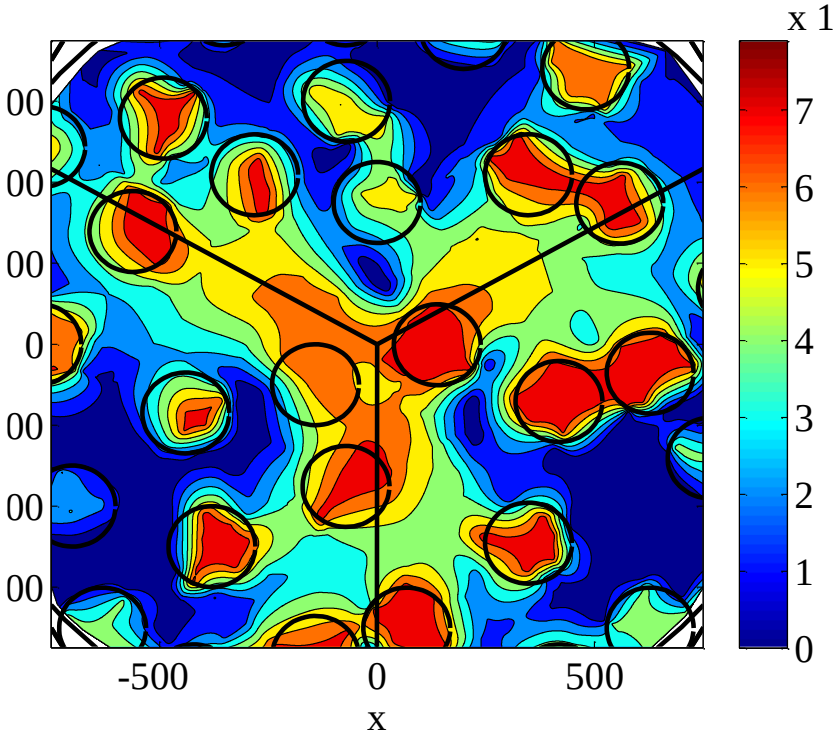


Figure 7.12 Allocated rate distribution throughout the cell: MiFi and an intermediate scheduler (α=1)

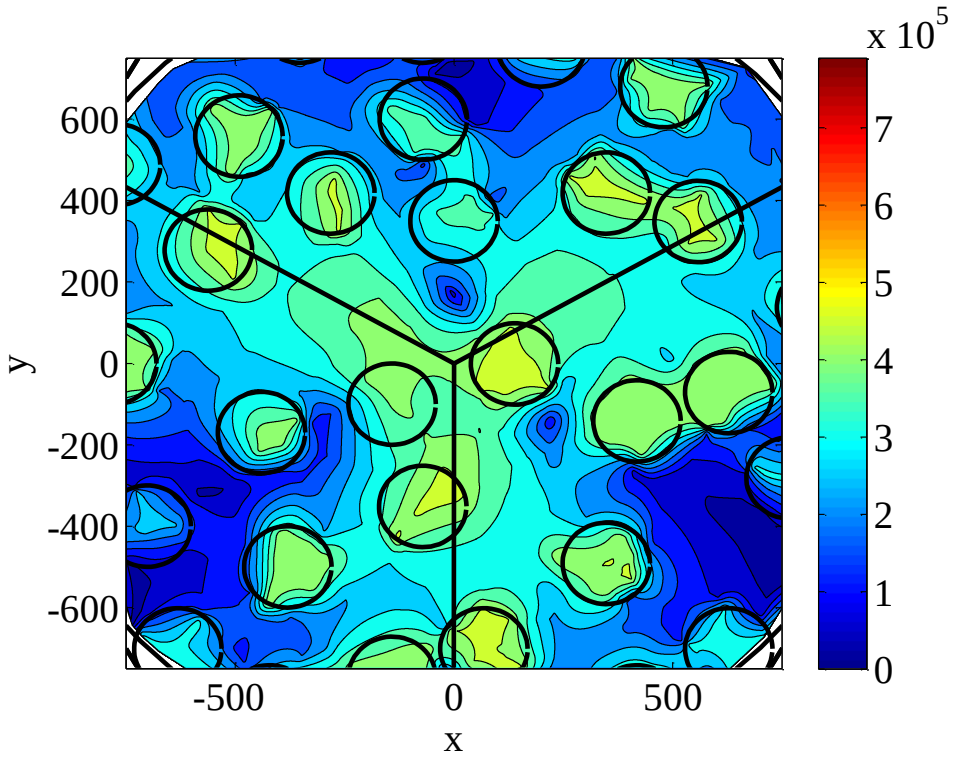


Figure 7.13 Allocated rate distribution throughout the cell: MiFi and a fair scheduler (α=10)

7.5 Summary

A rate redistribution scheme for traffic offloading through mobile WiFi devices with LTE backbone connection (MiFi), alongside the standard WiFi hotspots (I-WLAN), has been investigated in this chapter. The performance of various interworking scenarios, in terms of the allocated data rate per user, has been evaluated within a QoE driven and parametric scheduling algorithm. The distribution of the allocated data rate for a video streaming service in a macro-cell can be controlled to achieve the desired trade-off between fairness and efficiency of the mobile network. While WiFi-LTE interworking requires a fair scheduling policy to determine the rate distribution given the original geometric dependency in the cell, reducing the efficiency of the system for the sake of fairness will degrade achievable perceived quality in the cell. The rate redistribution algorithm proposed here is able to optimize the performance of WiFi offloading in the LTE network by applying the high-order PI metric, as described above, to achieve an aimed efficiency-fairness balance.

Chapter 8
Conclusion and Future Work

8.1 Summary of Conclusions

The main achievements of this work can be described into two areas:

1. A thorough analysis of the playback buffer behaviour through deriving a stochastic model and closed form formulae to characterize the 'pause-play' events during video playback, including the occurrence probability of the pause and play durations, average durations of pause and play during the playback, the frequency of the pause-play events and the intensity of the phenomena based on the defined quality metric, Pause Intensity; and
2. Optimisation of the resource utilisation through a quality driven approach by applying the Pause Intensity metric and in the context of wireless and mobile communications systems such as LTE and adaptive streaming services

The thesis has begun with providing technical background information required for the clarification of the streaming service studied and the chosen communications systems for the examination of the proposed ideas. The background study includes the review of the state of the art video delivery techniques and relevant quality assessment methods in Chapter 2. This provides an overview of the delivery network infrastructures, streaming protocols and services, in addition to the quality assessment techniques for video streaming. The support provided for the quality assessment methods has also been explained in relation to the new generations of mobile communication systems (e.g. 3GPP 3G and 4G-LTE). The publicly available standards, such as IETF-RTP and MPEG/3GPP-DASH, and the proprietary protocols such as HLS and IIS, have been reviewed and compared in this chapter.

An overview of the architecture, functionality and protocols of LTE has been provided in Chapter 3, covering the user (UE), the communication channel (air interface) and the base station as a network unit (eNodeB). The functions and protocols are explained based on the 3GPP standards, in

connection to the resource allocation unit and utilisation strategies, error control mechanisms, adaptive modulation and coding schemes. The implementation issues of LTE have also been compared with those used by the previous generations of the system, represented by HSPA. Our developed MATLAB-based simulator for LTE has been introduced in this chapter. This simulator has been used for the investigation and evaluation of the proposed analytical models and algorithms throughout the study. The main characteristics of the simulator have been explained on the overall performance of the simulated system and other features including the spatially correlated large scale.

Chapter 4 presents a comprehensive analysis of the playback buffer behaviour given the performance of the network (i.e. TCP throughput) and the demand of the user (i.e. video bitrate). The analysis provides the stochastic characteristic of the pause and play durations derived from the network performance which is represented by the packet loss rate and throughput probability distribution functions. It has been shown that the derived probability distribution for pause and play durations is independent of the type of the distribution function of throughput, though it depends on its mean and variance. This result has been used to simplify the model of the buffer occupancy in time domain and led to the derivation of the closed form formulae for pause duration, pause frequency and Pause Intensity (PI). Extensive simulation and subjective testing have been conducted to validate the accuracy of the proposed model and the correlation between PI and MOS, the well-known subjective metric for users' satisfaction. It has also been shown that PI outperforms other quality assessment metrics such as pause duration and pause frequency. Furthermore, the performance of PI is shown to be content independent in terms of the correlation with MOS.

In Chapter 5 the objective and no-reference quality metric, PI, is adopted in the resource utilisation principles of LTE to establish a quality-driven scheduling strategy, aimed to improve QoE in video service delivery in mobile wireless networks. The relationship between the demand of the users (video bitrates) and the provision by the network (allocated data rate) has also been investigated in terms of their correlation. The proposed model has been compared with other scheduling algorithm including *best-CQI*, *Proportional Fair* and *MaxMin throughput*. It has been shown that the proposed idea provides a tailored trade-off between efficiency and fairness when taking into account the correlation between the demand of the user and the supply of the network.

A parametric and adaptive scheduler based on the QoE-driven scheduling algorithm in Chapter 5 is proposed for adaptive video streaming services in Chapter 6. The proposed algorithm provides a flexible tool for achieving specific levels of fairness and efficiency. It has also been shown that the proposed scheduler can maintain a desired trade-off between fairness and efficiency through an online control algorithm. The implemented rate adaptation algorithm at the user-side employs a PI-based regulation process to control the video bitrate according to the channel status of the user. The QoE assessment through PI is shown to be feasible on both client and network sides without the need for extra communication between user and network.

Chapter 7 is dedicated to the rate redistribution mechanism across a macro-cell and through interworking between the base station (i.e. eNodeB in LTE) and wireless LAN access points. The performance of various interworking scenarios, in terms of the allocated data rate per user, has been studied in this chapter. Discussion, specifically, focused on the mobile WiFi devices with LTE backbone connection (also known as MiFi) alongside the standard WiFi hotspots (I-WLAN). Using the QoE-driven scheduling algorithm proposed in Chapter 6, the distribution of the allocated data rate for video streaming service in a macro-cell can be controlled to maintain the desired trade-off between fairness and efficiency of the mobile network. It has been shown that there is also a trade-off between the achievable QoE in the cell (based on PI) and the efficiency of the allocation algorithm in the base station of the macro cell.

8.2 Personal Contributions

Pause Intensity Model, Simulation and Subjective Testing

Playback buffer behaviour analysis based on the stochastic characteristics of the network performance has been employed for modelling a new no-reference quality metric (Pause Intensity), leading to an analytical model that has been verified by simulation and subjective testing. This work was initially presented at ICME 2012 and the extended paper has been published in IEEE Transactions on Circuit and Systems for Video Technology in 2013.

A Quality-Driven Scheduling for Video-Based Services in LTE

The proposed objective and no-reference quality metric (PI) is employed in the establishment of a quality driven scheduling algorithm in LTE to improve QoE in mobile wireless networks. The initial work was presented at the European Wireless conference 2013 (EW-2013) and the follow-on work was later presented at IEEE WCNC 2014.

A Parametric Adaptive Scheduler Joint with Adaptive Rate Video Streaming

Pause Intensity is capable of being evaluated in both user-side and network-side to represent the experienced quality of the service by client. This property has been used to establish a framework of parametric and adaptive Scheduler aimed to achieve a desired trade-off between fairness and efficiency and best perform in the presence of adaptive video streaming. This work was published in the IEEE IUCC-2014 proceeding.

Rate Redistribution Evaluation for LTE-WLAN Interworking

Interworking between a macro-cell base station and surrounding short-range WiFi access points, for offloading purposes, reshapes the distribution of the allocated rate to the users across the macro-

cell. An evaluation method, driven by Pause Intensity, is developed to quantify and compare the achieved QoE in such a network.

8.3 Future Work

The analytical modelling of buffer behaviour in this work has resulted in general expressions representing the stochastic attributes of pause and play durations in a video streaming service. However, the exact type of the probability distribution function has not been identified. It would be more beneficial to investigate the classification of the probability distribution function and its relationship with the stochastic properties of the network performance based on the network throughput or packet loss rate. Furthermore, the subjective testing and simulation for the evaluation of Pause Intensity was based on a fixed rate video streaming service. Therefore, further evaluations will be carried out to provide the PI-MOS correlation diagrams in the adaptive rate video streaming environment.

Our proposed quality-driven scheduling algorithm has been examined in the downlink of LTE and using the FDD technology. Nevertheless, the adopted characteristics of LTE in our simulator allow the results to be extended and generalised for covering the case of TDD, uplink and other contemporary mobile communication functions such as those in HSPA and LTE-Advanced. The differences and similarities between FDD and TDD as well as the HSPA and LTE have initially been studied and will be investigated further to understand to what extent these results are applicable to the state of the art wireless and mobile communication technologies

Considering the forthcoming resource utilisation technologies in LTE-Advanced such as carrier aggregation and the evolutionary changes of service provisioning in 5G including the new coding/streaming techniques for multimedia content distribution, more technical aspects based on the currently proposed model will be investigated and accommodated. The proposed utilisation model will also be extended from just for video content to the combination of video, voice and text information.

8.4 Related Publications

Journal:

- **Seyedebrahimi**, M.; Bailey, C.; Xiao-Hong Peng, "*Model and Performance of a No-Reference Quality Assessment Metric for Video Streaming,*" Circuits and Systems for Video Technology, IEEE Transactions on , vol.23, no.12, pp.2034,2043, Dec. 2013

Conferences:

- **Seyedebrahimi**, M.; Xiao-Hong Peng, "Ensuring QoE in Contemporary Mobile Networks for Video Content Distribution", IEEE INFOCOM Workshop, Apr. 2015.
- **Seyedebrahimi**, M.; Xiao-Hong Peng; Rob Harrison, "Adaptive Resource Allocation for QoE-Aware Mobile Communication Networks", 2014 IEEE 17th International Conference on Computational Science and Engineering (CSE), pp. 868-875, 19-21 Dec. 2014.
- **Seyedebrahimi**, M.; Xiao-Hong Peng; Rob Harrison, *"A Quality Driven Framework for Adaptive Video Streaming in Mobile Wireless Networks*", Wireless Communications and Networking Conference (WCNC), 2014 IEEE , pp.2994-2999, 6-9 April 2014.
- **Seyedebrahimi**, M; Peng, Xiao-Hong; Harrison, R, "*A New Scheduling Method for Enhanced Quality of Experience in LTE Systems*", European Wireless 2013 - 19th European Wireless Conference, Guildford, UK, 16-18 April 2013.
- Bailey, C.; **Seyedebrahimi**, M.; Xiao-Hong Peng, "Pause Intensity*: A No-Reference Quality Assessment Metric for Video Streaming in TCP Networks*," Multimedia and Expo (ICME), 2012 IEEE International Conference on, pp.818-823, 9-13 July 2012.

References

[1] T. Porter and X.-H. Peng, "An Objective Approach to Measuring Video Playback Quality in Lossy Networks using TCP," *Communications Letters, IEEE*, vol. 15, pp. 76-78, 2011.

[2] ITU-T, "Recommendation H.264, Advanced video coding for generic audiovisual services," International Telecommunication Union, Feb. 2014.

[3] ISO/IEC, "14496-10:2012 – Information technology — coding of audio-visual objects — Part 10: Advanced Video Coding," Retrieved 2012-06-11.

[4] ITU-T, "Recommendation H.265, High efficiency video coding," International Telecommunication Union, Apr. 2013.

[5] Cisco, "Visual Networking Index: Forecast and Methodology, 2013–2018," Retrieved 10-10-2014.

[6] Conviva. (Retrieved 10-10-2014). *2014 viewer experience report: http://www.conviva.com/vxr-home/*.

[7] T. Szigeti and C. Hattingh, "End-to-End QoS Network Design: Quality of Service in LANs, WANs, and VPNs (Networking Technology)," Cisco Press, 9 Nov 2004.

[8] T. Yufang, L. Xueming, L. Yang, L. Chenyu, and X. Yan, "Review of content distribution network architectures," in *Computer Science and Network Technology (ICCSNT), 2013 3rd International Conference on*, 2013, pp. 777-782.

[9] 3GPP, "TS 26.233: Transparent end-to-end packet switched streaming service (PSS); General description," Retrieved 10-10-2014.

[10] H. Nishiyama, H. Yamada, H. Yoshino, and N. Kato, "A Cooperative User-System Approach for Optimizing Performance in Content Distribution/Delivery Networks," *Selected Areas in Communications, IEEE Journal on*, vol. 30, pp. 476-483, 2012.

[11] T. Hai Anh, S. Hoceini, A. Mellouk, J. Perez, and S. Zeadally, "QoE-Based Server Selection for Content Distribution Networks," *Computers, IEEE Transactions on*, vol. 63, pp. 2803-2815, 2014.

[12] S. M. Y. Seyyedi and B. Akbari, "Hybrid CDN-P2P architectures for live video streaming: Comparative study of connected and unconnected meshes," in *Computer Networks and Distributed Systems (CNDS), 2011 International Symposium on*, 2011, pp. 175-180.

[13] 3GPP, "3GPP LTE and LTE-Advanced Technology, http://www.3gpp.org/ftp/Specs/html-info/36-series.htm," Retrieved 10-10-2014.

[14] ITU-R, "Report M.2134, Requirements related to technical performance for IMT-Advanced radio interface(s)," Approved in Nov 2008.

[15] 3GPP, "TS 26.234: Transparent end-to-end packet switched streaming service (PSS); Protocols and codecs," Retrieved 10-10-2014.

[16] 3GPP, "TS 22.233: Transparent end-to-end packet switched streaming service; Stage 1," Retrieved 10-10-2014.

[17] 3GPP, "TS 26.244: Transparent end-to-end packet switched streaming service (PSS); 3GPP file format (3GP)," Retrieved 10-10-2014.

[18] 3GPP, "TS 22.228: Service requirements for the Internet Protocol (IP) multimedia core network subsystem (IMS); Stage 1,".

[19] 3GPP, "TS 26.237: IP Multimedia Subsystem (IMS) based Packet Switch Streaming (PSS) and Multimedia Broadcast/Multicast Service (MBMS) User Service; Protocols,".

[20] 3GPP, "TS 26.247: Transparent end-to-end Packet-switched Streaming Service (PSS); Progressive Download and Dynamic Adaptive Streaming over HTTP (3GP-DASH),".

[21] C. Joonho, Y. Myungsik, and B. Mukherjee, "Efficient Video-on-Demand Streaming for Broadband Access Networks," *Optical Communications and Networking, IEEE/OSA Journal of*, vol. 2, pp. 38-50, 2010.

[22] C. Zhijia, L. Chuang, and W. Xiaogang, "Enabling on-demand internet video streaming services to multi-terminal users in large scale," *Consumer Electronics, IEEE Transactions on*, vol. 55, pp. 1988-1996, 2009.

[23] P. Sun Sik, H. Myung Jin, M. Sung Soo, and Y. Hee Yong, "An Efficient VoD Scheme Providing Service Continuity for Mobile IPTV in Heterogeneous Networks," in *Computer and Information Technology (CIT), 2010 IEEE 10th International Conference on*, 2010, pp. 2589-2595.

[24] M. M. Fouda, T. Taleb, M. Guizani, and N. Kato, "Towards efficient P2P-based VoD provisioning in future internet," in *Communications and Networking in China (CHINACOM), 2010 5th International ICST Conference on*, 2010, pp. 1-5.

[25] IETF, "HTTP Live Streaming, R. Pantos: https://tools.ietf.org/html/draft-pantos-http-live-streaming-14," Retrieved 20-11-2014.

[26] A. Systems, "HTTP Dynamic Streaming, Adobe Systems: http://www.adobe.com/products/hds-dynamic-streaming.html,".

[27] Microsoft, "IIS Smooth Streaming Technical Overview: http://www.microsoft.com/en-gb/download/details.aspx?id=17678,".

[28] Akamai, "Media Content Delivery: http://www.akamai.com/html/solutions/media-delivery.html,".

[29] Limelight, "Limelight Orchestrate, Content Delivery Network (CDN): http://www.limelight.com/services/orchestrate-content-delivery.html,".

[30] IETF, "RFC 3550, RTP: A Transport Protocol for Real-Time Applications,".

[31] ISO/IEC, "CD 23008-1, Information technology — High efficiency coding and media delivery in heterogeneous environments — Part 1: MPEG media transport (MMT)," Retrieved 20-11-2014.

[32] ISO/IEC, "23009-1:2012, Information technology -- Dynamic adaptive streaming over HTTP (DASH) -- Part 1: Media presentation description and segment formats,".

[33] IETF, "RFC 2616, Hypertext Transfer Protocol -- HTTP/1.1, R. Fielding,".

[34] T. Stockhammer, "Dynamic adaptive streaming over HTTP --: standards and design principles," presented at the Proceedings of the second annual ACM conference on Multimedia systems, San Jose, CA, USA, 2011.

[35] "The 5th revision of Hypertext Markup Language (HTML5), World Wide Web Consortium: http://www.w3.org/TR/html5/," Retrieved 02-12-2014.

[36] C. Bailey, "Video Quality Assessment for Modern Video Streaming," Doctor of Philosophy thesis, Aston University, October 2013.

[37] O. Yen-Fu, L. Tao, Z. Zhi, M. Zhan, and W. Yao, "Modeling the impact of frame rate on perceptual quality of video," in *Image Processing, 2008. ICIP 2008. 15th IEEE International Conference on*, 2008, pp. 689-692.

[38] C. Lee, J. Lee, S. Lee, K. Lee, H. Choi, G. Seo, *et al.*, "Full reference video quality assessment for multimedia applications," in *Proceeding of 10th WSEAS international conference on electronics, hardware, wireless and optical communications*, 2011, pp. 206-209.

[39] A. P. Hekstra, J. G. Beerends, D. Ledermann, F. E. de Caluwe, S. Kohler, R. H. Koenen, *et al.*, "PVQM – A perceptual video quality measure," *Signal Processing: Image Communication*, vol. 17, pp. 781-798, 11// 2002.

[40] A. Bhat, S. Kannangara, Z. Yafan, and I. Richardson, "A Full Reference Quality Metric for Compressed Video Based on Mean Squared Error and Video Content," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 22, pp. 165-173, 2012.

[41] A. Bhat, I. Richardson, and S. Kannangara, "A new perceptual quality metric for compressed video," in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, 2009, pp. 933-936.

[42] R. Dosselmann and X. Yang, "A comprehensive assessment of the structural similarity index," *Signal, Image and Video Processing*, vol. 5, pp. 81-91, 2011/03/01 2011.

[43] S. Chikkerur, V. Sundaram, M. Reisslein, and L. J. Karam, "Objective Video Quality Assessment Methods: A Classification, Review, and Performance Comparison," *Broadcasting, IEEE Transactions on*, vol. 57, pp. 165-182, 2011.

[44] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *Signals, Systems and Computers, 2004. Conference Record of the Thirty-Seventh Asilomar Conference on*, 2003, pp. 1398-1402 Vol.2.

[45] A. K. Moorthy and A. C. Bovik, "A motion compensated approach to video quality assessment," in *Signals, Systems and Computers, 2009 Conference Record of the Forty-Third Asilomar Conference on*, 2009, pp. 872-875.

[46] *The Handbook of MPEG Applications: Standards in Practice*: John Wiley & Sons, November 2010.

[47] A. B. Watson, Q. J. Hu, J. F. McGowan Iii, and J. B. Mulligan, "Design and performance of a digital video quality metric," in *SPIE 3644, Human Vision and Electronic Imaging IV*, 1999, pp. 168-174.

[48] X. H. Zhang, W. S. Lin, and P. Xue, "Improved estimation for just-noticeable visual distortion," *Signal Processing*, vol. 85, pp. 795-808, 4// 2005.

[49] ITU-T, "Recommendation J.246, Perceptual visual quality measurement techniques for multimedia services over digital cable television networks in the presence of a reduced bandwidth reference," International Telecommunication Union, 2008.

[50] P. Marziliano, F. Dufaux, S. Winkler, and T. Ebrahimi, "A no-reference perceptual blur metric," in *Image Processing. 2002. Proceedings. 2002 International Conference on*, 2002, pp. III-57-III-60 vol.3.

[51] F. Crete, T. Dolmiere, P. Ladret, and M. Nicolas, "The blur effect: perception and estimation with a new no-reference perceptual blur metric," 2007, pp. 64920I-64920I-11.

[52] K. T. Tan and M. Ghanbari, "Frequency domain measurement of blockiness in MPEG-2 coded video," in *Image Processing, 2000. Proceedings. 2000 International Conference on*, 2000, pp. 977-980 vol.3.

[53] T. Kim, N. Avadhanam, and S. Subramanian, "Dimensioning Receiver Buffer Requirement for Unidirectional VBR Video Streaming over TCP," in *Image Processing, 2006 IEEE International Conference on*, 2006, pp. 3061-3064.

[54] ITU-T, "J.247, Objective perceptual multimedia video quality measurement in the presence of a full reference," International Telecommunication Union, Aug. 2008.

[55] ITU-T, "J.341, Objective perceptual multimedia video quality measurement of HDTV for digital cable television in the presence of a full reference," International Telecommunication Union, Jan. 2011.

[56] 3GPP, "TS 23.401: General Packet Radio Service (GPRS) enhancements for Evolved Universal Terrestrial Radio Access Network (E-UTRAN) access,".

[57] 3GPP, "TS 36.401: Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Architecture description,".

[58] 3GPP, "TS 23.104: Evolved Universal Terrestrial Radio Access (E-UTRA); Base Station (BS) radio transmission and reception,".

[59] 3GPP, "TS 23.101: Evolved Universal Terrestrial Radio Access (E-UTRA); User Equipment (UE) radio transmission and reception,".

[60] 3GPP, "TS 36.420: Evolved Universal Terrestrial Radio Access Network (E-UTRAN); X2 general aspects and principles,".

[61] 3GPP, "TS 36.410: Evolved Universal Terrestrial Radio Access Network (E-UTRAN); S1 general aspects and principles,".

[62] 3GPP, "TS 36.211: Evolved Universal Terrestrial Radio Access (E-UTRA); Physical channels and modulation,".

[63] J. She, M. Jingqing, J. Ho, H. Pin-Han, and J. Hong, "Layered Adaptive Modulation and Coding for 4G Wireless Networks," in *Global Telecommunications Conference (GLOBECOM 2010), 2010 IEEE*, 2010, pp. 1-6.

[64] C. Fa-tang and T. Gen-lin, "A novel MCS selection criterion for supporting AMC in LTE system," in *Computer Application and System Modeling (ICCASM), 2010 International Conference on*, 2010, pp. V6-598-V6-603.

[65] ITU-R, "Recommendation M.1225, Guidelines for evaluation of radio transmission, Technologies for IMT-2000," 1997.

[66] H. Zhongqiu and Z. Fei, "Performance of HARQ with AMC Schemes in LTE Downlink," in *Communications and Mobile Computing (CMC), 2010 International Conference on*, 2010, pp. 250-254.

[67] 3GPP, "TS 36.212: Evolved Universal Terrestrial Radio Access (E-UTRA); Multiplexing and channel coding,".

[68] 3GPP, "TS 36.213: Evolved Universal Terrestrial Radio Access (E-UTRA); Physical layer procedures,".

[69] S. Lin and D. J. Costello, "Error Control Coding," 2nd ed., ISBN 0-13-042672-5, Upper Saddle River, NJ: Prentice-Hall, 2004.

[70] E. Dahlman, S. Parkvall, J. Sköld, and P. Beming, *3G Evolution, HSPA and LTE for Mobile Broadband*, 2nd ed., ISBN: 978-0-12-374538-5, Published by Elsevier Ltd, 2008.

[71] 3GPP, "TS 25.212: 3rd Generation Partnership Project; Technical Specification Group Radio Access Network; Multiplexing and channel coding (FDD)," Retrieved 10-10-2014.

[72] 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*, 2005, pp. 512-516.

[73] M. Gudmundson, "Correlation model for shadow fading in mobile radio systems," *Electronics Letters*, vol. 27, pp. 2145-2146, 1991.

[74] Q. Huynh-Thu and M. Ghanbari, "Scope of validity of PSNR in image/video quality assessment," *Electronics Letters*, vol. 44, pp. 800-801, 2008.

[75] Z. Yanan, G. Xiangyang, W. Wendong, and Q. Xirong, "A rate adaptive algorithm for HTTP streaming," in *Cloud Computing and Intelligent Systems (CCIS), 2012 IEEE 2nd International Conference on*, 2012, pp. 529-532.

[76] C. Bailey, M. Seyedebrahimi, and P. Xiao-Hong, "Pause Intensity: A No-Reference Quality Assessment Metric for Video Streaming in TCP Networks," in *Multimedia and Expo (ICME), 2012 IEEE International Conference on*, 2012, pp. 818-823.

[77] R. Feghali, W. Demin, F. Speranza, and A. Vincent, "Quality metric for video sequences with temporal scalability," in *Image Processing, 2005. ICIP 2005. IEEE International Conference on*, 2005, pp. III-137-40.

[78] C. An, Z. Kai, P. Mohapatra, L. Sung-Ju, and S. Banerjee, "Metrics for Evaluating Video Streaming Quality in Lossy IEEE 802.11 Wireless Networks," in *INFOCOM, 2010 Proceedings IEEE*, 2010, pp. 1-9.

[79] Z. Wang, L. Lu, and A. C. Bovik, *Video Quality Assessment Based on Structural Distortion Measurement*, 2004.

[80] M. H. Pinson and S. Wolf, "A new standardized method for objectively measuring video quality," *Broadcasting, IEEE Transactions on*, vol. 50, pp. 312-322, 2004.

[81] SwissQual, "Ensuring quality and consistency across your entire network, http://www.swissqual.com/," Retrieved 10-10-2014.

[82] K.WILL, "Objective Analysis of Your Video Quality In Real-Time, http://www.kwillcorporation.com/products/VP21H.html," Retrieved 10-10-2014.

[83] D. Schroeder, A. El Essaili, E. Steinbach, D. Staehle, and M. Shehada, "Low-Complexity No-Reference PSNR Estimation for H.264/AVC Encoded Video," in *Packet Video Workshop (PV), 2013 20th International*, 2013, pp. 1-6.

[84] R. Ferzli and L. J. Karam, "A No-Reference Objective Image Sharpness Metric Based on the Notion of Just Noticeable Blur (JNB)," *Image Processing, IEEE Transactions on,* vol. 18, pp. 717-728, 2009.

[85] H. Liu and I. Heynderickx, "A perceptually relevant no-reference blockiness metric based on local image characteristics," *EURASIP Journal on Advances in Signal Processing,* vol. 2009, p. 2, 01/01/2009 2009.

[86] M. Mitsumura, H. Masuyama, S. Kasahara, and Y. Takahashi, "Buffer-overflow and starvation probabilities for video streaming services with application-layer rate-control mechanism," in *Proceedings of the 6th International Conference on Queueing Theory and Network Applications*, 2011, pp. 134-138.

[87] C. Zhifeng and Y. Reznik, "Analysis of video codec buffer and delay under time-varying channel," in *Visual Communications and Image Processing (VCIP), 2012 IEEE*, 2012, pp. 1-6.

[88] M. Kalman, E. Steinbach, and B. Girod, "Adaptive media playout for low-delay video streaming over error-prone channels," *Circuits and Systems for Video Technology, IEEE Transactions on,* vol. 14, pp. 841-851, 2004.

[89] S. Lee and K. Chung, "Buffer-driven adaptive video streaming with TCP-friendliness," *Computer Communications,* vol. 31, pp. 2621-2630, 6/25/ 2008.

[90] Y. H. Jung and Y. Choe, "Resource-aware and quality-fair video-streaming using multiple adaptive TCP connections," *Computers & Electrical Engineering,* vol. 36, pp. 702-717, 7// 2010.

[91] A. Raghuveer, E. Kusmierek, and D. H. C. Du, "A Network-Aware Approach for Video and Metadata Streaming," *Circuits and Systems for Video Technology, IEEE Transactions on,* vol. 17, pp. 1028-1040, 2007.

[92] K. J. Ma, R. Bartoš, and S. Bhatia, "Review: A survey of schemes for Internet-based video delivery," *Journal of Network and Computer Applications,* vol. 34, pp. 1572-1586, 09/01/2011 2011.

[93] N. B. Yoma, C. Busso, and I. Soto, "Packet-loss modelling in IP networks with state-duration constraints," *Communications, IEE Proceedings-*, vol. 152, pp. 1-5, 2005.

[94] Y. Jinyao, W. Muhlbauer, and B. Plattner, "Analytical Framework for Improving the Quality of Streaming Over TCP," *Multimedia, IEEE Transactions on,* vol. 14, pp. 1579-1590, 2012.

[95] G. Hasslinger and O. Hohlfeld, "The Gilbert-Elliott Model for Packet Loss in Real Time Services on the Internet," in *Measuring, Modelling and Evaluation of Computer and Communication Systems (MMB), 2008 14th GI/ITG Conference* -, 2008, pp. 1-15.

[96] J. Padhye, V. Firoiu, D. F. Towsley, and J. F. Kurose, "Modeling TCP Reno performance: a simple model and its empirical validation," *Networking, IEEE/ACM Transactions on,* vol. 8, pp. 133-145, 2000.

[97] B. Wang, J. Kurose, P. Shenoy, and D. Towsley, "Multimedia streaming via TCP: An analytic performance study," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM),* vol. 4, p. 16, 05/01/2008 2008.

[98] ITU-T, "Recommendation P.911, Subjective audiovisual quality assessment methods for multimedia applications," 1998.

[99] "Subjective Testing of PI, http://www.baileyc1.ht.aston.ac.uk/new/start_page.php," Retrieved 10-10-2014.

[100] J. F. K. a. E. S. Keeping, *Mathematics of Statistics*, Pt. 1, 3rd ed. Princeton, NJ: Van Nostrand, pp. 252-285, 1962.

[101] M. Seyedebrahimi, C. Bailey, and P. Xiao-Hong, "Model and Performance of a No-Reference Quality Assessment Metric for Video Streaming," *Circuits and Systems for Video Technology, IEEE Transactions on,* vol. 23, pp. 2034-2043, 2013.

[102] 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,* vol. 15, pp. 678-700, 2013.

[103] E. Dahlman, S. Parkvall, and J. Skold, *4G: LTE/LTE-Advanced for Mobile Broadband*, Academic Press, 1 edition 22 Mar. 2011.

[104] A. El Essaili, Z. Liang, D. Schroeder, E. Steinbach, and W. Kellerer, "QoE-driven live and on-demand LTE uplink video transmission," in *Multimedia Signal Processing (MMSP), 2011 IEEE 13th International Workshop on*, 2011, pp. 1-6.

[105] G. Piro, L. A. Grieco, G. Boggia, R. Fortuna, and P. Camarda, "Two-Level Downlink Scheduling for Real-Time Multimedia Services in LTE Networks," *Multimedia, IEEE Transactions on,* vol. 13, pp. 1052-1065, 2011.

[106] S. Singh, O. Oyman, A. Papathanassiou, D. Chatterjee, and J. G. Andrews, "Video capacity and QoE enhancements over LTE," in *Communications (ICC), 2012 IEEE International Conference on*, 2012, pp. 7071-7076.

[107] J. Navarro-Ortiz, P. Ameigeiras, J. M. Lopez-Soler, J. Lorca-Hernando, Q. Perez-Tarrero, and R. Garcia-Perez, "A QoE-Aware Scheduler for HTTP Progressive Video in OFDMA Systems," *Communications Letters, IEEE,* vol. 17, pp. 677-680, 2013.

[108] M. Seyedebrahimi, X.-H. Peng, and R. Harrison, "A New Scheduling Method for Enhanced Quality of Experience in LTE Systems," in *Wireless Conference (EW), Proceedings of the 2013 19th European*, 2013, pp. 1-6.

[109] L. Hao and L. Huixi, "A Research of Resource Scheduling Strategy for Cloud Computing Based on Pareto Optimality M□ Production Model," in *Management and Service Science (MASS), 2011 International Conference on*, 2011, pp. 1-5.

[110] K. Hoon and H. Youngnam, "A proportional fair scheduling for multicarrier transmission systems," *Communications Letters, IEEE,* vol. 9, pp. 210-212, 2005.

[111] C. Mehlführer, M. Wrulich, J. C. Ikuno, D. Bosanska, and M. Rupp, "Simulating the Long Term Evolution Physical Layer," In Proceedings of the 17th European Signal Processing Conference (EUSIPCO 2009), (Aug. 2009), Glasgow, Scotland, pp.1471-1478.

[112] "Enhanced UMTS Radio Access Network Extension for ns-2, http://eurane.ti-wmc.nl/eurane/," Retrieved 15-7-2012.

[113] Cisco, "Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2013–2018," San Jose, CA, USA: Cisco Systems, Inc., Feb. 2014.

[114] L. De Cicco and S. Mascolo, "An Adaptive Video Streaming Control System: Modeling, Validation, and Performance Evaluation," *Networking, IEEE/ACM Transactions on,* vol. 22, pp. 526-539, 2014.

[115] W. Jingjing, T. Z. J. Fu, C. Dah Ming, and L. Zhibin, "Perceptual quality assessment on B-D tradeoff of P2P assisted layered video streaming," in *Visual Communications and Image Processing (VCIP), 2011 IEEE*, 2011, pp. 1-4.

[116] C. Songqing, S. Bo, S. Wee, and Z. Xiaodong, "SProxy: A Caching Infrastructure to Support Internet Streaming," *Multimedia, IEEE Transactions on,* vol. 9, pp. 1062-1072, 2007.

[117] R. K. P. Mok, E. W. W. Chan, and R. K. C. Chang, "Measuring the quality of experience of HTTP video streaming," in *Integrated Network Management (IM), 2011 IFIP/IEEE International Symposium on*, 2011, pp. 485-492.

[118] K. Seshadrinathan, R. Soundararajan, A. C. Bovik, and L. K. Cormack, "Study of Subjective and Objective Quality Assessment of Video," *Image Processing, IEEE Transactions on,* vol. 19, pp. 1427-1441, 2010.

[119] R. Huysegems, B. De Vleeschauwer, K. De Schepper, C. Hawinkel, W. Tingyao, K. Laevens, *et al.*, "Session reconstruction for HTTP adaptive streaming: Laying the foundation for network-based QoE monitoring," in *Quality of Service (IWQoS), 2012 IEEE 20th International Workshop on*, 2012, pp. 1-9.

[120] A. Vishwanath, P. Dutta, M. Chetlu, P. Gupta, S. Kalyanaraman, and A. Ghosh, "Perspectives on quality of experience for video streaming over WiMAX," *ACM SIGMOBILE Mobile Computing and Communications Review,* vol. 13, pp. 15-25, 03/12/2010 2010.

[121] R. Kuschnig, I. Kofler, and H. Hellwagner, "Evaluation of HTTP-based request-response streams for internet video streaming," in *Proceedings of the second annual ACM conference on Multimedia systems*, 2011, pp. 245-256.

[122] S. Schwarz, C. Mehlfuhrer, and M. Rupp, "Throughput Maximizing Multiuser Scheduling with Adjustable Fairness," in *Communications (ICC), 2011 IEEE International Conference on*, 2011, pp. 1-5.

[123] S. Zhishui, Y. Changchuan, and Y. Guangxin, "Reduced-Complexity Proportional Fair Scheduling for OFDMA Systems," in *Communications, Circuits and Systems Proceedings, 2006 International Conference on*, 2006, pp. 1221-1225.

[124] IEEE, "Std 802.11u-2011 - Part 11: Wireless LAN Medium Access Control and Physical Layer Specifications: Amendment 9," 2011.

[125] 3GPP, "TS 23.234, 3GPP System to Wireless Local Area Network (WLAN) interworking; System Description," Retrieved 10-10-2014.

[126] 3GPP, "TS 23.261, IP flow mobility and seamless Wireless Local Area Network (WLAN) offload; Stage 2,".

[127] A. Awada, B. Wegmann, I. Viering, and A. Klein, "A game-theoretic approach to load balancing in cellular radio networks," in *Personal Indoor and Mobile Radio Communications (PIMRC), 2010 IEEE 21st International Symposium on*, 2010, pp. 1184-1189.

[128] L. Jiajia, Y. Kawamoto, H. Nishiyama, N. Kato, and N. Kadowaki, "Device-to-device communications achieve efficient load balancing in LTE-advanced networks," *Wireless Communications, IEEE,* vol. 21, pp. 57-65, 2014.

[129] Z. Xuejun, G. Wei, C. Guohong, and D. Yiqi, "Win-Coupon: An incentive framework for 3G traffic offloading," in *Network Protocols (ICNP), 2011 19th IEEE International Conference on*, 2011, pp. 206-215.

[130] M. Radenkovic and A. Grundy, "Framework for utility driven congestion control in delay tolerant opportunistic networks," in *Wireless Communications and Mobile Computing Conference (IWCMC), 2011 7th International*, 2011, pp. 448-454.

[131] Y. Wonyong and J. Beakcheol, "Enhanced Non-Seamless Offload for LTE and WLAN Networks," *Communications Letters, IEEE*, vol. 17, pp. 1960-1963, 2013.

[132] H. Liang, C. Coletti, H. Nguyen, Kova, x, I. Z. cs, *et al.*, "Realistic Indoor Wi-Fi and Femto Deployment Study as the Offloading Solution to LTE Macro Networks," in *Vehicular Technology Conference (VTC Fall), 2012 IEEE*, 2012, pp. 1-6.

[133] A. Pyattaev, K. Johnsson, S. Andreev, and Y. Koucheryavy, "3GPP LTE traffic offloading onto WiFi Direct," in *Wireless Communications and Networking Conference Workshops (WCNCW), 2013 IEEE*, 2013, pp. 135-140.

[134] M. Simsek, M. Bennis, M. Debbah, and A. Czylwik, "Rethinking offload: How to intelligently combine WiFi and small cells?," in *Communications (ICC), 2013 IEEE International Conference on*, 2013, pp. 5204-5208.

[135] C. Shengyang, Y. Zhenhui, and G. M. Muntean, "A traffic burstiness-based offload scheme for energy efficiency deliveries in heterogeneous wireless networks," in *Globecom Workshops (GC Wkshps), 2013 IEEE*, 2013, pp. 538-543.

[136] M. Seyedebrahimi, P. Xiao-Hong, and R. Harrison, "A quality driven framework for adaptive video streaming in mobile wireless networks," in *Wireless Communications and Networking Conference (WCNC), 2014 IEEE*, 2014, pp. 2994-2999.

[137] M. Seyedebrahimi, P. Xiao-Hong, and R. Harrison, "Adaptive Resource Allocation for QoE-Aware Mobile Communication Networks," in *Computational Science and Engineering (CSE), 2014 IEEE 17th International Conference on*, 2014, pp. 868-875.