Technische Universität München
Institute for Media Technology
Prof. Dr.-Ing. Eckehard Steinbach

PhD Thesis

Quality of Experience based Network Resource Allocation for Wireless Multimedia Delivery

Srisakul Thakolsri

TECHNISCHE UNIVERSITÄT MÜNCHEN
Lehrstuhl für Medientechnik

Quality of Experience based Network Resource Allocation for Wireless Multimedia Delivery

Srisakul Thakolsri
M.Sc.

Vollständiger Abdruck der von der Fakultät 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. Assoc. Prof. Maria G. Martini, Ph.D., Kingston University London/Großbritannien

Die Dissertation wurde am 24.10.2011 bei der Technischen Universität München eingereicht und durch die Fakultät für Elektrotechnik und Informationstechnik am 25.04.2012 angenommen.

Acknowledgement

This dissertation was written during my time as researcher in the Networking Research Group at DOCOMO Communications Laboratories Europe GmbH and as a PhD candidate at the Institute for Media Technology at Technische Universität München (TUM). I am really grateful to both institutes to give me the opportunity to do an academic and industrial research work. The successful outcome of this work could not be accomplished without the helps and supports from many people in various ways. I do apologize if I do not explicitly mention all of them here.

First, I would like to express my sincere appreciation to my PhD advisor, Prof. Dr.-Ing. Eckehard Steinbach, for accepting me as one of his PhD candidate at his research institute at TUM. I am grateful to Prof. Steinbach for his suggestions, guidances and many helpful and stimulating discussions during the various phases of the dissertation. In particular, I thank Prof. Steinbach for providing me all facilities for running a user test at the TUM. During the period of his supervision, I learned a lot from him on how to proceed step-by-step till the end of my PhD thesis and his attitudes of writing a good conference/journal paper.

I would also like to give a special thank to my Director at DOCOMO Euro-Labs, Dr.-Ing. Wolfgang Kellerer, for his valuable comments, fruitful discussions and motivation during the whole period of this work. Especially, I am really appreciated for his constant support on providing any facilities that I need in order to get results for this PhD thesis.

I am greatly indebted to all my DOCOMO Euro-Labs colleagues and all PhD candidates/staff at Prof. Steinbach's institute for sharing a good time and providing a friendly and pleasant atmosphere during this work. Specifically, I would like to thank ex-colleagues from DOCOMO Euro-Labs (Dr. Zoran Despotovic, Dr. Dan Jurca, Daniele Munaretto, Dr. Gerald Kunzmann), ex-colleagues from TUM (Shoaib Khan, Yang Peng), and all my students (Anas Al-Nuaimi, Shahzad Rasool, Serdar Cokbulan, Mohammed Shehada and Wang Zibin), who had contributed to the success of this thesis. For a smooth administrative preparation of submitting this thesis, I also thank Dr. Martin Maier.

Finally, I would not have passed all difficulties while carrying out this work, if there is no support from my private life at home. Hence, I would like to sincerely thank my dear parents (Papa and Mama), my siblings (Chok, Vit and Pin) and my lovely wife (Valli) for their encouragement and their continuous support.

Munich, May 2012 *Srisakul Thakolsri*

Abstract

This dissertation discusses a Quality of Experience (QoE) driven cross-layer optimization framework for efficient network resource allocation in the High Speed Downlink Packet Access (HSDPA) system. The proposed scheme jointly optimizes the application layer and the lower layers of the wireless protocol stack to determine the application data rate and the network resources with the aim of improving the service quality perceived by the user. The Mean Opinion Score (MOS) is used as a unified utility metric that encompasses the user-perceived quality under certain receiving conditions for the user application. Various QoE-based optimization schemes taking into account different criteria are proposed and compared to a throughput maximization scheme and a non-optimized system. Results show that the QoE-based approaches lead to significant improvements of user perceived quality.

Zusammenfassung

In der vorliegenden Arbeit wird ein Konzept für die effiziente Zuteilung von Ressourcen in Mobilfunksystemen beschrieben, das den Einfluss der empfundenen Qualität von Diensten (Quality of Experience, QoE) berücksichtigt. Das vorgeschlagene Verfahren optimiert gemeinsam die Anwendungsschicht und die unteren Protokollschichten (Physical layer, Link layer), um die Datenrate der Anwendung und die bestmögliche Verteilung der Ressourcen zu ermitteln. Der vorgeschlagene Ansatz verwendet den Mean Opinion Score (MOS) als einheitliche Metrik für die Erfassung der Nutzerzufriedenheit. Verschiedene QoE-basierte Optimierungsverfahren werden vorgestellt und mit der Maximierung des Gesamtdurchsatzes und nicht-optimierten Systemen verglichen. Die Ergebnisse zeigen, dass die QoE-basierten Verfahren zu deutlichen Verbesserungen der Systemperformanz und insbesondere der Servicequalität führen.

Contents

Chapter 1

Introduction

As Internet multimedia communication moves rapidly into the wireless commercial realm, it has evolved and changed the way we work and live. End users of the Internet expect no longer to be exclusively connected to the Internet through a computer at home or in the office, but also through their handheld devices, and thus enabling a seamless multimedia communication across devices. For instance, a corporate employee is able to check his emails on a mobile device whilst travelling, or a consumer is able to watch his favourite movies on his mobile device whilst on holidays. Though allowing a user to enjoy a multimedia service from any place, anytime and on any devices regardless of the access network gives flexibility and convenience, it comes at a cost. In particular, this is a challenge for the wireless network operators, as wireless network resources are normally scarcer than wired resources, and become hence more expensive than wired ones.

This introductory chapter highlights multimedia communication in mobile networks, and provides motivation for research in the topic of mobile multimedia communication by addressing the problems of mobile multimedia delivery and associated challenges. Subsequently, we discuss the scope and contributions of this thesis. Finally, a road map of the following chapters is provided.

1.1 Mobile multimedia communication

In recent years, the demand for multimedia communication in mobile networks has increased due to the following factors. First, mobile network infrastructures are quickly evolving to support Internet Protocol (IP) and other transport types required by Internet multimedia communication. Second, mobile devices are improving with enhanced capabilities such as larger display size, long-life battery, high computational power and touch-screen feature. Some of them are even equipped with Global Positioning System (GPS) and camera functionality. The latter allows users to create their own multimedia content and share with others. Furthermore, the mobile radio access networks have been advanced and the available bandwidth has been increased significantly. For example, the Third Generation (3G) mobile network, Universal Mobile Telecommunications System (UMTS) [82], with the maximum downlink speed upto 384 Kbps is enhanced with the High-Speed Downlink Packet Access (HSDPA) [62] that allows upto 14 Mbps. This broadband wireless communication further facilitates the delivery of bandwidth-demanding multimedia applications such as video streaming or video conferencing on mobile terminals.

Besides the voice-only service, people are using their mobile devices for E-mailing, sending text messages, browsing Internet web-sites, sharing pictures to each other, playing games, and watching video clips, movies, or even live television. Regular cell phones are giving way to smart mobile phones (i.e. iPhone [12]), and people are spending more time on using their devices according to the survey done by Morgan Stanley [140]. In particular, this survey shows what activities and the duration of each activity that an average user will spend his time on using a normal cell phone or a smart phone as depicted in Figure 1.1. Obviously, the percentage of time spent on a smartphone for talking is getting much smaller, since more time is dedicated to web-based and other activities.

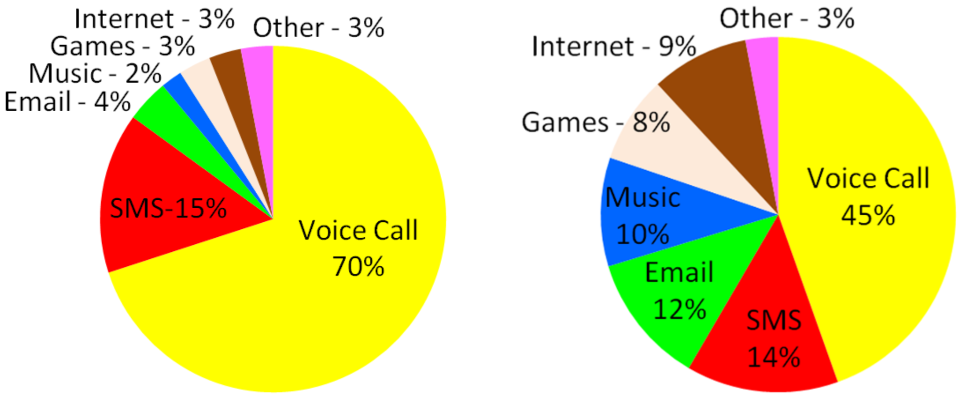


Figure 1.1: Mobile device daily usage breakdown showing a percentage of time on each activities on an average cell phone (left), and on the iPhone (right). (modified from [140])

From the emerging mobile multimedia communications mentioned above, one can classify them into different categories by considering different perspectives or view points [171], [34] as follows:

- *Content perspective*: One can classify multimedia communication into live content and stored content. Live content is the content that is encoded in real-time, and is transmitted to the receiver expecting to view and consume the content immediately. In contrast, stored content is off-line encoded content that is prepared ahead of consumption time and stored in a specific format. The stored content will be then transmitted at a later point in time to the receiver, for example, after a user has requested it.

- *Delivery perspective*: Downloading and streaming are sometimes used to classify multimedia communication. Downloading implies whenever the receiver only views and consumes the media content after finishing delivery all bit streams of the media content, whereas streaming refers to a transmission of media content that is split into separate independent chunks and thus allowing the receiver to play back already received parts of media content while other parts of bit stream are still being delivered to the receiver.

- *Network configuration perspective*: In terms of network model, multimedia communication can be classified into one of the two categories: client-server model or peer-to-peer (P2P) model. Client-server model describes the relationship between two computer programs, in which a central server hosts and transmits the media content to client(s). P2P model refers to an unstructured and distributed point-to-point communication among computers. Each computer acts as a peer that plays a role of both client and server to other peers.

- *Interaction perspective*: Each multimedia communication session can be classified into delay-tolerant sessions and delay-sensitive sessions. Sessions without interactivity such as video streaming or live broadcast are tolerant to delay and delay-jitter, whereas interactive sessions such as voice call, video conference and gaming require fast response to user interactions and hence are delay-sensitive.

- *Sender and receiver relationship perspective*: The relationship between the number of senders and receivers determines three different types of multimedia communication: unicast, multicast, and broadcast. A unicast communication is a point-to-point communication, in which only one sender and one receiver exist. A multicast communication consists of a single sender and a set of elected receivers participating in the same session. Alternatively, a single sender may transmit a media content to all receivers connected in a network, which forms a broadcast communication.

In practice, most of the mobile multimedia communication applications usually belong to multiple categories simultaneously. For example, video conferencing is an interactive multimedia communication that streams a live video content in real-time. In case of more than two users joining the conference bridge, it is a multicast communication. Otherwise, it is considered to be a unicast communication. Another example is IP-TV, which is a broadcast communication with streaming of live multimedia content. Since it has no interactivity, it is more tolerant to the delay during transmission compared to voice call communication.

Although mobile networks support as many types of multimedia communication as fixed networks, they are different, for example in terms of available transmission rate for wireless transmission and channel conditions. In the next section, we discuss the differences in detail and the challenges posed by mobile multimedia delivery.

1.2 Problem statement and contributions

According to the statistics released by vendors [39], [36], the global mobile data traffic continues to grow exponentially due to a tremendous demand on multimedia delivery over mobile networks. In particular, a study of visual networking index by Cisco [36] shows that voice communication will be a minority of mobile data traffic distribution share, and over two-thirds of the world's mobile traffic will be video by 2014 as shown in Figure 1.2. Even with mobile network upgrades to HSDPA or Long-Term Evolution (LTE) that allow a peak throughput of 14.4 Mbps and 173 Mbps respectively [11], a huge ramp in mobile data traffic driven by social networking, mobile Internet browsing, and video services is still considered to be a major reason causing network congestion. Hence, mobile access networks remain a bottleneck link of mobile multimedia communication when providing mobile multimedia services to a large number of users.

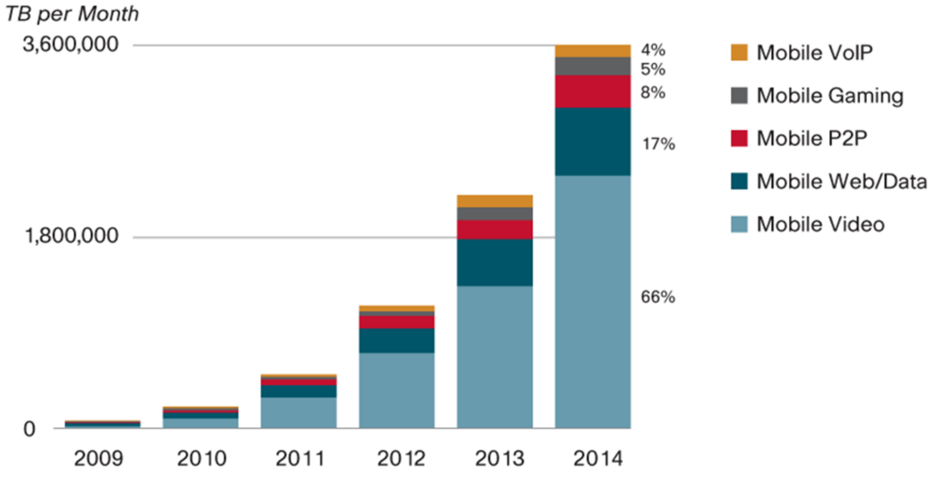


Figure 1.2: Forecast of global mobile data traffic growth [36].

Unlike a traditional Internet-based communication over fixed networks, mobile multimedia communication poses many challenges as follows.

- *Wireless lossy channels*: Packet losses in wireline networks are usually caused by congestions in intermediate routers, whereas the wireless channel possesses higher packet loss rate and bit error rate due to the signal fading from multipath effect, channel shadowing from urban obstacles, as well as the effects of noise and interferences from external sources [97], [171]. Mobility and hand-over across base-stations can also lead to packet loss due to out of order packet delivery as discussed in [122].
- *Bandwidth limitation and fluctuation*: Mobile networks are characterized by limited wireless network resources (or network capacity) that are shared among users. Users accessing the shared wireless medium (or served by the same base station) are usually mobile, which, in turn, will cause variability of the wireless channel condition. Hence,

the packet loss rate and the throughput from the time allocated to the respective user (allocated bandwidth) can also vary significantly over time.

- *Different Quality of Service (QoS)*[1] *requirements*: Different multimedia applications have different requirements in terms of bandwidth, data storage (buffer) prior to viewing/consuming media content, data transmission reliability, and deadline or timing for a continuous media playout. For instance, downloading an MP3 song to a portable device will take much less time and requires much less storage space than downloading a video clip file. Making a video call on a mobile phone requires a short delay of communication, and a good video quality can still be maintained under a certain amount of packet losses by using available error-resilience techniques. In contrast, a background service such as emailing and text messaging does not allow any packet loss, but has a large delay tolerance [165].
- *Different impact on Quality of Experience (QoE)*[2]: Packet loss stemming from an error-prone wireless channel and the application of rate-adaptation schemes to overcome time-varying wireless channel conditions (e.g. adaptive video coding, adaptive multirate speech codec) have a high impact on the user-perceived quality. The degree of this impact varies with the multimedia content and multimedia application type being transmitted. For example, missing a packet containing a scene of the video stream with dynamic scenes (e.g. sport video) can result into a huge gap of discontinuity of subsequent scenes, and thus makes it easier for a user to perceive a quality degradation. Whereas missing a scene of the video with static video content such as a paronama view or a news reporter, a user can often hardly recognize the difference, as the video content of subsequent scenes is quite similar.

In this thesis, we address the challenges above, however, we do not intend to provide a single solution that copes with the very complex and complicated problem. Our aim is to present a possible unified and formal approach to the problem of multimedia delivery over mobile networks. We consider several key aspects of how to efficiently use the limited wireless network resources, and focus on comprehensive traffic management and traffic engineering solutions that provide a high quality of service perceived by the end user to the possible largest number of users.

To achieve these goals, one approach is to collect information from all layers (e.g. 7 layers of the Open System Interconnection (OSI) model [69]) that are involved in a communication and jointly optimize them. However, this approach seems to be time consuming and not feasible in practice. We adopt a simplified Cross-Layer Optimization (CLO) framework that takes into account the link-layer and the application-layer information. The CLO framework used in this thesis is based on the previous work in [84, 79, 32, 33], in which only key parameters from the application layer and the lower layers of a wireless protocol

[1] In ITU-T E.800 [74], QoS is defined as "totality of characteristics of a telecommunications service that bear on its ability to satisfy stated and implied needs of the user of the service."

[2] In ITU-T P.10 [75], QoE is defined as "the overall acceptability of an application or service, as perceived subjectively by the end-user."

stack are abstracted. To cope with the problem of network resource allocation across multiple types of applications run by different users in a single cell scenario, the optimization scheme maps network and application parameters onto a common metric that quantifies the user perceived quality of service for the service delivery. For instance, in [87], the video utility function is obtained by varying the quantization steps for encoding a raw video and measures the average data rate and the video quality using the Peak Signal to Noise Ratio (PSNR), which is then linearly mapped to the Mean Opinion Score (MOS) [70] to capture user satisfaction. In this thesis, we update the video utility function by using the Structual SIMilarity (SSIM) [162] index and video SSIM index [164], as it matches well to human visual system quality perception that is highly adapted to extract a deformation of structural information instead of a pixel-based distortion. For the voice and file transfer application, we use the utility functions as described in [87]. We adopt the radio link layer parameter abstraction as proposed in [132], which determines the average maximum achievable data rate for each user experiencing different wireless channel conditions.

Within this CLO framework, we offer an in-depth discussion of the optimization problem with the goal of maximizing the overall user perceived quality of all users. In particular, we use a theoretical analysis to show that the optimum point of the objective function lies on the boundary of the utility space. Furthermore, we show that using a greedy search algorithm which starts from an arbitrary point on the boundary can quickly reach to the optimum. We extend the QoE-based optimization criteria with the max-min fairness for achieving similar perceived quality for all users. In addition, we apply a new constraint to the optimization objective, for example, a max-min fairness with a minimum guaranteed quality to all users. To substantiate some discussions, simulation results based on the HSDPA mobile network are included, showing that all QoE-based schemes proposed outperform the conventional mobile network system and the traditional throughput maximization approach.

Unlike work done in [87], we do not assume that the application server is located next to the base station. In order to avoid any delay of response in adapting the transmitted data rate due to signaling from the optimization module to the application server, we assume a rate adaptation module is available in the core network or is located at the base station, which uses a transcoding (transrating) or packet dropping technique. We investigate the complexity and the impact of applying different rate adaptation schemes on the user-perceived quality, which, in turn, allows us to derive a novel mechanism and algorithm for an intelligent rate adaptation scheme selection for resource-constrained wireless video transmission. The proposed scheme can be integrated with the QoE-based resource allocation optimization regardless of the optimization objective. In our example, we discuss how to integrate the proposed scheme with the objective of minimizing the media distortion perceived by the end user.

To further advance our optimization framework, we address a multi-objective optimization problem. We start with two objectives: the utility maximization and the max-min fairness. In this thesis, we define the utility as a degree of system efficiency describing how efficient the network resources are used in terms of the resulting average quality perceived by all users. Whereas, the fairness is defined as the difference of quality between the user experiencing the highest quality and another user experiencing the lowest quality. The two objectives are strongly depending on each other, since we can increase the system efficiency only when we decrease the degree of fairness balancing among all users. When the operation points for these contradicting between the system efficiency and the user fairness are too far apart, intermediary operation points may be preferred. We design a tuning mechanism, which allows a system operator to dynamically adjust its operation point between the extreme points of maximum system efficiency and fair resource allocation among all users. This mechanism is not limited to the case of static number of users joining the system, but it can also be applied for a dynamic environment, in which clients come online or go offline, access different applications, under variable channel conditions.

In addition to the user fairness and system efficiency, we formulate the CLO problem with a third objective. Here, we consider a constraint on the temporal fluctuation of video quality due to time-varying wireless channel conditions and the rate adaptation applied. The objective is to minimize the temporal change of the video quality that is perceivable by a user and negatively affects the overall user-perceived quality, while at the same time maintaining the average user-perceived quality of all users as high as possible. In fact, the novel QoE-based objective function is general, and can be combined with any other constraint. Furthermore, the proposed scheme gives flexibility to a mobile network operator to prioritize each of the two objectives according to its policy. The perceivable threshold of temporal video quality fluctuation is based on the Just Noticeable Difference (JND) concept [109], and is determined according to our extensive subjective tests that are performed in a room compliant with recommendation ITU-R BT.500 [72].

Finally, we consider all three criteria: system efficiency, user fairness, and temporal quality smoothness in our CLO framework. We propose a two-step optimization scheme with the aim of fulfilling both system efficiency and user fairness, while keeping perceivable quality fluctuation as low as possible. The proposed scheme is a practical approach searching for an optimal resource allocation taking into account all three criteria. First, we find an operating point that meets the constraints of system efficiency and user fairness that are assumed to be set by the operator prior to the optimization. The result of the first step optimization will be taken as a basis for finding a new operating point resulting to a smooth quality fluctuation with a possible highest average quality of all users under the user fairness constraint.

1.3 Thesis outline

The remainder of the thesis is organized as follows.

Chapter 2 first presents an overview of the state of the art in optimization for wireless multimedia delivery, which covers both the end-to-end rate control and the proxy-based rate control solutions. Some error concealment and scheduler-based techniques that improve the usage of wireless constrained resources are also discussed. The chapter proceeds to cover relevant basics of Cross-Layer Design (CLD) that enables several possibilities of interactions between the OSI layers. Following that, existing cross-layer optimization schemes, which take full advantage of adaptivity according to information sharing across layers, are discussed.

Fundamentals of our QoE-driven cross-layer optimization framework that is applied to the HSDPA mobile network are explained in Chapter 3. We start with background of HSDPA system and the long-term radio link-layer model that we use throughout this thesis. Next, we formulate multimedia QoE by constructing long-term utility functions (application-layer model), describe the multiuser utility space and derive its properties. We show analytically that the maximization of the sum of utility (max-MOS) can be efficiently solved by a fast greedy algorithm which searches only through the boundary of the utility space. We investigate two alternatives to the max-MOS approach, which introduce additional fairness in the system. We compare our proposed QoE-based cross layer optimization schemes to a system that is configured to maximize the overall throughput. For completeness, we also compare our approaches to a non-optimized HSDPA system. Besides the network resource allocation perspective, we investigate different rate adaptation schemes and its impact on the user-perceived quality. At the end, we discuss how to integrate a novel rate adaptation scheme selection with our QoE-based optimization framework, which, in turn, allow us to solve an optimization problem in presence of both constrained computational resources and constrained transmission resources.

In Chapter 4, the new strategies for QoE-based optimization are presented, adopting multi-objective optimization and making explicit the constraints and criteria. In particular, we consider the system efficiency, the user fairness, and the temporal quality fluctuation as an objective function for the resource allocation optimization. The chapter begins with the optimization problem taking into account the user fairness together with the average quality of all users. For this part, a tuning algorithm is presented that allows a mobile network operator to set constraints for the user fairness and the system efficiency. Next, we consider another multi-criteria combination of maximizing the average quality of all users (system efficiency) and minimizing a perceivable temporal video quality fluctuation. The latter utilizes the result of subjective tests to determine the average threshold of change in temporal video quality that is perceivable by human eyes. It also mitigates the effect of time-varying wireless channel condition. To complete this chapter, we discuss how to make use of the hybrid lexicographic method to solve the cross-layer optimization problem when taking all three criteria into account.

Chapter 5 concludes the thesis outlining the main lessons learned and pointing out potential future work.

It should be noted that parts of this dissertation have been published in [147, 89, 85, 149, 148, 150, 151, 146].

Chapter 2

State of the Art

With the fast growth of wireless networks and great success of Internet multimedia applications, future wireless networks are envisioned not just to provide a higher data rate to the mobile users, but also to serve various mobile terminals ranging from only-voice phone (dumb phone) to smart phone, and to support heterogeneous applications with different quality of service (QoS) requirements in terms of data rates, delay and packet loss [170]. QoS provisioning is a multidisciplinary topic comprising of several contingents, compassing from applications, terminals, networking architectures to network management, business models, and ultimately the end users [175]. The latter is crucial, as the network and service providers needs to know how their customers perceive the service provided. This enables the network and service providers to improve the service quality, and thus keeps their customers to stay in their business in the long term. Nevertheless, providing QoS or QoE in wireless networks is a challenging task, since the transmission condition of wireless channel dynamically changes over time as explained in the previous chapter.

In addition to the QoS support, the increased usage of a wide variety of cellular multimedia services through smart phones is putting an ever increasing demand for high data rates on the wireless interface. Although the traffic carrying capacity of wireless networks has increased significantly, the increase in actual user traffic continues to outpace. This has resulted in increased network congestion and many times in a degraded service experience for the user. Hence, in wireless networks, network resource management and resource allocation across multiple users together with QoS support for multiple applications have become a priority for network operators.

In this chapter, we provide an overview of related works addressing the aforementioned challenges and problems of wireless multimedia delivery. We start with a review of technologies available at each layer (independently) from lower layers (including both physical and link layer) of mobile radio access network upto the application-layer, and then discuss the Cross-Layer Design (CLD) based approaches. The technologies supporting circuit-switched services are out of the scope of this chapter of literature reviews, as the thesis concentrates on improving user-perceived quality for the Internet-based multimedia applications over wireless networks, which are considered as the packet-switched services.

2.1 Advances in mobile radio access networks

Over the last decade, considerable progress has been made in advancing the radio access technology in order to realize mobile broadband communication providing high system throughput, low round trip delay, and better QoS support. The High-Speed Downlink Packet Access (HSDPA) standardized by the 3rd Generation Partnership Project (3GPP) [3], [62] is one of the well-known radio access technologies that has been widely commercialized and deployed. In comparison to its predecessor, Universal Mobile Telecommunication System (UMTS), HSDPA is based on shared channel transmission, in which the channel codes and the transmission power available at each base station are commonly shared among users. This leads to a more efficient usage of available codes and transmission power resources when compared with the use of a dedicated channel. Furthermore, the downlink shared HSDPA channel (HS-DSCH) is a fast link adaptation channel that is based on Adaptive Modulation and Coding (AMC), Hybrid Automatic Repeat request (HARQ), and a short allocation time (Transmission Time Interval, TTI) of 2 msec. AMC adapts the modulation and coding scheme of the transmitted signal in accordance with variations in wireless channel conditions, which are periodically reported by the receiver. For instance, more information bearing bits are transmitted when the channel condition is good, and less information bearing bits are transmitted when the channel condition deteriorates. Whereas HARQ aims to control the error of the transmitted information bearing bits caused by the channel impairment in order to achieve an error-free transmission. HARQ uses both the Forward Error Correlation (FEC) for error protection and the Automatic Repeat request (ARQ) for retransmission of the errorneous received data. Another major change of HSDPA in constrast with UMTS is to move the packet scheduler from the centralized Radio Network Controller (RNC) to the base station (NodeB) and to embed the packet scheduling in its Medium Access Control (MAC) layer. With all these features, it enables high rate and robust data transmission, which in turn increases the total cell throughput up to approximately 14 Mbps.

In addition to the features introduced by both the PHY-layer and the MAC-layer as discussed above, HSDPA has another two sub-layers: Packet Data Convergence Protocol (PDCP) and Radio Link Control (RLC). The PDCP sublayer is used to compress the headers of higher layer protocols for efficient packet delivery, for example, IP header, whereas the RLC sublayer is used for fragmentation the IP packets to the small data unit to be transmitted over the HSDPA channel and for reassembling afterwards. The RLC also provides additional error detection, orderly packet delivery and link-layer retransmission, which leads to a low packet loss rate.

To differentiate QoS requirements of various types of services, HSDPA adopts the concept of UMTS traffic class (TC) and other UMTS bearer attributes (e.g. traffic handling priority (THP), allocation retention priority (ARP)) [37]. As specified in [2], four traffic classes have been defined, which are (1) conversational, (2) streaming, (3) interactive, and (4) background classes. The classification and the prioritized ordering are done according to their real-time needs (e.g. expected response in time from the user). For instance, delay sensitive services such as Voice over IP (VoIP) are given the high priority of the conversational class, whereas Web browsing is considered as an interactive class service, which has lower priority due to its delay tolerance.

The existing QoS control framework in UMTS has been modified and adapted to the HSDPA architecture due to the relocation of RNC functionality to the NodeB and the new features of HSDPA as discussed above. One example is a new QoS interface between the RNC and the NodeB (Iub interface), in which specific HSDPA QoS parameters set by the RNC are transferred to the NodeB. The Iub interface also allows the flow control, in which the NodeB is able to control the amount of data from the RNC in order to avoid packet loss due to buffer overflow at NodeB and at the same time to keep the delay on the NodeB buffer at low level. [108] demonstrated the impact of the Iub flow control on the system performance. The study showed that the IP packet delay increases significantly as the update period of the control loop becomes larger especially for the user mobility scenario.

In general, one can implement the QoS control in HSDPA [117] by implementing any combination of the three mechanisms as follows:

- *Dynamic transmission resource allocation*: HSDPA allows network operators to dynamically allocate transmission resources across multiple users in both time and code domains by adjusting the channelization codes available at each TTI. For example, upto 15 channelization codes can operate in the 5MHz WCDMA radio channel for a short TTI of 2 msec.
- *QoS-aware packet scheduling*: The behaviour of MAC-layer packet scheduler located at the NodeB is instructed by the HSDPA QoS parameters that are determined by the RNC based on the user's traffic class and other bearer attributes. These HSDPA QoS parameters are the Guaranteed Bit Rate (GBR), Scheduling Priority Indicator (SPI) and Discard Timer (DT). The scheduler uses the GBR parameter as a target average bit rate, whereas the SPI expresses the priority of the flow. The DT is used to avoid an unnecessary transmission of any outdated packets, which may be not useful at the receiver. The packet will be discarded, if the packet being buffered in NodeB stays longer than the allowed maximum time.
- *Quality-based admission control*: Admission control performed by the RNC plays an important role for efficient use of scarce radio resources. It determines whether a new user should be granted access or blocked depending on the user's QoS requirements and the current load in the cell, so as the QoS requirements of already admitted users remain fulfilled. The simple admission control uses the threshold-based mechanism [13], in which the system maintains the total allocated resources for all existing users

already served in the cell and for the new user less than the threshold. Admission control has been enhanced by taking into account, for example, the reservation of handoff calls [63], the estimated user mobility information [99], or the revenue of the service provider [66]. A comprehensive survey on the state-of-the-art of admission control solution approaches are presented in [110].

Details of the HSDPA QoS interface between the RNC and the NodeB are specified in 3GPP TS25.214 [6]. Nevertheless, the ultimate goal of applying the above QoS control mechanism is left opened in 3GPP specifications, as it is up to the NodeB manufacturer (vendor) and the operator to select the most appropriate solution taking into account their business models, for example, providing different service qualities for different classes of user subscription (e.g. premium user, flat-rate user).

2.2 Data transport and session control protocols

Delivery of Internet multimedia services is based on the Internet Protocol (IP) [119] and transport protocols standardized by the Internet Engineering Task Force (IETF). There are basically two transport protocols that are usually used to carry media content: Transmission Control Protocol (TCP) and User Data Protocol (UDP). TCP [120] is a connection-oriented protocol that offers an error correction and a reliable and orderly packet data transfer. It is mainly used for a multimedia application that is not delay sensitive but requires a guaranteed delivery that the received packet is identical to its original including the ordering of the data. TCP-based applications are for example web browsing, E-mailing and file transfer application. In contrast, UDP [118] is a lightweight connectionless protocol that aims to transmit the content to the receiver as fast as possible without guaranteeing the correctness. It is suitable for an application that has stringent real-time constraints such as VoIP or live video streaming. With real-time applications, people prefer to loose some audio frames or video packets rather than have to wait a few seconds for the network to recover and retransmit.

In practice, many commercialized multimedia applications often combine the above TCP or UDP protocol with other protocols standardized by the IETF. For instance, video streaming services use the UDP together with the Real-time Transport Protocol (RTP) [133] to carry the media stream and the Real-time Transport Protocol (RTCP) [133] to monitor transmission statistics (e.g. round-trip time, packet loss rate) and control the service quality. The Real Time Streaming Protocol (RTSP) [134] can also be used simultaneously for establishing or tearing down a media session and remotely controlling media stream sent from the server that is similar to VCR-like commands such as play and pause. Alternatively, end-points can establish a session by using the Session Initiation Protocol (SIP) [128] or the Hypertext Transport Protocol (HTTP) [43]. For web-browsing, it uses HTTP over TCP to carry the web page information that is usually coded in the Hypertext Markup Language (HTML) [124]. In addition, IETF specified the Session Description Protocol

(SDP) [59] providing a standard presentation for describing multimedia sessions for the purposes of session announcement, session invitation and session establishment. This includes, for example, the media and transport details (e.g. media type, media format, used transport protocol), a high-level session description (e.g. session name, purpose, privacy), etc. Selection of which transport protocol and session control protocol to be used and how to make use of them is implementation specific, as it depends on the application's needs and the service provider's requirements.

Originally, the above transport protocols were designed and optimized for wired networks. For example, TCP provides a transmission rate control based on the cumulative acknowledgements of transmitted packets with an aim to efficiently use the available transmission capacity in the network and to avoid network congestion, which then causes packet losses. The behaviour of sender-based rate adjustment uses the Additive Increase and Multiplicative Decrease (AIMD) [31] that increases the transmission rate in a step-like fashion in the absence of packet loss and reduces multiplicatively when congestion is detected. Details of TCP congestion control and TCP Friendly Rate Control (TFRC) are specified in RFC5681 [10] and RFC5348 [47] respectively. For the UDP, it provides error detection in the packet header or payload by using the cyclic redundancy check (CRC). However, the UDP does perform neither any error recovery nor congestion control mechanism. If an error is detected, it simply discards the whole packet. In [91], IETF specified the Datagram Congestion Control Protocol (DCCP) that provides an effective congestion control mechanism for unreliable datagram flows. The DCCP enables existing and new applications that are delay sensitive to easily use it to transfer timely data without destabilizing the network (avoiding network congestion).

Compared to wireline packet networks, multimedia delivery in wireless networks is characterized as low transmission rate and unreliable. The packet losses are induced by both error-prone wireless channel and network congestion. For TCP, many studies [27, 15, 98] have shown that the TCP has a poor performance in the context of mobile communications due to the inherent wireless transmission characteristics such as high bit-error rate, the user's mobility and the limited transmission bandwidth. One of the major reasons is because the TCP considers any lost packet as a signal of network congestion and adjusts its transmission rate accordingly. However, packet losses can also stem, for example, from the high bit error rate over wireless links, the call handoff across base stations, or the unpredictable disconnection while the user is in motion. Using the standard congestion control and error recovery through retransmission hence results in unnecessary degraded performance of bandwidth utilization and system throughput [153]. A number of efforts have been made to improve the performance of TCP in wireless environments, for example, freeze TCP [56], and TCP-Probing [152]. This includes efforts in the standardization communities, for example the IETF provides an Explicit Congestion Notification (ECN) mechanism [125] that allows any intermediate node located along the transmission path to inform the end hosts (receiver and sender) about an incipient congestion at its node, such that the sender invokes a congestion control algorithm and thus avoiding a packet dropping at the intermediate nodes. However, the ECN is still not widely deployed, since it requires that all intermediate nodes and end hosts must support ECN. In [129], Sardar *et al.* classified TCP enhancements into two categories: connection management related approaches and wireless loss related approaches. They also provided a qualitative comparative study of different existing solutions of TCP.

Transmission of UDP packets in error-prone network environments is inefficient, since most of the corrupted packets are discarded by a checksum at UDP protocol stack. This also applies for packets containing only a small part of corrupted data or even only single-bit errors, which are sometimes useful packets to the application layer [95]. Codecs for voice (e.g. AMR codec [138]) and video (e.g. H.264 [76]) are examples of application that benefit from having damaged data delivered rather than discarded by the network due to its build-in error-resilience capability. Consequently, the UDP-Lite [93] has been introduced to avoid unnecessary packet discarding by UDP, and thus reducing the excessive packet loss rate for UDP traffic. Instead of having all or none of packets being protected by a checksum, UDP-Lite provides flexibility in the form of partial checksum that allows senders to specify the coverage of the checksum on a per-packet basis and to define the payload as partially insensitive to bit errors. In [137], Singh *et al.* proved that the use of flexible checksumming scheme improves the overall performance in terms of delay and packet loss. However, UDP-Lite imposes the backward compatibility problem, as it has its own protocol identifier, which makes it compatible only with the devices/applications that are UDP-Lite capable. Another variance of UDP called UDP-Liter has been proposed in [92], which requires minor modifications to the traditional UDP and BSD socket API, but yet maintain backward compatibility. Through the socket call, the application specifies an option for the UDP-Liter to either retain the traditional UDP behaviour or simply pass packets to the application. Nonetheless, a drawback of the UDP-Liter is that it is unable to differentiate header errors from payload errors, which may then lead to a mishandling of packets at the application-layer.

Using RTP over UDP is resilient against packet losses to some extent, as it provides a restoration mechanism of packet re-ordering through packet sequence ID and timestamp, and a feedback mechanism through the exchanging of RTCP report that allows a sender to adapt the coding scheme and transmission behavior to the observed network Quality of Service (QoS) such as round-trip delay, jitter, etc. The adaptation can be done in the order of several seconds to minutes. Additional supplement statistics, which are mainly useful for real-time monitoring and diagnosis for VoIP application (e.g. average loss rate, burst duration, gap duration, average mean opinion score for voice quality, etc.), can be conveyed by the RTCP Extended Report (RTCP-XR) described in the RFC3611 [49]. However, RTP makes no provision for errorneous/distorted packet recovery and a timely feedback that would allow a sender to repair the media stream immediately. An extension of RTCP-based feedback for the Audio-Visual Profile (RTP/AVPF) has been proposed to address the aforementioned problems and standardized by the IETF in RFC4585 [113]. This early feedback profile (AVPF) is used to convey information about events observed at a receiver such as packet loss, packet reception, frame loss, etc. Having received the feedback from the receiver, the sender can then react accordingly. For instance, when the receiver sends a Picture Loss Indication (PLI) message to inform the sender about the loss of an intra-picture of the encoded video data, the sender becomes aware that the prediction chain may be broken and thus may react to a PLI by restransmitting the intra-picture to recover the error. IETF RFC4588 [127] has specified an RTP payload format for the RTP retransmitted packets that are sent in a separate stream from the original RTP stream.

The Packet-switched Streaming Service (PSS) defined by the 3GPP [7] specifies an end-to-end based bitrate adaptation mechanism allowing a sender to control the transmission bit rate in order to avoid packet losses caused by the network congestion or buffer overflow at the receiver and a buffer underflow that would interrupt the continuous playback, and thus providing best possible service quality to the end user. To achieve these goals, PSS uses several IETF standards such as RTCP-XR [49], RTP/AVPF [113]. Moreover, the PSS extends the SDP and RTSP in the form of attributes, option tags and headers, so as the sender is able to periodically monitor both link rate and client buffer status through the client's feedback. For example, the SDP attribute 'a = 3GPPAdaptation-Support' requests the client (receiver) to provide buffer status feedback and to configure how frequently it should be done. The RTSP header '3GPP-Adaptation' is used to inform the sender about the client's buffer size and minimum required buffering to ensure interrupt-free playback. In [50], results showed that the sender adapts the transmission bit rate with respect to the change of channel rate capacity, and thus the significant improved media quality is offered.

Lately, IETF has started standardizing how to make use of the ECN concept for the RTP flows running over UDP [166], which allows real-time applications to respond to the onset of the congestion (via the ECN flag) before an intermediate network node is forced to drop packets. The objective is to enable real-time applications to control their transmission rate, rather than trying to conceal the negative effects of unpredictable packet loss. In contrast to the conventional ECN for TCP flows [125], ECN for RTP over UDP/IP requires an extension of the RTP/AVPF feedback for urgent feedback of ECN information, an extension of the RTCP-XR for ECN summary report for the regular RTCP transmission period, and an extension of the SDP for negotiation of the ECN capability between the end hosts.

Media content delivery on the web usually uses HTTP over TCP [43]. In the past, viewing of the media content was only possible after having finished downloading an entire media content. But today, many content providers (e.g. YouTube) deliver their media content using the progressive download approach, which allows the users to view the content as soon as enough data is retrieved and buffered at the client while the download is still in progress. However, the progressive download still has several drawbacks. First, it does not allow the user to change the media content to be downloaded on the fly, if there are different versions for the same content. This is because the requested media content is seen as one big chunk from the client side. Consequently, this makes it inappropriate especially for the wireless network, in which the available transmission rate varies over time. The user will experience video stalling due to rebuffering, if the available channel bit rate is less than the data rate for longer period of time, which is required by the media content of the selected version. Second, the progressive download does not offer rich features of streaming (trick modes) such as play, rewind, etc. Adaptive HTTP streaming based approach [16] has been proposed to address these shortcomings while preserving the simplicity of progressive download.

Adaptive HTTP streaming is a hybrid of progressive download and streaming. It allows the user to selectively view the segment of media content (rather than one big progressing download) and dynamically select the short duration of media segment at different quality (different bit rate) that matches to the available transmission rate, while continuing to download the content from the server in the background. The latter is important, as this would allow an efficient usage of time-varying network resources. Since the media segments are short, it enables the client to use trick modes efficiently. As a result, the user would have an impression as of streaming applications. Currently, there are two different implementations of adaptive HTTP streaming based on the multi-bitrate fragments: Smooth Streaming from Microsoft [105] and HTTP Live Streaming from Apple [114]. Microsoft's implementation uses the Protected Interoperable File Format (PIFF) [20] as an extension of the MPEG4 (MP4) file format specification [78], whereas the Apple's variance uses the MPEG2 Transport Stream file format [77] for the fragmented content storage.

2.3 Adaptive multimedia applications

So far, we have discussed how the radio access networks and its transports are enhanced to cope with the QoS provisioning focusing on the layer1/2 and transport technologies respectively. In this section, we will discuss the advances in multimedia applications aiming to avoid congestion in a packet-switched network and to protect and recover errorneous or missing bits of media content due to a transmission over an error-prone wireless interface.

2.3.1 Application-level congestion control

Many TCP-based applications such as web-browsing, file downloading, exploit existing congestion control mechanisms in TCP as discussed in Section 2.1, since they work well for networks with heavy TCP traffic. With this, the TCP-based application is transparent to any changes of transmission rate done by the TCP. However, for real-time applications like VoIP and video conferencing that use UDP/RTP for its transport, the application itself is actively involved in congestion control. In particular, the application makes use of the QoS feedback from the receiver through, for example, the RTCP-suit protocols also discussed in Section 2.1 to control its transmission rate in order to avoid network congestions.

Due to the fact that the networks generally consist of both TCP flows and non-TCP flows, if we assume that the non-TCP flows exist in the networks and do not have any mechanisms to adapt their transmission rate, this situation can lead to starvation of TCP flows [45]. To avoid any performance degradation due to the aforementioned problem, for which the TCP flows and the non-TCP flows are competing for transmission on a shared wireless link, several TCP-friendly congestion/rate control algorithms for non-TCP flows have been proposed, for example, the Rate Control Scheme (RCS) [145], the Analytical Rate Control (ARC) [9]. The RCS suits for the typical satellite links, which have high bandwidth-delay products and high bit error rates, while the ARC is designed for the wireless networks with low access delay. Another effect of the two schemes is to reduce the negative impact of the wireless link error on the throughput. However, its drawback is that they do not adapt well to the abrupt increase of the bandwidth, as they do not consider the bandwidth variation over wireless networks. In [96], Lee *et al.* proposed a new TCP-friendly congestion control algorithm based on the ECN [125] for streaming real-time applications that takes into account both the wireless link error and the available bandwidth. The estimation of the available bandwidth of the bottleneck link is done by using the inter-arrival time of ECN ACK packets. With this, the proposed ECN-based scheme utilizes the time-varying bandwidth efficiently without penalizing TCP flows, and thus avoiding a network underutilization in presence of an abrupt increase of the bandwidth.

Other application-level congestion control algorithms are the TCP-Friendly Rate Adaptation Based on Loss (TRABOL) [14] and the congestion-threshold based approach [35]. The TRABOL employs the concept of the TCP congestion control at the application level. The sender's transmission rate is adapted according to the number of lost packets within an interval of time measurement at the receiver's side, but without delay requirements of real-time applications. In [35], Chua *et al.* proposed a new solution allowing the application to perform a congestion control taking into account the inherent characteristics of real-time applications (e.g. end-to-end delay requirement, impact of delay variation on the QoS). The proposed solution consists of the two independent components: congestion detection/notification and adaptive transmission control. The detection/notification is done at the receiver, whereas the adaptation is performed at the sender. To detect congestion, it computes the absolute value of the difference between the inter-packet arriving intervals and the original inter-packet transmission interval. Moreover, the detection algorithm tells the transmitter to what degree congestion is present by evaluating the computed absolute difference by using a simple first-order infinite impulse response (IIR) filter in order to obtain severity of the congestion in the network. Using the congestion-severity information, the transmitting endpoint selects the bandwidth-reduction method that is most appropriate for the level of congestion, thereby maintaining optimal link utilization and throughput while simultaneously curbing congestion. Two bandwidth-reduction techniques are discussed: compression switching approach and multi-packet merging approach. The former is simply to switch to a different compression algorithm (e.g. different encoding quantization parameters) resulting to a variation of bit rate, whereas the latter method is to transmit the same data with fewer transmissions by covering a longer period of time with the data it packs into each IP packet.

2.3.2 Application-layer error control and recovery

In addition to the error detection and retransmission (ARQ), and error correction (FEC) available at the physical and link-layers of the wireless protocol stack discussed in Section 2.1, many application-layer techniques have been proposed to make the multimedia content delivery more resilient to wireless transmission errors [55, 34]. One major reason is that the lower-layer error protection mechanisms only work for wireless link errors, whereas the application-layer error control mechanisms work for both packet loss due to network congestion and wireless link errors. Second, though the transport level provides a packet retransmission to guarantee a certain level of quality as discussed in Section 2.2. However, it has disadvantages, as retransmission introduces additional delay and thus it is not appropriate to some delay-sensitive applications such as voice call and video streaming. The application-layer error control avoids such unnecessary delay by using an error protection or an error correction technique that is fully relied on the end-points [160]. One can classify the application-layer error control techniques that are provided at the sender or at the receiver. A comprehensive survey of error resilience techniques can be found in [160] for a general overview for video application and in [141, 41] for the specific video codecs such as the H.264-AVC standard. Below, we briefly discuss each approach with the focus on its application to the delay-sensitive applications.

- *Sender-based error resilience*: To alleviate quality degradation due to packet losses during the transmission over a lossy channel, the sender exploits an error resilient source encoding technique, for example, by adding redundancy in the bitstream. This helps the receiver to recover from the transmission errors, but it also comes at a cost of coding efficiency due to the fact that it uses more bits to obtain the same video quality in the absence of any transmission errors. Hence, its design goal is to minimize the redundancy while achieving a desired level of resilience in presence of packet losses. There are several ways to introduce redundancy in the bistream. One simple but effective approach is to insert resynchronization codes periodically, which is already supported in MPEG-4 [48], in order to localize transmission errors and to limit the quality degradation within the same codewords, for example, within the same slice in a video frame. Instead of discarding the whole slice containing only a few corrupted bits, the Reversible Variable Length Code (RVLC) [26] further limits the errorneous region by allowing the receiver to decode in both forward and backward directions between the two resynchronization codewords. However, in advanced codecs such as H.264/AVC [168] that uses the temporal prediction to achieve a higher coding efficiency, an error of a single slice of a frame would lead to an error in decoding of the subsequent reconstructed frames. To stop such temporal error propagation, one can insert an intra-coded pictures or blocks [60], which are independently coded.
- *Receiver-based error resilience*: In contrast to the error resilient source encoding, error concealment has an advantage of not employing any additional bitrate, however, it adds computational complexity at the decoder due to the detection and concealment

of errors. In video applications, a simple error concealment approach is to copy either the whole previous frame or the slices of co-located positions in the reference frame. Alternatively, the decoder recovers damaged regions using the texture information from the surrounding regions in the same frame (spatial interpolation) or in nearby frames (temporal interpolation) [64]. In [130], it has been shown that a combination of spatial and temporal interpolation also leads to a significant gain for improving the video quality in error-prone environments. More sophisticated concealment approaches are for example the recovery of corrupted blocks by using the motion vector information of surrounding blocks [53], or replenishment of missing blocks by taking into account the perception of the human visual system (HVS) [17].

2.4 Cross layer design

Several existing approaches mainly address finding efficient wireless transmission techniques based on optimizing a single layer in the protocol stack, in which messages are interchanged between entities of the same layer as discussed in previous sections. Each layer is aware of its own messages and embeds its information into upper layer messages when they go down in the layer stack, while it discards the lower layers' information when messages go up. Although such a layered architecture offers simplicity and modularity of protocol design, recent results indicate that the traditional seven-layers of the Open Systems Interconnection (OSI) model [167], which was originally designed for wired networks, may not lead to a global optimization for wireless communication systems due to different characteristics and different requirements of the wired and wireless medium. For instance, one of the well-known assumptions in the TCP protocol is that packet loss is caused by network congestion. However, in wireless systems, packet loss often occurs due to corruption. Moreover, the performance in wireless networking is very sensitive to the mobility and the surrounding environment. This makes the wireless systems much more complex than wired communication networks, which are assumed to have high reliability and high communication capacity.

A new paradigm to design networks by optimizing across layers, so called "Cross-Layer Design (CLD)" has been proposed in [58, 79] to take full advantage of adaptivity according to information sharing across layers. Figure 2.1 shows the OSI layered model and a subset of the possible interactions that can be considered in the CLD. As an example of downward information sharing, the application would notify the lower layers of the expected communication load in the near future, thus allowing the network to reconfigure and move resources into areas of higher demand. For upward information sharing, the application, in response to its knowledge of the status of the network, could either modify (or even totally defer) the requested Web-page information, for instance, by avoiding fetching of high-volume data.

Nevertheless, the violation of the original OSI architecture brings also some disadvantages

[83]. For instance, with the CLD architecture, it sometimes does not represent the actual system and can lead to the complete loss of the meaning of the initial architecture. The level of modularity of the network and abstraction in the network implementation is reduced due to the cross layer design, which increases the complexity of the network. It is important to note that the layers are not independent anymore and any change in one layer could have impact onto other layers.

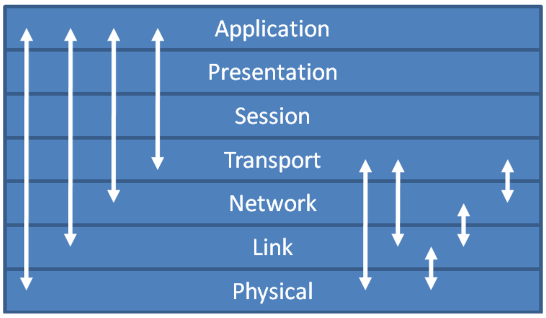


Figure 2.1: Subset of Cross-Layer Design in the OSI model.

Applying the CLD approach for the wireless networking was introduced in [135, 29, 139] to address the poor performance in wireless multimedia delivery and to efficiently allocate the scarce radio resources while still providing QoS to mobile users. To enable the communication between layers, the CLD requires cross-layer signaling, which can be classified into two groups as follows:

2.4.1 Internal cross-layer signalling

Internal cross-layer signalling is the message internally exchanged among different OSI layers inside of a physical entity. For example, due to a bad channel quality (e.g. very low signal strength between the mobile device and the base station), the physical layer on the mobile device may inform the conference call application running on that mobile device to send the video stream at low resolution and framerate.

The simplest way of sharing cross-layer information is to have a common signalling pipe across all layers as depicted in Figure 2.2-a. The internal signalling can be a dedicated protocol or a standard protocol. In [172], Wu *et al.* introduced an extension header for IPv6 header, called "Wireless Extension Header (WEH)". The WEH extension header is used as an internal storage of the cross-layer information, which is then transported through the general signalling pipe. Results show that sharing radio-link parameters (e.g. data rate, radio-link round trip delay, fading conditions) to the TCP will improve the system performance, since TCP can help the Radio Link Protocol (RLP) by lowering the Round-trip Time Out (RTO) upper bound and setting a smaller timer back-off factor, to the extent that it does not lead to congestion collapse.

Another approach is to use the selected holes method as depicted in Figure 2.2-b. This method is different than the common signalling pipe, as the internal signalling message that is used in one propagation path may not be the same as other paths, but still carrying similar content [142]. For example, the link-layer may use an IP specific additional header to carry information about the current wireless channel conditions, and the network-layer may extract this information and send to upper layers in another form such as using the standard ICMP header.

The Cross-LAyer Signalling Shortcuts (CLASS) proposed in [159] employ a direct signaling between non-neighboring layers. This scheme is more efficient, flexible and comprehensive than the two methods discussed above, as the intermediate layer(s) along the propagation path from the source layer to the destination layer is not involved, and therefore, avoiding unnecessary processing overhead and propagation latency. CLASS is a new light-weighted protocol, as it reduces additional headers (e.g. IP headers, common ICMP header), that are not necessary when performing cross-layer optimization, and simplifies the message format. The signalling shortcuts concept is depicted in Figure 2.2-c.

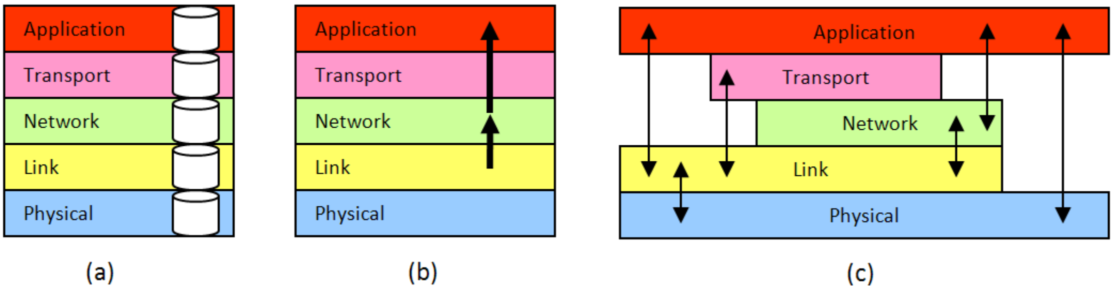


Figure 2.2: Options for internal cross-layer signalling: (a) General signalling pipe, (b) Selected signalling holes and (c) Shortcuts signalling.

2.4.2 External cross-layer signalling

External cross-layer signalling refers to the exchange of messages among different OSI layers across multiple physical entities. For example, the application server connected to the core network may inform the base station about the required data rate, so that the base station may reserve radio resources in order to fulfill the requirement.

Signaling for resource reservation prior to the data transmission has been extensively studied in the last decades. One of the first standardized resource reservation architecture was the Integrated Service (IntServ) [21], which uses the Resource reSerVation Protocol (RSVP) [22] as the underlying mechanism to signal the reservation information across the network. However, IntServ is not very popular due to its scalability issue in large size networks. It requires all nodes along the data path to be stateful, which continuously monitor their network resource utilization and calculate available resources for the incoming RSVP reservation requests. The Differentiated Services (DiffServ) [19] provides a simple and scalable QoS mechanism that prioritizes IP packets based on the traffic class (e.g. voice, streaming media, or web-based traffic). Each router treats each data packet based on its class that are set in the Type of Service (ToS) field of the IP header.

Besides the resource reservation, the external cross-layer signalling is also used to provide an intelligent multimedia content adaptation [90] or to recover from errors during the transmission [176]. In [90], an intelligent network service collects and stores the Wireless Channel Information (WCI) at the WCI server via the logical interfaces linked with the base stations and other network elements. The collected information is then abstracted in XML format that can be retrieved through signalling by the application server in order to adapt the multimedia content with respect to the wireless channel condition. In [176], Zheng and Boyce proposed a new UDP that utilizes a cross-layer technique to interchange information from the physical and link layers to the IP or transport layer in order to assist error recovery at the packet level.

Hints and Notifications (HAN) [94] allow the application and transport layers to communicate with wireless link layers in order to utilize the wireless network resources efficiently and improve the service quality for individual users. Hints contain useful application-layer information which is used to guide the link layer, e.g., the boundaries between parts of a packet and the acceptable error-rate and delay requirements for each part. Notifications are information from the link layer sent to the application layer informing the application layer how to react more accurately due to variations in the physical medium.

The IST project M-Pipe [144] proposed a uni-directional cross-layer signalling, which carries information about the application, so called ”Layer Independent Descriptor (LID)”, along the user plane path in order to guide the network elements for local resource management and rate adaptation. No feedback from the network elements is required. The LID contains three main information groups: traffic class, packet drop preference, and error protection preference. The traffic class identifies whether the application is loss-tolerant, adaptive and scalable. Whereas the packet drop dependency specifies the offsets of truncation points, which are used to specify the error protection preference due to the error-prone wireless transmission. The LID uses the standard signalling protocol such as RSVP as its transport.

2.4.3 Cross layer optimization in wireless networks

Exchanging key parameters across the layers as discussed earlier enables the network operator to perform the Cross-Layer Optimization (CLO). The CLO is basically done by collecting the correct information from various layers, manipulating them in a decision center and distributing the decisions to layers so that the overall optimization of the system is achieved instead of an individual and separate optimization in the layer. There are a number of researches on using the CLO technique to solve the problems in wireless multmedia delivery. For instance, the network operator implements the CLO with the aim of efficient packet scheduling [80, 18, 52, 51], efficient modulation scheme selection [157, 121], efficient joint channel and source coding [44, 23], or even efficient power management for mobile devices [174, 65]. For resource allocation optimization, most works are based on the conventional throughput maximization [100, 154], or a combination of throughput optimization with queue length information in scheduling to ensure fairness of resource allocation [40, 107, 103]. In [84], a joint optimization of application and link layers, including the physical and medium access layers, was proposed in order to efficiently use the limited network resource taking into account the impact on the video quality. Only key parameters from each layer are abstracted and used for the optimization. Later, a new way of abstraction of application-layer information for video streaming was proposed by using the Rate-Distortion (RD) model, which leads to a low number of application model parameters [33]. The proposed RD models are applicable for both the Mean Square Error (MSE) or the Peak Signal to Noise Ratio (PSNR). A QoS-aware scheduling algorithm for wireless video delivery was presented in [101], in which the PSNR was used as a video quality metric. She *et al.* [136] proposed an intelligent active packet dropping at the MAC-layer for real-time video streaming based on a new quality metric taking into account the frame rate when calculating the video quality.

A utility-based cross-layer optimization framework was first proposed in [86], where a concave utility function for an elastic traffic (e.g., webbrowsing, file transfer) is used to capture the user satisfaction as a function of data rate. Whereas, a non-concave sigmoidal utility functions [102] has been proposed for inelastic flows (e.g., voice or video). For a comprehensive overview of the Network Utility Maximization (NUM) framework please refer to [30] and the references therein. In [28, 123], a new fairness concept with utility max-min allocation has been introduced that corresponds to the satisfaction of each user in the system being equal regardless of the application type and the channel quality condition. They showed that the utility max-min allocation outperforms the bandwidth max-min allocation under a variety of utility functions for different application classes.

Unlike other previous works, which mainly concentrated on the system supporting only a single application type, Khan *et al.* [87] extended the CLO by taking into account the user-perceived quality aspect, and also made it a general framework which can be applied to any applications (e.g. voice calls, video streaming, file transfer). To achieve this, the Mean Opinion Score (MOS), which was originally used for voice call quality assessment [73], is used as a common metric in the optimization scheme to represent the user perceived quality of each application as a function of transmission data rate and packet loss rate. Similar to the utility maximization, the resource allocation optimization in [87] aims to achieve the maximum average user quality. The proposed MOS-based scheme was extended by targeting at different objective functions such as a modified max-min fairness that guarantees a minimum service quality to all users [131]. Their simulation results showed that the MOS-based optimization leads to remarkable improvements in terms of user-perceived quality respectively an increased number of users in a cell when compared to the conventional throughput-based optimization and the non-optimized system.

2.5 Justification and positioning

In this thesis, we focus on the QoE-driven optimization for network resource allocation in cellular networks based on the work presented in [87], which is important for an network operator due to several reasons. Firstly, current throughput-based optimization only makes sense in case of packet-based charging. High cellular bitrates, e.g. 3GPP Long Term Evolution (LTE), which is expected to approach a peak bit rate of 100 Mbps [38], would further push a flat-rate billing model or models based on quality guarantees. In that scenario, the operators would find a clear motivation to maximize the satisfaction of their customers, irrespective of the requirements of their services. Secondly, user satisfaction is gaining importance to the operators who realize that unsatisfied users would usually quit the network without ever complaining to the operator, and would possibly share their experience with other potential customers, resulting in increased customer churn rate and thus severe loss of revenues. Thirdly, QoE-based optimization allows potentially more customers to be served simultaneously without a loss of user perceived quality. Lastly, the proposed QoE-based resource allocation is performed at every second. It fills a gap between physical layer transmission intervals (2ms in HSDPA) and long term application layer mechanisms such as adaptive streaming or TCP congestion control (between 5 and 10 seconds).

Contrary to most of the NUM literatures, where only concave, continuously differentiable utility functions and theoretical link models are assumed, our scheme proposes a framework considering realistic utility functions and applies the framework to a standardized cellular network system such as the HSDPA system. Unlike [87], which only takes into account the network resource constraint, we consider additionally the hardware constraint of computational resources used for processing the in-network rate adaptation. Furthermore, we go one more step beyond the state of the art solutions by investigating a multi-criteria optimization that searches for an optimal resource allocation with the objectives of maximizing the average user-perceived quality of all users, minimizing the maximum quality difference among users, and minimizing the perceivable temporal quality fluctuation. The last objective function is newly proposed aiming to reduce a temporal quality change that would have a negative impact on the overall user-perceived quality.

Chapter 3

QoE-driven cross layer optimization

The increased usage of a wide variety of cellular multimedia services is putting an ever increasing demand for high data rates on the wireless interface. As the downlink of the cellular system often acts as the bottleneck link, an efficient usage of downlink wireless resources becomes essential in order to provide high quality of services to the largest possible number of users. The time varying transmission conditions of the wireless channel and the dynamic changes of application requirements of multimedia services make the optimization of downlink resources a challenging task. We have seen in the previous chapter that there have been many studies over the past years for optimizing resource allocation in wireless systems. The Cross-Layer Design (CLD) based approach is one of the promising techniques expected to be widely deployed.

The work presented in this chapter is based on the Cross-Layer Optimization (CLO) framework proposed in [87], which jointly optimizes the application layer and the lower layers of the wireless protocol stack with the aim of improving the user-perceived quality (Quality of Experience, QoE). In our work, we consider a more practical approach to deploying the CLO framework in mobile systems, which invalidates some of the assumptions given in [87]. For instance, previous work assumes that video streaming servers are located very near to the mobile base-station and thus allows data rate adaptation by changing the quantization at the video encoder. In a realistic scenario, video servers are located outside the mobile network. Furthermore, using explicit application-layer signalling for rate adaptation has several drawbacks. For example, it imposes an additional delay in response to the congestion problem. Moreover, it requires a video server end-system to support the signalling for rate adaptation, and thus leading to a backward-compatibility issue.

In our work, we consider a more realistic scenario, in which each user accesses multimedia contents that are encoded at high quality and are stored at the Application Server (AS) located outside the mobile network. To enable the proposed QoE-driven optimization, we introduce two main functional entities, the Traffic Management (TM) module and a Traffic Engineering (TE) module that are located inside the mobile core network as depicted in

Figure 3.1. The TM module acts as an optimizer for downlink resource allocation, whereas the TE module acts as a controller for rate adaptation. Though the figure shows the location of both modules in the mobile core network, in fact, they can also be placed in the Radio Access Network (RAN), e.g. at the base station. It is not necessary that both modules are co-located at the same place. For instance, an operator may place the TM at the base station and the TE at the gateway towards networks outside the mobile network such as the Internet. With this, the network operator can save unnecessary traffic coming from outside for the whole mobile network. Optimization is done based on lower layer information (e.g., average channel quality), the objective function, and the application utility functions, which are either stored in advance or sent along with the data stream.

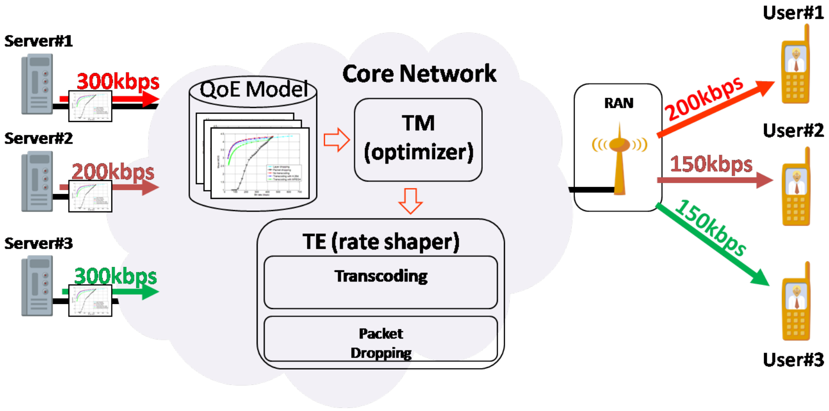


Figure 3.1: Target use case and network configuration considered in this chapter.

Throughout this thesis, we use the High-Speed Data Packet Access (HSDPA) system as an example of the target mobile network system for the practical approach validation. However, the QoE-driven optimization framework is general and hence can be applied for any mobile network systems.

In the following sections, we first quickly go through the basics of the HSDPA system and discuss the long-term models for both the application utility function and the radio link-layer model related to the provided functionalities and features of the HSDPA system. Next, we elaborate the architecture of the CLO framework applied to the HSDPA system. For determining the goal of the optimization, we discuss various objective functions implemented in an HSDPA simulator, which allows us to compare all proposed QoE-based schemes with other existing techniques such as the throughput maximization approach. We describe details of the greedy search algorithm, which is used as a heuristic approach to find a good solution. Finally, we investigate the impact of applying different rate adaptation techniques to the user-perceived quality and discuss how to integrate them into the proposed CLO framework for the constrained system, which is characterized by both limited network resources for data transmissions and limited computational resources for processing of rate adaptation.

3.1 HSDPA overview

The key concept of HSDPA is to increase the packet data throughput using link adaptation and fast retransmission from the base station (Node B). Link adaptation of HSDPA uses Adaptive Modulation and Coding (AMC) with two modulation schemes, QPSK and 16-QAM, and a rate 1/3 turbo code with variable amount of puncturing. AMC adapts to the radio condition based on the Channel Quality Indicator (CQI) report from the receiver every Transmission Time Interval (TTI) which is fixed at 2ms. For fast packet scheduling and retransmission, HSDPA employs the Hybrid Automatic Repeat-Request (HARQ), which is also dependent on the wireless channel quality Q.

Figure 3.2 depicts the HSDPA scenario considered in this chapter, with the three main network elements involved in HSDPA: Radio Network Controller (RNC), Base Station or NodeB, and the User Equipment (UE). The RNC is responsible for the control of the radio resources. The NodeB schedules the packets to the UEs, taking advantage of AMC and HARQ. These functionalities are embedded in the RNC's and the NodeB's protocol stack as shown in Figure 3.3(a). The packet transmissions over the RNC and the NodeB is illustrated as given in Figure 3.3(b). At the RNC, IP packets are received from the core network and each of them is encapsulated into one Radio Link Control (RLC) Service Data Unit (SDU). The RLC-SDU is then segmented into fixed-size RLC-PDUs of 40 Bytes [4]. At the Node B, one transport block (TB) is sent over the air each TTI. The number of information bits that can be sent in each TB is denoted as the Transport Block Size (TBS) which depends on the CQI of the user. We discuss typical sizes of the TBS as a function of the CQI in Section 3.2.

The process of estimating the TBS is shown in Figure 3.4. In each TTI, one or more users are selected to be scheduled. When multiple users are allowed to be scheduled within the same TTI, the available power and code resources are calculated using a resource-allocation algorithm. When user multiplexing is not used, all the available power and code resources can be allocated to a single user during the TTI. In this thesis, we assume that no user

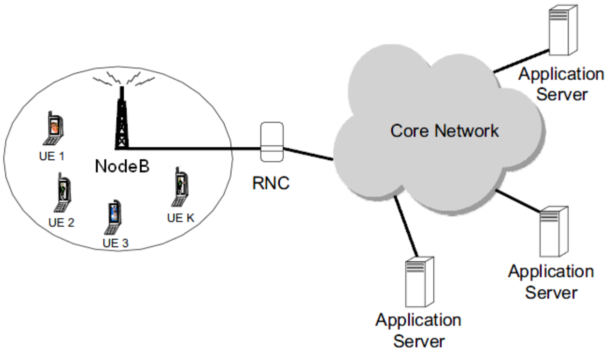
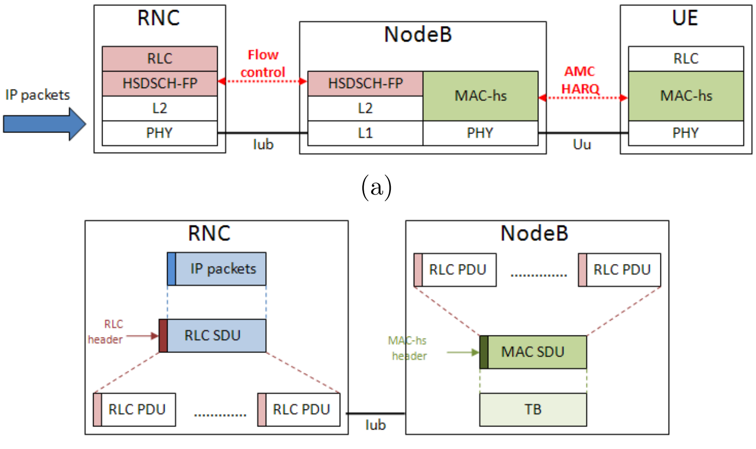


Figure 3.2: HSDPA scenario example. [148]



(a)

(b)

Figure 3.3: HSDPA overview: (a) protocol stack, (b) packet transmissions.

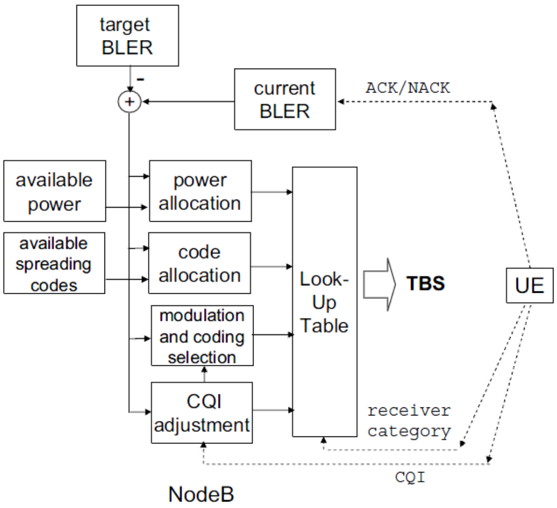


Figure 3.4: TBS estimation process adapted from the standard [5] and extended for the deployed HSDPA system.

multiplexing is used for allocating the HSDPA wireless resources. Look-Up Tables (LUT) are used to get the TBS, given the available power, code and CQI values as standardized by 3GPP in [5]. The UE also sends ACK/NACK messages associated to the previous TB. This helps to estimate the actual Block Error Rate (BLER) of the user. An appropriate TBS is chosen for a target BLER of 10%. The difference between the target and the current BLER is used to update the power, code and CQI values to be used in the LUT.

Let $\mathcal{S}$ be the set of users, $\mathcal{S} = \{1, 2, \cdots, N\}$. Let i^t be the user who is given access to the channel at time t, where t is the index of TTI, $i \in \mathcal{S}$, $t \in \mathcal{Z}_+$ with $\mathcal{Z}_+$ being the set of positive integers. Let $\mathcal{Q}$ be the set of possible CQI, $\mathcal{Q} = \{1, 2, \cdots, 30\}$, and Q_i^t be the CQI of user i at time t, $Q_i^1 \in \mathcal{Q}$.

Using AMC the Node B chooses a transmission format for a fixed target BLER resulting into a TBS of B_i^t which depends on Q_i^{t-d}:

$$B_i^t = g(Q_i^{t-d}) \tag{3.1}$$

where d is the link adaptation delay. The relationship in Eq. (3.1) is standardized by 3GPP [5].

3.2 Radio link-layer model

In this thesis, we adopt the long term radio link-layer model originally proposed in [132], which uses the average maximum achievable bandwidth for each user as a representative of average wireless channel condition experienced by the user. The abstraction model is less complex for the parameter estimation, parameter exchange, and performing the optimization. Hence, it allows a network operator to flexibly place the QoE optimization module anywhere in its network due to its low complexity. Like in [132], we estimate the long term average rate of each user i, denoted by $R_{max,i}$, $i \in \mathcal{S}$, that the user can support when all the wireless resources are allocated to the user. Let R_i be the long term data rate provided to user i, given the normalized resource share α_i. Then the radio-link layer is described as:

$$R_i = \alpha_i \cdot R_{max,i}, 0 \leq \alpha_i \leq 1, \forall i \tag{3.2}$$

Eq. (3.2) defines the HSDPA rate region. In the following the estimation of $R_{max,i}$ is performed for HSDPA. For the analysis, we consider an individual user at each time instance. Hence, we drop the user and the time index. Let r be the instantaneous data rate of the user. Assuming that the scheduler selects only the users who have packets to send, and Q is slowly varying, $r = B$, and from Eq. (3.1) follows:

$$r = g(Q) \tag{3.3}$$

Taking expected values on both sides of Eq. (3.3):

$$E\{r\} = E\{g(Q)\} \tag{3.4}$$

Assuming only one user is scheduled at a TTI, all the resources are allocated to the user, so that

$$E\{r\} = R_{max} \tag{3.5}$$

Due to the user mobility, signal fading from the multipath effect, channel shadowing from urban obstacles, as well as the effect of noise and interferences from external sources, the wireless channel condition Q is time varying, and thus changing the maximum achievable data rate R_{max}. However, if we are interested in the average of wireless channel condition

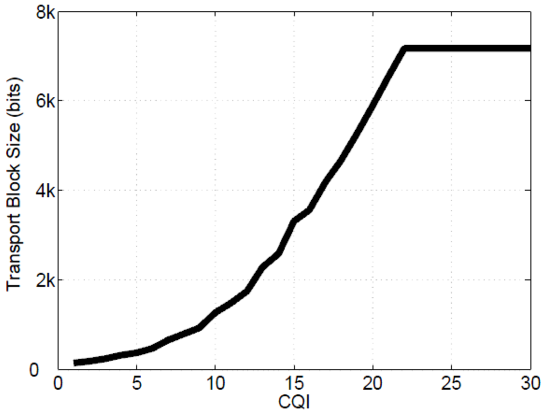


Figure 3.5: TBS vs. CQI for a category 6 receiver [5]. A variable TBS is attained by using adaptive modulation and coding, combined with a variable number of spreading codes. The category 6 receiver can use either QPSK or 16QAM and a maximum of five spreading codes.

within a time interval (e.g. 1 second period) rather than at each 2ms of TTI, it can be assumed that the mean CQI does not change considerably. Therefore, although Eq. (3.3) is a non-linear function when the whole domain of the function is taken into account, it can be approximated by a piecewise linear curve, as shown in Figure 3.5. Hence,

$$E\{g(Q)\} = g(E\{Q\}) = g(\bar{Q}) \tag{3.6}$$

where $E\{Q\} = \bar{Q}$. From Eq. (3.4), (3.5) and (3.6):

$$R_{max} = g(\bar{Q}). \tag{3.7}$$

Hence, R_{max} can be estimated by observing the mean CQI values over a period of time.

3.3 Application layer model

We use application utility functions to describe the Quality of Experience (QoE) for different applications as a function of lower or radio link-layer parameters, e.g., rate or throughput, time share, power, spreading code, bandwidth, etc. As proposed in [87], the Mean Opinion Score (MOS) is used to provide a common numerical measure of the QoE across different applications, and thus allows us to perform a cross-layer optimization for resource allocation across users accessing different types of application. Like in [143], the utility functions used therein are described as a function of transmission data rate and packet loss rate. However, due to the fact that the HSDPA link-layer provides a robust retransmission mechanism, we assume that all packets are transmitted successfully. Hence, in this thesis, the utility function U can be simplified as a function of transmission data rate R as given below:

$$U = f(R), f : \mathcal{R} \rightarrow MOS \tag{3.8}$$

where $\mathcal{R}$ is the set of possible rates, and $MOS = [1 : 4.5]$. As shown in Figure 3.6, MOS 4.5 means that nearly all users would rate the service with an excellent quality, while MOS 1 means the service is expected to be rated by all users with a very poor quality. Below we describe the derivation of the utility functions of different applications and the multiuser utility space in more details.

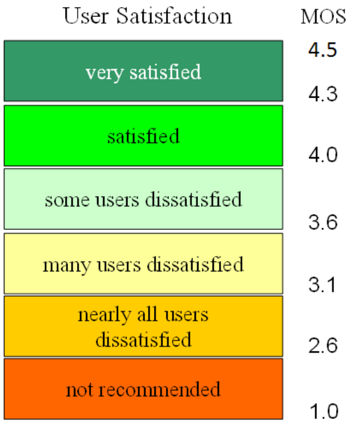


Figure 3.6: Relation between MOS and user satisfaction [73].

3.3.1 Voice call application

End-to-end rate adaptation

Traditionally, assessment of voice quality can be done by performing subjective tests with panels of human listeners, which is usually time consuming and not feasible for the network operator to monitor the delievered voice service to their customers in real-time. The ITU-T has standardized a model, Perceptual Evaluation of Speech Quality (PESQ) [71], which objectively measures and predicts the one-way voice quality score (MOS) that is likely to be given by the user. Still, such algorithms are computationally expensive and require the original speech signal to be compared with the degraded speech. Hence, they are not suitable for online system monitoring and optimization. To solve this we precompute voice utility functions that estimate MOS via the PESQ algorithm as a function of the transmission rate R that determines which voice codec to be used. In Figure 3.7, we show experimental curves for MOS estimation as a function of R. Each point represents a different codec (G.723, iLBC, SPEEX and G.711) used at the sender side. The MOS is measured from a set of speech files with different contents for the case of error-free transmission. Due to distortion imposed by the source codec, every voice codec leads to a different MOS value. This utility curve can be stored at the base station for information when performing QoE-driven optimization for resource allocation.

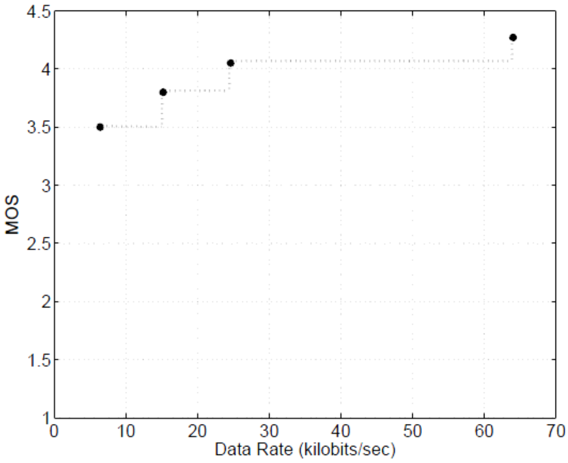


Figure 3.7: PESQ-based MOS as a function of the transmission data rate for different voice codecs. The utility curve consists of 4 discrete points representing the 4 codecs (G.723, iLBC, SPEEX and G.711) operating at fixed bit rates of 6.4kbps, 15.2kbps, 24.6kbps, and 64kbps respectively. [148]

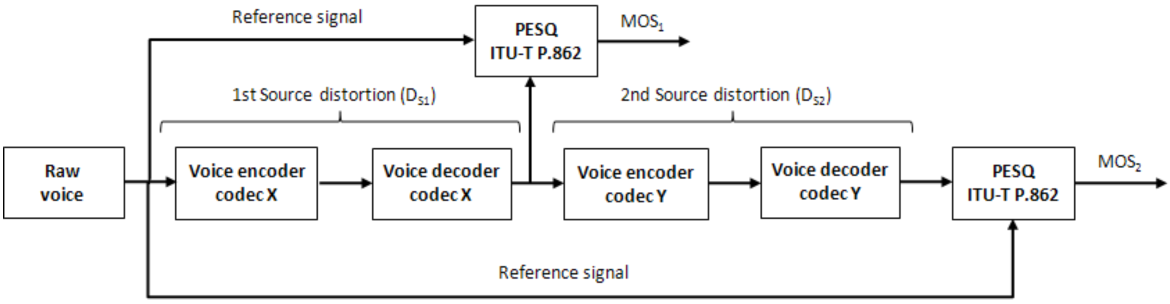


Figure 3.8: Evaluation scheme for voice transcoding using the PESQ algorithm.

In-network rate adaptation

In case the network operator performs a voice transcoding inside its network, there is an additional source distortion (2nd source distortion, D_{S2}) caused by the transcoding process from the codec 'X' to codec 'Y' as shown in Figure 3.8. Whereas, the voice utility functions discussed earlier only considers the encoding distortion at the sender (1st source distortion, D_{S1}), since the rate adaptation is done by changing to a different codec at the sender. For the evaluation and prediction of the transcoded voice quality MOS_2, we employ the PESQ algorithm with an assumption that the reference signal is available.

Figure 3.9 shows the results of different in-network voice transcoding possibilities from a large number of voice samples including both male and female voices available at [71], [115]. The evaluation is done by assuming that there is no packet loss while transmitting from the sender to the network entity responsible for the transcoding. This assumption is

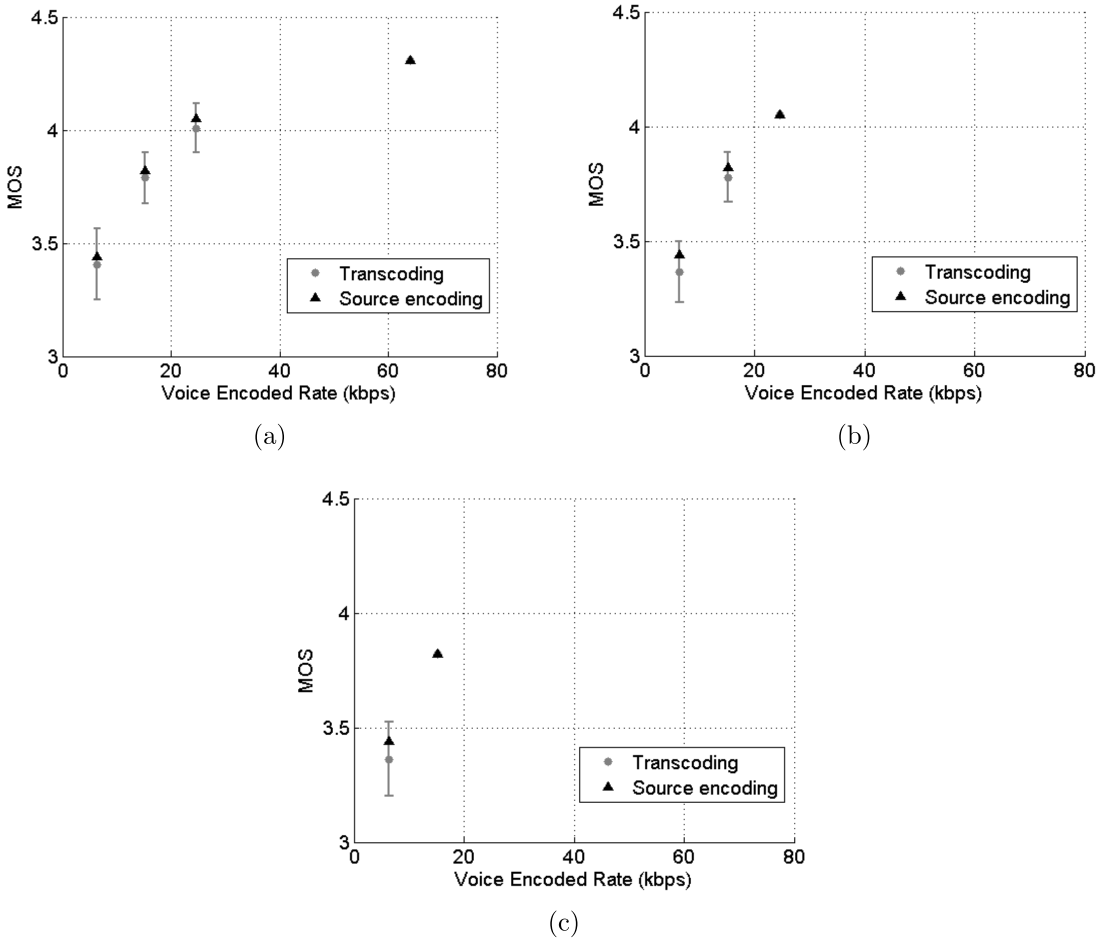


Figure 3.9: Comparison of voice quality from the source encoding distortion and the transcoding distortion: (a) when transcoding from the G711 codec to a lower encoded rate, (b) when transcoding from the SPEEX codec to a lower encoded rate, and (c) when transcoding from the iLBC to a lower encoded rate.

valid only when the sender is a fixed terminal, which trasmits voice packets over the fixed line. Otherwise, if the sender is a wireless terminal, the evaluation scheme in Figure 3.8 is to be updated with an additional packet loss distortion between the 1st and 2nd source distortion. It is to be noted that wireless voice sender is out of scope. In Figure 3.9-a, the sender transmits the voice content with a high data rate of 64kbps and the intermediate node in the network transcodes the voice stream to a different codec so as to achieve a lower data rate. We observe that the previous model marked with the black triangle fits well to the average transcoded voice quality. Also, the lower encoding rate we transcode, the higher standard deviation of the transcoded voice quality we receive. These conclusions also apply for the cases of the SPEEX transcoding and the iLBC transcoding to a lower encoding rate as shown in the Figure 3.9-b and Figure 3.9-c respectively. From all results, we see that the maximum of the standard deviation of the transcoded voice quality is roughly about 0.15 MOS, which is marginal. Hence, the average MOS model, which only takes into account the source encoding at the sender, can be used as a general model for any voice stream and the additional distortion caused by the transcoding is negligible. If the network operator would like to optimize its network resource allocation for voice call applications, it is sufficient to use an average model of voice utility function derived from the source encoding with different codecs. These models just have to be stored in advance in the network and no extra signalling from the end-points during the mid-session is needed.

3.3.2 File download application

File download or web-based applications are considered to be elastic services, for which the utility function is an increasing, strictly concave, and continuously differentiable function of throughput [86]. Based on this assumption, the file download utility function is assumed to be logarithmic with respect to the data rate R as follows:

$$MOS = a \cdot log_{10}\left(b \cdot R\right) \tag{3.9}$$

where a and b parameters are determined from the maximum and minimum user perceived quality and the user's service subscription to the network operator. For example, if a user has subscribed for a specific rate service R and receives this service rate R when downloading the file, then in case of no packet loss user satisfaction on the MOS scale should be maximum, i.e., 4.5. On the other hand, we define minimum transmission rate (e.g. 10kbps in Figure 3.10) and assign to it a MOS value of 1. Using these parameters, we fit the logarithmic curve in Eq. (3.9) for the estimated MOS. Figure 3.10 presents the MOS function by varying the R based on the assumption that the user subscribes to a service of 200kbps data rate. To enable the cross-layer optimization across flows discussed later, a signalling mechanism for a and b parameters is needed. This can be done in an end-to-end fashion or through the retrieval of user's subscription information from the subscription database in the network.

In fact, file download or web browsing applications use TCP as its transport protocol, which has its own end-to-end mechanisms between the sender and the receiver such as flow control and congestion control. The flow control adapts the sending data rate in order to prevent a fast sender from overrunning a slow receiver, while the congestion control keeps the data flow below a rate that would trigger a network congestion, which makes the network performance fall. In order to adapt the transmission rate to optimize the transmission in a base station, the cross-layer optimizer may contact the sender or simply, e.g., slow down the TCP flow. If this happens on a small time scale (e.g., seconds) TCP will not notice. If this situation pertains then TCP will react accordingly by adapting its sending rate. We assume that the TCP rate adaptation process, which is for example modeled by the TCP Friendly Rate Control (TFRC) equation [46], has no significant impact on the

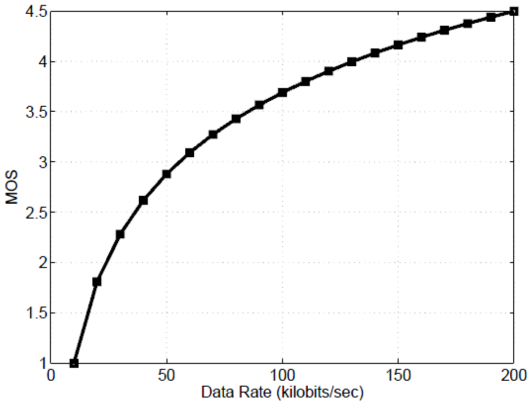


Figure 3.10: MOS as a function of transmission rate for file download applications. [148]

user perceived throughput (as shown in Eq. (3.9)). Alternatively, the proposed cross-layer optimization framework may leave the TCP connections untouched and only relies on the TCP end-to-end mechanisms as discussed above.

The above TCP-based application utility function might be different, if a TCP proxy-based solution is used for a TCP connection. For example, the wireless network operator may implement the Wireless TCP (WTCP) mechanism [126] in one of their network entities, which splits the TCP connection into two TCP connections, one from the sender to the TCP proxy and another from the TCP proxy to the mobile client, and thus gives flexibility to control the latter connection according to the wireless channel condition experienced by the mobile client. The works presented in this thesis do not cover such advanced TCP transport protocol. The file download utility function of the TCP proxy-based solution and its potential impact are out of scope and require hence further study.

3.3.3 Video streaming application

End-to-end rate adaptation

The schema for deriving the video utility function can be illustrated as in Figure 3.11. In this diagram, we only have the source distortion D stemmed from the source's encoding and the receiver's decoding process. The video distortion is calculated by the Mean Square Error (MSE) between the two $m \times n$ images A and B, where one of them is considered as a degraded image of the other. D is defined as follows:

$$D = \frac{1}{m \cdot n} \sum_{i=1}^{m} \sum_{j=1}^{n} [A(i,j) - B(i,j)]^2 \tag{3.10}$$

To see the relationship of D and the video encoded rate R, we change the setting of the quantization parameter for encoding the I-, P- and B-frames, and measure the corresponding source distortion as shown in Figure 3.12. Two different video contents

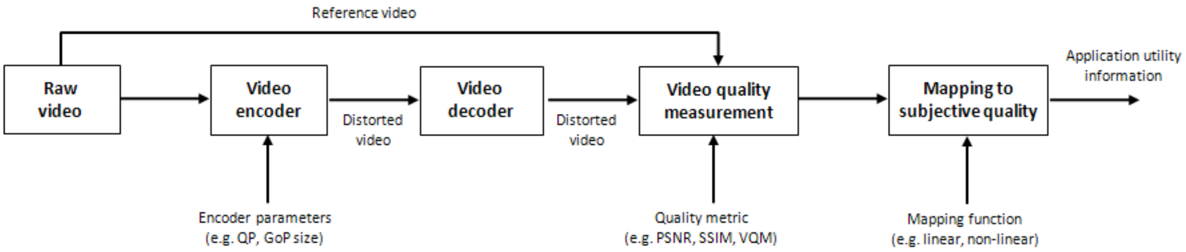

Reference video
Raw video
Video encoder
Distorted video
Video decoder
Distorted video
Video quality measurement
Mapping to subjective quality
Application utility information
Encoder parameters (e.g. QP, GoP size)
Quality metric (e.g. PSNR, SSIM, VQM)
Mapping function (e.g. linear, non-linear)

Figure 3.11: Schema of video utility derivation for an end-to-end rate adaptation.

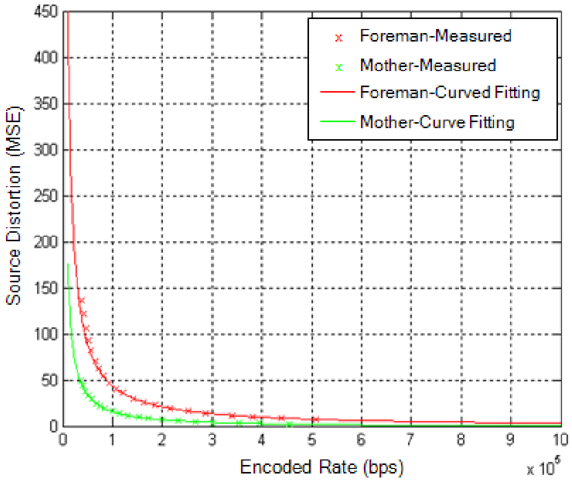


Figure 3.12: Measurement result and model for video source distortion.

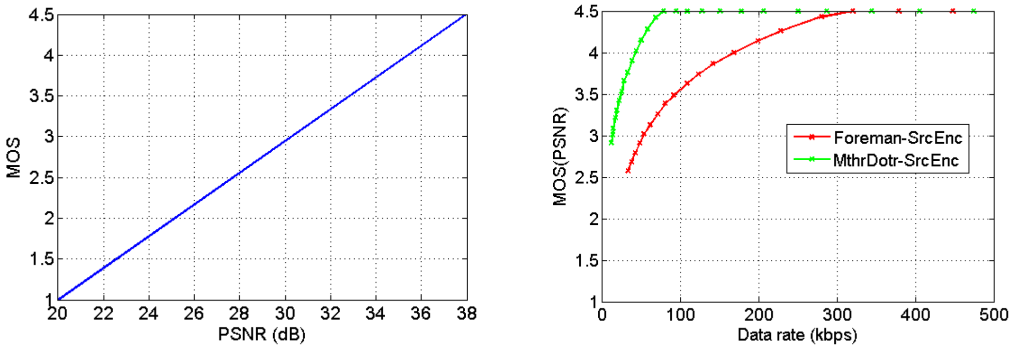
(Mother&Daughter (Mother) and Foreman) are used as an example. Both videos have QCIF resolution and are encoded with the H.264 AVC at 30 frames/sec. The Group-of-Pictures (GoP) size is set to 30 frames. In [32], an analytical model of source distortion is proposed, which requires only three reference points of source distortion from three different encoding rates to determine the two constant parameters χ and δ in Eq. (3.11).

$$D = \frac{\chi}{e^{R/\delta} - 1} \tag{3.11}$$

As shown in the measurement results in Figure 3.12, we see that the source distortion model proposed in Eq. (3.11) fits well with the experimental measurements for both video sequences. Moreover, the Foreman video, which contains a dynamic video content, is more sensitive to the change of encoded rate than the Mother video.

Alternative to the MSE, the Peak Signal to Noise Ratio (PSNR) is also used to measure the video quality, which is calculated from the source distortion D by using a logarithmic function as follows:

$$PSNR = 10 \cdot log_{10} \frac{255^2}{D} \tag{3.12}$$



(a)

(b)

Figure 3.13: Linear relationship between PSNR and MOS (a) Video utility functions based on the PSNR quality measure (b).

To get the MOS model for the video application, we use a linear mapping between the PSNR and the MOS as given below.

$$MOS = \upsilon \cdot PSNR + \varsigma \tag{3.13}$$

The two constant parameters υ and ς in Eq. (3.13) can be determined by specifying a minimum $PSNR_{min}$ and a maximum $PSNR_{max}$ including its corresponding minimal and maximal user-perceived quality MOS_{min} and MOS_{max} respectively. For example, if the PSNR is less than 20dB then the MOS will be 1, and if the PSNR is more than 38dB then the MOS will be 4.5. Figure 3.13(a) shows the linear mapping function of the given example. We apply this linear mapping for the rest of this chapter.

From Eq. (3.12) and (3.13), the video utility function for both 'Foreman' and 'MthrDotr' video sequences can be depicted as shown in Figure 3.13(b). We see that a video user will be more sensitive to a video quality change of the 'Foreman' video than a quality change of the 'MthrDotr' video. For the 'MthrDotr' video, when the encoded rate is about 90kbps, the user-perceived quality already reaches the maximum level of MOS 4.5. Further increasing the encoding rate for the 'MthrDotr' video will not lead to a higher QoE. This tells us that there is a range of video quality, in which it makes sense and makes no sense to increase the encoded rate in order to achieve a higher quality perceived by the user. In contrast, the 'Foreman' video requires much more data rate in order to reach the highest MOS of 4.5. This concludes that the user perceived quality is strongly dependent on the video content and one should encode the video appropriately in order to avoid network congestion and efficiently use the limited network resources if transmitted over the wireless networks.

In-network rate adaptation

The video utility function investigated so far only considers the rate adaptation at the server (e.g. changing the encoded rate at the video source). However, a network operator may prefer to perform a rate adaptation in its network, so that it has a control for changing the video data rate of all traffics in its network, and thus reacting properly to the network congestion or to a poor wireless channel condition experienced by the user. Furthermore, having such control also gives a flexibility to the network operator to adapt the data rate across flows, and not only on a single flow basis. The downside of doing a rate adaptation in the network is an additional distortion to the video quality, which is illustrated in the schematic diagram in Figure 3.14. Below, we discuss two in-network rate adaptation techniques: transcoding and packet dropping, and investigate its impact on the video utility function.

- *Transcoding*: Video transcoding can actually be done in several ways [173, 8]. The simplest method of transcoding is to decode the already encoded video stream to an intermediate format (i.e., YUV for video), and re-encode the resulting file into the target format [106] with the target data rate. Alternative is the sample rate conversion approach, which converts the digital signal from one sampling rate to another [156]. While converting, it also minimizes the change of information carried in the signal. Throughout this thesis, we use the simple re-encoding scheme as an example of transcoding due to its simplicity of implementation for evaluation. After performing a transcoding, we measure the average data rate and the video quality, which is done by comparing the transcoded video with the original video.

- *Packet dropping*: In H.264 AVC codec, a video frame can be encoded in three different frame types: I-frame, P-frame, and B-frame. The Intra coded picture or I-frame is the conventional, full size frame that contains all contents in the original frame. The

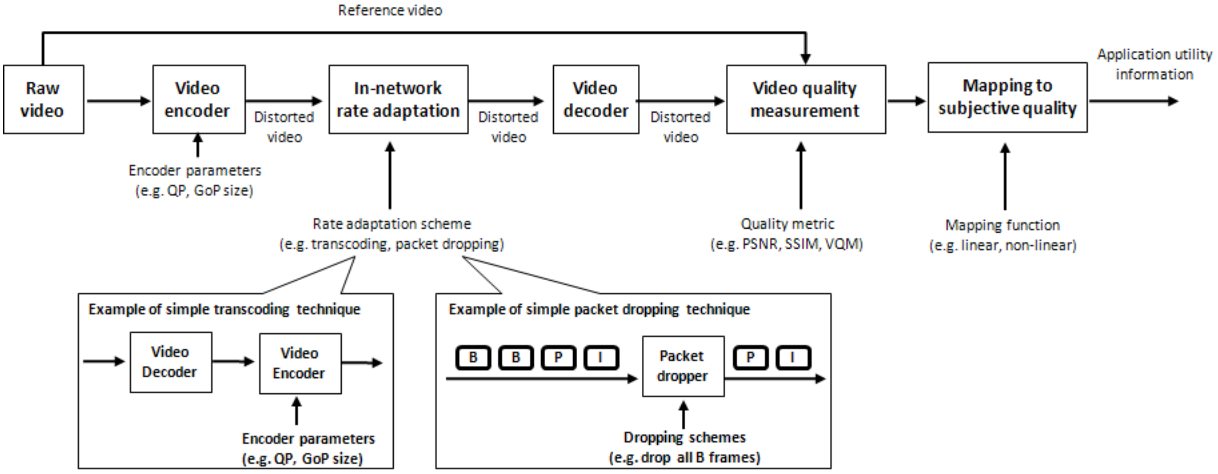
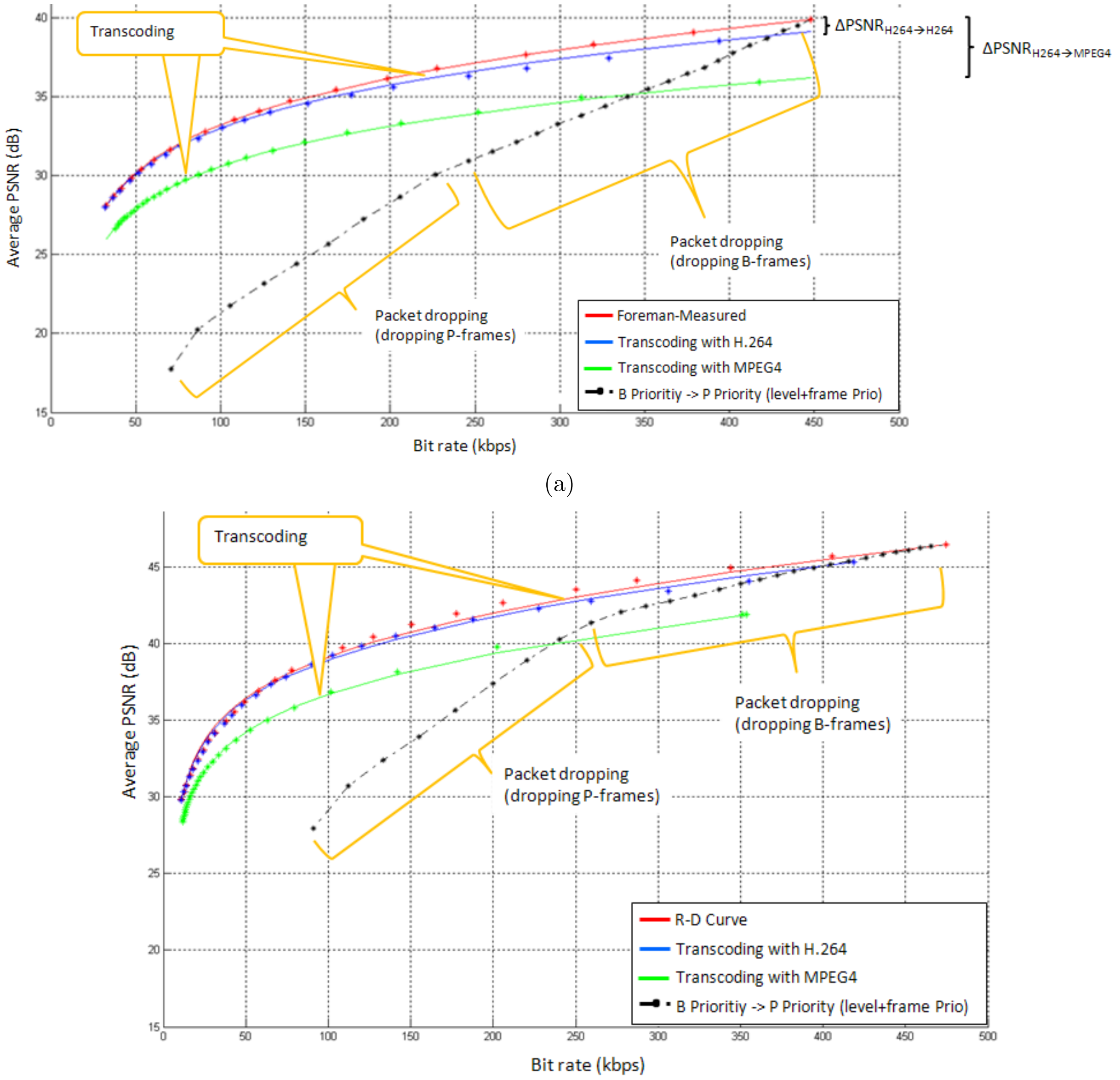

Reference video
Raw video
Video encoder
Distorted video
In-network rate adaptation
Distorted video
Video decoder
Distorted video
Video quality measurement
Mapping to subjective quality
Application utility information
Encoder parameters (e.g. QP, GoP size)
Rate adaptation scheme (e.g. transcoding, packet dropping)
Quality metric (e.g. PSNR, SSIM, VQM)
Mapping function (e.g. linear, non-linear)
Example of simple transcoding technique
Video Decoder
Video Encoder
Encoder parameters (e.g. QP, GoP size)
Example of simple packet dropping technique
B
B
P
I
Packet dropper
P
I
Dropping schemes (e.g. drop all B frames)

Figure 3.14: Schema of video utility derivation for an in-network rate adaptation..

predicted picture or P-frame is a delta frame, which stores only changes in respect to the previous I- or P-frame. And the bi-directional predicted picture or B-frame stores the data with reference to both the previous and precedent frames. Both P-frame and B-frame are used in order to improve the video compression efficiency. With these three frame types, the video can be encoded with any encoding schemes such as I-P-P scheme, or I-B-P scheme and thus allows a hierarchical frame encoding that tells about the importance of each frame type. In this case, the I-frame is the most important frame, as it is the head-end frame that are used as a reference for both P- and B- frames. The second most important frame is the P-frame, as the B-frame is dependent on it. And the least important one is the B-frame, since they are not referred by any other frames. To achieve rate adaptation, one can simply drop a packet or the whole frame from the video stream. However, this should be done carefully in order to minimize the overall quality degradation that may cause by the packet dropping due to the structure of frame dependencies when encoding the raw video. For instance, in our work, the video is encoded with the I-P-B-B-P structure. The typical way for frame dropping is to first drop all B frames. If further rate reduction is still needed, then we start dropping P-frames, until there are no P-frames left. As a last possibility, we then have to drop an I-frame causing a still image for the whole period of the Group of Pictures (GoP). For P-frame dropping in a single GoP, we start dropping from the last P-frame instead of dropping the first P-frame in the GoP or a random dropping to avoid a distortion propagation as explained in [155].

Figure 3.15(a) shows the Rate-Distortion (RD) curve for the Foreman video when applying transcoding and packet dropping for adapting video data rate. The upper bound curve is the source-encoding distortion caused by the H.264 AVC codec at the sender. Each red points represents the actual average video rate and the video quality that we measure after varying the quantization parameter at the sender (video streaming server). The red line is the RD model based on the function proposed by Choi *et al.* in [32]. Prior to performing a transcoding or a packet dropping, we assume that the sender transmits a video with high data rate (e.g. at 450kbps) to the receiver. In the network, we transcode the high bit rate video stream to a lower rate by decoding and re-encoding with the same H.264 codec but different quantization parameter. The RD measurement for the "H.264 ⇒ H.264" transcoding is shown by the blue dots and the blue line from Choi's RD model. In case, the mobile terminal does support only the MPEG4 codec and not the H.264 AVC codec, the network operator may transcode the original H.264 AVC video stream to the MPEG4 codec format. The result of an additional distortion from the MPEG4 transcoding is shown with the green dots and green line. We see that transcoding with the same codec (H.264 AVC ⇒ H.264 AVC) does not deviate much from the original source encoding RD curve, whereas transcoding to a different codec (e.g. H.264 AVC ⇒ MPEG4) would lead to a stronger degradation of the video quality.

The black dots in Figure 3.15(a) show the measurement results of the video quality when applying an intelligent packet dropping scheme based on the importance of different video



(a)

(b)

Figure 3.15: Rate-Distortion curve for Foreman video (a) and Mthr&Dotr video (b) when applying different rate adaptation techniques: source encoding (red line), transcoding (blue and green lines) and packet dropping scheme (black line).

frame types as discussed earlier. In this result, we use the frame copy for an error concealment in case of a lost frame. We see that dropping the B-frame causes less quality degradation than dropping the P-frame. Another observation is when dropping each B frame, the change of data rate and its corresponding video quality (PSNR) is quite similar. This means a linear function can be used to model the experimental result of dropping B frames. The RD curve for dropping each P frame can also be modelled by using a linear

Table 3.1: Statistical results of transcoding 25 video sequences

Transcoding scheme	Avg. ΔPSNR	Std. ΔPSNR	Max. ΔPSNR	Max. ΔMOS
H.264 ⇒ H.264	0.747	0.307	1.054	0.247
H.264 ⇒ MPEG4	3.205	1.06	4.211	0.99

function, but with different constant parameters.

In Figure 3.15(b), we show how the transcoding and packet dropping have an impact on the video quality for the Mother&Daughter video, which contains a low motion of video scenes. Like the Foreman video, we see that the transcoding outperforms the packet dropping. However, the reduction rate of video quality for the Mother&Daughter video is much less than the Foreman video when applying both rate adaptation schemes. This implies that a dynamic video is more sensitive to any rate adaptation scheme than the static video. Thus, the video content plays an important role and should be taken into account when performing rate adaptation in the network.

In addition to those two videos, we have done an extensive evaluation of video transcoding with 25 different video contents. Table 3.1 shows the statistical results from this experiment. In short summary, the maximum PSNR difference for the H.264 AVC ⇒ H.264 AVC transcoding ($\Delta PSNR_{H264 \Rightarrow H264}$) is about 1.054 dB, whereas the maximum PSNR difference for the H.264 AVC ⇒ MPEG4 transcoding ($\Delta PSNR_{H264 \Rightarrow MPEG4}$) is much higher (about 4.211 dB). When using the linear relationship between the PSNR and the MOS as presented in Figure 3.13(a), these can be translated to the maximum of MOS difference of 0.247 and 0.99 for the H.264 AVC ⇒ H.264 AVC transcoding and the H.264 AVC ⇒ MPEG4 transcoding, respectively.

Figure 3.16 depicts the video utility for 'Foreman' and 'Mother&Daughter' videos based on the linear mapping between the PSNR and the MOS as given in Eq. (3.13). In this Figure, we only depict the H.264 AVC ⇒ H.264 AVC transcoding, as we use it for the rest of this thesis and not the H.264 AVC ⇒ MPEG4 transcoding scheme.

HVS-based video utility function

Previously, we measured the video quality by using the PSNR due to its simplicity of calculation. However, many studies [54],[161],[163] show that such pixel-based distortion measure does not match well to user perceived visual quality due to the fact that human eyes are highly adapted to structural information. Structural SIMilarity (SSIM) index [163] was first used to measure an image quality based on the structural distortion. In principle, SSIM measures the similarity of two signals (the original signal and the distorted signal) by comparing the luminance, the contrast and the structure. The luminance is the mean intensity from the signal. The contrast is the standard deviation of the signal. The structure is the signal after luminance subtraction and variance normalization. These two signals

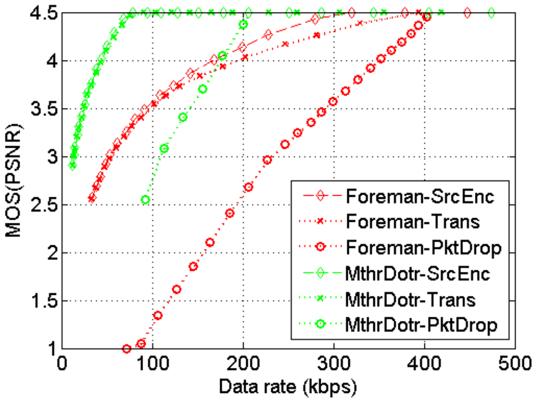


Figure 3.16: PSNR-based video utility functions for different rate adaptation schemes.

are taken from a local window, which is just a part of the whole image. In [164], Wang *et al.* extended the image SSIM to support video applications. The Video SSIM (VSSIM) considers the chrominance part of the image and employs some adjustment methods that assign different importances for different regions in a frame and for different frames in the video sequence. The adjustment improves the accuracy of the quality assessment algorithm due to two reasons. First, dark regions in a frame are assigned with smaller weighting values, since they do not attract fixations. Second, human eyes perceive the video quality differently in each frame depending on the degree of motion in the video sequence, which can be measured by using the motion vector information during the encoding process.

To obtain the video utility functions, we vary the quantization steps for encoding the raw video and measure the average data rate and the average VSSIM index. Transformation of the objective video quality (e.g. VSSIM) to the predicted user perception in quality degradation (DMOS) can be done in several ways. For example, the Video Quality Expert Group (VQEG) in ITU recommended to use a nonlinear regression function as shown in Figure 3.17. DMOS 0 means that the user does not see the quality degradation compared to the perfect video quality. Whereas, the higher DMOS refers to the lower video quality that the user would rate. Since all test video sequences that are used in the VQEG Phase I test [158] are not available publicly, for simplicity, we map the VSSIM value to the MOS scale using a linear function with an upper and lower bound as follows:

$$MOS = \begin{cases} 1, \text{if } VSSIM < 0.74 \\ a \cdot VSSIM + b, \text{if } 0.74 \leq VSSIM \leq 0.98 \\ 4.5, \text{if } VSSIM > 0.98 \end{cases} \tag{3.14}$$

Note that the upper and lower bound, and the constant parameters (a and b) of the linear function are determined such that it fits best to the scatter plot of objective and subjective measurements as depicted in Figure 3.17. Figure 3.18 depicts an example of a VSSIM-based video utility curve for the 'Foreman' and the 'Mother&Daughter' video

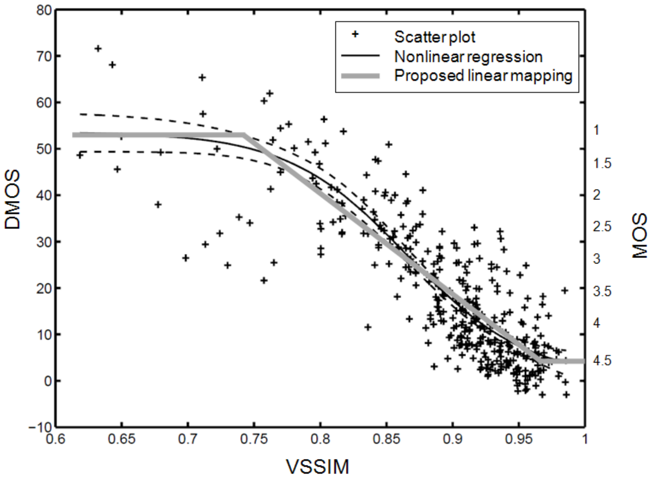


Figure 3.17: Scatter plot and linear/non-linear regression of the VSSIM-based video quality assessment model on VQEG Phase I test dataset. [164]

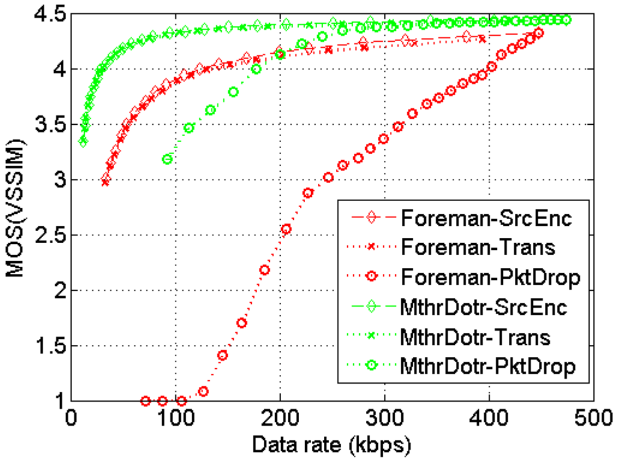


Figure 3.18: VSSIM-based video utility functions for different rate adaptation schemes.

sequences. In contrast to the PSNR-based video utility, the video quality will not degrade that much when transcoding the video. Whereas, for the packet dropping scheme, user-perceived quality will reach a minimum MOS of 1 earlier than the PSNR-based result for the 'Foreman' video. But for the 'Mthr&Dotr' video, the MOS will be higher even though all B- and P-frames are dropped in every GoP. This implies that human eyes are more sensitive for the dynamic video content than the static video content when performing a packet dropping for rate adaptation.

Complexity of in-network video rate adaptation

Complexity of performing a video transcoding or dropping video packets are different and has to be considered when performing rate adaptation in the network due to hardware constraints. To measure complexity, we have performed experiments by, for example, measuring the processing time (in second) needed to perform transcoding and packet dropping for the 'Foreman' and the 'MthrDotr' videos as depicted in Figure 3.19. In this experimental measurement, the computer used for the transcoding and packet dropping only has a CPU of Pentium 4 processor and a RAM of 750 Mbytes. If using a PC with better performance, it is expected that the processing time will be significantly reduced. Results show that transcoding is computationally more expensive than packet dropping. The time for transcoding ranges between 1 second and 4.2 seconds depending on the target data rate and the video content itself. Whereas, packet dropping only requires a 100 ms, and the required processing time is independent of the video content and the target data rate. In case, there is a computational constraint of hardware (e.g. processor), an intelligent selection of which video stream to be applied the transcoding is necessary, since video transcoding causes less video quality degradation than packet dropping. We discuss in details our proposed rate adaptation scheme selection across multiple videos in Section 3.9.

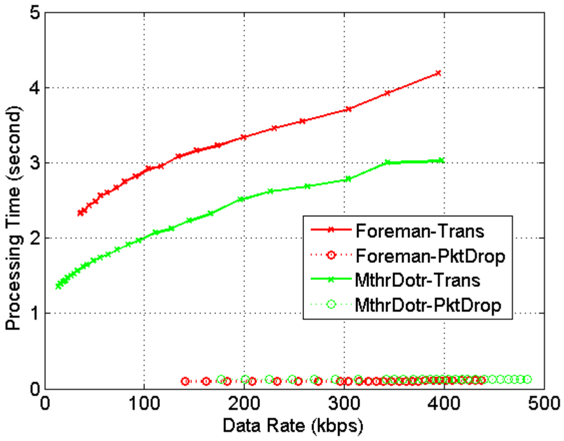


Figure 3.19: Experimental measurement of time consumption for different rate shaping schemes.

3.4 Multiuser utility space

The multiuser utility space, $\mathcal{U}$, defines a set of feasible utility vectors constrained by the total system resources:

$$\mathcal{U} \subseteq R^n, \sum_i \alpha_i \leq 1, \tag{3.15}$$

where R^n is the n dimensional Euclidean Space and α_i is a normalized resource share to user i. Considering a simple wireless access network system with a total symbol rate S^*, we can formulate a resource allocation constraint across users as $\sum_N^{i=1} S_i \leq S^*$, where S_i is the symbol rate assigned to user i. α_i, which is fraction of total symbol rate assigned to user i, can be calculated as $\alpha_i = S_i/S^*$.

In Figure 3.20, we show the utility functions U of video and file download applications for different channel conditions given S^*=100Ksymbol/sec. Each curve corresponds to the bound on the QoE of the user as a function of the given resource for a particular long-term receiver Signal-to-Noise Ratio (SNR) experienced by the user. We note that the utility functions are monotonically increasing with respect to α. In general, U is non-concave and non-differentiable. A multiuser utility space $\mathcal{U}$ can be formed by combining the transmission policies of every user under a certain wireless channel condition and a constraint of total resources shared among users. Figure 3.21(a) shows an example of $\mathcal{U}$ for a two user case: one video user and one FTP user. The receiver SNR for the two users are 15dB and 5dB, respectively. The individual points correspond to all possible combinations of (α_1,α_2) such that $\alpha_1 + \alpha_2 \leq 1$. Figure 3.21(b) shows the boundary of the $\mathcal{U}$. We denote this boundary as $U_{15,5}$, the subscripts specifying the channel state of the users. *a*–*b* and *d*–*e* correspond to user rates of $(0, R_{max,1})$ and $(0, R_{max,2})$. *c* is the optimum point with respect to the objective function described in the next section.

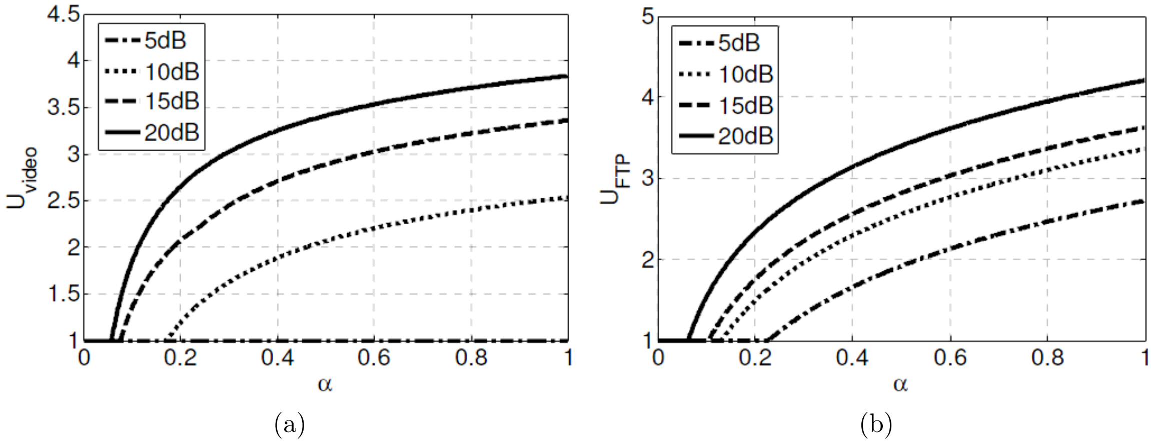
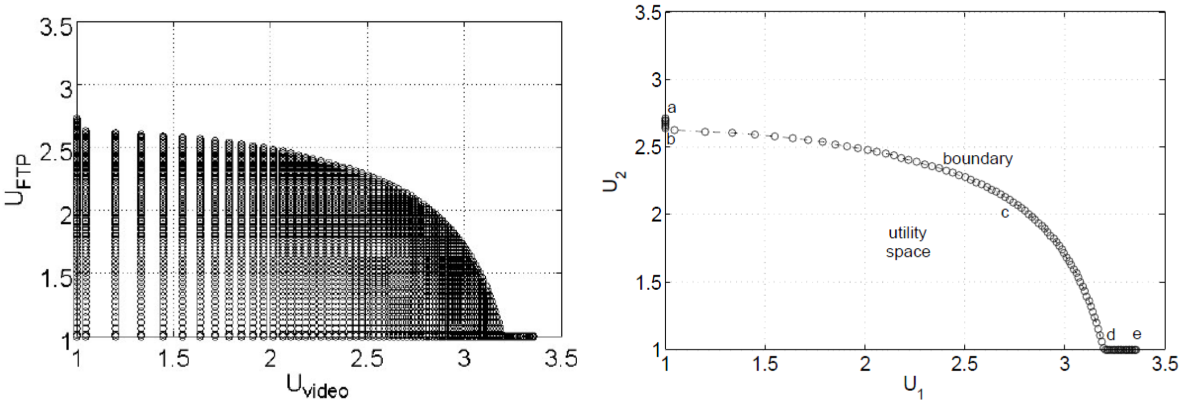


Figure 3.20: Utility function for different channel conditions: (a) video streaming application and (b) file download application.



(a) (b)

Figure 3.21: Utility space (a) and boundary of utility space (b) for the two user case; SNR(video) = 15dB, SNR(FTP) = 5dB. [89]

3.5 QoE-driven CLO framework

In this section, we discuss a Quality of Experience (QoE) driven optimization framework for resource allocation in HSDPA [1]. The framework is integrated into a HSDPA mobile system as shown in Figure 3.22. The long-term link layer and the application layer models discussed in previous sections are communicated to a QoE-based optimizer acting as a downlink resource allocator. For example, we use the maximum achievable data rate R_{max} for each user i when assuming that the total resources are allocated to the user as a key parameter from the link-layer, while in the application-layer we use the utility function U representing the user's QoE for different applications. Depending on the objective function that is set prior to the optimization, the optimizer finds an optimal resource allocation α_{opt} and then sets the applications-layer data rate R_{opt} for each user accordingly. In the following subsections, we discuss different objective functions and how the rate adaptation done by the application-layer can be realized at the HSDPA base station.

3.5.1 Utility-based objective functions

The utility functions introduced in Section 3.3 provide the information about the required transmission rate at the application-layer in order to achieve a certain level of QoE. The representation of the lower-layers depends on the channel quality of each user. Information about the channel quality is obtained through the CQI feedback from the UE as described in Section 3.2. Depending on the selected objective function, the optimizer allocates the wireless system resources differently. Below we discuss two QoE-based objective functions applied in our work.

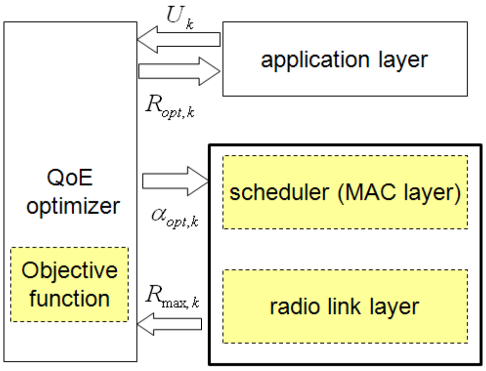


Figure 3.22: QoE-driven optimization framework for HSDPA

Utility maximization

With the objective function of utility maximization, which has been first proposed in [79, 32], the optimizer maximizes the average utility of all users as given in Eq. (3.16):

$$\mathcal{F}(\tilde{\mathbf{x}}) = \frac{1}{N} \cdot \sum_{i=1}^{N} U_i(\tilde{\mathbf{x}}) \tag{3.16}$$

where $\mathcal{F}(\tilde{\mathbf{x}})$ is the objective function with the cross-layer parameter tuple $\tilde{\mathbf{x}} \in \tilde{\mathbf{X}}$. N is the total number of users in the system, $\tilde{\mathbf{X}}$ is the set of possible parameter tuples abstracted from the protocol layers representing the set of candidate operation modes. The decision of the optimizer can be expressed as:

$$\tilde{\mathbf{x}}_{opt} = \arg\max_{\tilde{\mathbf{x}} \in \tilde{\mathbf{X}}} \mathcal{F}(\tilde{\mathbf{x}}) \tag{3.17}$$

where $\tilde{\mathbf{x}}_{opt}$ is the parameter tuple which maximizes the objective function. After selection of the optimal values of the parameters, those parameters are sent back to the individual layers, which are responsible for translating them back into actual layer-specific modes of operation. Further details of parameter abstraction can be found in [79], [33] and [88].

Depending on the type of application, we create different sets of transmission policies, which specify possible transmission data rates. We denote the set of transmission policies for a user i by T_i. With utility-based optimization, the optimizer chooses a combination of resource allocation that maximizes the following objective function:

$$\mathcal{F}(\tilde{\mathbf{x}}) = \sum_{i=1}^{N} \sum_{j=1}^{|T_i|} E\{\mathcal{I}_{ij} \cdot U_{ij}(\tilde{\mathbf{x}})\} \tag{3.18}$$

where i denotes the user index, j refers to the index of the transmission policy. $\mathcal{I}_{ij}$ is the indicator function. Its value is 1 when the transmission policy j is chosen for user i, and 0 otherwise.

In principle, with the utility maximization, all users are served with very high quality, if all are experiencing a good channel condition. Otherwise, in constrained systems, the optimizer will give less network resources to the user having a very bad channel condition or the user accessing a high-demand application.

Max-min utility

The max-min fairness concept [123] applied to our QoE-based cross layer optimization means that the optimizer allocates the resources such that all users experience the maximally possible same level of quality. The max-min objective function is defined as:

$$\tilde{\mathbf{x}}_{opt} = \arg \max_{\tilde{\mathbf{x}} \in \tilde{\mathbf{X}}} \left\{ \min_{i \in N} U_i(\tilde{\mathbf{x}}) \right\} \tag{3.19}$$

A drawback of using max-min fairness is the unequal quality loss. For instance, when a single user runs a very demanding application or has a very poor channel quality, the optimizer tries to give this user more resources and therefore forces all other users to share this poor experience fairness. A modified max-min technique [131] has been proposed to allow for setting a minimum guarantee of service quality. It first checks whether there is enough resources to provide all users with that guaranteed quality. If not, the system will drop the user with the highest resource consumption, meaning that no resources are given to this user until the next optimization loop. After checking the constraint, it performs a usual max-min utility based optimization as described in Eq. (3.19).

3.5.2 Realization of rate adaptation in HSDPA

At each TTI, a number of data blocks or RLC PDUs are passed from the higher layers to the radio link layer. The size of a data block to be transmitted in one TTI depends on the Channel Quality Indicator (CQI), which is carried via the uplink High Speed-Dedicated Physical Control Channel (HS-DPCCH). The TTI is set to 2ms, meaning that there are 500 TTI slots available in one second period to be shared among users.

The optimizer decides the best combination of all user's operation modes, which maximizes the selected objective function. To assure the data rate of each user, the number of TTI slots must be assigned correctly. Estimation of the required number of TTI is done by using the following equation:

$$S_i = \lceil \frac{A_{app,i} + OH_i}{\bar{B}_i} \rceil \tag{3.20}$$

where S_i is the number of transmission opportunities to be allocated to user i. $A_{app,i}$ is the number of bits to be sent in one second. We assume that the application is sending with a constant bit rate (CBR) during the time interval of interest. $\bar{B}_i$ is the mean size of a transport block. OH_i is the amount of overhead due to transport and network layer headers.

The use of the proposed framework does not exclude the possibility of setting Guaranteed Bit Rate (GBR), Scheduling and Priority Indicator (SPI) and Discard Timer (DT) for quality control, as proposed in [117]. GBR can be set at once as the values out of the optimization or periodically reconfigured during optimization. Setting SPI would be essential in order to ensure delay guarantees. In this paper we assume that the streaming and realtime traffic are prioritized with respect to file download traffic. The exact priority indices would largely depend on the scheduler used. The approach taken in this paper does not rely on any particular scheduling scheme, and hence can be used with any scheduler.

In order to work harmoniously with an admission control policy, it should be considered that for a relatively small number of users all the resources are not exhaustively distributed to the existing users. Otherwise, as each new user is admitted into the system, the existing users would be forced to lower their share of resources, resulting into lower quality and unsatisfied users. It should be noted that for audio-visual services users would usually prefer to keep the quality level fairly constant rather than being exposed to fluctuations of quality, even if the mean quality of the latter is higher. We propose an extension of QoE-driven optimization that addresses the issue of temporal quality fluctuation in Section 4.2.

3.6 Greedy search optimization

Greedy algorithm was first devised by Gross [57] for solving general discrete resource allocation problem, in which the objective is to minimize a separable convex function under a single budget constraint. The resource allocation problem is extended by considering upper bounds [42] or both upper and lower bounds [61]. In [61], Hochbaum discusses an application of greedy algorithm for a continuous nonlinear variable. A comprehensive review of algorithmic approaches including greedy algorithm for discrete and continuous resource allocation problem are provided in [67] and [116] respectively. In general, a greedy algorithm makes a locally optimal choice at each step with the hope of finding the global optimum, and therefore, cannot guaranteed to find the optimal solution since it does not operate exhaustively on the whole constraint space. Due to its low complexity, greedy algorithm has also been applied for other problems such as the knapsack problem [25], the outsourcing warranty repair service problem [111], etc.

In telecommunications, greedy algorithm has also been used to find an optimal solution for network resource allocation problem, for example, to minimize a total transmit power across multiple wireless users [169], or to minimize video quality degradation of all users in multipath networks [81]. It is to be noted that most works of greedy algorithm in resource allocation problem usually start its greedy algorithm with zero resource allocation. Brehmer *et al.* [24] propose to initialize a resource allocation for multiple users that lies along the Pareto efficient set. They analytically illustrate how the proposed Iterative Efficient set Approximation (IEA) develops step-by-step along the set of possible resource allocations that satisfy the equality of the constraint (Pareto efficient set) with an aim of maximizing the sum of utilities of all users. In principle, the concept of IEA algorithm is similar to the greedy algorithm. Namely, both IEA and greedy algorithms find the next best possible operating point of resource allocation by taking a small amount of resources from the user with the minimum quality degradation and assigning them to the user receiving the maximum benefit. To achieve this, the IEA algorithm uses the gradient projection method on the tangent space of the current operating point, whereas, the greedy algorithm uses a max-gain and min-loss ranking method across all users.

In this thesis, we use the greedy search algorithm to solve a discrete resource allocation problem and consider a linear budget constraint, which would then result to a flat plane of Pareto efficient set. In contrast, the IEA algorithm deals with non-linear Pareto efficient set for a continuous optimization problem. Nevertheless, both greedy and IEA algorithms have a commonality, in which they initialize its algorithm with an operating point that lies on the Pareto efficient set.

In the following sub-sections, we first derive some properties of the constraint space which we call the utility space $\mathcal{U}$. Then we elaborate the greedy search algorithm for discrete resource allocation optimization problem in detail. Lastly, we discuss the worst case properties of the algorithm and compare it with that of the full search approach.

3.6.1 Properties of utility space

It is to be noted that the Theorem and the Proof discussed below have been published earlier in [148].

Theorem 1. *Let P be a set of points in the utility space corresponding to $\sum_i \alpha_i(p) = 1$, $P = \{p \in \mathcal{U} \text{ s.t. } \sum \alpha_i(p) = 1\}$. Let $\tilde{x}^*$ be the optimum mode of operation: $\tilde{x}^* = \arg\max \sum_i U_i$. Then, $\tilde{x}^* \in P$.*

Theorem 2. *The optimum of the objective function, $\tilde{x}^*$ lies on the boundary of the utility space, i.e., $\tilde{x}^* \in B_{\mathcal{U}}$.*

Proof. Let p be an interior point of the utility space $\mathcal{U}$, $p \in \mathcal{U}$, $p \notin B_{\mathcal{U}}$ and let $d(x, y)$ denote the Euclidean distance between points x and y. Then there exists another point $q \in \mathcal{U}$, $d(q, 0) - d(p, 0) > 0$ such that $\mathcal{F}(p) < \mathcal{F}(q)$. The existence of q is guaranteed until q lies on the boundary of $\mathcal{U}$, i.e., $q \in B_{\mathcal{U}}$. But $\sum_i U_i(p) < \sum_i U_i(q)$, so that an interior point of $\mathcal{U}$ cannot be an optimum. In other words, the optimum must lie on the boundary: $\tilde{x}^* \in B_U$. □

Theorem 3. *Assume monotonically increasing utility functions, $U_i(\alpha)$ for $\forall i$. Let P be a set of points in the utility space corresponding to $\sum_i \alpha_i = 1$, $P = \{p \text{ s.t. } \sum_i \alpha_i = 1\}$. Then $P = B_{\mathcal{U}}$.*

Proof. First we show that $P \subseteq B_{\mathcal{U}}$. Let $q \in \mathcal{U}$, $q \in P$ and $q \notin B_{\mathcal{U}}$. Then there exists another point $r \in B_{\mathcal{U}}$ such that $d(r, 0) - d(q, 0) > 0$. Hence, $\sum U(q) < \sum U(r)$ and $U_i(q) < U_i(r)$ for some i. Since $U_i(\alpha_a) > U_i(\alpha_b)$ only if $\alpha_a > \alpha_b$ (non-decreasing utility functions), $\sum \alpha(q) < \sum \alpha(r)$ which implies $\sum \alpha(r) > 1$. But then $r \notin \mathcal{U}$ and hence, $r \notin B_{\mathcal{U}}$. Therefore, $q \in B_{\mathcal{U}}$ which implies $P \subseteq B_{\mathcal{U}}$. Similarly, $B_{\mathcal{U}} \subseteq P$ can be proved by using the fact that $\alpha_a > \alpha_b$ only if $U(\alpha_a) > U(\alpha_b)$ (strictly increasing utility functions). $P \subseteq B_{\mathcal{U}}$ and $B_{\mathcal{U}} \subseteq P$ implies that $P = B_{\mathcal{U}}$. □

The proof of Theorem 1 follows from results of Theorem 2 and Theorem 3.

Discussion: Theorem 1 implies that the optimum of the utility maximization problem lies on the boundary of the utility space, so that a search through the whole utility space is not required. Hence, any algorithm that performs an exhaustive search over the set $B_{\mathcal{U}}$ would eventually find the global optimum.

3.6.2 Algorithm description

We consider a time window of S_o TTI. Let S_i be the number of TTI assigned to user i. Then we have, $\sum_{i=1}^{N} S_i \leq S_o$.

The greedy algorithm for the utility maximization is described below. Throughput maximization is performed in a similar fashion. The algorithm is initialized by assigning an amount of resource for every user such that $\sum_{i=1}^{N} S_i = S_o$. At each subsequent iteration a small amount of resources is taken from the user with the lowest sensitivity with respect to decrease of utility and assigned to the user which receives the maximum benefit. This process is repeated until there is no further improvement in the objective function.

Let U_i denote the utility function and α_i the fraction of total TTI assigned to user i: $\alpha_i = \frac{S_i}{S_o}$, $\sum_{i=1}^{N} \alpha_i = 1$. We consider a discrete set of α_i:

$$\alpha_i \in \{n \cdot \Delta\alpha \ s.t. \ n \in \mathcal{Z}_o, 0 \leq \alpha_i \leq 1\}, \forall i \tag{3.21}$$

where $\mathcal{Z}_o$ denotes the set of non-negative integers. Let ΔU_i denote the change of utility for user i due to a change of its resource share, $\Delta\alpha$. The greedy algorithm can be expressed as an iterative maximization of the incremental utility values of two users i^+ and i^-, $i^+ \neq i^-$ such that

$$i^+ = \arg\max_i \{\Delta U_i | \alpha_i \leftarrow \alpha_i + \Delta\alpha\} \tag{3.22}$$

$$i^- = \arg\min_i \{\Delta U_i | \alpha_i \leftarrow \alpha_i - \Delta\alpha\} \tag{3.23}$$

The greedy algorithm for the utility maximization is summarized in Algorithm 1.

3.6.3 Complexity

The worst case complexity of the greedy algorithm described in the previous section depends on the number of users and the granularity of the sampling of α. It can be shown that the cardinality of the constraint set, and hence the number of points that have to be

Algorithm 1 Greedy Algorithm

```
Input: Utility function $U$, Transmission policies $T$, number of user $N$, resource budget $S_o$, step size $\Delta\alpha$, increase of step size $\Delta\alpha_{inc}$, minimum expected utility change $\Delta U_{min}$, maximum number of iterations $I_{max}$.
Output: Optimal operating mode $\tilde{x}_{opt}$;
Initialization: initial resource share: $\alpha = [1, 0, 0, \cdot, 0]$, set $\Delta U_{max,inc}$ to a value greater than $\Delta U_{min}$. Iteration index, $I = 0$.
for $i = 1$ to $N$ do
    get operating mode $\tilde{x}_i$ from $\alpha_i$, $\tilde{x}_i \in T_i$
    Compute $U_i$
end for
loop
    for $i = 1$ to $N$ do
        get operating mode $\tilde{x}_{inc,i}$ from $\alpha_i + \Delta\alpha$, where $\tilde{x}_{inc,i} \in T_i$;
        get operating mode $\tilde{x}_{dec,i}$ from $\alpha_i - \Delta\alpha$, where $\tilde{x}_{dec,i} \in T_i$;
        compute $\Delta U_i(\tilde{x}_{inc,i})$ and $\Delta U_i(\tilde{x}_{dec,i})$;
    end for
    $\Delta U_{max,inc} = \Delta U_i(\tilde{x}_{inc,i}) - \Delta U_i(\tilde{x}_{dec,i})$
    if $\Delta U_{max,inc} < \Delta U_{min}$ then
        set $\Delta\alpha$ to $\Delta\alpha + \Delta\alpha_{inc}$
    else
        find $i^+$, $i^-$ using equations 3.22 and 3.23
        set $\Delta\alpha$ to $\Delta\alpha_{inc}$
    end if
    $I + +$;
    if $I > I_{max}$ then
        break;
    end if
end loop
output: $\tilde{x}_{opt}$
```

searched in the worst case increases with both the number of users and the granularity of sampling. Specifically, it is shown that the cardinality of the constraint set stays constant when the number of users and the number of samples are interchanged.

Let h be the number of possible modes for each user, $h \in \{1, 2, \cdots\}$. We assume the modes to be equally spaced, so that $\Delta\alpha = 1/(h-1)$. Let P be a set of vectors such that $P = \{(p_1 \cdots p_N)' s.t. \sum_{i=1}^{N} p_i = h, p_i \in \{0, 1, \cdots, h\}\}$. Then P is the set of points corresponding to $\sum_{i=1}^{N} \alpha_i = 1$. Hence the cardinality of the set P, $|P|$ is the worst case number of iterations for the greedy algorithm.

Let $|P| = N_G(h, N)$. Then, $N_G(h, N) =$

$$\begin{pmatrix} h+N-1 \\ N-1 \end{pmatrix} = \begin{pmatrix} h+N-2 \\ N-1 \end{pmatrix} + \begin{pmatrix} h+N-2 \\ N-2 \end{pmatrix} = N_G(h-1, N) + N_G(h, N-1).$$

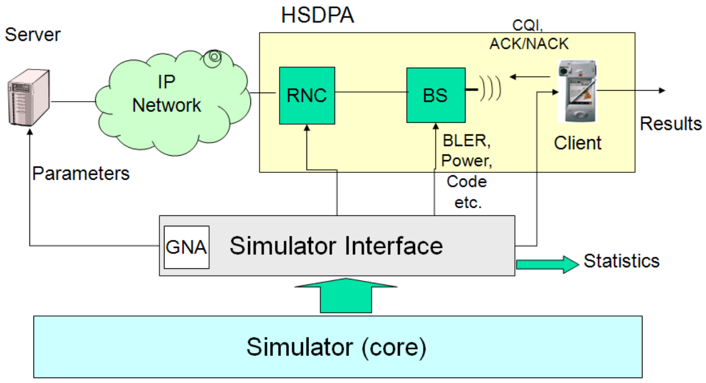


Figure 3.23: High-level architecture of OPNET HSDPA simulator.

This results in a 2D symmetric matrix of $N_G(h, N)$ which implies that we can interchange the number of users with the granularity of the sampling and yet the worst case number of iterations for the algorithm stays constant. This fact can be taken advantage of by using less granularity of sampling as the number of users grow, such that the real-time computation of the optimum remains feasible. In comparison, the number of iterations for a full search is h^N which becomes infeasible when $N \gg 1$.

3.7 OPNET HSDPA simulator overview

Throughout this thesis, we use the OPNET HSDPA module to emulate and simulate the HSDPA mobile system. The OPNET HSDPA provides complete and validated TCP/IP models, along with configurable settings that are needed to run a HSDPA simulation. The configurations include for example network node settings or application profiles of which the emulated HSDPA user is accessing. With this, it allows to repeat the same scenario compare applications of different schemes implemented in the HSDPA system.

Figure 3.23 depicts an architecture of the OPNET HSDPA simulator. The core simulator is the module that performs actual simulations based on the given simulation settings (e.g. random seeds, simulation period). The simulation interface allows us to configure any network nodes in the simulation project such as the application server, the Radio Network Controller (RNC) node, the Base Station (BS) or even the client in the HSDPA module. One can use the standard Generalized Network Application (GNA) library that contains a large number of standard network nodes (e.g. SUN work station) and interfaces (e.g. ethernet cable, or fiber cable) that are commercially available hardware and implemented by various vendors (e.g. CISCO, 3COM). Moreover, the simulation interface also collects simulation results and provides a number of statistics after finishing a simulation for an evaluation of system performance.

A screenshot of the Graphical User Interface (GUI) of OPNET HSDPA simulator is shown

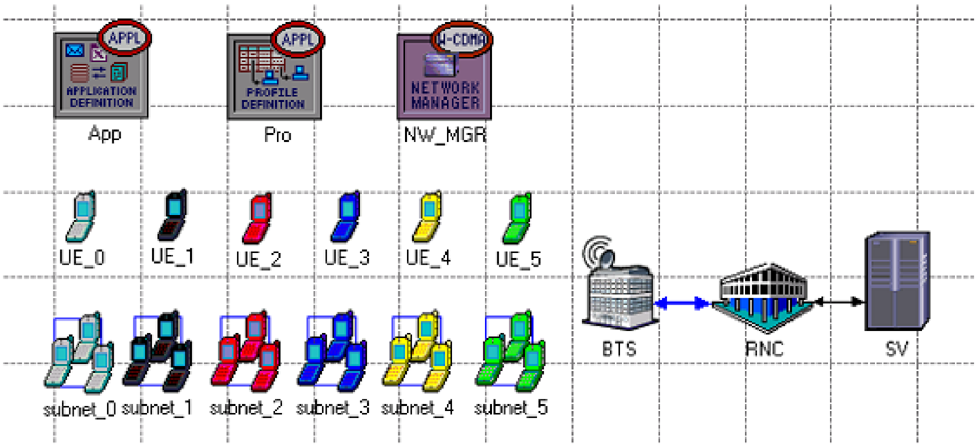


Figure 3.24: Screenshot of OPNET HSDPA simulator.

in Figure 3.24. In this example, we consider a single HSDPA base station (BTS) that is connected with the core network entity (RNC) responsible for radio network resource control across multiple users. The standardized Iub interface is used to connect between the BTS node and the RNC node. There are two types of mobile users: a single user (UE) and a group of mobile users (subnet). The latter mobile user type can be used, for example, in a scenario, where a group of mobile users is moving together or staying exactly at the same place, and thus experiencing the same wireless channel condition. The characteristics of each application type can be defined in the "Application Definition (App)" such as the packet interarrival time or the packet size. Configuration of how each user runs an application on his/her mobile terminal (e.g. start time or duration of application) can be specified in the "Application Profile Definition (Pro)". The "Network Manager (NWMGR)" is a place where all parameters related to network resource management can be configured such as a maximum transmission power and a maximum transmission code used for the high-speed downlink shared channel (HSDSCH).

3.8 Performance of QoE-driven CLO for network resource constrained system

3.8.1 Simulation setup

A HSDPA single cell scenario, where limited resources are shared among 10 users running different applications (voice, video streaming, video conferencing and file transfer), is considered in our simulation as depicted in Figure 3.2. Data traffic of all users is transmitted over the shared channel HS-DSCH and no dedicated channels are considered. For performance evaluation, five schemes to be compared are implemented in the HSDPA simulator as follows:

1. *No-adaptation*: This is the default HSDPA mode. The system is left to run into overloaded situations (network congestion), as no application-layer rate adaptation is supported. If the data rate measured at the UE is lower than the data rate sent by the application server and the packet delay exceeds the playout buffer time (e.g. 2 seconds), then we assume that the user experiences a minimal quality level (MOS 1). This is because the user cannot enjoy watching the video continuously, as the video player will stop displaying the video due to late packet arrival.

2. *Max-Rate*: Based on the channel conditions, adaptation is done so as to achieve maximum cell throughput regardless of the application type and the content. In this case, the utility function is based on the average data rate $\bar{R}_i$ as defined below:

   $$U_i = \bar{R}_i, \forall i \in \mathcal{S} \tag{3.24}$$

   Resources will be given to a user with good channel condition to achieve the highest data rate for the video that a user is accessing. If there are resources left, they will be given to the user, who has the next best channel conditions.

3. *Max-MOS*: Adaptation is done to maximize the mean user-perceived quality over all users. The utility function is a function of MOS as described as follows:

   $$U_i = MOS_i(\bar{R}_i), \forall i \in \mathcal{S} \tag{3.25}$$

   Resources are first given to a user having a good channel condition and accessing a low-demand application.

4. *MaxMin-MOS*: Based on the utility function defined in Eq. (3.25), the max-min fairness allocates resources such that all users experience the same perceived quality regardless of channel condition and application sensitivity.

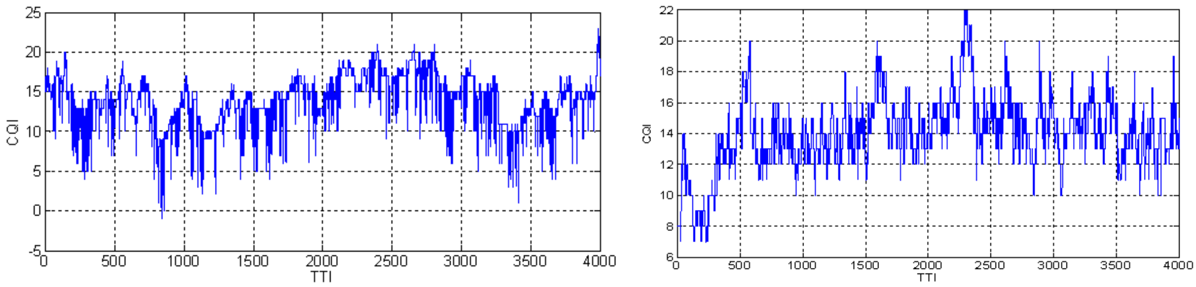
5. *MaxMin-MinMOSX.Y-MOS*: Similar to the max-min fairness approach, this scheme first sets a minimum guarantee of MOS $X.Y$ for all users and then adapts the resource allocation so as to achieve the same MOS that is equal or higher than the guarantee MOS. If the system cannot provide all users with the guaranteed MOS, a user or more requiring the highest amount of resources is dropped.

It should be noted that schemes 2) to 5) are application-aware, and a simple transcoding is used for rate adaptation in order to avoid network congestion. The optimization is performed every second.

The parameters used in our simulations are given in Table 3.2. The wireless channel model in the HSDPA simulator is based on the measured CQI trace representing different mobility schemes under different environments. Examples of CQI trace for a HSDPA static user and a HSDPA mobile user are shown in Figure 3.25. It is obvious that the mobile user experiences more dynamic and drastic changes of wireless channel conditions than the static user. For the packet scheduler, we use a proportional fair scheduler, and we assign lower priority to FTP with respect to other services. A set of possible rates, $\mathcal{R}_{vs}$, $\mathcal{R}_{vc}$, $\mathcal{R}_{voice}$, and $\mathcal{R}_{FTP}$ for video streaming, video conferencing, voice and FTP services, respectively are chosen as shown in Table 3.2. Discard timer, DT are set as shown in the table.

Table 3.2: Simulation Parameters

Total transmit power	15.8W
Power allocated to HS-DSCH	11W
Carrier Frequency	2GHz
User speed	3km/h
Distance from Node B	500m – 1.8km
UE category	6
Target BLER	10%
CQI averaging cycle	1sec
RLC PDU size	40byte
Scheduler	Proportional Fair
$\mathcal{R}_{vs}$	$\{0, 30, ..., < 500\}$kbps
$\mathcal{R}_{vc}$	$\{0, 96\}$kbps
$\mathcal{R}_{voice}$	$\{0, 6.4, 15.2, 24.6, 64\}$kbps
$\mathcal{R}_{FTP}$	$\{0, 50, 100, \cdots, 250\}$kbps
Video codec used	H.264
Voice codec used	G.723, iLBC, SPEEX, G.711
Loss concealment	Copy previous frame
Video/Voice rate shaping	Transcoding
DT_{vs}, DT_{FTP}	2sec
DT_{vc}, DT_{voice}	150ms
Simulator	OPNET 9.1 with NTT DoCoMo HSDPA plugin



(a)

(b)

Figure 3.25: HSDPA CQI trace examples for (a) a static user, (b) a mobile user.

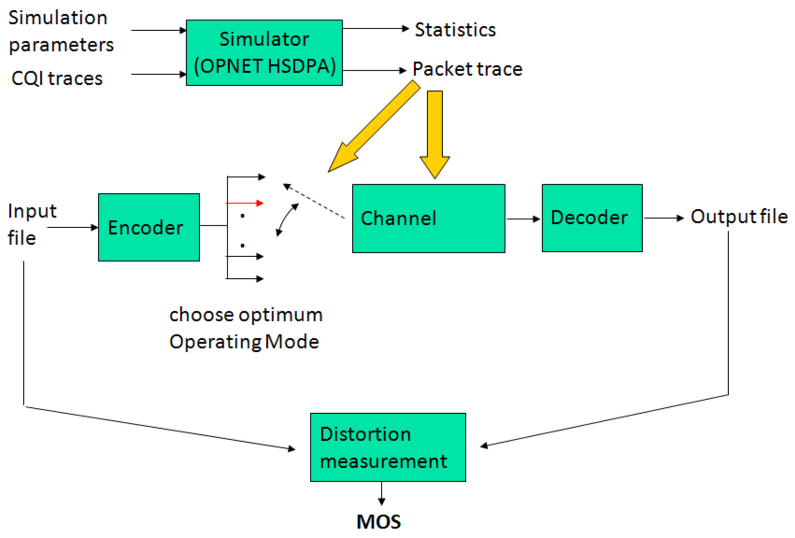


Figure 3.26: Schematic diagram of off-line evaluation methodology.

3.8.2 Evaluation methodology

The evaluation methodology is of particular importance to the quality-aware optimization framework, as we are interested in characterizing the system performance in terms of user perceived quality instead of only network-related parameters.

The simulation of a particular scenario produces packet traces which contain the time of generation and arrival of each packet and the chosen rate/operating mode corresponding to the packet. From this information, an offline evaluation is performed. Each media type is encoded into a set of possible rates. The packet trace file is used to infer the rates chosen for each user. Errors introduced to the bit-stream due to late arrival of the packets are simulated using the packet arrival times. This distorted bitstream is then decoded by the audio/video decoder with error-concealment enabled. The distortion between the original input stream and the output distorted stream is measured and converted to MOS following the approach outlined in Section 3.3. Figure 3.26 depicts an overview of the aforementioned methodology for evaluating the user-perceived quality given the simulation results.

3.8.3 Simulation results

Figure 3.27 shows the mean utility of all the users over the simulation period of 3 min. From time 10sec to 35 sec, users join the system one by one. The rate-based scheme and all utility-based schemes start around 40 sec. We see a significant performance gain between the no-adaptation scheme and the other schemes. Also, all MOS-based utility optimization schemes lead to an additional gain compared to the rate-based scheme. The Max-MOS scheme is the scheme resulting to the highest mean MOS under a constraint of network

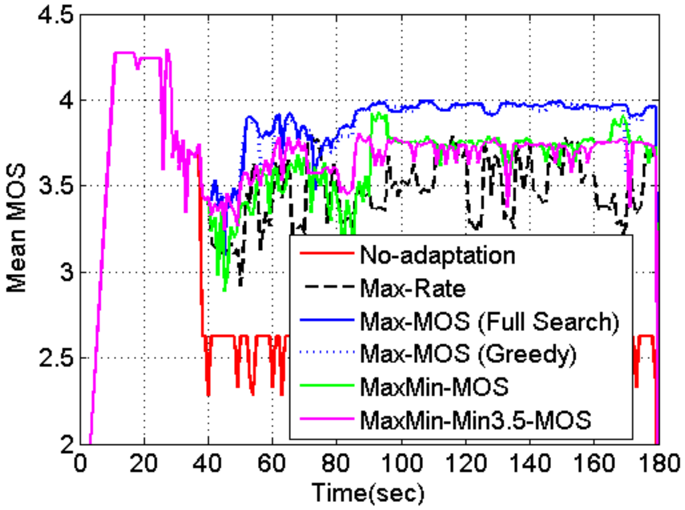


Figure 3.27: Mean utility for the 10 user case as a function of simulation time.

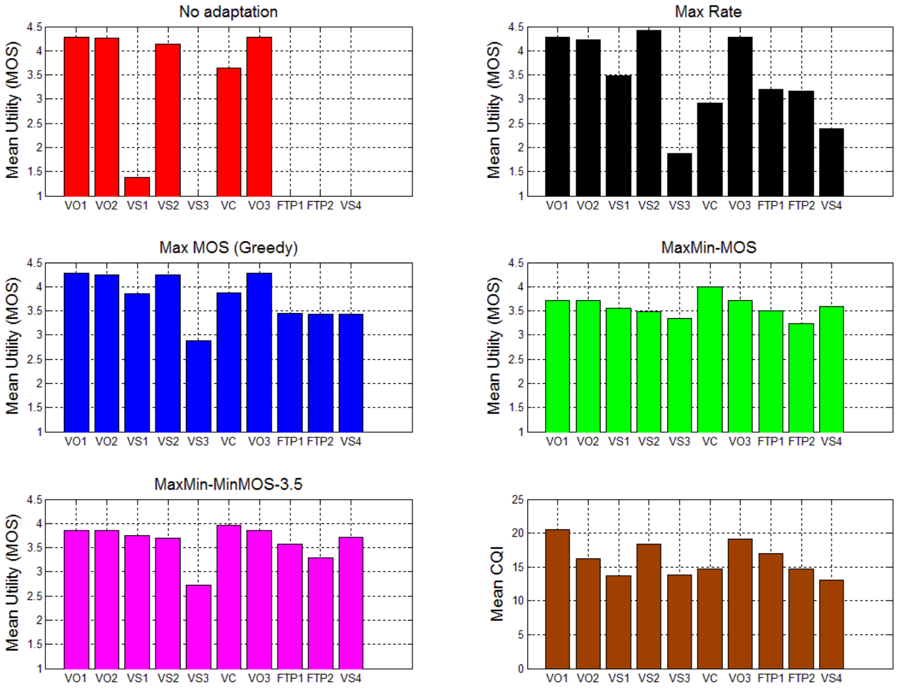


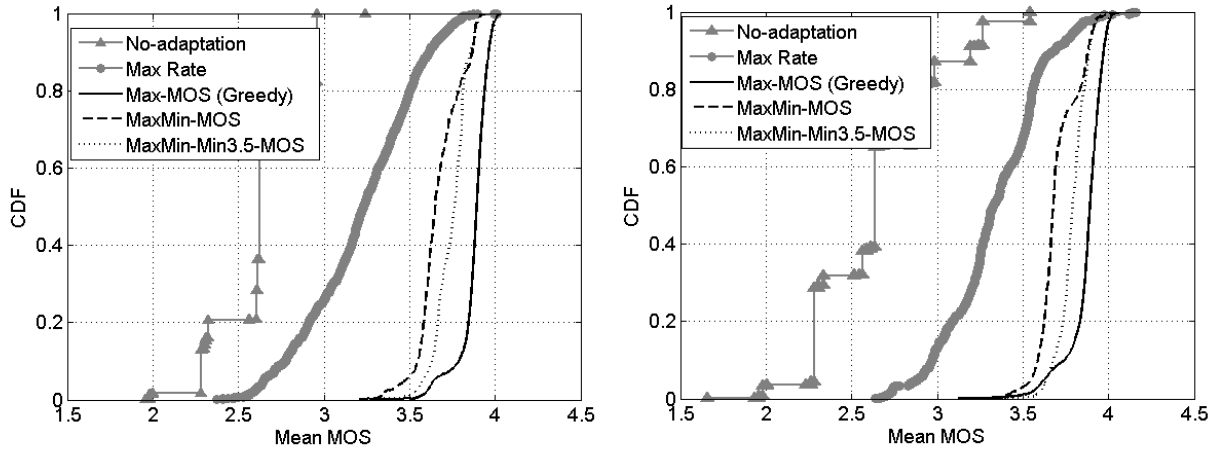
Figure 3.28: Mean utility and the corresponding mean CQI values for 10 users. [148]

resources shared among all users.

The average MOS of each user from one simulation is shown in Figure 3.28. Most of the gain for Max-MOS schemes comes from the users experiencing relatively bad channel conditions and demanding applications, e.g. video streaming user 3 (VS3). With the MaxMin-MOS approach, all users experience a similar service quality (around MOS 3.4) except the video conference (VC) user, who perceives higher quality than others due to the fact that its utility function consists of only one data rate (96kbps). By setting a minimum guarantee of service quality with MOS 3.5, the VS3 user suffers the most, since it requires the most resource to achieve the guaranteed MOS. But the overall quality of other users compared to no adaptation or rate-based scheme is still better. Note that the average perceived service quality for the two FTP users is slightly lower than other users due to a lower priority setting at the scheduler and the TCP slow-start behaviour.

Figure 3.29(a) shows the Cumulative Distribution Function (CDF) of mean MOS over all users over 300 simulation runs, each consisting of three minutes of simulation time. To avoid the effects of startup, in which users join the system one by one, we take the results of only the last two minutes. For clarity of the picture, we have left out the results of MaxMOS with full search, since the MaxMOS with greedy optimization performs as good as MaxMOS with full search as shown in Figure 3.27. Also, it is not feasible to do full search for each simulation run due to its complexity. We see that the rate-based optimization outperforms the no adaptation scheme with an average of 0.6 MOS. A further gain of 0.4 is achieved when using the MOS-based utility optimization scheme. The results also show that the MOS-based schemes have less dispersion around the mean value, which results in more stable user perceived quality compared to the other schemes. Although using Max-Rate approach would lead to the best result in terms of throughput, it does not guarantee that the quality perceived by the end-user will be the best. With the MOS-based schemes, the resources are allocated taking into account the cost and the gain in terms of user-perceived quality when allocating more or less resources to the users, who are running different multimedia applications and experiencing different wireless channel conditions.

By having more voice call users in the cell and less users for other applications, the gain



(a)

(b)

Figure 3.29: CDF of mean utility for the 10 user case using VSSIM (a) and PSNR (b) as a video quality assessment.

of MOS-based schemes is expected to be less due to the fact that the voice utility function only has a smaller number of steps and the voice call application is usually not a high-demanding application. Consequently, there are fewer operating points to adjust the data rate in the network, and less possibility to find an operating point that improves the quality for other users.

All the results that we have discussed so far are based on the MOS derived from the VSSIM-based video quality assessment as described in Section 3.3.3. We have also run similar simulations for the PSNR-based video quality measurement, and the CDF curve depicted in Figure 3.29(b) shows a similar result as for the case of the VSSIM-based CDF results. We conclude from the similarity of these two results that whichever video quality assessment type we use, the MOS-based utility optimization schemes always lead to a noticeable improvement of overall user satisfaction compared to the no-adaptation and the rate-based optimization scheme.

So far, we only see the advantage of the MOS-based scheme in terms of the mean utility (mean MOS) of all user. However, the proposed schemes also has another advantage that the operator can serve more users in the cell. Figure 3.30 shows the average increase in number of users N_{inc} of the system as a function of target mean MOS MOS_{target}. This result is drawn from a large number of simulations with the MaxMOS and the Max Rate schemes using the same random set of channel realizations for a ten user scenario as previous simulations. In another set of runs, we arbitrarily add from 0 to 10 new voice users. Out of all the simulations from the first and second set, we find the set of simulation which have a mean MOS that is more than or equal to a target mean MOS of MOS_{target}. We calculate the average increase in number of user N_{inc} by taking the average of the number of users within this set of simulations. From Figure 3.30, we observe that the MaxMOS scheme can add more than four users on the average when the MOS_{target} is set to 3.5 MOS, while the Max-Rate scheme can only add an average of 0.8 users. The fractional increase of user number less than unity for the Max-Rate case indicates that in many cases no user can be

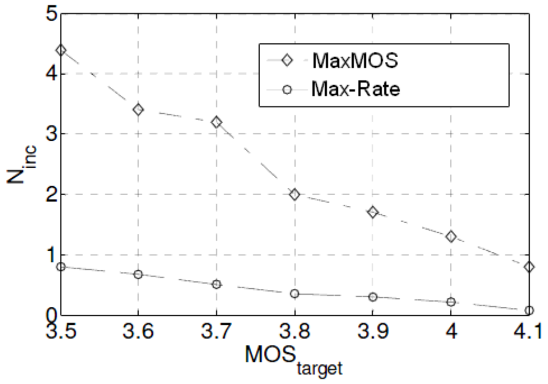


Figure 3.30: Target mean MOS vs. the increase in number of users for maxMOS and maxThroughput scheme. [89]

added in order to keep the MOS_{target}. For both schemes, the number of users that can be added decreases slowly with the increase of the MOS_{target}. We observe that the MaxMOS scheme can admit significantly more number of users than the Max-Rate scheme.

3.9 QoE-driven optimization for computational- and network resource constrained system

In this section, we consider a wireless network system for which a network entity performing a rate adaptation has a constraint of computational resource (processing power). To achieve an optimal rate adaptation with the new constraint, the QoE-based optimization framework discussed in Section 3.5 is now consisting of two steps: (1) selecting an optimal scheme for rate adaptation taking into account the resulting user-perceived quality (QoE) and the limited computational resources, and (2) finding an optimal network resource allocation with a constraint of limited network resources. The additional step prior to the QoE-driven optimization for resource allocation does an intelligent selection of rate adaptation scheme to be applied to each data stream taking into account the resulting user-perceived quality and the limited computational resources. To enable this, we assume that the video utility functions as shown in Figure 3.18 are precomputed at the streaming server and are signalled as side information along with the video bitstream, which will be then extracted in the core network for performing the QoE-driven rate adaptation scheme selection and the QoE-driven optimization. Due to the fact that video applications will become a majority of mobile traffic, we do consider the scenario, in which users only access different video content from server providing a high video quality with high data rate.

3.9.1 Rate adaptation scheme selection

In general, a node performing rate adaptation in the network has limited computational resources. Also as discussed in Section 3.3.3, transcoding is the technique that takes most of the computational resources. As a result, by knowing the hardware specification, one could assume a maximum number of video bitstreams to be transcoded simultaneously N_{max}. In case there are more video bitstreams than N_{max}, an algorithm for rate adaptation scheme selection to be applied for each video stream is needed. We propose a novel algorithm using the gradient between the user perceived quality and the network parameter from the packet dropping scheme as a measure of video sensitivity. With this, we avoid applying the packet dropping scheme to the videos that are more sensitive to packet dropping.

Two examples of network parameter are used in our work: the data rate R, or the required network resource α to provide the certain data rate. Using the data rate, the gradient is fixed for the whole period of the simulation, whereas the latter alternative takes the impact of channel quality into account, and hence, the gradient changes for each optimization cycle.

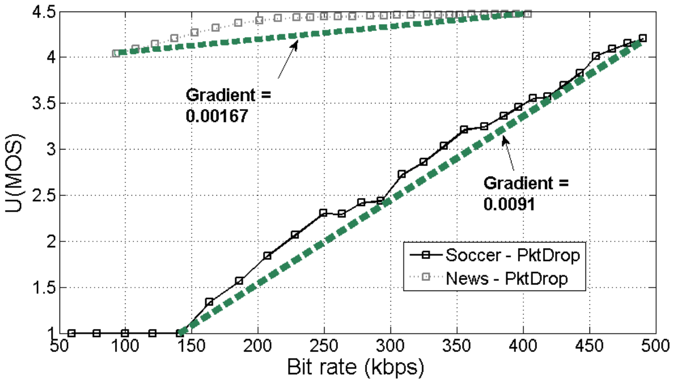


Figure 3.31: Gradient calculation for 'News' and 'Soccer' video sequences. [150]

Below, we describe how to calculate the gradient (η) using the data rate as the input for network parameter.

$$\eta = \frac{U_{max} - U_{min}}{R_{U_{max}} - R_{U_{min}}} \tag{3.26}$$

where U_{max} and $R_{U_{max}}$ are the maximum of utility and data rate from the utility function of packet dropping respectively, in which no packet is dropped. U_{min} and $R_{U_{min}}$ are the minimum of utility and data rate from the packet dropping scheme. Alternatively, the minimum could be a point, where the utility has first reached the minimum value. Figure 3.31 depicts an example of how the gradient is calculated for two video sequences. Since the Soccer video has a higher gradient, the operator first applies video transcoding to the Soccer video if there are sufficient computational resources available to perform transcoding for only one video. To calculate channel-based gradient, we apply the radio link layer model Eq. (3.2) and Eq. (3.7) as described in Section 3.2 and substitute $R_{U_{max}}$ and $R_{U_{min}}$ in Eq. (3.26) with $\alpha_{U_{max}}$ and $\alpha_{U_{min}}$ respectively. In this case, the $\alpha_{U_{max}}$ is the fraction of network resource required in order to achieve the maximum utility U_{max}, and the $\alpha_{U_{min}}$ is for the fraction of network resource that results to the minimum utility U_{min}.

3.9.2 Utility-based network resource allocation

After selecting the rate adaptation scheme to be applied for each video stream, we optimize the network resource allocation across users. As discussed in Section 3.5, the optimization can be done in several ways depending on the objective function set by the network operator. To validate the proposed scheme for rate adaptation scheme selection, we use the utility maximization as an example for the objective function, which maximizes the average utility of all K users. (see Section 3.5.1 for details) Also, we use a greedy search algorithm (GR) that leads to close-to-optimal results but is feasible in practice as shown in the simulation results in Section 3.8.3.

3.9.3 Simulation results

In our simulation, six video users located in a single cell are accessing different video content through HSDPA. To emulate a constraint on the computational resources, we assume a maximum of three video streams can be transcoded simultaneously. For the remaining videos, packet dropping is used due to its low complexity. Except the parameters related to the voice and FTP applications, other HSDPA simulation parameters given in Table 3.2 are used for the simulations. The two variants of the proposed scheme described below is compared with the *No-adaptation* scheme, the *MaxRate* scheme and the *MaxMOS* scheme (GR-Worst). The details of the latter three schemes are described in Section 3.8.1.

- *MaxMOS-RateSelect (GR-Rate)*: In addition to the MaxMOS scheme, the rate-based gradient is used to prioritize among multiple video streams and to decide which rate adaptation scheme to be applied for each single video stream.
- *MaxMOS-ChanSelect (GR-Channel)*: Before performing QoE-based resource allocation, we use the channel-based gradient for rate adaptation scheme selection.

To show the advantage of applying our rate adaption scheme selection, we fix the rate adaptation schemes for the optimization scheme 1, 2 and 3. For instance, we apply transcoding for the low-motion video and packet dropping for the high-motion video. Figure 3.32 gives an overview of the quality perceived by each user. VS1, VS2 and VS5 are users that access low-motion video content (e.g. 'News') and experience better channel quality, and therefore receive better video quality. The gain for the Max-MOS scheme comes from VS1, VS4 and VS6 by taking the application-layer knowledge into account. By performing an appropriate rate adaptation scheme selection prior to network resource allocation, we see

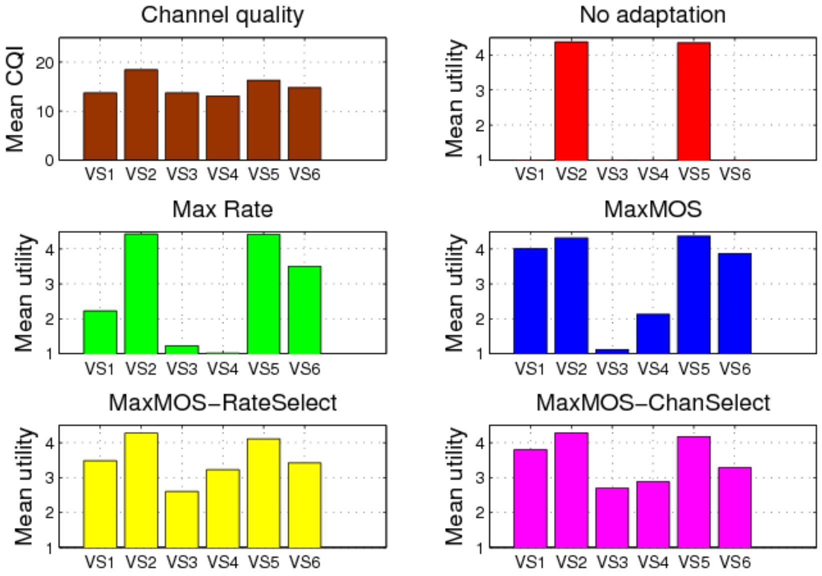


Figure 3.32: Mean utility and the corresponding CQI for each user. [150]

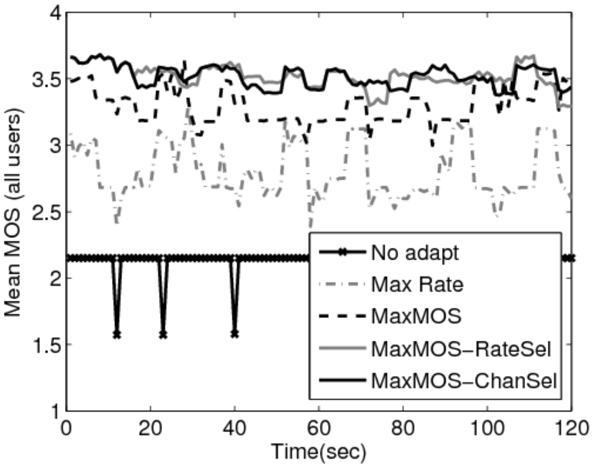


Figure 3.33: Mean utility for 6 users over time for a single simulation run. [150]

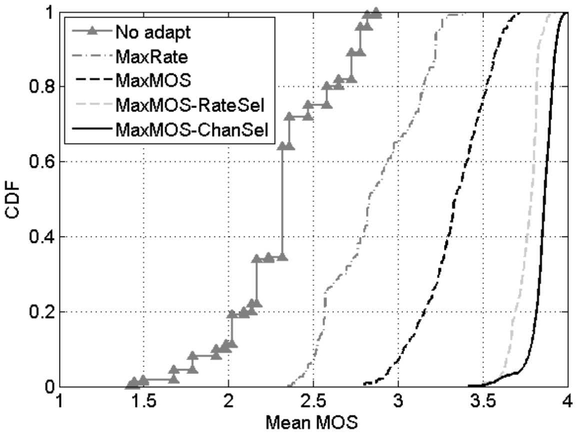


Figure 3.34: CDF of mean utility for 6 video users from 30 simulation runs.

a significant gain coming from the user VS3 and VS4 accessing a high motion video (e.g. 'Soccer' or 'Football').

Figure 3.33 depicts the mean utility of all 6 users over the simulation period of 2 min. We see a significant gain between the no adaptation case and the other schemes. All MOS-based schemes further improve the average mean MOS of all users when compared to the rate-based adaptation scheme. It is obvious that applying the QoE-based rate adaptation scheme selection prior to the QoE-based optimization for resource allocation leads to an additional gain of 0.25. The gain is achieved by transcoding the video, which is more sensitive to packet dropping, and thus avoiding the most drastic video quality degradation when performing an in-network rate adaptation.

Figure 3.34 shows the CDF of mean MOS of all 6 users over 30 simulation runs, for which each simulation last for two minutes period. We see an average improvement of 0.5 MOS for the Max-Rate optimization scheme when compared to the no-adaptation case. All

QoE-based approaches lead to additional improvements of user perceived quality. However, in case the QoE-based rate adaptation is applied prior to the QoE-based optimization for network resource allocation, the results are further improved in average of 0.4 MOS when compared to the utility maximization without the QoE-based rate adaptation scheme selection. Also, we observe that using the information of wireless channel condition experiencing by each user when selecting the rate adaptation scheme to be applied for each video flow (MaxSum-ChanSel) is better than the rate-based scheme selection (MaxSum-RateSel). However, the difference between the two schemes is marginal (0.07 MOS in average).

3.10 Summary

This chapter discusses and evaluates key problems faced by the multimedia communications over wireless networks. We focus on the loaded single cell scenario, in which there is no enough resources to support all users accessing different applications and experiencing different wireless channel conditions. We introduce a general QoE-driven optimization framework for network resource allocation across multiple users, which alleviates congestion at the base station. Conventionally, this situation is avoided by a strict admission control policy. But by doing this, users would suffer from high blocking probability and operators would loose revenue. We propose that the applications be re-adapted, taking into account the wireless channel conditions and the application utility functions that describes the relationship between the estimated user-perceived quality to the network performance parameters. This policy results in better mean quality of experience for given system resources and for a fixed number of users, and the admission of more users for a given target quality.

We use the HSDPA system as an example of a mobile access network in order to evaluate the proposed scheme and compare it with other existing approaches. Simulation results show that all QoE-driven optimization schemes (maximization mean MOS, max-min MOS fairness and max-min MOS with minimum MOS guarantee) outperform the no adaptation and the throughput maximization. Nevertheless, we do not conclude which QoE-based objective function is the best, as selection of the QoE-based objective function depends solely on the network operator's needs and their policy on allocating their network resources.

Furthermore, we propose a QoE-based rate adaptation scheme selection for video applications that allows a network operator to easily handle multiple video streams with various contents and to select dynamically an appropriate rate adaptation scheme to be applied to each video stream. The proposed selection algorithm is easily integrated with the QoE-driven network resource allocation optimization. Results have shown a significant improvement of having such additional intelligent selection of the rate adaptation scheme.

Although the proposed QoE-based framework and the QoE-based rate adaptation scheme selection are implemented in the HSDPA system, it can be also integrated into future packet-based systems, e.g. in LTE, due to its generality.

Chapter 4

Multi-objective QoE-driven optimization

Due to a hugh ramp in video traffic, the mobile access networks is expected to remain a bottleneck link when providing video services to a large number of users. In Chapter 3, we discuss the QoE-driven Cross-Layer Optimization (CLO) schemes that optimize wireless resource allocation and perform application-layer rate adaptation according to the pre-defined objective function. For example, the resources are allocated so as to achieve maximum average user-perceived quality of the whole system (utility maximization) or to satisfy all users with the same quality regardless of the application type and the channel quality condition (utility maxmin fairness). Moreover, it has been shown that applying any of the utility-based approaches leads to significant improvements of user perceived quality compared to other approaches including the throughput maximization. However, we do not deal with the optimization of wireless resource allocation addressing multiple objectives simultaneously.

In this chapter, we discuss the QoE-driven optimization for wireless resource allocation taking into account multiple objectives. First, we introduce a tuning mechanism that allows the network operator to prioritize their resource allocation policy from the two utility-based objectives (utility maximization and utility maxmin fairness) already mentioned in Chapter 3. Then we combine the utility maximization with the novel objective function that minimizes the perceivable temporal change of the video quality negatively affecting the overall quality of experience. The combined objective function aims at achieving minimal perceivable change of temporal quality while at the same time maintaining the average perceived quality of all users as high as possible. Lastly, we talk about a practical approach for the multi-criteria QoE-driven optimization that includes all three objective functions. Throughout this chapter, we consider the scenario, in which multiple users only access video contents and not other application types due to the fact that video applications are expected to contribute to a majority of user-plane traffic in mobile networks. Also, we assume that

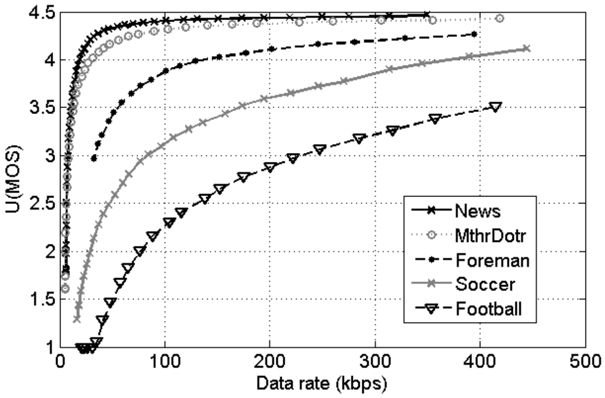


Figure 4.1: VSSIM-based video utility functions for different video sequences obtained with transcoding. [151]

- the VSSIM-based video utility functions discussed in Section 3.3.3 are precomputed at the streaming server and are signalled as side information along with the video bitstream, which will be then extracted in the core network for performing the QoE-driven optimization and necessary rate adaptation; and
- the base station sends periodically the long-term link-layer information as discussed in Section 3.2 to the QoE optimizer module, which is located nearby or at the base station.

Figure 4.1 depicts video utility functions for five different video contents that we use in all simulations presented in this chapter.

4.1 QoE-driven optimization addressing system efficiency and user fairness

In this section, we focus on a QoE-driven CLO for wireless video delivery that takes into account two objectives: utility maximization and utility max-min fairness. Like in Chapter 3, the utility is defined as a degree of user-perceived quality of the service delivered by the network operator. The first objective emphasizes achieving a maximum average perceived quality of all users which can be interpreted as how efficient the network resources are used and distributed to all users. Whereas for the second objective, its goal is to achieve a similar perceived quality among all users. It emphasizes minimizing the quality difference between the user experiencing the highest quality and another user experiencing the lowest quality. We formulate and solve the optimization problem which allocates the total system resources among the active clients so as to satisfy the chosen optimization criteria. While in perfect systems with users having good channel conditions, all applications can be served with very high quality, in constrained systems, the resource allocation must be carefully

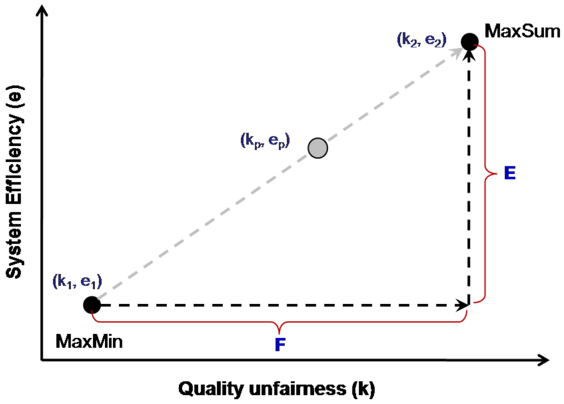


Figure 4.2: Schematic depiction of the proposed tuning mechanism between system efficiency e and quality unfairness k.

performed as a consequence of the trade-off between the system efficiency of the allocated resources and the fairness balancing among all users. When these two operation points are too far apart, the network operator may prefer to have an intermediary operation point. We propose a tuning mechanism allowing a system operator to dynamically adjust its operation point between the extreme points of maximum system efficiency and maximum fairness of perceived quality among all users. Its purpose is to provide flexibility to the network operator, and not to show whether the tuning algorithm outperforms the existing QoE-based optimization schemes, for example, in terms of the mean utility of all users.

4.1.1 Efficiency vs Fairness trade-off problem formulation

To formulate the tuning problem, we define the system efficiency and the quality unfairness among users as follows:

Definition 1: The system efficiency e is the total sum of the MOS values perceived by all N users as given in Eq. (4.1).

$$e = \sum_{i=1}^{N} MOS_i \tag{4.1}$$

Definition 2: The quality unfairness k is the maximum MOS deviation among users in a user group S, which is the perceived quality difference between the user experiencing the highest MOS (MOS_{max}) and another user experiencing the lowest MOS (MOS_{min}) at each time instance t.

$$k = MOS_{max} - MOS_{min} \tag{4.2}$$

where $MOS_{max} = \max_{i\in S}\{U_i(t)\}$ and $MOS_{min} = \min_{i\in S}\{U_i(t)\}$.

Figure 4.2 shows how the two objectives (utility maximization and utility max-min) can be mapped on a two-dimensional diagram capturing the system efficiency and the quality unfairness. In this example, the utility maximization results to the highest system efficiency and the lowest quality fairness. Whereas, the highest fairness can be achieved by using the utility maxmin, however, it comes at a cost of system efficiency. The operation points of maxmin and maxsum are denoted with index 1 and 2 respectively, which then specifies the fairness interval F and the system efficiency interval E. Alternative to achieving maximum system efficiency or maximum quality fairness, the network operator may prefer to have any intermediary operation point or to set a desired fairness level and a targeted system efficiency. Hence, an algorithm that enables a control of operation point is necessary. The desired operation point p is assumed to lie within the fairness interval F and the system efficiency interval E as shown in the figure.

4.1.2 Influence factors for fairness and efficiency tuning problem

Investigation of the trade-off problem (e.g. determination of the fairness interval F and the efficiency interval E) is important in order to construct a tuning algorithm that works well within these borders. Some of the important factors that influence the fairness and efficiency of a system are, for example, the variance of wireless channel quality among users σ^2, the number of users accessing the same wireless medium N, and the video content types V the users are watching. We use the maximum achievable data rate R_{max} defined in Eq. (3.7) as a measure of wireless channel quality. The higher variance of channel quality σ^2 means the R_{max} among users is more diverging. In other words, some users experience a very good channel condition and some experience a very bad channel condition. Taking into account these three factors, the fairness interval F and the efficiency interval E can be formulated as follows:

$$F = f_1(\sigma^2, N, V), \quad E = f_2(\sigma^2, N, V) \tag{4.3}$$

Through simulations of a mobile access system, we examine these relations to comprehend the factors that influence the tunable range of the resource allocation. Figure 4.3(a) and Figure 4.3(b) show an impact of the three influence factors on the F and the E intervals respectively. Obviously, both intervals increase when the wireless channel condition among users is getting more diverging (higher σ^2) regardless of the video content type and the number of users in the system. The same observation can be made when there are more users joining the system. This is due to the fact that when there are less number of users, the system becomes less loaded and has more resources to be allocated to all users. For a comparison of having different video contents, we see that when users access a video with dynamic scenes ('Soccer'), both F and E intervals are larger than the case of users accessing static video content ('News'). The reason is that the 'Soccer' video requires more resources to achieve the same QoE as of the 'News' video. Also, the 'Soccer' video is more sensitive to the data rate as shown in Figure 4.1.

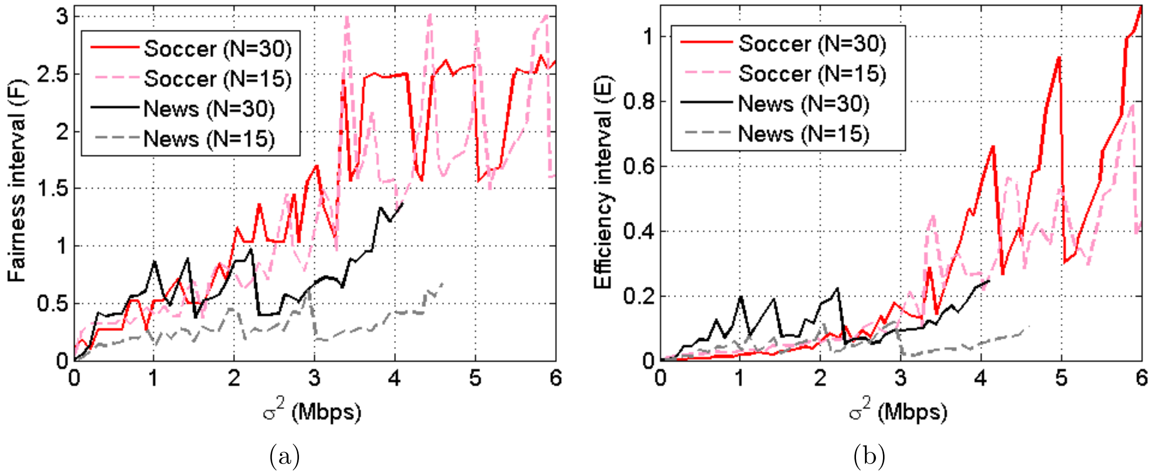


Figure 4.3: Impact of influence factors on the fairness interval F (a), and the system efficiency interval E (b).

4.1.3 Efficiency and fairness tuning mechanism

From the results shown in previous section, we see that it is not easy to construct and derive a mathematical formular that is applicable to all kinds of users and video groups, since the tuning parameters (e.g. E, F) are strongly dependent on the scenario specified by the number of users, channel variation among users, and application utility functions. Hence, we propose a tuning algorithm based on the heuristic and iterative solution. From the two criteria as discussed in Section 4.1.1, we devise three tuning algorithms allowing the network operator to find a desired operation point of resource allocation by specifying A) only e-constraint, B) only k-constraint, or C) both e and k constraints at the same time. The search for optimal resource allocation that meets the given constraint(s) is determined by using the greedy search algorithm. We elaborate each of the tuning algorithm in the following subsections.

Sum-MOS algorithm (e-constraint)

The Sum-MOS algorithm enables a full control of resource allocation in order to deliver the target sum of quality e_{req} (or target mean quality) of all users that is pre-defined by the network operator. To avoid large quality differences among the users, the Sum-MOS algorithm starts allocating the resources based on the utility max-min resource distribution [28]. From this point, the algorithm improves the total sum of quality of all users by using the greedy-based utility maximization (Max-Sum MOS) as given in Algorithm 1 in Section 3.6.2. However, we do add a e_{req} constraint to the optimization problem, which can then

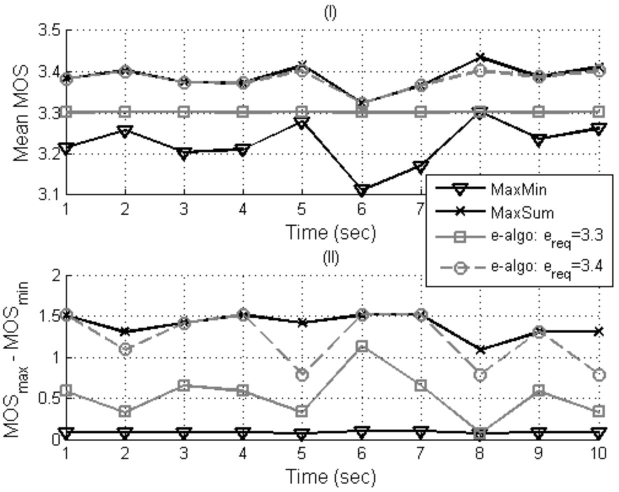


Figure 4.4: Comparisons of the Sum-MOS algorithm and the MaxSum/MaxMin algorithms: (I) Mean MOS of all users over time, (II) Unfairness over time.

be formulated as follows:

$$\tilde{\alpha}_{opt} = \arg\max_{\tilde{\alpha}\in\alpha_S} \frac{1}{N}\sum_{i=1}^{N} U_i(\tilde{\alpha}_i) \tag{4.4}$$

subject to $\sum_{i=1}^{N} \tilde{\alpha}_i = 1$ and $\frac{1}{N}\sum_{i=1}^{N} U_i(\tilde{\alpha}_{opt}) \geq e_{req}$.

Comparing with the objective function in Eq. (3.17), the cross-layer parameter tuple in Eq. (4.4) is the fraction of allocated resources $\tilde{\alpha} \in \alpha_S$, where α_S is the set of possible resource allocations from operation modes. In ideal case and under a continuous utility function U, the algorithm is able to stop when the target sum of quality e_{req} is achieved. Unless, if the e_{req} is set too high, the algorithm will stop at the maximum sum of perceived quality of all users that is achievable as of the Max-Sum MOS algorithm. In case U is a discrete function, the algorithm will stop when the average utility of all users is higher than the e_{req}. In brief, the Sum-MOS algorithm is slightly different than the Max-Sum MOS algorithm, as there is the aforementioned additional constraint and condition after the line 20 in Algorithm 1.

Figure 4.4 shows the system efficiency e and the unfairness k as a function of time, when applying the utility maximization (MaxSum), the utility max-min optimization (MaxMin), and the Sum-MOS algorithm with different settings of e_{req} constraints. In this scenario, multiple video users access different video contents and experience different wireless channel condition at a time instance. Obviously, with the MaxSum, the differences in perceived quality among all users are much larger than applying the MaxMin, however, the MaxSum has a higher mean MOS of all users in the system. With the Sum-MOS algorithm, the target mean MOS e_{req} is maintained during the simulation period, for example at mean MOS 3.3. In case, the e_{req} is set too high and cannot be achieved, it allocates resources so as to achieve a possible maximum mean MOS as result in the MaxSum case. This can be seen in Figure 4.4, for example, the maximum achievable mean MOS at 6th second is about 3.2 although the e_{req} is set to 3.4.

k-algorithm (k-constraint)

As its name indicates, the k-algorithm focuses on the quality (un)fairness k. It allows the network operator to apply a strict fairness constraint k_{req} that is set in advance and to allocate its network resource accordingly, while maintaining the system efficiency e as high as possible under this fairness condition. The optimization problem is similar as defined in Eq. (4.4), however the constraint is now changed to $\sum_{i=1}^{N} \tilde{\alpha}_i = 1$ and $MOS_{max}(\tilde{\alpha}_{opt}) - MOS_{min}(\tilde{\alpha}_{opt}) \leq k_{req}$. The $MOS_{max}(\tilde{\alpha}_{opt})$ and $MOS_{min}(\tilde{\alpha}_{opt})$ are the maximum and the minimum of utility among all users given the optimal resource allocation $\tilde{\alpha}_{opt}$ as defined below.

$$MOS_{max}(\tilde{\alpha}_{opt}) = \max_{i \in S} \{U_i(\tilde{\alpha}_{opt})\}$$

$$MOS_{min}(\tilde{\alpha}_{opt}) = \min_{i \in S} \{U_i(\tilde{\alpha}_{opt})\}$$

In case, the algorithm cannot find an operation point that meets the fairness constraint, for example, due to the discrete application utility function, it will find the operation point closest to the desired fairness operation point. The k-algorithm is summarized in Algorithm 2. Figure 4.5 compares the k-algorithm for different k_{req} settings with other utility-based schemes. The same scenario as mentioned in the Sum-MOS algorithm is applied here. In contrast to the Sum-MOS algorithm, the k-algorithm maintains the difference in quality among users over time according to the pre-defined k_{req}. However, this results to a variation of the mean MOS over time. Moreover, we observe that the higher k_{req} we set, the higher mean MOS of the system we could achieve.

Advanced k-algorithm (e and k constraints)

Advanced k-algorithm allows the network operator to allocate its network resource with pre-defined e_{req} and k_{req} constraints. Since it is possible that both constraints may not be met from any feasible set of resource allocation, we propose a two-step optimization that combines the Sum-MOS algorithm with the k-algorithm. We start with the k-algorithm as described in Algorithm 2 in order to fulfill the fairness criteria. From this point, we check whether the mean quality of all users is equal or greater than the desired mean quality e_{req}. If this is the case, the resource is allocated as of the output of k-algorithm. Otherwise, based on the result of k-algorithm, we proceed with the 2nd-step optimization aiming at utility maximization (Max-Sum MOS) as given in Algorithm 1 in Section 3.6.2. The optimization stops either when the sum of perceived quality of all users reaches at the level of e_{req} or when the maximum sum of perceived quality of all users is achieved. The latter case will happen when the e_{req} is set too high. The objective function for the

Algorithm 2 k-algorithm

```
Input: Utility function $U$, number of users $N$, quality unfairness constraint $k_{req}$.
Output: Optimal operation mode $\tilde{\alpha}_{opt}$
Initialization: zero resource share: $\tilde{\alpha} = [0, 0, ..., 0]$, $U = [1, 1, ..., 1]$.
loop
    for $i = 1$ to $N$ do
        Calculate the difference to next MOS level and the required resource: $\Delta U_i$, $\Delta\tilde{\alpha}_i$
        Compute utility gain $G_i = \frac{\Delta U_i}{\Delta\tilde{\alpha}_i}$
    end for
    Ordering all users by utility gain $G$
    Find the user giving highest utility gain ($i_{max} = \arg\max_{i \in N}\{G_i | \tilde{\alpha}_i \leftarrow \tilde{\alpha}_i + \Delta\tilde{\alpha}\}$)
    Precalculate the fairness $k_{new}$ and the total allocated resources $\tilde{\alpha}_{new} = \sum_{i=1}^{N} \tilde{\alpha}_i$ by assuming
    that resource share is given to the user $i_{max}$
    if $k_{new} \leq k_{req}$ and $\tilde{\alpha}_{new} \leq 1$ then
        Allocate $\Delta\tilde{\alpha}$ to the user $i_{max}$
    else
        Continue searching for the user $i_{next}$ with smaller $G$ but satisfying the $k_{req}$ and resource
        constraints
    end if
    break if there is no resource left ($\tilde{\alpha}_{new} \geq 1$)
end loop
```

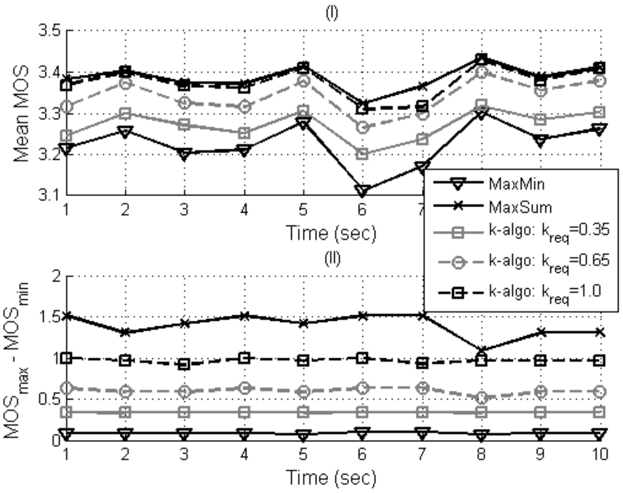


Figure 4.5: Comparisons of the k-algorithm and the MaxSum/MaxMin algorithms: (I) Mean MOS of all users over time, (II) Unfairness over time.

advanced k-algorithm is similar to the objective function defined in Eq. (4.4), except that there are now three constraints to be considered as given below.

$$\sum_{i=1}^{N} \tilde{\alpha}_i = 1$$

$$MOS_{max}(\tilde{\alpha}_{opt}) - MOS_{min}(\tilde{\alpha}_{opt}) \leq k_{req}$$

$$\tfrac{1}{N}\sum_{i=1}^{N} U_i(\tilde{\alpha}_{opt}) = e_{req}$$

In case the result of the 2nd-step optimization dissatisfies the k_{req}, we find a new operation point that is based on the priority of system efficiency c_e and quality fairness c_k, which are assumed to be specified in advance by the network operator. The advanced k-algorithm is summarized in Algorithm 3.

We illustrate the optimization progress of the advanced k-algorithm by using two hypothetical examples as depicted in Figure 4.6. In this example, the k_{req} and the e_{req} are set to 1 and 4.2 respectively. The operation mode in an ideal case will be the desired point. However, as depicted in Figure 4.6, it is possible that there is no feasible operation mode that fulfills both requirements at the same time. By using the k-algorithm and the Sum-MOS algorithm, we can achieve the reference points (denoted as p_1 and p_2) respectively. With the reference points and the prioritization coefficients c_e and c_k, a new desired point is determined, for which we find the closest feasible operation mode as denoted with the point p_{opt}.

Figure 4.7 shows how the mean MOS and the unfairness develop over time when applying the advanced k-algorithm with e_{req} and k_{req} set to 3.3 and 0.35 respectively. We see that when both requirements cannot be met at the same time (e.g. at 6th second), the priority parameters c_e and c_k play an important role. For instance, if the c_e is set to a higher

Algorithm 3 Advanced k-algorithm

```
Input: Utility function $U$, number of users $N$, step size of resource $\Delta\alpha$, quality unfairness constraint $k_{req}$, system efficiency constraint $e_{req}$, weighting coefficients for fairness and efficiency $c_k$ and $c_e$
Output: Optimal operating mode $\tilde{\alpha}_{opt}$
Initialization: zero resource share: zero resource share: $\tilde{\alpha} = [0, 0, ..., 0]$, $U = [1, 1, ..., 1]$.
call: k-algorithm (Algorithm 2).
Output of k-algorithm $\rightarrow p1$ operating mode $\tilde{\alpha}_{opt,p1}$
if $e_{p1} \geq e_{req}$ then
  return: $\tilde{\alpha}_{opt} \leftarrow \tilde{\alpha}_{opt,p1}$
else
  Run Max-Sum algorithm (Algorithm 1) until $e(\tilde{\alpha}) \geq e_{req}$.
  Output of Max-Sum algorithm $\rightarrow p2$ operating mode $\tilde{\alpha}_{opt,p2}$
  if $k_{p2} \geq k_{req}$ then
    $e_{end} = e_{p1} + c_e(e_{p2} - e_{p1})$
    $k_{end} = k_{p1} + c_k(k_{p2} - k_{p1})$
    recall: Max-Sum algorithm starting from $p1$ until $e(\tilde{\alpha}) \geq e_{end}$
    Output of Max-Sum algorithm $\leftarrow p3$ operating mode
    return: $\tilde{\alpha}_{opt} \leftarrow \tilde{\alpha}_{opt,p3}$
  else
    return: $\tilde{\alpha}_{opt} \leftarrow \tilde{\alpha}_{opt,p2}$
  end if
end if
```

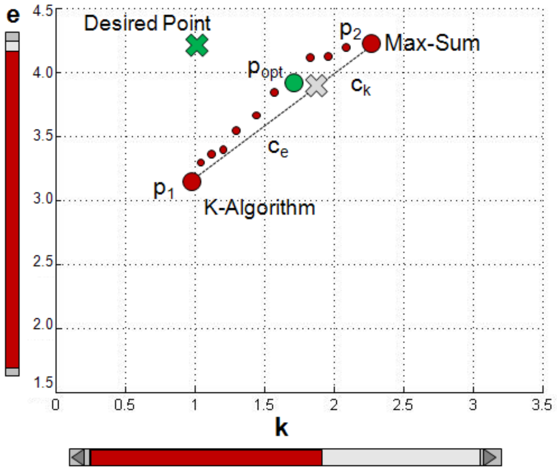


Figure 4.6: Optimization progress with the advanced k-algorithm from a hypothetical example.

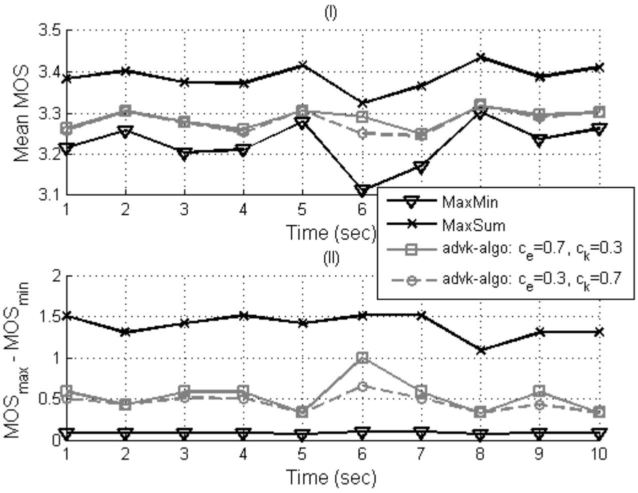


Figure 4.7: Comparisons of the advanced k-algorithm (with e_{req}=3.3 and k_{req}=0.35) and the MaxSum/MaxMin algorithms: (I) Mean MOS of all users over time, (II) Unfairness over time.

priority, the resulting mean MOS is getting closer to the e_{req} and vice versa.

Concluding remarks

From the results of all three tuning algorithms, one can observe that the user fairness comes at a cost of the system efficiency. It should be also noted that all tuning mechanisms do not intend to get a better result than the MaxSum and the MaxMin in terms of the mean MOS and the degree of user fairness respectively. In fact, they are used to provide the network operator a flexibility in specifying its policy for allocating its network resources in different ways based on the pre-defined system efficiency e_{req} and user fairness k_{req} constraints.

4.2 QoE-driven optimization with unperceivable temporal video quality fluctuation

The QoE-driven CLO discussed so far does allocate network resources and adapt video data rate by considering the video content sensitivity and the wireless channel condition experiencing by the user at each time instance. But it does not deal with the problem of how to avoid noticeable quality fluctuations over time. Even if the user-perceived quality is good on average, drastic quality changes can lead to a negative impression of the service quality. Hence, temporal quality fluctuation is an important factor for wireless video transmission.

In this section, we extend the previous QoE-driven CLO framework by introducing a new objective function incorporating the temporal video quality fluctuation. Let us explain this extension along with the drawbacks of two typical utility-based objective functions (a) maximization of average user-perceived quality and (b) maxmin fairness. With the utility maximization, when the channel quality changes drastically over time, the user experiences perceivable changes of video quality and may be annoyed while watching the video. Specifically, when the channel quality condition is getting poor, the optimizer allocates less network resources to the user. Or in the worst case, if the system is very congested, no resource is given to the user. But when the channel quality is very good, the optimizer will give a higher priority for network resource allocation to the user. Thus, the service quality perceived by the user strongly depends on the channel condition. In fact, early work in the area of Variable Bit Rate (VBR) versus Constant Bit Rate (CBR) encoding of video has shown that users prefer constant quality compared to temporally fluctuating quality even if the average quality is lower [112]. In contrary, with the maxmin fairness objective function, all users perceive the same quality, which would make the temporal quality smooth. However, this leads to a minimum of system efficiency in terms of network resource utilization, as most of the network resources are given to the user having a poor channel quality or to the user accessing a high-demand application. Considering the impact of temporal quality changes perceived by the user in the objective function of the network resource allocation problem hence has the potential to improve the overall user perceived quality. We apply the Just Noticeable Difference (JND) concept [109] to find a threshold of the temporal video quality change perceivable by the human eyes.

In the following subsections, we discuss how we determine the JND for video quality fluctuation through subjective tests, and how to apply the JND concept to the QoE-driven optimization problem in order to achieve a smooth temporal video quality while keeping the network resource usage (system efficiency) as high as possible.

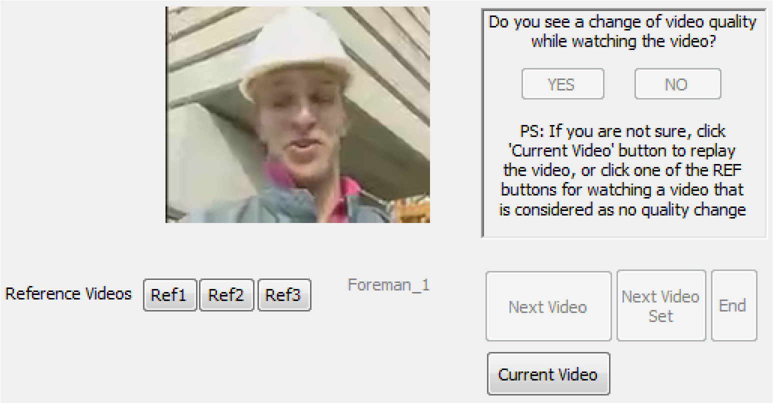
4.2.1 Perceivable temporal video quality fluctuation

Similar to any other instances in which a human is able to perceive the change of a stimulus, for example the perception of the weight change of carried objects, which is only perceivable if the weight change exceeds a certain threshold, there is also a threshold for the temporal video quality change humans are able to recognize. Incorporating such a threshold into the objective function of network resource allocation improves the overall user-perceived quality of the whole period of accessing the service/application. Also, it gives more flexibility for network resource allocation. For instance, within the range of unperceivable video quality change, some of the network resources allocated to the user accessing a low-demand video or to the user having a good channel condition may be given to the user accessing a high-demand video or to the user having a bad channel condition, while the user giving the resources to others is not aware of any quality change.

Subjective test

To find the Just Noticeable Difference (JND) for the change of temporal video quality, we have performed a subjective test with 30 persons using the forced-choice method as specified in [68]. Note that all persons participating the subjective test are mainly students at the Technische Universität München (TUM) whose age ranges from 23 to 28 year's old. In this recognition testing, we present a stimulus (test video sequence) to the subject (user), and ask him/her "Do you recognize any changes of video quality?" As specified in the ITU standard, it is required that each subject should not perform user tests longer than 30 min., as human eyes will be tired and the results will not be reliable. Due to the time constraint, two video sequences: 'Mother and Daughter' (static scene) and 'Foreman' (dynamic scene) are used in our test. For each video, we create 16 test video sequences that are encoded at two different levels of video quality. For instance, if the length of a video is 8 sec, the video is encoded at MOS 3.0 quality for the first 4 seconds, and at MOS 4.0 quality for the last 4 seconds. In this case, we have a magnitude of quality change ξ of MOS 1.0 quality, and the time of quality change τ at 4 second. All test video sequences are encoded with H.264/AVC at 30 frames/sec, and have QCIF resolution. Table 4.1 gives an overview of the test conditions for all test sequences. Note that we use the Variable Bit Rate (VBR) techniques such as encoding the video by fixing a quantization step to maintain the video quality for a certain period. Also, we calculate the video quality in terms of MOS using the VSSIM index as discussed in Section 3.3.3.

Figure 4.8 depicts the GUI design that is used in our subjective test. From the control panel of the subjective test, a participant has a possibility to check three reference videos that are constantly encoded at three different video quality levels for the whole period of


Do you see a change of video quality while watching the video?

PS: If you are not sure, click 'Current Video' button to replay the video, or click one of the REF buttons for watching a video that is considered as no quality change

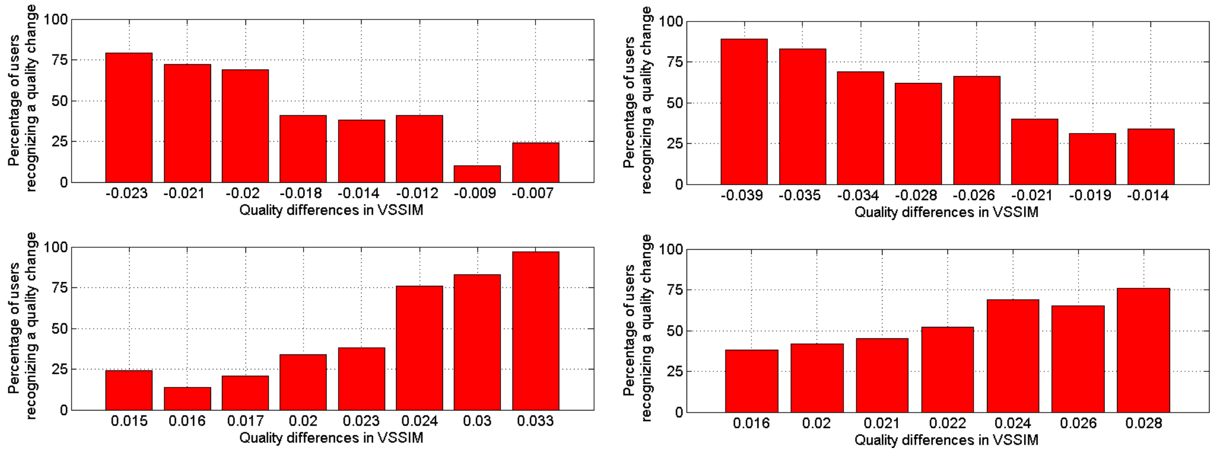
Figure 4.8: Screenshot of GUI used in the subjective test. [151]

Table 4.1: Test conditions for video sequences in user test. [151]

Video content	Rate (kbps)	ξ (MOS)	τ (sec)
Mother and Daughter	[20;120]	[-0.23;0.33]	[2;4]
Foreman	[65;200]	[-0.39;0.28]	[2;4]

the video. In particular, these references help participants to decide whether the test video sequence being watched should be regarded as a perceivable change of video quality, in case the participant is uncertain on his/her decisions. In order to control the period of time that each participant needs to finish their subjective test, we allow each participant to view each test video sequence at maximum of 3 times. To replay the current test video sequence, he/she just clicks on the 'Current Video' button. With this, the participant can reconsider their opinion prior to giving their answer for each test video sequence. Also, we do not employ a time limit (in seconds) for each test video sequence, so as the participant can carefully give their answers without being under pressure due to the time constraint. As soon as the participant is sure about giving their answers, he/she continues watching the next test video sequence by clicking the 'Next Video' button. If all test video sequences of the same video content have been voted, the participant watches the next video set with different video contents. With the 'End' button, the participant indicates that he/she has finished the subjective test and we then save all his/her votes for the statistical analysis of JND.

From the subjective tests, we collect the statistics of how many users recognize the temporal video quality change with different magnitude of ξ as depicted in Figure 4.9. Results show that the JND for 'Mother and Daughter' video is -0.02 and 0.024 for negative and positive change of video quality respectively. Whereas, for 'Foreman' video, we have JND of -0.026 for negative change and 0.022 for positive change of video quality. These JNDs are the threshold at which 15 persons (50 percent of all subjects) are able to recognize the change of temporal video quality. From the results, we conclude that the JND for both video sequences in negative and positive changes are pretty close.



(a) (b)

Figure 4.9: Percentage of users recognizing a video quality change for (a) 'Mother and Daughter', and (b) 'Foreman' video sequence.

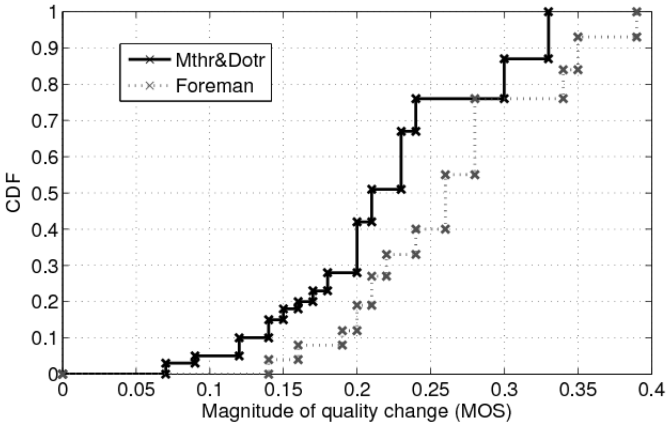


Figure 4.10: CDF of magnitude of perceivable quality change. [151]

Another observation from the result of the user test is that the JND in VSSIM or MOS scale seems not to deviate much, when comparing the two cases of encoding the video with very good quality and with an intermediate quality. This is different than the Weber's law, which states that for people to really perceive a difference, the stimuli (magnitude of change in tested video sequence) must differ by a constant "proportion" not a constant "amount". Taking Weber's law into account, the JND in MOS scale from our subjective test should be different depending on the original video quality level prior to the change. However, this is not the case for our subjective test as shown in Figure 4.9. One of the reason is that the VSSIM has presumably taken into account such influence of initial intensity.

To incorporate these findings in the framework of QoE-driven resource allocation optimization, we consider the JND of the absolute value of ξ from all persons participating in the subjective test. Using the linear mapping between the VSSIM and the MOS in Eq. (3.14), we plot the CDF of the absolute value of ξ from all persons participating in the subjective test as shown in Figure 4.10. We see that the JND for 'Mother and Daughter' and for 'Foreman' is 0.21 MOS and 0.26 MOS respectively. We apply the average JND threshold of ξ_{th}= 0.23 MOS to our QoE-driven optimization framework so as to achieve unperceivable change of temporal video quality as much as possible when allocating network resources and adapting the video data rate as a result of optimization.

4.2.2 Temporal quality smoothness maximization

We enhance the objective function in Eq. (4.4) by taking into account the threshold of perceivable quality change ξ_{th} as:

$$\tilde{\alpha}_{opt} = \arg\max_{\tilde{\alpha}\in\alpha_S} \left[\frac{1}{N} \left(\sum_{i=1}^{N} U_i(\tilde{\alpha}_i) - \beta \sum_{i=1}^{N} (\xi_i - \xi_{th}) \right) \right] \tag{4.5}$$

subject to $\sum_{i=1}^{N} \tilde{\alpha}_i = 1$

where

$$\xi_i = |U_i(\tilde{\alpha}_i(t-1)) - U_i(\tilde{\alpha}_i(t))| \tag{4.6}$$

$$\beta = \begin{cases} 0, \text{ if } \xi_i < \xi_{th} \\ 1, \text{ if } \xi_i \geq \xi_{th} \end{cases} \tag{4.7}$$

t is the notion of time scale in second, and $U_i(\tilde{\alpha}_i(t-1))$ is the average MOS for user i with the fraction of allocated resource $\tilde{\alpha}_i$ during the last 1 second in the past. The subtracted element in Eq. (4.5) is regarded as a penalty parameter, which negatively affects the overall perceived quality, if the temporal change of video quality exceeds the threshold ξ_{th}. β is a weighting factor used for giving priority for the smoothness of temporal video quality. Using the enhanced objective function in Eq. (4.5), the optimizer allocates resources such that all active users (if possible) experience a smooth and unperceivable change of temporal video quality even in the presence of a drastic change of wireless channel condition, while maintaining the system efficiency in terms of network resource utilization as high as possible. With the β defined in Eq. (4.7), we use a linear function to penalize the conventional utility maximization. To increase the impact of the penalty term, an operator may use for example a quadratic function by defining β as follows:

$$\beta = \begin{cases} 0, \text{ if } \xi_i < \xi_{th} \\ \sigma \cdot \xi_i + \rho, \text{ if } \xi_i \geq \xi_{th} \end{cases} \tag{4.8}$$

where σ and ρ are constant parameters that are set prior to network resource allocation optimization. For example, we use the value 100 and -22 for σ and ρ respectively. The optimal resource allocation with the new objective function is achieved by using a greedy search algorithm as summarized in Algorithm 4.

Algorithm 4 Greedy Algorithm for Smoothness Maximization

```
Input: Utility function $U$, number of users $N$, step size of resource $\Delta\alpha$, increase of step size $\Delta\alpha_{inc}$, minimum change of utility gain $\Delta G_{min}$, maximum number of iterations $I_{max}$, perceivable threshold $\xi_{th}$, prioritized weighting factor for smoothness $\beta$.
Output: Optimal operating mode $\tilde{\alpha}_{opt}$;
Initialization: initial resource share: $\tilde{\alpha} = [1, 0, 0, ..., 0]$, set $\Delta G_{max,inc}$ to a value greater than $\Delta G_{min}$. Iteration index, $I = 0$.
for $i = 1$ to $N$ do
    Compute $U_i(\tilde{\alpha}_i(t))$,
    Compute $\xi_i(\tilde{\alpha}_i(t)) = |U_i(\tilde{\alpha}_i(t-1)) - U_i(\tilde{\alpha}_i(t))|$
    Compute utility gain $G_i(\tilde{\alpha}_i(t)) = U_i(\tilde{\alpha}_i(t)) - \beta \cdot (\xi_i(\tilde{\alpha}_i(t)) - \xi_{th})$
end for
loop
    for $i = 1$ to $N$ do
        get operating mode $\tilde{\alpha}_{inc,i}$ from $\tilde{\alpha}_i + \Delta\alpha$, where $\tilde{\alpha}_{inc,i} \in \alpha_{S,i}$;
        get operating mode $\tilde{\alpha}_{dec,i}$ from $\tilde{\alpha}_i - \Delta\alpha$, where $\tilde{\alpha}_{dec,i} \in \alpha_{S,i}$;
        compute $\Delta U_i(\tilde{\alpha}_{inc,i})$ , $\Delta U_i(\tilde{\alpha}_{dec,i})$ , $\Delta\xi_i(\tilde{\alpha}_{inc,i})$ , $\Delta\xi_i(\tilde{\alpha}_{dec,i})$ , $\Delta G_i(\tilde{\alpha}_{inc,i})$ and $\Delta G_i(\tilde{\alpha}_{dec,i})$;
    end for
    if $\Delta G_{max,inc} < \Delta G_{min}$ then
        set $\Delta\tilde{\alpha}$ to $\Delta\tilde{\alpha} + \Delta\tilde{\alpha}_{inc}$
    else
        find $i^+ = \arg\max_{i \in N}\{\Delta G_i(\tilde{\alpha}_{inc,i}) | \tilde{\alpha}_i \leftarrow \tilde{\alpha}_i + \Delta\alpha\}$
        find $i^- = \arg\min_{i \in N}\{\Delta G_i(\tilde{\alpha}_{dec,i}) | \tilde{\alpha}_i \leftarrow \tilde{\alpha}_i - \Delta\alpha\}$
        $\Delta G_{max,inc} = \Delta G_i(\tilde{\alpha}_{inc,i^+}) - \Delta G_i(\tilde{\alpha}_{dec,i^-})$
        set $\Delta\tilde{\alpha}$ to $\Delta\tilde{\alpha}_{inc}$
    end if
    $I++$;
    if $I > I_{max}$ then
        break;
    end if
end loop
output: $\tilde{\alpha}_{opt}$
```

Performance evaluations

We use the HSDPA simulator to emulate a resource-constrained single cell scenario, in which six users access different video contents at high data rates and experience different wireless channel conditions. The parameters used in our simulations are the same as already given in Table 3.2. The wireless channel model in the HSDPA simulator is based on the measured CQI trace representing different mobility schemes under different environments. To evaluate the performance of the new objective function given in Eq. (4.5), we compared the proposed scheme with the other four schemes (*No-adaptation*, *MaxRate*, *MaxMOS* and MaxMin-MOS), which have already been described in details in Section 3.8.1. The proposed scheme is briefly summarized as following:

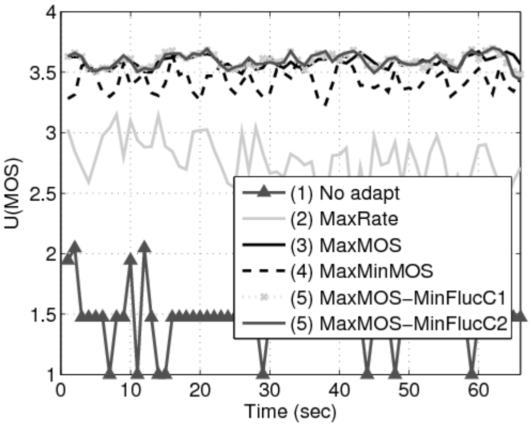
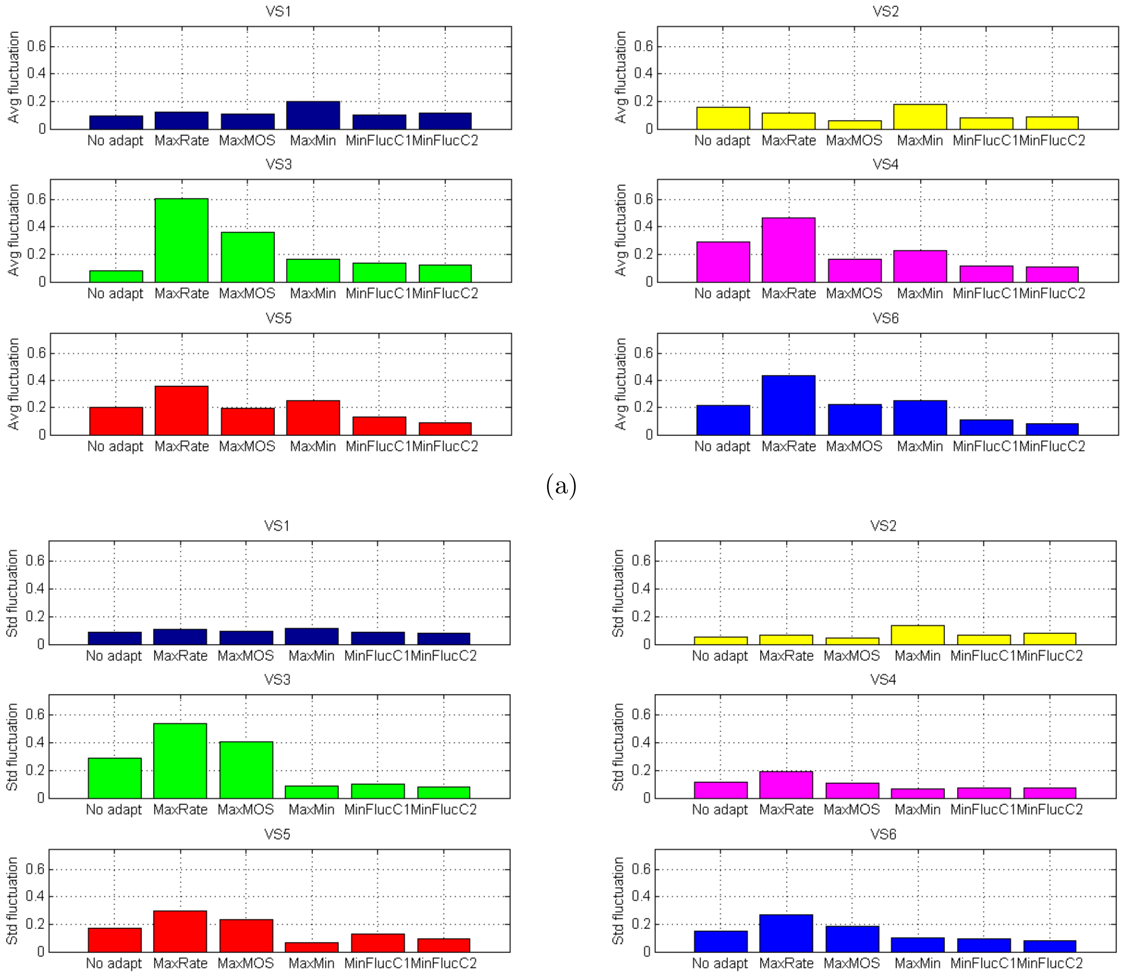


Figure 4.11: Plot of mean utility of all users as a function of simulation time. [151]

- *MaxMOS-MinFluc*: In addition to utility maximization as done in *MaxMOS*, the proposed scheme in Eq. (4.5) performs the resource allocation such that the change of temporal video quality lies within the unperceivable threshold ξ_{th}. Based on the results of the subjective test, we assume that users are able to perceive a change of video quality with the average threshold of 0.23 on the MOS scale for all video contents. The two cases of β described in Eq. (4.7) and (4.8) are denoted as 'Case 1 (C1)' and 'Case 2 (C2)' respectively. These two cases are used to see its effect when giving higher priority to the temporal quality fluctuation objective.

Figure 4.11 shows the mean MOS over all users over the simulation period of 1 min. Like in other simulation results, we see a significant gain between all application aware schemes and the no-adaptation scheme. A further gain of 0.5 on the MOS scale is achieved when applying our proposed MOS-based optimization (scheme 3 to 5) when compared to the throughput maximization scheme. Among all MOS-based schemes, maxmin fairness is the worst. The MaxMOS and our proposed scheme perform approximately the same.

Figure 4.12(a) and 4.12(b) depict the results of each UE for the average temporal quality change ξ and its standard deviation. In this scenario, VS1 and VS2 are users experiencing a good and stable wireless channel condition compared to other video users. Also, their videos contain static video scenes, whereas others watch a dynamic video content and experience worse and dynamic channel quality. It is obvious that the throughput maximization (MaxRate) performs the worst, as the resources are allocated based on the wireless channel condition. This results into a high fluctuation of user-perceived quality in presence of drastic change of wireless channel condition. The MaxMOS scheme partly follows the MaxRate-based strategy with respect to the wireless channel condition, however it considers whether the allocated resources contribute to the maximally increment of average mean MOS of all users. Hence, the average quality fluctuation for the MaxMOS scheme is less than the MaxRate scheme. In contrast, the MaxMinMOS scheme performs fairly equal for all users due to the fact that all users perceive similar quality. Thus, if one



(a)

(b)

Figure 4.12: Simulation results of each UE: (a) average video temporal quality fluctuation ξ, (b) standard deviation of ξ from a simulation run of 6 video users.

user experiences a bad wireless channel condition at one moment due to his mobility or his location such as at the edge of cell coverage, all other users would experience a drop of their perceived quality so as to achieve similar quality for all users and thus causing quality change to all of them. Nevertheless, the average ξ and the standard deviation of ξ are slightly different for each user due to the discrete operation mode from different video utility functions. The proposed schemes (MinFlucC1 and MinFlucC2) outperform all other approaches by achieving less temporal quality fluctuation in both the average and the standard deviation. It should be noted that the no-adaptation scheme results in a lower quality fluctuation than the MaxSum and the MaxMin schemes, as many users perceive minimal quality for almost the whole period of the simulation time.

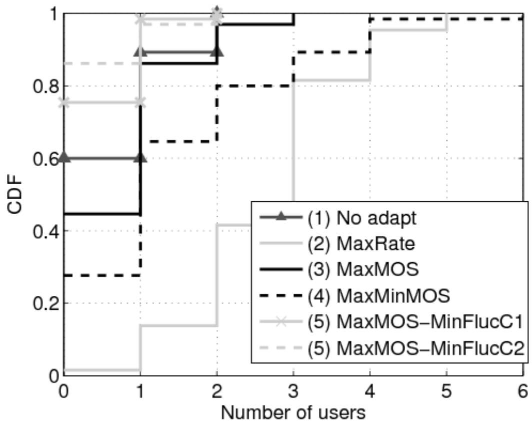


Figure 4.13: CDF of number of users experiencing quality changes exceeding the experimentally determined perception threshold. [151]

To complement the advantages of the proposed scheme, we perform a number of simulation runs and plot the CDF of the number of users experiencing the changes of temporal video quality exceeding the threshold ξ_{th} as shown in Figure 4.13. We see that three video users in average (50 percent of all users) perceive the temporal quality change when the rate-based optimization is applied. With the MaxMOS and the MaxMin scheme, less number of users perceive the change of temporal video quality. The proposed scheme achieves the best user-perceived quality compared to other schemes, as the users experience a smooth change of quality even in the presence of drastic changes of the wireless channel condition. By applying the *MaxMOS-MinFluc* scheme with Case 2, the result is further improved, as fewer users are able to perceive a temporal video quality fluctuation due to the fact that more priority is given to the temporal quality smoothness.

4.3 Optimization with system efficiency, quality fluctuation and fairness

In previous sections, we discussed how to combine two criteria into the optimization problem of network resource allocation, for example, a combination of mean utility of all users (system efficiency e) and quality differences among users (user fairness k), or a combination of system efficiency and temporal quality fluctuation ξ. In addition, we showed that there are several factors that influence allocating resources across multiple users, and it is difficult to come up with a mathematical formular covering all issues. Hence, we propose to use a heuristic and iterative approach to solve the two-criteria optimization problem.

In this section, we discuss how to integrate all three criteria (the system efficiency e, the user quality fairness k and the temporal quality fluctuation ξ) into the QoE-driven optimization. Solving multiple criteria optimization problem is actually not a new dimension.

In the last decades, there has been a large number of literatures addressing and solving multiple criteria optimization problem. One of the good literature reviews is summarized by Marler *et al.* [104]. Therein, authors stated that there is no simple answer to which method to use for a particular problem, as each optimization method has different properties suited for different type of problems. By looking at our resource allocation optimization problem in wireless networks with three criteria aforementioned, we have chosen the priori-articulation of preferences approach assuming that the network operator wants to be flexible in specifying its preferences among multiple criteria prior to the optimization.

4.3.1 Weighted-sum method

Since the multiobjective optimization problem is usually characterized by the presence of many conflicting objectives, it is unlikely that there exist a solution, which simultaneously fulfills all the objectives. Furthermore, there exist sometimes no clear relationship between the objectives, and thus making the optimization even more complicated.

To see a relationship between the three criteria in our QoE-driven optimization problem, we have simulated a single HSDPA cell scenario with 8 video users by applying the normalized weighted-sum method to our multi-criteria optimization problem, which can be aggregated to one single figure of merit as follows:

$$\tilde{\alpha}_{opt} = \arg\min_{\tilde{\alpha}\in\alpha_S} \left[\left(\frac{\lambda_1 \cdot (4.5 - f_1(\tilde{\alpha}))}{3.5} \right) + \left(\frac{\lambda_2 \cdot f_2(\tilde{\alpha})}{3.5} \right) + \left(\frac{\lambda_3 \cdot f_3(\tilde{\alpha})}{3.5} \right) \right] \tag{4.9}$$

subject to

$$\textstyle\sum_{i=1}^{N} \tilde{\alpha}_i = 1$$

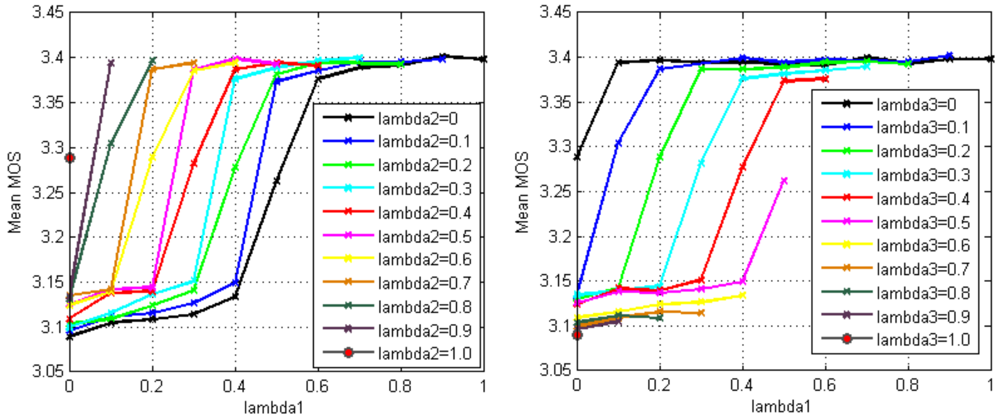
$$\lambda \in R|\lambda_j \geq 0 \ , \ \textstyle\sum_{j=1}^{3} \lambda_j = 1$$

where $f_j(\tilde{\alpha})$ is the j-th objective function that depends on the resource distribution $\tilde{\alpha}$ across N users and λ_j is a weighting factor that allows a network operator to prioritize the j-th objective function. In this work, $f_1(\tilde{\alpha})$ aims to maximize the average utility of all users. $f_2(\tilde{\alpha})$ focuses on minimizing the perceivable temporal quality fluctuation and $f_3(\tilde{\alpha})$ aims to minimize the quality difference among all users. We formulate the objective functions $f_1(\tilde{\alpha})$, $f_2(\tilde{\alpha})$ and $f_3(\tilde{\alpha})$ as follows:

$$f_1(\tilde{\alpha}) = \frac{1}{N} \sum_{i=1}^{N} U_i(\tilde{\alpha}) \tag{4.10}$$

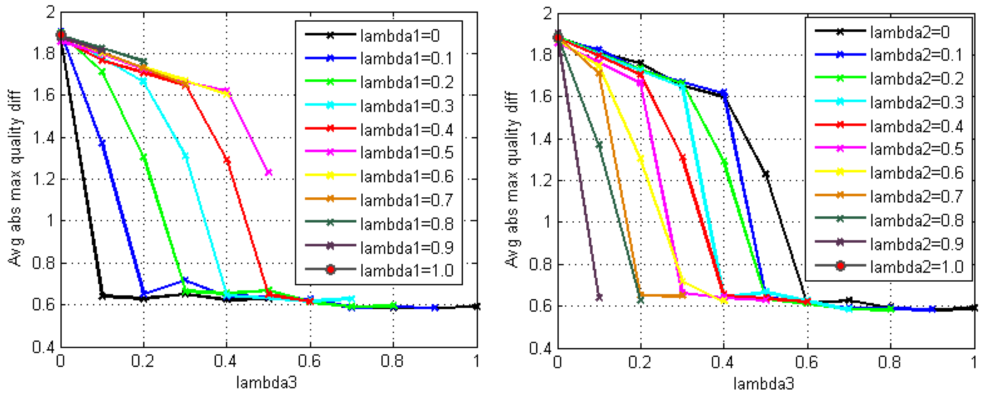
$$f_2(\tilde{\alpha}) = \frac{1}{N} \sum_{i=1}^{N} |\xi_i(\tilde{\alpha}) - \xi_{th}| \tag{4.11}$$

$$f_3(\tilde{\alpha}) = \frac{1}{N} \sum_{i=1}^{N} |U_i(\tilde{\alpha}) - U_{avg}| \tag{4.12}$$



(a) (b)

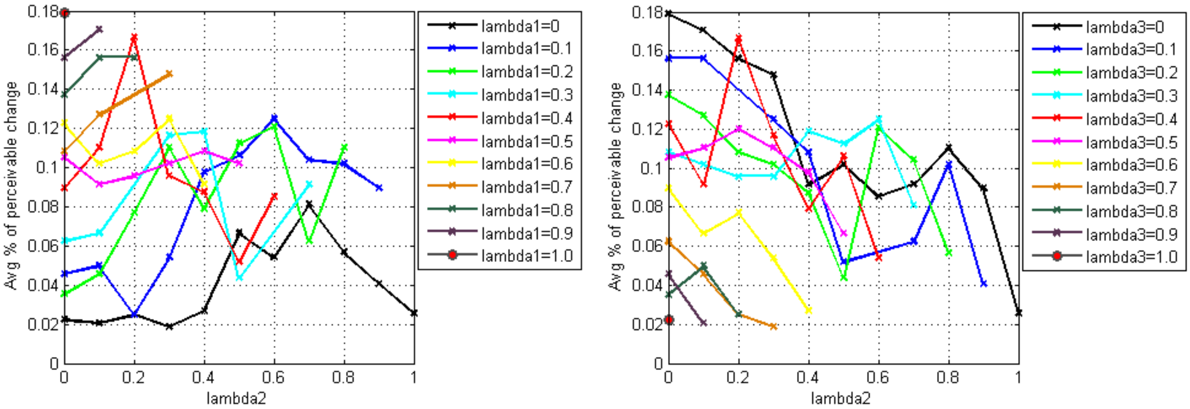
Figure 4.14: Mean utility as a function of λ_1 by fixing λ_2 (a), and λ_3 (b).



(a) (b)

Figure 4.15: Average maximum quality difference k_{avg} as a function of λ_3 by fixing λ_1 (a), and λ_2 (b).

where U_{avg} is the average utility of all users based on the allocated resource allocation as defined in Eq. (4.10). We run simulations for a single HSDPA cell scenario with 8 video users by varying λ_1, λ_2 and λ_3 with the step size of 0.1. Figure 4.14, 4.15 and 4.16 show the results in terms of average utility U_{avg}, average maximum quality difference k_{avg} and average percentage of perceivable temporal quality fluctuation ξ_{avg} respectively. Each point represents a combination of λ_1, λ_2 and λ_3. In Figure 4.14 and 4.15, we see that the maximum average quality of all users is about 3.4. The maximum mean MOS is achieved by almost all of the case of fixing λ_2, except when λ_2 is equal to 1 as depicted in Figure 4.14(a). In contrast, by fixing λ_3, the maximum mean MOS can also be achieved but only for the case of having λ_3 setting from 0 to 0.3. This results show that setting a higher priority for the quality fairness objective comes at a cost of the mean MOS. In particular, by setting the λ_3 to 1, we achieve a minimum average quality of all users about 3.09 MOS.



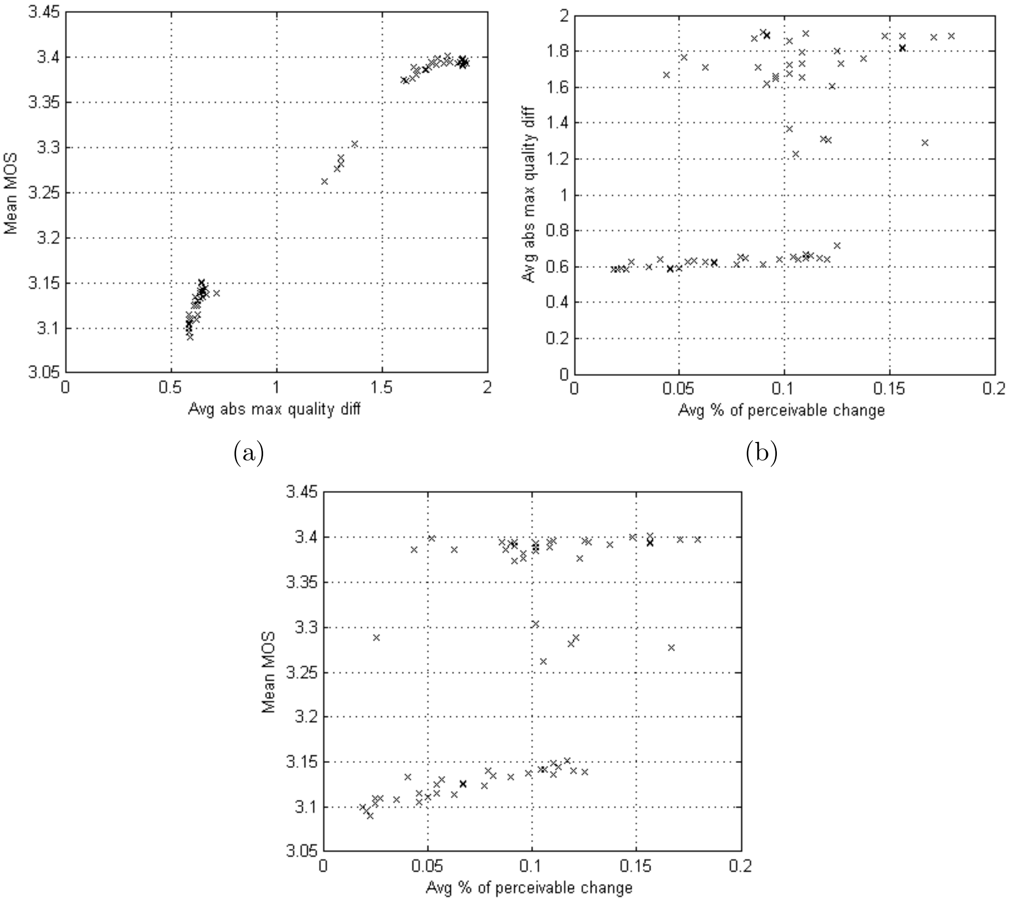
(a) (b)

Figure 4.16: Mean percentage of perceivable change of temporal quality as a function of λ_2 by fixing λ_1 (a), and λ_3 (b).

Figure 4.15(a) and 4.15(b) show the maximum quality difference among users when fixing λ_1 and λ_2 respectively. Obviously, in both cases, when the λ_3 increases, the quality difference among users is decreasing, and thus resulting to more quality fairness. The minimum quality difference from the considered scenario is about 0.6. Fixing higher λ_1 results to less maximum achievable quality fairness, but a higher maximum mean MOS can be achieved as shown in Figure 4.14(b).

Figure 4.16(a) and 4.16(b) show the result of the average percentage of number of users that perceive the temporal quality fluctuation as a function of λ_2 when fixing the weighting factor for the average mean MOS λ_1 and for the quality fairness λ_3. Unlike the previous results, we observe that increasing λ_2 does not guarantee that we will receive less users perceiving temporal change of video quality. One of the reasons is that both objectives of quality fairness maximization and utility maximization do not have a clear relationship with the temporal quality smoothness. For example, results discussed in Section 4.2.2 show that it is possible to find a resource allocation so as to minimize the perceivable temporal quality fluctuation while maintaining the average quality as high as possbile as of achieving by the utility maximization scheme. But having more perceivable temporal change does not necessarily mean that the mean MOS will decrease.

This observation becomes more obvious when looking at the scatter plot of the result from all combinations of λ_1, λ_2 and λ_3 as depicted in Figure 4.17(a), 4.17(b) and 4.17(c). In Figure 4.17(a), we see that there is a clear relationship between U_{avg} and k_{avg}. A higher average MOS is achieved when the average quality difference among users is getting larger. Whereas, no relationship can be concluded for the other two results shown in Figure 4.17(b) and 4.17(c). For instance, allocating resources with an aim of having a minimal perceivable temporal quality fluctuation could result to both high and low fairness.



(a) (b)

(c)

Figure 4.17: Relationship among three criteria: (a) average maximum quality difference and average MOS, (b) average perceivable quality fluctuation and average maximum quality difference, and (c) average perceivable quality fluctuation and average MOS.

4.3.2 Hybrid-lexicographic method

To avoid uncertainty of allocating resources in presence of having a clear policy, we decouple the fairness and the quality fluctuation and apply the hybrid-lexicographic method for the multi-criteria QoE-driven optimization. It is hybrid-lexicographic, as we perform the sequential optimization of two steps, and each step consists of two criteria as described below.

1. Fairness+Efficiency: Finding an optimal resource allocation given that the network operator has specified its requirement of the average MOS e_{req} and the fairness k_{req} through the advanced k-algorithm as described in Algorithm 3.
2. Smoothness+Efficiency: Finding an optimal allocation resulting to a smooth temporal quality fluctuation while keeping the average MOS as high as possible as done in

Algorithm 5 2loopKcon algorithm

```
**Input**: Utility function $U$, number of user $N$, step size of resource $\Delta\alpha$, increase of step size $\Delta\alpha_{inc}$, minimum change of utility gain $\Delta G_{min}$, maximum number of iterations $I_{max}$, perceivable threshold $\xi_{th}$, prioritized weighting factor for smoothness $\beta$, quality unfairness constraint $k_{req}$, system efficiency constraint $e_{req}$, weighting coefficients for fairness and efficiency $c_k$ and $c_e$.
**Output**: Optimal operating mode $\tilde{\alpha}_{opt}$;
**Initialization**: zero resource share: zero resource share: $\tilde{\alpha} = [0, 0, ..., 0]$, $U = [1, 1, ..., 1]$, set $\Delta G_{max,inc}$ to a value greater than $\Delta G_{min}$. Iteration index, $I = 0$.
**call:** advanced k-algorithm (Algorithm 3).
Output of advanced k-algorithm $\rightarrow p1$ operating mode $\tilde{\alpha}_{opt,p1}$, and $k_{end}$
**call:** smoothness maximization algorithm (Algorithm 4). Imposing a fairness constraint of $k_{end}$ prior to taking resource $\Delta\tilde{\alpha}_{dec}$ from user $i_{-}$ and allocating $\Delta\tilde{\alpha}_{inc}$ to user $i_{+}$.
Output of smoothness maximization $\rightarrow p2$ operating mode $\tilde{\alpha}_{opt,p2}$
**output**: $\tilde{\alpha}_{opt} \leftarrow p2$
```

the MaxMOS-MinFluc algorithm described in Algorithm 4.

Note that we consider the average utility in both steps, as it is an important measure of the whole system performance. Also, the network operator usually aims to achieve their system performance as high as possible. The first-step ensures that the resource allocation satisfies both e_{req} and k_{req} constraints, whereas the second-step ensures that the temporal quality fluctuation is kept as minimal as possible while maximizing the average user-perceived quality of all users. Initialization of the 2nd-step optimization is the result of optimal resource alocation from the 1st-step optimization. This affects the search for the 2nd-step optimal solution such that the quality difference among users is not far apart from the result of the 1st-step optimization when compared with other initialization points. We name the two-step optimization as "2loop" optimization.

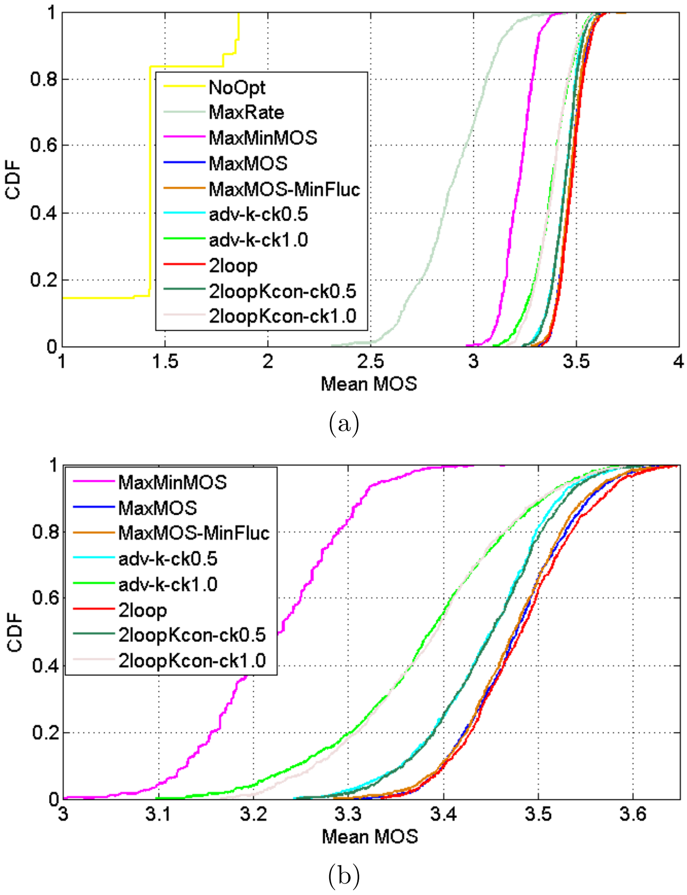
In case, the network operator would like to make the fairness constraint more important, the k_{req} can be added into the 2nd-step of "2loop" optimization. We call this variant of "2loop" as the "2loopKcon" optimization scheme whose the algorithm is summarized as shown in Alogirthm 5.

4.3.3 Simulation results

We do simulations of a HSDPA single cell scenario, in which eight users are accessing videos over the HSDPA downlink shared channel. To evaluate the performance of the proposed "2loop" and "2loopKcon" schemes, we compare them with the "No adaptation (NoOpt)", "Throughput Maximization (MaxRate)", "Max-Min utility (MaxMinMOS)", "Utility Maximization (MaxMOS)", "Smoothness Maximization (MaxMOS-MinFluc)" and "advanced k-algorithm (adv-k)". Details of the first four schemes, the adv-k and the MaxMOS-MinFluc are already described in Section 3.8.1, 4.1 and 4.2.2 respectively.

For the adv-k, 2loop and 2loop-kcon schemes, we set the target mean utility e_{req} and the target quality fairness k_{req} to 3.5 and 1 in MOS scale. In case, both requirements can't be met at the same time, we apply the prioritization coefficients c_e and c_k.

Figure 4.18(a) shows the CDF plot of the mean MOS of all eight users from 30 simulation runs for each scheme. Like other simulation results presented in this thesis, all QoE-driven optimization approaches lead to significant improvements of user-perceived quality when compared with the non-optimized HSDPA system and the rate-based adaptation scheme. The least improvement among all QoE-based schemes is the MaxMinMOS, as the resources are distributed so as all users perceive a similar quality regardless of the video utility function and the wireless channel condition. Three QoE-based schemes (MaxMOS, MaxMOS-MinFluc and 2loop) result to the most improvement, as all of them focus on achieving a maximum mean MOS. By considering the fairness constraint, it comes at a cost of the mean MOS, e.g. the more importance we put on the fairness, the less mean MOS we receive. This effect can be seen when looking at the "adv-k" and "2loopKcon" results



(a)

(b)

Figure 4.18: CDF of mean utility for 8 video users: (a) comparison of all 8 schemes, (b) zoom view of all QoE-based schemes.

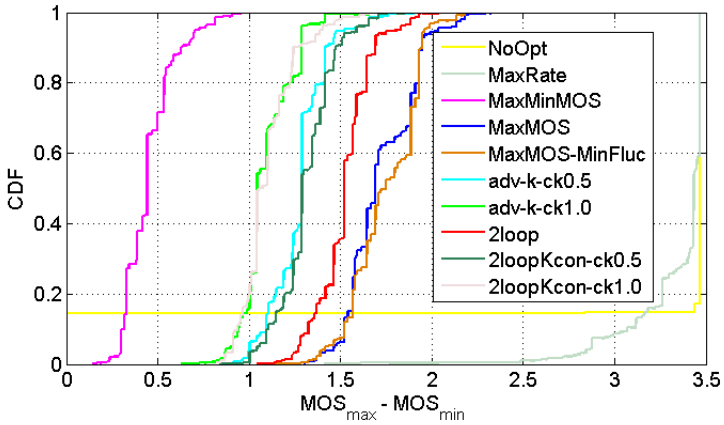


Figure 4.19: CDF of maximum quality difference among users.

with different values of c_k. In average, having a hard k_{req} constraint (with $c_k = 1$), the mean MOS is reduced by an average of 0.1 in MOS. Whereas with a softer k_{req} constraint (with $c_k = 0.5$), we only have 0.03 of MOS reduction from the maximum achievable mean MOS. It is to be noted that the mean MOS for both "adv-k" and "2loopKcon" schemes are similar, though the "2loopKcon" algorithm employs the MaxMOS-MinFluc in the 2nd-step optimization as discussed earlier.

Figure 4.19 shows a CDF comparison of all schemes for the average of maximum quality difference among users representing the system performance in terms of quality fairness. Both non-optimized HSDPA system and throughput maximization schemes have a minimal fairness, as their adaptation is only based on the wireless channel condition. The resources are hence often given to the users experiencing a good channel quality. Whereas the QoE-based schemes optimally allocate the network resources by taking into account the gain of user-perceived quality (utility) prior to network resource allocation. Among all QoE-based schemes, the most fair approach is the Max-Min utility scheme. Both MaxMOS and MaxMOS-MinFluc similarly perform the worst in terms of fairness due to the fact that they mainly concentrate on other objectives of utility maximization and smoothness maximization without considering the fairness. For instance, the MaxMOS first allocates its resources to the user having better channel quality and accessing the video that is low demanding but results in high user-perceived quality such as a static video. Until the gain of giving resource to the user is less than others, the MaxMOS starts allocating network resources to a user watching a high-demanding video or experiencing a poor channel condition. All schemes that consist of k-algorithm ("adv-k", "2loop" and "2loopKcon") take into account the fairness constraint while optimizing its resource allocation, and thus resulting to a higher fairness when compared to the MaxMOS and the MaxMOS-MinFluc schemes. By increasing the value prioritization coefficient c_k, we could achieve less quality difference among users (high quality fairness).

Results of number of users recognizing the temporal quality change that exceeds the per-

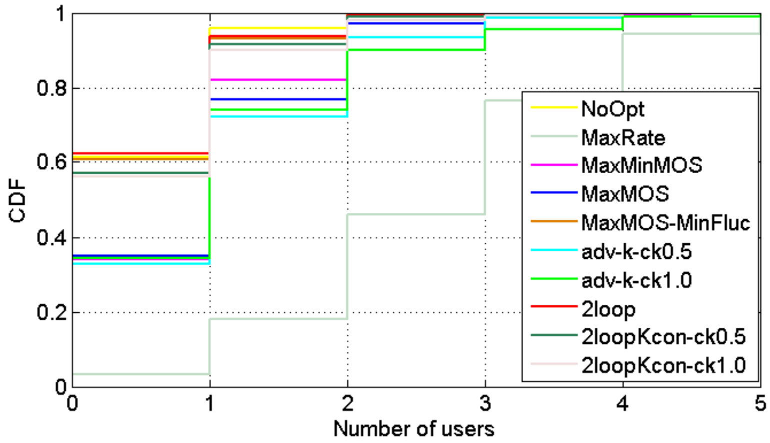


Figure 4.20: CDF of number of users experiencing quality changes exceeding the perceivable threshold.

ceivable threshold ξ_{th} are shown in Figure 4.20. We see that the rate-based optimization results to the most perceivable fluctuation as it adapts the application data rate and allocates resources based on the wireless channel condition. With the MaxMOS, MaxMinMOS and adv-k schemes, there are less number of users perceiving the temporal video quality fluctuation. Both MaxMOS-MinFluc and 2loop perform the best in terms of unperceivable change of temporal video quality as their objective is to minimize the perceivable temporal quality fluctuation. By applying the k_{req} constraint to the 2loop optimization scheme (2loopKcon), we receive a marginal performance degradation of smooth temporal quality change. This implies that if we allow a little increase of number of users being able to perceive temporal quality change, we can get closer to the desired maximum quality difference among user k_{req}. Nevertheless, we observe that changing the c_k for both adv-k and 2loopKcon scheme does not have any impact on the system performance in terms of temporal quality fluctuation. For the result of no-adaptation scheme, it should be noted that there is also less number of users perceiving temporal quality fluctuation, as many users are perceiving minimal quality for almost the whole period of the simulation time due to the congestion at the base station and no application-layer adaptation being applied when network gets congested.

4.4 Summary

This chapter has addressed the multi-criteria optimization problem of network resource allocation in a loaded single cell scenario where users consume only video applications on their mobile devices. First, we introduce a novel tuning mechanism that allows the network operator to adapt its network resource allocation based on its policy in terms of the mean perceived quality of all users (system efficiency) and the quality fairness among users. The proposed algorithm makes use of the utility maximization and the utility max-min fairness, and extends them by introducing additional constraints of system efficiency and user fairness that are defined by the network operator prior to the optimization of network resource allocation. Three variants of tuning algorithms are introduced, which allows the network operator to set their requirements differently, e.g., only the system efficiency, or only the user fairness, or both of them. Simulation results show the feasibility of the proposed tuning algorithms.

Next, we discuss a novel QoE-based objective function for cross-layer optimization of wireless video, which considers the change of temporal video quality and the corresponding human perception threshold into the overall user-perceived quality rating. The threshold is based on the Just Noticeable Difference (JND) concept and is derived by performing subjective tests. The proposed scheme is implemented in the resource allocation optimization across multiple users accessing different video contents and being present in the same wireless cell. The goal of the optimization is to achieve minimal perceivable change of temporal quality while at the same time maintaining the average perceived quality of all users as high as possible. Results show that our proposed scheme achieves a better user-perceived quality compared to other schemes, as the users experience a smooth change of quality even in presence of drastic changes of the wireless channel conditions.

Lastly, we consider all three criteria (the system efficiency, the user-perceived quality fairness and the smooth temporal quality fluctuation) in the QoE-driven optimization problem. Due to the fact that one criteria has an impact on another, we propose a practical implementation of multi-criteria optimization based on the hybrid-lexicographic method. The proposed two-step optimization enables us to decouple the interdependency between the quality fairness and the temporal quality fluctuation. The 1st-step optimization considers both the mean perceived quality of all users and the quality fairness among users by applying the tuning algorithms as discussed earlier. In the 2nd-step optimization, we apply the aggregate objective function consisting of the utility maximization and the perceivable temporal quality fluctuation minimization. Alternative, one can add the fairness constraint in the 2nd-step optimization in order to direct the optimal resource allocation such that the maximum quality difference among users is getting close to the fairness constraint as much as possible. Results show that it is feasible to find an optimal resource allocation with the two-step optimization that tries to fulfill all three criteria as much as possible without degrading a lot of system performance in terms of mean utility, quality fairness and unperceivable temporal quality fluctuation.

Chapter 5

Conclusion and Outlook

The continued growth of data traffic on mobile networks driven by the rapidly increasing use of 'always on' smart devices (e.g., smart phones, tablets, PCs using dongles, etc.) that support various and numerous applications is congesting the mobile networks, and thus degrading user service experience. To cope with the growing demands, the network operator may add more base stations and enhance its network facilities in its core/backhaul network, however, such investments are usually expensive and not cost effective. Moreover, although the data capacity of networks has been increased significantly, the observed increase in traffic continues to outpace the growth in capacity. Hence, the operators are looking at ways of managing the growth of traffic efficiently to deliver acceptable level of Quality of Experience (QoE) to their users in the presence of constrained network resources, which result in reduction of their expenditures during the downturn of economy.

In this thesis, a QoE-driven optimization for resource allocation and rate adaptation has been proposed for mobile multimedia communication. The optimization framework consists of two main functional entities: the Traffic Management (TM) and the Traffic Engineering (TE). The TM acts as a resource allocator that jointly optimizes the application layer and the lower layers of wireless protocol stack with an aim of improving the user perceived quality. The optimizer periodically reviews the total system resources and makes an estimate of the time-share needed for each user for each possible application-layer rate. If necessary, the optimizer suggests re-adaptation of the application rates. The TE performs actual rate shaping and adaptation according to the instruction given by the TM. In order to compare the proposed QoE-based optimization schemes with other approaches such as throughput maximization and non-optimized system, we perform simulations using a software implementation of a developed High-Speed Data Packet Access (HSDPA) system. In the following, we summarize the major results, draw conclusions, and give directions for future work.

Application Utility Functions

We formulate multimedia QoE by constructing long-term utility functions, which use the Mean Opinion Score (MOS) as a unified utility metric that encompasses the user-perceived quality under certain receiving conditions for all application types. The utility function is simplified as a function of transmission data rate by assuming that all packets are most likely to be transmitted successfully due to the HSDPA MAC-layer retransmission mechanism providing a more reliable transmission over the wireless interface. Three application types (voice call, video streaming and file transfer) are considered and their utility functions were presented, when changing the transmission data rate at the server or sender side. For voice and video applications, we investigate an impact of in-network rate adaptation, for example, by transcoding or dropping packets in the mobile network. Our conclusions are that our voice utility based on the source encoding rate is sufficient enough, since the impact of transcoding is marginal and can be negligible. Whereas, transcoding the video stream in the network should be taken into account as it has more impact on the user-perceived quality. Our investigation shows that dropping video packets results in much worse video quality than the video transcoding. This tells us that the network operator shall employ transcoding technique and apply it to all video streams in order to minimize the degraded service quality delivered to the user when network gets congested. If not possible due to the hardware constraints, transcoding shall be first applied to the video stream with dynamic content, as it is more sensitive to the packet dropping than the video with static scenes.

Single-criteria QoE-driven Optimization

In a single loaded cell scenario, where limited resources are shared among users running different applications and experiencing different wireless channel quality, we describe the multiuser utility space and derive its properties. Depending on the objective function, the QoE optimizer finds an optimal resource allocation and then sets the applications-layer data rate for each user accordingly. Two conventional utility-based schemes are implemented: utility maximization and max-min utility. The former targets at the maximum average user quality, whereas the latter aims to achieve a similar quality for all users regardless of the application type and the channel quality condition. We show analytically that the maximization of the sum of utility can be efficiently solved by a fast greedy algorithm which searches only through the boundary of the utility space. The QoE-based schemes are compared to the non-optimized HSDPA system and the system configured to maximize the overall throughput. Results show that our QoE-based approach leads to significantly improved user perceived quality compared to the other approaches. In addition to the network resource constrained system, we integrate the novel QoE-based rate adaptation scheme selection in the optimization framework taking into account the hardware computational constraints when performing rate adaptation for multiple flows simultaneously that is computationally expensive such as transcoding. From the obtained simulation results, we see that further improvement of user-perceived quality is achieved by optimally selecting the rate adaptation scheme to be applied for each media flow prior to the QoE-driven resource allocation optimization.

Multi-criteria QoE-driven Optimization

Network operators sometimes prefer to have flexibility in specifying their policies for managing their network resources by inclusion of multiple criteria rather than having only one target goal as discussed earlier. For example, instead of the utility maximization or the max-min utility, an operator may prefer to allocate network resources in order to meet the target average quality of all users and/or the desired maximum quality difference among users. Due to the fact that the optimization problem is influenced by several factors such as the number of users in the cell, the channel variation among users and the application type, we propose to solve the problem by using a heuristic and iterative algorithm which allows the network operator to change dynamically its operating point of resource allocation based on its pre-determined constraints of the average user-perceived quality and the quality fairness among users. From the simulation results of tuning algorithms, we learned that the mean utility and the utility fairness are contrary to each other.

In addition to the two objectives, a novel objective function that minimizes the temporal change of the video quality as perceivable quality fluctuations negatively affect the overall quality of experience has been proposed and combined with the utility maximization. The new aggregated objective function aims to allocate resources such that the fluctuations lie within the range of unperceivable changes that is determined via extensive subjective tests, while maintaining the average quality of all users as high as possible. Our results show that the proposed scheme leads to a noticeable improvement of temporal quality fluctuation, and a similar average utility as for the utility maximization.

Finally, we address the QoE-driven optimization problem for multi-user wireless video delivery with all three criteria aforementioned. Through simulations with the weighted-sum method, we see a clear relationship between the average quality of all users and the fairness, but not the relationship for other objectives such as between the quality fairness and the temporal quality change. To avoid complication and uncertainty when optimizing network resources, a practical two-step optimization based on the hybrid-lexicographic method is used which decouples the fairness and the temporal quality fluctuation. The obtained simulation results show that the proposed two-step optimization scheme is able to search for an optimal resource allocation for all users taking into account all three criteria. In particular, the network resources can be allocated so as to meet the constraints of average utility and quality fairness as much as possible, while keeping the perceivable temporal quality fluctuation at minimum.

Outlook

In this thesis, we have proposed a QoE-driven optimization framework including several algorithms and methodologies for performance evaluation, which can serve as a basis for further research. In the following paragraphs, we briefly discuss interesting subjects for future work in the area of user-plane traffic management and traffic engineering.

The video utility presented in this work is based on the H.264 AVC, which is now widely deployed and supported in many mobile devices. In the near future, it is foreseen that more advanced video codecs such as the H.264 Scalable Video Coding (H.264 SVC) will become available. With the H.264 SVC, the encoded video bitstream contains one or more subset bitstreams (layers) providing different spatial resolution (picture size), different temporal resolution (frame rate) and different video quality. User perception of different scalability modalities should be investigated, so that a SVC utility function can be constructed and used for the optimization framework. For instance, how a user perceives the video quality for a video with low spatial resolution and good video quality and for a video with high spatial resolution but with poor video quality. Though transmitting a scalable video bitstream gives the network operator more flexibility to reduce the bandwidth required for the bitstream through packet (layer) dropping, it comes at a cost of increased complexity in finding an optimal resource allocation and requires a modification of the search algorithm or even the objective function.

The scenario considered in this work assumes that the network operator is in control of application servers. Performing a QoE-driven optimization for legacy application servers, which do not provide any utility functions and the operator is not in control, is important and should be taken into account. The network operator may apply a default utility function, if it knows about the application type running on the user's devices. Nevertheless, the loss in gain of applying such default utility function has to be investigated. Alternatively, the network operator may employ Deep Packet Inspection (DPI) at its gateway network entity to get additional information out of the media flow. An example is the motion vector information that is useful for deriving the video characteristic whether the video content is dynamic or static.

Another challenge that has not been addressed yet is the QoE-driven resource allocation optimization for the user moving across cells. In this case, it is necessary that all QoE optimizers located at each base station communicate and collaborate with each other. For example, if the objective is to achieve a low perceivable temporal quality fluctuation, a history of user-perceived quality at the source base station has to be forwarded to the target base station. Furthermore, prediction of user's mobility trace may be useful information and incorporated into the optimization framework, so as to have a smooth transition for both the handover users and the other existing users served by the target base station.

Chapter 6

Abbreviations

3G	Third Generation
AMC	Adaptive Modulation and Coding
ARQ	Automatic Repeat request
AS	Application Server
CDF	Cumulative Density Function
CLD	Cross-Layer Design
CLO	Cross-Layer Optimization
CQI	Channel Quality Indicator
DMOS	Differential MOS
FEC	Forward Error Correlation
FTP	File Transfer Protocol
GoP	Group of Pictures
GPS	Global Positioning System
HARQ	Hybrid ARQ
HSDPA	High-Speed Downlink Packet Access
IP	Internet Protocol
ITU	International Telecommunication Union
JND	Just Noticeable Difference
LTE	Long-Term Evolution
MAC	Medium Access Control
MOS	Mean Opinion Score
MSE	Mean Square Error
P2P	Peer-to-Peer
PESQ	Perceptual Evaluation of Speech Quality
PSNR	Peak Signal to Noise Ratio
QoE	Quality of Experience
QoS	Quality of Service
RNC	Radio Network Controller

SDU	Service Data Unit
SSIM	Structual SIMilarity
TB	Transport Block
TBS	Transport Block Size
TCP	Transmission Control Protocol
TE	Traffic Engineering
TM	Traffic Management
TTI	Transmission Time Interval
UDP	User Datagram Protocol
UE	User Equipment
UMTS	Universal Mobile Telecommunications System
VQEG	Video Quality Expertise Group
VS	Video Streaming
VSSIM	Video Structual SIMilarity

List of Figures

List of Tables

Bibliography

[1] 3rd Generation Partnership Project. Physical layer aspects of utra high speed downlink packet access. TS 25.848, 3GPP, Mar. 2001.

[2] 3rd Generation Partnership Project. Quality of service (QoS) concept and architecture. TS 23.107, Dec. 2003.

[3] 3rd Generation Partnership Project. High Speed Downlink Packet Access (HSDPA); Overall description. TS 25.308, 3GPP, Dec. 2004.

[4] 3rd Generation Partnership Project. Radio Link Control (RLC) protocol specification. TS 25.322, 3GPP, Dec. 2005.

[5] 3rd Generation Partnership Project. Physical layer procedures (FDD). TS 25.214 V7.1.0, 3GPP, Jun. 2006.

[6] 3rd Generation Partnership Project. UTRAN Iub interface Node B Application Part (NBAP) signalling. TS 25.433, Sep. 2006.

[7] 3rd Generation Partnership Project. Transparent End-to-End Packet-Switched Streaming Service (PSS); Protocols and Codecs. TS 26.234, Dec. 2010.

[8] I. Ahmad, X. Wei, Y. Sun, and Y. Zhang. Video transcoding an overview of various techniques and research issues. *IEEE Transactions on Multimedia*, 7(5):793–804, Oct. 2005.

[9] O. B. Akan and I. F. Akyildiz. ARC : The analytical rate control scheme for real-time traffic in wireless networks. *IEEE/ACM Transactions on Networking*, 12(4):634–644, Aug. 2004.

[10] M. Allman, V. Paxson, and E. Blanton. TCP Congestion Control. RFC 5681, IETF, Sep. 2009.

[11] G. Americas. Mobile Broadband: EDGE, HSPA and LTE. `http://www.3gamericas.org/documents/2006_Rysavy_Data_Paper_FINAL_09.15.06.pdf`, 2006.

[12] Apple. iPhone. `http://www.apple.com/iphone/`, 2010.

[13] L. Badia, M. Zorzi, and A. Gazzini. A Model for Threshold Comparison Call Admission Control in Third Generation Cellular System. In *IEEE International Conference on Communications (ICC)*, Alaska, USA, May 11-15, 2003.

[14] S. Bangolae, A. Jayasumana, and V. Chandrasekar. TCP-Friendly Congestion Control Mechanism for a UDP-Based High-Speed Radar Application and Characterization of Fairness. In *International Conference on Communication Systems (ICCS)*, Singapore, Nov. 25-28, 2002.

[15] C. Barakat, E. Altman, and W. Dabbous. On TCP performance in a heterogeneous network: a survey. *IEEE Communications Magazine*, 38(1):40–46, Jan. 2000.

[16] A. Begen, T. Akgul, and M. Baugher. Watching Video over the Web: Part 1: Streaming Protocols. *IEEE Internet Computing*, 15(2):54–63, Mar. 2011.

[17] S. Belfiore, L. Crisà, M. Grangetto, E. Magli, and G. Olmo. Robust and Edge-Preserving Video Error Concealment by Coarse-to-Fine Block Replenishment. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Orlando, Florida, USA, May 13-17, 2002.

[18] R. A. Berry and E. M. Yeh. Cross-layer wireless resource allocation. *IEEE Signal Processing Magazine*, 21(5):59–68, Sep. 2004.

[19] S. Blake, D. Black, M. Carlson, E. Davies, Z. Wang, and W. Weiss. An Architecture for Differentiated Services. RFC 2475, IETF, Dec. 1998.

[20] J. A. Bocharov, Q. Burns, F. Folta, K. Hughes, A. Murching, L. Olson, P. Schnell, and J. Simmons. Portable encoding of audio-video objects: The Protected Interoperable File Format (PIFF). Technical report, Microsoft, Mar. 2010.

[21] R. Braden, D. Clark, and S. Shenker. Integrated Services in the Internet Architecture: an Overview. RFC 1633, IETF, Jun. 1994.

[22] R. Braden, L. Zhang, S. Berson, S. Herzog, and S. Jamin. Resource ReSerVation Protocol (RSVP). RFC 2205, IETF, Sep. 1997.

[23] T. Breddermann, H. Lüders, P. Vary, I. Aktas, and F. Schmidt. Iterative Source-Channel Decoding with Cross-Layer Support for Wireless VoIP. In *ITG Conference on Source and Channel Coding*, Siegen, Germany, Jan. 18-21, 2010.

[24] J. Brehmer and W. Utschick. A decomposition of the downlink utility maximization problem. In *Proceedings of the Conference on Information Sciences and Systems*, Baltimore, USA, Mar. 14-16, 2007. Johns Hopkins University.

[25] K. M. Bretthauer and B. Shetty. The nonlinear knapsack problem - algorithms and approaches. *European Journal of Operations Research*, 138(3):459–472, 2002.

[26] M. Budagavi, W. R. Heinzelman, J. Webb, and R. Talluri. Wireless MPEG-4 video communiction on DSP chips. *IEEE Signal Processing Magazine*, 17(1):36–53, Jan. 2000.

[27] R. Caceres and L. Iftode. Improving the Performance of Reliable Transport Protocols in Mobile Computing Environments. *IEEE Journal on Selected Areas in Communications*, 13(5):850–857, Jun. 1996.

[28] Z. Cao and E. W. Zegura. Utility max-min: An application-oriented bandwidth allocation scheme. In *IEEE Conference on Computer Communications (INFOCOM)*, New York, NY, USA, Mar. 21-25, 1999.

[29] G. Carneiro, J. Ruela, and M. Ricardo. Cross-Layer Design in 4G Wireless Terminals. *IEEE Wireless Communications*, 11(2):7–13, Apr. 2004.

[30] M. Chiang, S. H. Low, R. Calderbank, and J. C. Doyle. Layering as optimization decomposition: a mathematical theory on network architectures. *Proceedings of the IEEE*, 95(1):255–312, 2007.

[31] D. Chiu and R. Jain. Analysis of the increase/decrease algorithms for congestion avoidance in computer networks. *Journal of Computer Networks and ISDN Systems*, 17(1):1–14, Jun. 1989.

[32] L. U. Choi, W. Kellerer, and E. Steinbach. Cross-Layer Optimization for Wireless Multi-user Video Streaming. In *IEEE International Conference on Image Processing (ICIP)*, Singapore, Oct. 24-27, 2004.

[33] L. U. Choi, W. Kellerer, and E. Steinbach. On Cross-Layer Design for Streaming Video Delivery in Multi-User Wireless Environments. *EURASIP Journal on Wireless Communications and Networking, special issue on Radio Resource Management in 3G+ Systems*, 2006:1–10, Aug. 2006.

[34] A. P. Chou and M. van der Schaar. *Multimedia over IP and Wireless Networks.* Academic Press, 2007.

[35] T. K. Chua and D. C. Pheanis. Application-Level Adaptive Congestion Detection and Control for VoIP. In *International Conference on Networking and Services (ICNS)*, Athen, Greece, Jun. 19-25, 2007.

[36] Cisco. Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2009-2014. `http://www.cisco.com/en/US/solutions/collateral/ns341/ns525/ns537/ns705/ns827/white_paper_c11-520862.html`, 2010.

[37] S. Dixit, Y. Guo, and Z. Antoniou. Resource Management and Quality of Service in Third-Generation Wireless Networks. *IEEE Communications Magazine*, 39(2):125–133, Feb. 2001.

[38] H. Ekstrom, A. Furuskar, J. Karlsson, M. Meyer, S. Parkvall, J. Torsner, and M. Wahlqvist. Technical solutions for the 3G Long-Term Evolution. *IEEE Communications Magazine*, 44(3):38–45, Mar. 2006.

[39] Ericsson. Global mobile data traffic nearly triples in 1 year. `http://www.ericsson.com/news/1437680`, 2010.

[40] A. Eryilmaz and R. Srikant. Fair Resource Allocation in Wireless Networks Using Queue-length-based Scheduling and Congestion Control. In *IEEE Conference on Computer Communications (INFOCOM)*, Miami, Florida, USA, Mar. 13-17, 2005.

[41] M. Etoh and T. Yoshimura. Advances in Wireless Video Delivery. *Proceedings of the IEEE*, 93(1):111–122, Jan. 2005.

[42] A. Federgruen and P. Zipkin. Solution techniques for some allocation problems. *Mathematical Programming*, 25(1):13–24, Nov. 1983.

[43] R. Fielding, J. Gettys, J. Mogul, H. Frystyk, L. Masinter, P. Leach, and T. B. Lee. Hypertext Transfer Protocol – HTTP/1.1. RFC 2616, IETF, Jun. 1999.

[44] T. Fingscheidt, T. Hindelang, R. V. Cox, and N. Seshadri. Joint Source-Channel (De)Coding for Mobile Communications. *IEEE Transactions on Communications*, 50(2):200–212, Feb. 2002.

[45] S. Floyd and K. Fall. Promoting the use of end-to-end congestion control in the Internet. *IEEE/ACM Transaction Networking*, 7(4):458–472, Aug. 1999.

[46] S. Floyd, J. Padhye, and J. Widmer. TCP Friendly Rate Control (TFRC): Protocol Specification. RFC 3448, IETF, Jan. 2003.

[47] S. Floyd, M. H. J. Padhye, and J. Widmer. TCP Friendly Rate Control (TFRC): Protocol Specification. RFC 5348, IETF, Sep. 2008.

[48] I. O. for Standardization. Information technology – Coding of audio-visual objects – Part 2: Visual. Technical Report ISO/IEC 14496-2, ISO, 2004.

[49] T. Friedman, R. Caceres, and A. Clark. RTP Control Protocol Extended Reports (RTCP XR). RFC 3611, IETF, Nov. 2003.

[50] P. Froejdh, U. Horn, M. Kampmann, A. Nohlgren, and M. Westerlund. Adaptive Streaming within the 3GPP Packet-Switched Streaming Service. *IEEE Network*, 20(2):34–40, Mar. 2006.

[51] F. Fu and M. van der Schaar. Decomposition Principles and Online Learning in Cross-Layer Optimization for Delay-Sensitive Applications. *IEEE Transactions on Signal Processing*, 58(3):1401–1415, Mar. 2010.

[52] L. Georgiadis, M. J. Neely, and L. Tassiulas. Resource allocation and cross-layer control in wireless networks. *Foundations and Trends in Networking*, 1(1):1–146, 2006.

[53] S. Ghanbari, L. Cieplinski, and M. Bober. Recovery of Lost Motion Vectors for Error Concealment in Video Coding. In *Picture Coding Symposium*, Saint-Malo, France, Apr. 23-25, 2003.

[54] B. Girod. What's wrong with mean-squared error? *Digital Images and Human Vision*, pages 207–220, May 1993. the MIT press.

[55] B. Girod and N. Faerber. *Wireless Video in Compressed Video over Networks.* New York: Marcel Dekker, 2001.

[56] T. Goff, J. Moronski, D. S. Phatak, and V. Gupta. Freeze-TCP: A True End-to-End TCP Enhancement Mechanism for Mobile Environments. In *IEEE Conference on Computer Communications (INFOCOM)*, Tel Aviv, Israel, Mar. 26-30, 2000.

[57] O. Gross. A class of discrete type minimization problems. Technical Report RM-1644, RAND Corporation, 1956.

[58] Z. J. Haas. Design Methodologies for Adaptive and Multimedia Networks. *IEEE Communications Magazine*, 39(11):106–107, Nov. 2001.

[59] M. Handley, V. Jacobson, and C. Perkins. SDP: Session Description Protocol. RFC 4566, IETF, Jul. 2006.

[60] P. Haskell and D. Messerschmitt. Resynchronization of motion compensated video affected by ATM cell loss. San Francisco, California, USA, Mar. 23-26, 1992.

[61] D. S. Hochbaum. Lower and upper bounds for the allocation problem and other nonlinear optimization problems. *Mathematics of Operations Research*, 19(2):390–409, 1994.

[62] H. Holma and A. Toskala. *HSDPA/HSUPA for UMTS.* Wiley, Apr. 2006.

[63] D. Hong and S. S. Rappaport. Traffic Model and Performance Analysis for Cellular Mobile Radio Telephone Systems with Prioritized and Nonprioritized Handoff. *IEEE Transaction on Vehicular Technology*, 35(3):77–92, Aug. 1986.

[64] M. C. Hong, L. Kondi, H. Schwab, and A. K. Katsaggelos. Video Error Concealment Techniques. *Signal Processing: Image Communication*, 14(6-8):473–492, 1999.

[65] K. Hongseok, C. Chan-Byoung, G. de Veciana, and R. W. Heath. A cross-layer approach to energy efficiency for adaptive MIMO systems exploiting spare capacity. *IEEE Transactions on Wireless Communications*, 8(8):4264–4275, Aug. 2009.

[66] J. Hou, J. Yang, and S. Papavassilou. Integration of Pricing with Call Admission Control to Meet QoS Requirements in Cellular Networks. *IEEE Transaction on Parallel and Distributed Systems*, 19(9):898–910, Sep. 2002.

[67] T. Ibaraki and N. Katoh. *Resource Allocation Problems: Algorithmic Approaches.* The MIT Press, Boston, MA, 1988.

[68] International Telecommunication Union. Studies toward the unification of picture assessment methodology. ITU-R Report BT.1082-1, Jan. 1990.

[69] International Telecommunication Union. Information technology - Open System Interconnection - Basic reference model: The basic model. ITU-T Recommendation X.200, 1994.

[70] International Telecommunication Union. Method for subjective determination of transmission quality. ITU-T Recommendation P.800, 1996.

[71] International Telecommunication Union. Perceptual evaluation of speech quality (PESQ): an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs. ITU-T Recommendation P.862, 2001.

[72] International Telecommunication Union. Methodology for the Subjective Assessment of the Quality for Television Pictures. ITU-T Recommendation BT.500, 2002.

[73] International Telecommunication Union. The E-model, a computational model for use in transmission planning. ITU-T Recommendation G.107, 2005.

[74] International Telecommunication Union. Terms and definitions related to quality of service and network performance including dependability. ITU-T Recommendation E.800, Sep. 2008.

[75] International Telecommunication Union. Vocabulary for performance and quality of service. ITU-T Recommendation P.10/G.100 Amendment 2, Jul. 2008.

[76] International Telecommunication Union. Advanced video coding for generic audiovisual services. ITU-T Recommendation H.264, Mar. 2010.

[77] ISO. ISO/IEC 13818-1: Information technology: Generic coding of moving pictures and associated audio information: Systems. `http://www.iso.org/iso/catalogue_detail.htm?csnumber=51533`, 2007.

[78] ISO. ISO/IEC 14496: Information technology: Coding of audio-visual objects: Part 12: ISO base media file format. `http://www.iso.org/iso/catalogue_detail.htm?csnumber=51533`, 2008.

[79] M. T. Ivrlac and J. A. Nossek. Cross Layer Optimization-an Equivalence Class Approach. In *ITG Workshop Smart Antennas*, Munich, Germany, Mar. 18-19, 2004.

[80] H. Jiang, W. Zhuang, and X. Shen. Cross-layer design for resource allocation in 3G wireless networks and beyond. *IEEE Communications Magazine*, 43(12):120–126, Dec. 2005.

[81] D. Jurca and P. Frossard. Media-Specific Rate Allocation in Heterogeneous Wireless Networks. In *IEEE Packet Video Workshop*, Hangzhou, China, Apr. 20-21, 2006.

[82] H. Kaaranen, A. Ahitainen, L. Laitinen, S. Naghian, and V. Niemi. *UMTS Networks: Architecture, Mobility, and Services.* John Wiley and Sons, 2001.

[83] V. Kawadia and P. Kumar. A Cautionary Perspective on Cross Layer Design. *IEEE Wireless Communications*, 12(1):3–11, Feb. 2005.

[84] W. Kellerer, L. U. Choi, and E. Steinbach. Cross-layer adaptation for optimized B3G service provisioning. In *International Symposium on Wireless Personal Multimedia Communications (WPMC)*, Yokosuka, Japan, Oct. 21-22, 2003.

[85] W. Kellerer, S. Thakolsri, S. Khan, and E. Steinbach. Quality of Experience Driven Cross-Layer Optimization for the Future Mobile Internet. In *2nd GI/ITG KuVS Workshop on The Future Internet*, Karlsruhe, Germany, Nov. 11, 2008.

[86] F. P. Kelly. Charging and rate control for elastic traffic. *European Transactions of Telecommunication*, 8:33–37, Jan. 1997.

[87] S. Khan, S. Duhovnikov, E. Steinbach, and W. Kellerer. MOS-Based Multiuser Multiapplication Cross-Layer Optimization for Mobile Multimedia Communication. *Advances in Multimedia*, (doi:10.1155/2007/94918), 2007.

[88] S. Khan, M. Sgroi, Y. Peng, E. Steinbach, and W. Kellerer. Application-driven cross-layer optimization for video streaming over wireless networks. *IEEE Communications Magazine*, pages 122–130, Jan. 2006.

[89] S. Khan, S. Thakolsri, E. Steinbach, and W. Kellerer. QoE-based Cross-layer Optimization for Wireless Multiuser Systems. In *18th ITC Specialist Seminar on Quality of Experience*, Karlskrona, Sweden, May 29-30, 2008.

[90] B. J. Kim. A network service providing wireless channel information for adaptive mobile applications: part I: proposal. In *IEEE International Conference on Communications (ICC)*, Beijing, China, May 19-23, 2001.

[91] E. Kohler, M. Handley, and S. Floyd. Datagram Congestion Control Protocol (DCCP). RFC 4340, IETF, Mar. 2006.

[92] P. P. K. Lam and S. C. Liew. UDP-Liter: an improved UDP protocol for real-time multimedia applications over wireless links. In *International Symposium on Wireless Communication Systems*, Mauritius, Sep. 20-22, 2004.

[93] L. Larzon, M. Degermark, and S. Pink. The Lightweight User Datagram Protocol (UDP-Lite). RFC 3828, IETF, Jul. 2004.

[94] L. A. Larzon, U. Bodin, and O. Schelen. Hints and Notifications. In *IEEE Wireless Communications and Networking Conference (WCNC)*, Orlando, Florida, USA, Mar. 17-21, 2002.

[95] L. A. Larzon, M. Degermark, and S. Pink. Efficient use of wireless bandwidth for multimedia applications. In *IEEE International Workshop on Mobile Multimedia Communications (MoMuC)*, San Diego, CA, USA, Nov. 15-17, 1999.

[96] H. J. Lee, J. H. Jeon, and J. T. Lim. On Congestion Control for Streaming Real-time Applications over Wireless Networks with Bandwidth Variation. In *14th Asia-Pacific Conference on Communications (APCC)*, Tokyo, Japan, Oct. 14-16, 2008.

[97] W. C. P. Lee. *Mobile Communincations Design Fundamentals.* John Wiley and Sons, 1993.

[98] K. Leung and V. O. K. Li. Transmission control protocol (TCP) in wireless networks: issues, approaches, and challenges. *IEEE Communications Surveys and Tutorials*, 8(4):64–79, Oct. 2006.

[99] D. A. Levine, I. F. Akyildiz, and M. Naghshineh. A resource Estimation and Call Admission Algorithm for Wireless Multimedia Networks Using the Shadow Cluster Concept. *IEEE/ACM Transaction on Networking*, 5(1):1–12, Feb. 1997.

[100] X. Liu, E. K. P. Chong, and N. B. Schroff. Transmission scheduling for efficient wireless utilization. In *IEEE Conference on Computer Communications (INFOCOM)*, Anchorage, Alaska, USA, Apr. 22-26, 2001.

[101] H. Luo, S. Ci, D. Wu, J. Wu, and H. Tang. Quality-driven cross-layer optimized video delivery over LTE. *IEEE Communications Magazine*, 48(2):102–109, Feb. 2010.

[102] S. Z. M. Chiang and P. Hande. Distributed rate allocation for inelastic flows: Optimization frameworks, optimality conditions, and optimal algorithms. In *IEEE Conference on Computer Communications (INFOCOM)*, Miami, Florida, USA, Mar. 13-17, 2005.

[103] G. Manfredi, P. Annese, and U. Spagnolini. A channel aware scheduling algorithm for hsdpa system. In *16th IEEE Internation Symposium on Personal, Indoor and Mobile Radio Communications*, Berlin, Germany, Sep. 11-14, 2005.

[104] R. Marler and J. Arora. Survey of multi-objective optimization methods for engineering. *Structural and Multidisciplinary Optimization*, 26(6):369–395, Apr. 2004.

[105] Microsoft. Smooth Streaming. `http://www.iis.net/download/SmoothStreaming`, 2011.

[106] Y. Nakajima, H. Hori, and T. Kanoh. Rate conversion of mpeg coded video by re-quantization process. In *IEEE International Conference on Image Processing (ICIP)*, Washington, DC , USA, Oct. 23-26, 1995.

[107] N. Nasser, Al-Manthari, and H. Hassanein. A performance comparison of class-based scheduling algorithms in future umts access. In *IEEE International Performance, Computing and Communications Conference*, Phoenix, Arizona, USA, Apr. 7-9, 2005.

[108] M. C. Necker and A. Weber. Impact of Iub Flow Control on HSDPA System Performance. In *16th IEEE Internation Symposium on Personal, Indoor and Mobile Radio Communications*, Berlin, Germany, Sep. 11-14, 2005.

[109] M. T. Nietzel. *Introduction of clinical psychology.* Prentice Hall, 1991.

[110] D. Niyato and E. Hossain. Call Admission Control for QoS Provisioning in 4G Wireless Networks: Issues and Approaches. *IEEE Network*, 19(5):5–11, Sep. 2005.

[111] M. Opp. Outsourcing Warranty Repair Services. Phd thesis, University of North Carolina at Chapel Hill, 2003.

[112] A. Ortega and M. Khansari. Rate control for video coding over variable bit rate channels with applications to wireless transmission. In *IEEE International Conference on Image Processing (ICIP)*, Washington, DC , USA, Oct. 23-26, 1995.

[113] J. Ott, S. Wenger, N. Sato, C. Burmeister, and J. Rey. Extended RTP Profile for Real-time Transport Control Protocol (RTCP)-Based Feedback (RTP/AVPF). RFC 4585, IETF, Jul. 2006.

[114] R. Pantos and W. May. HTTP Live Streaming. Draft draft-pantos-http-live-streaming-06, IETF, Mar. 2011.

[115] A. D. Patel. The harvard-haskins database of regularly-timed speech. `http://vesicle.nsi.edu/users/patel/download.html`.

[116] M. Patriksson. A survey on the continuous nonlinear resource allocation problem. *European Journal of Operational Research*, 185(1):1–46, 2008.

[117] K. I. Pedersen, P. E. Mogensen, and T. E. Kolding. Overview of QoS options for HSDPA. *IEEE Communications Magazine*, 44(7):100–105, Jul. 2006.

[118] J. Postel. User Datagram Protocol. RFC 768, IETF, Aug. 1980.

[119] J. Postel. Internet Protocol. RFC 791, IETF, Sep. 1981.

[120] J. Postel. Transmission Control Protocol. RFC 793, IETF, Sep. 1981.

[121] L. Qingwen, Z. Shengli, and G. B. Giannakis. Queuing with adaptive modulation and coding over wireless links: cross-Layer analysis and design. *IEEE Transactions on Wireless Communications*, 4(3):1142–1153, May 2005.

[122] A. Racz, A. Temesvary, and N. Reider. Handover Performance in 3GPP Long Term Evoluation (LTE) Systems. In *16th IST Mobile and Wireless Communications Summit*, Budapest, Jul. 1-5, 2007.

[123] B. Radunovic and J.-Y. L. Boudec. A unified framework for max-min and min-max fairness with applications. *IEEE/ACM Trans. on Networking*, 15(5):1073–1083, Oct. 2007.

[124] D. Raggett, A. L. Hors, and I. Jacobs. HTML 4.0 Specification. Recommendation REC-html40-971218, W3C, Dec. 1997.

[125] K. Ramakrishnan, S. Floyd, and D. Black. The Addition of Explicit Congestion Notification (ECN) to IP. RFC 3168, IETF, Sep. 2001.

[126] K. Ratnam and I. Matta. WTCP: an efficient mechanism for improving TCP performance over wireless links. In *IEEE Symposium on Computers and Communications (ISCC)*, Athens, Greece, Jun. 30 - Jul. 2, 1998.

[127] J. Rey, D. Leon, A. Miyazaki, V. Varsa, and R. Hakenberg. RTP Retransmission Payload Format. RFC 4588, IETF, Jul. 2006.

[128] J. Rosenberg, H. Schulzrinne, G. Camarillo, A. Johnston, J. Peterson, R. Sparks, M. Handley, and E. Schooler. SIP: Session Initiation Protocol. RFC 3261, IETF, Jun. 2002.

[129] B. Sardar and D. Saha. A survey of TCP enhancements for last-hop wireless networks. *IEEE Communications Surveys and Tutorials*, 8(3):20–34, Jul. 2006.

[130] B. Sardar and D. Saha. Spatial and Temporal Error Concealment Techniques for Video Transmission Over Noisy Channels. *IEEE Transactions on Circuits and Systems for Video Technology*, 16(7):789 – 803, Jul. 2006.

[131] A. Saul. Wireless Resource Allocation With Perceived Quality Fairness. In *IEEE Annual Asilomar Conference on Signals, Systems, and Computers*, PACIFIC GROVE, CA, USA, Nov. 1-4, 2008.

[132] A. Saul, S. Khan, G. Auer, W. Kellerer, and E. Steinbach. Cross-Layer Optimization With Model-Based Parameter Exchange. In *International Conference on Communications (ICC)*, Glasgow, Scotland, Jun. 24-28, 2007.

[133] H. Schulzrinne, S. Casner, R. Frederick, and V. Jacobson. RTP: A transport protocol for real-time applications. RFC 3550, IETF, Jul. 2003.

[134] H. Schulzrinne, A. Rao, and R. Lanphier. Real Time Streaming Protocol (RTSP). RFC 2326, IETF, Apr. 1998.

[135] S. Shakkottai, T. S. Rappport, and P. C. Karlsson. Cross-Layer Design for Wireless Networks. *IEEE Communications Magazine*, 41(10):74–80, Oct. 2003.

[136] J. She, F. Hou, B. Shihada, and P. H. Ho. MAC-layer active dropping for real-time video streaming in 4G access networks. *IEEE Systems Journal*, 4(4):561–572, Dec. 2010.

[137] A. Singh, A. Konrad, and A. D. Joseph. Performance Evaluation of UDP Lite for Cellular Video. In *International Workshop on Network and Operating Systems Support for Digital Audio and Video (NOSSDAV)*, New York, USA, Jun. 25-26, 2001.

[138] J. Sjoberg, M. Westerlund, A. Lakeaniemi, and Q. Xie. Real-Time Transport Protocol (RTP) Payload Format and File Storage Format for the Adaptive Multi-Rate (AMR) and Adaptive Multi-Rate Wideband (AMR-WB) Audio Codecs. RFC 3267, IETF, Jun. 2002.

[139] C. Soon-Hyeok, D. Perry, and S. M. Nettles. A Software Architecture for Cross-Layer Wireless Network Adaptations. In *Seventh Working IEEE/IFIP Conference on Software Architecture (WISCA 2008)*, pages 281–284, Vancouver, Canada, Feb. 18-22, 2008.

[140] M. Stanley. The mobile internet, consumer usage and implications for media and marketing brands. `http://www.viralhousingfix.com/2010/02/12/the-mobile-internet-consumer-usage-and-implications-for-media/-and-marketing-brands/`, 2010.

[141] T. Stockhammer, M. M. Hannuksela, and T. Wiegang. H.264/AVC in Wireless Environments. *IEEE Transaction on Circuits and Systems for Video Technology*, 13(7), Jul. 2003.

[142] P. Sudame and B. R. Badrinath. On Providing Support for Protocol Adaptation in Mobile Wireless Networks. *Mobile Networks and Applications*, 6(1):43–55, Jan. 2001.

[143] G. J. Sullivan and T. Wiegand. Rate-distortion optimization for video compression. *IEEE Signal Processing Magazine*, 15(6):74–90, Nov. 1998.

[144] A. Takács, . Kovács, I. Gódor, F. Kalleitner, H. Brand, M. Ek, T. Stefansson, and F. Sjöberg. The Layer-Independent Descriptor Concept. *Journal of computers*, 1(2):23–32, May 2006.

[145] J. Tang, G. Morabito, F. Akyildiz, and M. Johnson. RCS: A rate control scheme for real-time traffic in networks with high bandwidth-delay products and high bit error rates. In *IEEE Conference on Computer Communications (INFOCOM)*, Anchorage, Alaska, USA, Apr. 22-26, 2001.

[146] S. Thakolsri, S. Cokbulan, D. Jurca, Z. Despotovic, and W. Kellerer. QoE-driven Cross-Layer Optimization in Wireless Networks Addressing System Efficiency and Utility fairness. In *2nd IEEE Workshop on Multimedia Communications and Services (MCS) in conjunction with IEEE International Conference on Global Communications (GLOBECOM 2011)*, Houston, Texas, USA, Dec. 5-9, 2011. Accepted for publication.

[147] S. Thakolsri, W. Kellerer, S. Khan, and E. Steinbach. Application-driven Cross-layer Optimization in Wireless Networks. In *2nd Seminar on Service Quality Evaluation in Wireless Networks supported by COST 290*, Stuttgart, Germany, Jun. 12, 2007.

[148] S. Thakolsri, W. Kellerer, S. Khan, and E. Steinbach. QoE-Driven Cross-Layer Optimization for High Speed Downlink Packet Access. *Journal of Communications, Special Issue on Multimedia Communications, Networking and Applications*, 4(9):669–680, Oct. 2009.

[149] S. Thakolsri, W. Kellerer, and E. Steinbach. Application-Driven Cross Layer Optimization for Wireless Networks using MOS-based Utility Functions. In *Fourth International Conference on Communications and Networking in China (ChinaCom*

09), Workshop on Application-driven Cross-layer Design for Multimedia Communications (ACDMC), Xi An, China, Aug. 26-28, 2009.

[150] S. Thakolsri, W. Kellerer, and E. Steinbach. QoE-based Rate Adaptation Scheme Selection for Resource-constrained Wireless Video Transmission. In *ACM Multimedia (short paper)*, Firenze, Italy, Oct. 25-29, 2010.

[151] S. Thakolsri, W. Kellerer, and E. Steinbach. QoE-based Cross-Layer Optimization of Wireless Video with Unperceivable Temporal video Quality Fluctuation. In *IEEE International Conference on Communications (ICC 2011)*, Kyoto, Japan, Jun. 5-9, 2011.

[152] V. Tsaoussidis and H. Badr. TCP-probing: towards an error control schema with energy and throughput performance gains. In *International Conference on Network Protocols (ICNP)*, Osaka, Japan, Nov. 14-17, 2000.

[153] V. Tsaoussidis and I. Matta. Open Issues on TCP for Mobile Computing. *Journal of Wireless Communications Mobile Computing*, 2(1):3–20, Feb. 2002.

[154] V. Tsibonis, L. Georgiadis, and L. Tassiulas. Exploiting wireless channel state information for throughput maximization. In *IEEE Conference on Computer Communications (INFOCOM)*, San Francisco, California, USA, Mar. 30 - Apr. 3, 2003.

[155] W. Tu, J. Chakareski, and E. Steinbach. Rate-Distortion Optimized Frame Dropping for Multiuser Streaming and Conversational Videos. *Advances in Multimedia*, (628970), 2008.

[156] P. N. Tudor and O. H. Werner. Real-Time Transcoding of MPEG-2 Video Bit Streams. In *Proceedings of International Broadcasting Convention*, Amsterdam, Netherlands, Sep. 12-16, 1997.

[157] M. van der Schaar and N. S. Shankar. Cross-layer wireless multimedia transmission: challenges, principles, and new paradigms. *IEEE Wireless Communications*, 12(4):50–58, Aug. 2005.

[158] Video Quality Expert Group (VQEG). Final report from the video quality experts group on the validation of objective models of video quality assessment. Technical report, ITU, Mar. 2000.

[159] Q. Wang and M. A. Abu-Rgheff. Cross Layer Signaling for Next-Generation Wireless Systems. In *IEEE Wireless Communications and Networking Conference (WCNC)*, New Orleans, Louisiana, USA, Mar. 16-20, 2003.

[160] Y. Wang, S. Wenger, J. Wen, and A. K. Katsaggelos. Error Resilient Video Coding Techniques. *IEEE Signal Processing Magazine*, 17(4):61–82, Jul. 2000.

[161] Z. Wang, A. C. Bovik, and L. Lu. Why is image quality assessment so difficult? In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Orlando, Florida, USA, May 13-17, 2002.

[162] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, Apr. 2004.

[163] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image Quality Assessment: From Error Visibility to Structural Similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, Apr. 2004.

[164] Z. Wang, L. Lu, and A. C. Bovik. Video Quality Assessment Based on Structural Distortion Measurement. *Signal Processing: Image Communication*, 19(1):121–132, Jan. 2004.

[165] I. Weissberger, I. Kostanic, and C. E. Otero. Background service QoS in a UMTS network. In *Proceeding of IEEE Southeastcon*, pages 230–233, Concord, NC, Mar. 18-21, 2010.

[166] M. Westerlund, I. Johansson, C. Perkins, P. O'Hanlon, and K. Carlberg. TCP Congestion Control. I-Draft draft-ietf-avtcore-ecn-for-rtp-02, IETF, May. 2011.

[167] D. Wetterroth. *OSI Reference Model for Telecommunication.* McGraw-Hill, 2001.

[168] T. Wiegang, G. J. Sullivan, G. Bjontegaard, and A. Luthra. Overview of the H.264/AVC Video Coding Standard. *IEEE Transaction on Circuits and Systems for Video Technology*, 13(7), Jul. 2003.

[169] C. W. Wong, C. Y. Tsui, R. S. Cheng, and K. B. Letaief. A real-time sub-carrier allocation scheme for multiple access downlink OFDM transmission. In *IEEE Vehicular Technology Conference*, Amsterdam, Netherland, Sep. 19-22, 1999.

[170] D. Wu, Y. T. Hou, and Y. Q. Zhang. Transporting real-time video over the Internet: Challenges and approaches. *Proceedings of the IEEE*, 88(12):1855–1877, Dec. 2000.

[171] F. Wu, G. Shen, K. Tan, K. Yang, and S. Li. Next generation mobile multimedia communications: Media codec and media transport perspectives. *China Communications Magazine*, pages 30–44, Oct. 2006.

[172] G. Wu, Y. Bai, J. Lai, and A. Ogielski. Interaction between TCP and RLP in wireless Internet. In *IEEE Global Communications Conference (GlobeCom)*, Rio de Janeiro, Brazil, Dec. 5-9, 1999.

[173] J. Xin, C. Lin, and M. Sun. Digital Video Transcoding. *Proceedings of the IEEE*, 93(1):84–97, Jan. 2005.

[174] J. Zhang, D. Wu, S. Ci, H. Wang, and A. K. Katsaggelos. Power-Aware Mobile Multimedia: a Survey. *Journal of Communications*, 4(9):600–613, Oct. 2009.

[175] Q. Zhang, W. Zhu, and Y. Q. Zhang. End-to-End QoS for Video Delivery Over Wireless Internet. *Proceedings of the IEEE*, 93(1):123–134, Jan. 2005.

[176] H. Zheng and J. Boyce. An Improved UDP Protocol for Video Transmission Over Internet-to-Wireless Networks. *IEEE Transactions on Multimedia*, 3(3):356–365, Sep. 2001.