Technische Universität München
Lehrstuhl für Medientechnik

# Towards User-centric Video Transmission in Next Generation Mobile Networks

Ali El Essaili, M.Sc.

Vollständiger Abdruck der von der Fakultät für Elektrotechnik und Informationstechnik der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktor-Ingenieurs (Dr.-Ing.)

genehmigten Dissertation.

| | | |
|---|---|---|
| Vorsitzender: | | Univ.-Prof. Dr.-Ing. Wolfgang Utschick |
| Prüfer der Dissertation: | 1. | Univ.-Prof. Dr.-Ing. Eckehard Steinbach |
| | 2. | Univ.-Prof. Dr. techn. Hermann Hellwagner<br>(Alpen-Adria-Universität Klagenfurt/Österreich) |

Die Dissertation wurde am 12.11.2013 bei der Technischen Universität München eingereicht und durch die Fakultät für Elektrotechnik und Informationstechnik am 06.02.2014 angenommen.

*To Abbass, you will always live in our hearts*

# Abstract

At a time where video communication is the dominating traffic source in mobile networks, providing a satisfactory user experience is crucial for mobile network operators. On one hand, this is challenged by the mobile users' demand for high-quality real-time video content with continuous playout. On the other hand, the access network is suffering from resource crunch with an increasing number of smartphone users.

The focus of this thesis is to design, implement and evaluate novel user-centric approaches for the distribution of the limited wireless network resources across multiple users in a mobile cell. Different from Quality of Service (QoS)-driven resource allocation schemes, the goal is to maximize the Quality of Experience (QoE) by considering the individual application and content characteristics of the mobile users.

The first part of this thesis presents a service-centric concept for uplink distribution of user-generated video content over fourth generation mobile networks. Challenging in this respect is to understand the end-user requirements who are the actual consumers of the uploaded videos. To this end, the proposed approach incorporates knowledge about the video consumers, in terms of video popularity and live versus on-demand consumption, in order to prioritise the uplink resources. Furthermore, given the active involvement of the mobile users, distributed approaches for uplink video transmission with minimal signaling overhead are desired. Hence, the uplink optimization process is decomposed into two parts: a distributed QoE optimization which is performed at the mobile terminals to decide on the video coding parameters and packet transmission, and a centralized QoE optimization in the mobile network to allocate the network resources among the video producers. Both theoretical and practical performance analysis are conducted which indicate substantial user perceived gains compared to state-of-the-art schedulers in LTE systems.

The second part of the thesis explores the benefit of QoE-based traffic and resource management in the mobile network in the context of adaptive HTTP downlink video delivery. Indeed, there is a pragmatic shift in mobile multimedia streaming from RTP/UDP-based into HTTP/TCP-based streaming services to ensure higher reliability. Dynamic Adaptive Streaming over HTTP (DASH) is poised to be the future standard for mobile multimedia delivery.

Most notably, it allows for intra-session rate adaptation to deal with the variability in mobile networks and supports both live and on-demand streaming. Whereas DASH has been mainly studied from an end-to-end perspective, this thesis is the first to propose in-network management approaches for multi-user adaptive HTTP streaming. In particular, it exploits the DASH specificities to develop low-complexity adaptation approaches which are both suitable for Over the Top (OTT) streaming services and can be deployed in a real mobile network. The approaches presented in this thesis are validated by extensive experimental results, subjective laboratory tests and simulation results in the LTE OPNET simulator. The conducted performance evaluation shows that the proposed QoE framework can provide a better video quality, faster adaptation to network variations and a fairer resource allocation in a mobile cell compared to standard adaptive HTTP streaming.

# Kurzfassung

In einer Zeit in der Videoanwendungen das Verkehrsaufkommen in Mobilfunknetzen dominieren, ist es für Mobilfunkbetreiber entscheidend, ein zufriedenstellendes Benutzererlebnis anzubieten. Einerseits wird dies durch die Nachfrage von mobilen Nutzern nach qualitativ hochwertigen Echtzeit-Videoinhalten mit kontinuierliche Playout herausgefordert. Andererseits leidet das Zugangsnetz unter Ressourcenmangel aufgrund der wachsenden Zahl von Smartphone-Nutzern.

Der Schwerpunkt dieser Arbeit liegt auf der Konzeption, Umsetzung und Evaluierung von nutzerzentrierten Ansätzen für die Aufteilung der begrenzten Übertragungsressourcen auf mehrere drahtlose mobile Nutzer in einer Zelle. In Gegensatz zur Quality of Service (QoS)-gesteuerten Ressourcenallokation, ist es das Ziel dieser Arbeit, die Nutzerzufriedenheit (Quality of Experience (QoE)) unter Berücksichtigung der Eigenschaften der einzelnen mobilen Anwender zu maximieren.

Der erste Teil dieser Arbeit stellt ein dienstorientiertes Konzept zur Uplink-Übertragung von benutzergenerierten Videoinhalten über Mobilfunknetze der vierten Generation vor. Hier ist die Herausforderung, die Anforderungen der Endnutzer zu verstehen, die die tatsächlichen Verbraucher der hochgeladenen Videos sind. Zu diesem Zweck enthält der vorgeschlagene Ansatz Wissen über die Video-Konsumenten in Bezug auf die Popularität und den Live vs. Video-on-Demand Verbrauch, um die Uplink-Ressourcen zu priorisieren. Da die mobilen Nutzer außerdem aktiv beteiligt werden, sind verteilte Ansätze für die Uplink-Videoübertragung mit minimalen Signalisierungsaufwand wünschenswert. Auf diesem Grund wird die Uplink-Optimierung in zwei Teile aufgeteilt: eine verteilte QoE-Optimierung, die auf den mobilen Endgeräten ausgeführt wird, um die Parameter für die Videocodierung und die Paketübertragungsstrategie zu bestimmen, und eine zentrale QoE-Optimierung im Mobilfunknetz, um die Netzwerk-Ressourcen zwischen den Videoproduzenten aufzuteilen. Sowohl theoretische als praktische Performanz-Analysen werden durchgeführt, die erhebliche Gewinne in der Nutzerzufriedenheit im Vergleich zum Stand-der-Technik in LTE-Systemen zeigen.

Der zweite Teil der Arbeit untersucht den Nutzen eines QoE-basierten Verkehrs- und

Ressourcenmanagements im Mobilfunknetz im Rahmen der adaptiven HTTP-basierten Downlink Videoübertragung. Aktuell findet ein Paradigmenwechsel im mobilen Multimedia Streaming von RTP/UDP-basierten zu HTTP/TCP-basierten Streamingdiensten statt, um eine höhere Zuverlässigkeit zu gewährleisten. Dynamic Adaptive Streaming over HTTP (DASH) wird voraussichtlich das Standardprotokoll für die mobile Multimedia-Übertragung. DASH ermöglicht eine intra-Session Anpassung der Datenrate und unterstützt sowohl Live- als On-Demand-Streaming. Während DASH bisher hauptsächlich aus einer Ende-zu-Ende Sicht untersucht wurde, ist diese Dissertation die Erste, die Netzmanagementansätze für adaptives HTTP Mehrbenutzer Streaming vorschlägt. Insbesondere werden die DASH Besonderheiten zur Entwicklung von Anpassungsverfahren mit geringer Komplexität genutzt, die für Over-the-Top (OTT) Streaming-Dienste geeignet sind und in einem echten Mobilfunknetz eingesetzt werden können. Die in dieser Arbeit präsentierten Ansätze werden durch umfangreiche experimentelle Ergebnisse, subjektive Labortests und Simulationsergebnisse im LTE OPNET Simulator überprüft. Die durchgeführte Leistungsbewertung zeigt, dass das vorgeschlagene QoE-Framework eine bessere Videoqualität, eine schnellere Anpassung an Veränderungen im Netz und eine gerechtere Ressourcenaufteilung in einer mobilen Zelle im Vergleich zum standard HTTP-adaptiven Streaming bietet.

# Acknowledgements

First of all, I would like to express my sincere gratitude to Prof. Dr.-Ing. Eckehard Steinbach for giving me the chance to conduct my PhD under his supervision and work on this timely and exciting topic. His guidance, commitment and expertise have been instrumental throughout my thesis work.

Furthermore, I would like to thank the second examiner, Prof. Dr. techn. Hermann Hellwagner, for coming to Munich for my thesis defense, and Prof. Dr.-Ing. Wolfgang Utschick for chairing the thesis committee.

A special thanks goes to my colleagues at DOCOMO Communications Laboratories for the fruitful collaboration and for giving me access to the LTE OPNET facilities at the lab's premises. In particular, I am thankful to Prof. Dr.-Ing. Wolfgang Kellerer and Dr. Dirk Staehle for their helpful and insightful suggestions.

Moreover, I would like to thank all my colleagues at the Institute for Media Technology (LMT) for the enjoyable and friendly working environment. I am specially thankful to Damien Schroeder for his contribution to this thesis and the excellent teamwork on the DOCOMO project. I also wish to thank Dr. Liang Zhou and Chen-Chi Wu for the joint work during their time at LMT. Furthermore, I would like to acknowledge the helpful discussions on LTE with Jan Ellenbeck. Finally, I am indebted to all my students for their support especially with the OPNET simulations.

The deepest gratitude goes to my parents, my wife Nada and my son Hassan. They have always endured me with the love and energy I need to pursue my thesis. Their unlimited support and care have made my PhD journey much more comfortable and pleasant.

Munich, March 2014 *Ali El Essaili*

# Contents

# List of Figures

# List of Tables

# List of Abbreviations

| Abbreviation | Description | Definition |
|---|---|---|
| LTE | Long Term Evolution | page 1 |
| QoE | Quality of Experience | page 2 |
| QoS | Quality of Service | page 2 |
| 3GPP | 3G Partnership Project | page 7 |
| OFDMA | Orthogonal Frequency Division Multiple Access | page 7 |
| SC-FDMA | Single Carrier Frequency Division Multiple Access | page 8 |
| WiMAX | Worldwide Interoperability for Microwave Access | page 1 |
| UE | User Equipment | page 8 |
| RRM | Radio Resource Management | page 8 |
| CQI | Channel Quality Indicator | page 8 |
| SRSs | Sounding Reference Signals | page 8 |
| eNodeB | Evolved Node B | page 8 |
| NUM | Network Utility Maximization | page 22 |
| MOS | Mean Opinion Score | page 13 |
| GBR | Guaranteed Bit Rate | page 9 |
| CLO | Cross-Layer Optimization | page 21 |
| DASH | Dynamic Adaptive Streaming over HTTP | page 2 |
| HSDPA | High-Speed Downlink Packet Access | page 23 |
| HSPA | High-Speed Packet Access | page 7 |
| PRB | Physical Resource Block | page 8 |
| OTT | Over the Top | page 2 |
| MPEG | Moving Picture Experts Group | page 18 |
| MPD | Media Presentation Description | page 3 |
| UDP | User Datagram Protocol | page 2 |
| LMSC | LAN/MAN Standard Committee | page 7 |
| OFDM | Orthogonal Frequency Division Multiplexing | page 8 |
| VQA | Video Quality Assessment | page 13 |
| UGC | User Generated Content | page 1 |
| VoIP | Voice over IP | page 14 |

| Abbreviation | Description | Definition |
|---|---|---|
| PF | Proportional Fair | page 8 |
| PSNR | Peak Signal to Noise Ratio | page 10 |
| SSIM | Structural Similarity Index Metric | page 10 |
| VQM | Video Quality Metric | page 13 |
| SVC | Scalable Video Coding | page 10 |
| AVC | Advanced Video Coding | page 10 |
| HTTP | Hypertext Transfer Protocol | page 2 |
| RTP | Real-time Transport Protocol | page 2 |
| TCP | Transmission Control Protocol | page 2 |
| UDP | User Datagram Protocol | page 2 |
| SAMVIQ | Subjective Assessment of Multimedia VIdeo Quality | page 72 |
| QP | Quantization Parameter | page 11 |
| HVS | Human Visual System | page 10 |
| FTP | File Transfer Protocol | page 16 |
| JND | Just Noticeable Difference | page 65 |
| MAC | Medium Access Control | page 14 |
| GoP | Group of Picture | page 12 |
| ECN | Explicit Congestion Notification | page 17 |
| TFRC | TCP Friendly Rate Control | page 16 |
| MSE | Mean Square Error | page 10 |
| MDC | Multiple Description Coding | page 11 |
| RTSP | Real Time Streaming Protocol | page 16 |
| RTCP | RTP Control Protocol | page 16 |
| R-D | Rate-Distortion | page 9 |
| RR | Round Robin | page 8 |
| EDF | Earliest Deadline First | page 42 |
| TTI | Transmission Time Interval | page 8 |
| QCI | QoS Class Identifier | page 9 |
| PM | Parametric Model | page 63 |

# Chapter 1

# Introduction

Next generation mobile networks will offer improved throughput and lower delays. The evolution towards the 4G standards Long Term Evolution (LTE) and the Worldwide Interoperability for Microwave Access (WiMAX) has been mainly driven by a rapid increase in resource-demanding multimedia applications. Whereas the first cellular networks only supported voice and short text messaging, mobile video is eating up more than half of the wireless bandwidth nowadays [San13]. According to Cisco's traffic forecast, mobile video traffic is expected to further grow by 90% between 2011 and 2016 [Cis12].

Despite the capacity enhancements in wireless networks, mobile communication is challenged by an increasing number of smartphone users. There is a proliferation of powerful mobile devices (phones, tablets) with improved camera capabilities and which are also supported with large screens capable of displaying high quality video content. This has indeed changed the context of use of mobile phones. Mobile users increasingly tend to use social networking websites for sharing their experiences. Upstreaming of rich multimedia content from mobile devices to video portals is more common nowadays. Camcorders, in-car cameras and mobile users on the street are generating large volumes of video data which are being uploaded over the mobile network. In addition, cloud backup services are being integrated into the mobile phones to upload photos and videos for storing into the cloud. The downlink channel has been considered the main bottleneck in the past. Nevertheless, the increasing availability of smartphones and tablets will tighten the burden on the capacity-limited uplink distribution channel.

Moreover, the consumption of multimedia content has changed tremendously over the last few years with the advent of community portals and video sharing websites. Consumers are driven towards the consumption of on-demand video and user-generated content (UGC) available through video portals (e.g., YouTube, Netflix) which could eventually replace the traditional static TV broadcast [GKLP10]. Thus, we are rapidly moving from a managed network architecture where mobile operators possess full control over the streamed content towards a distributed one, dominated by over the top (OTT) media services. In addition, new streaming protocols are being standardized to cope with the dynamic nature and the resource constraints of wireless networks. Most notable is the pragmatic shift from RTP/UDP to HTTP/TCP based streaming. The lion share of mobile traffic today is TCP/IP based, dominated by streaming of video and audio [San13]. According to [San13], YouTube alone accounts for 27% of the mobile downlink traffic in North America at peak hours. Whereas progressive download, used for example in YouTube nowadays, suffers from playout interruptions and initial buffering delays, Dynamic Adaptive Streaming over HTTP (DASH [Sto11]) is emerging as the new standard which utilizes TCP/IP and offers intra session rate adaptation capable to deal with the variability of wireless networks. Adaptive streaming is expected to account for 51% of internet video by 2015 [TDG11].

Mobile networks will have to deal with a vast increase in video traffic given the quite limited resources in the wireless spectrum [OFYjSC10]. The admission control mechanisms implemented in mobile networks are based on static Quality of Service (QoS) constraints and are not suited for the dynamic and variable characteristics of video contents. Moreover, QoS-aware schedulers [CPG$^+$13] that map the QoS constraints into QoS parameters such as minimum delay, guaranteed bit-rate or packet loss rate are content agnostic and thus do not truly reflect the user satisfaction. It has become increasingly important for mobile operators to look for user-centric approaches that complement the QoS mechanisms in the network. More specifically, the ultimate goal of a mobile network operator is to maximize the users' quality of experience (QoE). This thesis builds on previous work on QoE-driven resource allocation in mobile networks [KPS$^+$06] [KDSK07] [TKSK09], which focused on RTP-based downlink video streaming services. Specifically, this thesis extends the previous works in two main directions, which represent the core parts of this thesis. Table 1.1 summarizes the basic differences between this thesis and the previous work in [TKSK09].

The first part studies the problem of resource-efficient uplink distribution of user-generated video content over fourth generation mobile networks. QoE-based resource allocation is an approach for allocating the network resources among multiple users in a mobile cell such that the overall QoE is maximized [TKSK09]. The main concept is to exploit the variable content characteristics of the video users and their individual channel conditions to find an optimal allocation policy in terms of user satisfaction. Meanwhile, uplink resource management brings along several challenges compared to the downlink case. First, for downlink-based services, a pure network-centric approach for optimization is possible as both the network including the base station and the content source are in front of the bottleneck wireless link. However, for uplink-based services the bottleneck is between the content source, i.e., the terminal and the network. Hence, a distributed optimization approach has to be considered involving entities in

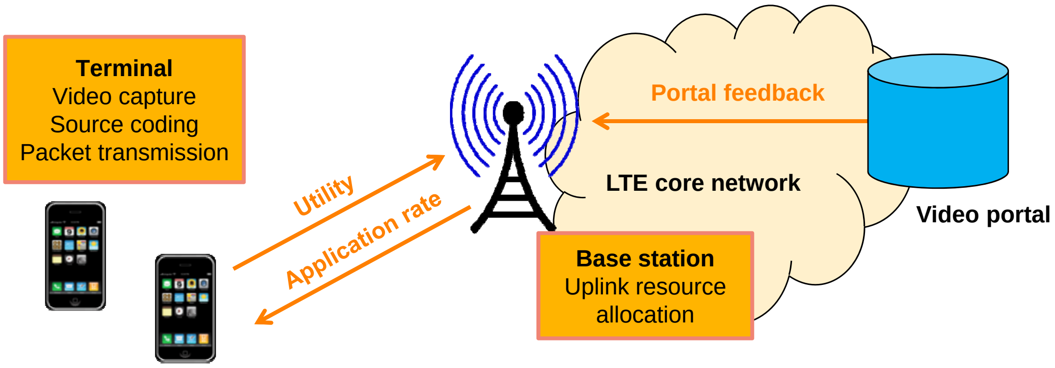


Figure 1.1: Schematic overview of QoE-based uplink resource allocation based on video portal feedback. The mobile terminals provide their utility functions to the base station. The base station determines the optimal rates of the mobile users in a cell by considering their utility and channel characteristics and the consumers' requirements. The terminals then adapt their video transmission to the available transmission capacity.

the network and in the terminal. More specifically, the base station collects channel and utility information about all the users competing for the resources and determines their uplink rates. The terminal can directly influence the source coding and packet transmission (Figure 1.1). Moreover, different from the downlink where the resource allocation maximizes the QoE of the mobile users downloading the video content, the end goal of the uplink resource allocation is to maximize the perceived quality of the users watching the upstreamed videos. Consequently, some knowledge about the video consumers is required to optimize the uplink distribution. For instance, not all videos uploaded to sharing portals have the same popularity or will be viewed at the same time. Here, the approach is to include the consumers' requirements into the uplink optimization problem, which is available through feedback from the video portal.

The second part focuses on OTT downlink adaptive HTTP streaming and proposes novel traffic and resource management approaches to maximize the user QoE. In DASH, a standard HTTP server encodes the media stream at different representations and provides the client with a list of the available representations in a media presentation description (MPD). The client uses HTTP requests for downloading the representation that matches its transmission capacity. This thesis investigates how adaptive HTTP video streaming can benefit from QoE-driven optimization in the mobile network (Figure 1.2). This is different from the downlink resource allocation considered in [TKSK09] in two aspects: First, transcoding is typically applied for video adaptation in RTP/UDP-based optimizations, which is costly in terms of computational resources and induces additional delays. Meanwhile, adaptive HTTP streaming provides inherent adaptivity by encoding the same content at multiple bitrates. Second, HTTP/TCP-based streaming is mostly delivered OTT. This means that the operator has less control on the transmitted data compared to a managed network scenario. This thesis

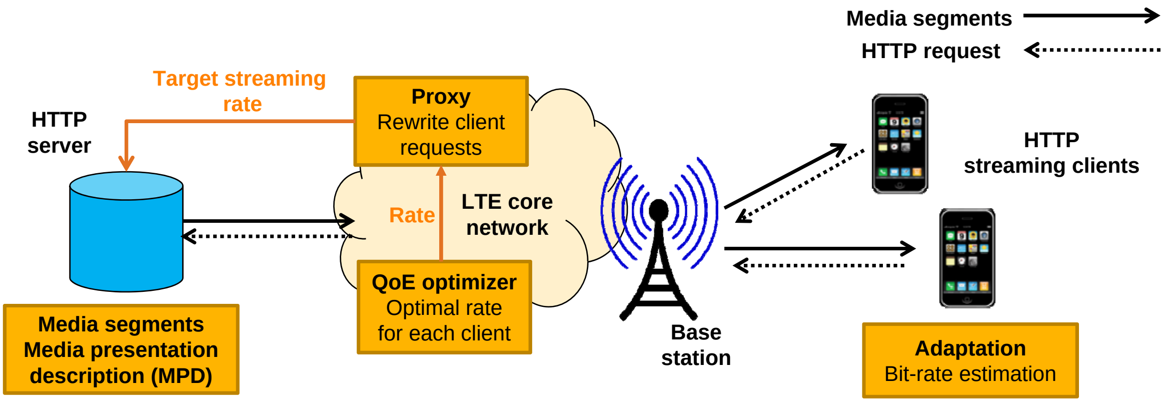


Figure 1.2: System overview for QoE-based adaptive HTTP streaming over mobile networks. A QoE optimizer in the mobile network determines the optimal transmission rate for each user. A proxy rewrites the client requests based on the feedback from the QoE optimizer and forwards the target streaming rate to the adaptive HTTP server.

is the first to address the resource allocation and streaming rate selection of the adaptive HTTP streaming users within a joint QoE optimization framework. More specifically, the optimization involves two processes: a QoE optimization in the mobile network to determine the optimal transmission rate for each user, and a proxy-based method to match the streaming rate of each user to the QoE optimization result. The proposed adaptation approach at the proxy makes use of the multiple bitrate encodings within the adaptive HTTP content and thus requires no further adaptation to the video content. This makes it particularly suitable for OTT streaming services.

## 1.1 Contributions of the thesis

This thesis envisions novel optimization approaches to improve the video delivery in next generation mobile networks. The main contributions are summarized as follows:

1) **Service-centric approach for uplink resource allocation**: This thesis proposes a service-centric approach for uplink resource allocation among multiple mobile video producers by incorporating feedback from a video portal on the consumers' preferences. Two use-cases are illustrated in Chapters 3 and 4 which consider the popularity of the uploaded content (i.e., number of followers) and the type of video delivery (live or on-demand), respectively. In both cases, the uplink resource allocation is optimized to provide video consumers with the best possible Quality of Experience (QoE) taking the limited and time-varying uplink resources into account. In particular, for uplink-based video communications, incorporating the feedback from the portal results in improved video quality and efficient usage of the network resources, compared to a QoE-driven resource allocation which is ignorant of the consumers' consumption patterns.

Table 1.1: Comparison between this thesis and the previous work on QoE-driven resource allocation in mobile networks [TKSK09]

| | **[TKSK09]** | **Chapters 3 and 4** | **Chapters 5 and 6** |
|---|---|---|---|
| Objective function | QoE-based | QoE-based with video portal feedback | QoE-based and playout buffer-aware |
| Streaming type | RTP | RTP | Adaptive HTTP |
| Mobile System | Downlink HSPA | Uplink LTE | Downlink LTE |
| Application | Video, audio, file download | Video (Live, live and on-demand) | Video (Live) |
| Optimization | Centralized (eNodeB) | Centralized (eNodeB), Distributed (Terminal) | Centralized (eNodeB), Distributed (Proxy) |
| Solution | Network (Greedy) | Network (Greedy), Distributed (Analytical, Iterative) | Network (Greedy), Distributed (Heuristic, Optimal) |

2) **Analytical model for joint live and on-demand optimization**: Chapter 4 presents a QoE-based distributed optimization approach for joint live and on-demand uplink video distribution. This is conceptually different from state-of-the-art work where optimization is performed for either live or on-demand consumption. Particularly, an optimal de-centralized approach is presented for scalable video scheduling which is independently performed at each mobile terminal. The proposed approach defines deadlines both from the consumer perspective (video request time) and the producer perspective (upload time), and both constraints are included into the distributed optimization problem. Furthermore, an analytical solution that finds a global optimum for a set of video layers with different deadline constraints is provided.

3) **Uplink optimization architecture**: To provide an optimal user experience, both the mobile terminals and the network have to be QoE-aware. The benefit for the mobile network operator is a potentially substantial saving in used network resources or the ability to provide higher user satisfaction for a given set of uplink resources. In addition, by distributing the adaptation functionalities across the mobile terminals and making use of the processing and caching capabilities of the tablets and smartphones, the signaling information is highly reduced compared to a fully centralized scheme. The conducted performance evaluation shows that the proposed uplink optimization can substantially improve the user QoE compared to standard scheduling mechanisms in LTE networks.

4) **Proactive approach for adaptation of OTT adaptive HTTP content**: The proactive approach for rewriting the client HTTP requests presented in this thesis gives control of the video content adaptation to the network operator. This provides three key improvement aspects while also allowing a standard unmodified DASH client to decode and play the redirected media segments. First, it results in improved traffic management in the

mobile network. OTT DASH provides a stalling-free playout at the expense of lower video quality. Here, the network resource allocation is optimized to provide a better user experience or support more clients at the same video quality. Second, it ensures a fairer user experience in the cell. Recent studies on adaptive HTTP streaming show unfair throughput when multiple clients are competing for shared resources due to the video player behavior [ABD11]. The objective and subjective results presented in Chapter 5 indicate a smaller QoE range among the different users. Third, clients react late to congestion and channel variations. The proposed approach exploits the network operator's information on the traffic load and radio conditions in the cell which allows for a faster adaptation to network dynamics.

5) **Joint traffic and resource management for adaptive HTTP video delivery**: A main contribution of this thesis is the joint transmission and streaming rate allocation for adaptive HTTP mobile video delivery. Specifically, both optimal and heuristic approaches for solving the corresponding rate allocation problems are investigated.

## 1.2 Organization

The rest of this thesis is structured as follows. The next chapter provides background material for this thesis. It starts by reviewing the recent advances in mobile communications, video coding and quality of experience assessment, and then surveys the state-of-the-art work on resource allocation approaches for optimizing the video delivery in wireless networks. Chapter 3 presents a service-centric approach for uplink resource allocation based on popularity feedback from a video portal. Specifically, it addresses the network resource allocation for live video transmission in the LTE uplink. Chapter 4 studies the simultaneous upstreaming of real-time and on-demand user-generated video content from the mobile terminals to a video portal. More specifically, a systematic resource allocation and transmission optimization approach is presented, which considers both the QoE-based multi-user resource allocation in the network and the distributed QoE optimization for packet scheduling at the mobile terminals. In Chapter 5, QoE-based traffic management for OTT downlink adaptive HTTP video delivery is considered. Chapter 6 additionally considers the buffered media time at the clients and studies both traffic and resource management for adaptive HTTP streaming. Finally, Chapter 7 concludes this thesis and points out some limitations and future research directions.

Parts of this dissertation have been published in [ESM$^+$11, EZS$^+$11, ESS$^+$13, KMT$^+$13, SSK$^+$14].

## Chapter 2

# Background

This chapter first reviews some basic concepts on modern mobile communication systems, multimedia content preparation and delivery, and quality of experience for multimedia applications that form prerequisite material for this thesis. Then, it covers state-of-the-art work on optimization techniques for mobile multimedia communications.

## 2.1 Modern mobile communications

With the unprecedented growth in mobile traffic, particularly driven by audio and video applications, there is an increasing demand for higher data rates and lower latencies. Meanwhile, mobile networks have also evolved to address these stringent requirements. Both the 3G Partnership Project (3GPP) and the IEEE 802 LMSC (LAN/MAN Standard Committee) have been working towards the fourth generation (4G) standards, which meet the International Mobile Telecommunications-Advanced (IMT-Advanced) requirements for future mobile systems. More specifically, 3GPP introduced Long Term Evolution (LTE) in its Release 8 [3GP10b]. LTE provides improved data rates for cell-edge users and for high mobility scenarios. In a 20 MHz spectrum, LTE supports peak rates of up to 300 Mbps for the downlink and 75 Mbps for the uplink and a radio delay of less than 5 ms [ADF$^+$09]. In 3GPP Release 10 (LTE-Advanced [3GP08a]) additional support for local area radio is provided with peak data rates of 1 Gbps. The commercial deployment of LTE networks has just started. There are already 175 LTE networks and more than 90 million LTE subscribers nowadays [Glo13].

Compared to the third generation (3G) networks (HSPA), LTE provides an enhanced radio interface with flexible system bandwidth and support for higher spectral efficiencies. Downlink access in LTE is based on orthogonal frequency division multiple access (OFDMA), which is based on multi-carrier transmission compared to single carrier transmission in WCDMA (Wideband Code Division Multiple Access). OFDMA provides orthogonal access among multiple users in a cell and combines narrowband subcarriers with a cyclic prefix and is thus insensitive to time dispersions on the radio channel [DPSB08]. The uplink access is based on single carrier frequency division multiple access (SC-FDMA) which assigns the users adjacent subcarriers (one virtual subcarrier) to compensate for the power variations in the transmitted signal which come from the multi-carrier transmission [DPSB08]. It results in a smaller peak-to-average power ratio than orthogonal frequency division multiplexing (OFDM) which is important for battery-powered mobile devices in the uplink.

The user equipment (UE) and the base station (eNodeB) are the basic elements in the LTE access network. Furthermore, LTE exhibits a distributed architecture where the eNodeB is solely in charge of radio resource management (RRM) in a mobile cell [HT09]. The RRM functionalities can be further classified into static (admission control and QoS management) which are carried out during the setup of the data transmissions and dynamic (scheduling and resource allocation) which are executed during transmission on a millisecond (ms) basis.

Channel-dependent scheduling is considered to exploit channel diversities in the time and frequency domains [ADF+09]. The Physical Resource Block (PRB) is the smallest resource allocation entity in LTE that can be allocated to a user. More specifically, each LTE radio frame has a length of 10 ms, divided into 10 sub-frames of length 1 ms (1 Transmission Time Interval (TTI)) each. In the frequency domain, a sub-frame is divided into chunks of subcarriers (typically 12 OFDM subcarriers with 15 kHz subcarrier spacing, i.e., granularity of 180 kHz). For a system bandwidth of 5 MHz or 10 MHz, this results in 25 or 50 PRBs, respectively. The scheduler decides on the allocation of the PRBs in each sub-frame by selecting the users with better channel conditions. For uplink transmissions, the UEs transmit Sounding Reference Signals (SRSs) to the eNodeB which are used to determine the channel state of each user. In the downlink, each UE computes a Channel Quality Indicator (CQI) and sends it back to the eNodeB.

Proportional Fair (PF) scheduling is most commonly used in LTE systems. It combines instantaneous channel information with the throughput history of each user to provide a fair resource allocation [JPP00]. As a scheduling metric, it divides the instantaneous throughput per PRB of user $i$ ($T_i^{PRB}(t)$) by its average throughput ($\bar{T}_i(t)$). The average throughput of each user is updated in each TTI by considering a fairness threshold $\alpha_{PF}$, where $\bar{T}_i(t+1) = (1 - \alpha_{PF}) \cdot \bar{T}_i(t) + \alpha_{PF} \cdot T_i^{TTI}(t)$, where $T_i^{TTI}(t)$ is the current transmission rate of user $i$, which corresponds to all the PRBs allocated to user $i$ in this TTI. When $\alpha_{PF} = 0$, the current throughput is not included in the average throughput calculation and the scheduler will act in a greedy way by trying to serve the users with maximum throughput. On the other hand, when $\alpha_{PF} = 1$, the average throughput is equal to the current one and the scheduler allocates the PRBs in a completely fair manner. In this case, it is equivalent to the Round Robin (RR) scheduling policy that allocates the same number of PRBs to every user.

Decisions about user admission are made by the eNodeB. More specifically, a radio bearer defines a logical connection between the eNodeB and the UE which is used for QoS provisioning in the radio access network. There are two types of bearers, a Guaranteed bit-rate (GBR) or non-guaranteed bit-rate (non-GBR) bearer. For GBR bearers, the eNodeB guarantees upon admission a bit-rate by reserving the resources for the user [Eks09]. Furthermore, a maximum bit-rate is defined for a GBR bearer. Non-GBR bearers, on the other hand, are admitted without reserving the resources. Instead, an aggregate maximum bit rate (AMBR) is specified which represents the total amount of bit-rate of a group of non-GBR bearers. This means that non-GBR bearers may experience congestion in high traffic load scenarios. Furthermore, each bearer is assigned a QoS class identifier (QCI) which is used for the admission control. A total of 9 QCIs are specified in the LTE standard [3GP12], which define a bearer type (GBR or non-GBR), QoS parameters (delay budget, packet loss rate), and priority levels depending on the user's application. A new user is admitted if its QoS requirements can be fulfilled, given the cell load and considering the QoS requirements of the other users in the cell.

## 2.2 Video coding and Quality of Experience

### 2.2.1 Rate-distortion models

Media transmission is lossy due to source coding and channel impairments which result in variable delays, throughput or packet losses that influence the user perceived quality. Specifically, a video is compressed to minimize its transmission rate by exploiting perceptual redundancies in the input data stream [WR05]. This induces some level of distortion due to quantization errors in the encoded stream which can not be fully recovered at the decoder. A tradeoff exists between the desired rate and the corresponding distortion. R-D optimization is applied at the encoder to minimize the distortion subject to the rate constraint [Ber71].

Moreover, the R-D performance is content-dependent, and thus it varies depending on the amount of spatial and temporal activities in the captured video. The R-D optimization can be performed offline (on-demand media) or online (live media). Calculating all the operating R-D points induces high complexity at the encoder. For delay-sensitive media applications, this would result in additional delays. Also, for in-network optimizations, meta information that describes the R-D characteristics of the content should be transmitted to a central controller with minimal overhead. Therefore, different abstraction models have been proposed to model the R-D performance of a video (e.g., [SFLG00] [CISN05]). Instead of calculating the distortion for all operational R-D pairs, this allows to predict the R-D characteristics from a few operating points. This is particularly relevant for the scope of this thesis where optimization problems involving multiple users need to be performed in real-time at the base station.

Distortion is commonly measured by the mean squared error (MSE) (2.1) or peak-signal-to-noise-ratio (PSNR) (2.2). The MSE is calculated by performing a pixel-by-pixel comparison between a reference image $X$ and a reconstructed image $X_r$, where M and N are the height and width of the image, respectively.

$$MSE = \frac{1}{M \cdot N} \sum_{m=1}^{M} \sum_{n=1}^{N} (X(m,n) - X_r(m,n))^2 \tag{2.1}$$

$$PSNR = 10 \log_{10} \frac{255^2}{MSE} \tag{2.2}$$

Despite its simplicity, the PSNR is criticized for not considering the characteristics of the human visual system (HVS) [Gir93]. This means that different types of distortion will be quantified equally if they have the same distortion measure. The structural similarity index metric (SSIM) uses structural distortions as a perceptual measure, which correlates with the human visual perception of distortions [WBSS04]. Specifically, it measures the similarity between two images by comparing the luminance (mean intensity), the contrast (standard deviation), and the structure after mean subtraction and variance normalization.

Both metrics require the availability of a reference image and can be used at the video encoder to generate the R-D curves. In [CISN05], a sequence-level model which requires three pairs of rate and distortion measures to model the R-D function of a video is proposed:

$$PSNR(R) = a + b \cdot \sqrt{\frac{R}{c}}(1 - \frac{c}{R}) \tag{2.3}$$

where $(a, b, c)$ are three content-dependent coefficients which can be fitted for each video sequence. Using three measurement pairs of source distortion $PSNR_i$ and source rate $R_i$, $i = \{1, 2, 3\}$, the coefficients $a, b$ and $c$ can be analytically expressed by:

$$c = R_1 \cdot \frac{(1 - v_{2,1})\mu + v_{3,1} - 1}{(1 - v_{1,2})\mu + v_{1,3} - 1},\ b = \frac{PSNR_2 - PSNR_1}{\xi_2 - \xi_1},\ a = PSNR_1 - b \cdot \xi_1 \tag{2.4}$$

where

$$v_{i,j} = \sqrt{\frac{R_i}{R_j}},\ \xi_i = \sqrt{\frac{R_i}{c}} \left(1 - \frac{c}{R_i}\right),\ \mu = \frac{PSNR_1 - PSNR_3}{PSNR_1 - PSNR_2} \tag{2.5}$$

### 2.2.2 Video coding techniques

Video streams can be classified as scalable, where the same video is encoded into multiple bitrates or quality levels, or nonscalable, where it is available at a single bitrate. The H.264/SVC (Scalable Video Coding [SMW07]) and the H.264/MPEG-4 AVC (Advanced Video Coding) [AVC12] are the corresponding state-of-the-art video coding standards, respectively. On top of this, a video stream is prepared for transmission over dynamic and resource-limited channels. To this end, different coding and rate adaptation (also referred to as traffic shaping) approaches have been considered [vC07] [CF07]. A schematic illustration of different rate adaptation techniques is given in Figure 2.1.

For nonscalable video, rate adaptation can be done at the encoder by adjusting the video coding parameters (e.g., quantization parameter (QP) [HOK99]) to meet a target encoding rate. Many encoder rate control methods have been proposed to optimize the H.264 video delivery (e.g., [CCLC11] [OHC11]). Another form of dynamic rate adaptation is transcoding. Rate transcoding is commonly used where the original stream is transcoded to a lower rate that matches the available network capacity [BC98]. This typically requires to first decode the input stream and re-encode at a target rate, which induces additional complexity and encoding delays.

Scalable video coding [SMW07], on the other hand, introduces inherent rate scalability in the encoded bitstream. It encodes the video into a base layer and several enhancement layers. A base layer can be independently decoded at the receiver and provides a coarse video quality. The enhancement layers can be utilized once the base layer is received to enhance the perceived video quality. To generate multiple video layers, either temporal, spatial or SNR scalability is applied whereby the base layer is encoded at a lower frame rate, a lower spatial resolution or at the same spatio-temporal resolution but a lower fidelity, respectively. A combination thereof can be also considered. Another alternative for rate scalable video coding is multiple description coding (MDC), where the stream is encoded at multiple representations which can be independently decoded at the receiver [Goy01]. Loss adaptive coding that combines multi-stream video with forward error correction has been studied as well and results in higher resilience against network variations (e.g., [HSX05]).

Scalable video coding provides rate adaptivity with minimal redundancy at the expense of additional coding complexity. An alternative mechanism is to provide redundant representations of the video stream at multiple bitrates. This can be done, for instance, by offering alternative encodings of the same content (e.g., low, medium and high rates) and switching among the streams depending on the available transmission rate [FHK$^{+}$06]. Furthermore, the video files can be segmented into small fragments which are encoded at different rates. To deal with drift effects due to the mismatch of reference frames in the decoder, the video streams can be aligned at a segment level, typically starting with an I-frame. This allows seamless switching between the segments depending on the available bandwidth. A most recent example is adaptive HTTP streaming which is explained in detail in Section 2.3.3.

In addition to the source coding mechanisms, receivers perform their own adaptation policies to cope with the channel variabilities. Adaptation can be applied at the decoder by adjusting the initial playout time [SJK04], the playout rate [KSG04] or by applying error

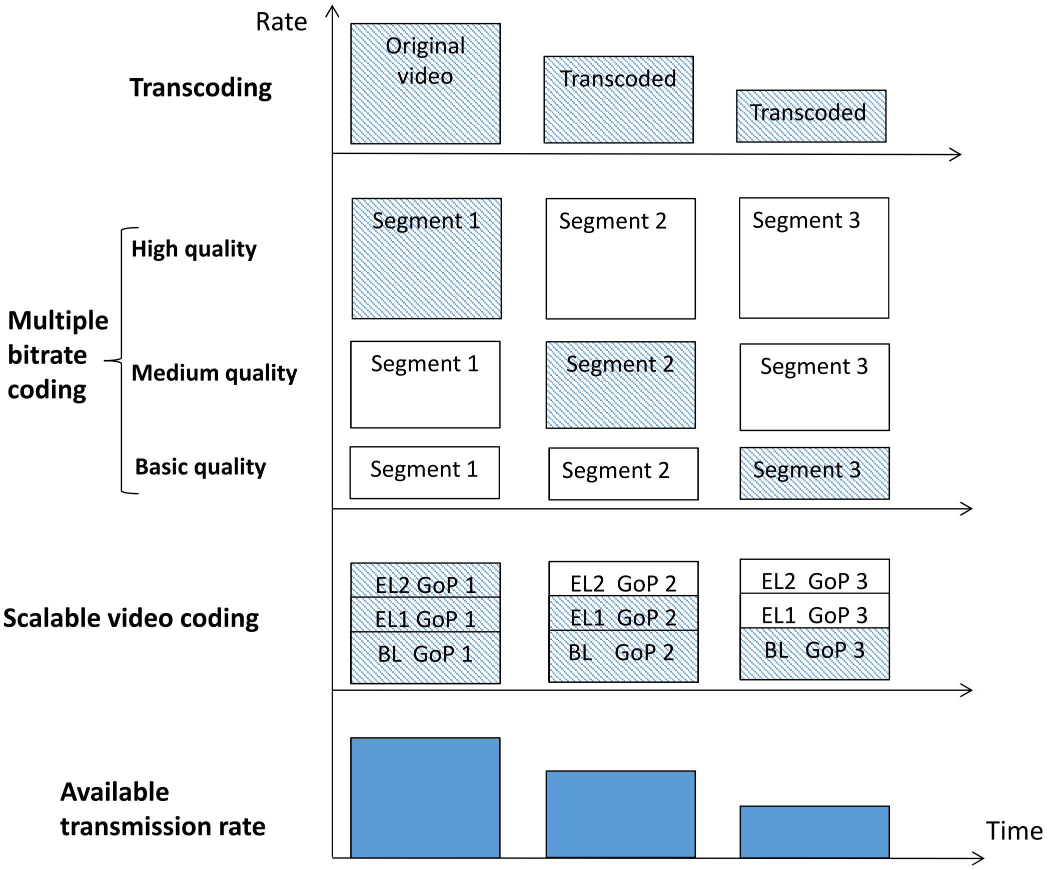


Figure 2.1: Illustration of different rate adaptation approaches: Transcoding dynamically adapts the video bitrate to the available transmission rate. Multiple bitrate coding provides different quality levels for the same content, and can switch at the segment boundaries. Scalable video coding encodes each Group of Picture (GoP) into a base layer (BL) and multiple enhancement layers (EL). The shaded regions in the figure correspond to the transmitted videos.

concealment [WZ98] in the case of lost frames.

### 2.2.3 Quality of Experience

The network Quality of Service (QoS) is measured using parameters such as delay, jitter, packet loss, and throughput [Hal00]. The application quality can be assessed by evaluating the impact of these parameters on the perceived quality. For instance, packet losses may result in lost frames that degrade the video quality. The impact level, however, depends on the relative importance of the lost frame. If the reference frame is available, the corresponding distortion can be measured. In this respect, quality is a measure of fidelity, i.e., frame-by-frame comparison, between the two videos. More generally, in the context of communication systems, the application QoS has been defined in terms of requirements such as required bit-rate or maximum end-to-end delay [Hal00]. So, the quality has been linked with the degree of satisfying these requirements. According to the ITU definition, QoS is a measure of the application performance from a system perspective [IT94].

Quality of experience, on the other hand, is a subjective measure of the user perceivability of an application [IT07]. This depends on several aspects such as the content characteristics, user expectations, viewing conditions and the context of use of the application. According to [FKR09], QoE represents the degree of delight of the user of an application. [Moe10] elaborates the dimensions of QoE by considering the individual elements that affect the user experience. These can be grouped into communication, service and contextual factors. These factors are translated into quality aspects (e.g., communication efficiency, comfort, service efficiency, and economical benefits) which provide an overall utility measure.

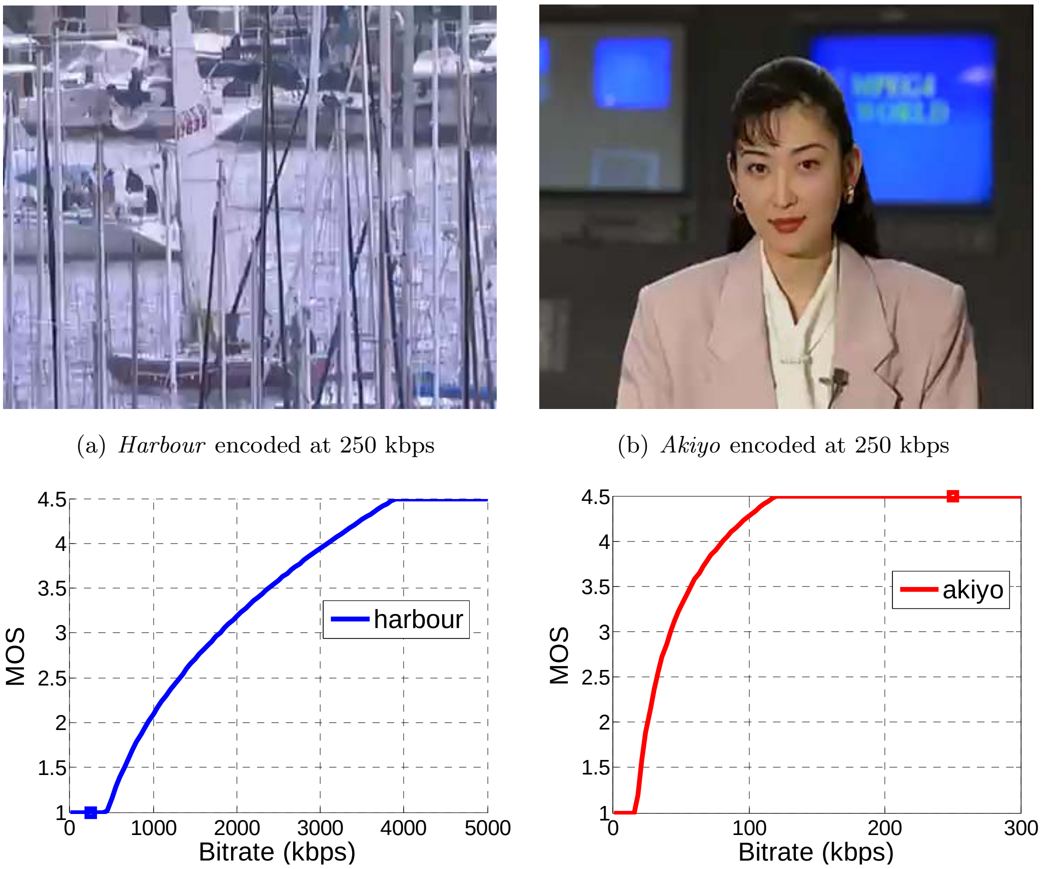
Subjective quality assessment is the most reliable measure of the user satisfaction with a certain service. The user QoE is typically expressed on a Mean Opinion Score (MOS) scale [IR98]. For instance, on a 5-scale rating which is commonly used for audio and video assessment, a subjective score of 1 stands for bad quality while a score of 5 represents excellent quality. Moreover, QoE assessment results should be reproducible. Therefore, ITU has defined several requirements for performing the subjective evaluations, which cover the test design, the selection of test material and the appropriate test methods for a given application [IR98] [IT99] [IR07]. Subjective tests are typically conducted in a laboratory where individual users participate in a subjective experiment and assess the quality of a new algorithm or system approach. Nevertheless, this approach is best suited for offline evaluations since it is time-consuming and requires the involvement of the end-user.

Instead, objective measures are typically considered for assessing the user perceived quality in real-time [WSB03]. Meanwhile, subjective test results are used for developing and validating new objective metrics. Generally, objective video quality assessment (VQA) metrics can be classified into three categories: full reference metrics require the availability of the original video which is compared on a frame-by-frame basis with the reconstructed one to determine the video quality (e.g., PSNR, SSIM [WBSS04] and the motion-based video integrity evaluation (MOVIE) [SSBC10]). This represents the most accurate quality estimation but is not suitable for real-time assessment at the end user. A survey and comparison of full reference VQA metrics can be found in [CSRK11]. No-reference metrics do not require a reference video. Instead, the video quality is deduced by processing the bit-stream (e.g., [SES$^+$13]) or measuring certain types of distortions (e.g., blockiness [WY97], blurriness [MDWE02]) in the received video stream. These metrics are most suitable for in-service monitoring. Reduced reference metrics rely on spatial or temporal features which are extracted from the original video and signaled along the video stream (e.g., Video Quality Metric (VQM) [PW04]). They require less overhead compared to the full reference metrics.

### 2.2.4 QoE-based utility functions

QoE models which consider the channel and video coding properties are typically used to represent the user experience with a multimedia service. Specifically, for resource allocation in mobile networks, utility functions that map the network performance parameters (e.g., throughput, delay, and packet loss) into a QoE level are desired. Utility functions provide a mathematical means for representing the user's satisfaction for a given set of resources [RTS10]. [She95] outlines various types of utility functions that correspond to different types of traffic. Elastic utility curves are characterized by their concavity and differentiability. That is, the marginal utility gain decreases as the rate increases. They nicely match the R-D characteristics of video streams where the quality improvements are more noticeable in lower rate regions. Discrete utility curves, on the other hand, operate at discrete levels. Examples are audio and video codecs that operate at predefined rates. These curves are more challenging to model as their properties are not known in advance (e.g., layered video coding and adaptive bitrate encoding).

QoE-based utility functions offer a natural extension of utilities by additionally taking the human perception into account. For instance, Ameigeiras et al. [AMO$^+$10] define QoE for web browsing as a function of delay and packet loss rate. In [GLGP13b] different MOS utility functions are considered for web browsing, YouTube, VoIP applications. The utility functions in [AMO$^+$10] and [GLGP13b] are further used for multi-user resource allocation in LTE networks. [BSK13] surveys state-of-the-art work on QoE modeling approaches and their application for resource management in mobile networks. Most relevant for this thesis are the application-driven utility functions defined in [KDSK07] [TKSK09]. In [KDSK07], QoE metrics for different applications (voice communication, file download, and video streaming) are defined by mapping the transmission rate and packet error probability into a common utility space. The utility function for video streaming has been extended by [TKSK09] to exploit the enhanced transmission characteristics in modern mobile networks. Due to the fast and reliable retransmissions at the medium access control (MAC) layer in wireless networks, it can be assumed that all packets are successfully transmitted. Therefore, the video utility can be described as a function of the data rate. Figure 2.2 shows utility curves for two videos with different content characteristics and their corresponding sample images. The utility curves are obtained by encoding the videos at three different QPs and then applying the model from (2.3) to reconstruct the whole R-D function. For each encoding setting, the average rate of the video stream is measured with the average PSNR, which is linearly mapped on the MOS scale. It can be observed that for the same data rate (i.e., 250 kbps) the perceived quality is different. While the *akiyo* sequence is characterized by a static background, the *harbor* sequence includes structural information (e.g., waves) which can not be well perceived at a low bitrate.



(a) *Harbour* encoded at 250 kbps

(b) *Akiyo* encoded at 250 kbps

(c) MOS of 1 at 250 kbps

(d) MOS of 4.5 at 250 kbps

Figure 2.2: Sample images from the *harbor* and *akiyo* sequences and their corresponding MOS-Rate curves [KDSK07].

## 2.3 Mobile multimedia streaming

Mobile multimedia streaming refers to the transmission of live or pre-encoded media content from a producer (server) to one or more consumers (clients) over a wireless network. Meanwhile, the clients are able to concurrently decode and play the media while retrieving it from the server. The server and the client maintain a logical end-to-end connection to control the sending rate according to the buffer level at the client and the network conditions.

Furthermore, the media content can be classified as managed if the server and the clients are inside the mobile operator network. In this case, the operator can adapt and prioritize the media transmission to provide QoS guarantees to the mobile clients. On the other hand, most dominant nowadays is over-the-top (OTT) content where the server is out of control of the mobile operator. Typical examples of OTT content are professional and user-generated content which are posted on third party servers such as YouTube.

Mobile multimedia streaming protocols provide end-to-end transmission and control mechanisms between the server and the clients. On the application layer, mobile streaming services use the Hypertext Transfer Protocol (HTTP) [FGM+99] or the Real-time Transport Protocol (RTP) [SCFJ03]. They are typically considered in combination with the Transmission Control Protocol (TCP) [Jac88] and the User Datagram Protocol (UDP) [Pos80] at the transport layer, respectively. The TCP protocol uses flow and error control to provide an error-free delivery at the expense of higher transport delay due to its congestion control and retransmission mechanisms. The UDP protocol provides a best effort service with low transport delays. Historically, HTTP/TCP has been used for web browsing and FTP (File Transfer Protocol) applications where as RTP/UDP was mainly considered for real-time streaming services. With the advances in mobile streaming protocols and the increasing need for higher transmission reliability, the attention shifted again to streaming over HTTP/TCP. This section reviews the different paradigms in mobile multimedia streaming, which are standardized by the 3GPP [GKLP10].

### 2.3.1 RTP-based streaming

RTP is a push-based streaming protocol used with UDP for real-time media delivery. On the application layer, a session control server (commonly Real Time Streaming Protocol (RTSP) [SRL98]) establishes and maintains a session with the client. Once the connection is established, the RTP/UDP protocol is used for pushing the media stream to the client. The application layer determines the transmission rate because of the absence of rate control mechanisms in UDP. That is, the stream is pushed to the lower radio layers at a rate which is equal to the source rate of the media file, referred to as streaming rate.

The actual streaming rate is adapted to the transmission characteristics between the client and the server. More specifically, the RTP/UDP server reacts to the allocated resources by the scheduler in the mobile network. Although UDP is less delay critical than TCP, it can not provide guarantees on timely delivery of the packets due to buffering delays in the access network. Moreover, RTP over UDP transmission can suffer from packet losses due to congestion if the available transmission rate is below the streaming rate. As a result, the estimated transmission rate might differ from the actually received rate at the client. This variability in the wireless network requires proper buffer control between the server and the client to avoid buffer underflow and buffer overflow events at the client. For this purpose, dynamic video bitrate adaptation is used to ensure a continuous playout [FHK+06]. Specifically, a client provides periodic reports on its observed session statistics (e.g., round trip time, packet loss rate) over the RTP Control Protocol (RTCP).

The TCP Friendly Rate Control (TFRC) [FHPW08] is commonly used with RTP to control the streaming rate in a fair manner with respect to other TCP flows that share the same resources [FHPW00]. TFRC relies on average throughput estimates and the adaptation is typically done in the order of seconds. An alternative approach is to use the explicit congestion notification (ECN) bit in the IP header [RFB01] to notify the application about imminent congestion in order to adapt its streaming rate. In a typical RTP/UDP streaming session, the server will strive to adapt the streaming rate to match the media playout rate at the client [BAB11a].

Packet losses are the main reason for video quality degradation in RTP/UDP streaming. Packet losses can correspond to packets dropped in the network or packets arriving after the playout deadline. A media client typically applies error concealment strategies to compensate for missing frames during playout which result in a degraded video quality at the receiver.

Despite the popularity RTP/UDP streaming has enjoyed, it has several limitations [Sto11]: First, it requires a specialized RTSP server instead of a standard web server, and the UDP ports are often blocked by firewalls. Also, the server keeps track of the client state until it disconnects. Finally, it leaves the flow control, congestion control, and error recovery responsibilities to the application layer. This leads to an additional complexity at the server which needs to manage and control the streaming of all the clients.

### 2.3.2 Progressive download over HTTP

The HTTP [FGM$^+$99] protocol is used for file download, progressive download and adaptive streaming applications. Progressive download was introduced in the 3GPP Release 6 [GKLP10]. Different from RTP/UDP streaming, progressive download over HTTP is pull-based. A client uses HTTP/TCP to download chunks of video before the actual playout begins. Once the client buffer reaches a predefined threshold, video playout can start. In this respect, progressive download provides a compromise between streaming and download.

Different from RTP/UDP streaming, packet losses are handled by TCP. Progressive download mainly suffers from the TCP rate fluctuations and the untimely delivery of the packets, which cause playout interruptions. The client buffer and the initial playout time are typically used to mitigate the rate variations and minimize the re-buffering occurrences, respectively.

Progressive download is widely deployed nowadays (e.g., YouTube). Nevertheless, it has several shortcomings [Sto11]: First, no live streaming is supported. Also, the client decides on the desired rate or resolution at the beginning of the streaming session, but no rate adaptivity is supported once the streaming starts. As a result, playout freezes can happen in case the bandwidth is limited. Recent studies [WKST08] show that the TCP throughput should be twice the video bit-rate to ensure a good streaming performance, which is difficult in resource-constrained wireless networks.

### 2.3.3 Adaptive HTTP streaming

Similar to progressive download, adaptive HTTP streaming is a pull-based streaming protocol that runs over TCP. In addition, it allows for dynamic rate adaptation during streaming and also provides support for live streaming. In adaptive HTTP streaming, the server encodes the same media content at multiple bitrates. Moreover, each media stream is divided into segments of equal duration (typically 2 to 10 seconds). A client can request the media segments that match its available throughput and also switch among different bitrates during the streaming session.

In adaptive HTTP streaming, the client requests the media segments using HTTP URL identifiers. Persistent connections are supported by default in HTTP/1.1 where multiple URLs are fetched in a single TCP connection. In addition, the HTTP requests can be pipelined which allows a client to send multiple HTTP requests without waiting for each response, thus allowing the TCP connection to be used more efficiently [FGM$^{+}$99].

The HTTP server treats each requested segment as a media file and sends it to the lower layers at a high rate. Different from RTP/UDP based streaming, the congestion control mechanism in TCP controls the transmission rate of the media segments. Meanwhile, a media client measures its TCP throughput and monitors its buffer state to decide on the streaming rate. The actual client control mechanisms are vendor specific. A typical adaptive HTTP client will switch to a lower streaming rate once the buffered media time drops below a certain threshold. Therefore, playout interruptions are less frequent compared to progressive download. The main cause of quality degradation is reduced video quality when switching to lower rate representations. Also, frequent and uncontrolled switches among the different representations can result in negative effects on the user perceived quality.

Proprietary solutions for adaptive HTTP streaming are widely deployed nowadays. Examples include the Microsoft Smooth Streaming, Apple Live Streaming and the Adobe Dynamic Streaming [ABD11]. For instance, BBC used the adaptive HTTP streaming technologies from Apple and Adobe for live coverage of the 2012 Olympics. To provide interoperability among servers and clients, both the 3GPP and MPEG communities have been working towards a standard adaptive HTTP streaming protocol. DASH, also referred to as MPEG-DASH [ISO12] and 3GP-DASH [3GP13], was first adopted in Release 10 by 3GPP in 2011 and standardized by MPEG in 2012, respectively. In addition, there is an ongoing work by MPEG to develop a reference software for DASH [ISO13].

The DASH protocol defines a media presentation description (MPD) for communication between an HTTP server and a streaming client (Figure 2.3). Each MPD is composed of one or more presentation periods. A period can include multiple representations of the same video content which correspond to different encoding characteristics (bitrate, resolution, codec, etc.). Each representation consists of an initialization segment which provides the client with the metadata that describes the content and one or more media segments. A client can seamlessly switch between the different segments during the streaming session by adaptively adjusting the streaming rate to its estimated transmission capacity.

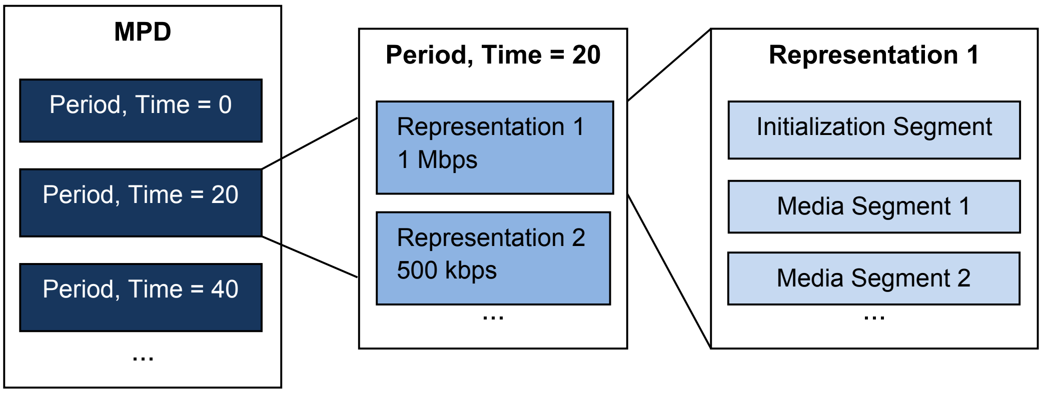


Figure 2.3: Media Presentation Description (MPD) for adaptive HTTP streaming [Sto11].

In adaptive HTTP streaming, both live and on-demand adaptive video streaming are supported. In the case of live video, a subset of the media segments (initial MPD) is generated on the fly by the server. The initial MPD can contain different representations which allow for dynamic rate adaptation at the client. Once new segments are available, they are communicated to the client through MPD updates. The MPD and segment formats are defined by the DASH specification. The delivery of the MPD, client control and media players, however, are not defined within the standard [Sod11].

DASH offers several advantages [Sto11]: First, the DASH representations can correspond to different quality levels or spatial and temporal resolutions. Thus, adaptive HTTP streaming can support a wide range of clients e.g., mobile phones, TVs, laptops. Also, different from RTP/UDP, DASH is HTTP-based which allows it to re-use the existing HTTP servers, caches and is firewall friendly. Moreover, being a client-driven protocol, this allows for the accommodation of a large number of clients without increasing the load on the server. Furthermore, DASH provides formats for media delivery and it can work on top of standard codecs (e.g., H.264, MPEG). Finally, its support for both live and on-demand media services makes it convenient for real-time media applications.

## 2.4 Optimization techniques for media transmission in mobile networks

Despite the capacity improvements in mobile networks, they continue to suffer from overload and congestion caused by an increasing number of mobile users. This is coupled with an increased mobile traffic with the emergence of resource-hungry applications such as high resolution or 3D movies, gaming and the increasing popularity of real-time streaming services. Specifically, the access network still represents the main bottleneck for video delivery. The QoS mechanisms implemented in wireless networks fall short in accommodating all the users and providing them with a satisfactory user experience. Besides the limited wireless resources, radio signals are characterized by high variability due to a non-line-of-sight radio path between the eNodeB and the UEs, multi-path propagation, channel fading and mobility effects. Although channel-aware scheduling is used in LTE to compensate for that, the adaptation cycles are on a very short time-scale which is not optimal from a media perspective. Consequently, long-term adaptation approaches that take into consideration the channel variations' influence on the perceived video quality have to be considered.

Moreover, mobile video delivery has to cope with the variability in the transported media content. Video streams are characterized by their variable bit rate (VBR) properties due to the varying spatiotemporal complexities of the video scenes. Also, the demand for rich high definition (HD) video content in the future will further increase the mobile traffic load. Modern video codecs offer high compression efficiency (H.264/AVC) and high rate scalability (H.264/SVC), but they are optimized for single-user transmission. Furthermore, in a multi-user scenario mobile users will be running applications with different delay requirements and themselves have diverse device capabilities and expectations. With the limited resources in the shared wireless spectrum and the increasing number of mobile users, there is a growing need for resource allocation approaches that can be performed in real-time and take into consideration the variable content and channel characteristics.

Generally, resource allocation for multimedia applications involves two complementary steps: On one hand, defining a resource allocation function for determining the resource shares of multiple users in a network. On the other hand, exploiting the result of the resource allocation for adapting the network resources, the media transmission, or the combination thereof. This depends on the optimization system's objective and the availability of the media stream. In this section, different optimization approaches for resource allocation are discussed based on 1) where the adaptation is carried out (network, end-to-end), 2) how (physical layer, application, or by involving multiple layers) and 3) when (short-term, long-term) the adaptation is performed. A summary is provided in Table 2.1.

### 2.4.1 Application-layer techniques

The state-of-the-art H.264 video codec provides a flexible syntax which is suitable for video transmission [WSJ$^+$03]. That is, the R-D information is exploited during the encoding process to determine the operating modes that maximize the video quality. Mobile media streaming, reviewed in Section 2.3, has evolved from simple RTP/UDP to adaptive HTTP/TCP to cope with the variability of the mobile networks [GKLP10]. Despite these major advances in multimedia delivery protocols, they generally provide an end-to-end adaptation mechanism whereas multi-access and network awareness remain an open issue [BAB11b]. The TCP or TFRC congestion control mechanisms, used with HTTP and RTP, try to provide some fair-

Table 2.1: Comparison of different adaptation mechanisms for mobile media delivery

| Mechanism | Location | Layer | Adaptation cycle |
|---|---|---|---|
| Adaptive streaming protocols (HTTP/TCP, RTP/TFRC) [GKLP10] | End-to-end (Server-client) | Application | 2-10 sec |
| SVC, MDC, packet dropping [CM06] [CF06] | Network | Application | Frame |
| Conventional CLO [vS05] [MMLB$^+$07] | Network (e.g., scheduling) or at the terminal (e.g., source coding) | Different layers of the protocol stack | order of ms |
| Packet scheduling, link adaptation, resource allocation [3GP06] | Network (eNodeB) | Radio link | 1-2 ms |
| QoE-based resource allocation [TKSK09] | Network (eNodeB) | Application | 1-2 sec |

ness among the multiple flows that share the network resources. Nevertheless, the adaptation is still based on the end-to-end flow characteristics and is network-blind, i.e., it only reacts to the network. Multi-stream video (e.g., Multiple Description Coding [Goy01] and Scalable Video Coding [SMW07] [SSW07]) encodes the same content into multiple representations and thus provides a more flexible framework for in-network adaptation (e.g., description or layer dropping). Furthermore, in-network adaptation can be performed by embedding the R-D characteristics of a video in the transmitted bitstream. The pioneer work is presented by [CM06] where R-D optimization is used for prioritizing the video frames or packets based on their delay dependencies and their loss impact on the user perceived quality. Packets with lower priority are dropped in case of congestion or rate limitation on the communication channel. R-D optimization has also been considered for multi-user resource allocation. In [TKS04] a centralized approach for R-D optimized streaming is considered where frame dropping is performed at an intermediate node to match the available rate constraint. In [CF06] a distributed approach that exploits the R-D information for packet prioritizing and scheduling across multiple streams is considered.

### 2.4.2 Cross-layer optimization

For multi-user resource allocation in mobile networks, it is important to understand the capabilities of the network and combine them with the video characteristics of the mobile users. Cross-layer optimization (CLO) approaches for resource allocation have been considered to improve the quality of service by exchanging information across the different protocol layers (e.g., [BY04], [MMLB$^+$07]). The objective, therein, is to maximize a system utility such as throughput, fairness, or video quality. Utility-based optimization for resource allocation across multiple flows has been first studied by [Kel97]. [Kel97] presented an optimization

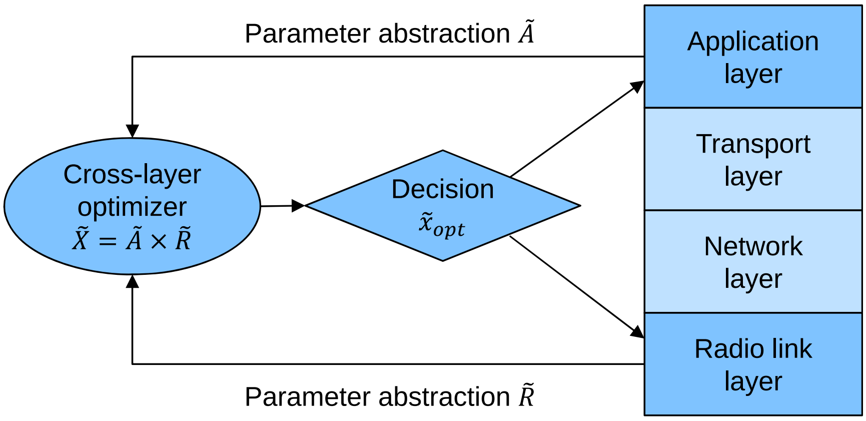

Parameter abstraction $\tilde{A}$
Application layer
Transport layer
Network layer
Radio link layer
Cross-layer optimizer $\tilde{X} = \tilde{A} \times \tilde{R}$
Decision $\tilde{x}_{opt}$
Parameter abstraction $\tilde{R}$

Figure 2.4: Application-driven cross-layer optimization concept [KPS+06].

framework for allocating the link rates across multiple TCP flows such that the overall utility is maximized. This framework, known as Network Utility Maximization (NUM), has been used to solve different optimization problems in communication systems, which are surveyed in [CLCD07].

In general, cross-layer optimization approaches can be classified into top-down approaches, where the resource management at the radio link layer is done by considering the video characteristics at the application layer or bottom-up approaches where the video source coding rate is adapted to the channel conditions [IN04]. [GNT06] uses abstracts models for the physical to transport layers and performs a CLO to maximize the QoS using different objective functions (e.g., throughput maximization, rate-utility maximization and energy consumption minimization). Conventional CLO adapts the instantaneous transmission parameters on a short timescale which is not optimal from a multimedia quality perspective [vS05]. Application-driven cross-layer optimization, on the other hand, directly maximizes an application-specific objective function while using abstracted models for the application and radio link layers [KPS+06] (Figure 2.4). A cross-layer optimizer jointly considers abstracted application (e.g., R-D profile) and radio (e.g., transmission rate) parameters. It determines the combination of parameters that maximizes an objective function which reflects the user satisfaction. The optimal decisions are then distributed to the corresponding layers where they are translated into layer-specific modes of operation (e.g., video source rate, modulation scheme).

### 2.4.3 Resource allocation in LTE networks

Resource allocation has also been widely investigated for optimizing the users' performance in next generation mobile networks. Specifically, resource allocation in LTE networks has been thoroughly studied for both uplink (e.g., [RdTBFM08]) and downlink (e.g., [SHL03])). In [LCW+10], the PRB assignment among multiple users is done with the objective of maximizing the real-time video streaming quality. These works, however, focus on short-term resource allocation solutions such as adaptive transmission rate (channel dependent scheduling), transmission power control, adaptive modulation and channel coding [3GP06]. From a multimedia perspective, utility-based resource allocation that aims at maximizing the user QoE is desired. [KL07] presents an utility-based maximization framework for both best effort and hard QoS (guaranteed service) traffic. The authors conclude that channel-dependent-only approaches which are ignorant of the traffic demands result in sub-optimal resource allocation. With respect to QoE-based resource allocation, utility optimization for radio resource management in the access network with the objective of maximizing the application QoE has been well studied (e.g., [AKK+10]). [GLGP13b] further estimates several performance indicators which are measured at the mobile terminals and provided to a central controller for QoE management in the wireless network. The feedback from the mobile users is considered for simultaneously controlling the QoE for VoIP, YouTube and web browsing applications. [FMM+13] combines QoE-based resource allocation with optimal path selection. First, it selects a video cache through either the LTE network or WLAN in order to reduce the transmission delay. Then, a traffic optimization module in the access network is considered for overall utility maximization by dropping scalable video layers to match the target data rate of each user. A detailed discussion on QoE-based management approaches can be found in [BSK13].

### 2.4.4 QoE-based resource allocation

A specific approach for QoE-based resource allocation is to proactively adapt the network resources on a long-term scale in order to meet the user requirements without interfering with the radio resource management mechanisms in the access network. For this, the application-driven CLO concepts [KPS+06] have been extended to solve resource allocation problems in mobile networks. Both centralized ([KDSK07] [TKSK09]) and distributed approaches (e.g., based on auction theory [SES+12]) for resource allocation have been studied in this context. More specifically, in [KDSK07] a multi-user multi-application cross-layer design framework is described. A MOS-maximization function that allocates the resources among users running multiple types of applications is formulated, and a sub-optimal greedy resource allocation algorithm is presented. A generic mobile scenario that reflects the user mobility is considered. [TKSK09] has further extended the work and applied it for downlink resource allocation in High Speed Downlink Packet Access (HSDPA) networks (Figure 2.5). The objective of the QoE-based optimization is to find a long-term resource allocation which maximizes the overall utility based on the application and channel conditions of the users in the cell. In-network content adaptation is then used to shape the transmitted video streams according to the QoE optimization result. In [TKS10] the authors show that transcoding can yield a better QoE compared to packet dropping at the cost of a higher computational complexity.

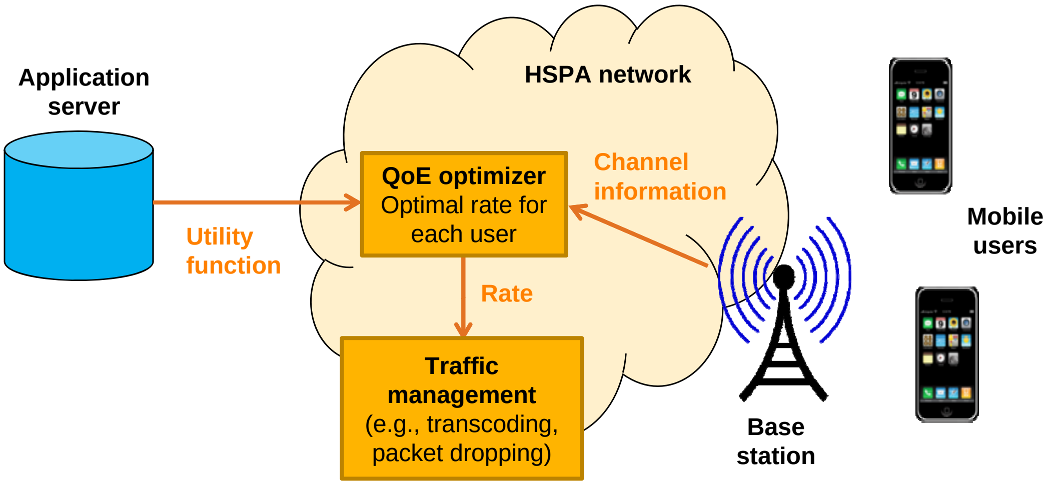


Figure 2.5: QoE optimization framework in [TKSK09] which involves a QoE optimizer that determines the target rate of each user and a traffic shaping module to adapt the application rate of the users in the mobile network.

The adaptation is carried out in the order of seconds which is less than the admission control functions, realized during the setup of data transmission, and by far larger than the scheduling cycles, which are executed on a ms basis. This can be further explained by:

1) The admission control mechanisms that are implemented in mobile networks are not sufficient for mobile multimedia applications. First, the users' streaming data rates vary over time, and thus, the QoS requirements which are considered during admission control do not truly reflect the actual cell load. Second, the QoS provided by the network is not aware of the multimedia content. In the QoE-based resource allocation, non-GBR bearers are considered. The goal is to optimize the system performance in loaded scenarios by admitting more users instead of blocking them. At the same time, by exploring the content information the quality of some users can be degraded in a controlled way while maximizing the average user satisfaction in the cell.

2) The objective of the QoE-based optimization is to control the system load without affecting the user quality of experience. That is, the different video streams are shaped such that the system does not run into overload. Meanwhile, given that the adaptation is performed every 1-2 seconds, this does not conflict with the scheduling mechanisms deployed in mobile systems. Overall, the QoE-based optimization can significantly improve the QoE in a mobile cell compared to throughput-based optimization and to standard non-optimized HSDPA systems [TKSK09].

## Chapter 3

# QoE-based resource optimization for live uplink video transmission in LTE

With the prevalence of video traffic in the uplink of next generation mobile networks, a need arises for optimizing the network resource allocation while preserving the user satisfaction. In this chapter, a service-centric approach is proposed for uplink distribution of real-time user-generated content based on the Quality of Experience (QoE) and popularity of the video content. In case of limited network resources, the proposed approach assigns more resources for popular contents while maintaining a minimum guaranteed QoE for the less popular ones. Furthermore, the service-centric optimization problem is compared with a QoE-driven one that does not consider video popularity and both approaches are evaluated for the uplink of an LTE system. The simulation results show that a significant gain in terms of average user satisfaction can be achieved.

## 3.1 Introduction

There is a proliferation of mobile phones equipped with digital cameras that allow the upstreaming of high quality multimedia content. Users capture real-time events and share them with other users, for instance, on video portals. The analysis of large-scale User-Generated Content (UGC) shows that the users' requests are highly skewed towards popular videos [CKR$^{+}$09]. In their work on YouTube traffic characterization, the authors of [GALM07] find that the video popularity is Zipf-like.

This chapter tries to answer the following question: Given that not all upstreamed videos have the same popularity, can the average user satisfaction be improved by optimizing the uplink resource allocation for the live captured videos? Despite the growing interest in UGC systems and services, there is no prior work that addresses this problem. Popularity has been traditionally exploited in cache management for proxy servers, whereby a proxy stores the initial frames of popular videos [SRT99]. In [BSW09], an analytical model for the design of coding strategies for time-shifted personalized video content is described. The authors show that a higher multicast gain can be achieved by considering the content popularity. In [RCFW10], the skewed popularity distribution in file sharing is utilized for optimal placement of the resources in structured Peer-to-Peer networks. By accelerating the search for popular contents, the average search cost for the whole system is reduced. A popularity-aware scheduling for network coding based content distribution in ad hoc networks is presented in [XDBC07]. Network-coded blocks are assigned a popularity value based on the requests from neighboring nodes. The transmission efficiency can be improved by assigning higher channel access priorities to popular blocks. Class-based resource allocation has been intensively studied in the literature (e.g., [GC02]). This is fundamentally different from this work which studies the popularity of UGC within the same video streaming service class.

The rest of this chapter is organized as follows. The next section first outlines the system model. Section 3.3 then describes the proposed service-centric resource allocation approach for the LTE uplink. In Section 3.4 the simulation results are presented and Section 3.5 concludes this chapter.

## 3.2 System model

### 3.2.1 Application model

This thesis focuses on characterizing the QoE in video streaming applications. It adopts the model from [TKSK09] where the utility function for video streaming is defined as a function of the application data rate $R$ by:

$$U = f(R), f : R \to MOS \tag{3.1}$$

where MOS represents the user satisfaction [ITU96]. The MOS is defined in the E-model over a continuous range between 1 and 4.5, which stand for unsatisfactory and very satisfactory streaming experiences, respectively [BM03].

Throughout this thesis, a simple linear mapping between the PSNR and the MOS is used [Vid00]. In addition, the considered mapping is validated with subjective experiments (in Chapter 5) which show high correlation between the objective results and the users' ratings. Moreover, other objective QoE models, which are based on the SSIM measure and more complex PSNR-based mappings, are provided for comparison in Appendix A.

In a CLO context, varying the transmission rate at the radio link layer allows the video application to adjust its encoding parameters (e.g., rate, quantization parameter) to maximize its utility function. Different from the downlink where the video application is constrained by the applied transcoding mechanism, an arbitrary set of encoding parameters can be defined in the uplink (i.e., at the video encoder). Figure 3.1 shows the utility functions for 10 different video sequences encoded with the H.264/AVC video codec [h26] at QCIF resolution and a frame rate of 30 frames/sec. To produce the utility curves, each video is encoded at three different quantization parameter settings which correspond to low, medium and high bit-rates. For each encoded video, the average rate and the average PSNR are measured. The application model from [CISN05] is then used to generate an arbitrary set of points for each sequence. The PSNR is linearly mapped to the MOS, where MOS can take on any value between 1.0 (30 dB) and 4.5 (42 dB), which represent the worst and best QoE, respectively. Each video sequence exhibits a different MOS-Rate granularity. In a multi-user scenario where each user is upstreaming a different video, the transmission rates of the different users can be determined such that the overall QoE is maximized.

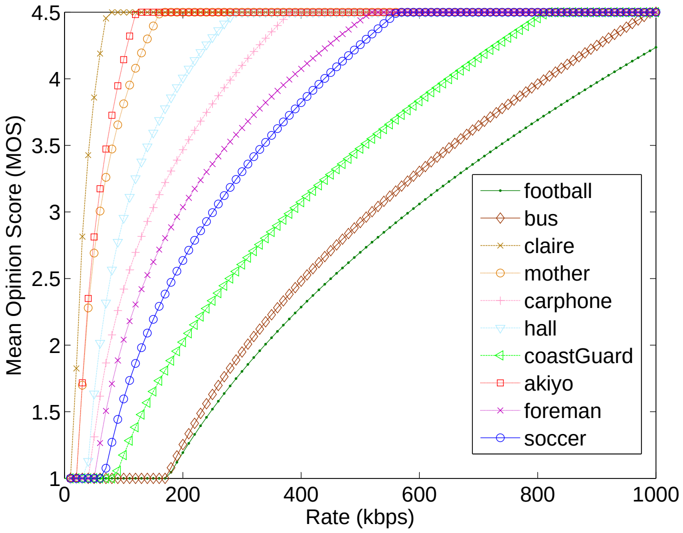


Figure 3.1: MOS as a function of data rate for different test video sequences. Videos are encoded with the H.264/AVC video codec [h26] at QCIF resolution and 30 frames/sec.

### 3.2.2 LTE model

Throughout this thesis a long-term radio link layer model with optimization periods in the order of seconds is considered. The objective is to determine the average resource share (i.e., number of PRBs) of each user in each optimization round irrespective of the individual assignment of the PRBs, which is carried out by the LTE scheduler. This allows for integrating the QoE-based optimization with any of the state-of-the-art LTE schedulers without the need to modify the scheduling mechanisms already deployed.

Specifically, an LTE simulator is developed in this thesis which follows the 3GPP LTE recommendations [3GP06]. To determine the achievable throughput per PRB for a given

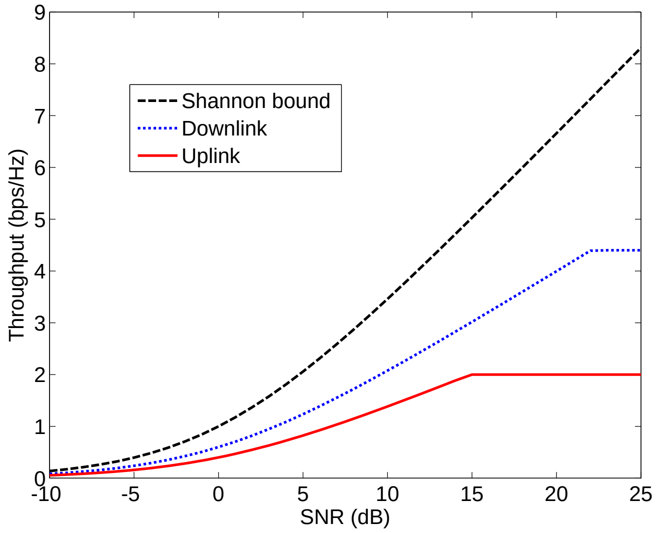


Figure 3.2: Link performance level mapping of Signal-to-Noise ratio to throughput [3GP08b]. The average throughput in the uplink and downlink can be approximated by an attenuation factor $\beta = 0.4$ and 0.6, with respect to the Shannon capacity, respectively.

Signal-to-Noise ratio ($\gamma$), the LTE link layer model from [3GP08b] is used. The model approximates the throughput $T$ for both uplink and downlink, after link adaptation and hybrid automatic repeat request (HARQ), by an attenuation factor $\beta$ compared to the Shannon capacity (3.2) (Figure 3.2).

As baseline uplink parameters, [3GP08b] defines $\beta = 0.4$, a $\gamma_{min}$ of -10 dB, a $\gamma_{max}$ of 15 dB and a maximum throughput $T_{max}$ of 2 $bps/Hz$. For the downlink, it defines $\beta = 0.6$, a $\gamma_{min}$ of -10 dB, a $\gamma_{max}$ of 23 dB and a maximum throughput $T_{max}$ of 4.4 bps/Hz, cf. (3.2).

$$T = \begin{cases} 0 & \text{for } \gamma < \gamma_{min} \\ \beta \log_2(1+\gamma) & \text{for } \gamma_{min} \leq \gamma < \gamma_{max} \\ T_{max} & \text{for } \gamma \geq \gamma_{max} \end{cases} \tag{3.2}$$

The objective of the QoE optimization is to adapt the video transmission of the mobile terminals on timescales of seconds. As a result, no instantaneous channel quality indicator (CQI) is required but rather a long-term CQI update for each user. At each optimization cycle, an average channel statistic for each user is calculated based on its observed channel conditions in the previous second. The channel realizations are further tuned to follow the typical SNR distribution of the users in an urban macrocell [3GP10a] (Figure 3.3). These distributions represent the average SNR a UE will experience in the uplink/downlink and are defined as calibration guidelines so that different LTE simulators produce comparable results. Although this model does not capture short-term channel effects, it provides an estimate of the average SNR for each user which is sufficient for the link-layer model in (3.2).

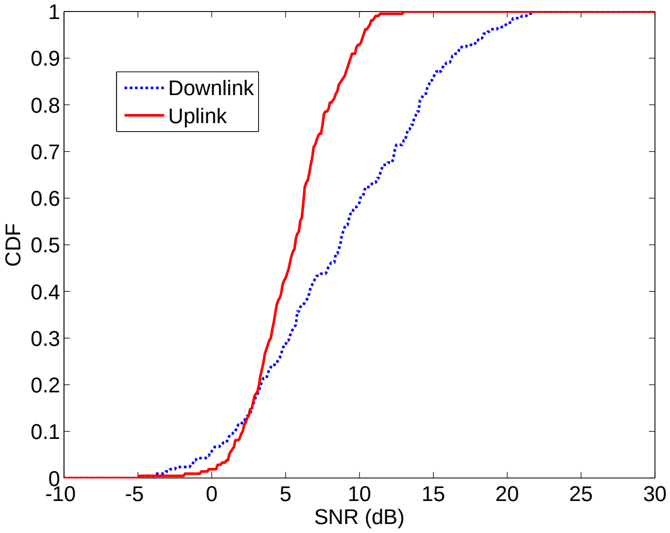


Figure 3.3: Distributions of downlink and uplink SNR in an urban macrocell [3GP10a]. The distributions are used for the calibration of simulations so that different simulators meet standard evaluation references. The figure is reproduced using the LTE simulator from [Ell12].

## 3.3 Service-centric resource allocation

### 3.3.1 Architecture overview

This chapter presents a method for uplink distribution of live video contents by addressing the popularity of the content (i.e., number of followers), the video characteristics, and the available network resources. Specifically, a service-centric concept that is based on video consumer-producer coordination through a video portal is introduced (Figure 3.4). In such a scenario, multiple users are simultaneously connecting to a video portal for sharing their captured live video content. The video portal ranks the video contents based on the consumers' requests and provides a central entity in the operator's network (e.g., eNodeB) with a standardized feedback about the popularity of the videos. This entity is then responsible for scheduling and resource allocation among multiple video producers. A big potential gain is foreseen by optimizing the uplink resources based on the popularity of video contents. Knowing that not all the live content uploaded to the portal has the same number of followers, the presented approach allows for better user satisfaction among the video consumers and provides efficient use of the wireless medium.

### 3.3.2 Objective function and solution

This chapter extends the QoE-driven resource allocation from [TKSK09] (Section 2.4.4) into a service-centric one by incorporating the feedback from a live video portal into the uplink optimization. The video portal provides live streaming for a list of video contents which are

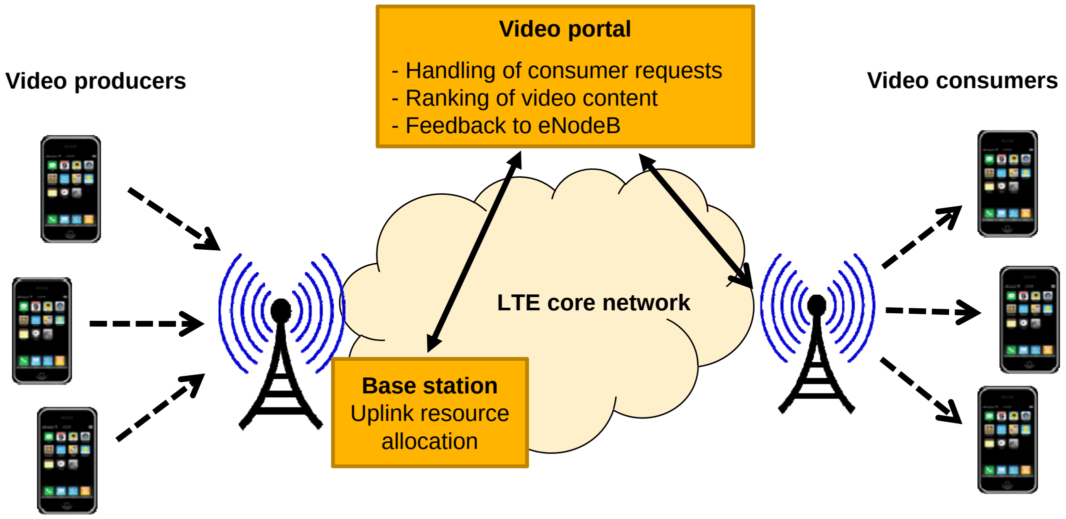


Figure 3.4: Schematic depiction of the proposed service-centric approach: A video portal collects the consumers' requests and provides the eNodeB with feedback on the popularity of the video contents; the eNodeB allocates the uplink resources among the video producers.

dynamically sorted by the number of downstream requests. The popularity of a video content follows a Zipf-Mandelbrot law [New05]. Specifically, the popularity of a content of rank $k$ out of a population of $C$ contents is defined by:

$$p_k = \frac{1/(k+q)^s}{H_{C,q,s}} \quad \text{with} \quad H_{C,q,s} = \sum_{i=1}^{C} \frac{1}{(i+q)^s} \tag{3.3}$$

where $q$ and $s$ are the shift and shape parameters of the distribution, respectively. By setting $s$ to 0, all contents have the same popularity. As $s$ increases, more requests are made for popular contents. Given a population of $N$ video upstreaming users, each uploading one content at a time, the utility-based maximization is defined by:

$$\tilde{\boldsymbol{x}}_{\boldsymbol{opt}} = \arg\max_{\tilde{\boldsymbol{x}} \in \tilde{\boldsymbol{X}}} \sum_{k=1}^{N} U_k(\tilde{\boldsymbol{x}}) \cdot p_k \quad \text{where} \sum_{k=1}^{N} p_k = 1 \tag{3.4}$$

$\tilde{\boldsymbol{x}}_{\boldsymbol{opt}}$ and $\tilde{\boldsymbol{X}}$ are the optimal and the set of possible optimization parameters abstracted from the protocol layers, respectively [IN04]. Please note that the formulation in (3.4) maximizes the sum of all users' objective functions. Alternative formulations are also possible (e.g., max-min utility [Sau08]). The content popularity, $p_k$, is a weighting factor that shapes the utility of user $k$ according to the importance of the uploaded content, where $\sum_{k=1}^{N} p_k = 1$. Hence, the convexity of the utility-based optimization problem in (3.4) is not affected.

To abstract the application parameters (i.e., utility) and the link layer parameters (i.e., data rate) the long-term cross-layer abstraction model from [SKA$^+$07] is considered. The

**Algorithm 1:** Greedy algorithm based on [JF07]

```
input  : Utility $U_k$, minimum utility improvement $\Delta u_{min}$, number of users $N$,
         maximum number of iterations $IT$, iteration step size $\Delta_\alpha$, iteration step
         size increment $\Delta_{\alpha,inc}$
output: Optimal allocation $\underline{\alpha}_{opt}$
Initialize starting allocation $\underline{\alpha}$ in a Round Robin way;
Set iteration index $t = 0$;
while $t < IT$ do
    for $k = 1$ to $N$ do
        calculate $\Delta u_{k,inc,t} = U_k(\alpha_k + \Delta_\alpha) - U_k(\alpha_k)$;
        calculate $\Delta u_{k,dec,t} = U_k(\alpha_k) - U_k(\alpha_k - \Delta_\alpha)$;
    find $\Delta u_{inc,max,t}$ for user $k^+$ which maximizes $\Delta u_{k,inc,t}$;
    find $\Delta u_{dec,min,t}$ for user $k^-$ which minimizes $\Delta u_{k,dec,t}$;
    where $k^+ \neq k^-$;
    $\Delta u_{inc,t} = \Delta u_{inc,max,t} - \Delta u_{dec,min,t}$;
    if $\Delta u_{inc,t} < \Delta u_{min}$ and $\Delta_\alpha < 1$ then
        $\Delta_\alpha = \Delta_\alpha + \Delta_{\alpha,inc}$;
    else
        $\alpha_{k^+} = \alpha_{k^+} + \Delta_\alpha$;
        $\alpha_{k^-} = \alpha_{k^-} - \Delta_\alpha$;
    $t = t + 1$;
Output allocation $\underline{\alpha}_{opt}$;
```

model defines the data rate $R_k$ for user $k$ as a function of its resource share $\alpha_k$ and its maximum achievable rate $R_{max,k}$ if all the PRBs are allocated exclusively to user $k$, cf. (3.5).

$$R_k = f_k(\alpha_k) = \alpha_k R_{max,k} \quad 0 \leq \alpha_k \leq 1, \forall k \tag{3.5}$$

The utility is defined as a function of the data rate as in (3.1). The utility function in (3.4), that maximizes the sum of utilities of $N$ users given each's content popularity, can then be described by:

$$\underset{(\alpha_1,\ldots,\alpha_N)}{\arg\max} \sum_{k=1}^{N} U_k(\alpha_k) \cdot p_k \quad \text{subject to} \sum_{k=1}^{N} \alpha_k = 1 \tag{3.6}$$

Each $\alpha_k$ value corresponds to the fraction of total PRBs assigned to user $k$. A greedy algorithm, similar to the work in [TKSK09], is used to determine the value of $\alpha_k$ (Algorithm 1). It is based on the approach originally proposed in [JF07]. The algorithm is initialized by assigning an equal amount of resources to every user. In each iteration, the algorithm iteratively takes a small amount of resources ($\Delta_\alpha$) from the user who is the least sensitive to the decrease in resources and assigns it to the user who gets the maximum benefit, until no further improvement in (3.6) is possible. The minimum resource allocation granularity in LTE is 1

PRB per TTI. For a 5 MHz system bandwidth with 25 PRBs, $\Delta_\alpha = \frac{1}{25} = 0.04$. In this case, the algorithm has little granularity to assign the PRBs for a large number of users. Given that the optimization is performed once every second, the algorithm determines the average number of PRBs for each user in a 1 second interval (i.e., 1000 TTI). Thus, $\Delta_\alpha$ can be chosen as small as $1/25000 = 0.00004$. In this work, $\Delta_\alpha = 0.004$ is used which ensures that the algorithm quickly converges to the solution. Moreover, the following parameters are defined inside the algorithm: $\Delta_{\alpha,inc} = \Delta_\alpha$, $\Delta u_{min} = 0.0005$ and $IT = 1000$.

## 3.4 Simulation results

### 3.4.1 Performance study of the service-centric optimization approach

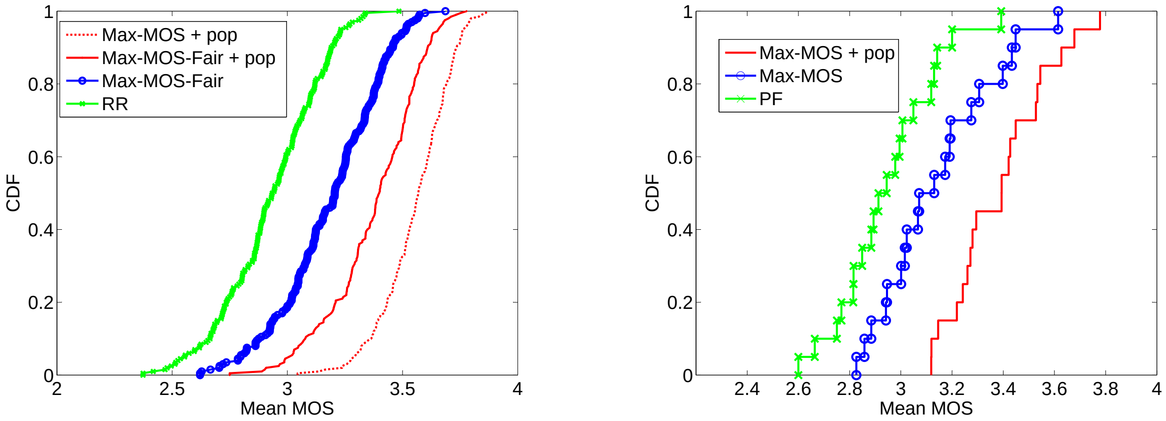
The service-centric approach is first evaluated in a Matlab-based simulated LTE environment. More specifically, a system-level LTE simulator which abstracts the application layer and radio layer characteristics of the mobile users is developed in this thesis. As input, it considers the average channel conditions over the last second to estimate the achievable throughput per PRB for each user (Section 3.2.2). On the application layer, it uses the average utility curves for different video sequences and assigns them to the mobile users (Figure 3.1). Furthermore, it combines the abstracted information from both layers and focuses on maximizing a long-term objective function (cf. (3.6)). As output, it returns a target rate for each user which can be used for encoding the video stream for the next optimization round (Section 3.2.1). An optimization cycle of 1 second is considered here.

In the following simulations, a single LTE cell scenario with multiple users upstreaming their videos is considered. Specifically, 25 test video sequences with different content characteristics are used. At each simulation run, the assignment of the sequences to upstream users is shuffled to guarantee no particular sequence enjoys higher popularity. Furthermore, the Zipf shape parameter for video popularity is initially set to 1.0 [CKR+09]. The simulation parameters are summarized in Table 3.1.

Two variants of the proposed service-centric approach are considered: 1) A scheme that maximizes the overall user satisfaction and does not provide any guarantees for the less popular contents (Max-MOS+pop). 2) A scheme that defines a minimum guaranteed QoE for all upstreamed videos (Max-MOS-Fair+pop). The Round Robin (RR) resource allocation scheme is first run as a baseline to determine the minimum QoE a user should get. The minimum QoE of each user is then added as an additional constraint to solve the optimization problem in (3.6). Generally, an explicit guaranteed QoE value for each user or group of users can be defined (e.g., [Sau08]). The service-centric approach is also compared with a QoE-driven one that does not consider video popularity (Max-MOS-Fair), and a reference RR scheme that allocates to each user an equal number of PRBs (RR).

Table 3.1: Simulation parameters for the live uplink scenario

| **Simulation parameters** | |
|---|---|
| Number of sequences | 25 |
| Number of upstream users | 5...50 |
| Number of downstream users | 1000 |
| Zipf shape parameter | 1 |
| PSNR-MOS mapping | Linear: (1,30 dB),(4.5,42 dB) |
| Application type | Video streaming |
| PSNR-Rate model | from [CISN05] |
| Simulation time | 30 sec |
| Simulation runs | 200 |
| **LTE parameters** | |
| System bandwidth | 5 MHz |
| Number of PRBs | 25 |
| Number of subcarriers | 300 |
| Bandwidth per PRB | 180 KHz |
| Link layer model [3GP08b] | see (3.2) |
| Channel model | Urban macrocell [3GP10a] |
| CQI averaging cycle | 1 sec |

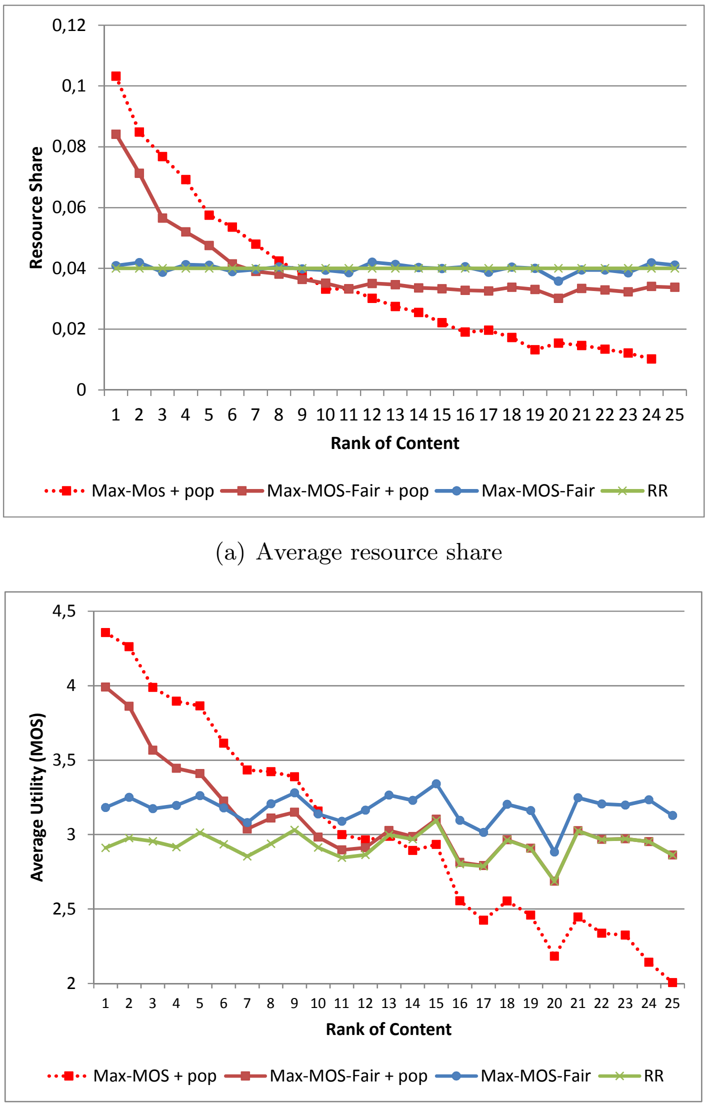


(a) LTE Matlab simulator.

(b) LTE OPNET simulator.

Figure 3.5: CDF of the mean MOS for 25 upstream users. The figures compare the QoE-driven approaches, with and without popularity, with a no-adaptation approach (default LTE mode); the round robin (RR) and proportional fair (PF) schedulers are considered for the Matlab and OPNET simulations, respectively.

The number of upstream users is initially set to 25, each uploading a different content. The number of downstream users is fixed to 1000. Figure 3.5(a) shows the cumulative distribution function (CDF) of the mean MOS for the different schemes. The mean MOS is computed by averaging the MOS for all downstream users over 200 simulation runs, 30 sec each. Please note that this is one way to measure the average user satisfaction while considering the downstream users' requests across different contents. The Max-MOS-Fair approach improves the mean MOS compared to the reference RR scheme. Both proposed service-centric approaches show an additional gain compared to the QoE-driven approach as they take the popularity of upstreamed contents into account. Meanwhile, the gain decreases in the Max-MOS-Fair+pop approach as a result of the constraint on minimum guaranteed QoE for the less popular contents.



(a) Average resource share

(b) Average utility

Figure 3.6: Individual performance for the upstreamed videos sorted in descending order of priority. The popularity-aware approaches follow the Zipf distribution whereas the Max-MOS-Fair and the RR schemes result on average in an equal resource share and utility for each content, independent of its popularity.

The above results can be further explained by inspecting the distribution of utilities and resource shares for each content. Figure 3.6(a) shows the average resource share per content

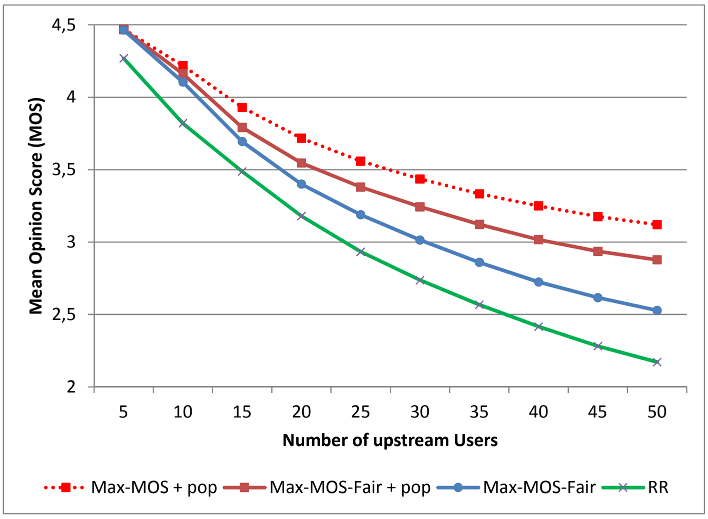
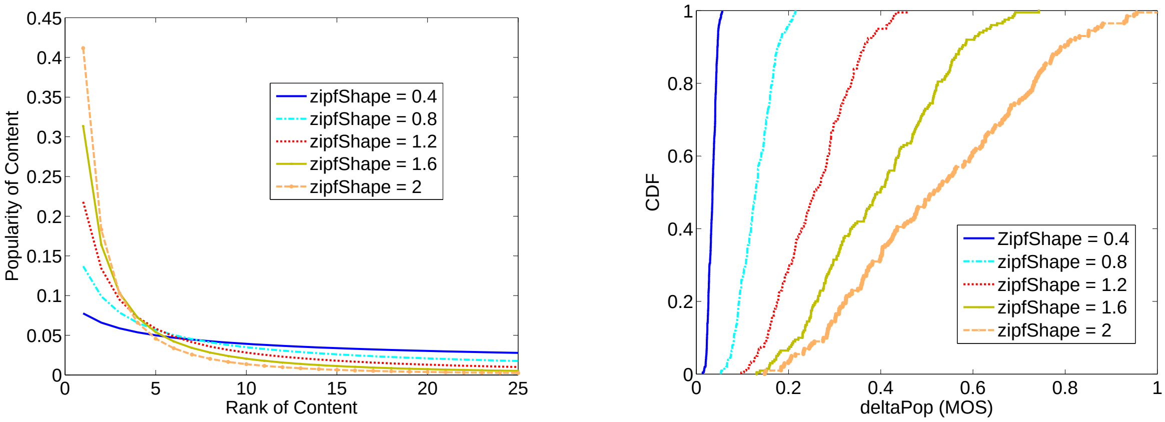


Figure 3.7: Mean MOS as a function of the number of upstream users in the cell, averaged over 200 simulation runs.



(a) Popularity of a content vs. rank of a content for different Zipf shape parameters.

(b) CDF of the MOS gain compared to a QoE-driven CLO scheme for 25 upstream users.

Figure 3.8: Comparison for different Zipf shape parameters.

as a function of the rank of the content. Contents are indexed from 1 to 25 which represent the most and least popular contents, respectively. Both service-centric approaches allocate more resources for popular contents and the distribution of resource shares much reflects the Zipf popularity distribution. Again, the Max-MOS-Fair+pop scheme shows a less skewed distribution due to the fairness constraint. The other two schemes will allocate on average equal resources for each content as they do not consider the content popularity. Figure 3.6(b) shows the distribution of average utilities of the 25 uploaded contents, sorted by their popularity. The Max-MOS-Fair approach provides an average gain compared to RR for all contents, irrespective of the content popularity. The Max-MOS+pop scheme improves the utility of popular contents dramatically. Less popular contents (i.e., contents which receive fewer downstream requests) suffer from a decrease in their individual utilities at the expense of higher overall user experience. The Max-MOS-Fair+pop scheme provides a slightly lower improvement in utility for popular contents compared to Max-MOS+pop, but it still guarantees a minimum QoE for the less popular contents (contents 13 to 25 get the same utility as in the RR scheme).

**System dimensioning gain**

Figure 3.7 shows the average MOS experienced by the downstream users as a function of the number of upstream users. When the number of upstream users is low, there are enough uplink resources and the performance of the QoE-driven and the service-centric approaches is similar. As the number of upstream users increases, the competition for resources is tighter and the service-centric approach improves the average MOS by prioritizing the popular contents. For a target MOS of 3.0, the Max-MOS-Fair+pop approach can admit 40 upstream users whereas the Max-MOS-Fair scheme can admit 30 users compared to 23 users for an RR scheme. This results in a system dimensioning gain of 25% and 42% compared to the QoE-driven and RR approaches respectively, while maintaining a minimum QoE for the less popular contents. The Max-MOS+pop scheme can even achieve a larger average gain, in particular when the network resources are more limited.

**Influence of the popularity factor**

Throughout this chapter, a Zipf shape parameter of 1.0 is considered. In Figure 3.8(a), the shape parameter of the Zipf-Mandelbrot distribution is varied between 0.4 and 2.0. As a result, the skewness of the video popularity distribution is altered. As the shape parameter increases, more downstream requests are made for popular contents. Furthermore, $deltaPop(MOS)$ is defined as the difference in MOS between the Max-MOS-Fair+pop service-centric approach and the Max-MOS-Fair scheme. Figure 3.8(b) shows the CDF of $deltaPop(MOS)$ for a 25 upstream users' case for different Zipf shape parameters. The more skewed the popularity distribution gets, the larger the popularity induced performance gain is.

### 3.4.2 LTE OPNET simulator results

Besides the Matlab-based LTE simulations, the service-centric optimization approach is further implemented and evaluated in an LTE OPNET simulator. Different from the Matlab simulations, the OPNET simulator provides detailed modules for customizing the application, transport and radio access functionalities. For each UE, different traffic patterns can be specified in an application profile and updated in real-time during the simulation (e.g., packet size, packet inter-arrival time). Also, at the eNodeB, the physical layer attributes (e.g., pathloss, channel models, power settings) and the MAC layer parameters (scheduler, random access) can be defined in configuration files. In addition, the simulations are event-driven with random seeds which allow for reproducing the same simulation (e.g., user mobility patterns) to evaluate different optimization schemes. At the end of each simulation, detailed simulation statistics are available through the simulation interface. For the QoE-based schemes, the greedy allocation approach in Algorithm 1 is implemented to determine a target transmission rate for each user. Specifically, the mean SNR is calculated in each optimization round and used as an input to the greedy algorithm. Furthermore, the result of the QoE optimization is used to adapt the uplink streaming rate at the UE without interfering with the scheduler at the eNodeB.

For the conducted performance evaluation, a single cell scenario is considered with 25 mobile users simultaneously upstreaming their video contents. In the simulation setup, admission control is disabled (i.e., all users use non-GBR bearers). The simulation parameters are the same as in Table 3.1. Furthermore, the following schemes are compared:

1) PF (Proportional Fair): The actual streaming rate is only determined by the PF scheduler. The application profile of each user defines an initial uplink target rate of 700 kbps. This represents the default LTE scheme with no adaptation.

2) Max-MOS: The application rate of each user is adjusted to the result of the QoE optimization at the beginning of each optimization round. The objective is to maximize the overall user satisfaction in the cell.

3) Max-MOS + pop: The goal of the optimization is to maximize the overall quality of experience by additionally including the popularity information. Each user adapts its streaming rate to the optimization result.

Figure 3.5(b) shows the CDF of the mean MOS for the 25 upstream users, which corresponds for 20 simulation runs. The Max-MOS scheme improves the mean MOS compared to the PF scheme as it adapts the application rate of each user to the channel and video characteristics. The Max-MOS+pop scheme provides an additional improvement in the overall user satisfaction by assigning higher priority for popular contents. Please note that in the OPNET implementation, there is no implicit guaranteed MOS for the users in the optimization problem. In the Matlab-based simulations in Section 3.4.1, however, a Max-MOS-Fair+pop scheme is additionally considered which provides a minimum guaranteed MOS for the less popular contents.

Moreover, by comparing the OPNET results to the Matlab results in Figure 3.5(a) similar gains in perceived video quality are observed. Specifically, the Max-MOS scheme improves the MOS by about 0.25 and 0.2 compared to the RR and PF schemes for the Matlab and OPNET results, respectively. Meanwhile, the Max-MOS+pop scheme achieves a gain of about 0.25 compared to the Max-MOS scheme for the OPNET results. In the Matlab results, the popularity aware optimizer manages to improve the MOS by about 0.2 and 0.37 for the Max-MOS-fair+pop and the Max-MOS+pop, respectively. Besides indicating similar gains, the above results also show that the OPNET simulation scenario is providing some inherent fairness to the less popular contents.

## 3.5 Chapter summary

This chapter presents a service-centric approach that incorporates popularity feedback from a video portal into the QoE-based uplink resource allocation. The objective is to improve the overall user satisfaction in loaded network situations by considering the asymmetry in the users' consumption of video contents. The fairness issue that could result from prioritizing the video contents is addressed by setting a minimum assured QoE for the less popular contents. The proposed approach is evaluated for the uplink of an LTE system and compared to a QoE-driven resource allocation scheme which does not consider the video popularity. A significant and consistent gain is observed by including the popularity information into the uplink resource allocation under various simulation scenarios.

The main contribution of this chapter is the service-centric resource allocation which distributes the uplink resources according to the popularity distribution of the upstreamed videos. Indeed, the approach exploits the fact that the majority of live video streams are watched by a few users and only a small number of the uploaded videos have a high popularity. The proposed approach, however, has a few limitations. First, it assumes no prior knowledge of the semantics of the uploaded videos. For instance, in the case of an emergency video, the popularity needs to be dynamically updated irrespective of the number of registered consumers. Moreover, this work assumes a reliable feedback channel between the video portal and the access network. This becomes more challenging when the consumers are served by different portals and as the popularity changes dynamically over time.

# Chapter 4

# QoE-based live and on-demand uplink video transmission in LTE

In Chapter 3, live video transmission on the uplink is studied using a service-centric QoE-based approach. This chapter considers the joint upstreaming of live and on-demand user-generated video content. This is challenged by 1) the capacity-limited and time-variant uplink spectrum, 2) the resource-hungry upstreamed videos and their dynamically changing complexity, and 3) the different playout times of the video consumers. To address these issues, a systematic approach for QoE-based scheduling and resource allocation of multi-user generated video content is proposed. First, the base station is responsible for optimizing the scarce uplink resources among multiple video producers in a cell. Second, a video portal provides feedback to the mobile terminals on the playout deadlines (live, on-demand, or both live and time-shifted) of the consumers of the content. Third, the mobile producers jointly optimize the transmission of live and time-shifted video by considering the available capacity, the deadline constraints and the video characteristics of the uploaded content. More specifically, a multi-constraint analytical model for de-centralized scalable video transmission at the mobile terminals is presented. Furthermore, a low-complexity iterative algorithm is proposed which is then used for real-time scheduling at mobile terminals in an emulated LTE network. Simulation results using the LTE OPNET simulator show that significant gains in perceived video quality can be achieved by the QoE-based resource optimization scheme when compared to the state-of-the-art proportional fair scheduler. In addition, the distributed optimization at the mobile terminals can further improve the user experience across the different types of consumers.

## 4.1 Introduction

Mobile media streaming, which has been largely dominated by static TV content, is moving into a user-defined streaming architecture where users are both producers and consumers of the media content [GKLP10]. The evolution of mobile social networks and the penetration of video portals for storing of user generated content (UGC) further contribute to this trend. As a result, mobile networks will have to deal with this vast increase in user-generated video content given the quite limited resources in the wireless uplink.

To address these issues, a service-centric approach for uplink resource allocation which jointly considers the consumer preferences and the uplink streaming conditions is proposed in Chapter 3. While Chapter 3 focuses on the live video transmission in the uplink, this chapter addresses the problem of uplink video transmission for both live and on-demand consumption. Indeed, the majority of uploaded videos to a video portal are expected to be retrieved for on-demand consumption at different time instants. The objective here is to optimize the uplink video distribution to provide the video consumers with the best possible QoE within the scheduled playout time taking the limited and time-varying uplink resources into account.

In the proposed solution, a video portal plays a central role in informing the mobile video producers about the intended consumers of the uploaded content. The portal supports different types of video delivery which include live streaming, time-shifted viewing and uploading/archiving for on-demand consumption. It collects feedback from consumers on their type of video delivery (live or on-demand) and playout times and provides this feedback to the video producers. A key novelty stems from the fact that consumers can simultaneously subscribe to different video services which pertain to the same content. For example, one user may be interested in watching a live video stream on its mobile phone and later with a refined and improved quality at home. Scalable video coding (SVC) [SMW07] provides this encoding flexibility by decomposing a video stream into multiple video layers where the base layer (BL) yields a basic video quality and the remaining enhancement layers (EL) provide a refined video quality. This chapter considers both live and time-shifted upstreaming of scalable video from a video producer to a video portal which acts as an intermediate node for both live consumption and archiving of video streams for on-demand retrieval. Video layers which can not be uploaded in real-time to the portal are cached at the mobile terminal and transmitted later for time-shifted consumption. The goal of the optimization is to provide the best possible resource allocation to the video producers while at the same time maximizing the user perceived quality across the different types of consumers.

Moreover, a pure network-centric optimization approach can be applied for downlink video communications as both the network including the base station and the content source (i.e., streaming server) are in front of the bottleneck wireless link (Section 2.4.4). For uplink-based video services, however, the bottleneck is between the content source (i.e., the terminal) and

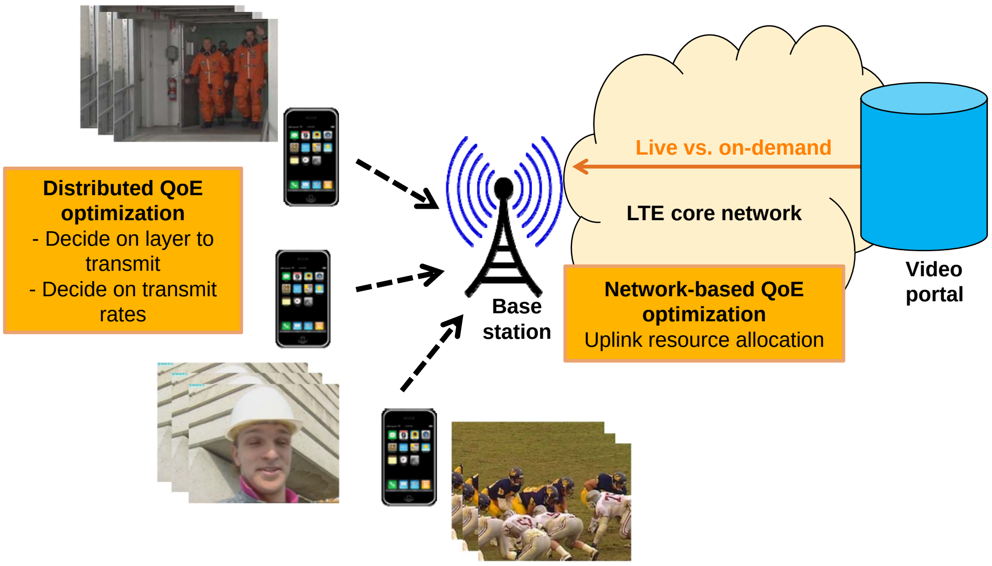

Distributed QoE optimization
- Decide on layer to transmit
- Decide on transmit rates
Base station
Live vs. on-demand
LTE core network
Network-based QoE optimization
Uplink resource allocation
Video portal

Figure 4.1: System image for QoE-driven uplink resource optimization, which includes network-based QoE optimization in the mobile network and distributed QoE optimization at the mobile terminals.

the network. More specifically, the optimization approach presented in this chapter distinguishes between the centralized multi-user resource allocation problem in the network and the distributed optimization for video layer selection at the mobile terminal. The base station has the most updated channel quality information of all the users competing for the resources, and it can determine the optimal resource distribution among the video producers. Mobile terminals can directly influence the source coding and packet transmission, and therefore a distributed optimization approach can be followed where each terminal determines its optimal scheduling decisions. The proposed approach thus represents a paradigmatic separation of responsibilities among the video producers and the network.

A schematic overview of the proposed system for uplink resource optimization is further illustrated in Figure 4.1. Multiple mobile video producers are upstreaming their content to a video portal. In the operator network, the base station is responsible for allocating the resources for uplink transmission among the video producers in one cell such that the overall QoE is maximized. Complementing it, a distributed QoE-based optimization is performed by each video producer to decide on which video layers to transmit and their respective rates.

The rest of the chapter is organized as follows. The next section surveys related work on deadline-aware scheduling for multimedia communications. In Section 4.3, the distributed QoE optimization which is performed by all terminals individually is presented. First, an optimal analytical solution is provided, and then a low-complexity iterative algorithm which is used for real-time scheduling of scalable video at the mobile terminals is proposed. Section 4.4 describes the network-based QoE optimization for uplink resource allocation among the video

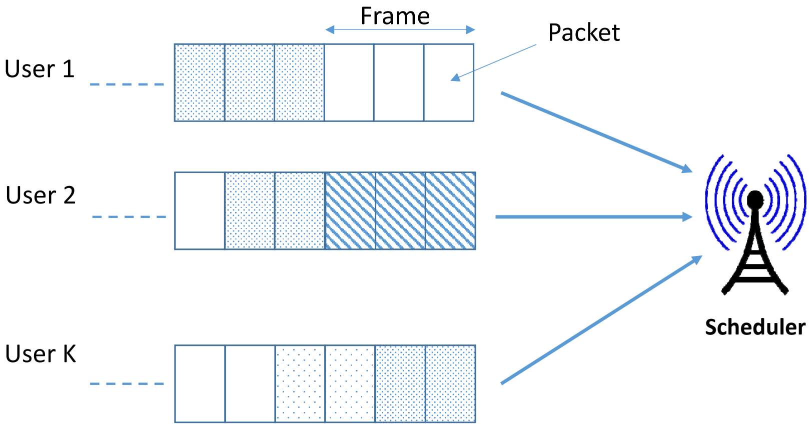


Figure 4.2: Schematic depiction of queued packets for $K$ users sharing the wireless network resources. A group of packets, which constitute a video frame, is associated with a common decoding deadline. (Reproduced for an uplink scenario from [DCBA10]).

producers. Sections 4.5 and 4.6 present the simulation results using the Matlab-based and OPNET LTE simulators, respectively.

## 4.2 Related work

The work presented in this chapter spans different directions of research on multi-user scheduling and delay-sensitive media transmission in mobile networks. An abstract system setting, widely studied in literature, is depicted in Figure 4.2. The model is formed of $K$ streaming users competing for shared network resources on a wireless channel. Each user has an output queue of packets which are associated with a decoding deadline. In the following, the key approaches that address the scheduling and packet transmission problems are surveyed:

- Earliest Deadline First (EDF) [GGP97] is one of the first studied scheduling policies whereby the objective is to schedule the packets with the most imminent deadline. Packets delivered after their deadline are useless and therefore the scheduler aims at delivering each packet before its expiry time. [SS02] shows that the EDF scheduling policy is sub-optimal in mobile environments as it fails to exploit the channel variabilities. The Round Robin, maximum throughput and PF scheduling schemes (Section 2.1) have been considered to achieve different levels of fairness/throughput efficiency in a mobile network [DBHL11].

  The above scheduling mechanisms, however, are not optimized for handling real-time multimedia communications with strict delay constraints. On one side, the scheduler in the mobile network is typically ignorant of the individual packet requirements and on the other hand, the scheduling decisions are performed on a short-term TTI basis which is not optimal from a multimedia perspective.

- The seminal work on packetized media transmission, which takes into account the content characteristics and the decoding deadlines of the individual packets, goes back to [CM06]. In their approach, the dependencies among the media packets are captured by a directed acyclic graph. This allows to prioritize the packet transmission based on the R-D characteristics and the relative importance of the packets. Specifically, the induced distortion cost from packets missing their playout deadline is introduced into the objective function and a Lagrangian rate-distortion cost function $J = D + \lambda \cdot R$ is used, where $D$ is the expected distortion and $R$ is the expected transmission rate, to decide on the packet transmission for a given rate budget. In [CM06], the channel condition is modeled with a Markov Decision Process (MDP) and assumed to be time-invariant where the transmission policy for a group of packets can be determined apriori. This is different from the time-varying characteristics of wireless channels where the transmission policies should be determined at each scheduling interval. The original work from [CM06] has been extended in several directions to account for channel, deadline and distortion impacts in multi-user streaming scenarios (e.g., [CF06] [DCBA10]).

- A cross-layer packetization and retransmission approach for video transmission in wireless LANs that explicitly considers the actual transmission requirements is proposed in [vT07]. More specifically, the video packets are grouped by their decoding deadlines and the cross-layer optimization problem is solved independently for all packets with a common decoding deadline using a general Lagrangian optimization formulation.

  While the authors do not consider the potential benefit of transmitting some packets with a later deadline before those with an earlier deadline, it is favorable in our case to schedule layers with a later deadline to guarantee a basic quality in real-time before previously cached enhancement layers with an earlier deadline are transmitted for on-demand consumption. As a result, the optimization problem cannot be solved independently as proposed in [vT07].

- Several research works have studied deadline-aware media transmission in LTE networks. The authors in [JHC$^+$09] propose a multi-user gradient-based scheduling framework that considers the streaming of scalable videos with different playback deadlines in an OFDMA system. They present a dynamic weighting metric for mitigating the approaching deadline effect such that video layers with imminent deadlines are assigned a higher priority. Again, the idea is to enforce the scheduler to assign more resources for users with approaching deadlines without considering the possibility of storing and transmitting the video layers for on-demand retrieval. A two-level scheduling approach for

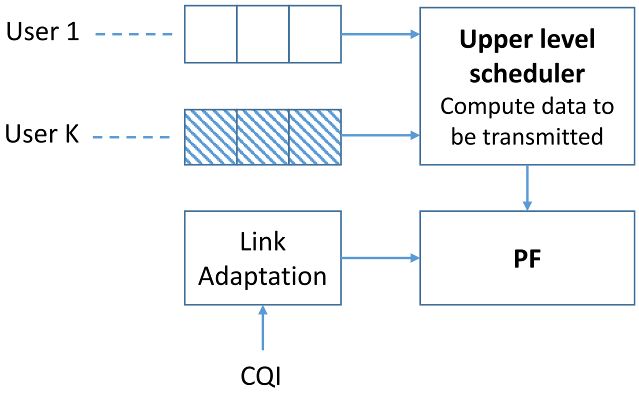


Figure 4.3: Two-level scheduling approach for real-time media transmission in LTE proposed in [PGB+11].

real-time media delivery in LTE networks is proposed in [PGB+11] (Figure 4.3). More specifically, a long-term resource allocation algorithm runs on top of the PF scheduler to decide the amount of data that each source should transmit in order to satisfy its delay constraint. The main drawback of the approach in [PGB+11] is that the upper level scheduler does not take into account the channel conditions of the users. Therefore, there is no guarantee that the PF scheduler can assign each user the amount of resources it needs to satisfy its delay requirements. Different from [PGB+11], the approach presented in this chapter leverages cross-layer information about the channel conditions and video characteristics of each user to determine its optimal rate budget.

## 4.3 Distributed QoE optimization at the mobile terminals

### 4.3.1 System model

Consider mobile terminals that are capable of generating scalable video streams which can be decomposed into several video layers (e.g., [SMW07]). In addition, the mobile producers can cache parts of the video stream and upload those parts later during good channel conditions and under lower cell load. As a result, the instantaneous user uplink transmission capacity is distributed for live and on-demand video layer transmission.

More precisely, a Group of Pictures (GoP) is encoded into a set of video layers with a common deadline (i.e., $BL$, $EL$, etc.). Please note that a GoP can be alternatively broken down into individual frames with different deadlines (e.g., [JHC+09]). At each scheduling round, a new GoP (with playout deadline $d_1$) and the layers from previous GoPs which have been cached at a mobile terminal (with playout deadline $(d_2 - j \cdot t_0)$, $1 \leq j \leq n$, $t_0$ is the scheduling period) are considered for upstreaming. $d_1$ and $d_2$ represent the playout times for live and on-demand video consumers, respectively. If the video is requested for on-demand

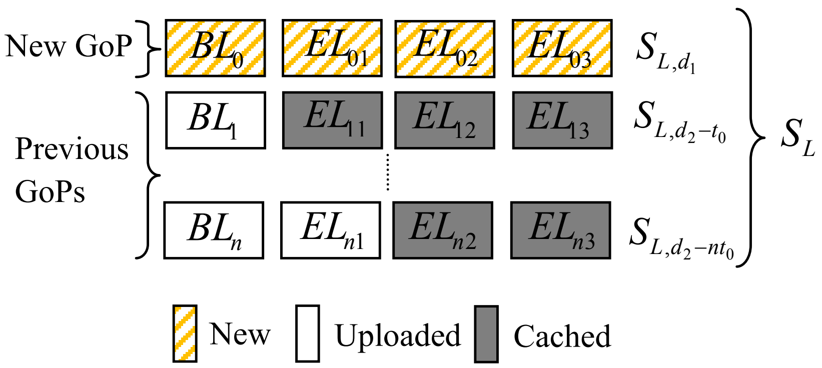


Figure 4.4: A set of video layers that correspond to a new GoP and cached layers from previous scheduling rounds at a mobile terminal. The new GoP has a playout deadline $d_1$, which represents the delay for live consumption. The first cached GoP has a playout deadline $d_2-t_0$, where $d_2$ represents the delay for on-demand video consumption and $t_0$ is the scheduling interval. The $n^{th}$ cached GoP has a playout deadline $d_2 - nt_0$.

Table 4.1: Nomenclature

| | |
|---|---|
| $K$ | total number of upstream users |
| $H_k$ | cache size of user $k$ |
| $c_k$ | uplink transmission capacity of user $k$ |
| $l, l^{'}$ | video layers |
| $l^{'} \prec l$ | $l$ is a descendant of layer $l^{'}$ |
| $L_l$ | size of layer $l$ |
| $R_l$ | instantaneous rate of layer $l$ |
| $t_l$ | transmission time of layer $l$ |
| $\Delta MOS_l$ | additional MOS improvement by transmitting layer $l$ |
| $d$ | playout deadline |
| $S_{L,d}$ | one set of video layers with a common deadline $d$ |
| $S_L$ | all sets of video layers |
| $a_l$ | binary: $= 1$ if layer $l$ is scheduled |
| $b_l$ | binary: $= 1$ if layer $l$ can be scheduled, i.e., $a_{l'} = 1, \forall\ l^{'} \prec l$ |

consumption only, then $d_1$ equals $d_2$. Moreover, $S_L$ represents the set of all candidate video layers at a mobile terminal (i.e., live and cached) and $S_{L,d}$ corresponds to one set of video layers with a common deadline $d$. These sets are illustrated in Figure 4.4 for a scalable video stream encoded into four layers.

In addition, the following notations are used: $K$ is the number of users who have data to send. $c_k$ is the instantaneous uplink transmission capacity, and $H_k$ is the cache size of user $k$. In addition, for each video layer $l \in S_{L,d}$: $L_l$ represents its size, $R_l$ is the instantaneous rate, $t_l$ is the required time to transmit the video layer, and $MOS_l$ is its utility value which is determined according to (3.1). Table 4.1 summarizes the notations used in this chapter.

### 4.3.2 Problem formulation

The distributed QoE optimization is performed independently at each mobile terminal. The goal is to maximize the QoE by determining the set of layers to be scheduled for transmission at each round, and the rate of each scheduled layer. If layer $l$ is scheduled, $a_l$ is equal to 1 and 0 otherwise. $b_l \in \{0, 1\}$, describes the layer dependencies inside a GoP, i.e., if layer $l$ is dependent on layer $l^{'}$ (i.e., $l^{'} \prec l$), then $b_l$ is equal to 1 if layer $l^{'}$ has been transmitted and 0 otherwise. That is, for layer $l$ to be scheduled, layer $l^{'}$ must also be scheduled. $\Delta MOS_l$ denotes the additional improvement of the QoE by transmitting layer $l$. The distributed optimization problem for user $k$ is formulated as:

$$\underset{a_l, R_l, \forall S_{L,d}, \forall l \in S_{L,d}}{\arg\max} \quad U_k^d = \sum_{S_{L,d}} \sum_{l \in S_{L,d}} a_l \cdot b_l \cdot \Delta MOS_l(R_l) \tag{4.1}$$

$$s.t. \quad \sum_{S_{L,d}} \sum_{l \in S_{L,d}} a_l R_l \leq c_k, \tag{4.2}$$

$$\sum_{S_{L,d}} \sum_{l \in S_{L,d}} L_l(1 - a_l) \leq H_k, \tag{4.3}$$

$$\sum_{l \in S_{L,d}} a_l t_l \leq d, \forall S_{L,d}, \tag{4.4}$$

where (4.1) is the QoE-based objective function; (4.2) corresponds to the rate constraint where the sum of rates of all transmitted layers should not exceed the available capacity; (4.3) is a necessary condition to avoid cache overflow; the size of cached layers should not exceed the cache size; (4.4) is a necessary condition to transmit layers by their deadlines; the time to transmit all layers with one common deadline should be less than the deadline.

For a given set of resources $c_k$ for user $k$, the problem in (4.1)-(4.4) can be solved locally for each user. In Appendix B, it is shown that the independent de-centralized optimization can lead to an optimal solution for each user. Moreover, an upper bound on the convergence is provided.

### 4.3.3 Low-complexity iterative algorithm

The distributed QoE optimization problem from (4.1)-(4.4) can be solved using the interior point method [BV04]. Nevertheless, the time complexity for determining the optimal solution will increase as a function of the number of cached video layers. As a result, searching through all the possible transmission policies of the newly generated and cached GoPs could become infeasible for real-time scheduling. In this section, a low-complexity iterative greedy algorithm is proposed to solve the problem in (4.1)-(4.4) at each mobile terminal. The optimization problem can be further simplified by considering that: 1) the GoPs are independent and thus the layer scheduling decisions in one GoP do not affect the subsequent GoPs. 2) The dependencies

**Algorithm 2:** Low-complexity distributed algorithm for scalable video transmission at mobile terminals

```
input  : available transmission capacity $c_k$ for user $k$
output: transmission vector $\underline{a}$ $\forall S_{L,d}, \forall l \in S_{L,d}$
Initialize transmission vector $\underline{a} = \{0\}$;
Drop layers with expired delay;
while $c_k > 0$ do
    if this is the first iteration then
        for all candidate GoPs do
            if $c_k$ is enough to send the layer $l$ in GoP $j$ that has not been sent yet
            and with the highest priority then
                calculate utility improvement $U_j = \Delta MOS_l(R_l)/(R_l)$;
            else
                $U_j = 0$;
    else
        only update $U_n = \Delta MOS_l(R_l)/(R_l)$, $l$ is the layer in GoP $n$ with the second
        highest priority;
    find $\arg\max_n U_j$ where $n$ is the GoP with the maximum utility improvement;
    set $a_l = 1$;
    $c_k = c_k - R_l$;
Check cache overflow, drop layers if necessary;
Output transmission vector $\underline{a}$;
```

between the video layers inside a GoP can be described by an acyclic graph [CM06]. In this case, the layers are represented by a sequential acyclic graph ($BL$, $EL$, etc.). An iterative greedy approach is considered for allocating the resources among the different GoPs. The objective is to determine which video layers should be scheduled/cached at each optimization round.

A description of the proposed low-complexity distributed algorithm is given in Algorithm 2. At a given scheduling round, there are $L$ video layers whose transmission policies are defined by $a_l \in \{0, 1\}, \forall S_{L,d}, \forall l \in S_{L,d}$. The transmission vector $\underline{a}$ is initialized to $\underline{0}$. At each iteration, the GoP which provides the maximum utility improvement is allocated a fraction of resources necessary to transmit a new layer $l$. Inside a GoP, the transmission priorities are determined by the acyclic graph. As a result, the iteration complexity is determined by the number of candidate GoPs. The above procedure is repeated until no more layers can be transmitted with the rate constraint. At the end of the optimization, the layer with the minimum MOS improvement is dropped in the case of cache overflow. Please note that the available capacity $c_k$ for user $k$ is determined by the network-based QoE optimization.

*Remark:* Acyclic graphs have been previously studied for the transmission of packetized media in a rate-distortion optimized way [CM06]. While the background of the proposed iterative algorithm and the work of [CM06] are similar, the approaches differ in a number of aspects. In [CM06], 1) the channel condition is assumed to be time-invariant and the transmission policy for a group of data units can be determined apriori; 2) only one deadline is considered for each data unit; 3) data units can get lost during transmission. In this approach, a time-varying wireless scenario is considered and therefore the transmission policies are determined at each optimization round. In addition, in each round a new GoP is generated for live transmission. Also, the proposed approach assumes that the video layers are transmitted successfully due to the fast retransmissions at the MAC layer in LTE.

## 4.4 Centralized QoE optimization at the base station

The objective of the network-based QoE optimization is to determine the uplink resources (i.e., $c_k$ in (4.2)) provided to each mobile terminal that maximizes the overall QoE. Let $U_k^d$ be the utility function of user $k$ as the resulting distributed QoE-based objective function in (4.1) at the mobile terminal for a given rate constraint $c_k$. The network-based objective function, that maximizes the sum of utilities of $K$ users, can then be described by:

$$\arg\max_{(\alpha_1,\ldots,\alpha_K)} \sum_{k=1}^{K} U_k^d(c_k) \quad \text{subject to} \sum_{k=1}^{K} \alpha_k = 1 \tag{4.5}$$

To determine the resource share $\alpha_k$ of each user, a similar greedy algorithm as described in Section 3.3.2 is considered. Each optimization round (1 GoP interval), the optimization problem in (4.5) is executed and the resulting $c_k$ is used for distributed layer scheduling at the mobile terminal $k$. Please note that this provides an upper limit on the system performance if network-based QoE optimization is considered. The uplink resource allocation, however, can be decoupled from the distributed QoE optimization at the mobile terminal. In this case, the actual data rate $c_k$ can be directly determined by the deployed scheduler at the base station, for instance. The optimization problem in (4.1)-(4.4) is still solved independently by each mobile terminal.

## 4.5 Simulation results

### 4.5.1 Comparison of QoE-based and earliest deadline first scheduling approaches

In the following simulation, three users upstreaming the *Football*, *Bus* and *Soccer* test video sequences with CIF resolution, 30 frames/sec and a GoP size of 8 frames are considered. Each sequence is encoded with the H.264 scalable video codec [SMW07] into four video layers using SNR scalability. For simplicity, a fixed transmission rate for each layer per GoP is used. In addition, a linear mapping between Peak Signal to Noise Ratio (PSNR) and MOS [KDSK07]

Table 4.2: Simulation parameters for the live and on-demand uplink scenario

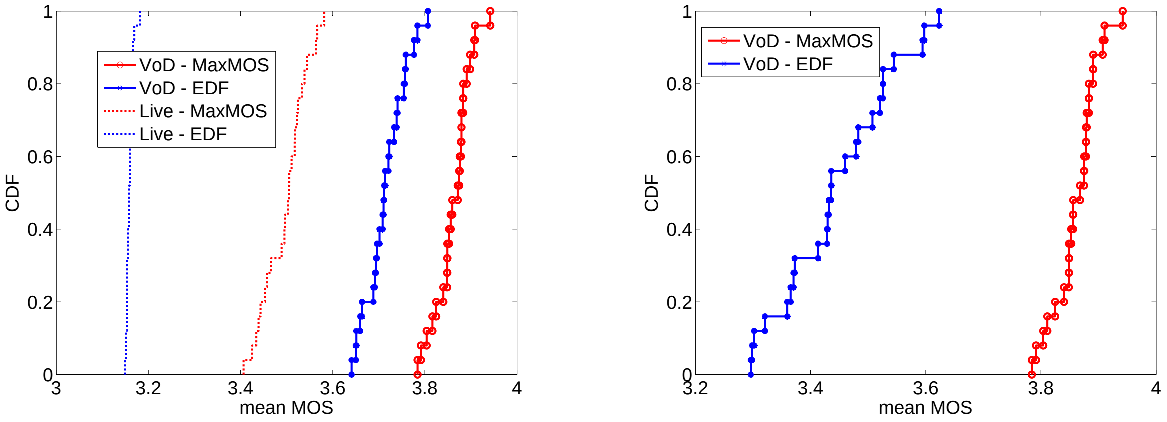
| Parameter | Value |
|---|---|
| LTE bandwidth | 5 MHz |
| Number of PRBs | 25 |
| Number of subcarriers | 300 |
| PRB size | 12 subcarriers |
| Link layer model [3GP08b] | see (3.2) |
| Channel model | Urban macrocell [3GP10a] |
| CQI averaging cycle | 266 msec |
| Number of upstream users | 3 |
| Video sequences | *Football*, *Bus* and *Soccer* |
| Relative cache size | 0.3 |
| (Live, VoD) deadlines | (266 msec, 10 sec) |
| Video codec | H.264/SVC, CIF, 30 fps, 4 layers |
| Simulation time | 10 sec |
| Simulation runs | 25 |

is considered. MOS can take on any value between 1.0 (25 dB) and 4.5 (40 dB), which represent the worst and best QoE, respectively. A simulation run lasts for 10 sec and the relative cache size is 0.3. In other words, a user can cache 30% of its video without dropping any layer. The simulation results are based on 25 simulation runs. The simulation parameters are summarized in Table 4.2.

The video transmission is optimized for two scenarios: 1) Live + time-shifted: the uploaded videos are optimized for both live and on-demand consumption. 2) Video-on-Demand (VoD): videos are intended for on-demand consumption only. In both scenarios, the VoD delay is equal to 10 sec for all users. That is, the VoD users will request the uploaded videos from a video portal after 10 sec. Meanwhile, the deadline for live video consumption in scenario 1) is 1 GoP (i.e., 266 ms). To evaluate the proposed approach, both the uplink resource allocation at the eNodeB (i.e., to find $c_k$) and the distributed optimization at the mobile terminals (i.e., (4.1)-(4.4)) are simulated. In the following results, an upper limit on the achievable QoE is provided when optimal resource allocation can be made by the eNodeB among the three users. More specifically, the two following schemes are compared:

- Max-MOS: the eNodeB is responsible for multi-user resource allocation. For an allocated bit rate, each user performs a de-centralized video layer optimization with the objective of transmitting layers which provide the best MOS improvement. The eNodeB assigns the resources in a greedy way until the sum of QoEs of all users is maximized.

- Priority-aware Earliest Deadline First (EDF): the eNodeB serves the user with the earliest deadline. Priority-aware means that each user knows about the video dependencies and orders its layers accordingly. Each user performs a de-centralized video layer optimization with the objective of transmitting the layers with the most imminent deadline first.



(a) Scenario 1 (live + on-demand video).

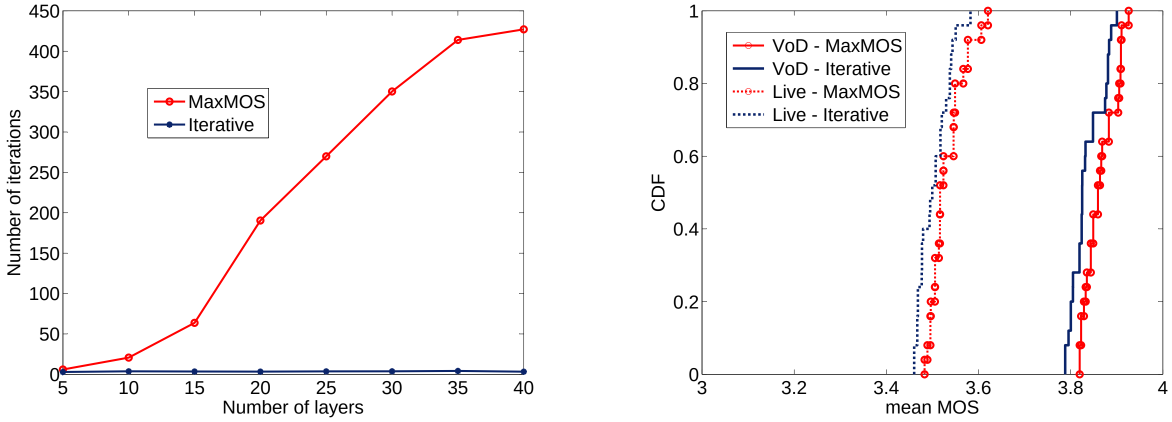
(b) Scenario 2 (on-demand video only).

Figure 4.5: CDF of the mean MOS for all users. The figure compares the performance of the MaxMOS and EDF schemes for both live and on-demand video consumption.

Figure 4.5(a) shows the cumulative distribution function (CDF) of the mean MOS of all users for the first scenario. The base layer is always sent in real-time and the enhancement layers are transmitted according to the optimization criteria for both schemes. Comparing the dotted curves (live) with the solid curves (VoD) for both schemes, a substantial gain can be achieved by uploading additional cached video layers before playout. In addition, the Max-MOS scheme improves the mean MOS compared to the EDF-based one as it selectively transmits layers which provide highest MOS improvement first. In fact, both schemes should converge to a maximum mean MOS, if the videos are requested after some sufficient time and enough cache is available for layers which are not sent in real-time.

Figure 4.5(b) shows the CDF of the mean MOS of all users for the second scenario. More specifically, it shows the average MOS if the videos are requested directly after the uplink transmission stops (VoD delay of 10 sec). Different from the first scenario, the uplink transmission of both base layers and enhancement layers is optimized for on-demand consumption for both schemes. In this case, the EDF-based scheme assigns the same transmission priority for all video layers which are transmitted according to the earliest deadline criteria. This is different from the first scenario where the base layers are assigned a higher priority to serve the live users. Therefore, not all the base layers can be duly transmitted and the EDF-based scheme fails to provide a sustainable video quality for the whole session. By comparing with the Max-MOS scheme, a significant gain can be achieved if the video layers are transmitted according to the MOS criteria.

Appendix A.1 further compares the performance of both schemes using an SSIM-based QoE-model, which shows similar QoE gains for the Max-MOS approach compared to the EDF-based scheme for both live and on-demand consumers.



(a) Mean number of iterations per optimization round as a function of the number of video layers at the mobile terminal. The number of iterations to achieve the solution can be drastically reduced as the number of cached video layers increases.

(b) Mean MOS of all users. The performance of the iterative solution is close to the optimal one for both live and on-demand consumption.

Figure 4.6: Comparison of the optimal solution from (4.1)-(4.4) and the approximate solution from the proposed iterative scheme.

### 4.5.2 Complexity analysis

The performance of the iterative greedy algorithm is compared with the optimal solution using the interior point method [BV04]. It is clear that the computational complexity can be significantly reduced with the iterative approach (Figure 4.6(a)). In fact, it is enough to look for the GoP with the highest utility gain and then use the acyclic graph to pick the corresponding layer inside the GoP to be scheduled. As a result, the number of iterations per optimization round is proportional to the number of scheduled layers for each user. Figure 4.6(b) further depicts the cumulative distribution function (CDF) of the mean MOS of all users. The MOS for the proposed scheme is close to the optimal solution for both live and VoD transmission. The slight MOS reduction depends on the choice of enhancements layers that are uploaded for on-demand consumption. Table 4.3 summarizes the average performance statistics using both schemes.

Table 4.3: Average statistics per optimization round

| | Optimal | Iterative |
|---|---|---|
| Number of iterations | 295 | 4 |
| Execution time (msec) | 69 | 1 |
| Mean MOS (Live) | 3.531 | 3.503 |
| Mean MOS (VoD) | 3.865 | 3.835 |

Table 4.4: LTE OPNET simulation setup for the live and on-demand uplink scenario

| **Application parameters** | |
|---|---|
| Video codec | H.264 SVC [SMW07] |
| SVC | SNR scalability, 4 layers at CIF resolution and 30 fps |
| GoP size | 8 frames |
| Video sequences | *Football*, *Bus*, *Soccer* and *Foreman* |
| Relative cache size | 0.3 |
| Live deadline | 1 GoP (266 msec) |
| VoD deadline | 20 sec |
| **LTE parameters** | |
| Transmission scheme | SC-FDMA |
| Base frequency | 1920 MHz |
| Bandwidth | 5 MHz |
| Bandwidth per PRB | 180 KHz |
| Subcarrier spacing | 15 KHz |
| eNodeB antenna gain | 15 dBi |
| UE speed | 30 km/h |
| UE antenna gain | 0 dBi |
| Max Tx power per user | 24 dBm |
| UE receiver sensitivity | -160 dBm |
| Scheduler | Proportional fair |
| Channel model | Urban macrocell |
| Shadowing | disabled |
| CQI averaging cycle | 266 msec |
| **Simulation parameters** | |
| Number of users | 4 |
| Simulation runs | 50 |
| Simulation time | 100 sec |
| Simulator | LTE OPNET 16.0 |

## 4.6 LTE OPNET simulator results

The benefits of the proposed QoE optimization framework are further evaluated in the LTE OPNET simulator. The purpose of the study is twofold: 1) to compare the network-based QoE resource allocation approach to the content-agnostic one by a standard LTE PF scheduler, and 2) to analyze the potential gains of applying the distributed QoE optimization at a mobile terminal for joint scheduling of live and on-demand video streams.

A single simulated LTE cell is considered with 4 users upstreaming the *Bus*, *Soccer*, *Football* and *Foreman* test video sequences. The videos are encoded with the H.264 scalable video codec [SMW07] into 4 layers using SNR scalability and have a CIF resolution and a frame rate of 30 frames per second. Furthermore, a relative cache size of 30% is considered at the mobile terminals. Moreover, the live and on-demand consumption deadlines are set to 1 GoP (i.e., 266 ms) and 20 sec, respectively. Please note that for each new GoP, the average rate and distortion characteristics of the different layers within this GoP are considered. The simulation parameters are summarized in Table 4.4.

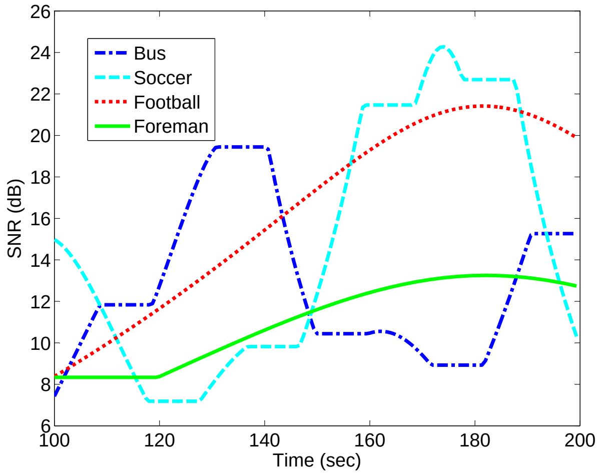


Figure 4.7: OPNET channel traces of the 4 video producers as a function of simulation time. The users are moving at a speed of 30 km/h in an urban macrocell.

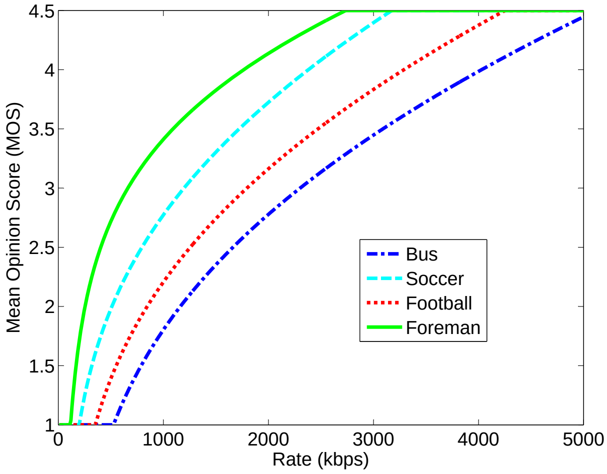
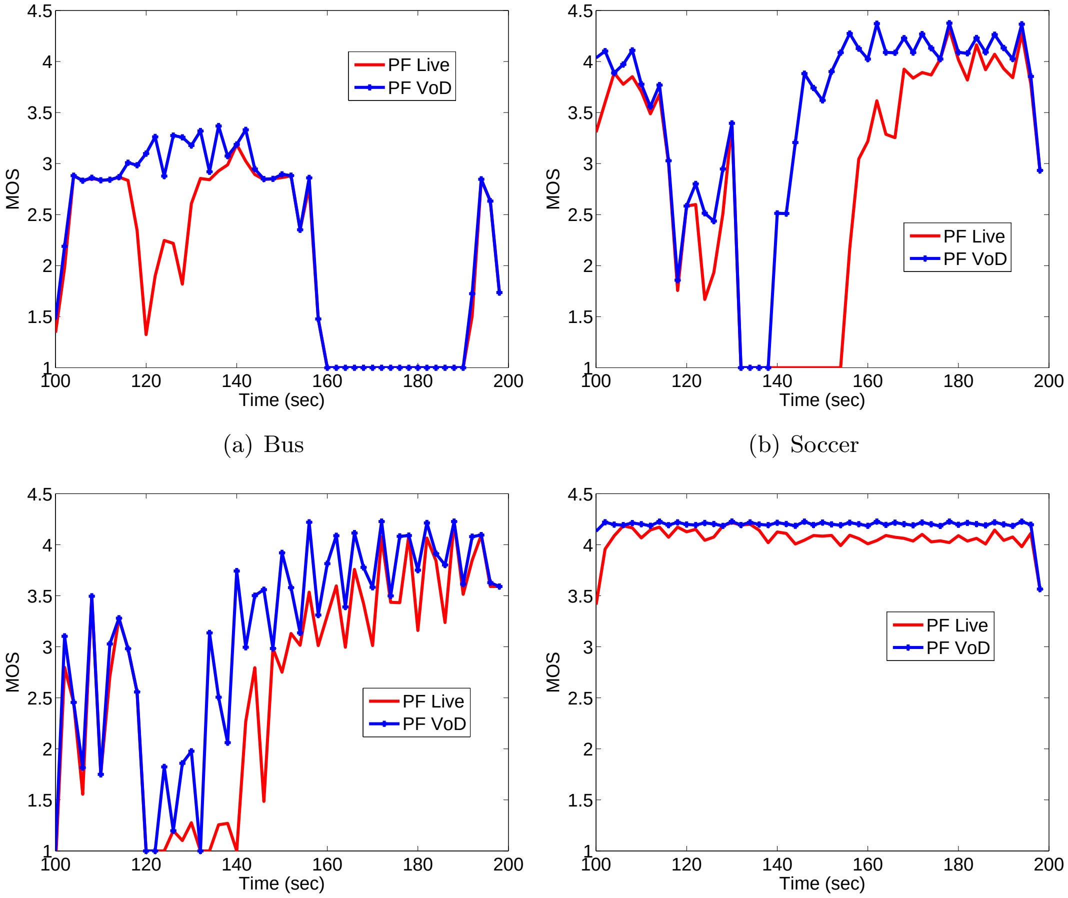


Figure 4.8: Utility functions for the upstreamed video content at the base station fitted using the parametric model from [CISN05].

### 4.6.1 Comparison of network-based QoE optimization and proportional fair scheduling

First, a single simulation run with the 4 video producers undergoing dynamic channel variations as depicted in Figure 4.7 is considered. The temporal quality of the upstreamed videos is analyzed for both network-based QoE optimization and standard non-optimized proportional fair scheduling. In both schemes, a distributed QoE optimization is applied at each mobile terminal for joint live and on-demand video layer scheduling, given the available network capacity. The low-complexity distributed approach from Section 4.3.3 is considered herein. For the network-based QoE optimization, it is assumed that a parametric utility function for each upstreamed video content is available at the base station [CISN05] (Figure 4.8). This means



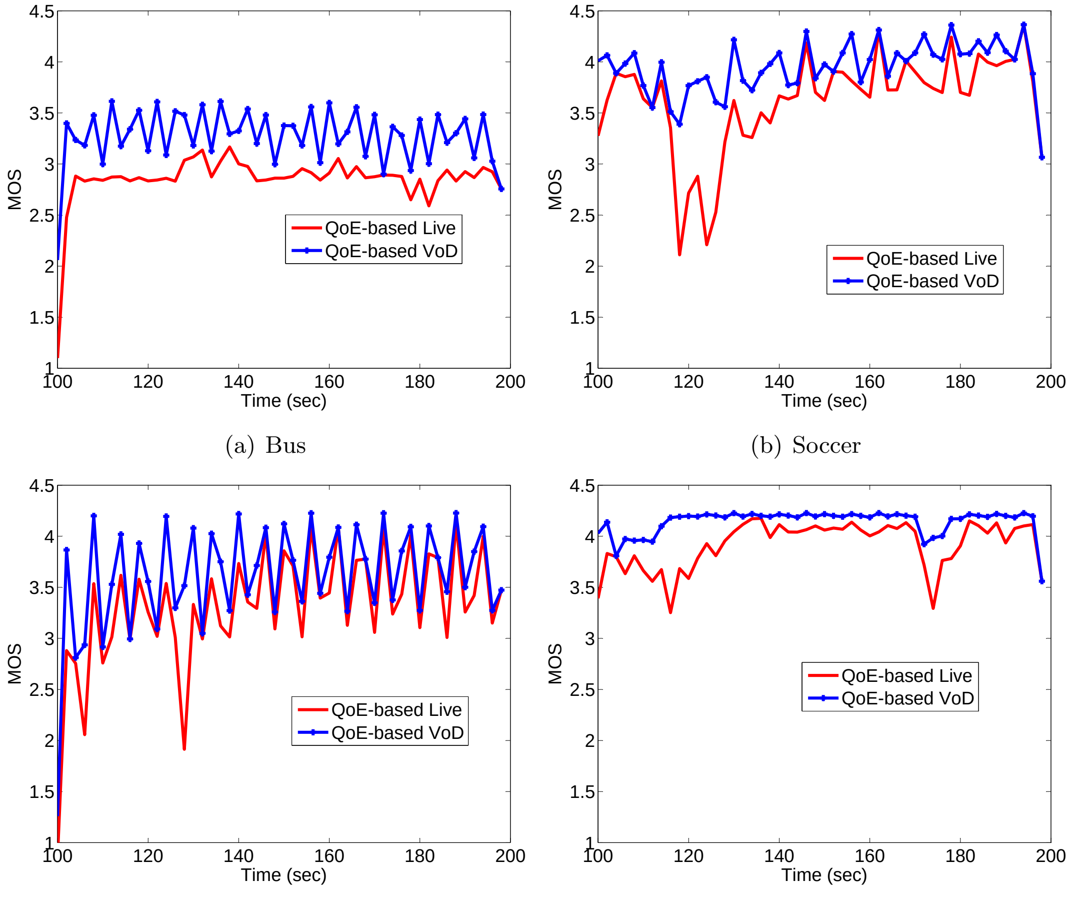
(a) Bus (b) Soccer

(c) Football (d) Foreman

Figure 4.9: MOS of the different users as a function of simulation time applying distributed QoE optimization at the mobile producers and proportional fair scheduling in the network. The PF Live and VoD curves show the MOS a video consumer will achieve when requesting the video from the portal for live and on-demand consumption, respectively.

that the mobile terminals do not need to exchange their instantaneous encoding parameters. This reduces the signaling of media-specific information between the base station and the mobile terminals and thus reduces the overall optimization complexity.

The PF scheduler (Figure 4.9(a)-4.9(d)) assigns the resources by considering the channel characteristics and irrespective of the uploaded video content. This results in high MOS for the less demanding videos while more demanding ones or users moving to the edge of the cell will suffer in streaming performance. For instance, the *Bus* and *Soccer* producers undergo bad channel situations between 150-180 sec and 120-150 sec, respectively. This leads to a severe degradation in the perceived video quality of the uploaded videos. Also, for the *Football* producer, the MOS will only improve after 140 sec as the user moves towards the center of the cell. Meanwhile, the less demanding *Foreman* producer who is in less favorable channel conditions can maintain a satisfactory video quality throughout the streaming session.



(a) Bus

(b) Soccer

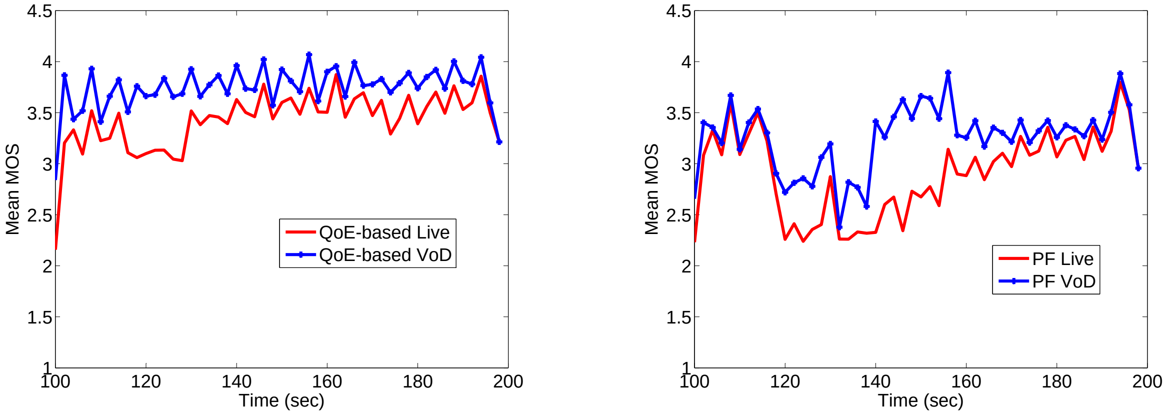
(c) Football

(d) Foreman

Figure 4.10: MOS of the different users as a function of simulation time applying both distributed and network-based QoE optimizations at the mobile terminals and the eNodeB, respectively. The QoE-based Live and VoD curves show the MOS a video consumer will achieve when requesting the video from the portal for live and on-demand consumption, respectively.

The network-based QoE optimization (Figure 4.10(a)-4.10(d)) distributes the resources by considering the channel and the video characteristics of each producer. This means that users undergoing bad channel conditions (e.g., *Bus* and *Soccer*) can benefit from the QoE optimization to upload a better video quality during channel fades. Furthermore, the demanding *Football* producer can achieve a higher MOS compared to the PF non-optimized case. This comes at the expense of a slight degradation in the MOS performance for the *Foreman* producer.

Figures 4.9 and 4.10 further illustrate the gains of the distributed QoE optimization at the mobile terminals for PF scheduling and network-based QoE optimization, respectively. In both cases a refined streaming performance is observed for all on-demand video consumers. Interestingly, the live quality of the *soccer* video in Figure 4.10(b) drops as the upstream user moves towards the edge (115 sec). In this case, only the base layer is uploaded for live



(a) Network-based QoE optimization.

(b) PF scheduler.

Figure 4.11: Mean MOS of all users as a function of simulation time. The network-based QoE optimization provides a better and smoother video quality compared to the LTE PF scheduler.

consumption. As the user moves again to the center of the cell (135 sec), the live video quality improves again and additionally the enhancement layers of previous GoPs are uploaded for on-demand consumption. For a VoD consumer, who requests the video from the portal after 20 seconds, the MOS will be substantially ameliorated (MOS increases by up to 1.5).

Figure 4.11 depicts the mean MOS of all the uploaded videos for the same simulation run. The network-based QoE optimized scheme sustains the average video quality and distributes the resources among the video producers to maximize the overall quality of experience. Meanwhile, the PF scheduler only adapts to the channel conditions which results in a fluctuating video quality. Moreover, both schemes show the potential improvement of applying the distributed QoE optimization at the mobile terminals. By caching and uploading the enhancement layers, the overall user experience for video on-demand consumption is enhanced.

Figure 4.12 shows the average mean MOS and the standard deviation over 50 simulation runs. As can be seen, for both schemes, the mean MOS for on-demand viewers can be improved by uploading additional enhancement layers before the playout deadline. In addition, the QoE-based optimization scheme manages to improve the mean MOS substantially for both live and on-demand consumers, when compared to the PF scheme. In fact, by adapting the transmission rates of the different producers, the network-based QoE optimization prevents the system from running into overload and thus allows for an improved overall user satisfaction. Besides, it results in less variation in the resulting mean MOS across the different simulation runs. Again, by shaping the transmission capacity of each terminal a more consistent user experience can be achieved.

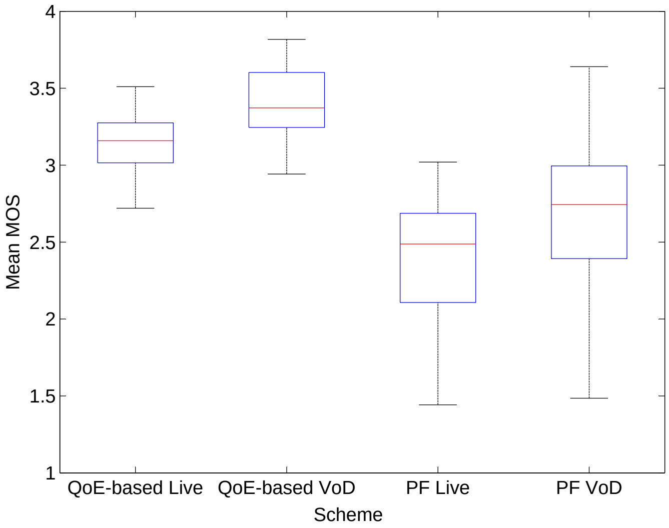


Figure 4.12: Average mean MOS of all users and the standard deviation of the mean MOS over 50 simulation runs.

## 4.7 Chapter summary

This chapter presents a systematic approach for live and on-demand uplink video transmission from video producers to sharing portals over next generation LTE networks. More specifically, a QoE-based uplink resource optimization system which involves both the network and the mobile producers is considered. Additionally, a video portal plays a key role by providing feedback to mobile terminals on the desired playout deadlines for the uploaded video content. The chapter proposes an optimal distributed model and a low-complexity iterative algorithm for scalable video transmission at mobile terminals, which adapts the uplink video transmission to the variable channel characteristics to provide optimal user experience for different types of consumers. The proposed QoE optimization approach is validated both theoretically and practically in an LTE OPNET simulator.

The experimental evaluation shows that the state-of-the-art proportional fair LTE scheduler results in suboptimal performance, in particular for resource-demanding videos. By acquiring general knowledge about the content characteristics, the base station can more efficiently shape the resources of each user to improve the level of user satisfaction in the mobile cell. The mobile terminals make the scheduling decisions based on their individual GoP characteristics and the available transmission rate. Indeed, both the network-based and the distributed QoE optimizations presented in this chapter can be implemented in real-time and require minimal signaling between the base station and the mobile equipments.

Moreover, this work exploits an important use-case for scalable video transmission over wireless networks. It is an enabler for different video communication services. In particular, live and time-shifted mobile video have been considered in this work. By properly adapting the media transmission to the available network resources, significant improvements in perceived video quality can be realized for on-demand video consumers while providing the live consumers with the best possible quality in real-time. A main challenge for the practical application of the proposed approach is the availability of real-time SVC encoders in the mobile devices. This should be realistic in the near future given the rapid improvements in the processing and storage capabilities of mobile phones.

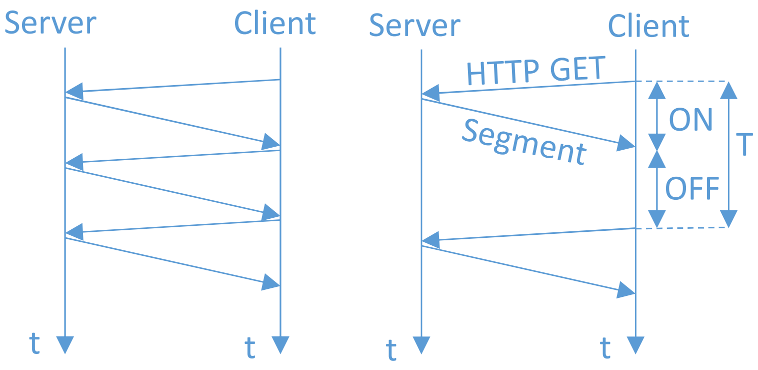
Chapter 5

# QoE-based adaptive HTTP downlink video delivery

This chapter focuses on the downlink transmission of adaptive HTTP video content over next generation mobile networks. It presents a QoE-driven approach for multi-user resource optimization in Dynamic Adaptive Streaming over HTTP (DASH) over LTE. The objective is to enhance the user experience in adaptive HTTP streaming by jointly considering the characteristics of the media content and the available wireless resources in the operator network. Specifically, a proactive QoE-based approach for rewriting the client HTTP requests at a proxy in the mobile network is proposed. The advantage of the proposed approach is its applicability for OTT streaming as it requires no adaptation of the media content. Furthermore, the proposed scheme is compared to both reactive QoE-optimized and to standard-DASH HTTP streaming. The performance evaluation, carried out using both objective and subjective experiments, shows that by taking a proactive role in determining the transmission and representation rates, the network operator can provide a better video quality and a fairer QoE across the streaming users.

## 5.1 Introduction

RTP/UDP-based streaming requires a specialized streaming server and is often blocked by firewalls. On the other hand, traditional HTTP/TCP progressive download is widely deployed nowadays (e.g., YouTube). Nevertheless, it does not support intra-session rate adaptation which results in frequent stallings under throughput limitations. Due to the TCP rate fluctuations, over-provisioning of the network resources is often required in conventional HTTP/TCP streaming. Indeed, [WKST08] concludes that the TCP throughput should be twice the video bit-rate to ensure a good streaming performance.



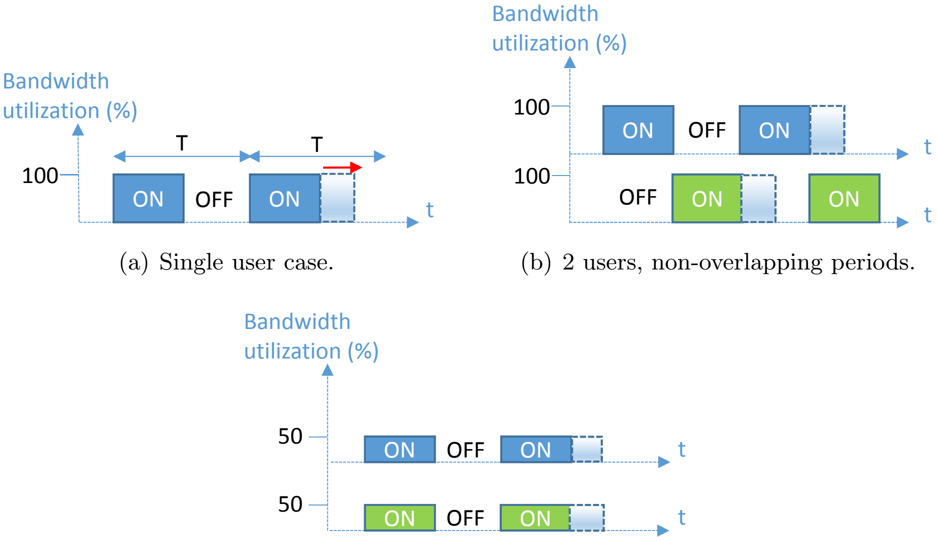
(a) Buffering state. (b) Steady state.

Figure 5.1: Typical client request and response timings during an adaptive HTTP streaming session [AABD12]. In the buffering state, a client builds up its buffer by requesting new segments after the current download finishes until a maximum buffer size is reached. In the steady state, a client aims to maintain a constant buffer size which results in ON-OFF periods.

DASH is specially designed to adapt the video quality to mobile networks with limited and highly variable resource availability. It provides multiple bitrate encodings of the same content which allows a client to continuously adjust its video rate during a streaming session to the current channel conditions. Nevertheless, DASH faces several challenges when it comes to video transport in multi-user mobile environments.

First, DASH gives the control of the streaming rate to the client. Recent studies show a fluctuating throughput and an underutilization of the network resources when multiple clients are competing for shared resources [ABD11] [AABD12]. Figure 5.1 explains a conventional client behaviour when requesting media segments from the server. At the beginning, a client tries to quickly build up its buffer by requesting a new segment immediately after the current segment is received until a maximum buffer size is reached (buffering state). Afterwards, a client aims to maintain a constant playout buffer size (steady state). In this case, the inter-request time is set to be equal to the video segment duration ($T$) which results in active ($ON$) periods, where the client is downloading a video segment, and idle ($OFF$) periods.

A client measures its TCP throughput during the ON intervals and then determines the streaming rate for the next segment. Figure 5.2 illustrates some cases where the ON-OFF intervals can be problematic when estimating the TCP throughput by a client. Figure 5.2(a) first considers a single user scenario who observes a TCP throughput which is higher than its video rate (i.e., $ON < T$) and consequently attempts to increase its streaming rate. In Figure 5.2(b), a second user is sharing the same resources and the ON-OFF intervals of the two users are non-overlapping. Subsequently, increasing the video rate of both users will lead to congestion as the network resources are already fully utilized in this case. Figure 5.2(c) shows the case when the ON-OFF intervals of the two users are fully overlapping. The


Bandwidth utilization (%)
100
T
T
ON
OFF
ON
t
Bandwidth utilization (%)
100
ON
OFF
ON
t
100
OFF
ON
ON
t
Bandwidth utilization (%)
50
ON
OFF
ON
t
50
ON
OFF
ON
t

(a) Single user case.

(b) 2 users, non-overlapping periods.

(c) 2 users, fully overlapping periods.

Figure 5.2: Illustrative example of ON-OFF periods at a client. The network resources are distributed in the time and frequency domain [AABD12] [LZG+13].

network resources are only half utilized in this case, and while the clients can reach an efficient throughput allocation at the end by increasing their streaming rates, the reactive behaviour of the client indicates a considerable delay to reach the optimal solution. A deeper discussion on the client behaviour can be found in [AABD12] [LZG+13].

Consequently, relying on the client decisions barely results in an optimal user experience. To address these issues, recent research focuses on improving the rate control logic at the client [JSZ12] or server [AADB13] for stabilizing the client's behaviour and efficiently using the network resources. Another promising approach is to transmit meta information that describes the quality of the representations such that the clients are made QoE-aware [THKP12]. All these approaches, however, optimize the HTTP streaming of a single client without further considering the influence on other DASH users sharing the same network resources.

To this end, there is a growing research interest in exploiting the multi-user adaptive HTTP streaming scenario. Instead of individually adapting the streaming rate of each user, the objective herein is to jointly optimize the transmission of multiple adaptive HTTP streaming users that share common resources. So far, the adaptive HTTP-based video delivery has been mainly studied from an end-to-end server-client perspective and the mobile network is treated as a black box [BAB11b] [MBBN11]. In the wireless network this means that DASH adapts the video quality individually for every user to the resources allocated by the scheduler in the eNodeB. The eNodeB, however, is typically not content-aware and the scheduler assigns resources only based on the channel conditions and without considering the characteristics of the transported content.

QoE-based resource allocation over wireless networks has been proposed for traditional RTP/UDP streaming (e.g., [TKSK09]) but has not yet been studied for adaptive HTTP media delivery. In-network content adaptation (e.g., transcoding [LG05]) is used to shape the transmitted video streams according to the QoE optimization result. This, however, is costly in terms of computational resources and induces additional delays. Meanwhile, DASH provides inherent adaptivity by encoding the same content at multiple bit-rates which simplifies the video adaptation compared to RTP/UDP based optimizations.

With respect to multi-user resource allocation for adaptive HTTP video delivery, it has been studied in [MB11] [HG12] [WSHS12]. In [MB11], network traffic management for adaptive HTTP video delivery across multiple clients is considered. The target bitrate is determined by the network based on available throughput estimates of all users. The authors in [HG12] conclude that a simple rate shaping policy in a residential gateway can improve the adaptive HTTP experience among two competing clients. [WSHS12] studies adaptive HTTP streaming in LTE networks with the aim of minimizing the number of interruptions. The investigated approaches, however, can be classified as reactive, i.e., they focus on optimizing the network resource allocation to meet an objective criteria and the clients react to the assigned resources. Recently, [PZC12] proposed a rate adaptation algorithm, WiDASH, for optimizing the adaptive HTTP streaming across multiple wireless clients. The proposed approach, however, does not consider the individual content characteristics of the different clients and aims at stabilizing the user throughput. Also, different from the approach presented in this chapter, the authors of [PZC12] propose to transcode the DASH stream, similar to typical RTP/UDP based optimizations (e.g., [TKSK09]), which is costly and may react too late.

The goal of this chapter is to evaluate the benefits of QoE-based traffic and resource management in the mobile network in the context of adaptive HTTP streaming. It exploits the mobile operator's knowledge about the radio conditions in the cell and the streaming users in terms of their content characteristics to maximize the overall user satisfaction. Different from previous work, the objective here is to proactively adapt the streaming rate and the network resources to the user perceived quality in adaptive HTTP streaming. Also, different from rate adaptation schemes which adjust to throughput variations, the QoE-based multi-user resource allocation approach directly considers the impact on the user quality of experience given that the streamed contents exhibit different rate-distortion characteristics.

The remainder of the chapter is structured as follows. First, Section 5.2 presents the proposed QoE-based optimization system for adaptive HTTP streaming. The experimental and the subjective evaluation results are then described in Section 5.3 and Section 5.4, respectively. These results are further validated in an LTE OPNET simulator as described in Section 5.5.

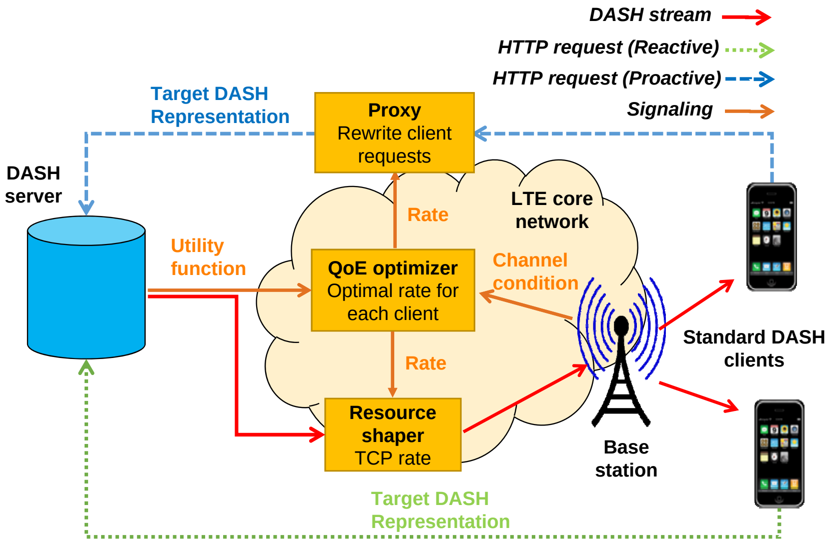


Figure 5.3: System image of QoE-driven adaptive HTTP mobile video delivery, illustrating both proactive and reactive optimization approaches.

## 5.2 QoE-based adaptive HTTP streaming system

A schematic depiction of the proposed QoE-based adaptive HTTP system is given in Figure 5.3. Multiple mobile clients are simultaneously downstreaming different DASH content over an LTE network. At the DASH server, the utility information of each content is first extracted and added to the MPD. The envisioned approaches for signaling meta information are explained in Section 5.2.1. At the base station or close to it, a QoE optimizer collects utility and channel information about the different clients, and then determines the target transmission rates using utility maximization, as described in Section 5.2.2. Furthermore, proactive and reactive approaches are considered for adapting the streaming rates, which are presented in Section 5.2.3.

### 5.2.1 Utility curve signaling

This study considers two options for providing meta information about the streamed DASH content. In the first case, the parametric model (PM) from [CISN05] is used which requires three pairs of rate and distortion to represent each video sequence. Please note that this represents a generic MOS-Rate function which can be delivered for instance at the beginning of the streaming session. The model from [CISN05] can be then used to generate an arbitrary set of MOS-Rate operating points. In the second case, the utility information is provided in the form of MOS-Rate pairs for each representation in the MPD. Although the DASH protocol does not explicitly define how to transmit utility information, it provides various options for this [ISO12]. The utility can, for example, be signaled in the initialization segments of the

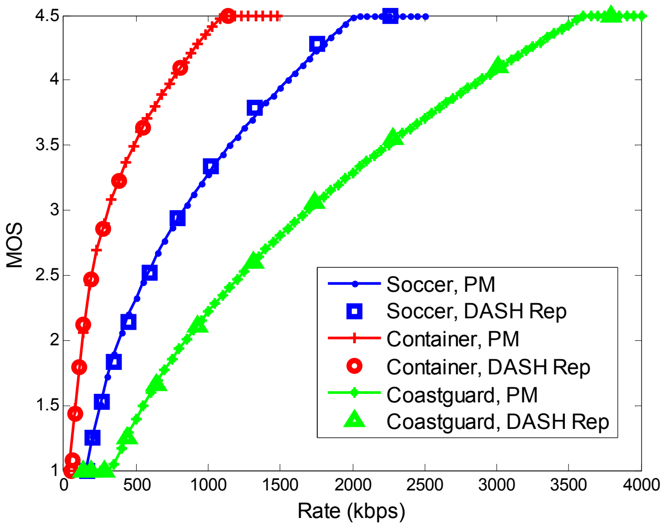


Figure 5.4: Utility curves using the actual DASH representations and by fitting using the parametric model from [CISN05].

MPD. In DASH, one initialization segment is allowed per representation. Alternatively, the utility of different representations within a program period can be added to the *Subset* element of DASH [THKP12].

To generate the DASH representations, a video sequence is encoded at different quality levels with the H.264/AVC video codec. Specifically, a total of 11 different quantization parameters ranging from 20 to 40 are used at the encoder to generate 11 representations of the same video. Then, the average bitrate and average MOS is calculated for each representation. The MOS is computed using the same linear mapping with the PSNR as in Section 3.2.1. Figure 5.4 shows the utility curves for three different video sequences which correspond to the discrete MOS values of the actual DASH representations and the interpolated values using the PM from [CISN05]. For the parametric model, three MOS-rate pairs that correspond to the quantization parameters of 40, 28 and 20 are used. The fitting of other MOS-Rate points is performed according to (2.3).

### 5.2.2 Optimization objective

The objective of the QoE-based resource allocation is to determine the transmission rates of all clients that maximize the overall user satisfaction. This work uses the objective function originally proposed in [TKS11]. The optimization problem for $K$ clients is given by:

$$\underset{(\alpha_1,\ldots,\alpha_K)}{\arg\max} \sum_{k=1}^{K} U_k(\alpha_k) - P_k \tag{5.1}$$

$$subject\ to \quad \sum_{k=1}^{K} \alpha_k = 1, \quad R_k \geq R_{min,k} \tag{5.2}$$

$$where\ P_k = \min(0, |U_k(\alpha_k)_t - U_k(\alpha_k)_{t-1}| - 0.23) \tag{5.3}$$

where (5.1) determines the network resource share of each user that maximizes the sum of utilities. It additionally penalizes the temporal fluctuations of the video quality which are perceivable by the users. Specifically, it adds a penalty term $P_k$ if the quality change between two successive optimization rounds (denoted by $t$ and $t-1$) exceeds a just noticeable difference (JND) threshold. In this work, an average JND threshold for all users is considered which has been derived using subjective tests and is equal to 0.23 MOS [TKS11]. (5.2) constrains on the available resources and defines a minimum rate that should be allocated to each user (e.g., lowest representation).

Each $\alpha_k$ value corresponds to the fraction of total PRBs assigned to user $k$ in each optimization round. A gradient-based greedy algorithm, similar to the work in [JF07], is used to determine the values of $\alpha_k$ (see Algorithm 1 in Section 3.3.2). Depending on the utility information two types of optimization are applied:

1) Continuous QoE optimization (**QoE**): In the case parametric meta data about the streamed content is available, the algorithm will search for the set of $\alpha_k$ values that maximizes (5.1). For arbitrary small $\alpha_k \rightarrow 0$, the algorithm can choose from a continuous set of rates for each user. The optimal bit-rate is returned by the QoE optimizer.

2) Discrete QoE optimization (**QoE-d**): When the actual MOS-Rate values of the DASH representations are available, the algorithm will choose from a discrete set of operating points. The set of $\alpha_k$ values corresponds to the encoding rates which are defined in the MPD. The actual target representation rate from the MPD is returned by the QoE optimizer.

### 5.2.3 Enforcement of video quality adaptation

Knowing the target transmission rate of each user as described in (5.1)-(5.2), the objective is to adapt the application to the data rates supported at the lower layers. When the DASH server and the mobile clients are contained in the operator network, the mobile operator can fully optimize the video delivery. This corresponds to the case of managed content where the server can adapt the streaming rate of each client to the target rate returned by the QoE optimizer. Internet video streaming, on the other hand, is dominated by OTT content where the server lies outside the operators' network. Specifically, two different paradigms for determining the streaming or representation rate of each user, given its target transmission rate are considered:

(A) **Proactive optimization**: Consider a proxy (e.g., at the edge of the wireless network) which intercepts the client HTTP requests and rewrites them according to the feedback from the QoE optimizer. In the proactive approach (Figure 5.3), the target rate of each client is signaled to a resource shaper and the proxy server. The resource shaper limits the TCP throughput of each client. In addition, the proxy rewrites the client HTTP requests and forwards them to the DASH server. Specifically, it rewrites the client requests to the closest lower representation. The DASH server and the DASH clients are unaware of the proxy operation, i.e., each client will decode and play an optimized representation for its requested segment. This approach exploits the multiple bit-rate encodings within the MPD and requires no adaptation to the media content, which makes it particularly suitable for OTT video streaming. It additionally allows a mobile operator to control the streaming rates of the different clients based on the cell load and the individual radio conditions.

Proactive optimization has been considered in the earlier work in [TKSK09] to shape the video streams so that the system does not go into congestion. This work makes use of the rate scalability in adaptive HTTP streaming to adapt the video streaming rates without the need to access and decode the transported video content.

(B) **Reactive optimization**: Alternatively, a mobile operator can adapt the network resource allocation without interfering with the client decisions. In the reactive approach, each client gets a TCP throughput equal to the target rate determined by the QoE optimizer (again enforced by the resource shaper in Figure 5.3). The representation rate, however, is only determined by the media streaming client which reacts to the throughput changes.

In both cases, a standard unmodified DASH client is considered. The approaches only differ in how the QoE-based resource allocation result is exploited for dynamic rate adaptation for overall QoE optimization.

## 5.3 Experimental results

### 5.3.1 Experimental setup

In the following, a single LTE cell with 8 clients requesting different DASH videos is considered. Specifically, 11 representations are available for each video as explained in Section 5.2.1. For the experimental evaluations, the Microsoft Smooth Streaming client [ABD11] and the DASH-enabled VLC client [MT11] are chosen because DASH clients are still under development. The two clients are used without any modification. Furthermore, a standard HTTP server is used and the wireless network is emulated. In other words, a resource shaper is placed between the server and the clients that limits the data rates per client to the output of the QoE optimizer. In order to emulate an LTE network, the Dummynet [Riz97] software is used on the server which allows enforcing bandwidth/bitrate limitations at TCP level by creating pipes. Several pipes outgoing from the server can be created and assigned to different users, so multiple downstream users will experience different bitrates. 50 simulation runs are considered in order to study the impact of different mobility patterns. All users start streaming at the same time. Table 5.1 summarizes the simulation parameters.

Moreover, the following schemes are compared:

- **QoE-Server**: The server encodes the video stream at the optimal rate returned by the QoE optimizer. This provides an upper limit on the achievable QoE in the case of managed video content.
- **QoE-Proxy**: The optimal rate is first signaled by the QoE optimizer to the proxy. The proxy then chooses the closest lower available streaming rate from the MPD, rewrites the client request and forwards it to the DASH server. This represents approach (A) in Section 5.2.3 where the DASH server is outside the control of the network operator.
- **QoE-d-Proxy**: Similar to the QoE-Proxy scheme. However, the optimizer uses the discrete utility representation and returns the target rate (MPD compatible) to the proxy which again rewrites the client request. This also represents approach (A).
- **QoE-Reactive**: Each client gets a TCP throughput equal to the optimal rate determined by the QoE optimizer. The streaming rate, however, is only determined by the media streaming client. This corresponds to approach (B).
- **Non-Opt**: Standard OTT DASH streaming where the transmission rate is determined by the content-agnostic LTE scheduler, and the streaming rate is dynamically decided by the standard DASH client. This represents the reference scheme for comparison.

### 5.3.2 Comparative evaluation of the different schemes

The Microsoft Smooth Streaming client is first considered for assessing the different approaches. Moreover, the MOS is measured from 20 to 60 seconds to exclude the client dependent start-up behaviour. In Figure 5.5, the distribution of average mean MOS for the different schemes is depicted in order to highlight the differences in the overall performance. The QoE-Server approach represents the optimal performance when all users are streaming at their optimal rates. The MOS for the QoE-Proxy approach will drop compared to the QoE-Server scheme as only a discrete set of representations is available. Meanwhile, the MOS degradation is less noticeable in the QoE-d-Proxy approach and is close to the optimal value (QoE-Server) as the QoE optimizer considers the actual DASH representations in the optimization problem. The QoE-Reactive scheme improves the perceived video quality compared

Table 5.1: Experimental and simulation setup parameters for the adaptive HTTP downlink streaming scenario

| | |
|---|---|
| **Application parameters** | |
| Video codec | H.264 AVC, CIF, 30 fps |
| Application type | Adaptive HTTP streaming |
| Segment size | 2 sec |
| Number of clients | 8 |
| Client software | MS Smooth Streaming, DASH-enabled VLC [MT11] |
| **LTE parameters** | |
| System Bandwidth | 5 MHz |
| Number of PRBs | 25 |
| PRB size | 12 subcarriers |
| Subcarrier spacing | 15 KHz |
| Bandwidth per PRB | 180 KHz |
| SNR averaging cycle | 2 sec |
| Link layer model [3GP08b] | see (3.2) |
| Channel model | Urban macrocell |
| Shadowing | disabled |
| User speed | 30 km/h |
| Default scheduler | Round Robin (RR) |
| **Experimental parameters** | |
| Number of runs | 50 |
| Experiment length | 60 sec |

to the non-optimized DASH scheme (Non-Opt). The MOS for the QoE-Reactive scheme, however, drops compared to the proactive approaches as the client reacts late to throughput changes and does not always converge to the best representation level. Please note that for the QoE-Reactive and Non-Opt schemes the streaming rate is only determined by the client. The QoE-Reactive, QoE-Proxy, QoE-d-Proxy and QoE-Server schemes improve the average user satisfaction by 0.2, 0.36, 0.48 and 0.57 on the MOS scale compared to the Non-Opt scheme, respectively.

Furthermore, the mean MOS for the individual videos is evaluated in order to explain and further highlight the benefits of the QoE optimization that considers the content characteristics (Figure 5.6). The Non-Opt scheme provides a very good performance for the less demanding videos but fails for more demanding ones like *bus*, *coastguard* and *harbour*. Meanwhile, the QoE-based schemes allocate the resources among the users such that the overall user satisfaction is maximized. This results in substantial gains in perceived video quality for the demanding users while maintaining the MOS for the less demanding videos.

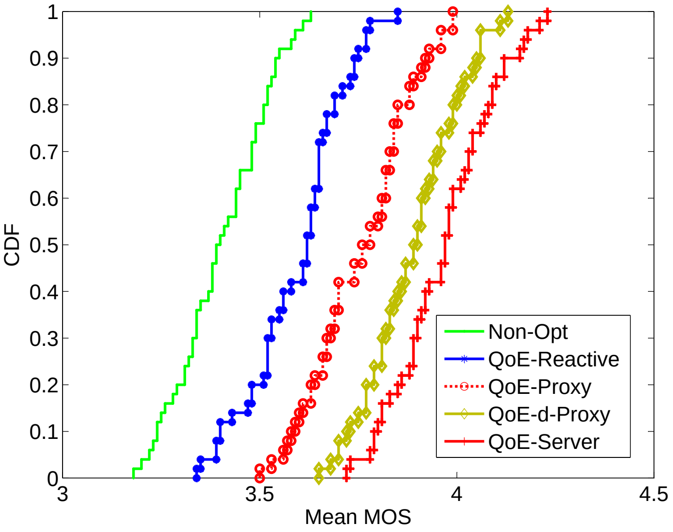


Figure 5.5: CDF of the mean MOS for 8 users. The figure illustrates the MOS gains for different QoE-based optimization approaches compared to non-optimized standard adaptive HTTP streaming. QoE-Server, QoE-Proxy, and QoE-Reactive, which consider generic utility curves, show the respective MOS gains by encoding the video at the optimal rate, adapting the streaming rate to an available representation rate and by only shaping the network resources, respectively. QoE-d-Proxy, which considers the actual MOS-Rate points, directly returns an MPD compatible representation rate thereby closing the gap to QoE-Server.

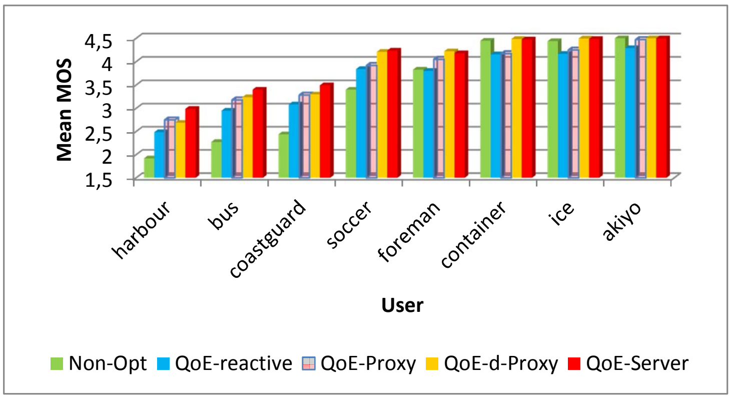


Figure 5.6: Individual performance averaged over 50 simulation runs. MOS gains are more perceivable for demanding videos such as *bus*, *coastguard*, and *harbour*. Less resource-demanding users achieve approximately the same MOS performance for the different optimization schemes.

### Temporal quality analysis

In the following, the temporal video quality of the QoE-Proxy approach is assessed and compared to the QoE-Reactive and Non-Opt schemes. Specifically, two separate experiments are conducted, one with the Microsoft Smooth Streaming client and one with the DASH-enabled VLC client [MT11]. Please note that in both experiments the clients remain unmodi-

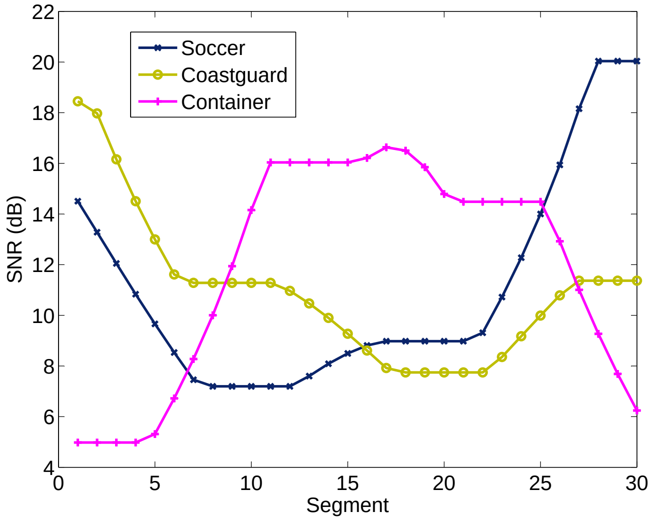


Figure 5.7: OPNET channel traces of 3 users as a function of time (segment = 2 seconds).

fied and thus the control mechanisms and buffering behaviour depend on the respective client implementation.

Figure 5.7 shows the signal-to-noise ratio (SNR) of three users who are undergoing dynamic channel variations. Figure 5.12 (a)-(f) show the requested representations by the users for the different schemes. The available transmission capacity for each user is shown for the non-optimized (Throughput RR) and the optimized (Throughput QoE) cases, as determined by the default LTE scheduler and the QoE optimizer, respectively. The Non-Opt and QoE-Reactive schemes indicate the corresponding representations as requested by the media client in both cases, respectively. For the QoE-Proxy approach, the client is unaware of the rewriting of the HTTP requests and can decode the redirected segments in both experiments.

In Figure 5.12 (a) and (b), both clients start streaming at a low bit-rate whereas the QoE-Proxy approach can start streaming at a much higher rate. Indeed, the user is experiencing favorable channel conditions and rewriting the requests cancels the slow-start behaviour of the standard clients. The Microsoft Smooth Streaming client smoothly increases the representation rate while the DASH-enabled VLC client immediately switches to a higher rate after the start-up phase (5 segments). As the channel conditions of the *Soccer* user deteriorates, the QoE-Proxy approach switches to a lower representation while the other approaches can continue playout at a higher rate for some time, as they have already buffered these representations during the start-up phase. As the channel conditions improve again (after 15 segments), QoE-Proxy can quickly switch to a higher rate whereas the other approaches are late to react.

Figure 5.12 (c) and (d) consider the demanding *Coastguard* user who undergoes some bad channel conditions after a good start-up phase. In this case, the DASH-enabled VLC client will run into a sequence of "rebuffering" events (e.g., after 14, 19, and 25 segments for the QoE-Reactive approach). The Microsoft Smooth Streaming client can continue to play the video at a higher rate before it switches to a lower representation. This is again explained by the buffered segments when the user was in an excellent channel condition.

Figure 5.12 (e) and (f) consider the *Container* user whose channel quality is bad at the beginning of the streaming session and starts improving afterwards. The Microsoft Smooth Streaming client and the DASH-enabled VLC client slowly adapt to the channel improvements and would only converge to the rate of the QoE-Proxy scheme after 27 and 24 segments, respectively.

These results show that the DASH-enabled VLC client provides the worst user experience as it fails to maintain quality during deep channel fades and runs into "rebuffering" mode. The Microsoft Smooth Streaming client keeps a large buffer which allows it to maintain a good quality when the channel degrades but often reacts very late and does not converge to the best representation. Meanwhile, the proposed QoE-Proxy approach can fully utilize the network resources and provide the best possible representation for each client under different channel conditions.

#### Perceptual video quality assessment

Furthermore, the perceptual gains of the proposed QoE-Proxy approach with respect to the QoE-Reactive scheme are illustrated using the Microsoft Smooth Streaming client. Figure 5.13 shows snapshots captured at different time instants during the start-up phase and after the quality has stabilized. The figure shows remarkable gains in perceived video quality that can be achieved by proactively adapting the video transmission rate by the QoE-Proxy scheme.

## 5.4 Subjective quality assessment

In order to assess how the different schemes impact the video perception, a subjective evaluation is performed with human subjects. For the subjective tests, the Microsoft Smooth Streaming client is selected because it performed better than the DASH-enabled VLC client [MT11] in the objective experiments. Without loss of generality, the following results can be applicable to any DASH system. Furthermore, a standard HTTP server is used and the wireless network is emulated similar to the experiments in Section 5.3.1.

Two scenarios each with 8 adaptive HTTP streaming users in one LTE cell are simulated. The simulation parameters are presented in Table 5.1. Each user is downstreaming a different video with specific rate-distortion characteristics (*soccer, ice, bus, coastguard, foreman, akiyo, container, harbour*). In the first scenario (scenario 1), users move in the cell with a speed of 30km/h. The second scenario (scenario 2) presents more rapidly changing channel conditions as all users move with a speed of 120km/h. Moreover, the QoE-Proxy and the

QoE-Reactive approaches are assessed. These methods are also compared to a non-optimized scheme (Non-Opt) where the PRBs are equally shared among the users, and the streaming rate is dynamically decided by the client. As a result, there are 6 cases to evaluate.

### 5.4.1 Test methodology

The subjective test is conducted using the SAMVIQ (Subjective Assessment of Multimedia Video Quality) method [IR07], which is specifically designed for assessing multimedia applications. The viewer is given access to several versions of a video which he can select and play through a computer graphic interface and then score as desired (Figure 5.8). The test videos are presented to the subjects one-by-one. That is, for each video sequence, the subjects can score the different approaches (buttons A to H) and can switch to the next video after rating the previous one.

A 10 seconds long sequence extracted from the simulated 60 seconds is presented to the test subjects (viewers) for each scenario and each optimization method. More precisely, seconds 30 to 40 from the sequences are extracted. This allows avoiding the typical poor quality start-up phase of the adaptive streaming client (Microsoft Smooth Streaming). Additionally, 10 seconds of videos are a recommended duration for conducting subjective tests [IR07]. Besides the evaluated 6 cases, viewers also rate a reference sequence (best possible quality), a hidden reference and a poor quality sequence. The quality anchors are recommended to stabilize the subjective results [IR07]. Viewers rate the videos on a continuous scale from 0 to 100. After the screening procedure described in [IR07], the data of 20 test subjects was verified to be valid. For each sequence, the average rating over the 20 viewers is taken and a differential quality score (DMOS) [IT99] value is then computed by:

$$DMOS = \overline{R}(sequence) - \overline{R}(hidden\ reference) + 100 \tag{5.4}$$

where $\overline{R}$ is the average rating over the 20 viewers. The DMOS value is used in the data analysis as the subjective quality rating.

### 5.4.2 Test results

Figure 5.9 shows the mean DMOS over the 8 users in the cell. Additionally, the boxplot illustrates the distribution of the DMOS values for the 8 users.

The first observation is that for both scenarios, QoE-Proxy achieves the best mean DMOS, that is, provides the best mean QoE for the users. The QoE-Reactive approach achieves a better mean DMOS than a non-optimized approach. That is, both QoE-based optimization approaches improve the mean QoE compared to a non-optimized approach. In the more dynamic scenario (scenario 2), the gains in term of mean QoE are more important compared to the less dynamic scenario (scenario 1).

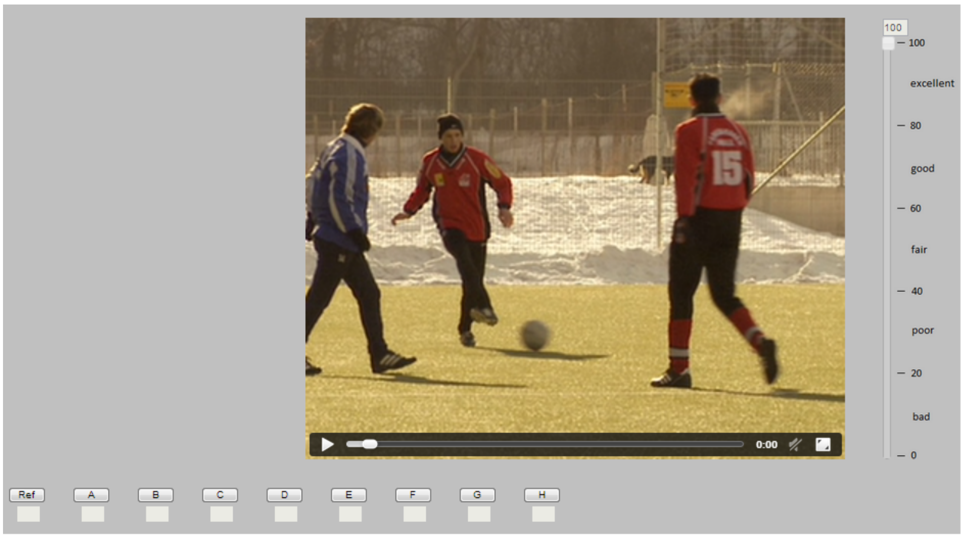


Figure 5.8: Graphical user interface, used in the subjective test, implemented using the SAMVIQ method [IR07].

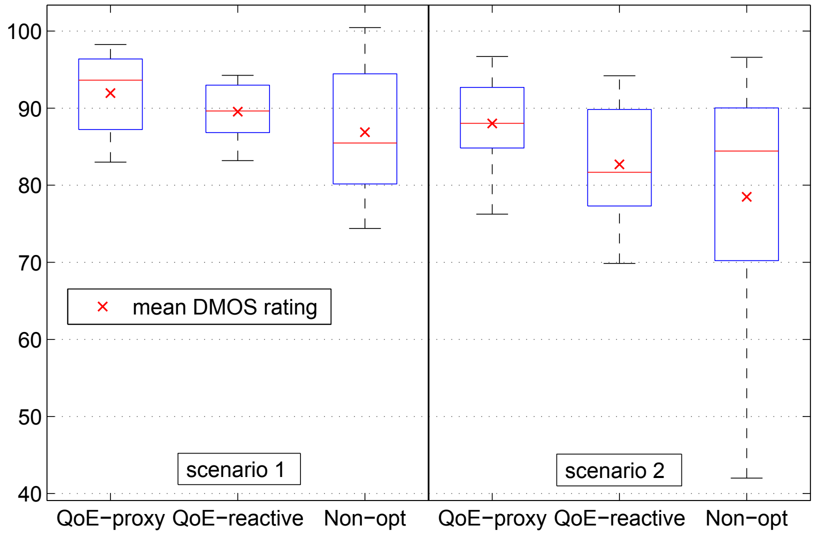


Figure 5.9: DMOS of 8 users for 30 km/h (scenario 1) and 120 km/h (scenario 2). The error bars represent the worst and best user ratings. The boxplot shows the median, 25*th* and the 75*th* percentiles.

Second, the boxplot indicates a larger spread of the users' DMOS for the non-optimized scheme. This indicates that in a non-optimized case, some users will experience an excellent quality (e.g., low-demanding video users) and some may experience a bad quality due, for example, to bad channel conditions and a high-demanding video. On the contrary, both QoE-based approaches present a lower variance, which indicates that most users will experience a similar quality, i.e., the fairness is improved by using a QoE-based method.

A detailed description of the individual results for each video and comparison with other objective QoE models is given in Appendix A.2.

## 5.5 LTE OPNET simulator results

### 5.5.1 Simulation setup

The adaptive HTTP optimization framework is further integrated and evaluated in an LTE OPNET simulator. The simulator provides transport control between the HTTP server and the HTTP clients using the TCP protocol. Nevertheless, adaptive HTTP is not directly supported in the OPNET simulator. Therefore, custom application profiles are used to set up an adaptive HTTP session between an HTTP server and multiple clients in a single LTE cell. Moreover, a client-driven rate adaptation algorithm [LBG11b] is implemented to decide on the segment switching at the mobile users. In addition, a rate controller module is implemented to limit the transmission rate of each client to the output of the QoE optimizer.

**Client-driven adaptive HTTP algorithm [LBG11b]**

A client-driven rate adaptation technique for adaptive HTTP streaming is implemented based on the approach from [LBG11b]. In this algorithm, the client compares the segment fetch time (SFT) of the segment with the media segment duration (MSD) to estimate the available data rate. So the client switches up/down to different bitrates between the available representations after receiving the last segment and before sending the next request. The SFT is defined as the time between sending the segment request and receiving the last bit of that segment. Moreover, the ratio of the media segment duration to the segment fetch time is used as a metric to detect the network capacity.

$$\mu = \frac{MSD}{SFT} \tag{5.5}$$

In this algorithm, switching to different representations is performed as follows:

*Switch up*: If $\mu > 1+\epsilon$ then the client switches up to the next higher bitrate representation level, where $\epsilon$ is the switch up factor and can be determined as

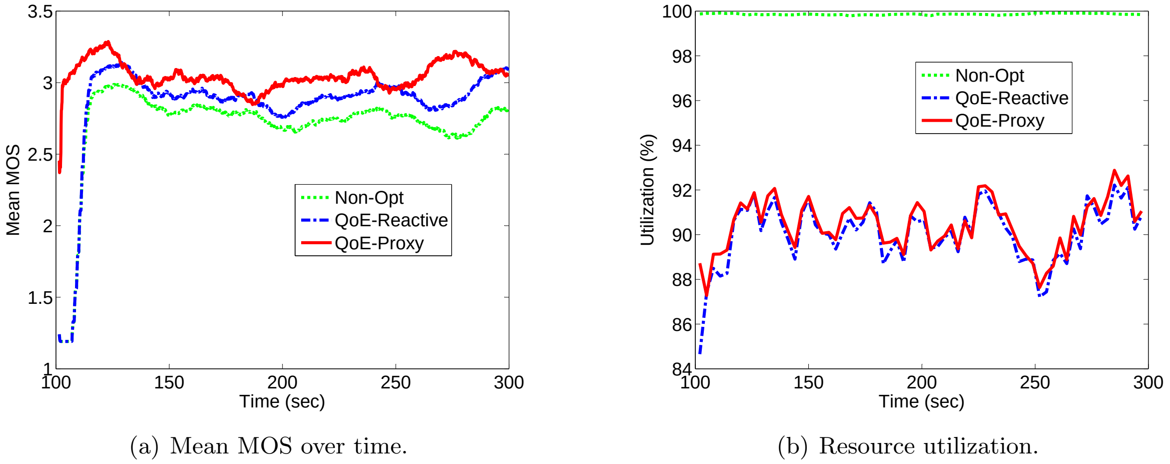
$$\epsilon = \max_{i=1\ldots N-1} \frac{br_{i+1} - br_i}{br_i} \tag{5.6}$$

where $br_i$ denotes the bitrate of representation $i$ and $N$ is the highest representation level.

*Switch down*: If $\mu < 1$ the client will switch to a lower representation. In the switch down case, an aggressive switch down will be performed and the highest bitrate representation that satisfies inequality (5.7) will be selected.

$$br_i < \mu \cdot b_c \tag{5.7}$$

where $b_c$ denotes the bitrate of the current representation.



(a) Mean MOS over time.

(b) Resource utilization.

Figure 5.10: Performance comparison of the different schemes for 8 users streaming adaptive HTTP content in a simulated LTE cell. The results are averaged over 20 simulation runs.

### Rate controller

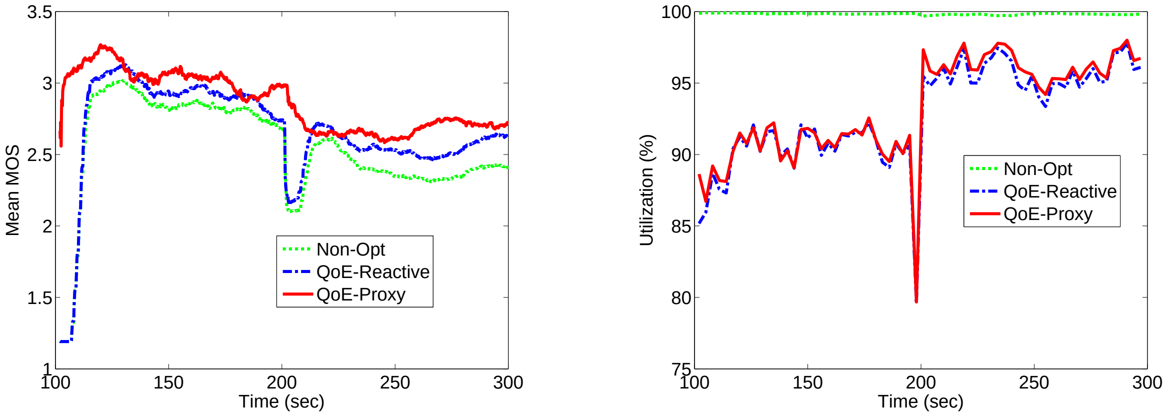
A rate controller is implemented at the eNodeB that reshapes the transmission rate of each user based on the output rate of the QoE optimizer. This is similar in function to the rate shaper used in the previous experiments. In adaptive HTTP streaming, the server sends the requested segments into the network at very high speed. This means that mobile users streaming at low data rates (e.g., non resource-demanding videos) will fetch much data within a short interval and quickly fill up their buffers while mobile users streaming at high data rates will experience larger delays. The rate controller is implemented at the eNodeB scheduler. It takes the optimized data rates for each bitstream as input from the QoE optimizer. The input traffic from higher layers is placed into the temporary buffers at the scheduler and the rate controller regulates the output rate. In short, the rate controller module limits the data rate of each bitstream to the QoE optimizer rate.

## 5.5.2 Simulation results

An LTE macrocell is considered with the adaptive HTTP custom application deployed. The mobile users are HTTP streaming clients and the server is an HTTP server. A total of 11 representations are available for each video. A segment duration of 1 second is considered. The clients start playout immediately after receiving the first segment. The simulation setup parameters are the same as in Table 5.1. Moreover, the QoE-Proxy, QoE-Reactive and Non-Opt schemes are compared (Section 5.3.1). For the Non-Opt scheme, the assigned resources are determined by the proportional fair scheduler in the LTE network.

### Eight user scenario

In the first simulation setup, 8 users are considered in a simulated LTE cell. Users start



(a) Mean QoE over time.

(b) Resource utilization.

Figure 5.11: Performance comparison of the different schemes with 8 users starting streaming at 100 sec and 5 more users joining at 200 sec. The results are averaged over 20 simulation runs.

streaming after 100 seconds and are continuously streaming for 200 seconds. Figure 5.10(a) shows the resulting QoE averaged over the 8 users and 20 simulation runs. The QoE-Proxy provides the best user experience, which confirms the previous results. The QoE-Reactive approach also performs better than the non-optimized case. Specifically, QoE-Reactive provides a mean gain of 0.16 on the MOS scale compared to the non-optimized case and QoE-Proxy provides a mean gain of 0.28 on the MOS scale. Furthermore, the number of buffer underflows are analyzed for the QoE-Proxy approach. Over the 20 simulations with 8 users requesting each 200 segments, which makes at total of 32000 segments, there are 33 playout interruptions. This corresponds to 0.1% of the segments. This shows that the QoE-Proxy approach not only offers the best visual quality, it also provides an almost interruption-free playout.

Figure 5.10(b) shows the percentage of resource utilization in the LTE cell over time for the different methods. In the non-optimized case, the resource utilization is constantly 100% over time. There is no rate-adaptation in this case, thus there is no limitation on the data that comes from the server to the eNodeB. The buffers at the eNodeB are filling without control, which can lead to congestion. On the other hand, in the QoE-based approaches, the resource utilization is about 90%. This is due to the rate control at eNodeB (Section 5.5.1) where the rate for each user is limited to the output of the QoE-optimizer. In this case, congestion is avoided, and although the radio resources are not totally exhausted, the QoE can be improved for the users as shown in Figure 5.10(a).

### Thirteen user scenario

In the second simulation setup, 8 users (*soccer*, *ice*, *bus*, *coast*, *foreman*, *akiyo*, *container*, *harbour*) start streaming at time t=100s. Additionally, 5 streaming users (*mother*, *crew*, *city*, *football*, *bridge*) join the cell at t=200s. Figure 5.11(a) shows the QoE on a MOS scale averaged over all users and 20 simulations. The QoE-Proxy approach provides the best QoE over time. The QoE-Reactive also shows an improved QoE compared to the non-optimized case. When the 5 additional users join, the mean QoE over all users drops sharply for the non-optimized case and the QoE-Reactive method. This is due to the start behavior of the client, which starts requesting video at a low quality. On the other hand, the QoE-Proxy method can provide a good quality from the beginning of the streaming session, which explains that the mean QoE is not dropping sharply as in the QoE-Proxy case when the additional users join. In addition, the number of playout interruptions are examined for the QoE-Proxy method. In total there are 42000 segments transmitted during the 20 simulations, and 88 interruptions are observed. This corresponds to 0.2% of the total number of segments. This confirms that the QoE-Proxy method is able to provide an almost interruption-free playout.

Figure 5.11(b) shows the mean resource utilization in the cell over the 20 simulations for the different approaches. The resource utilization is 100% for the non-optimized scenario, as there is no rate control in the cell. In both QoE-based approaches, the cell load is about 90% when 8 users are streaming, similar to the case presented in Figure 5.10(b). When 13 users are streaming, the cell load is about 95%. This means that the rate control allows to avoid congestion in the cell, while providing a better QoE if the rate is controlled according to the QoE-optimizer. A minor drop in the resource utilization is observed at t=200s when the 5 additional users join the cell. In the first second where the users join, no estimation of their channel conditions has been performed yet, and they are assigned a low bit-rate.

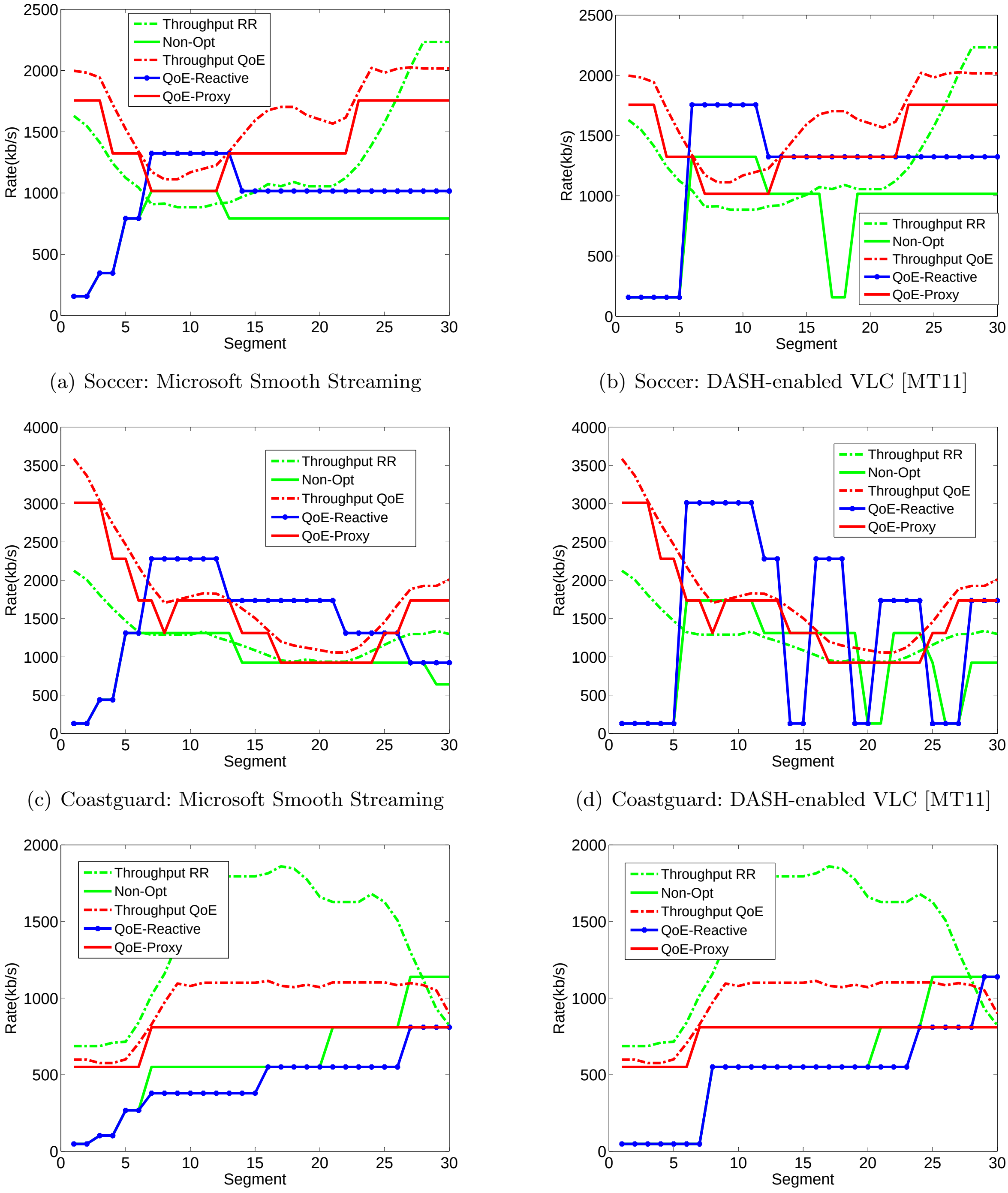
## 5.6 Chapter summary

This chapter presents a QoE-based approach for jointly optimizing the adaptive HTTP media delivery across multiple clients in a wireless cell. By considering the media characteristics of each content and the channel conditions of the mobile users the optimal streaming rate is determined. The experimental results show that QoE-based multi-user resource allocation improves the user experience compared to non-optimized OTT adaptive HTTP streaming, when using the same client adaptation algorithm. In addition, a QoE-based proxy approach is proposed for redirecting the HTTP client requests to the optimal streaming rate, and which can be still decoded by a standard DASH client. While client-based adaptation approaches react late to channel variations and fail to stream at the optimal rate, significant gains in user perceived video quality can be achieved by proactively adjusting the streaming rate. For the performance evaluations, two standard adaptive HTTP clients are considered, namely the Microsoft Smooth Streaming and the DASH-enabled VLC [MT11].

The experimental results are further corroborated by subjective tests, that reveal a higher MOS, with improvements more notable in dynamic scenarios. Additionally, a fairer user experience, up to 35% MOS increase for the worst-case user, compared to non-optimized

OTT DASH is reported. Finally, the LTE OPNET simulator results confirm the superiority of the proactive approach compared to the reactive approaches in terms of mean MOS over the users. Additionally, the investigation of playout interruptions shows that the proxy scheme can provide an almost stalling-free experience for the users.

In this chapter, both discrete and continuous QoE-based approaches for resource allocation are investigated. One limitation of the continuous optimization is the requirement of concave utility functions for modeling the R-D behaviour of the videos. In this work, the different DASH representations are generated by varying the quantization parameter settings. More generally, when varying the frame rate or spatial resolution the concavity property might not hold. The discrete optimization, however, directly operates on the actual MOS/Rate pairs and can be still applied in this case. Nevertheless, more complex mapping models which characterize the MOS as a function of the different encoding parameters need to be considered.



(a) Soccer: Microsoft Smooth Streaming

(b) Soccer: DASH-enabled VLC [MT11]

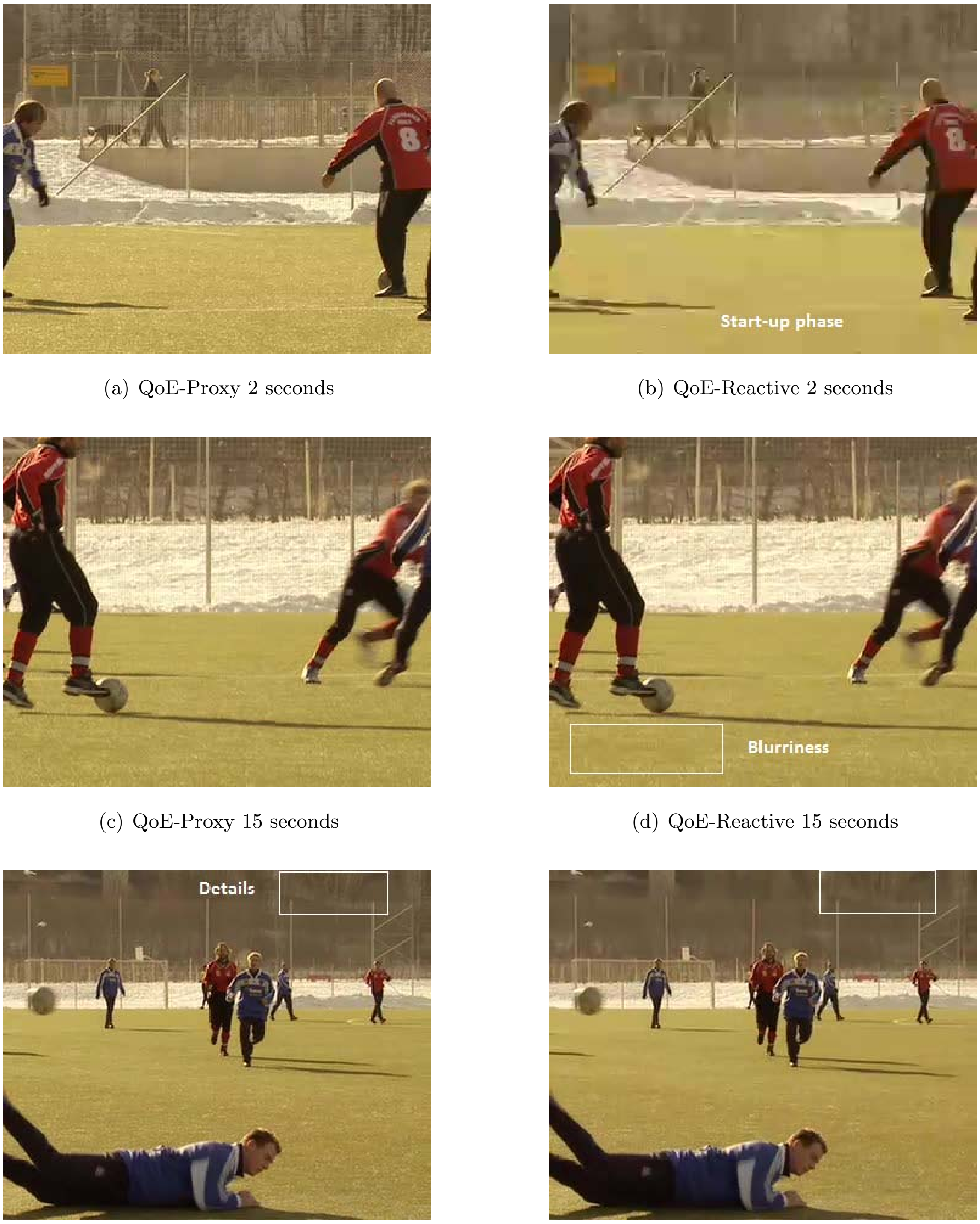
(c) Coastguard: Microsoft Smooth Streaming

(d) Coastguard: DASH-enabled VLC [MT11]

(e) Container: Microsoft Smooth Streaming

(f) Container: DASH-enabled VLC [MT11]

Figure 5.12: Temporal assessment of the different optimization approaches. Throughput QoE and throughput RR are the available rates for each user as determined by the QoE optimizer and RR scheduler, respectively.



(a) QoE-Proxy 2 seconds

(b) QoE-Reactive 2 seconds

(c) QoE-Proxy 15 seconds

(d) QoE-Reactive 15 seconds

(e) QoE-Proxy 29 seconds

(f) QoE-Reactive 29 seconds

Figure 5.13: Comparison of the visual quality for the QoE-Proxy approach and the QoE-Reactive scheme.

## Chapter 6

# QoE-based traffic and resource management for adaptive HTTP streaming

Despite the growing interest in adaptive HTTP streaming, there is still a lack of understanding of the mobile network's potential in enhancing the user satisfaction, particularly for OTT services. This chapter provides an insightful study of the multi-user resource allocation in the network and the representation selection of the individual clients. Specifically, it extends the proposed proxy approach in Chapter 5 to consider the playout buffer level of the clients, and then considers distributed and joint QoE optimization approaches for solving the corresponding rate allocation problems. A playout buffer-dependent approach is proposed that determines the representation rate of each client based on its buffer time and the achievable QoE under current channel conditions, which smoothes the video playout over time. Furthermore, the simulation results show that by jointly solving for the representation and transmission rates, the wireless network resources are more efficiently allocated among the users and substantial gains in the user perceived video quality can be achieved.

## 6.1 Introduction

In Chapter 5, different QoE-based traffic management approaches are investigated for enhancing the adaptive HTTP video delivery in next generation mobile networks. In particular, the proposed proxy approach proactively adapts the streaming rates of the clients to the result of a network resource allocation process that maximizes the overall utility. Indeed, by giving control of rewriting the client decisions to a network proxy, the responsiveness to channel

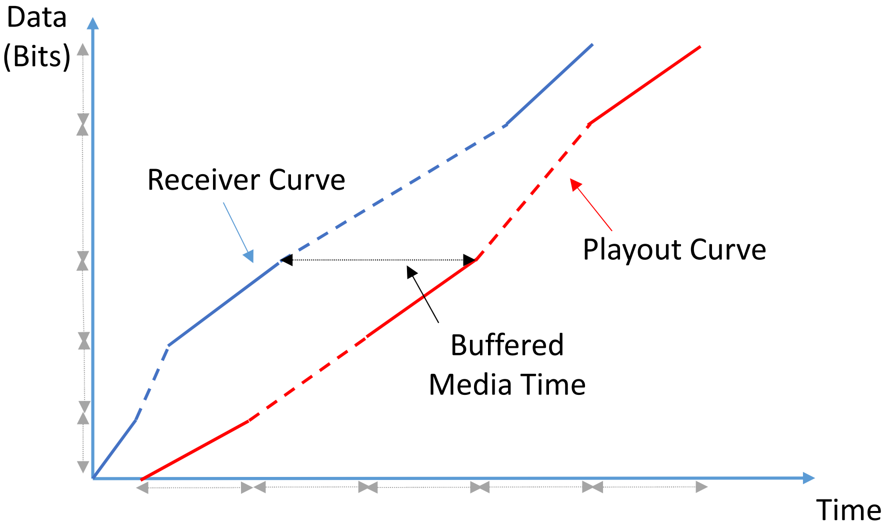


Figure 6.1: Receiver and playout curves of an adaptive HTTP video stream. Video segments have a variable size and the same playback time.

variations is improved and a higher perceived video quality is achieved.

This chapter extends the optimization approach in Chapter 5 to additionally consider the playout buffered video time at the client. In non-adaptive streaming, the buffer size (in bits) is commonly used to measure the amount of buffered video at the client [SJK04]. In adaptive HTTP streaming, the video segments correspond to representations with different bitrates and that are characterized by a constant playback time (Figure 6.1). Therefore, the buffered video time is used as a measure of the buffer fullness at the client.

The research on adaptation approaches for HTTP-based media delivery that relies on client buffer information can be classified into two main categories:

- *Client-based:* The buffered media time is considered as a key parameter in designing client-based adaptation strategies for adaptive HTTP streaming. Recently, much research work studied the combination of rate, utility and buffer information to improve the playback experience at the client (e.g., [ZXD$^+$13], [TL12]). In their evaluation of the end-to-end QoE in adaptive HTTP streaming over LTE, the authors in [OS12] conclude that playout interruptions can be largely reduced compared to HTTP-based progressive download. A client will typically switch to a lower streaming rate once its buffered media time drops below a certain threshold in order to avoid playback stalls. In addition, experimental results have shown that mature adaptive clients eliminate playout interruptions in real scenarios [MLT12]. Indeed, recent studies show that it is important to address the multi-level rate switches as a measure of user dissatisfaction in adaptive HTTP streaming (e.g., [KAS12], [MCLC11]). In this respect, [MLCC12] concludes that gradual switches in the representation rates are preferred in adaptive HTTP streaming.

- *Controller-based:* The buffered media time feedback has been also considered for optimizing the resource allocation in the mobile network. The majority of prior work focuses on minimizing the stalling events (e.g., [WSP$^+$12] [DSA$^+$12]). [WSP$^+$12] proposes a traffic prioritization approach at the scheduler in the mobile network that relies on playout buffer level feedback from YouTube videos. [DSA$^+$12] presents an analytical model for the multiplexing of multiple variable bit-rate videos over a time-varying wireless channel with the goal of minimizing the number of playout stalls.

  [CSK12] proposes a predictive encoder and buffer control algorithm for statistical multiplexing of video contents that considers the quality, fairness and temporal smoothness of the video streams. The approach in [CSK12] is categorically different from the work presented in this chapter as it assumes that the channel dynamics are represented by a Markov sequence and that future data rates are known a priori. This is not suitable for real-time video streaming over wireless networks where the video characteristics and the channel conditions are not known in advance.

  Recently, [WSHS12] [SOP$^+$12] proposed new metrics for optimizing the radio resource management in adaptive HTTP streaming over LTE networks. The approaches modify the resource allocation at eNodeB by considering the available DASH rates in the MPD [WSHS12] and the reported buffered time by the client [SOP$^+$12], respectively.

In this chapter, the feedback from the client on its buffered video time is considered to refine the selection of the representation rate at the proxy. As a further step, the playout buffer information of all clients is incorporated into the network resource allocation. The objective here is to jointly optimize the streaming and transmission rates by considering the instantaneous channel conditions, the content characteristics and the buffered video time of the DASH users. Compared to the proxy approach in Chapter 5, this allows a mobile operator to further tune the network resource allocation to the buffer demands of the mobile users. Compared to prior work, this one is the first that addresses the joint network and application rate allocation in the context of adaptive HTTP streaming.

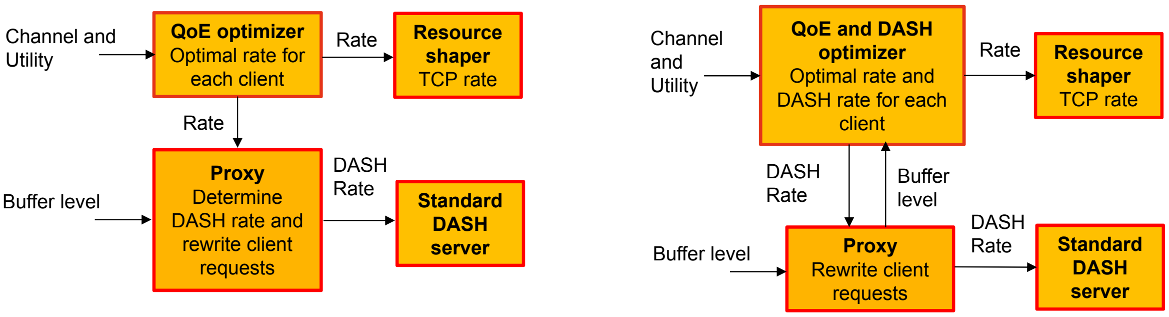
The remainder of this chapter is organised as follows. Section 6.2 introduces the playout buffer-aware QoE optimization approach. The enforcement of DASH rate adaptation and the joint streaming and transmission rate allocation approaches are described in Section 6.3 and Section 6.4, respectively. Then, the simulation results are presented in Section 6.5. Section 6.6 points out the limitations of this approach and Section 6.7 concludes the chapter.

## 6.2 Playout buffer-aware resource allocation

The QoE-based proxy adaptation scheme presented in Section 5.2 optimizes the adaptive HTTP video delivery by considering the channel and content characteristics of the streaming users. This section investigates whether the playout buffer information can be utilized to further improve the QoE. The buffer level is defined as the length of buffered media time at the client which can be either estimated at the proxy by analyzing the timestamps of the client HTTP requests or is reported by the client. Indeed, the DASH specification [ISO12] defines a buffer level metric where the client reports the playout duration for which media data is available.

In this section, two directions are considered where a) the buffer levels of the mobile clients are only available at the proxy and used to enhance the streaming rate selection of the individual clients and b) the buffer levels are additionally signaled to the QoE optimizer and considered in the multi-user resource allocation. The first case is referred to as the distributed DASH and network rate allocation, because the streaming and transmission rate problems are solved separately and the later as the joint DASH and network rate allocation, where the QoE optimizer has a global picture on the users' conditions (channel, utility and buffer levels) and jointly solves the two rate allocation problems.

In both cases, as with the proxy approach in Section 5.2, no modifications are required to a standard DASH client which can decode and play the data from a rewritten HTTP request. In the subsequent sections, $R_k$ and $Q_k$ are referred to as the transmission and representation rate for user $k$, as a result of the network and DASH rate allocation problems, respectively.



(a) Distributed DASH and network rate allocation.

(b) Joint DASH and network rate allocation.

Figure 6.2: Investigated approaches for traffic and resource management for adaptive HTTP video delivery.

## 6.3 Distributed DASH and network rate allocation

In this case, the QoE optimizer first determines the optimal rate for each client as in (5.1)-(5.2) and provides the rates to the resource shaper and the proxy (Figure 6.2(a)). The proxy then considers both the returned rate and the buffer level of each client when deciding on the representation rate. Specifically, two approaches are studied here: 1) the first one is to solve for the optimal representation that considers the throughput and buffer time of each client without any additional constraints. This represents the highest possible streaming rate given the buffer and transmission rate constraints. 2) The second approach aims at smoothing the playout video quality by considering playout buffer-aware quality bounds for selecting the representation rate of each client.

### 6.3.1 Optimal DASH rate selection

Given the available transmission capacity feedback from the QoE optimizer, the proxy solves for the highest representation for each user, given its current buffer level. The transmission capacity and buffer level are updated periodically (e.g., each 1 second). This approach is similar to the QoE-Proxy scheme and additionally allows to stream at a higher rate than the current transmission capacity if there is enough buffered media time at the client. The objective of user $k$ can be written as:

$$\underset{Q_k}{\arg\max} \quad OptQ(R_k, B_k) = U_k(Q_k) \tag{6.1}$$

$$s.t. \quad Q_k \leq R_k(1 + \frac{B_k}{T_{seg}}) \tag{6.2}$$

$$where: B_k = \text{buffer level (s)}, T_{seg} = \text{segment size (s)}$$

where $R_k$ is its transmission rate as determined by the QoE optimizer, $Q_k$ is the representation rate that maximizes the objective function in (6.1) and that satisfies the continuous playout constraint at the client (6.2). If the current buffer level at the client $B_k > 0$, then the user can stream at a representation rate which is higher than its transmission capacity and less than $R_k(1 + \frac{B_k}{T_{seg}})$, where $T_{seg}$ is the segment duration. This represents the highest possible representation that does not violate the buffer underflow constraint.

### 6.3.2 Playout buffer-dependent quality bounds

Instead of solving for the optimal DASH representation (6.1)-(6.2), the benefit of allowing clients to build-up buffer that can later compensate for dynamic channel variations is investigated. The main idea is to introduce upper and lower quality bounds which constrain on the representation quality as a function of the playout buffer time at the client (Figure 6.3). To select the target representation, both the current buffer level and the achievable QoE according to the instantaneous transmission capacity feedback from the QoE optimizer are considered. This is illustrated in Figure 6.4 which shows the available representation qualities and the actually selected representations as a function of the playout buffer time. Users who have few segments in their buffer can transmit at a lower representation rate in order to build up their buffers. Also, users with enough buffer can switch to a representation quality which is higher than the one at the instantaneous transmission rate. The overall objective is to improve the user experience by observing the buffer demands and the decision impact on the perceived user video quality.

The optimization is realized in two steps:

1) The MOS value ($MOS_{R_k}$) that corresponds to the returned rate by the QoE optimizer from (5.1)-(5.2) is determined. This represents the highest representation rate which is lower than the target rate from the QoE optimizer (cf. (6.3)).

$$MOS_{R_k} = \sup\{U_k(Q_k) : Q_k \leq R_k\} \tag{6.3}$$

2) Upper bound (UB) and lower bound (LB) thresholds are defined for finding the representation rate given the buffer level $B_k$ and the current MOS value (i.e., $MOS_{R_k}$). For simplicity, in this work, linear thresholds are considered (cf. (6.7)). An example of these boundary conditions is given in Figure 6.5. The values of $a$ and $b$ determine how fast the users will deplete and build up their buffers, respectively. The values of $T_L$ and $T_H$ represent the lower and upper QoE bounds, when the buffer level of the client is equal to zero. The DASH rate selection problem for user $k$ is:

$$\underset{Q_k}{\arg\max} \quad OptQ(R_k, B_k) = U_k(Q_k) \quad s.t. \tag{6.4}$$

$$U_k(Q_k) \leq \min\left(UB, MOS_{R_k}\right) \text{ if } MOS_{R_k} \geq UB \tag{6.5}$$

$$U_k(Q_k) \geq \max\left(LB, MOS_{R_k}\right) \text{ otherwise} \tag{6.6}$$

$$UB = T_H + B_k \cdot b, LB = T_L + B_k \cdot a \tag{6.7}$$

where (6.4) is the objective function for user $k$, (6.5) and (6.6) constrain on the desired region of representation quality.

In this work, the value of $T_L$ is set to 1 on the MOS scale. That is, for users with empty buffers, the requested representation rate should not exceed the available transmission capacity. Also, the value of $T_H$ is set to 4 on the MOS scale, so that the representation quality of the users with favorable conditions is not much degraded. For determining the $a$ and $b$ values (0.35 and 0.05, respectively), the assumption is that users with enough buffered segments (10 segments in this case) can request the highest quality representation. Please note that the upper and lower bound thresholds have been selected to control the QoE in a reasonable way. Nevertheless, other threshold values and in general more complex quality bound definitions could be considered.

## 6.4 Joint DASH and network rate allocation

The objective is to optimize the network resource allocation based on the content and channel characteristics and the client playout buffer levels. A big potential is foreseen in redistributing

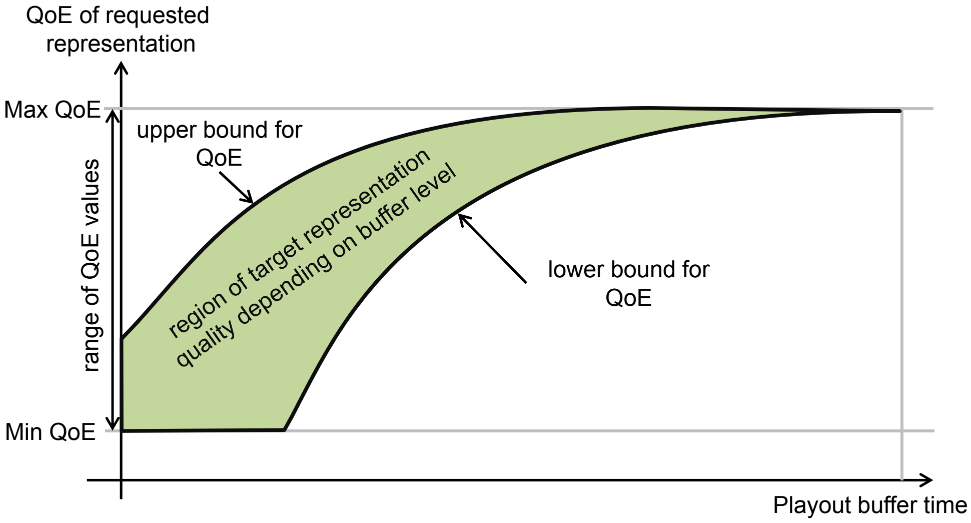

QoE of requested representation
Max QoE
range of QoE values
Min QoE
upper bound for QoE
region of target representation quality depending on buffer level
lower bound for QoE
Playout buffer time

Figure 6.3: Illustration of the playout buffer-dependent quality bounds approach for selecting the representation rate. The quality of requested representation is constrained to a target region depending on the current buffer level.

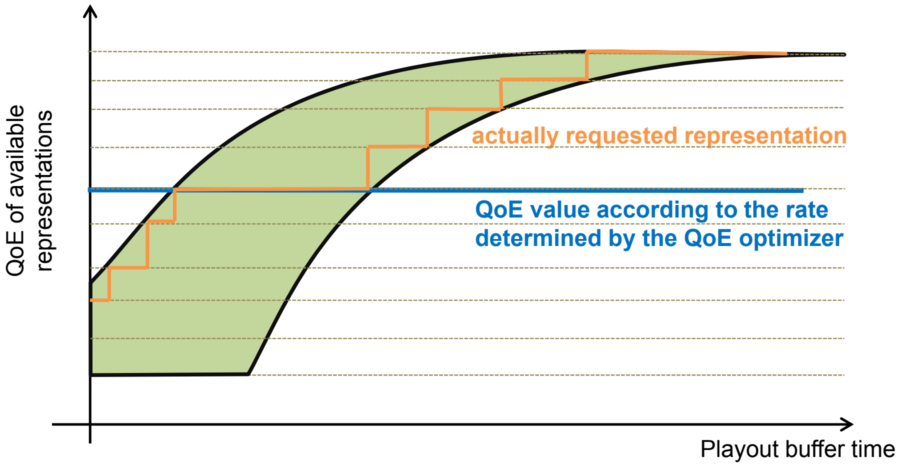

QoE of available representations
actually requested representation
QoE value according to the rate determined by the QoE optimizer
Playout buffer time

Figure 6.4: The representation is selected based on the buffer level and achievable QoE according to the instantaneous transmission capacity feedback from the QoE optimizer.

the network resources while considering the buffer information. This allows, for instance, to take some physical resources from a user without degrading the video quality of future representations if it has buffered enough segments and to assign these resources to another user with low buffer level which permits it to stream at a higher quality. Specifically, different from Sections 5.2 and 6.3, where the resource allocation and the enforcement of streaming rates were determined separately, both rate allocation problems are jointly solved now. In this case, the proxy first provides the buffer level of each client to the QoE optimizer. The optimizer then solves for the optimal transmission rate and the representation rate of each client. The resulting rates are signaled back to the resource shaper and the proxy, respectively.

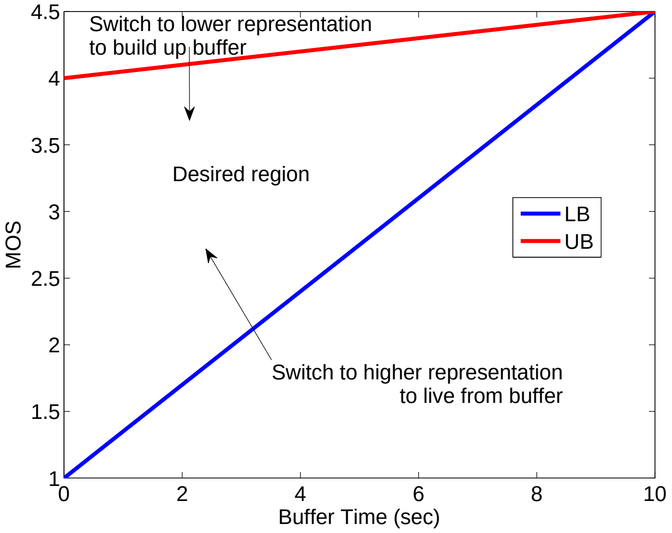


Figure 6.5: Example of buffer-aware boundary conditions for selecting the DASH representation rate. Values for ($T_L$, $a$) equal to (1, 0.35) and values of ($T_H$, $b$) equal to (4, 0.05) are considered in this work.

This is illustrated in Figure 6.2(b).

Thus, the objective of the joint optimization is:

$$\underset{(\alpha_1,\ldots,\alpha_K)}{\arg\max} \sum_{k=1}^{K} OptQ(\alpha_k, B_k) - P_k, \tag{6.8}$$

$$s.t. \quad \sum_{k=1}^{K} \alpha_k = 1 \tag{6.9}$$

where $OptQ(\alpha_k, B_k)$ is the utility value of user $k$, as the resulting distributed DASH optimization function in (6.1), respectively (6.4), for a given network resource share $\alpha_k$ ($R_k$ via (3.5)) and buffer level $B_k$.

To solve this problem, an iterative descent algorithm is considered. Let $((R_1)^0, ..., (R_K)^0)$ be the initial rate allocation vector and $((Q_1)^0, ..., (Q_K)^0)$ the corresponding DASH rates as determined in (6.1)-(6.2), respectively (6.4)-(6.6). At each iteration $m$, the algorithm searches for the users $i$ and $j$, where increasing the transmission rate $(R_i^+)^m$ and decreasing $(R_j^-)^m$ results in the maximum increase in the objective function (6.8). The corresponding representation rates $(Q_i)^m$ and $(Q_j)^m$ are updated. The above procedure is repeated until no further improvement in (6.8) is possible.

At each iteration, the objective function in (6.8) is maximized by determining the sensitivity of users $i$ and $j$ to gaining or losing a certain resource proportion, while keeping the optimization variables of the other users fixed, until convergence is achieved. The decision on the representation rate $Q_k$ of user $k$ for a given transmission rate $R_k$ is done locally independent of the other users. This holds for both distributed optimizations in (6.1)-(6.2) and (6.4)-(6.6). Hence, the optimality of the optimization approach depends on the set of achievable rates $R_k$ that are considered inside the algorithm. For arbitrary small $\alpha_k \rightarrow 0$, the algorithm can determine an optimal rate allocation.

## 6.5 Simulation results

A resource-constrained LTE cell is considered with 8 clients streaming adaptive HTTP video content from a DASH server. Each content is encoded into 11 representations by varying the encoder's quantization parameter. In the following, it is assumed that a client requests a new segment immediately after the previous segment is downloaded. Furthermore, each client provides a periodic feedback (each 1 sec) on its playout buffer level. The representation and transmission rates for each client are determined at a time scale of 1 sec as well. The simulation parameters are further illustrated in Table 6.1.

Figure 6.6 shows the individual channel traces and the mean SNR of the 8 users. In each simulation run, the requested content and the channel condition of each user are shuffled. Each content is limited to 20 segments. During the simulation, a user will periodically request a new content type after it has downloaded the previous one. A pool of 12 videos with different content characteristics is considered.

Moreover, the following schemes are compared:

- **QoE-Proxy**: This represents the buffer unaware proxy approach introduced in Section 5.2 and is used as a baseline for comparison with the buffer-aware schemes.
- **QoE-Opt**: This corresponds to the optimal DASH selection approach in Section 6.3.1.
- **QoE-QB**: This represents the playout buffer-dependent quality bounds approach in Section 6.3.2.
- **QoE-Opt-Joint**: This corresponds to the joint DASH and network rate allocation approach in Section 6.4. The utilities are determined according to the optimal DASH rate selection in (6.1).
- **QoE-QB-Joint**: This also represents the joint DASH and network rate allocation approach in Section 6.4. The utilities are determined according to the playout buffer-dependent quality bounds approach in (6.4).

Figures 6.7 and 6.8 show the mean MOS and the mean buffer level of all users for the different schemes. Results are averaged over 50 simulation runs. The QoE-Proxy scheme always selects the highest possible representation below the available throughput. Consequently, the buffer level at the client will increase over time. Please note that buffer overflow is not considered here and it is assumed that clients have enough buffer depth. The QoE-Opt scheme

Table 6.1: Simulation parameters for the buffer-aware adaptive HTTP downlink streaming scenario

| | |
|---|---|
| **Application parameters** | |
| Video codec | H.264 AVC, CIF, 30 fps |
| Application type | Adaptive HTTP streaming |
| Segment size | 1 sec |
| **LTE parameters** | |
| Carrier frequency | 2 GHz |
| System bandwidth | 5 MHz |
| Number of PRBs | 25 |
| Bandwidth per PRB | 180 KHz |
| SNR averaging cycle | 1 sec |
| Link layer model [3GP08b] | see (3.2) |
| Channel model | Urban macrocell |
| Shadowing standard deviation | 8 dB |
| Correlation distance of Shadowing | 50 m |
| **Simulation parameters** | |
| Number of users | 8 |
| Number of video sequences | 12 videos |
| Simulation runs | 50 |
| Simulation time | 300 sec |

utilizes the buffer feedback to request a higher quality representation in case the buffer is non-empty. Nevertheless, similar to the QoE-Proxy scheme, it adapts to the instantaneous rate. The QoE-QB provides a smoother mean MOS compared to the two other schemes. It builds up buffer at the client by requesting at a lower representation rate than the available throughput and then utilizes the buffer to smooth the playout curve over time. It also leads to a reasonably higher minimum mean MOS in the cell (3.08 for QoE-QB and 2.79 for QoE-Proxy and QoE-Opt schemes).

The QoE-Opt-Joint and the QoE-QB-Joint refer to the case when joint throughput and DASH rate optimization is considered. The QoE-Opt-Joint allows for a higher quality level compared to the QoE-Opt scheme but it still runs the client buffers to their limits and ends in a fluctuating mean MOS over time. The QoE-QB-Joint, on the other hand, provides the best overall video quality. The gain comes from the building/depleting of the client buffers and the time multiplexing of the representation rates and the transmission rates of the clients. For instance, between 200 and 250 sec, users build up enough buffer which is used afterwards to stream at a higher quality when the channel deteriorates. In the QoE-QB scheme, however, the transmission rate only depends on the content and the channel properties of the different clients. This means that users with good channels will get a high transmission rate, irrespective of their buffer level. As a result, the gap between the representation and transmission rates is reduced, overall less buffer is built up and the perceived quality will drop faster compared to the joint optimization scheme.

Figure 6.9 describes the mean MOS gain compared to the buffer unaware QoE-Proxy

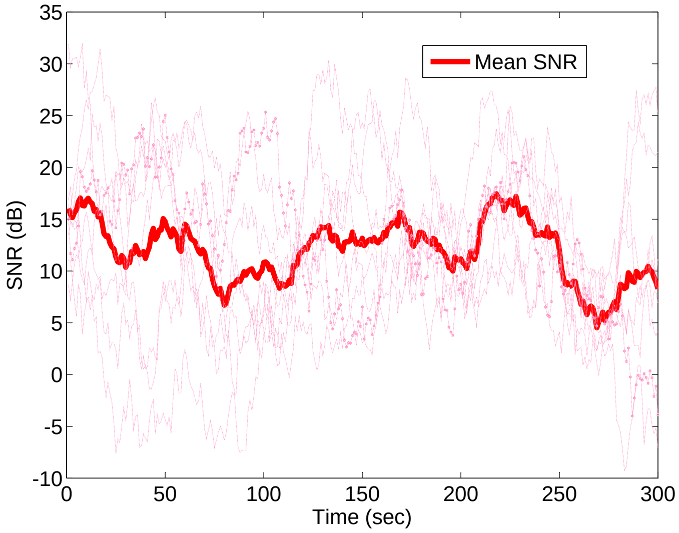


Figure 6.6: Mean SNR of all users as a function of time. SNRs of individual users are shuffled at each simulation run.

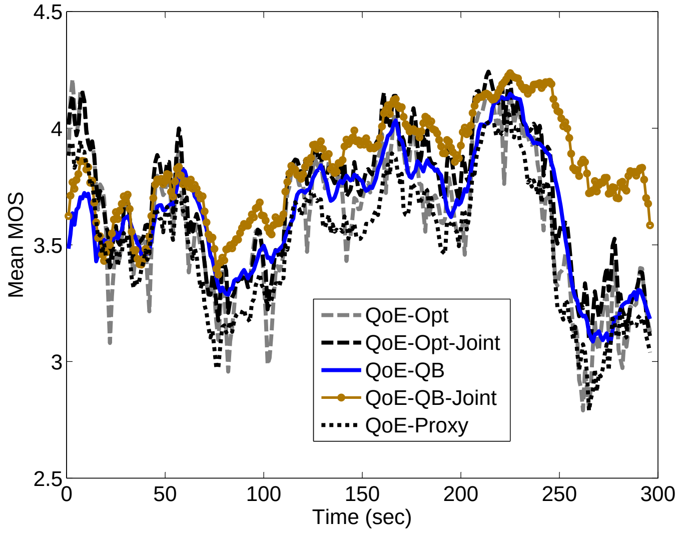


Figure 6.7: Mean MOS of 8 users as a function of simulation time averaged over 50 simulation runs.

approach. The QoE-Opt and QoE-QB, which use the same resource allocation approach as the QoE-Proxy, can result in a higher MOS by utilizing the buffer feedback when selecting the DASH representation. The joint optimization schemes can further improve the MOS level. Particularly, the proposed QoE-QB-Joint scheme results in up to 0.6 improvement on the MOS scale.

In addition, the impact of the different schemes on the perceived video quality of the individual users is analyzed. In Figure 6.10, the quality switches which exceed one representation level are considered as a measure of the temporal unsmoothness. The QoE-Proxy scheme

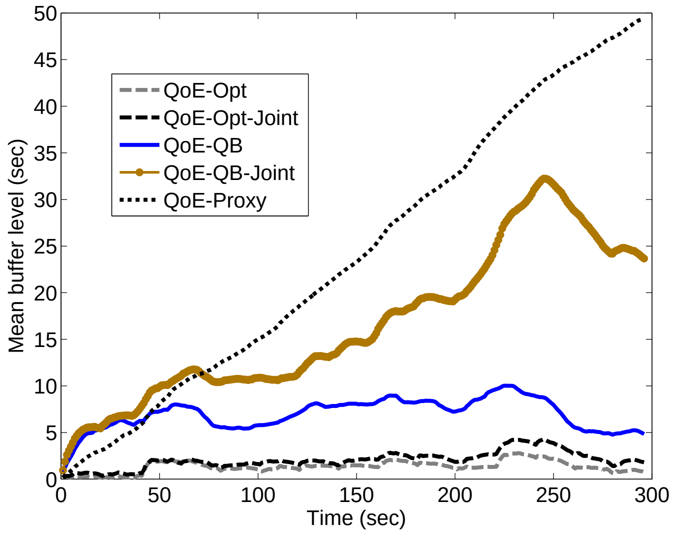


Figure 6.8: Mean buffer level of 8 users as a function of simulation time averaged over 50 simulation runs.

optimizes the resources to minimize temporal quality fluctuations between two optimization rounds and leads to very few unsmooth quality switches. Similarly, the QoE-QB scheme smoothly increases and decreases the DASH representation rate based on its boundary conditions and results in similar performance compared to the QoE-Proxy approach. The QoE-Opt scheme, however, always adapts the representation rate to the instantaneous buffer level and the available transmission rate resulting in more abrupt quality switches. Finally, in the joint optimization function, a penalty term is used to penalize quality switches across two successive rounds. Subsequently, the quality switches for the QoE-Opt-Joint and QoE-QB-Joint schemes are reduced.

## 6.6 Discussion

The buffer-aware rate allocation approach presented in this chapter requires no additional signaling between the mobile users and the proxy. In the DASH specification the clients periodically report their buffered media time, which is enough for performing the optimization. Nevertheless, there are a few challenges for the practical implementation of the proposed approach in a mobile network.

First, the joint network and rate allocation scheme assumes that the proxy and the QoE optimizer are synchronized. That is, in each optimization round, the proxy provides feedback on the buffer levels of the clients and gets a target DASH rate from the optimizer. On the other hand, in the distributed rate allocation approach, the representation rate can be determined independently by the proxy for each user based on its buffer level and the returned rate from the QoE optimizer. As a result, less signaling information is required. In both approaches,

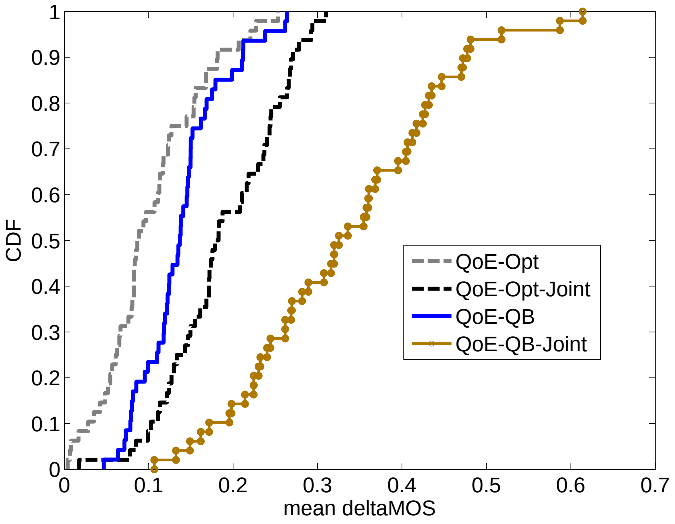


Figure 6.9: CDF of the mean MOS gain compared to QoE-Proxy scheme for 8 users. Results correspond to 50 simulation runs, 300 sec each.

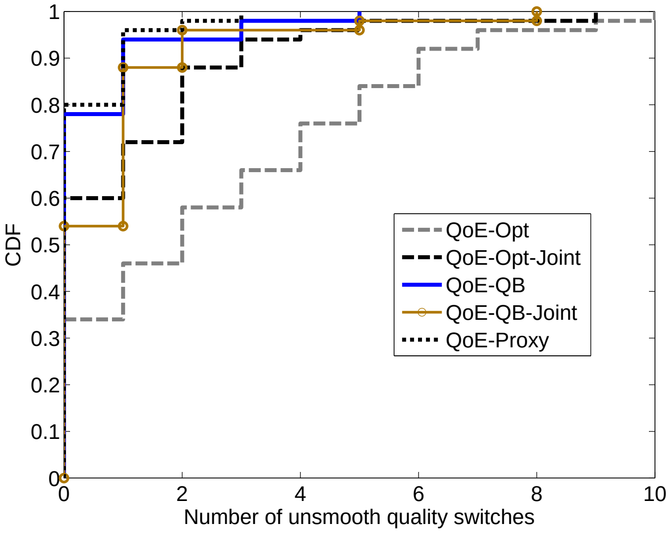


Figure 6.10: CDF of the mean number of non-smooth (jump more than 1 representation level) quality switches for each user. Simulation time is 300 sec. Results are averaged over 50 simulation runs.

however, it will be more practical to have the proxy close to the base station.

Furthermore, in the simulations, the content characteristics of the downloaded video segments by a client vary over time. The QoE optimizer then accordingly considers the utility information of the current requested segments when solving the multi-user resource allocation problem. This requires the availability of the actual MOS-bitrate profiles of each segment in the MPD.

This study presents the potential gains that can be achieved if QoE-based traffic and resource management is considered for adaptive HTTP streaming scenarios. The proposed proxy-based approach (Chapter 5) and the distributed and joint rate allocation approaches (this chapter) show different dimensions of QoE gains when the content, channel and buffer time information of the users is available to a central controller in the mobile network.

## 6.7 Chapter summary

This chapter explores the improvements that QoE-based traffic and resource management in the mobile network can offer in the context of multi-user adaptive HTTP streaming. Compared to RTP/UDP based streaming, adaptive HTTP streaming simplifies the adaptation process by providing multiple bit-rate encodings for the same content. Inspired by this a proactive QoE-based approach is proposed in Chapter 5 that rewrites the client HTTP requests at a proxy to the result of an overall network utility maximization. In this chapter, the buffered media time at the client is further incorporated into the optimization function. This allows the proxy to control the streaming rate over time to ensure a smoother playout experience.

Moreover, the main contribution of this paper is the joint optimization of the transmission and representation rates of the mobile DASH users taking into account their buffer levels. Trading off the resources among the users allows a mobile network operator to allocate higher throughput for those running at a low buffer level. At the same time, reducing the data rates of users with enough buffered media time still allows them to request high quality representations. This leads to an additional mean gain of 0.3 on the MOS scale to the 0.35 gain that the proxy approach achieves compared to standard DASH with end-to-end adaptation.

## Chapter 7

# Conclusion

The improved capabilities of mobile devices and the emergence of resource-demanding multimedia applications are fueling the demand for higher data rates and lower transport delays in future mobile networks. Despite the capacity upgrades in LTE networks, they are outpaced by an increasing number of smartphone users. Video streaming applications, in particular OTT services, will continue to dominate and grow exponentially in the upcoming years. Compared to a managed network scenario, operators will have much less control over the transported content. Hence, providing end-to-end QoS guarantees becomes more challenging. Moreover, maximizing the average user satisfaction of the mobile users is a key for mobile network operators. The strong content dependencies of multimedia services also means that traditional content agnostic optimization metrics will not provide an optimal user experience.

QoE-based resource management optimizes the network resource allocation to enhance the media experience of the mobile users. This is done by considering the user specific media characteristics and channel conditions and aims at maximizing an application objective function (e.g., average video quality in a cell). Assessing the user perceived quality in real-time is challenging. Subjective or objective measures are typically used to evaluate the user QoE. Subjective quality evaluation can not be used during online optimization. Therefore, mobile operators resort to objective measures for describing the user quality. Specifically, utility information which characterizes the user content is provided to the eNodeB along the bitstream and is considered for real-time optimization. Furthermore, the QoE-based optimization result can be exploited to reallocate the network resources among the users and also to adapt the media transport. In this thesis, different transport or traffic management approaches are investigated. While previous works have focused on transcoding or frame dropping schemes which result in additional complexity or require to inspect the media stream, respectively, one important contribution of this thesis is to simplify the rate adaptation process and provide alternatives which can be used in online system optimization at a lower cost. Specifically, for the uplink, the mobile users can control their encoding rates or decide on the packet transmission as they are the producers of their own content. Also, for the downlink, the inherent adaptivity in adaptive HTTP streaming is exploited to redirect the video streaming rate without requiring any further adaptation to the content. In the network, similar utility representations as in the previous works are used. It is also shown that generic utility functions of the video content are often enough for performing the QoE optimization in the network.

A main challenge for the success of QoE-based resource management schemes is to standardize the transport of such meta information along the media stream. The potential gains for network operators is high and incorporating these functions in the transported RTP or adaptive HTTP content can be easily done. Nevertheless, given the diversity of generated content nowadays, standardization efforts should focus on adding standard fields for utility representation which can be adopted by OTT content producers.

## 7.1 Summary of contributions

### Uplink video delivery

The uplink of next generation mobile networks has to cope with community portals for the upload and upstream of live media data. Moreover, video portal services such as YouTube and social web platforms that today serve mostly prerecorded videos will be enhanced with functionality for offering live video streams. Hence, the popularity of live video capturing and streaming with high quality and resolutions beyond today's phones is expected to challenge in particular the uplink channel of mobile data communication. Chapter 3 addresses the above issues by proposing a service-centric approach for uplink resource allocation among multiple mobile video producers. In particular, it exploits the heterogeneity in terms of the number of viewers of the uploaded video content. Optimizing the QoE-based uplink resource allocation by incorporating the popularity feedback from a video portal can substantially improve the overall QoE in a mobile cell.

Moreover, upstreamed video content to video portals can be consumed in real-time but might also be archived at the portal for on-demand retrieval. As a result, considering the type of video delivery (live or on-demand) and playout time can be used to further optimize the uplink resource allocation to provide video consumers with the best possible QoE within the scheduled playout time taking the limited and time varying uplink resources into account. Chapter 4 is concerned with the joint upstreaming of live and on-demand user-generated video content. It targets future mobile terminals that are capable of generating scalable video streams, and also assumes that these devices can cache parts of the stream for later upload. For a practical mobile environment where the available uplink resources and the channel state of each mobile terminal vary over time, the gap between the total available resources and the required resources for live video transmission is used for the upload of previously cached on-demand data. The proposed approach contributes to the state-of-the-art work on multimedia scheduling in three aspects: 1) The transmission of live and time-shifted video under scarce uplink resources is jointly optimized by transmitting a basic quality in real-time and uploading a refined quality for on-demand consumption. 2) A producer-consumer deadline-aware scheduling algorithm, that incorporates both the physical state of the mobile producer (e.g., cache fullness) and the scheduled playout time at the end-user, is proposed. 3) The scheduling decisions in 1) and 2) can be determined locally for each mobile producer.

The performance evaluation in an LTE OPNET simulator shows that the proposed QoE-based resource management scheme can notably improve the user satisfaction when compared to a standard proportional fair LTE scheduler. Indeed, the content agnostic LTE scheduler assigns the wireless resources based on the channel conditions and without considering the characteristics of the uploaded videos. This leads to non-optimal perceptual quality in particular for cell-edge users with demanding videos. In addition, the simulation results show that the mutual interaction between the video producers, the network and the target consumers can serve an enhanced user experience.

**Downlink video delivery**

The second part of thesis focuses on the downlink video delivery in fourth generation mobile networks. Particularly, it follows the pragmatic shift towards streaming over HTTP/TCP and proposes solutions for optimizing the OTT adaptive HTTP video delivery. Specifically, a proxy-based approach for redirecting the client HTTP requests according the result of an overall QoE optimization is first proposed. The proposed approach is validated by experimental results with state-of-the-art adaptive HTTP codecs and also using laboratory tests with human viewers. The main results show that standard OTT DASH leads to unsatisfactory performance since the content agnostic resource allocation by the LTE scheduler is far from optimal, and thus a clear QoE improvement can be achieved when considering the content characteristics. In addition, proactively adapting the video streaming rates of the clients gives control of the video content adaptation to the network operator which has better information than the client on the load and radio conditions in the cell. This allows faster reaction to network variations which yields additional gains in the user perceived video quality. Furthermore, the subjective results indicate a fairer resource allocation among the mobile users when QoE optimization is considered. It should be also noted that a standard unmodified DASH client remains unaware of the proposed rewriting of the HTTP requests and can decode and play the redirected media segments. Hence, the proposed approach is independent of the client control and adaptation strategies and fully complies with the DASH specifications.

In Chapter 6, a comprehensive study of the network and application rate allocation problems in the context of adaptive HTTP streaming is presented. While earlier work has studied both problems separately, this investigation shows that by jointly optimizing the streaming and transmission rates of the mobile users, a network operator has more flexibility in allocating the network resources also considering the buffer levels of the users. This leads to additional gains in user perceived video quality.

## 7.2 Future directions

With the foreseen growth in mobile video streaming, resource management in mobile networks remains a key challenge to ensure on-time high quality video distribution. Future work can build on some of the limitations of this thesis and extend in several directions:

1) This thesis focuses on resource allocation in a mobile cell where all the users are connected to one base station. With the introduction of LTE-Advanced, indoor wireless access points (APs) are supported and more users will be connected to smaller pico or femtocells in the future. Also, device-to-device (D2D) communication which has been long studied in ad hoc networks is being proposed as a potential solution for improving the real-time video communication in mobile networks. The QoE gains for D2D communication, however, are not clear yet. Given this hybrid architecture, it will be interesting to investigate distributed approaches for resource allocation that can be implemented at the AP or the mobile device, and to analyze the potential QoE gains across the multiple users. This is also challenging given that the knowledge about the other mobile users may not be locally available.

2) Multimedia streaming is going further apart from the managed server-client architecture into a distributed one, dominated by OTT and UGC content. This thesis has addressed the challenges in distributing un-managed video content over mobile networks. Nevertheless, the assumption is that some QoE knowledge about the streamed content is available. One alternative approach is to use no-reference QoE estimation models or to learn the QoE characteristics over time. The practical application of these methods to real-time resource allocation problems has been limited so far. In addition, understanding the application requirements and the user expectations will be crucial in this case. For instance, instead of simply considering real-time streaming, cloud backup and sharing to social networks as video streams, the data transmission can be prioritized based on the delay characteristics of each application. In this case, the mobile phone would be first to know once a mobile user starts the application.

3) Energy management on mobile devices and multimedia streaming have been often studied independently. Let's take adaptive HTTP streaming as an example. Actually, switching from an HD representation to an SD representation may result in unnoticeable degradation in the perceived video quality but might significantly prolong the battery life time. Appropriate real-time optimization strategies which jointly consider the application characteristics and the battery level of a mobile device are yet to be investigated. A first approach in this direction can be found in [KSES14].

# Appendix A

# Comparison of different objective mapping metrics

## A.1 Comparison with the SSIM-based QoE model

In Chapter 4, the uplink transmission of scalable video content is optimized for both live and on-demand consumption. Besides the network-based QoE optimization carried out at the base station and which is responsible for allocating the wireless resources among multiple video producers, a distributed QoE optimization is realized at each mobile terminal to determine the video layers to be transmitted in each scheduling round and their respective rates.

Throughout this thesis, a simple linear mapping between the MOS and PSNR is considered to represent the user perceived video quality. In this section, the results from Chapter 4 are additionally validated using an SSIM-based mapping model. The same simulation setup as in Section 4.5 is considered with 3 users upstreaming different video streams which are encoded with the H.264 scalable video codec [SMW07]. The simulation parameters are summarized in Table 4.2.

To calculate the SSIM values, the MSU Video Quality Measurement Tool is used [MSU]. Specifically, the average SSIM over all frames in a GoP is calculated as a measure of the video quality. A linear mapping between the SSIM and MOS, similar to [TKS11], is considered, which is based on the scatter plot of subjective and objective results for the SSIM video quality metric [WLB04]:

$$MOS = \begin{cases} 1 & \text{if } SSIM < 0.76 \\ 15.91 \cdot SSIM - 11.09 & \text{if } 0.76 \leq SSIM \leq 0.98 \\ 4.5 & \text{if } SSIM > 0.98 \end{cases} \tag{A.1}$$

Figures A.1(a) and A.1(b) show the CDF of the mean MOS for the PSNR-based and the SSIM-based mapping models, respectively. Moreover, the two QoE-based, namely the MaxMOS and its iterative solution, and the EDF distributed optimization schemes are compared. In all three cases, network-based QoE optimization is performed at the base station. The results from the SSIM-based mapping conform with the PSNR-based ones: First, both figures show similar MOS gains for the QoE-aware approaches compared to the EDF scheduling scheme, for both live and on-demand consumption cases. In addition, noticeable MOS improvements are noted for each scheme for the on-demand versus live transmission. Also, the performance of the iterative approach is very close to the optimal MaxMOS scheme for both mapping models. The average gains are further summarized in Table A.1.

The presented results show that the QoE optimization will improve the user satisfaction irrespective of the used QoE model. Nevertheless, it should be noted that the actual MOS values and the relative gains will depend on the used mapping model. It can be seen from Table A.1 that the relative gains are comparable for both mapping models, in this case.

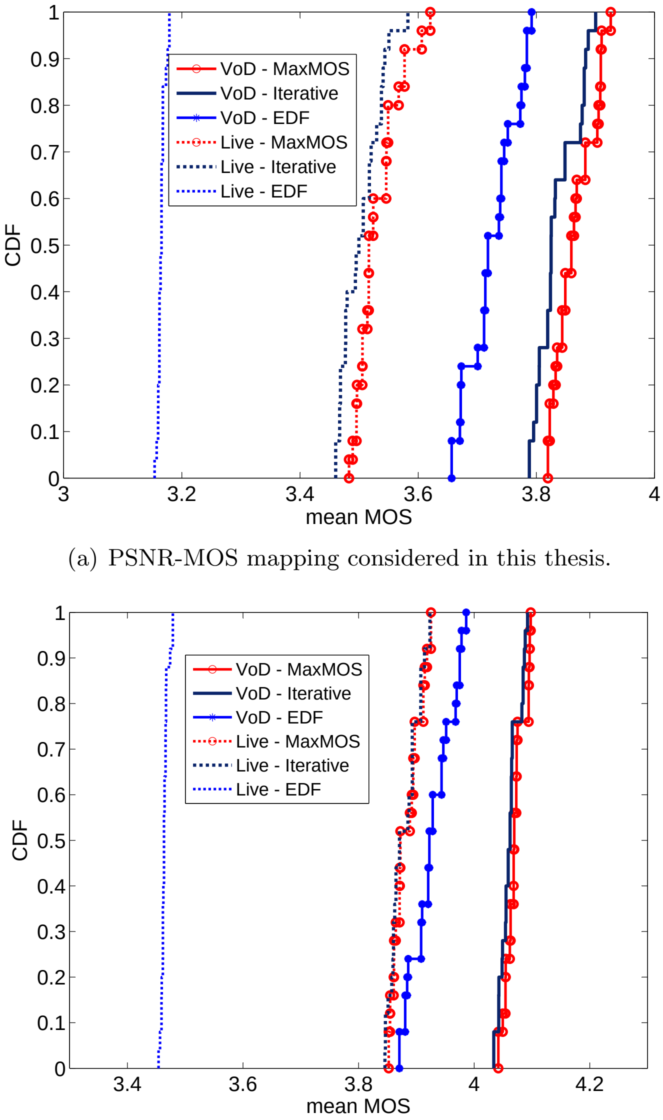
## A.2 Assessment of the prediction model

In Chapter 5, a subjective test is conducted to assess the performance of the different optimization approaches in adaptive HTTP video delivery. Specifically, the proactive (QoE-Proxy), reactive (QoE-Reactive) and standard non-optimized DASH (Non-Opt) schemes are compared.

This section first describes the individual results for the 8 different videos from the subjective experiment in Section 5.4. Then, the linear mapping used in the optimization function is validated with the subjective results and compared to other objective mappings. Furthermore, the correlation between these metrics and the subjective results is analyzed.

### Individual results

Figure A.2 presents the mean subjective rating for each video sequence as well as the standard deviation of the ratings (error bars) over the 20 viewers. Additionally, the predicted objective rating based on the linear PSNR-MOS mapping is presented. The predicted MOS values from the linear model are converted on the 100 rating scale as described in Appendix I of [BM03]. The results show a high correlation with the subjective results for all optimization schemes and in both scenarios, which indicates that the used mapping model is able to predict the tendency of the rating. Specifically, demanding videos such as *soccer*, *bus*, *coastguard* and *harbour* present a gap between the absolute predicted rating and the actual mean rating but show consistent rating behavior across the different schemes. For the the low-demanding video *ice* the predicted ratings are very close to the actual ones. Also, the *Foreman*, *akiyo* and the *container* videos present a very high correlation between the objective and subjective scores and the absolute DMOS values are very close as well.



(a) PSNR-MOS mapping considered in this thesis.

(b) SSIM-MOS mapping [TKS11].

Figure A.1: Comparison of the MaxMOS scheme and its corresponding low-complexity iterative approach with the EDF scheme for different QoE mapping models.

### Comparison with other objective metrics

Furthermore, the applicability of the linear PSNR/MOS model used for the optimization problem is assessed. Therefore, a post-analysis is conducted to measure the correlation between the subjective results and different objective quality metrics. Specifically, the linear mapping is compared with two non-linear metrics based on the PSNR, namely the PSNR based Video Quality Metric (VQM$_P$) proposed in [WP02] and the STVQM proposed in [PS11]. The Pearson correlation between the subjective ratings and the predicted ratings for the 8 videos

Table A.1: Average gains on the MOS scale

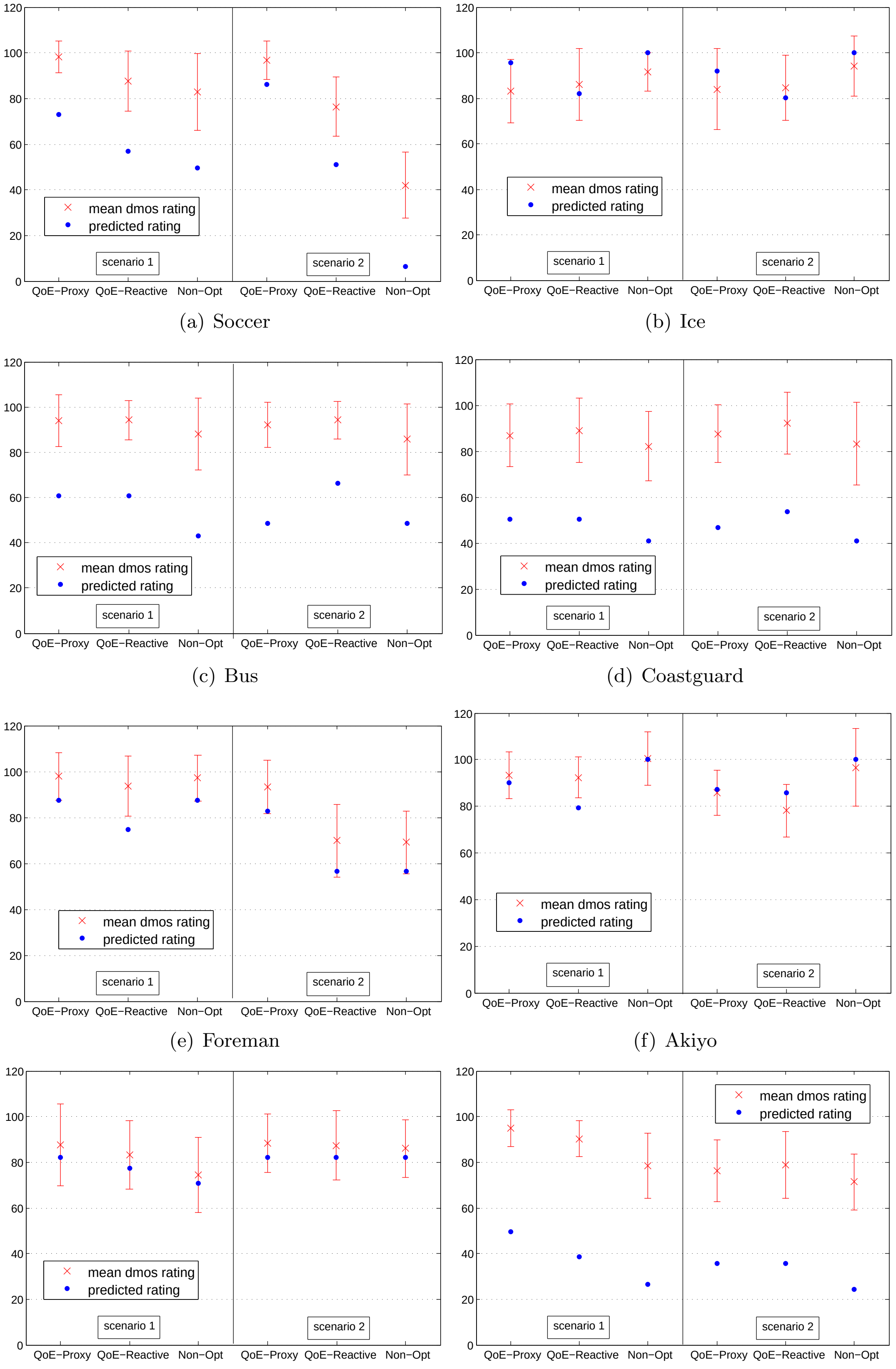
| | PSNR-MOS model | SSIM-MOS model |
|---|---|---|
| Live (EDF → QoE) | 0.36 | 0.4 |
| VoD (EDF → QoE) | 0.11 | 0.14 |
| QoE (Live → VoD) | 0.33 | 0.2 |
| EDF (Live → VoD) | 0.56 | 0.46 |

Table A.2: Pearson correlation for each video

| Metric | Linear | VQM$_P$ [WP02] | STVQM [PS11] |
|---|---|---|---|
| soccer | 0.9618 | 0.9866 | 0.9697 |
| ice | 0.4762 | 0.8537 | 0.7792 |
| bus | 0.8241 | 0.8241 | 0.8241 |
| coastguard | 0.9489 | 0.9466 | 0.9412 |
| foreman | 0.9787 | 0.9976 | 0.9988 |
| akiyo | 0.9115 | 0.9077 | 0.9401 |
| container | 0.9826 | 0.9872 | 0.9884 |
| harbour | 0.8959 | 0.9262 | 0.9172 |
| **mean** | **0.8725** | **0.9287** | **0.9198** |

is presented in Table A.2.

The main difference between the metrics is observed for the *ice* video, where the linear model achieves a low Pearson correlation. This is due mostly to the low resource-demanding characteristic of this video, which in our scenario leads to a very high perceived quality for all cases. The predicted ratings all being in a small range, the Pearson correlation is more sensitive to a variation of the subjective ratings. However, the mean Pearson correlation over the 8 videos is 0.8725 for the linear model. This is close to the highest mean Pearson correlation achieved by the VQM$_P$ (0.9287). This shows that although the used linear PSNR/MOS model is very simple, it performs almost as well as a more complex non-linear video quality metric in the case of an optimization over multiple videos.



(a) Soccer

(b) Ice

(c) Bus

(d) Coastguard

(e) Foreman

(f) Akiyo

(g) Container

(h) Harbour

Figure A.2: Individual ratings for each video. The figure shows the mean and standard deviation of the subjective DMOS values from the 20 participants. The dots represent the predicted objective DMOS values as a result of the QoE optimization in (5.1)-(5.2).

# Appendix B

# Optimality and convergence analysis

In Chapter 4, an analytical framework for scalable video transmission at a mobile terminal is presented. This section provides a detailed analysis on the optimality and convergence properties of the distributed scheme in (4.1)-(4.4). A good reference on dual methods for solving optimization problems is Chapter 6 of [Ber99].

## B.1 Optimality analysis

The optimality analysis is composed of three steps. First, the optimization problem in (4.1)-(4.4) is transformed to a Lagrange dual problem. Then, it is shown that the dual problem can be solved by an iteration method. Finally, a possible way to guarantee that the proposed solution converges to the optimal value is provided.

*Step 1:* Substitute $R_l$ by $\frac{L_l}{t_l}$ in (4.1)-(4.4) and solve for the optimal values of $a_l^*$, $t_l^*$ and $L_l^*$. From the definition of the objective function, it can be seen that (4.1) is a classical convex optimization problem satisfying Slater's condition [HBCR07]. Specifically, the slater condition states that strong duality holds if the inequality constraints in (4.2)-(4.4) are strictly feasible, i.e., they hold with strict inequalities [BV04]. In this case, there exists an allocation policy ($a_l^*$, $t_l^*$ and $L_l^*$), where

$$\sum_{S_{L,d}} \sum_{l \in S_{L,d}} a_l^* \frac{L_l^*}{t_l^*} < c_k, \tag{B.1}$$

$$\sum_{S_{L,d}} \sum_{l \in S_{L,d}} L_l^*(1 - a_l^*) < H_k, \tag{B.2}$$

$$\sum_{l \in S_{L,d}} a_l^* t_l^* < d, \forall S_{L,d}, \tag{B.3}$$

As a consequence of strong duality, the optimal duality gap is zero [BV04]. In other words, the optimal solution from the primal problem in (4.1)-(4.4) is equal to the best lower bound that can be obtained from a Lagrange dual function. As a next step, a Lagrange dual problem to achieve the optimal value is described.

*Step 2:* The Lagrange dual problem can be designed by augmenting the constraints from (4.2)-(4.4) into the objective function:

$$\begin{gathered} L\left(a_l, L_l, t_l, \phi_l, \theta_l, \delta_l\right) = \sum_{S_{L,d}} \sum_{l \in S_{L,d}} a_l b_l \Delta MOS_l \left(\frac{L_l}{t_l}\right) \\ - \sum_{S_{L,d}} \sum_{l \in S_{L,d}} \phi_l\left(\tau\right)\left(a_l L_l - c_k t_l\right) - \\ \sum_{S_{L,d}} \sum_{l \in S_{L,d}} \theta_l\left(\tau\right)\left[L_l\left(1 - a_l\right) - H_k\right] - \sum_{l \in S_{L,d}} \delta_l\left(\tau\right)\left[a_l t_l - d\right], \end{gathered} \tag{B.4}$$

where, $\phi_l$, $\theta_l$, and $\delta_l$ are the Lagrange multipliers associated with the constraints of (4.2), (4.3) and (4.4), and $\tau$ indicates the time index ($\tau \in \mathbb{N}$).

In order to simplify the optimization problem, the Lagrangian dual problem in (B.4) can be further decomposed into three sub-problems [Ber99]:

$$\max : \sum_{S_{L,d}} \sum_{l \in S_{L,d}} \phi_l\left(\tau\right) c_k t_l. \tag{B.5}$$

$$\max : -\sum_{S_{L,d}} \sum_{l \in S_{L,d}} \theta_l\left(\tau\right) L_l. \tag{B.6}$$

$$\begin{gathered} \max : \sum_{S_{L,d}} \sum_{l \in S_{L,d}} a_l b_l \Delta MOS_l \left(\frac{L_l}{t_l}\right) - \sum_{S_{L,d}} \sum_{l \in S_{L,d}} \phi_l\left(\tau\right) a_l L_l \\ + \sum_{S_{L,d}} \sum_{l \in S_{L,d}} \theta_l\left(\tau\right) a_l L_l - \sum_{l \in S_{L,d}} \delta_l\left(\tau\right) a_l t_l. \end{gathered} \tag{B.7}$$

Decomposition methods are used to divide an optimization problem into sub-problems which can be solved iteratively or sequentially. In this case, the original problem from (B.4) is solved by iteratively solving the sub-problems in (B.5)-(B.7). Specifically, at each iteration, the two sub-problems in (B.5)-(B.6) can be solved independently to find the values of the variables $t_l$ and $L_l$, respectively. Furthermore, sub-problem (B.7) can be solved to find the value of the variable $a_l$ for a fixed $t_l$ and $L_l$.

In particular, the Lagrange dual function $L_d\left(\phi, \theta, \delta\right)$ is defined as the maximum of the Lagrangian $L\left(a_l, L_l, t_l, \phi_l, \theta_l, \delta_l\right)$ over $a_l$, $L_l$ and $t_l$ for given $\phi$, $\theta$, and $\delta$. Hence, the Lagrange dual problem can be formulated by:

$$L_d\left(\phi_l, \theta_l, \delta_l\right) = L\left(\begin{array}{c} a_l^*\left(\phi_l, \theta_l, \delta_l\right), L_l^*\left(\phi_l, \theta_l, \delta_l\right), \\ t_l^*\left(\phi_l, \theta_l, \delta_l\right), \phi_l, \theta_l, \delta_l \end{array}\right), \tag{B.8}$$

where $\left(\phi_l, \theta_l, \delta_l\right)$ are the dual variables, and $a_l^*\left(\phi, \theta, \delta\right)$, $L_l^*\left(\phi, \theta, \delta\right)$, and $t_l^*\left(\phi, \theta, \delta\right)$ correspond to the values $a_l$, $t_l$ and $L_l$ when achieving the optimal solution. As explained in Step 1, since the duality gap is equal to zero, the Lagrange dual function returns the optimal solution for the optimization problem in (4.1)-(4.4).

*Step 3:* Since $L_d$ may be non-differentiable, an iterative sub-gradient method can be used to solve it by introducing the variables $\rho_\phi$, $\rho_\theta$, and $\rho_\delta$. Please note that $L_d$ is a convex optimization problem since the objective to be maximized is concave and can be solved by a conventional KKT method [Chi05]. For each dual variable, the sub-gradient method generates a sequence of feasible points which are updated at each iteration $\tau$. Specifically, from (B.5):

$$\phi_l\left(\tau+1\right) = \left[\phi_l\left(\tau\right) + \rho_\phi\left(\tau\right) g_\phi\left(\tau\right)\right]^+ = \left[\phi_l\left(\tau\right) + \rho_\phi\left(\tau\right) \sum_{S_{L,d}} \sum_{l \in S_{L,d}} c_k t_l\right]^+, \tag{B.9}$$

where $g_\phi\left(\tau\right)$ is the sub-gradient of (B.5) at $\phi_l\left(\tau\right)$, and $\rho_\phi\left(\tau\right)$ represents the iteration step size.

Similarly, from (B.6):

$$\theta_l\left(\tau+1\right) = \left[\theta_l\left(\tau\right) + \rho_\theta\left(\tau\right) g_\theta\left(\tau\right)\right]^+ = \left[\theta_l\left(\tau\right) - \rho_\theta\left(\tau\right) \sum_{S_{L,d}} L_l\right]^+. \tag{B.10}$$

where $g_\theta\left(\tau\right)$ is the sub-gradient of (B.6) at $\theta_l\left(\tau\right)$, and $\rho_\theta\left(\tau\right)$ represents the iteration step size.

In addition, $\frac{\partial(B.7)}{\partial a_l} = 0$ at the optimal solution,

$$\begin{aligned} &\sum_{S_{L,d}} \sum_{l \in S_{L,d}} b_l \Delta MOS_l\left(\tfrac{L_l}{t_l}\right) - \sum_{S_{L,d}} \sum_{l \in S_{L,d}} \phi_l\left(\tau\right) L_l + \\ &\sum_{S_{L,d}} \sum_{l \in S_{L,d}} \theta_l\left(\tau\right) L_l - \sum_{l \in S_{L,d}} \delta_l\left(\tau\right) t_l = 0. \end{aligned} \tag{B.11}$$

Hence, $\delta_l\left(\tau\right)$ can be determined as a function of the iteration step size $\rho_\delta\left(\tau\right)$:

$$\begin{aligned} &\delta_l\left(\tau+1\right) = \\ &\left[\delta_l\left(\tau\right) + \rho_\delta\left(\tau\right) \sum_{S_{L,d}} \sum_{l \in S_{L,d}} \begin{pmatrix} b_l \Delta MOS_l\left(\frac{L_l}{t_l}\right) \\ -c_k t_l + L_l - t_l \end{pmatrix}\right]^+ \end{aligned} \tag{B.12}$$

$\rho_\phi\left(\tau\right)$, $\rho_\theta\left(\tau\right)$ and $\rho_\delta\left(\tau\right)$ represent the iteration step size. Certain choices of step sizes, such as $\rho_\phi\left(\tau\right) = \frac{\rho_1}{\tau}$, $\rho_\theta\left(\tau\right) = \frac{\rho_2}{\tau}$, and $\rho_\delta\left(\tau\right) = \frac{\rho_3}{\tau}$, where $\rho_1, \rho_2, \rho_3 > 0$ guarantee that this algorithm will converge to the result obtained via joint optimization. An exemplary illustration of the choices of the iteration step sizes for determining the upper bound of the convergence step is provided in the next section.

## B.2 Convergence bound

The relationship between the convergence steps and $\rho_1, \rho_2, \rho_3$ is studied by analyzing their dependencies when reaching the optimal solution. Specifically,

1) When $\frac{\partial (B.5)}{\partial t_l} = 0$:

$$\sum_{S_{L,d}} \sum_{l \in S_{L,d}} \phi_l\left(\tau\right) c_k = 0. \tag{B.13}$$

2) When $\frac{\partial (B.6)}{\partial L_l} = 0$:

$$\sum_{S_{L,d}} \sum_{l \in S_{L,d}} \theta_l\left(\tau\right) = 0. \tag{B.14}$$

3) When $\frac{\partial (B.7)}{\partial t_l} = 0$:

$$\sum_{S_{L,d}} \sum_{l \in S_{L,d}} -a_l b_l \Delta MOS\left(\frac{L_l}{t_l}\right) L_l \frac{1}{t_l^2} - \sum_{l \in S_{L,d}} \delta_l\left(\tau\right) a_l = 0. \tag{B.15}$$

More precisely, from (B.13) and (B.14), $\phi_l\left(\tau\right) = 0$ and $\theta_l\left(\tau\right) = 0$ at the optimal solution, respectively. Combing with (B.9) and (B.10), the convergence bound is independent of the parameters $\rho_1$ and $\rho_2$.

Furthermore, by replacing $\rho_\delta\left(\tau\right) = \frac{\rho_3}{\tau}$ and including $\delta_l\left(\tau\right)$ from (B.12) into (B.15):

$$\begin{gathered} \sum_{S_{L,d}} \sum_{l \in S_{L,d}} -a_l b_l \Delta MOS\left(\tfrac{L_l}{t_l}\right) \tfrac{L_l}{t_l^2} - \\ \sum_{l \in S_{L,d}} \left[ \begin{array}{c} \delta_l\left(\tau-1\right) + \frac{\rho_3}{\tau} \sum\limits_{S_{L,d}} \sum\limits_{l \in S_{L,d}} \\ \left(b_l \Delta MOS\left(\tfrac{L_l}{t_l}\right) - c_k t_l + L_l - t_l\right) \end{array} \right]^+ a_l = 0. \end{gathered} \tag{B.16}$$

Moreover, when optimality is achieved, $\delta_l\left(\tau\right) = \delta_l\left(\tau-1\right) = C$, where $C$ is a constant. So,

$$\sum_{S_{L,d}} \sum_{l \in S_{L,d}} -a_l b_l \Delta MOS\left(\tfrac{L_l}{t_l}\right) \tfrac{L_l}{t_l^2} - \ \rho_3 \sum_{S_{L,d}} \sum_{l \in S_{L,d}} \left(b_l \Delta MOS\left(\tfrac{L_l}{t_l}\right) - c_k t_l + L_l - t_l\right) \ = C \tag{B.17}$$

$$\Longrightarrow \sum_{S_{L,d}} \sum_{l \in S_{L,d}} -a_l b_l \Delta MOS\left(\tfrac{L_l}{t_l}\right) \tfrac{L_l}{t_l^2} - C = \ \rho_3 \sum_{S_{L,d}} \sum_{l \in S_{L,d}} \left(b_l \Delta MOS\left(\tfrac{L_l}{t_l}\right) - c_k t_l + L_l - t_l\right) \tag{B.18}$$

$$\Longrightarrow \rho_3 \propto \frac{\sum\limits_{S_{L,d}} \sum\limits_{l \in S_{L,d}} -a_l b_l \Delta MOS\left(\frac{L_l}{t_l}\right) \frac{L_l}{t_l^2}}{\sum\limits_{S_{L,d}} \sum\limits_{l \in S_{L,d}} \left(b_l \Delta MOS\left(\frac{L_l}{t_l}\right) - c_k t_l + L_l - t_l\right)} \tag{B.19}$$

Furthermore, the MOS improvement is bounded and $b_l \Delta MOS\left(\frac{L_l}{t_l}\right) \leq 4.5$. Therefore, the upper bound of the convergence step is $O\left(\frac{\sum\limits_{S_{L,d}} \sum\limits_{l \in S_{L,d}} a_l \frac{L_l}{t_l^2}}{\sum\limits_{S_{L,d}} \sum\limits_{l \in S_{L,d}} \left(c_k t_l - L_l\right)}\right)$.

# Bibliography

## Publications by the author

[ESM$^+$11] A. El Essaili, E. Steinbach, D. Munaretto, S. Thakolsri, and W. Kellerer. QoE-driven resource optimization for user generated video content in next generation mobile networks. In *Proc. IEEE International Conference on Image Processing*, Brussels, Belgium, September 2011. [cited at p. 6]

[ESS$^+$13] A. El Essaili, D. Schroeder, D. Staehle, M. Shehada, W. Kellerer, and E. Steinbach. Quality-of-Experience driven adaptive HTTP media delivery. In *Proc. IEEE International Conference on Communications*, Budapest, Hungary, June 2013. [cited at p. 6]

[EZS$^+$11] A. El Essaili, L. Zhou, D. Schroeder, E. Steinbach, and W. Kellerer. QoE-driven live and on-demand LTE uplink video transmission. In *Proc. IEEE International Workshop on Multimedia Signal Processing*, Hangzhou, China, October 2011. [cited at p. 6]

[KMT$^+$13] W. Kellerer, D. Munaretto, S. Thakolsri, E. Steinbach, and A. El Essaili. Method and evaluation server for evaluating a plurality of videos. European Patent EP 2479684 B1, US Patent US 2012192242 A1, September 2013. [cited at p. 6]

[KSES14] S. Khan, D. Schroeder, A. El Essaili, and E. Steinbach. Energy-efficient and QoE-driven adaptive HTTP streaming over LTE. In *Proc. IEEE Wireless Communications and Networking Conference*, Istanbul, Turkey, April 2014. [cited at p. 98]

[SES$^+$12] D. Schroeder, A. El Essaili, E. Steinbach, Z. Despotovic, and W. Kellerer. A quality-of-experience driven bidding game for uplink video transmission in next generation mobile networks. In *Proc. IEEE International Conference on Image Processing*, Orlando, Florida, USA, September 2012. [cited at p. 23]

[SES$^+$13] 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 *Proc. International Packet Video Workshop*, San Jose, CA, USA, December 2013. [cited at p. 13]

[SSK$^+$14] M. Shehada, D. Staehle, W. Kellerer, E. Steinbach, A. El Essaili, and D. Schroeder. Method, system and network for transmitting multimedia data to a plurality of clients. European Patent EP 2696552 A1, US Patent US 2014047071 A1, February 2014. [cited at p. 6]

## General publications

[3GP06] 3GPP TR 25.814 V.7.1.0. Physical layer aspects for Evolved Universal Terrestrial Radio Access (E-UTRA). September 2006. [cited at p. 21, 23, 27]

[3GP08a] 3GPP 36.913 v8.0.0. Requirements for further advancements for evolved universal terrestrial radio access e-utra (lte-advanced). June 2008. [cited at p. 7]

[3GP08b] 3GPP TR 36.942 V.8.1.0. Evolved Universal Terrestrial Radio Access (E-UTRA); Radio Frequency (RF) system scenarios. December 2008. [cited at p. vi, 28, 33, 49, 68, 90]

[3GP10a] 3GPP TR 36.814 V.9.0.0. Evolved Universal Terrestrial Radio Access (E-UTRA); Further advancements for E-UTRA physical layer aspects. March 2010. [cited at p. vi, 28, 29, 33, 49]

[3GP10b] 3GPP TS 36.300 v10.0.0. Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN): Overall Description. June 2010. [cited at p. 7]

[3GP12] 3GPP TS 23.203 v9.12.0. Technical specification group services and system aspects; policy and charging control architecture. June 2012. [cited at p. 9]

[3GP13] 3GPP TS 26.247 V11.2.0. Technical specification group services and system aspects; transparent end-to-end packet-switched streaming service (PSS); progressive download and dynamic adaptive streaming over HTTP (3GP-DASH). March 2013. [cited at p. 18]

[AABD12] S. Akhshabi, L. Anantakrishnan, A. C. Begen, and C. Dovrolis. What happens when HTTP adaptive streaming players compete for bandwidth? In *Proceedings NOSSDAV '12, Toronto, Canada*, June 2012. [cited at p. vii, 60, 61]

[AADB13] S. Akhshabi, L. Anantakrishnan, C. Dovrolis, and A. C. Begen. Server-based traffic shaping for stabilizing oscillating adaptive streaming players. In *Proceeding NOSSDAV '13, Oslo, Norway*, February 2013. [cited at p. 61]

[ABD11] S. Akhshabi, A. C. Begen, and C. Dovrolis. An experimental evaluation of rate-adaptation algorithms in adaptive streaming over HTTP. *Proc. MMSys 2011, California, USA*, February 2011. [cited at p. 6, 18, 60, 66]

[ADF$^+$09] D. Astely, E. Dahlman, A. Furuskar, A. Kangas, M. Lindstrom, and S. Parkvall. LTE: the evolution of mobile broadband. *IEEE Communications Magazine*, 47(4):44–51, April 2009. [cited at p. 7, 8]

[AKK$^+$10] G. Aristomenopoulos, T. Kastrinogiannis, V. Kaldanis, G. Karantonis, and Papavassiliou. A novel framework for dynamic utility-based QoE provisioning in wireless networks. *Proc. IEEE Globecom, Miami, USA*, 2010. [cited at p. 23]

[AMO$^+$10] P. Ameigeiras, J. Ramos Munoz, J. Navarro Ortiz, P. Mogensen, and J.M. Lopez Soler. QoE oriented cross-layer design of a resource allocation algorithm in beyond 3G systems. *Computer Communications*, 33(5):571–582, March 2010. [cited at p. 14]

[AVC12] Advanced video coding for generic audiovisual services. *International Telecommunication Union Std. H.264, Rev. 16*, January 2012. [cited at p. 10]

[BAB11a] A. Begen, T. Akgul, and M. Baugher. Watching video over the web: Part 1: Streaming protocols. *IEEE Internet Computing*, March-April 2011. [cited at p. 17]

[BAB11b] A. Begen, T. Akgul, and M. Baugher. Watching video over the web: Part 2: Applications, standardization, and open issues. *IEEE Internet Computing*, 15(3):59–63, May-June 2011. [cited at p. 20, 61]

[BC98] N. Bjork and C. Christopoulos. Transcoder architectures for video coding. *IEEE Transactions on Consumer Electronics*, 44(1):88–98, February 1998. [cited at p. 11]

[Ber71] T. Berger. *Rate-Distortion Theory.* Englewood Cliffs, NJ: Prentice Hall, 1971. [cited at p. 9]

[Ber99] D. P. Bertsekas. *Nonlinear Programming.* Athena Scientific, 1999. [cited at p. 105, 106]

[BM03] J. A. Bergstra and C. A. Middelburg. ITU-T Recommendation G.107 : The E-Model, a computational model for use in transmission planning. Technical report, 2003. [cited at p. 26, 100]

[BSK13] S. Barakovi and L. Skorin-Kapov. Survey and challenges of QoE management issues in wireless networks. *Journal of Computer Networks and Communications, Journal of Computer Networks and Communications, Article ID 165146*, 2013. [cited at p. 14, 23]

[BSW09] S. Borst, I. Saniee, and A. Walid. An analytical model for provisioning of emerging personalized content distribution services. *Proc. ITC'09, Paris, France*, September 2009. [cited at p. 26]

[BV04] S. Boyd and L. Vandenberghe. Convex Optimization. *Cambridge University Press*, 2004. [cited at p. 46, 51, 105, 106]

[BY04] R. Berry and E Yeh. Cross-layer wireless resource allocation. *IEEE Signal Processing Magazine*, 21(5):59–68, September 2004. [cited at p. 21]

[CCLC11] Jen-Yeu Chen, Chia-Wen Chiu, Gwo-Long Li, and Mei-Juan Chen. Burst-aware dynamic rate control for H.264/AVC video streaming. *IEEE Transactions on Broadcasting*, 57(1):89 – 93, March 2011. [cited at p. 11]

[CF06] J. Chakareski and P. Frossard. Rate-distortion optimized distributed packet scheduling of multiple video streams over shared communication resources. *IEEE Transactions on Multimedia*, 8(2):1011–1020, April 2006. [cited at p. 21, 43]

[CF07] J. Chakareski and P. Frossard. Adaptive systems for improved media streaming experience. *IEEE Communications Magazine*, January 2007. [cited at p. 11]

[Chi05] Mung Chiang. Balancing transport and physical layers in wireless multihop networks: jointly optimal congestion control and power control. *IEEE Journal on Selected Areas in Communications*, 23(1):104–116, January 2005. [cited at p. 107]

[Cis12] Cisco white paper. Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2011-2016. Technical report, 2012. [cited at p. 1]

[CISN05] L. Choi, M.T. Ivrlac, E. Steinbach, and J.A. Nossek. Sequence-level methods for distortion-rate behavior of compressed video. *Proc. IEEE ICIP'05, Genova, Italy*, September 2005. [cited at p. vii, 9, 10, 27, 33, 53, 63, 64]

[CKR+09] M. Cha, H. Kwak, P. Rodriguez, Y. Ahn, and S. Moon. Analyzing the video popularity characteristics of large-scale user generated content systems. *IEEE/ACM Transactions on Networking*, 17(5):1357–1370, October 2009. [cited at p. 25, 32]

[CLCD07] Mung Chiang, S.H. Low, A.R. Calderbank, and J.C. Doyle. Layering as optimization decomposition: A mathematical theory of network architectures. *Proceedings of the IEEE*, 95(1):255–312, January 2007. [cited at p. 22]

[CM06] P.A. Chou and Z. Miao. Rate-distortion optimized streaming of packetized media. *IEEE Transactions on Multimedia*, 8(2):390– 404, April 2006. [cited at p. 21, 43, 47, 48]

[CPG+13] F. Capozzi, G. Piro, L.A. Grieco, G. Boggia, and P. Camarda. Downlink packet scheduling in LTE cellular networks: Key design issues and a survey. *IEEE Communications Surveys Tutorials*, 15(2):678–700, Second Quarter 2013. [cited at p. 2]

[CSK12] N. Changuel, B. Sayadi, and M. Kieffer. Predictive encoder and buffer control for statistical multiplexing of multimedia contents. *IEEE Transactions on Broadcasting*, 58(3):401–416, September 2012. [cited at p. 83]

[CSRK11] S. Chikkerur, V. Sundaram, M. Reisslein, and L.J. Karam. Objective video quality assessment methods: A classification, review, and performance comparison. *IEEE Transactions on Broadcasting*, 57(2):165–182, June 2011. [cited at p. 13]

[DBHL11] D.C. Dimitrova, J.L. Berg, G. Heijenk, and R. Litjens. LTE uplink scheduling - flow level analysis. pages 181–192. Springer Berlin Heidelberg, 2011. [cited at p. 42]

[DCBA10] A. Dua, C.W. Chan, N. Bambos, and J. Apostolopoulos. Channel, deadline, and distortion (CD2) aware scheduling for video streams over wireless. *IEEE Transactions on Wireless Communications*, 9(3):1001–1011, March 2010. [cited at p. vi, 42, 43]

[DPSB08] Erik Dahlman, Stefan Parkvall, Johan Skold, and Per Beming. *3G Evolution, Second Edition: HSPA and LTE for Mobile Broadband.* Academic Press, 2 edition, 2008. [cited at p. 8]

[DSA+12] P. Dutta, A. Seetharam, V. Arya, M. Chetlur, S. Kalyanaraman, and J. Kurose. On managing quality of experience of multiple video streams in wireless networks. In *Proceedings IEEE INFOCOM, Orlando, USA*, March 2012. [cited at p. 83]

[Eks09] H. Ekstrom. QoS control in the 3GPP evolved packet system. *IEEE Communications Magazine*, 47(2):76–83, Februray 2009. [cited at p. 9]

[Ell12] J. Ellenbeck. IMTAphy source code hosted on launchpad.net. *[Online]. Available: http://launchpad.net/imtaphy*, January 2012. [cited at p. vi, 29]

[FGM+99] R. Fielding, J. Gettys, J. Mogul, H. Frystyk, L. Masinter, P. Leach, and T. B. Lee. Hypertext Transfer Protocol – HTTP/1.1. RFC 2616, June 1999. [cited at p. 16, 17, 18]

[FHK+06] P. Frojdh, U. Horn, M. Kampmann, A. Nohlgren, and M. Westerlund. Adaptive streaming within the 3GPP packet-switched streaming service. *IEEE Network*, 20(2):34–40, 2006. [cited at p. 11, 16]

[FHPW00] Sally Floyd, Mark Handley, Jitendra Padhye, and Jörg Widmer. Equation-based congestion control for unicast applications. In *Proc. SIGCOMM 2000*, August 2000. [cited at p. 16]

[FHPW08] S. Floyd, M. Handley, J. Padhye, and J. Widmer. TCP Friendly Rate Control (TFRC): Protocol Specification. RFC 5348, September 2008. [cited at p. 16]

[FKR09] M. Fiedler, K. Kilkki, and P. Reichl. From quality of service to quality of experience. *Executive Summary, Seminar 09192, Dagstuhl Seminar Proceedings, available at URL: http://drops.dagstuhl.de/opus/volltexte/2009/2235/pdf/09192.SWM.Paper.2235.pdf.*, 2009. [cited at p. 13]

[FMM+13] B. Fu, D. Munaretto, T. Melia, B. Sayadi, and W. Kellerer. Analyzing the combination of different approaches for video transport optimization for next generation cellular networks. *IEEE Network Magazine, Special Issue on Video over Mobile Networks*, 27(2), March/April 2013. [cited at p. 23]

[GALM07] P. Gill, M. Arlitt, Z. Li, and A. Mahanti. YouTube traffic characterization: a view from the edge. *Proc. ACM IMC'07, San Diego, USA*, pages 15–28, October 2007. [cited at p. 25]

[GC02] Y. Guo and H. Chaskar. Class-based quality of service over air interfaces in 4G mobile networks. *IEEE Communications Magazine*, 40(3):132–137, March 2002. [cited at p. 26]

[GGP97] L. Georgiadis, R. Guerin, and A. Parekh. Optimal multiplexing on a single link: delay and buffer requirements. *IEEE Transactions on Information Theory*, 43(5):1518–1535, September 1997. [cited at p. 42]

[Gir93] Bernd Girod. What's wrong with mean-squared error? *Visual Factors of Electronic Image Communications, Cambridge, MA: MIT Press*, 1993. [cited at p. 10]

[GKLP10] F. Gabin, M. Kampmann, T. Lohmar, and C. Priddle. 3GPP mobile multimedia streaming standards. *IEEE Signal Processing Magazine*, 27(6):134–138, November 2010. [cited at p. 1, 16, 17, 20, 21, 40]

[GLGP13b] Gerardo Gomez, Javier Lorca, Raquel Garcia, and Quiliano Perez. Towards a QoE-driven resource control in LTE and LTE-A networks. *Journal of Computer Networks and Communications*, Article ID 505910, January 2013. [cited at p. 14, 23]

[Glo13] Global mobile Suppliers Association (GSA). LTE is mainstream. Technical report, May 2013. [cited at p. 7]

[GNT06] L. Georgiadis, M. J. Neely, and L. Tassiulas. Resource allocation and cross-layer control in wireless networks. *Foundations and Trends in Networking*, 1(1):1–144, 2006. [cited at p. 22]

[Goy01] V.K. Goyal. Multiple description coding: compression meets the network. *IEEE Signal Processing Magazine*, 18(5):74–93, September 2001. [cited at p. 11, 21]

[h26] H.264/AVC JM reference software, online: http://iphome.hhi.de/suehring/tml/. [cited at p. v, 27]

[Hal00] Fred Halsall. *Multimedia Communications: Applications, Networks, Protocols, and Standards*. Addison-Wesley, 2000. [cited at p. 12]

[HBCR07] J. He, M. Bresler, M. Chiang, and J. Rexford. Towards robust multi-layer traffic engineering: Optimization of congestion control and routing. *IEEE Journal on Selected Areas in Communications*, 25(5):868–880, June 2007. [cited at p. 105]

[HG12] Remi Houdaille and Stephane Gouache. Shaping HTTP adaptive streams for a better user experience. *Proc. MMSys 2012, North Carolina, USA*, February 2012. [cited at p. 62]

[HOK99] Chi-Yuan Hsu, A. Ortega, and M. Khansari. Rate control for robust video transmission over burst-error wireless channels. *IEEE Journal on Selected Areas in Communications*, 17(5):756–773, May 1999. [cited at p. 11]

[HSX05] R. Hamzaoui, V. Stankovic, and Z. Xiong. Optimized error protection of scalable image bit streams [advances in joint source-channel coding for images]. *IEEE Signal Processing Magazine*, 22(6):91–107, November 2005. [cited at p. 11]

[HT09] H. Holma and A. Toskala. *LTE for UMTS - OFDMA and SC-FDMA Based Radio Access*. Wiley, 2009. [cited at p. 8]

[IN04] M. Ivrlac and J. Nossek. Cross layer design - an equivalence class approach. *Proc. IEEE ISSSE '04, Linz, Austria*, August 2004. [cited at p. 22, 30]

[IR98] ITU-R. ITU-R BT.500-9, Methodology for the subjective assessment of the quality of television pictures. 1998. [cited at p. 13]

[IR07] ITU-R. ITU-R BT.1788, Methodology for the subjective assessment of video quality in multimedia applications. 2007. [cited at p. viii, 13, 72, 73]

[ISO12] ISO/IEC IS 23009-1. Information technology- Dynamic adaptive streaming over HTTP (DASH)- Part I: Media presentation description and segment formats. Technical report, April 2012. [cited at p. 18, 63, 84]

[ISO13] ISO/IEC DIS 23009-2. Information technology – Dynamic adaptive streaming over HTTP (DASH) – Part 2: Conformance and reference software. Technical report, March 2013. [cited at p. 18]

[IT94] ITU-T. Rec. E.800 Terms and definitions related to quality of service and network performance including dependability. 1994. [cited at p. 13]

[IT99] ITU-T. ITU-T P.910, Subjective video quality assessment methods for multimedia applications. 1999. [cited at p. 13, 72]

[IT07] ITU-T. Rec. P.10 Vocabulary for performance and quality of service. 2007. [cited at p. 13]

[ITU96] ITU. *Methods for subjective determination of transmission quality (ITU-T Recommendation P.800).* International Telecommunication Union, 1996. [cited at p. 26]

[Jac88] V. Jacobson. Congestion avoidance and control. *ACM Computer Communication Review*, 18:314–329, 1988. [cited at p. 16]

[JF07] D. Jurca and P. Frossard. Media flow rate allocation in multipath networks. *IEEE Trans. on Multimedia*, 9(6):1227 –1240, October 2007. [cited at p. 31, 65]

[JHC$^+$09] X. Ji, J. Huang, M. Chiang, G. Lafruit, and F. Catthoor. Scheduling and resource allocation for SVC streaming over OFDM downlink systems. *IEEE Transactions on Circuits and Systems for Video Technology*, 19(10):1549–1555, October 2009. [cited at p. 43, 44]

[JPP00] A. Jalali, R. Padovani, and R. Pankaj. Data throughput of CDMA-HDR a high efficiency-high data rate personal communication wireless system. *Proc. IEEE VTC 2000-Spring, Tokyo, Japan*, May 2000. [cited at p. 8]

[JSZ12] Junchen Jiang, Vyas Sekar, and Hui Zhang. Improving fairness, efficiency, and stability in HTTP-based adaptive video streaming with FESTIVE. In *Proceedings CoNEXT '12, Nice, France*, December 2012. [cited at p. 61]

[KAS12] B. Krogfoss, A. Agrawal, and L. Sofman. Analytical method for objective scoring of HTTP Adaptive Streaming (HAS). In *IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB), Soeul, Korea*, June 2012. [cited at p. 82]

[KDSK07] S. Khan, S. Duhovnikov, E. Steinbach, and W. Kellerer. MOS-based multiuser multiapplication cross-layer optimization for mobile multimedia communication. *Advances in Multimedia, article ID 94918*, January 2007. [cited at p. v, 2, 14, 15, 23, 48]

[Kel97] F. P. Kelly. Charging and rate control for elastic traffic. *European Transactionson Telecommunications*, 8:33–37, January-February 1997. [cited at p. 21]

[KL07] W. Kuo and W. Liao. Utility-based resource allocation in wireless networks. *IEEE Transactions on Wireless Communications*, 6(10):3600 –3606, October 2007. [cited at p. 23]

[KPS$^+$06] S. Khan, Y. Peng, E. Steinbach, M. Sgroi, and W. Kellerer. Application-driven cross-layer optimization for video streaming over wireless networks. *IEEE Communications Magazine*, 44(1):122–130, January 2006. [cited at p. v, 2, 22, 23]

[KSG04] M. Kalman, E. Steinbach, and B. Girod. Adaptive media playout for low-delay video streaming over error-prone channels. *IEEE Transactions on Circuits and Systems for Video Technology*, 14(6):841–851, June 2004. [cited at p. 11]

[LBG11b] C. Liu, I. Bouazizi, and M. Gabbouj. Rate adaptation for adaptive HTTP streaming. In *Proceedings of the second annual ACM conference on Multimedia systems*, MMSys '11, February 2011. [cited at p. 74]

[LCW$^+$10] Haiyan Luo, Song Ci, Dalei Wu, Jianjun Wu, and Hui Tang. Quality-driven cross-layer optimized video delivery over LTE. *IEEE Communications Magazine*, 48(2):102–109, Februray 2010. [cited at p. 22]

[LG05] Zhijun Lei and Nicolas D. Georganas. Adaptive video transcoding and streaming over wireless channels. *Journal of Systems and Software*, 75(3):253–270, March 2005. [cited at p. 62]

[LZG$^+$13] Zhi Li, Xiaoqing Zhu, Josh Gahm, Rong Pan, Hao Hu, Ali C. Begen, and Dave Oran. Probe and adapt: Rate adaptation for HTTP video streaming at scale. *CoRR*, abs/1305.0510, July 2013. [cited at p. vii, 61]

[MB11] K.J. Ma and R. Bartos. HTTP live streaming bandwidth management using intelligent segment selection. *Proc. IEEE Globecom 2011, Texas, USA*, December 2011. [cited at p. 62]

[MBBN11] K. Ma, R. Bartos, S. Bhatia, and R. Nair. Mobile video delivery with HTTP. *IEEE Communications Magazine*, 49(4):166–175, April 2011. [cited at p. 61]

[MCLC11] Ricky K.P. Mok, Edmond W.W. Chan, Xiapu Luo, and Rocky K.C. Chang. Inferring the QoE of HTTP video streaming from user-viewing activities. In *Proceedings ACM SIGCOMM W-MUST, Toronto, Ontario, Canada*, August 2011. [cited at p. 82]

[MDWE02] P. Marziliano, F. Dufaux, S. Winkler, and T. Ebrahimi. A no-reference perceptual blur metric. In *Proc. ICIP'02, Rochester, New York, USA*, September 2002. [cited at p. 13]

[MLCC12] Ricky K. P. Mok, Xiapu Luo, Edmond W. W. Chan, and Rocky K. C. Chang. QDASH: a QoE-aware DASH system. In *Proceedings MMSys '12, Chapel Hill, North Carolina, USA*, February 2012. [cited at p. 82]

[MLT12] Christopher Mueller, Stefan Lederer, and Christian Timmerer. An evaluation of dynamic adaptive streaming over HTTP in vehicular environments. In *Proc. ACM MoVid '12, Chapel Hill, NC, USA*, February 2012. [cited at p. 82]

[MMLB$^+$07] M.G. Martini, M. Mazzotti, C. Lamy-Bergot, J. Huusko, and P. Amon. Content adaptive network aware joint optimization of wireless video transmission. *IEEE Communications Magazine*, 45(1):84–90, January 2007. [cited at p. 21]

[Moe10] S. Moeller. *Quality Engineering: Qualität Kommunikationstechnischer System.* Springer London, Limited, 2010. [cited at p. 13]

[MSU] MSU Video Quality Measurement Tool. Graphics and Moscow State University Media Lab, CMC department, http://compression.graphicon.ru/video/. [cited at p. 99]

[MT11] C. Mueller and C. Timmerer. A VLC media player plugin enabling dynamic adaptive streaming over HTTP. *Proc. ACM Multimedia 2011, Arizona, USA*, November 2011. [cited at p. 66, 68, 69, 71, 77, 79]

[New05] M. E. J. Newman. Power laws, Pareto distributions and Zipfs law. *Contemporary Physics*, 46:323, 2005. [cited at p. 30]

[OFYjSC10] O. Oyman, J. Foerster, T. Yong-joo, and L. Seong-Choon. Toward enhanced mobile video services over WiMAX and LTE. *IEEE Communications Magazine*, 48(8):68–76, August 2010. [cited at p. 2]

[OHC11] Tao-Sheng Ou, Yi-Hsin Huang, and H.H. Chen. SSIM-based perceptual rate control for video coding. *IEEE Transactions on Circuits and Systems for Video Technology*, 21(5):682–691, May 2011. [cited at p. 11]

[OS12] O. Oyman and S. Singh. Quality of experience for HTTP adaptive streaming services. *IEEE Comm. Magazine*, 50(4):20–27, April 2012. [cited at p. 82]

[PGB$^+$11] G. Piro, L.A. Grieco, G. Boggia, R. Fortuna, and P. Camarda. Two-level downlink scheduling for real-time multimedia services in LTE networks. *IEEE Transactions on Multimedia*, 13(5):1052–1065, May 2011. [cited at p. vi, 44]

[Pos80] J. Postel. User datagram protocol. RFC 768, August 1980. [cited at p. 16]

[PS11] Y. Peng and E. Steinbach. A novel full-reference video quality metric and its application to wireless video transmission. In *IEEE International Conference on Image Processing (ICIP)*, Brussels, Belgium, September 2011. [cited at p. 101, 102]

[PW04] M.H. Pinson and S. Wolf. A new standardized method for objectively measuring video quality. *IEEE Transactions on Broadcasting*, 50(3):312–322, September 2004. [cited at p. 13]

[PZC12] W. Pu, Z. Zou, and C. W. Chen. Video adaptation proxy for wireless dynamic adaptive streaming over HTTP. *Proc. Packet Video Workshop 2012, Munich, Germany*, May 2012. [cited at p. 62]

[RCFW10] W. Rao, L. Chen, A. Fu, and G. Wang. Optimal resource placement in structured peer-to-peer networks. *IEEE TPDS*, 21(7):1011–1026, July 2010. [cited at p. 26]

[RdTBFM08] L. Ruiz de Temino, G. Berardinelli, S. Frattasi, and P. Mogensen. Channel-aware scheduling algorithms for SC-FDMA in LTE uplink. *Proc. IEEE PIMRC '08, Cannes, France*, September 2008. [cited at p. 22]

[RFB01] K. Ramakrishnan, S. Floyd, and D. Black. The addition of explicit congestion notifcation (ECN) to IP. *RFC 3168, IETF*, September 2001. [cited at p. 17]

[Riz97] L. Rizzo. Dummynet: a simple approach to the evaluation of network protocols. *Proc. ACM SIGCOMM Computer Communication Review*, January 1997. [cited at p. 66]

[RTS10] P. Reichl, B. Tuffin, and R. Schatz. Economics of logarithmic quality-of-experience in communication networks. In *Conference on Telecommunications Internet and Media Techno Economics (CTTE), Vienna, Austria*, June 2010. [cited at p. 14]

[San13] Sandvine. Global internet phenomena report. Technical report, 2013. [cited at p. 1, 2]

[Sau08] A. Saul. Wireless resource allocation with perceived quality fairness. *Proc. IEEE ACSSC'08, CA, USA*, November 2008. [cited at p. 30, 32]

[SCFJ03] H. Schulzrinne, S. Casner, R. Frederick, and V. Jacobson. RTP: A transport protocol for real-time applications. RFC 3550, July 2003. [cited at p. 16]

[SFLG00] Klaus Stuhlmüller, Niko Färber, Michael Link, and Bernd Girod. Analysis of video transmission over lossy channels. *IEEE Journal on Selected Areas in Communications*, 18:1012–1032, June 2000. [cited at p. 9]

[She95] Scott Shenker. Fundamental design issues for the future Internet. *IEEE Journal on Selected Areas in Communications*, September 1995. [cited at p. 14]

[SHL03] G. Song, Y. Han, and Y. Lee. Adaptive subcarrier and power allocation in OFDM based on maximizing utility. *IEEE Vehicular Technology Conference, VTC2003-Spring, Orlando, USA*, April 2003. [cited at p. 22]

[SJK04] T. Stockhammer, H. Jenkac, and G. Kuhn. Streaming video over variable bit-rate wireless channels. *IEEE Transactions on Multimedia*, 6(2):268–277, April 2004. [cited at p. 11, 82]

[SKA$^+$07] A. Saul, S. Khan, G. Auer, W. Kellerer, and E. Steinbach. Cross-layer optimization using model-based parameter exchange. *Proc. IEEE ICC'07, Glasgow, Scotland*, June 2007. [cited at p. 30]

[SMW07] H. Schwarz, D. Marpe, and T. Wiegand. Overview of the scalable video coding extension of the H.264/AVC standard. *IEEE Transactions on Circuits and Systems for Video Technology*, 17(9):1103–1120, September 2007. [cited at p. 10, 11, 21, 40, 44, 48, 52, 99]

[Sod11] I. Sodagar. The MPEG-DASH standard for multimedia streaming over the internet. *IEEE MultiMedia*, 18(4):62–67, April 2011. [cited at p. 19]

[SOP$^+$12] S. Singh, O. Oyman, A. Papathanassiou, D. Chatterjee, and J.G. Andrews. Video capacity and QoE enhancements over LTE. In *IEEE International Conference on Communications (ICC), Ottawa, Canada*, June 2012. [cited at p. 83]

[SRL98] H. Schulzrinne, A. Rao, and R. Lanphier. Real time streaming protocol (RTSP). RFC 2326, April 1998. [cited at p. 16]

[SRT99] S. Sen, J. Rexford, and D. Towsley. Proxy prefix caching for multimedia streams. *Proc. IEEE INFOCOM'99, NY, USA*, March 1999. [cited at p. 26]

[SS02] S. Shakkottai and R. Srikant. Scheduling real-time traffic with deadlines over a wireless channel. *ACM Wireless Networking*, 8(1):13–26, 2002. [cited at p. 42]

[SSBC10] K. Seshadrinathan, R. Soundararajan, A.C. Bovik, and L.K. Cormack. Study of subjective and objective quality assessment of video. *IEEE Transactions on Image Processing*, 19(6):1427–1441, June 2010. [cited at p. 13]

[SSW07] T. Schierl, T. Stockhammer, and T. Wiegand. Mobile video transmission using scalable video coding. *IEEE Transactions on Circuits and Systems for Video Technology*, 17(9):1204–1217, September 2007. [cited at p. 21]

[Sto11] Thomas Stockhammer. Dynamic adaptive streaming over HTTP - standards and design principles. *Proc. MMSys 2011, California, USA*, February 2011. [cited at p. v, 2, 17, 19]

[TDG11] TDG research. Adaptive Bitrate Technology: Driving Video to Three-Screens. Technical report, 2011. [cited at p. 2]

[THKP12] T.C. Thang, Q. Ho, J.W. Kang, and A.T. Pham. Adaptive streaming of audiovisual content using MPEG DASH. *IEEE Transactions on Consumer Electronics*, 58(1):78–85, February 2012. [cited at p. 61, 64]

[TKS04] Wei Tu, Wolfgang Kellerer, and Eckehard Steinbach. Rate-distortion optimized video frame dropping on active network nodes. In *Packet Video Workshop 2004*, Irvine, California, December 2004. [cited at p. 21]

[TKS10] S. Thakolsri, W. Kellerer, and E. Steinbach. QoE-based rate adaptation scheme selection for resource-constrained wireless video transmission. In *Proceedings of the International Conference on Multimedia*, MM '10, October 2010. [cited at p. 23]

[TKS11] S. Thakolsri, W. Kellerer, and E. Steinbach. QoE-based cross-layer optimization of wireless video with unperceivable temporal video quality fluctuation. *Proc. IEEE ICC'11, Kyoto, Japan*, June 2011. [cited at p. 64, 65, 99, 101]

[TKSK09] S. Thakolsri, S. Khan, E. Steinbach, and W. Kellerer. QoE-driven cross-layer optimization for high speed downlink packet access. *Journal of Communications*, 4(9):669–680, October 2009. [cited at p. v, xi, 2, 3, 5, 14, 21, 23, 24, 26, 29, 31, 62, 66]

[TL12] G. Tian and Y. Liu. Towards agile and smooth video adaptation in dynamic HTTP streaming. In *Proceedings of the 8th international conference on Emerging networking experiments and technologies, CoNEXT '12, Nice, France*, Dec. 2012. [cited at p. 82]

[vC07] M. van der Schaar and P. Chou. *Multimedia over IP and Wireless Networks: Compression, Networking, and Systems.* Academic Press, 2007. [cited at p. 11]

[Vid00] Video Quality Experts Group (VQEG). Final report from the video quality experts group on the validation of objective models of video quality assessment. *Technical report, ITU*, March 2000. [cited at p. 26]

[vS05] M. van der Schaar and N. Sai Shankar. Cross-layer wireless multimedia transmission: challenges, principles, and new paradigms. *IEEE Wireless Communications*, 12(4):50–58, August 2005. [cited at p. 21, 22]

[vT07] M. van der Schaar and D.S. Turaga. Cross-layer packetization and retransmission strategies for delay-sensitive wireless multimedia transmission. *IEEE Transactions on Multimedia*, 9(1):185–197, January 2007. [cited at p. 43]

[WBSS04] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, April 2004. [cited at p. 10, 13]

[WKST08] 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*, 4(2):16:1–16:22, May 2008. [cited at p. 17, 59]

[WLB04] Z. Wang, L. Lu, and A. C. Bovik. Video quality assessment based on structural distortion measurement. *Signal Processing: Image Communication, special issue on Objective Video Quality Metrics*, 19(2):121–132, February 2004. [cited at p. 99]

[WP02] S. Wolf and M. Pinson. Video quality measurement techniques. Technical report, NTIA Technical Report TR-02-392, June 2002. [cited at p. 101, 102]

[WR05] H. R. Wu and K. R. Rao. *Digital Video Image Quality and Perceptual Coding (Signal Processing and Communications).* CRC Press, Inc., Boca Raton, FL, USA, 2005. [cited at p. 9]

[WSB03] Zhou Wang, Hamid R. Sheikh, and Alan C. Bovik. Objective video quality assessment. In *The Handbook of Video Databases: Design and Applications*, pages 1041–1078. CRC Press, 2003. [cited at p. 13]

[WSHS12] Thomas Wirth, Yago Sánchez, Bernd Holfeld, and Thomas Schierl. Advanced downlink LTE radio resource management for HTTP-streaming. In *Proceedings ACM MM '12, Nara, Japan*, October 2012. [cited at p. 62, 83]

[WSJ$^+$03] T. Wiegand, H. Schwarz, A. Joch, F. Kossentini, and G. J. Sullivan. Rate-constrained coder control and comparison of video coding standards. *IEEE Transactions on Circuits and Systems for Video Technology*, 13(7):688–703, July 2003. [cited at p. 20]

[WSP+12] F. Wamser, D. Staehle, J. Prokopec, A. Maeder, and P. Tran-Gia. Utilizing buffered YouTube playtime for QoE-oriented scheduling in OFDMA networks. In *International Teletraffic Congress (ITC 24), Krakow, Poland*, September 2012. [cited at p. 83]

[WY97] H.R. Wu and M. Yuen. A generalized block-edge impairment metric for video coding. *IEEE Signal Processing Letters*, 4(11):317–320, November 1997. [cited at p. 13]

[WZ98] Yao Wang and Qin-Fan Zhu. Error control and concealment for video communication: a review. *Proceedings of the IEEE*, 86(5):974–997, May 1998. [cited at p. 12]

[XDBC07] Fang Xie, Lei Du, Yong Bai, and Lan Chen. Popularity aware scheduling for network coding based content distribution in ad hoc networks. *Proc. IEEE PIMRC'07, Athens, Greece*, September 2007. [cited at p. 26]

[ZXD+13] Shaobo Zhang, Yangpo Xu, Peiyun Di, Alex Giladi, Changquan Ai, and Xin Wang. Quality driven streaming using MPEG-DASH. *IEEE Communications Letters*, 8(2):34 –38, March 2013. [cited at p. 82]