Julius-Maximilians-Universität Würzburg
Institut für Informatik
Lehrstuhl für Kommunikationsnetze
Prof. Dr.-Ing. P. Tran-Gia

Performance Assessment of Resource Management Strategies for Cellular and Wireless Mesh Networks

Florian Wamser

Würzburger Beiträge zur
Leistungsbewertung Verteilter Systeme

Bericht 1/15

Würzburger Beiträge zur Leistungsbewertung Verteilter Systeme

Herausgeber

Prof. Dr.-Ing. P. Tran-Gia
Universität Würzburg
Institut für Informatik
Lehrstuhl für Kommunikationsnetze
Am Hubland
D-97074 Würzburg

Tel.: +49-931-31-86630
Fax.: +49-931-31-86632
email: trangia@informatik.uni-wuerzburg.de

Satz

Reproduktionsfähige Vorlage vom Autor.
Gesetzt in LaTeX Computer Modern 9pt.

ISSN 1432-8801

Performance Assessment of Resource Management Strategies for Cellular and Wireless Mesh Networks

Dissertation zur Erlangung des
naturwissenschaftlichen Doktorgrades
der Julius–Maximilians–Universität Würzburg

vorgelegt von

Florian Wamser

aus

Werbach

Würzburg 2015

Eingereicht am: 18.12.2014
bei der Fakultät für Mathematik und Informatik
1. Gutachter: Prof. Dr.-Ing. P. Tran-Gia
2. Gutachter: Prof. Dr.-Ing. W. Kellerer
3. Gutachter: Prof. Dr.-Ing. R. Steinmetz
Tag der mündlichen Prüfung: 17.02.2015

Danksagung

Nach Beendigung meiner Doktorarbeit möchte ich mich bei vielen Personen ganz herzlich bedanken. Dies betrifft vor allem meinen Doktorvater Prof. Phuoc Tran-Gia und das gesamte Umfeld des Lehrstuhls für Informatik III in Würzburg.

Bei meinem Doktorvater bedanke ich mich für die tolle Arbeitsatmosphäre. Er bietet in Würzburg das perfekte Umfeld mit einer Mischung aus Spaß unter Kollegen und interessanten internationalen Projekten. Er ermöglichte mir die Teilnahme an zahlreichen Konferenzen, Workshops und Projekttreffen und stand mir immer - fachlich als auch menschlich - mit Rat und Tat zur Seite.

Ebenfalls möchte ich mich bei meinem Zweitgutachter Prof. Wolfgang Kellerer bedanken. Er war offen für Fragen und wissenschaftliche Diskussionen im Vorfeld der Doktorarbeit. Mein weiterer Dank gilt Prof. Ralf Steinmetz, der als Drittgutachter meiner Doktorarbeit fungierte. Des Weiteren möchte ich mich bei den Mitgliedern der Prüfungskommission Prof. Reiner Kolla und Prof. Samuel Kounev ganz herzlich bedanken.

Mein weiterer Dank gilt Dirk Staehle, dem ehemaligen Leiter meiner Arbeitsgruppe am Lehrstuhl. Er hat mir alle Einzelheiten und wichtigen Hintergründe der mobilen Kommunikation erläutert. Er war mein erster fachlicher Ansprechpartner und hatte immer zahlreiche Ideen und weiterführende Ratschläge parat. Mit ihm zusammen habe ich zahlreiche Paper geschrieben und bei Projekten mitgearbeitet. Er ist der geistige Vater von YoMo, dem YouTube Monitor, der inzwischen in vielen unterschiedlichen Versionen für PCs, Heim-Router und Smartphones existiert. In diesem Zuge geht mein Dank auch an Barbara Staehle. Sie war das zweite Mitglied der Arbeitsgruppe und hat mit mir zusammen viele Male YoMo auf Konferenzen präsentiert. Mit ihr habe ich das *Aquarium* entworfen, das für uns ein Synonym für die Vielzahl an Tools und Programmen ist, die wir im Mesh-Testbed genutzt haben. Wir haben viele YouTube-Videos angesehen - unter anderem auch *Findet Nemo*: "Fischi? Fischi WACH AUF, du darfst jetzt nicht schlafen!!". Weiterhin möchte ich mich ganz herzlich bei Rastin Pries und Andreas Mäder bedanken. Sie waren das dritte und vierte Mitglied der Arbeitsgruppe *Wireless Networks* am Lehrstuhl. Bei Rastin habe ich meine Diplomarbeit geschrieben und meine ersten Studien und Paper verfasst.

Ganz besonderer Dank gilt Thomas Zinner, der nicht nur meine Doktorarbeit in großen Teilen korrigiert, sondern auch eng mit mir darüber hinaus zusammengearbeitet hat. Weiterer Dank geht an alle früheren und aktuellen Kollegen: Andreas Binzenhöfer, Kathrin Borchert, Valentin Burger, Lam Dinh-Xuan, Michael Duelli, David Hock, Steffen Gebert, Matthias Hartmann, Robert Henjes, Matthias Hirth, Michael Jarschel, Dominik Klein, Stanislav Lange, Frank Lehrieder, Rüdiger Martin, Christopher Metter, Jens Milbrandt, Anh Nguyen-Ngoc, Simon Oechsner, Daniel Schlosser, Christian Schwartz, Michael Seufert, Prof. Tobias Hossfeld, Prof. Kurt Tutschku, Prof. Michael Menth und Prof. Harald Wehnes. Ich möchte mich weiterhin auch bei allen Studenten und Mitautoren von gemeinsamen Papern bedanken. In diesem Zuge sind mindestens David Mittelstädt, Sebastian Deschner, Andreas Blenk, Bastian Blößl, David Stezenbach, Michael Denkler, Lukas Iffländer und Jing Zhu zu nennen. Weiterhin möchte ich Gisela Förster und Alison Wichmann für Ihre organisatorische Unterstützung bei der Verwaltungsarbeit danken.

Am Schluss geht mein besonderer Dank an meine Familie und an meiner Freundin. Ich bedanke mich bei meinen Eltern Birgit und Winfried und meinem Bruder Thomas für die immerwährende Unterstützung und bei Susanna für Ihre große Hilfe in allen Phasen der Arbeit.

Würzburg, im März 2015, Florian Wamser

Contents

1 Introduction

This monograph deals with the analysis and performance evaluation of resource management strategies in cellular and wireless mesh networks. In the following, a motivation is given on this topic. Thereafter, the scientific contribution of this work is described. At the end of this introductory chapter, the outline of the thesis is provided.

1.1 Motivation

The rapid growth in the field of communication networks has been truly amazing in the last decades. We are currently experiencing a continuation thereof with an increase in traffic and the emergence of new fields of application. In particular, the latter is interesting since due to advances in the networks and new devices, such as smartphones, tablet PCs, and all kinds of Internet-connected devices, new additional applications arise from different areas. These include personal cloud services, multiplayer game streaming, browser-based multimedia applications, commercial Internet services for transportation and manufacturing, ultra high definition multimedia applications, assistance systems for the elderly, machine-to-machine communications with electronic meters and mobile sensors for everyday use. What applies for all these services is that they come from very different directions and belong to different user groups. This results in a very heterogeneous application mix with different requirements and needs on the access networks.

The applications within these networks typically use the network technology as a matter of course, and expect that it works in all situations and for all sorts of purposes without any further intervention. Mobile TV, for example, assumes that the cellular networks support the streaming of video data. Likewise, mobile-connected electricity meters rely on the timely transmission of accounting data for electricity billing. From the perspective of the communication networks, this requires not only the technical realization for the individual case, but a broad consideration of all circumstances and all requirements of special devices and applications of the users. All this applies in addition to the usual challenges, such as energy savings in the network, costs, and electromagnetic interference as it occurs in wireless networks.

Such a comprehensive consideration of all eventualities can only be achieved by a dynamic, customized, and intelligent management of the transmission resources. This management requires to exploit the theoretical capacity as much as possible while also taking system and network architecture as well as user and application demands into account. Hence, for a high level of customer satisfaction, all requirements of the customers and the applications need to be considered, which requires a multi-faceted resource management, especially with respect to the multitude of different applications and Internet-enabled devices.

The prerequisite for supporting all devices and applications is consequently a holistic resource management at different levels. At the physical level, the technical possibilities provided by different access technologies, e.g., more transmission antennas, modulation and coding of data, possible cooperation between network elements, etc., need to be exploited on the one hand. On the other hand, interference and changing network conditions have to be counteracted at physical level. On the application and user level, the focus should be on the customer demands due to the currently increasing amount of different devices and diverse applications (medical, hobby, entertainment, business, civil protection, etc.). Overall, the goal is always a management that maximizes the total performance of the communication network defined by different operator-specific objectives. Critical aspects here are the consideration of all demands and the interaction between the different methods at different levels, always with respect to the balance between costs and benefits of each method in terms of the overall performance.

Accordingly, the development and assessment of methods for resource management in communication networks is an important and far-reaching activity that guarantees the success of the technology and defines its acceptance.

1.2 Scientific Contribution

The intention of this thesis is the development, investigation, and evaluation of a holistic resource management with respect to new application use cases and requirements for the networks. Therefore, different communication layers are investigated and corresponding approaches are developed using simulative methods as well as practical emulation in testbeds. The new approaches are designed with respect to different complexity and implementation levels in order to cover the design space of resource management in a systematic way. Since the approaches cannot be evaluated generally for all types of access networks, network-specific use cases and evaluations are finally carried out in addition to the conceptual design and the modeling of the scenario. We consider cellular networks and wireless mesh access networks in this thesis.

Figure 1.1 gives an overview of the different topics. The topics are sorted according to the main methodology covering simulation, modeling, implementation, and practical measurements. Additionally, research studies of related areas that are also conducted by the author, are included in the figure. In total, this monograph covers four topics that will be briefly summarized in the following.

The first part is concerned with management of resources at physical layer. We study distributed resource allocation approaches under different settings. Due to the ambiguous performance objectives, a high spectrum reuse is conducted in current cellular networks. This results in possible interference between cells that transmit on the same frequencies. The focus is on the identification of approaches that are able to mitigate such interference. We propose different novel coordination mechanisms which lead to a more efficient resource usage. Further on, new opportunities to partition or restrict the different transmission resources are

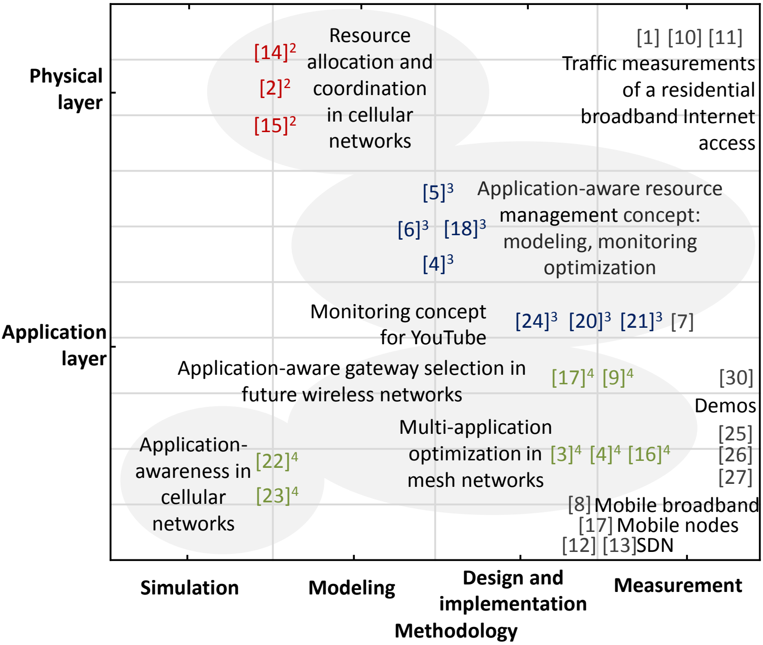

Physical layer
Application layer
[14]2 [2]2 [15]2 Resource allocation and coordination in cellular networks
[1] [10] [11] Traffic measurements of a residential broadband Internet access
[5]3 [6]3 [18]3 [4]3 Application-aware resource management concept: modeling, monitoring optimization
Monitoring concept for YouTube
[24]3 [20]3 [21]3 [7]
Application-aware gateway selection in future wireless networks
[17]4 [9]4
[30] Demos [25] [26] [27]
Application-awareness in cellular networks [22]4 [23]4
Multi-application optimization in mesh networks [3]4 [4]4 [16]4
[8] Mobile broadband
[17] Mobile nodes
[12] [13]SDN
Simulation
Modeling
Design and implementation
Measurement
Methodology

Figure 1.1: *Contribution of this work illustrated by a cartography of the research studies carried out. The notion* $[x]^y$ *indicates that the scientific publication* $[x]$ *is discussed in Chapter* y *of this monograph.*

evaluated in terms of the number of supported users. These evaluations are conducted for constant traffic demand. Another additional objective is furthermore to evaluate the different approaches with respect to current traffic patterns such as non-saturated traffic in the uplink. Such approaches are modeled and evaluated in detail with extensive simulations.

Due to the heterogeneity of the applications in the networks, increasingly different application-specific requirements are experienced by the networks. Consequently, the focus is shifted in the second part from optimization of network parameters to consideration and integration of the application and user needs by adjusting network parameters. Therefore, application-aware resource management is introduced to enable efficient and customized access networks. It uses information about the status of an application (e.g. active/idle, displaying a video, application buffer level, chatting, displaying information, calculating, processing information, interaction with user), and integrates this information in the network resource management. Based on appropriate literature, application information reflects the perceived quality of the user to a high degree if properly selected. An integral part of application-aware resource management is the monitoring. The performance of the resource management is highly dependent on the utilized information. To monitor the appropriate information, a dynamic approach is required which is designed and evaluated in the third chapter on the example of YouTube video streaming.

As indicated before, approaches cannot be evaluated generally for all types of access networks. Consequently, the third contribution is the definition and realization of the application-aware paradigm in different access networks. First, we address multi-hop wireless mesh networks. Wireless mesh networks maintain many different options and management possibilities to transmit traffic to the Internet, thus providing a flexible network structure that gives an excellent opportunity to quantify the benefits of application-aware resource management. The evaluation is done in a mesh testbed via empirical measurements. Such practical implementations illustrate the benefits of the concept and demonstrate the feasibility.

Finally, we focus with the fourth contribution on cellular networks again. Application-aware resource management is applied here to the air interface between user device and the base station. Especially in cellular networks, the intensive cost-driven competition among the different operators facilitates the usage of such a resource management to provide cost-efficient and customized networks with respect to the running applications. The numerical quantification of these benefits is subject to a simulative evaluation for YouTube, file downloads, and web browsing.

Overall, the results of this work fit into the broad field of resource management for access networks. Without such a resource management efficient, productive, and cost-effective networks would not be possible. The literature currently focuses heavily on network-related optimizations, such as those discussed in Chapter 2. Much more important in terms of economic aspects however is the optimization for specific applications and user requirements. To that effect, the work provides significant new ideas and concepts that are supported by implementations and simulations for certain types of networks.

The overall objective of this work is the evaluation of resource management approaches on different layers. The entire work can be seen as a contribution to the resource management research of mobile communication networks and wireless mesh networks, which gives clear conclusions supported by simulation results and real implementations of resource management strategies of different complexity in order to assess and justify the use of such strategies.

1.3 Outline of This Thesis

The organization of this monograph is shown in Figure 1.2. Each chapter contains a section that shows the background and related work of the covered topics. Additionally, we summarize lessons learned at the end of each chapter. The three columns cover from left to right (1) the problems and challenges, (2) the proposed solution with algorithms or mechanisms to cope with the problems, and (3) the results which present the impact of the applied mechanisms on the network performance. The arrows between the sections show their relation, background, and findings that are used in later sections. The section numbers of the building blocks are given in parentheses.

The remainder of this thesis is organized as follows. Chapter 2 contains all approaches that operate close to the physical layer. These include the approaches that coordinate resource allocation according to interference levels, uplink traffic, and frequency reuse strategy. Chapter 3 focuses on application-aware resource management. Here, the concept is introduced and discussed that allows for the integration of application layer information within the resource management.

Based on this definition, use cases are given in Chapter 4. We first investigate in Chapter 4 application-aware algorithms within wireless mesh networks. In the second part, we focus on cellular networks. An application-aware packet scheduling is defined that allocates resources for transmission according to application layer information. Finally, this monograph is concluded in Chapter 5 by a summary of the presented results and achievements.

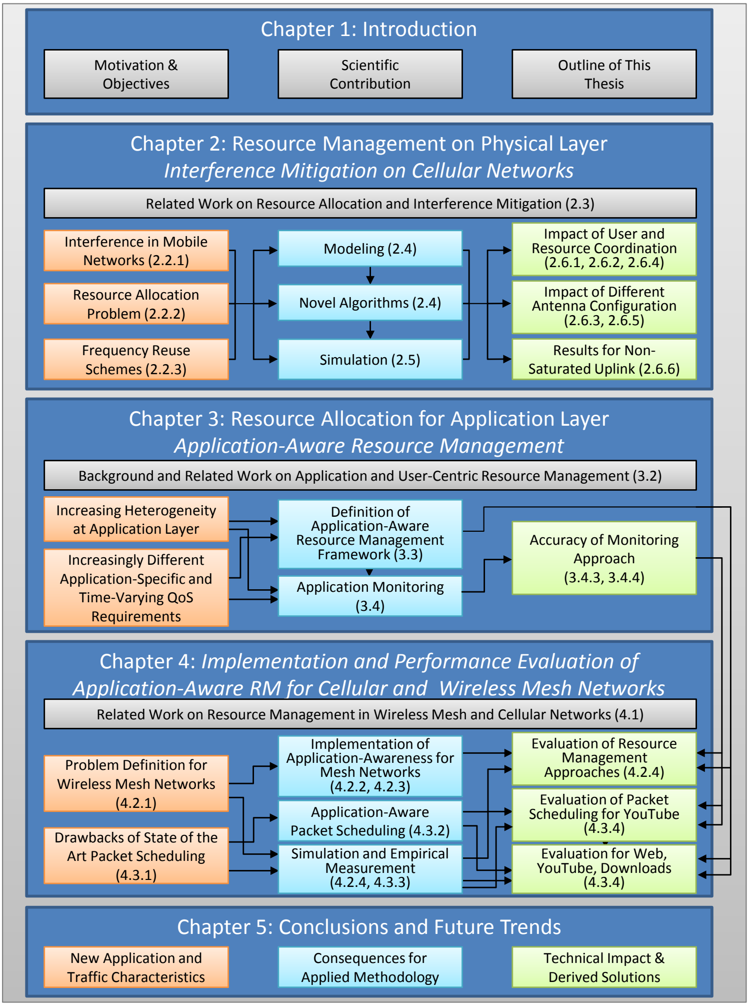

Chapter 1: Introduction
Motivation & Objectives
Scientific Contribution
Outline of This Thesis
Chapter 2: Resource Management on Physical Layer
Interference Mitigation on Cellular Networks
Related Work on Resource Allocation and Interference Mitigation (2.3)
Interference in Mobile Networks (2.2.1)
Resource Allocation Problem (2.2.2)
Frequency Reuse Schemes (2.2.3)
Modeling (2.4)
Novel Algorithms (2.4)
Simulation (2.5)
Impact of User and Resource Coordination (2.6.1, 2.6.2, 2.6.4)
Impact of Different Antenna Configuration (2.6.3, 2.6.5)
Results for Non-Saturated Uplink (2.6.6)
Chapter 3: Resource Allocation for Application Layer
Application-Aware Resource Management
Background and Related Work on Application and User-Centric Resource Management (3.2)
Increasing Heterogeneity at Application Layer
Increasingly Different Application-Specific and Time-Varying QoS Requirements
Definition of Application-Aware Resource Management Framework (3.3)
Application Monitoring (3.4)
Accuracy of Monitoring Approach (3.4.3, 3.4.4)
Chapter 4: Implementation and Performance Evaluation of Application-Aware RM for Cellular and Wireless Mesh Networks
Related Work on Resource Management in Wireless Mesh and Cellular Networks (4.1)
Problem Definition for Wireless Mesh Networks (4.2.1)
Drawbacks of State of the Art Packet Scheduling (4.3.1)
Implementation of Application-Awareness for Mesh Networks (4.2.2, 4.2.3)
Application-Aware Packet Scheduling (4.3.2)
Simulation and Empirical Measurement (4.2.4, 4.3.3)
Evaluation of Resource Management Approaches (4.2.4)
Evaluation of Packet Scheduling for YouTube (4.3.4)
Evaluation for Web, YouTube, Downloads (4.3.4)
Chapter 5: Conclusions and Future Trends
New Application and Traffic Characteristics
Consequences for Applied Methodology
Technical Impact & Derived Solutions

Figure 1.2: *Organization and contribution of this monograph according to problems and challenges, algorithms and methodologies, and impact of the applied mechanisms on the investigated systems.*

2 Resource Management on Physical Layer: Interference Mitigation in Cellular Networks

The first chapter is concerned with resource management on the physical layer. At this layer, the technical possibilities provided by different access technologies need to be exploited on the one hand. These include the modulation and coding of data, the adjustment of the transmission power, the choice of frequency or transmission wavelength, the choice of transmitter or transmit antenna, and possible cooperation between network elements. On the other hand, interference and changing network conditions have to be counteracted at physical layer.

We focus in this chapter on mobile communication networks and the topics *inter-cell interference mitigation* and *resource management* with frequency reuse schemes. We propose different coordination mechanisms which lead to a more efficient resource usage.

2.1 Motivation and Objectives

One of the most crucial tasks in mobile communications is the mitigation of *inter-cell interference* at the physical layer. From the beginning of mobile communications, there have been efforts to reduce such to an acceptable level. The reason for this is based on the fact that the transmission resources in a mobile communication network are limited. For the transmission usually only a restricted number of frequencies are available which are exclusively registered and purchased for dedicated use by the network operator. This consequently results in the need to reuse these resources to a high degree in order to achieve a high data transmission rate and quality in the network. However, if the frequency spectrum is reused aggressively, current systems encounter interference between cells, which necessitates the use of resource management approaches for *inter-cell interference mitigation*.

From a technical point of view, Orthogonal Frequency Division Multiple Access (OFDMA) is the current technology of choice for mobile networks. Both Worldwide Interoperability for Microwave Access (WiMAX) and 3GPP Long Term Evolution (LTE) use OFDMA as multi-user access scheme. The underlying principle is that OFDMA regulates the access of individual users within a cell. Simultaneous transmissions in the same frequency band of adjacent cells, however, result in interference that degrades the transmission signal.

A number of techniques have been proposed to counteract inter-cell interference [31–34]. They differ in technical complexity and according to the benefits for the network. First, there is *interference averaging*. The task of inter-cell interference averaging is to distribute the generated interference evenly across all users. *Frequency hopping*, as it is used in 2G Global System for Mobile Communications (GSM) systems, is classified as inter-cell interference averaging. Within the WiMAX IEEE 802.16e standard, similar techniques are proposed, e.g. *random sub-carrier permutations*. A second technique to mitigate interference is the use of smart antennas. A smart antenna is an arrangement of antenna elements. It can be used to achieve *beamforming*, which controls the directivity of the antenna, to narrow the caused interference to a specific geographical area. *Inter-cell interference avoidance* furthermore is another class of inter-cell interference mitigation techniques. Part of *interference avoidance* is the problem of optimal resource allocation. Additionally, a further part of *interference avoidance* is *resource partitioning and management*. It addresses the partitioning of resources into groups serving different purposes and the management of the resources, e.g. by applying transmit power limitations to the groups.

Frequency reuse schemes are part of the *resource partitioning and management*. A frequency reuse defines how the frequency resources are shared among neighboring cells. A network-wide universal reuse as it is used in Universal Mobile Telecommunications System (UMTS) is also known under the term *Frequency Reuse 1*. A *Frequency Reuse 1* results in a high amount of available resources, but at the same time causes a high level of interference resulting from the concurrent transmissions in neighboring cells of these resources. *Frequency Reuse 3* is similar to the approach of frequency planning in GSM. Here, the frequency resources are divided into three orthogonal sets. Each cell is assigned to one of the sets in a fashion that minimizes the interference. This results in the amount of available resources to be divided by three, whereas the interference is at a much lower level than with *Frequency Reuse 1*. *Fractional Frequency Reuse* (FFR) is an extension of both reuse concepts. The cell area is divided into cell edge with *Frequency Reuse 3* to achieve a high channel quality, while *Frequency Reuse 1* is used in the cell center to maximize the resource utilization.

Fractional frequency reuse is a popular topic in modern wireless communications and is considered in many works. Most of these works focus on the downlink transmission. The interference in the uplink and the downlink, however, have differing characteristics. A base station antenna uses a fixed transmit power for the downlink transmission that is shared among the users. In the uplink, however, each user has a specific transmit power available. This results in the available transmit power scaling with the number of users. The capacity of FFR in the uplink therefore needs to be evaluated with suitable measures.

Another challenge are the changing traffic characteristics through new applications and different devices. Many previous works consider the downlink direction with saturated users [35–38]. In the downlink from the Internet to the user, most of the data is transmitted. In contrast, we concentrate in this chapter on the uplink transmission. Moreover, we especially focus on the uplink with non-saturated users. Such type of users seems to be the more realistic choice for the uplink since nowadays the typical uplink traffic consists of constantly-recurring, short TCP acknowledgements, HTTP requests, voice and video traffic, and control traffic, due to the large amount of downlink traffic. TCP connections with large data volume on the uplink are expected to occur only sporadically, e.g. due to the transmission of user-generated content.

In this chapter, FFR approaches for the resource allocation in the uplink are investigated. The aim is to consider the uplink in a non-saturated case. First, we provide a model for the scenario and carry out a performance evaluation with respect to current traffic characteristics. Furthermore, the fact is taken into account that due to the dense network structure and the resulting reuse of frequencies, interference is generated. The second objective is therefore to avoid such interference by an appropriate resource allocation. In the end, this work should provide recommendations for network providers how a decentralized resource allocation needs to look like and what points should be considered to i) take the new traffic characteristics into account and ii) mitigate interference between cells, since these two points are seen as some of the main performance degradation issues for current networks.

The content of this chapter is mainly taken from [2, 14, 15]. Its remainder is structured as follows. We start with a section on the basics related to the subject of this chapter. Herein, an introduction to interference in mobile networks is first given in Section 2.2.1 and different types of mitigation are discussed. Afterwards, technical details on the physical resource allocation in mobile networks are briefly described in Section 2.2.2. Especially, the flexibility of the OFDMA multi-user access scheme is outlined. In the end, an introduction to frequency reuse schemes for resource partitioning is given in Section 2.2.3. Based on that, related work on the most important classical resource allocation constraints and objectives is summarized and discussed in Section 2.3. The intention is to provide an overview of the most relevant resource management objectives on the physical layer in this work. Thereafter, related work especially about interference mitigation is enumerated to provide an overview of existing interference mitigation approaches in Section 2.3.2. Prior to the investigation of interference mitigation approaches, the simulation model is defined in Section 2.4 and the simulation is introduced in Section 2.5. The evaluation is then finally laid out in six subsections in Section 2.6. Lastly, results and findings are summarized, and basing on such, recommendations are made on resource allocation of today's mobile networks.

2.2 Background

This section provides background information on resource management and interference mitigation. First, inter-cell interference is explained and different types of mitigation are discussed. Afterwards, we briefly introduce the resource allocation process in OFDMA mobile networks. Thereafter, frequency reuse schemes are introduced since they influence the resource allocation process to mitigate inter-cell interference.

2.2.1 Interference in Mobile Networks

In mobile communications, the resources for the data transmission are scarce and limited. Therefore, the frequencies at which the data is transmitted are often re-used in multiple cells. This principle is essential in order to meet the traffic demands in cellular networks.

Through the reuse of frequencies within multiple cells, however, interference between cells may be induced. This interference significantly affects the transmission capacity in the individual cells and is commonly addressed by appropriate resource management. If the interference significantly affects system performance, the system is referred to as *interference-limited*.

Figure 2.1 illustrates two mobile users that are using the same frequency band and interfere with each other. As both their signals are sent to the base station antennas, they also reach other mobile users. This causes interference for both users in the different cells. The intensity of the interference in the uplink is dependent on i) the location of the users in the cell and ii) the current transmit power of the devices.

In order to cope with interference, either a different frequency band has to be assigned to conflicting users or the transmission power must be reduced, such that the other party is not disturbed. Traditionally, in GSM mobile communication system, inter-cell interference is avoided by the assignment of disjoint frequency bands. This assignment is however static and is done in the planning phase of

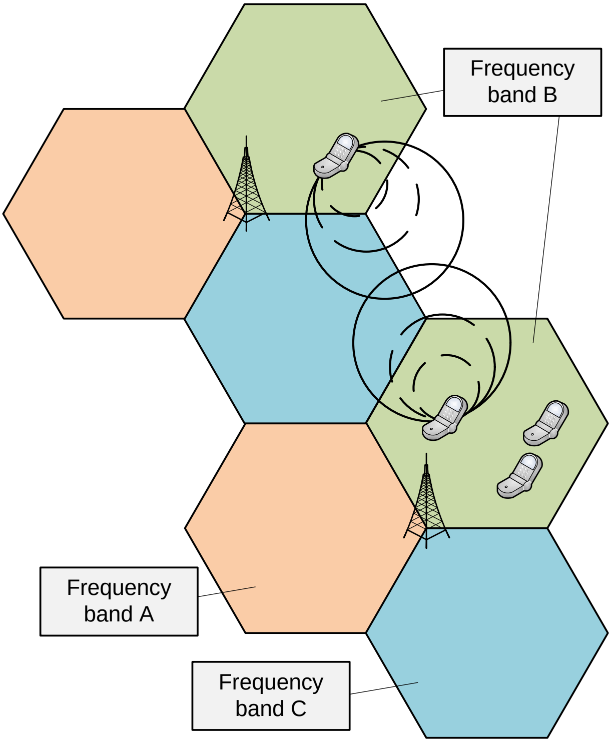


Figure 2.1: *Illustration for the occurrence of uplink inter-cell interference: mobiles in different cells, but on the same frequency band.*

the network, which considerably complicates subsequent extensions or modifications. All participants in 3G WCDMA systems such as UMTS in contrast use the same frequency band. Inter- and intra-cell interference is reduced by orthogonal and pseudo-orthogonal coding. An explicit resource management to mitigate interference is not necessary. However, this also results in the fact that part of the potential system capacity that is caused by the high variability of the interference cannot be exploited. In the 3.5G UMTS enhanced uplink, inter-cell interference is locally mitigated by threshold-based signaling to intercept random load peaks, and better utilize the system capacity. Finally, in 4G OFDMA systems inter-cell interference may occur due to the orthogonal resource allocation strategy.[1] The flexibility of the OFDMA access scheme allows different methods for explicit reduction or mitigation of interference:

- Static or semi-static resource allocation methods assign the available subcarriers to distinct cells or disjoint areas inside cells. Such approaches are called *Soft Frequency Reuse* (SFR) or *Fractional Frequency Reuse* (FFR). In the semi-static case, the allocation may be adjusted according to the load situation in the cell. In Section 2.2.3, an overview on frequency reuse strategies including FFR and SFR is given.

- Randomized methods try to distribute the inter-cell interference as evenly as possible across the frequency spectrum in order to avoid a reduction in signal quality of individual subcarriers. Such approaches, like frequency hopping, belong to the class of interference averaging. They are generally not suitable for frequency selective scheduling. In Section 2.3.2, related work on interference averaging is given.

- Dynamic allocation methods assign the resources in such a way that interference is mitigated and an objective function is optimized at the same time.

- Beamforming approaches avoid the interference by directing the transmission in certain directions. For this purpose, the transmitter uses an array of at least two antennas and directs the transfer to a geographically defined area, so that the unwanted interference to other cells is omitted. This approach requires multiple antenna technology and detailed information about users, see also related work at Section 2.3.1.

[1] A detailed description of OFDMA-based resource allocation follows in the next subsection.

The approaches can further be distinguished whether the interference management is centrally controlled or distributed to cells. A central approach with a global knowledge of the system state may achieve theoretically optimal allocations. However, more realistic are distributed or hybrid methods that reduce the signaling load.

The challenge is to combine the interference mitigation objectives with the objectives of other constraints such as frequency selective scheduling. An optimal resource allocation can be achieved theoretically, if the allocation is done dynamically throughout the whole system at the beginning of each transmission frame. However, this requires a central unit with global knowledge about the system state and synchronized base stations. The disadvantages and difficulties of such a centralized approach are firstly an increased signaling traffic load at the relatively expensive fixed connections between the base stations. Secondly, the signaling with a central controller leads to a certain delay, which in turn reduces the gain of a high-speed frequency selective scheduling. In addition, the computational complexity increases for larger networks and therefore good solutions for the decoupling of a network and the multiple control entities are required in this case.

In this work, we investigate static and semi-static FFR schemes and provide a performance evaluation of such systems. We stick to decentralized approaches that do not require signaling and compare them with conventional schemes.

2.2.2 Resource Allocation in Mobile Networks

The resource allocation in mobile networks is the process for the coordinated allocation of transmission resources, i.e. frequency, time and transmit power, to the users. The challenge is to allocate resources based on different network specifications and different user characteristics, and meet various and sometimes conflicting performance objectives of the overall network.

In 3GPP LTE mobile networks, the minimum transmission unit is a *Resource Element* (RE). The number of REs associated with the mobile user, is directly proportional to the data rate experienced by the user. Orthogonal Frequency-Division Multiplexing (OFDM) is the digital modulation scheme that is used to send the data on multiple carrier frequencies. An RE is defined in time as one OFDM symbol and in frequency domain as one subcarrier with either 15 kHz or 24 kHz. Depending on the length of the cyclic prefix, 6 or 7 OFDM symbols with 12 subcarriers in a row form a *Resource Block* (RB). The RB is the smallest resource allocation unit. Adaptive Modulation and Coding (AMC) is done by RB and selects the right modulation with respect to the current wireless channel between the base station antenna and the end user.

The assignment of user data on the transmission resources is often made on the basis of subframes since a too fine-grained allocation per RB is more complex and requires additional knowledge. A subframe comprises two RBs resulting in a general scheduling interval of 1 ms.

In order to support multiple users on multiple frequencies, the OFDMA scheme is used. It is based on OFDM. It divides the transmission resources in time and allocates RBs or subframes to each user for transmission. OFDMA is the technology of choice for multi-user access schemes for today's mobile communication networks. The IEEE 802.16 Worldwide Interoperability for Microwave Access (WiMAX) [39] standard uses an OFDMA-based physical layer specified in the IEEE 802.16 standard [40]. 3GPP LTE [41] uses OFDMA on the downlink.

One advantage of OFDMA is the high flexibility which results from the two-dimensional structure (time and frequency) of the physical resources. The users can be dynamically assigned to resources in such a way that high data rates are possible. To achieve high data rates, wide frequency bands are needed (in LTE up to 20 MHz and up to 40 MHz in IEEE 802.16m are specified). Wide frequency bands, however, are not alone sufficient for high data rates. The corresponding transmission channel undergoes frequency selective fading and receives interference from other cells and users, i.e., the instantaneous channel amplitudes of the individual narrowband subcarriers depend on the time of observation and their frequency. Assuming that the majority of users experience fading and interference statistically independent from each other, this can be exploited in order to increase system capacity with scheduling and resource allocation at appropriate times and frequencies with respect to interference by the resource management. Technically, this multi-user diversity (MUD) can be realized through channel-dependent allocation of transmission resources, interference-reducing allocation of transmission power, and appropriate modulation and coding. The prerequisite therefore is a good estimate of the transmission channel on the one hand and a signaling of the channel conditions to the base station and back on the other hand.

2.2.3 Frequency Reuse Schemes

Since resource allocation with respect to interference mitigation is evaluated in this chapter, we now provide a technical description of frequency reuse techniques for interference mitigation. They belong to the class of decentralized static or semi-static interference mitigation methods as described in Section 2.2.1.

Frequency reuse strategies address the trade-off between resource utilization and resulting interference. The principle is to plan and restrict the use of resources by an appropriate frequency planning strategy throughout the entire network, so that an overlap in the use of resources, and thus interference, does not occur.

Frequency Reuse 1 and Frequency Reuse 3

If no frequency planning is done at all in the network, the reuse strategy is called *Frequency Reuse 1*. In every cell and sector the entire frequency spectrum is used at any time. In Figure 2.2, both a transmission frame as well as three neighboring cells of a corresponding *Frequency Reuse 1* network are presented. The advantage of this approach is that all resources are always available for all cells. The disadvantage is that inter-cell interference occurs which may reduce the overall capacity of the network.

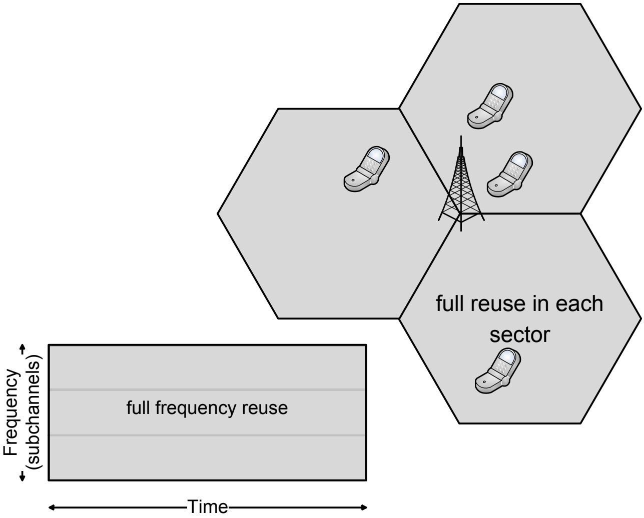


Figure 2.2: *Frequency reuse scheme: Frequency Reuse 1.*

In contrast, with *Frequency Reuse 3*, adjacent cells are assigned disjoint frequency bands, thus preventing interference. The problem here is that only a part of the resources becomes available for each cell. For *Frequency Reuse 3*, the spectrum is divided into three parts. Consequently, only a third of the resources are available. This case is illustrated in Figure 2.3. *Frequency Reuse 3* mitigates inter-cell interference quite effectively due to the large distance between sectors using the same frequency band. However, the resulting higher signal-to-interference plus noise values are achieved on behalf of a significant loss in resources.

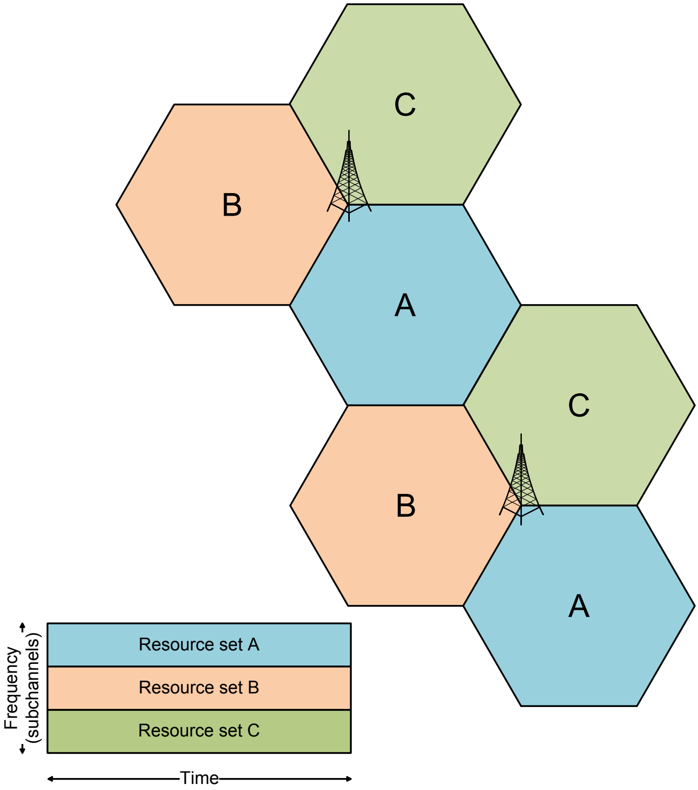


Figure 2.3: *Frequency reuse scheme: Frequency Reuse 3.*

Fractional Frequency Reuse

Fractional Frequency Reuse (FFR) schemes represent a combination of the two approaches mentioned above and make a partial allocation of transmission resources. It is assumed that users within the cell center transmit with less transmission power to the antenna and vice versa since they are closer to the antenna. Due to the low transmit power, consequently, less interference occurs for other cells. Thus, a full frequency reuse is allowed for the users within the cell center. In contrast, users at the cell edge send with high transmit power, since they are far away from the antenna. Moreover, they are also close to the neighboring cell,

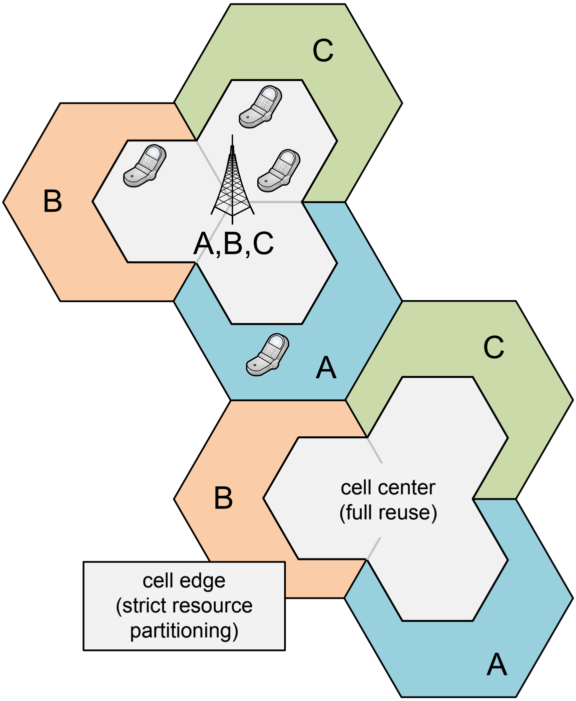
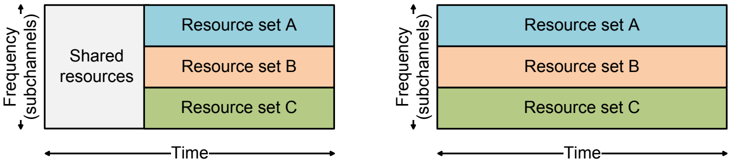


Figure 2.4: *Fractional Frequency Reuse with full reuse in the cell center and Frequency Reuse 3 at the cell border.*

thus causing even more interference to others. For users at the cell edge, a strict *Frequency Reuse 3* is set, so that they do not cause harmful interference despite the high transmission power and proximity to the neighboring cell. The resulting spatial allocation is illustrated in Figure 2.4.

The challenge is to define the areas for the *Frequency Reuse 3* users at the cell edge and the *Frequency Reuse 1* users inside the cell center in the transmission frame. Commonly, this is achieved by restricting the allowed power for transmission of particular frequency resources. Further improvement can be achieved by dynamically adapting the FFR assignments according to the channel quality measurements (CQI) or the path loss of the users.

The FFR schemes can be classified into *Partial Frequency Reuse* (PFR) [37,42, 43] and *Soft Frequency Reuse* (SFR) [37, 44] schemes. They differ in the reservation of the transmission resources within the transmission frame. PFR divides the transmission resources in four distinct areas. In particular, it provides a separate frequency reuse zone for cell center users within the transmission frame. In Figure 2.5(a), the allocation at the transmission frame is depicted. The frequency band on left is used of all sectors as shared resources for cell center users. The other part of the transmission frame is divided into three equal frequency bands that are assigned to cell edge users.



(a) Partial Reuse.

(b) Soft Reuse.

Figure 2.5: *Frequency arrangement of the FFR schemes.*

In contrast, SFR does not rely on a shared reuse zone. The resources are split into three sets. Randomization is applied to each set of resources so that the interference is averaged over the resource set. Each sector is associated to a *home-band* which is the band that the sector would be assigned to when using *Frequency Reuse 3*. The remaining two bands are referred to as the *side-bands* of the sector. The cell-center users are allowed to use the home-band and the side-bands, whereas the cell-edge users are restricted to use only the home-band. Thus, the effective overall frequency reuse factor is still close to one which guarantees a high spectral efficiency. The restriction for the side-band users is usually implemented by defining transmit power limitations (*power mask*) for the certain frequency band. In this way, users that are located at the cell-edge and require to send with high power, are only allowed to use the home-band of a sector. Figure 2.5(b) shows a transmission frame divided in several frequency resource blocks.

2.3 Related Work

In this section, related work is summarized. First, common resource allocation approaches are enumerated. The approaches are grouped by the resource allocation objectives. We review approaches that optimize (1) the transmission power, (2) the fairness within the network, (3) that satisfy Quality of Service (QoS) requirements of applications, (4) that handle the resource allocation of multiple antenna systems, and (5) that control the network load with load and admission control. After this subsection, related work for interference mitigation is summarized.

2.3.1 Resource Allocation Objectives and Approaches

Resource allocation is an essential part of an OFDMA mobile communication network. It takes advantage of the flexible allocation of resources of the physical layer and allocates resources in a way that the system capacity is increased and common problems are tackled. On the one hand, the system capacity is increased by the resource management. On the other hand, however, the approaches may require complex algorithms and signaling in the network which may again diminish system performance. The difficulty to design a good resource allocation algorithm is always increased by a variety of boundary conditions and objectives that must be all considered within a holistic resource management for a mobile network. In the following, we enumerate the most important ones and provide related work:

Transmission power In downlink direction, the available transmit power of the base station has to be distributed between all the users according to their propagation loss and/or traffic demands. In uplink direction, mainly interference issues must be considered, see next subsection.

Fairness The resource management should consider a kind of fairness for the users. Otherwise, users at the cell edge or far away from the base station might get only insufficient throughput.

Quality of Service (QoS) requirements Applications usually have certain demands on the network (minimum throughput, maximum latency, etc.) which should be considered by the resource management.

Multiple-antenna systems Current base stations as well as current mobile phones may make use of several transmit and receive antennas. This represents an additional degree of freedom for the resource allocation. In fact, the resource management has to define which data should be sent over which antenna.

Load and admission control Resource management should define an absolute threshold below which a reliable operation of the network can be guaranteed. By contrast, new users should be rejected in case of an operation above this threshold since the network cannot support them in any meaningful way.

A holistic resource management should address as many points as possible that are mentioned above. The following section discusses the individual points and provides related work.

Transmission Power One of the most fundamental issues tackled by resource management is the constraint that only a limited transmission power for all users is available at the base station. In general, this means that the transmission power has to be distributed between the users at different distances and different transmission channels to the base station. There are two approaches in the literature that deal with this problem. With the *margin adaptive approach*, the transmission power of the base station is minimized while the data rate requirements of users are met. In contrast, the *rate adaptive approach* utilizes the full transmit power and tries to maximize the data rates of the users. In [45], the margin adaptive approach is defined and the resource allocation is formulated as an optimization problem to minimize the transmission power to the users. The premise is that all the data rate requirements of users are fulfilled. In contrast, in [46], a rate adaptive approach is presented where the data rate is maximized for a certain transmit power. Consequently, in [46], the overall system capacity is limited by the transmit power of the base station. To compare the results of the algorithm with the optimal solution with respect to spectral efficiency, a water-filling algorithm is used to calculate the optimal solution in the paper. In [47], the bit error probability at physical layer is additionally taken into account, which is intended to prevent resource allocations with very high bit error rate. All in all, the resource management of all work mentioned above is based on solving a nonlinear optimization problem for a single cell, under the assumption that the resource management perfectly knows the channel conditions of the users and does not care about resulting interference. Further on, they all assume saturated traffic conditions, i.e., it is assumed that all users always want to transmit data at all times. Thus, these approaches are all rather theoretical work.

Hence, other research focuses on realistic approaches with fewer assumptions and practical realizations of the resource management problem in order to ensure the applicability. In [48], a two-step algorithm is proposed to reduce the computational effort. The objective is to maximize the overall data rate for a limited maximum transmission power. In the first step, the number of sub-carriers are determined that are required for the users according to the average channel quality.

In the second step, the sub-carriers are then allocated to the users according to the actual signal quality of the users. In [49], a similar approach is used to minimize the transmit power while maintaining a minimum data rate. A comparison of these sub-optimal strategies against the results of the nonlinear optimization approach, as well as a comparison of the required runtimes of the algorithms, is given in [50].

Fairness In addition to the problem of optimal allocation of transmit power, another significant aspect in mobile communications is to maintain the fairness property of the system. In general, users are disadvantaged at the cell edge with respect to the users in the cell center, as they have worse channel conditions. In order to also achieve an acceptable data transfer for those users, approaches such as max-min fairness are proposed, in which the lowest data rate in the cell is maximized. The disadvantage is that the increase of the minimal data rate is at the expense of the total capacity. An alternative to max-min fairness is the proportional fairness. Here, the data rate is distributed proportional to the previous throughput and the channel quality. In [51], the subchannels are allocated to the users with the best channel-to-noise ratio, while a set of proportional fairness constraints is imposed to ensure that each user can achieve the required data rate as requested. Overall, the algorithm is based on the theoretical concept in [47]. It is shown that the total capacity is maximized when each subchannel is assigned to a user with the best subchannel gain and when the power is distributed according to the water-filling algorithm. However, in this approach a fixed data rate per user is assumed, which is usually in sharp contrast to the real situation in a mobile network. A similar method is described in [52] and [53]. In [54], in contrast, also heterogeneous traffic is considered. A *Generalized Processor Sharing* (GPS) scheduling is presented that maximizes the throughput for the different flows in the system taking into account the fairness and the total power constraint. Finally, in [55] a two-stage frequency-selective scheduling method is proposed. First, the users are ordered according to a metric such as the relative throughput. The resources are subsequently assigned to a certain number of users with respect to a frequency-selective metric that is not necessarily identical to the ordering metric. It is found that the first step, i.e. the selection of the user, is critical to the characteristics and performance of the scheduler.

Quality of Service Requirements On the one hand, it is useful to maintain the fairness in terms of throughput in the network. The reason therefore is that users with constant poor channel conditions on average should also have the ability to transfer data. On the other hand, if the user is however considered in more detail, one can usually recognize that the user's requirements are more specific than just a certain average data rate over time. A classic example of this are voice calls. In voice communication, a static codec is usually used which has a fixed bit rate. In order to ensure a good voice communication, the network should guarantee at least this particular bit rate.

All specific application requirements are commonly summarized under the term *Quality of Service (QoS)*. A network provider tries to establish a certain QoS for a group of users in such a way that a catalog of network transmission requirements is defined and enforced within the network. Typical requirements include minimum required data rate, maximum packet latency, or even average throughput over a certain time. In the network, these criteria are enforced by resource scheduling according to the resource management in order to guarantee a certain QoS setting for traffic. The difficulty, however, is to schedule the data with respect to the changing radio channels and other criteria of different users. In [56], a first step towards a QoS scheduling is done by defining a class-based scheduling algorithm that prefers users with the largest weighted delay at the base station. The weight is set according to the QoS delay requirements. Further, a token bucket filter is used in conjunction with this scheduling algorithm to additionally ensure a certain minimum throughput in a 3G mobile network. In [57], a three-step heuristic is proposed in order to additionally include fairness. First, the number of subcarriers that a user should get within a time instant, is derived. This is done with respect to three factors: (1) delay, (2) the user's average rate (fairness), and (3) current transmission channel. Afterwards, if there are still resources available, the rest is shared between the active users according to their previous waiting time. In the third step, the so-called Head of Line (HoL)-blocking is considered. The algorithm sorts the active users in a descending order according to a specific HoL packet time to expire. For the scheduling, it iterates over the users in that order. The authors assume that perfect channel information is known at both the transmitter and the receiver. In [58], a scheduler with different QoS service classes is presented. The allocation of sub-carriers is done here by maximizing a utility function that distinguishes between delay-critical video traffic and best-effort traffic. With the classification into different QoS categories, in general, groups of users with similar needs are prioritized equally or forwarding of different groups is done according to the quality of service definitions. For instance, video traffic is especially supported in modern mobile communication systems by special QoS classes. In IEEE 802.16e WiMAX [59], in addition to a best-effort QoS class for default traffic, an Unsolicited Grant Service (UGS) and a Real-Time Polling Service (rtPS) is defined. UGS is a constant-bitrate service. rtPS in contrast is designed to support real-time service flows that generate variable size data packets on a periodic basis, such as MPEG video. However, with the definition of QoS criteria, the complexity of resource management is considerably increased.

One problem with QoS scheduling in general is that the QoS requirements must be signaled to the network. In [60], a scheduling according to QoS service classes is presented. In contrast to [58], however, the trade-off between accurate scheduling and signaling load is addressed. The QoS requirements are enforced on much larger time scales. Due to the longer scheduling periods, the theoretical capacity is not fully utilized, but the signaling load is reduced.

Multiple-Antenna Systems In LTE Release 8 and WiMAX, multi-antenna techniques are introduced. Several antennas at the base station and at the user equipment are used simultaneously to send and receive data. By special coding of the data, the bit error probability is decreased or the data rate is improved. With respect to the resource allocation, this availability of several antennas represent an additional degree of freedom since the data can now be transmitted over several geographically separated antennas with different channel conditions. Such methods usually require a high degree of coordination between multiple transmit and receive antennas, further good channel knowledge in order to increase the data rate, but also to gain robustness against interference.

Beamforming is one multi-antenna technique that can also be used for interference avoidance [38, 61–63]. With beamforming, the majority of the transmission power of an array of antennas can be focused in a certain direction. This is the simplest case of antenna precoding used to exploit transmit diversity by weighting data streams at different antennas. Using the channel state information, the transmitter encodes the data streams per antenna and sends it to the receiver in order to maximize the throughput of the receiver. There are different versions of beamforming: autonomous single-layer beamforming, precoded/codebook-based multi-layer beamforming.

Interference cancelation [32, 36] is another technique to reduce interference in radio resource management which can make use of several antennas. In the so-called *Joint Detection* not only the own signal is evaluated, but also the data of other mobile users is considered and decoded. With this knowledge about other users, it is possible to subtract the signal of other users from the received signal, thus reducing the interference. In order to decode the signal of other users, the mobile device must know the time delay of each user. This can be achieved with large computational efforts or by simple signaling. Moreover, for future OFDMA cellular networks, other techniques are proposed that require an even larger level of cooperation [63–65]. In contrast to *Joint Detection*, the *Coordinated Multi-Point* (CoMP) transmission technology uses *Joint Processing* or *Joint Transmission*. The transmissions to and from base stations of multiple antennas is coordinated and combined throughout the whole network. CoMP is a centralized approach where mobile users with poor or even negative signal-to-noise ratio, e.g. at the cell edge, are allowed to use the full frequency spectrum at the expense of more signaling and coordination.

CoMP techniques can be categorized according to the complexity of coordination and the required communication between the base stations. One approach is to encode, decode, and process the signal from all spatially separated antennas of different cells. This approach is known in the literature as *Network MIMO*. It promises the highest benefits, but needs a lot of computational cost and requires a high capacity for the control channel between the base stations of the mobile network [64]. Another possibility is that base stations exchange only a few information since the connection of base stations has little capacity or high latency, such as in femtocell scenarios. The information in this approach is not sufficient for joint processing, but allows a coordinated allocation of resources over multiple antennas in multiple cells.

Load and Admission Control Relevant works on load and admission control mainly deal with WiMAX mobile networks. The idea is to reject users if the network is overcrowded and not able to support them in a meaningful way. The resource management should define an absolute threshold below which a reliable operation of the network can be guaranteed. New users are then rejected in case the system runs in a state above this threshold. A stationary scenario is often assumed, so that the effects of user mobility can be neglected. Examples thereof are [66–68]. The work of [67] deals mainly with the quality of service architecture of WiMAX. In [68], a stationary scenario is assumed without band-AMC, which allows for the use of fixed bandwidths. In [69], a measurement-based admission control algorithm is proposed. The method is based on adaptive thresholds which refer to the measured frame error rate of the VoIP traffic.

2.3.2 Interference Mitigation Strategies

In addition to the interference mitigation strategies mentioned in the introduction, there are many types of interference mitigation strategies in literature. In [70], a distributed scheme is proposed that disables sub-carriers in cells, when the performance loss due to the external interference is larger than the expected increase in capacity. However, this approach is based on the assumption that interference is dependent on the position of the receiver within the network, which is only the case in very dense network scenarios.

Xiang et al. [37] compare different fractional reuse schemes in OFDMA-based networks and their parametrization. They study SFR and PFR in the downlink in comparison to simple *Frequency Reuse 1* and *Frequency Reuse 3* schemes. They use a static power allocation on the available frequency band. Nevertheless, simulation is done with a sophisticated simulator which uses a packetized constant bit rate traffic model. The scheduling follows a channel-aware *Round-Robin* strategy. The cell edge/cell center users are differentiated by a geometry factor which is SINR based. Furthermore, they concentrate on *Single-Input-Single-Output* (SISO) antenna transmissions with fixed antenna patterns. Doppler et al. [71] also simulate SFR and PFR in the downlink. They use an SINR-based metric to determine an order to allocate the users. Critical users are allocated first. A scheduler is employed which is either a Round-Robin one or it schedules based on an equal throughput time domain fairness criteria. In contrast to other work, Doppler uses a Poisson arrival traffic model in a metropolitan Manhattan-like area as simulation scenario. Rahman et al. [72] does a comprehensive investigation of downlink FFR based on coordination with a utility function and a central controller. The central controller manages the allocation of mobiles to resources. Furthermore, they included TCP traffic in their simulation.

Bohge et al. [73] investigate the impact of four different power mask configurations which result in *Frequency Reuse 1*, *Frequency Reuse 3*, SFR, and PFR. They compare the downlink performance of a network with a central controller against a network without a central controller. They prove that there is a significant gap in performance between a locally optimal scheduler and the results of a global scheduler. Further on, in [74], Zhou and Zein simulate a mobile WiMAX system with OFDMA and PFR. They conclude that coverage and throughput increase compared to *Frequency Reuse 1* and *Frequency Reuse 3*. The work is based on a simulated WiMAX system, but considers only the downlink, a full buffer, and PFR system. Simonsson [75] evaluates reuse schemes in the downlink and uplink of a 3GPP LTE network using a snapshot simulation with full buffer model. *Frequency Reuse 1*, *Frequency Reuse 3*, PFR, and SFR are considered. In the downlink a static transmit power is assumed. In the uplink, power control is employed and compensates for noise and path-loss. Multiple antenna configurations are tested. The best performing configurations for the reuse schemes are compared. Link quality, spatial distribution, and service bandwidth impact are discussed. It is concluded that a simple *Frequency Reuse 1* performs best of the studied reuse schemes. It is further noted that dynamic coordination schemes are required to improve the performance for wideband packet data services.

Within the *WINNER*, *WINNER II*, and *WINNER+* projects of the *Information Society Technologies* (IST), also several static and dynamic approaches for interference coordination and management with and without frequency-selective scheduling are investigated and evaluated [31–34]. In this case, the system architecture and signaling load is additionally taken into account. Among the static approaches, a static SFR and PFR is investigated for the downlink by restricting the transmit power at certain subcarriers throughout the network.

Further, a semi-static variant of FFR for the downlink is investigated in [31] which makes use of a flexible frequency reuse scheme. Here, the size of the individual cell areas is dynamically adjusted to the current load situation. The adaptation is based on preset threshold values for the ratio between the received signal strength of the broadcast channels of their own and with that of the neighboring cells. The advantages of the static or semi-static concepts are the low complexity and easy implementation. The disadvantage is the lack of flexibility, since the frequency reuse scheme must already be defined in the planning phase of the network, and, in comparison to the dynamic method, it achieves lower system performance. In general, FFR approaches are simple and easy to deploy. In fact, they do not rely on signaling. However, if signaling is additionally used, approaches like inter-cell interference coordination [76] or optimal interference mitigation solutions with a central controller [72, 73] can be applied.

In [32], approaches to average or to cancel the inter-cell interference are considered. The goal for interference averaging is to distribute the interference as evenly as possible to avoid excessive stress on individual subcarriers. In OFDMA systems, this can be realized for example by a permutation of the subcarriers. This however collides with the objectives of frequency selective scheduling.

Interference cancellation, in contrast, tries to exploit the channel response of the interfering signal to subtract the actual interference from the desired signal. This requires an accurate estimate of the channel response, and increases the complexity of the receiver at the mobile device. Finally, in [33], beamforming techniques are investigated for interference reduction and there are also graph theoretic approaches like [77, 78]. All the work uses homogeneous user distributions for simulations.

2.4 System Model for Resource Allocation in OFDMA Uplink

This section provides a model for the resource allocation process for uplink transmission in cellular OFDMA networks. There is a significant potential to increase the capacity of the network due to the flexible OFDMA resource allocation. The model provides the fundamental basis for a simulative analysis of resource management strategies.

We consider a network with a set $\mathcal{M}$ of M users connected to a set $\mathcal{S}$ of S sectors. $\mathcal{M}_x$ denotes the set of users connected to sector x and x_i is the sector user i is connected to. For each cell, we define different frequency bands for the possible use of FFR. We distinguish between PFR and SFR such that in case of PFR, there is a shared frequency band in time that is shared between all cells. This definition is consistent with the description in Section 2.2.3. A frequency partition pattern consisting of four digits w:x:y:z with $w, x, y, z \in \{0, 1\}$ is defined that specifies in the first digit w, if a shared frequency band exists. In the case of SFR no shared frequency band is required, which is indicated by $w = 0$

in the pattern. Further, we denote a distinct frequency band for each cell as *home-band*. The other frequency bands are called *side-bands* representing in fact home-bands of other cells. In PFR, the side-bands are strictly allocated to other cells. In SFR, users that do not generate considerable interference, i.e. likely users that are located in the cell center, are allowed to use these bands. For example, using frequency partition pattern *0:1:1:1*, we have three frequency bands per sector and denote the home-band as A and the side-bands as B and C. The power factor $\hat{p}_{x,b}$ defines the fraction of the maximum mobile transmit power P_{max} which is allowed for transmissions in frequency band $b \in Z$, $Z = \{A, B, C\}$, i.e., the maximum transmit power is $P_{x,b}^{max} = P_{max} \cdot \hat{p}_{x,b}$. In the following, we assume that the power factor is equal for all sectors and the maximum power P_x^{max} depends on the frequency band only, i.e. $P_b^{max} = P_{x,b}^{max} \ \ \forall x \in \mathcal{S}$.

2.4.1 Allocation of Transmission Power, Modulation and Coding

Let us now consider a user $i \in \mathcal{M}_x$ that has to transmit V bits of data. Further, let $I_{x,b}$ be the average interference for frequency band b at sector x and $L_{i,x}$ be the average propagation gain from i to x. If user i uses *Modulation and Coding Scheme* (MCS) k, it occupies $R_k(V)$ resource elements (REs) and requires an SINR γ_k^* for transmission. The power $P_{k,b}^{sc}(V)$ that a user can spend per subcarrier depends first, on the power factor specified by the interference mitigation scheme and second, on the maximum number $C_k(V)$ of parallel subcarriers that the $R_k(V)$ REs occupy. Consequently, the power per subcarrier is $P_{k,b}^{sc}(V) = P_b^{max}/C_k(V)$. We consider open loop power control as specified in [79]. The power control adjusts the transmit power to the MCS specific SINR target γ_k^* such that the required power per subcarrier for MCS k is

$$P_k^*(I, L) = \frac{\gamma_k^* \cdot (N_0 + I)}{L} \quad , \tag{2.1}$$

where N_0 is the noise power. Consequently, on frequency band $b \in Z$, a user i may use all MCSs that require less than the maximum power, i.e.

$$\mathcal{K}_{b,i} = \{k | P_k^*(I_{x_i,b}, L_{i,x_i}) \leq P_b^{max}\} \tag{2.2}$$

is the set of available MCSs. If the standard adaptive modulation and coding (AMC) strategy is used that chooses the most resource-efficient MCS, then the MCS $k_{i,b}$, selected for user i when allocated to frequency band b, is

$$k_{i,b} = \arg \min_{k \in \mathcal{K}_{b,i}} \{R_k(V)\}. \tag{2.3}$$

The required transmit power on frequency band b is

$$P_{i,b} = P_{k_{i,b}}^*(I_{x_i,b}, L_{i,x_i}) \ , \tag{2.4}$$

and the number of required REs is $R_{i,b} = R_{k_{i,b}}(V)$.

2.4.2 User Allocation Metric

After modeling of power control and modulation, the next step is to allocate the users to the transmission frame. Whether they generate interference to other sectors or not depends on their placement in the frame. The resource allocation strategy is executed at each sector independently.

For each sector $x \in \mathcal{S}$, the users are considered according to a resource allocation ordering metric O. It allows to decide which users to put into the cell-specific home-band. The basic principle is to preferably allocate critical users to the home-band. In order to do this, the users within a sector are made comparable by defining the metric $O(u)$ that estimates the criticalness of a user u. The criticalness may depend on the generated interference or on other factors. Various possible metrics are defined and evaluated later on. Higher values of $O(u)$ indicate lower criticalness.

Figure 2.6 shows how the user allocation order affects the placement within the transmission frame. In this depicted example, *User 4* is considered as the final user according to the allocation metric. Thus, and due to the resource demands of the other users (see transmission frame example in Figure 2.6), he needs to be placed on a foreign side-band of another sector when using FFR. In the case of *Frequency Reuse 3*, the user is even blocked, since there are no more free resource available.

We now define different user allocation metrics. The first method $O_{Random}(u) = U(0,1)$ is called *RandomOrder*. Here, the users are assigned to the transmission frame in random order according to a uniform value $U(0,1)$ between 0 and 1.

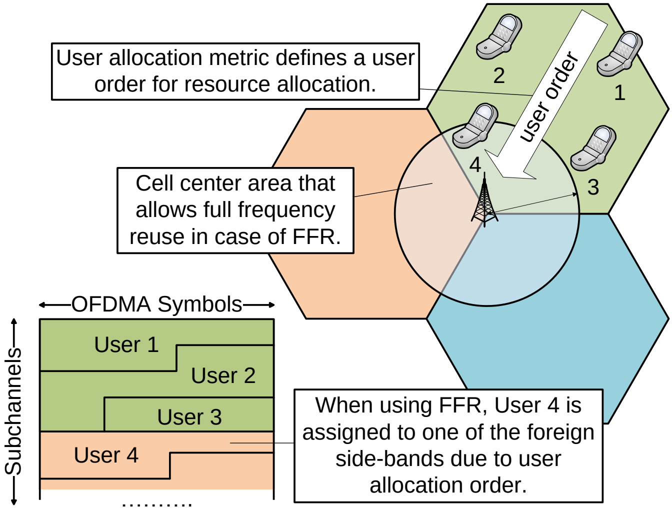

User allocation metric defines a user order for resource allocation.
2
1
user order
4
3
Cell center area that allows full frequency reuse in case of FFR.
OFDMA Symbols
Subchannels
User 1
User 2
User 3
User 4
....
When using FFR, User 4 is assigned to one of the foreign side-bands due to user allocation order.

Figure 2.6: *User allocation metric.*

The second allocation order is called *PropGain* which uses the propagation gain $L_{u,\bar{x}}$ of a mobile u to its assigned sector $\bar{x}$ as allocation metric

$$O_{PropGain}(u) = L_{u,\bar{x}} \quad . \tag{2.5}$$

The mobile with the highest propagation loss, i.e. lowest propagation gain, far from the base station, is scheduled first since it requires more resources than a mobile next to the base station. Furthermore, the mobile is also expected to generate a high interference due to the large distance to the base station, which additionally supports the early assignment of such a mobile to the transmission frame.

In addition to *PropGain* allocation order, *PropGainRatio* additionally considers the propagation gain to other sectors $x \neq \bar{x}, x \in \mathcal{S}$. The propagation gain $L_{u,\bar{x}}$ of a mobile u belonging to sector $\bar{x}$ divided by the propagation gain to all other sectors $\sum_{x,x\neq\bar{x}} L_{n,x}$ is derived for all frequency bands $b \in Z$. The largest ratio of all frequency bands

$$O_{PropGainRatio}(u) = \max_{b\in Z}\left[\sum_{x_b,x_b\neq\bar{x}} \frac{L_{u,\bar{x}}}{L_{u,x_b}}\right] \tag{2.6}$$

defines the priority of a mobile.

It is also possible to consider the generated interference of users to other sectors for the allocation order metric. *IntfSum* is defined as the maximum of all frequency bands of the following term. The average required transmit power to other sectors $E[P^*_{u,b}]$ that transmit at frequency band b for user u is weighted by the propagation gain L_{u,x_b} to these sectors. This transmit power is used as an indicator for the generated interference,

$$O_{IntfSum}(u) = \max_{b\in Z}\left[\sum_{x_b,x_b\neq\bar{x}} \frac{1/E[P^*_{u,b}]}{L_{u,x_b}}\right] . \tag{2.7}$$

2.4.3 Limitation Strategy for Frequency Bands

The *Limitation Strategy* defines the resource limitations for particular resources within the sector. Due to this, a resource partitioning such as *Frequency Reuse 1*, *Frequency Reuse 3*, or FFR can be implemented by restricting a certain frequency band.

A *Limitation Strategy* $Q(u, k, b)$ is defined to restrict the users that are being allocated to the frequency band b. During allocation, the user u with MCS k is checked whether the allocation to this frequency band is allowed or not. If $Q(u, k, b)$ returns 0, the user may use the given resources with transmit power resulting from given MCS. If $Q(u, k, b)$ returns 1, the user is limited and the resource allocation algorithm may not allocate him to the given frequency band. To do so, especially the number of users allocated to the side-bands can be restricted to ensure a good network performance. Implicitly, this also defines the area for the full reuse of frequencies, cf. Figure 2.7. When using the distance to the base station as a metric, the so-called cell center can be determined. However, the distance is often difficult to measure, therefore metrics such as the propagation gain or the signal strength are more common. In the following we define some metrics that are evaluated later on.

If there is no limitation at all for each frequency band b, the strategy is called $Q_{Unlimited} = 0$.

Furthermore, if $h_{\bar{x}} \in Z$ corresponds to the home-band of sector $\bar{x}$, where user u is assigned to, and the following applies

$$Q_{zero}(u, k, b) = \begin{cases} 0, & \text{if } b = h_{\bar{x}} \\ 1, & \text{else} \end{cases}$$

only the home-band can be used which corresponds to a *Frequency Reuse 3* in case of SFR.

To restrict the users in the side-bands, two metrics are evaluated in this thesis. First, $Q_{AggregatePower}$ allows a group of users to transmit at the side-

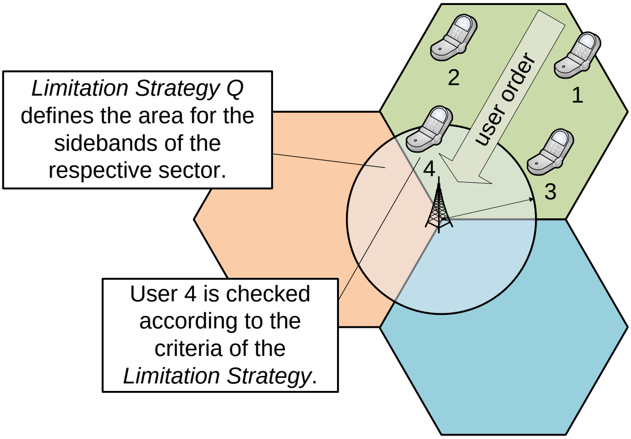

Limitation Strategy Q defines the area for the sidebands of the respective sector.
User 4 is checked according to the criteria of the Limitation Strategy.
user order
1
2
3
4

Figure 2.7: *Limitation strategy, which defines the area for the side-bands.*

bands if their aggregated transmit power is lower than a threshold T. Second, $Q_{PowerLimitation}$ works the other way round and individually limits the power of side-band users which results in a low generated interference of these users. This is equal to the power profile proposed for the downlink in [37].

2.4.4 User Selection Metric in Case of Outage

After allocation metric and limitation strategy, outage selection $B(u)$ assigns a probability to each user u that the user is blocked if the resources are not sufficient for the allocation of all mobiles. The idea is not to block the user who does not get resources anymore, but specifically select a user according to this probability, so that the network performance can be improved. If not all users are assigned to resources, a user is selected and blocked according to this metric. The complete resource allocation is then repeated without this user until all remaining users can be assigned to the transmission frame.

Let $\mathcal{M}_x$ be the set of users assigned to sector x. $B_{Random}(u) = \frac{1}{|\mathcal{M}_x|}, \forall u \in \mathcal{M}_x$, ensures fairness. The users are all blocked with the same probability.

$B_{OutageMin}$ blocks the user u with the lowest propagation gain $L_{u,\bar{x}}$. It is defined as

$$B_{OutageMin} = \begin{cases} 1, & \text{if } u = \arg\min_{v \in \mathcal{M}_x}(L_{v,\bar{x}}) \\ 0, & \text{else} \end{cases} .$$

Finally, $B_{WeightedRandom}$ provides a mixture of both which achieves a trade-off between fairness and performance. The users are weighted according to weightening factor α. The weightening factor can be set so that it can be decided between random outage and outage minimization.

For $\alpha = 0$ the user is selected randomly. For $\alpha = \inf$ the user with the lowest propagation gain is selected:

$$B_{WeightedRandom}(u) = \frac{(1/L_{n,\bar{x}})^{\alpha}}{\sum_{v \in \mathcal{M}_x}(1/L_{v,\bar{x}})^{\alpha}} .$$

2.4.5 Interference-Efficient Allocation on the Side-Bands

Adaptive Modulation and Coding (AMC) typically minimizes the number of resources required for transmitting a certain amount of data. This is achieved by using all available power in order to use the most resource-efficient MCS. A low-order MCS, in contrast, is more robust, requires a lower SINR, and according to Eq. 2.1, a lower transmit power per symbol. The disadvantage is that it requires more resources. In the following, we exemplarily investigate the total power of a data transmission, i.e. not the power per symbol but the power to transmit the entire data volume, depending on the used MCS.

To transmit V bits with MCS k, $R_k(V)$ REs are required. Further on, $P_k^*(I, L)$ denotes the required transmit power per subcarrier and OFDMA symbol. With N_{RE} as the total number of symbols per RE, the cumulative transmit power of a mobile in a frame is

$$P_k^{cumul} = P_k^*(I, L) \cdot R_k(V) \cdot N_{RE}. \tag{2.8}$$

The power P_k^{cumul} is proportional to the interference the mobile generates in neighboring sectors. Figure 2.8 shows P_k^{cumul} as a function of the used MCS. Interference from other mobiles is not assumed, i.e. $I_{x_i} = 0$, and the propagation gain L is set to -100 dB. Obviously, the cumulative transmit power is by far higher if higher-order MCSs are used. Additionally, the difference between two consecutive MCSs is increasing. This becomes clear when considering that the Shannon capacity for wireless transmissions in general increases with the logarithm of the SINR [80]. In contrast, the cumulative transmit power increases only linear with the number of resources.

For the same data volume, consequently, the cumulative power required to send the data is considerably lower for a low-order MCS compared to a high-order MCS. This fact makes it favorable to select lower MCSs if possible. As a consequence, all available REs within a transmission frame should be utilized by assigning low-order MCSs instead of transmitting on a few REs with high-order MCS.

Application of the Findings in Resource Management

A trade-off exists between using a low-order MCS which is *power-efficient* and thus, also inter-cell interference reducing, and a high-order MCS which allows to support a large number of users, i.e., a high-order MCS is *resource-efficient*.

With the MCS optimization that we propose in this subsection, the resource allocation is able to optimize the frame according to both objectives. As usual, the first objective is a resource-efficient allocation. Power control determines the power required for a transmission at MCS k and propagation gain L. MCS k is chosen by the AMC function. Thus, the primary goal of AMC and power control is the minimization of resources. This method works well for systems that are rather resource than interference-limited. This applies for a *Frequency Reuse 3*

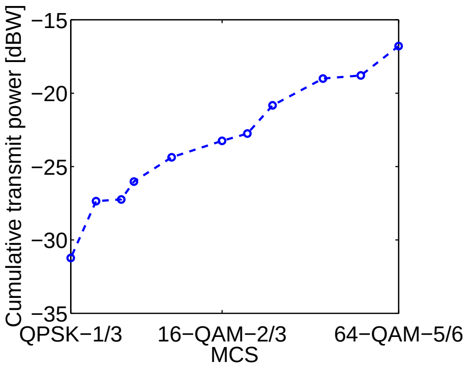


Figure 2.8: *Relation of cumulative transmit power and MCS for a fixed data volume.*

scheme in the home-band of an SFR, for instance. The allocation strategy for the home-band is to serve as many users as possible as the number of users is limited by the available resources only. Additionally, the inter-cell interference is not too critical since in the neighboring sector, this frequency band is used as side-band and serves users with probably high propagation gain.

With a *Frequency Reuse 1* scheme or in the side-bands when using SFR, the situation is different. All resources can be used with the consequence that inter-cell interference occurs. The system may become interference-limited depending on the number of users. In such a case, reducing the MCS results in a more power-efficient transmission and, hence, decreases the total inter-cell interference. As mentioned above, the effect of a power-efficient resource allocation is in particular meaningful if there are available resources in some sectors while neighboring sectors are crowded. This happens with heterogeneous user distributions, due to under-utilized cells, or inefficiently scheduled frames. Especially in the uplink, where the typical traffic consists of acknowledgments and HTTP requests, the case that not all resources are utilized is rather frequent. If not all resources are used, it is better to choose a low-order MCS to spread the users over all available resources. With such an allocation, the transmit powers of the users become lower and consequently, the inter-cell interference is reduced and the overall throughput of the network is increased. This principle is illustrated in Figure 2.9.

The side-bands of an SFR are interference-critical since the users allocated to the side-bands produce interference to the home-band of the neighboring sectors. The maximum allowed power of the side-band is commonly configured with the objective to maximize the system capacity when all cells are fully loaded. If some cells experience a lower load, they will not entirely utilize the resources available on the side-band and there is room for selecting low-order MCSs. This decreases the interference in the home-bands of the neighboring sectors such that more users can be allocated to these home-bands.

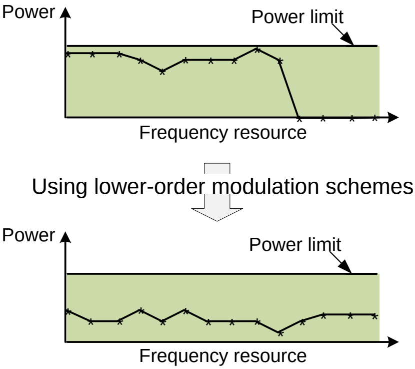


Figure 2.9: *MCS optimization concept in the uplink.*

Definition of the MCS Optimization

We introduce in the following an algorithm to change a *resource-efficient resource allocation* to a *power-efficient resource allocation*. The general idea is to consider all side-bands separately and to start with a resource-efficient resource allocation according to the standard AMC as described in Section 2.4.1. Afterwards, all users in the side-band are considered iteratively and the next MCS with low-order is chosen if possible, i.e. if free resources are available. Three metrics, i) *Highest MCS First*, ii) *Highest Cumulative Power First*, and iii) *Highest Inter-Cell Interference First*, are considered.

The MCS optimization is defined as follows:

Let R_b be the number of resources occupied in frequency band $b \in Z$ and R_b^{max} be the number of available resources. Then, we define the next lower MCS k_u^{next} of user u as

$$k_u^{next} = \max\left\{k | R_k(V) > R_{k_{u,b}}(V)\right\}. \tag{2.9}$$

The function

$$r(u) = \begin{cases} 0 & \text{if } R_b - R_{k_{u,b}}(V) + R_{k_u^{next}}(V) \leq R_b^{max} \\ 1 & \text{else} \end{cases} \tag{2.10}$$

determines the possibility to allocate user u to the MCS with next lower order. The corresponding algorithm is outlined in Algorithm 1.

Algorithm 1 MCS optimization algorithm

```
$\mathcal{Q} = \mathcal{M}_{x,b}$
while $\mathcal{Q} \neq \emptyset$ do
    Choose user $u$ according to metric defined below
    if $r(u) = 1$ then
        $R_b = R_b - R_{k_{u,b}}(V) + R_{k_u^{next}}(V)$
        $k_{u,b} = k_u^{next}$
    else
        $\mathcal{Q} \leftarrow \mathcal{Q} \setminus u$
    end if
end while
```

The following metrics are defined to choose the next user to consider for MCS optimization:

i) Reduce Maximum MCS First

The idea is to choose the user with highest MCS first. If the MCS is equal for two users, the metric $O(u)$ is used. *Reduce Maximum MCS First* yields the highest gain per symbol. The next user u is selected by

$$l = \max_{v \in \mathcal{M}_x} k_{v,b} \quad , \tag{2.11}$$

$$u = \arg \min_{\left\{v \in \mathcal{M}_{x,b} | k_{v,b} = l\right\}} O(v) \quad . \tag{2.12}$$

ii) Reduce Maximum Power First

The idea is to choose the user with maximum cumulative power first.

$$u = \arg \max_{v \in \mathcal{M}_{x,b}} P^{*}_{k_{v,b}}(I_{x_v,b}, L_{v,x_v}) \cdot R_{k_{v,b}}(V) \cdot N_{RU} \quad . \tag{2.13}$$

iii) Reduce Maximum Interference First

The idea is to reduce the MCS of the user that leads to the highest reduction in inter-cell interference per additional resource. The difference in inter-cell interference is defined as

$$T(j) = \frac{I^{oc}_{j,k_{j,b},b} - I^{oc}_{j,k_j^{next},b}}{R_{k_{j,b}}(V) - R_{k_j^{next}}(V)} \tag{2.14}$$

with

$$I^{oc}_{j,k_{j,b},b} = P^{*}_{k_{j,b}}(I_{x_j,b}, L_{j,x_j}) \sum_{y \in \mathcal{S} \setminus x} \frac{1}{L_{j,y}} \tag{2.15}$$

The next user is the one with maximum $T(j)$.

2.5 Simulation Methodology

In this section, the simulation is described. It discusses the simulation methodology and presents the evaluation scenario.

2.5.1 System-Level Simulation Technique

The system level simulations of the OFDMA uplink are carried out using a time-invariant *Monte Carlo simulator* [81]. It is based on fundamentals described in the evaluation methodology document for mobile communications of the IEEE 802.16m standard [82]. The uplink resource allocation of one transmission frame is simulated with a fixed traffic demand of all users. This means that each user is constantly trying to send v bits and the simulation calculates a possible solution according to the power and resource allocation algorithms.

The cell simulation case is based on a 5x5 deployment with hexagonal 3-sector sites. In order to avoid bounding effects and thus, an overestimation of the system performance, wrap around is applied which ensures that all cells experience the same interference characteristics. The simulator is able to process non-MIMO antenna configurations including different downtilts, diverse antenna patterns, and different constant traffic volumes per user. In all simulations, error free feedback from the user to the base station is assumed. Shadowing is included according to the urban macrocell path loss model [82].

The *Monte Carlo simulation technique* is used which means an iterative stochastic simulation over one uplink frame is evaluated. It works as follows. Each cell serves a constant number of users which are distributed according to a spatial random process. The users try to transmit a fixed number of bits while the power control tries to compensate interference by increasing power and instructing the adequate modulation and coding scheme. Afterwards, the users are scheduled and assigned to uplink slots according to the different algorithms. The interference of all users is calculated per sector which is the basis for the next iteration. The steps are now repeated as long as the interference level does not

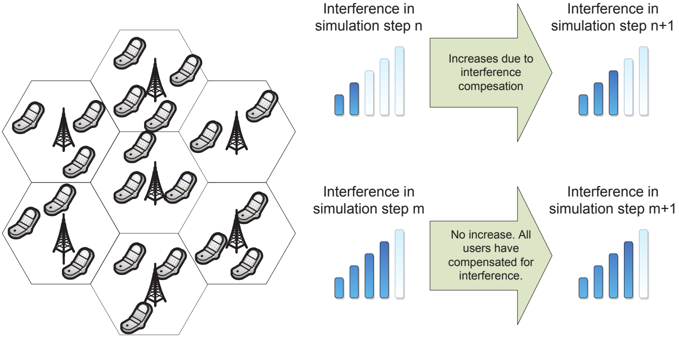

Interference in simulation step n
Increases due to interference compesation
Interference in simulation step n+1
Interference in simulation step m
No increase. All users have compensated for interference.
Interference in simulation step m+1

Figure 2.10: *Simplified illustration of the simulation process.*

converge from iteration n to iteration $n+1$. If the interference level stabilizes at iteration $m > n$, the simulation process is stopped. A simplified illustration of the simulation process is given in Figure 2.10. The final state of the simulation provides a scenario with fixed transmit powers and fixed interference and noise.

2.5.2 Antenna Patterns

The simulator is able to process different antenna patterns. The performance of a mobile communication network is dependent on the antenna configuration. The configuration has the purpose to, on the one hand, support all users in the cell, which means it should provide sufficient cell coverage. However, on the other hand, it also has to avoid a cell overlap and in turn, provide good cell isolation to avoid inter-cell interference.

There are two different mechanisms available. Either the antenna can be tilted mechanically or the antenna tilt can be achieved by electrically changing the phases of antennas which are closely together in an antenna array. The first method is called *mechanical downtilt*. The second method is called *electrical downtilt*. The electrical downtilt is more complex but also more easy to adapt.

We further distinguish in the following between the gain on the vertical plane due to downtilt and uptilt, and the gain on the horizontal plane defined by the user location within the sector. First, we describe the pattern on the vertical plane to the user.

In [83], a *vertical electrical downtilt* antenna radiation model is given. It is verified with measured data of real networks. The vertical deviation from antenna boresight is calculated as illustrated in Figure 2.11. Angle β is calculated as the difference between the downtilt angle and the vertical angle of the direct line between the mobile and the sector antenna. Angle θ_{tilt} describes the vertical tilt of the base station antenna while θ_{3dB} defines the angle between an attenuation of 3dB when deviating from the tilt direction into both directions. For this model, the vertical antenna gain is independent of a horizontal angle α between the mobile and the base station. This may have an impact on the performance of a system as users in the center of a neighboring sector may become interference critical. The vertical electrical downtilt pattern is given by

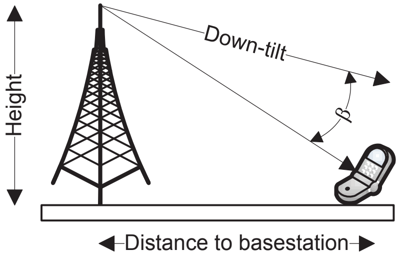


Figure 2.11: *Illustration of the downtilt angle.*

$$G_{el.}(\beta) = -\min\left[12 \cdot \left(\frac{\beta - \theta_{tilt}}{\theta_{3dB}}\right)^2, \phi_{SLL}\right] , \qquad (2.16)$$

where $\phi_{SLL} = 20\,dB$ is the side lope level, relative to the maximal gain of the main beam. Figure 2.12 shows the attenuation of the electrical downtilt as a function of the mobile's distance from the base station. In this figure, the base station antenna height h_b is assumed to be 30 m. The height of the user antenna h_m is assumed to be 1.5 m. The three curves each represent a different angle θ_{tilt}. The value of θ_{3dB} is 6.2 °.

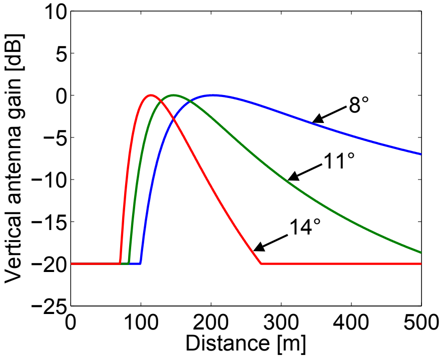


Figure 2.12: *Vertical electrical antenna pattern under different downtilts.*

The mechanical downtilt in contrast considers the effect of mechanically tilting the base station antenna by an angle θ_{tilt}. The vertical attenuation depends on the angle α between horizontal antenna direction and user direction. It is given by

$$G_{me.}(\alpha, \beta) = -\min\left[12 \cdot \left(\frac{\beta - \theta_{tilt} \cdot \cos\alpha}{\theta_{3dB}}\right)^2, \phi_{SLL}\right]. \tag{2.17}$$

Additionally to the vertical downtilt, an antenna has also got a horizontal radiation pattern. The directivity in the horizontal plane (azimuth) for the antenna model is proposed by many system evaluation documents to simulate a realistic scenario [82, 84]. In [83, 84] the following model is recommended:

$$G_{horiz.}(\alpha) = -\min\left[12 \cdot \left(\frac{\alpha}{\alpha_{3dB}}\right)^2, \phi_{FRB}\right] , \tag{2.18}$$

where the angle in which a user will experience an attenuation of 3 dB is defined as α_{3dB}. Furthermore, $\phi_{FRB} = 20\,dB$ is the front back ratio of the antenna. Figure 2.13 shows the attenuation of a signal in dB as a function of the horizontal angle between the base station antenna and the mobile as defined in Equation 2.18. The value of α_{3dB} is 70 °. This results in a gain of $-3\,dB$ at an angle α of $\pm 35°$.

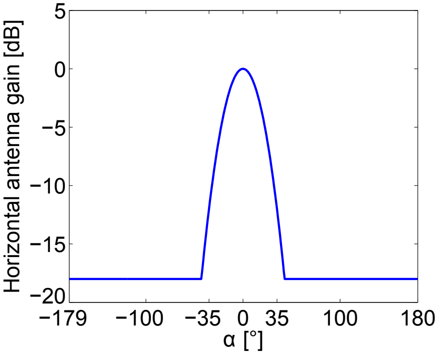


Figure 2.13: *Horizontal antenna pattern.*

2.5.3 Simulation Parameters

Table 2.1 shows the central modeling parameters and assumptions which have been used in the simulation studies. The physical OFDMA access scheme is designed for frequency bands below 11 GHz. For the simulations, frequencies in the 3.5 GHz band are chosen. We consider an OFDMA system with fast Fourier transform (FFT) size of 512 subcarriers. After eliminating the guard subcarriers, we effectively use 432 data subcarriers. The simulation also supports various spatial user distributions by random point processes such as the Matérn cluster process [85] or a Poisson process [86].

2.5.4 Scenario Description

For the system evaluation, several parameters are used which are enumerated in the following subsection. Mainly, the results are generated with 1 to 26 users per sector. Each user transmits 1024 bits per frame. If not specified otherwise, the optimal vertical downtilt of the base station antenna was determined prior the evaluations. In most simulations, it is set to 11°. The downtilt is highly dependent on the cell size. If not properly configured, the cell is either not fully covered or the base station causes interference to other sectors.

2.5.5 Performance Metric

To evaluate the performance of the network at different algorithms and reuse schemes, a performance measure is necessary. In this work, we consider the total outage percentage as performance measure, i.e., the relative number of users in a cell whose transmission is being blocked. Since we simulate a fixed amount of v bits for each user, it also corresponds to the actual possible throughput of the sector due to the particular interference situation depending on the user current locations with this traffic demand.

Every plotted curve of the results shows the mean value of at least 30 samples. Additionally, the 95 % confidence intervals are drawn to ensure the reliability of the stochastic results.

Table 2.1: *Simulation parameters*

Cellular layout	25 hexagonal, trisectorized cells
Site-to-site distance	0.5 km
BS/UE antenna height	30 m, 1.5 m
Carrier frequency, bandwidth	3.5 GHz, 5 MHz
FFT size, # subchannels	512, 24
Frame length	5 ms
OFDMA symbols per frame	48
Subchannel mode	Additional optional symbol structure for PUSC [59]
Modulation and coding schemes	QPSK-1/3, QPSK-1/2, QPSK-2/3, QPSK-3/4, 16-QAM-1/2, 16-QAM-2/3, 16-QAM-3/4, 16-QAM-5/6, 64-QAM-2/3, 64-QAM-3/4, 64-QAM-5/6
Antenna configuration	Single-Input-Single-Output
Antenna horizontal pattern *(unless otherwise specified)*	$G(\alpha) = -min\left[12 \cdot (\frac{\alpha}{\alpha_{3dB}})^2, \phi_{FRB}\right],$ $\alpha_{3dB} = 70^\circ, \phi_{FRB} = 20\,dB$
Antenna vertical pattern *(unless otherwise specified)*	$G(\beta) = -min\left[12 \cdot (\frac{\beta-\theta_{tilt}}{\theta_{3dB}})^2, \phi_{SLL}\right],$ $\theta_{3dB} = 11^\circ, \phi_{SLL} = 20\,dB$
BS/UE antenna gain	14 dBi/0 dBi
UE thermal noise density	-174 dBm/Hz
Path loss model	Urban Macrocell [82], $PL[dB] = 35.2 + 35\log_{10}(d) + 26\log_{10}(f/2)$
UE maximal power	200 mW
Channel State Information (CSI)	Perfect CSI

2.6 Performance Evaluation of Resource Allocation Strategies in the OFDMA Uplink

We now provide a performance evaluation of resource allocation strategies in the uplink of an OFDMA network. The performance evaluation is divided into six subsections and covers three different research directions in the area of resource management.

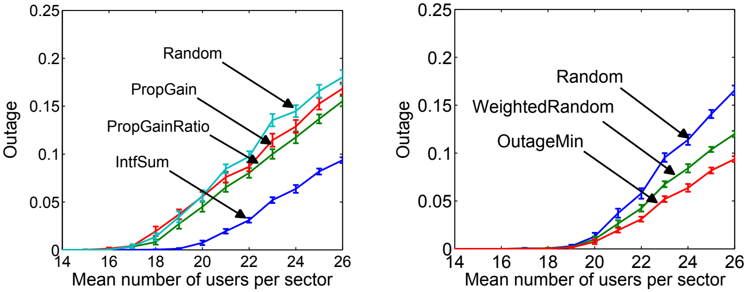
The first three subsections address the research about *user and resource coordination*. We show results on user allocation strategies (Section 2.6.1) and user blocking strategies (Section 2.6.2, outage selection). For these evaluations, we restrict ourselves to full frequency reuse (*Frequency Reuse 1*) and *Frequency Reuse 3*. On this basis, the performance of *SFR* is investigated afterwards for the uplink of a mobile communication network (Section 2.6.4). Subsequently, the influence of the antenna configuration is examined for all investigated approaches (Section 2.6.3 and 2.6.5). This also represents the second research direction. Thereafter, research related to resource management according to different traffic characteristics is conducted and the performance of the resource allocation for non-saturated uplink traffic conditions is evaluated (Section 2.6.6).

2.6.1 Impact of User Allocation Metric on Network Performance

In the following, we consider the influences of the ordering metrics for resource allocation on the outage of a cell. Figure 2.14(a) shows the results for the four strategies that are proposed in Section 2.4.2. The figure shows on the y-axis the relative number of users that experiences outage. The x-axis refers to the mean number of users per sector. In this case, full frequency reuse is used in all 25 cells.

Starting from approximately 16 users on average, outage occurs in this scenario and users are blocked. The system is in overload and either the interference prevents the transmission of some users or the resources in the cell are no longer sufficient for this number of users. Since this is the *Frequency Reuse 1* case and all resources are available, the interference is the limiting factor up to 23 users.

For the area where users are blocked, Figure 2.14(a) further shows that the situation can be improved by the use of an ordering metric. The random metric, which selects an arbitrary user for allocation, performs worst. In contrast, the metric that considers the interference to other sectors (*IntfSum*) is significantly better than all other metrics. With this metric, users that cause a lot of interference are set to opposite areas in the transmission frame with respect to neighboring cells. Accordingly, the allocation should be done with regard to expected interference so that quality degradations to other sectors are avoided, if possible. In contrast, if only the propagation gain is taken into account, as with the *PropGain* or *PropGainRatio* metric, there is no significant increase in performance. In both of these strategies, some users are assigned to parts in the transmission frame on which other users would generate less inter-cell interference.



(a) *Frequency Reuse 1* outage with different allocation metrics.

(b) *Frequency Reuse 1* with different outage selection strategies.

Figure 2.14: *Investigation of Frequency Reuse 1 at different user allocation order metrics and outage strategies.*

2.6.2 Impact of Outage Strategy on Network Performance

In Figure 2.14(b), the impact of the outage selection strategies defined in Section 2.4.4 is shown. The axes are kept to outage percentage on the y-axis and average users per sector on the x-axis. The allocation metric is set to *IntfSum* due to the previous results. The strategy *OutageMin* that blocks the user with the lowest propagation gain gives the best results since, here, the cell size decreases for higher loads resulting in less interference. This is because the users at the edge, which generate the highest interference, are preferably blocked. At an outage of 5 %, this strategy supports on average one additional user per cell sector or 3 users more per cell compared to the random strategy.

From the user point of view, this strategy is however unfair in the sense that more users are blocked at the cell edge. A compromise represents the weighted outage strategy *WeightedRandom*. With it, the provider can choose whether he wants to block the users randomly or at the expense of fairness. In Figure 2.14(b), this strategy lies exactly between the other two strategies as expected. The fairest strategy is the random blocking. This is also the most common one in today's networks, as the providers usually does not want to prefer a certain group of users.

2.6.3 Impact of Different Downtilt Configurations

In this subsection, we study the impact of different antenna configuration on the network performance. The aim is to quantify the effects of the different configurations.

To motivate the investigation of different antenna settings, Figure 2.15 shows the spatial distribution of the inter-cell interference level for the *IntfSum* user ordering metric for the home-band. The *electrical downtilt* is depicted with θ_{3dB} = 6.2 °in the left sub-figure. The right sub-figure shows the interference distribution at *mechanical downtilt* with θ_{3dB} = 6.2 °. The positions in dark color, which are marked with an arrow, have low values of interference whereas the dark areas at the border indicate high interference. Therefore, the areas at the border show interference critical positions in a sector. When comparing the two figures, it can be seen that the sector inter-cell interference is changing significantly with the downtilt type.

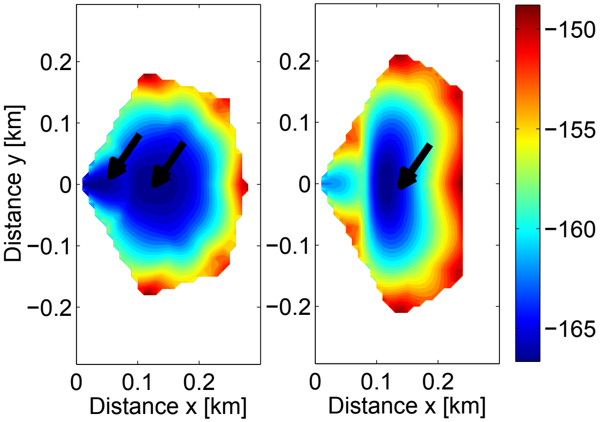
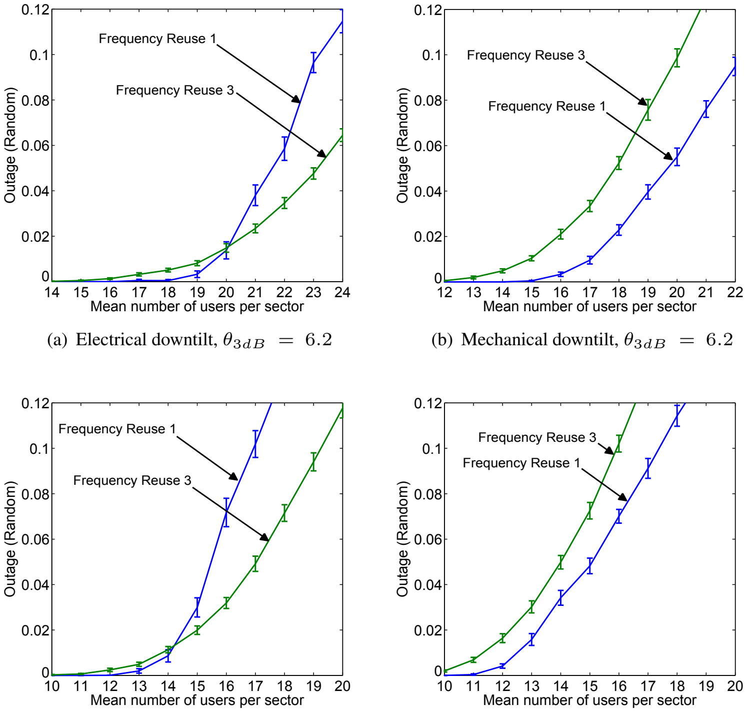


Figure 2.15: *Spatial distribution of the interference level under electrical and mechanical downtilt.*

In Figure 2.16 different reuse schemes for electrical and mechanical downtilt with θ_{3dB} at either 6.2 ° or 15 ° are evaluated. Overall, we can deduce that for the cell size of 0.5 km in our scenario, a downtilt of 6.2 ° outperforms the downtilt of 15 °. This result depends on the cell size and on the frequency reuse scheme. For *Frequency Reuse 1*, where interference is generated, the outage of the users increases with a larger vertical opening angle of the antenna, i.e. higher values of θ_{3dB}. The outage of the *Frequency Reuse 1* intersects *Frequency Reuse 3* at an average load of 16 users for electrical downtilt with θ_{3dB} = 6.2 °. Thus, a *Frequency Reuse 3* should be used in such a case with a large number of users.



(a) Electrical downtilt, $\theta_{3dB} \;=\; 6.2$

(b) Mechanical downtilt, $\theta_{3dB} \;=\; 6.2$

(c) Electrical downtilt, $\theta_{3dB} \;=\; 15$

(d) Mechanical downtilt, $\theta_{3dB} \;=\; 15$

Figure 2.16: *Frequency reuse schemes under different downtilt configurations.*

At lower load in the sector and up to 16 users, a *Frequency Reuse 1* performs better, which accepts nearly all users for transmission. For $\theta_{3dB} = 15\,^{\circ}$, the point of intersection for electrical downtilt is at an average load of 14 users. The users experience outage more earlier since the inter-cell interference is higher for this configuration. For mechanical downtilt, *Frequency Reuse 1* always performs better than *Frequency Reuse 3*.

Overall, the comparison of *Frequency Reuse 1* and *Frequency Reuse 3* shows that *Frequency Reuse 1* works better especially at a lower number of users in the system. The main reason is that sufficient resources are available and the interference does not significantly affect the performance. However, when the load increases and more users are in the system, the interference exceeds a critical point, and the interference becomes the limiting factor. In such a case, a *Frequency Reuse 3* works substantially better, see electrical downtilt. This trade-off also indicates that there is potential for further sophisticated frequency reuse strategies which both provide resources, as well as limit the inter-cell interference.

2.6.4 Network Performance with Soft Frequency Reuse

In this subsection, the SFR scheme is finally examined and compared to the previous approaches. SFR addresses the aforementioned problems and provides more resources as well as a limitation for inter-cell interference.

Figure 2.17 shows the impact of the SFR scheme onto the outage probability of the cell. The outage strategy *OutageMin* is used for this investigation. For reference, *Frequency Reuse 1* is included with both *OutageMin* and *Random* outage strategy. Two different versions of SFR are investigated. First, the limitation strategy *PowerLimitation* is applied. This means that the power limitation is done according to a value defined per user. In this investigation, this factor is set to -10 dB for each user. Second, the limitation strategy *AggregatePower* is used to restrict the transmit power for a group of users. The upper limit for the cumulated transmit power for all users in the side-bands is set to -6 dBW.

Both versions of SFR allow more users on average per cell compared to *Frequency Reuse 1* or *Frequency Reuse 3*. At an outage of 5 %, SFR with *AggregatePower* strategy performs at about 16 %, or 4 users per sector on average, better than *Frequency Reuse 1* or *Frequency Reuse 3*. The performance gain comes mainly through the coordination of the users. Users with less transmit power are allowed to use the side-bands and thus, the sector is able to serve more users than in the *Frequency Reuse 3* case. Furthermore, the interference situation is

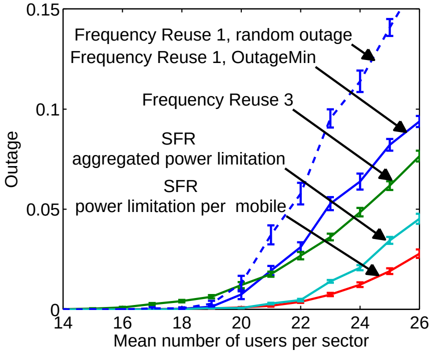


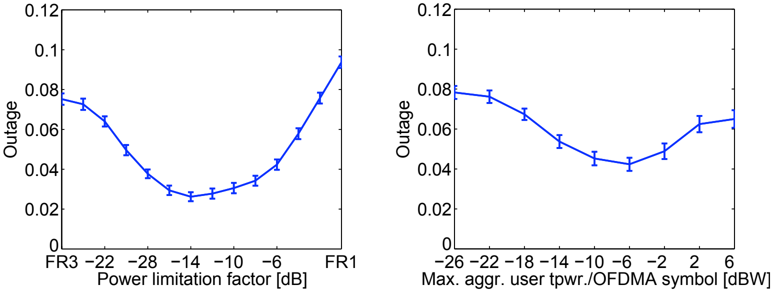
Figure 2.17: *SFR with two different limitation strategies in comparison with Frequency Reuse 1 and Frequency Reuse 3 frequency scheme.*

improved, as the users with potentially high inter-cell interference are placed at the cell's own home-band.

For the results in Figure 2.17, the ordering metric *IntfSum* is used. Comparing the results of various other ordering metrics, similar result as in Section 2.6.1 are obtained, respectively. An ordering metric is also critical in this case since it specifies a sequence after that the users are assigned to the transmission resources.

Another important performance parameter for SFR is the selection of the type of limitation strategy. Besides SFR with *PowerLimitation* strategy which limits individually the power of a mobile, the *AggregatePower* limitation strategy is evaluated. The SFR with this limitation strategy also performs better than *Frequency Reuse 3* but worse than SFR with *PowerLimitation* strategy. This is due to the fact that a limitation according to the aggregated transmit power of a group, allows a single user to use a significantly high transmit power, resulting in much interference. In contrast, a consideration of the individual user as with *PowerLimitation* strategy, prevents this by restricting the highest value for transmission power per user.

The input parameters of the SFR strategies are further investigated in the following. *PowerLimitation* restricts the maximum power of a user if the user wants to send on a side-band. Consequently, the users which require few power and thus, are located next to the antenna are candidates for this zone in the SFR scenario. A parameter study is performed in the following to determine the optimal value. If the parameter is set too low, no user is able to transmit on the side-bands. Thus, the scenario degenerates to a *Frequency Reuse 3* scenario. If the parameter is set too high, every user can transmit and thus, can generate inter-cell interference. Hence, the scenario degenerates to a *Frequency Reuse 1* scenario. Simulations are done for 25 users per sector to determine the optimal *PowerLimitation* parameter. The results are shown in Figure 2.18(a). The x-axis shows the values of the parameter. The y-axis displays the outage with this parameter. The minimum is at -14 dB which results in an optimal *PowerLimitation* parameter. The factor should be adapted along with the increasing number of users (cf. Figure 2.22(a) and Figure 2.23(a)). With more users, the power limit has to be decreased to keep the interference low. For the *AggregatePower* limitation strategy, a similar study is presented in Figure 2.18(b). The tuning parameter for *SFR AggregatePower* is the maximum aggregated transmit power per OFDMA symbol. The optimal threshold was determined to -6 dBW for this cell configuration.



(a) SFR with power limitation per mobile

(b) SFR with aggregated power limitation

Figure 2.18: *SFR limitation parameter optimization.*

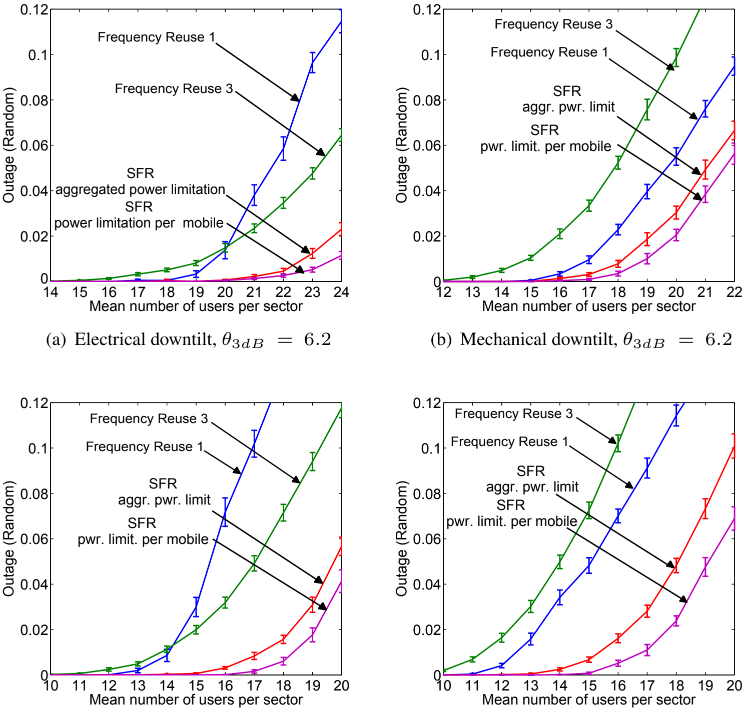
2.6.5 Impact of Different Antenna Configurations on Soft Frequency Reuse

After the presentation of the results on SFR, there is the question of the influence of various antenna configurations on SFR. This is especially true with regard to the results in Section 2.6.3 for *Frequency Reuse 1* and *Frequency Reuse 3*.

The results for SFR are depicted in Figure 2.19. The same way as in the subsection before, we distinguish between SFR with power limitation per mobile (*PowerLimitation*) and aggregated power limitation (*AggregatePower*). Comparing the different sub-figures, it can be observed that the performance of SFR depends on the antenna settings. Although the situation is improved with SFR at all investigated antenna settings, there are significant differences in the performance when the configuration of the antenna is changed.

With electrical downtilt and an opening angle with θ_{3dB} = 6.2 °, an additional gain of about 3 users per sector on average is achieved by *SFR AggregatePower*. This corresponds to a gain of 15 % in terms of allocated users at 1 % outage. For mechanical downtilt, the gain at θ_{3dB} = 6.2 ° is lower. In this case, about 2 more users on average can be supported. If a larger, and according to the results of Section 2.6.3, a worse opening angle is chosen, the performance of SFR also deteriorates. These results are consistent with the results for *Frequency Reuse 1* and *Frequency Reuse 3* in Section 2.6.3. The reason for the performance degradation of SFR is the larger opening angle that allows for more inter-cell interference. These results underline the fact that it is essential to choose a suitable downtilt configuration for SFR in current systems.

To evaluate the impact of the downtilt on SFR in detail, several downtilt angles are compared under varying transmit power limitation factors. The average number of users per sector is 24. Figure 2.20 shows the random outage of the network. The four curves represent the outage under a downtilt of 9 ° to 12 °. The optimal parameters and the achieved outage vary strongly for the four downtilt angles. In the considered scenario, the downtilt of 11 ° performs best.



(a) Electrical downtilt, $\theta_{3dB} = 6.2$

(b) Mechanical downtilt, $\theta_{3dB} = 6.2$

(c) Electrical downtilt, $\theta_{3dB} = 15$

(d) Mechanical downtilt, $\theta_{3dB} = 15$

Figure 2.19: *SFR scheme under different downtilt configurations.*

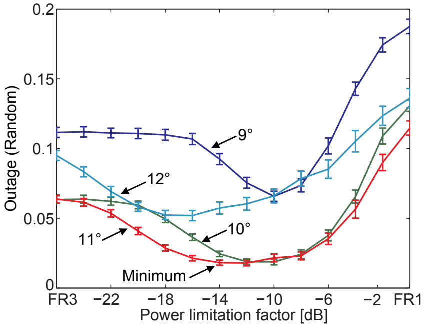


Figure 2.20: *SFR transmission power parameter study under different varying downtilt angles.*

2.6.6 Resource Allocation for Non-Saturated Uplink Traffic Conditions

The previous results have shown that it is essential to choose the right parameter settings for SFR and other resource allocation schemes. Among other parameters, also time varying factors such as the current *load* on the network have an impact on performance. It is therefore desirable to develop a scheme which is less load-dependent and in which the performance is less dependent on network settings and parameters. It should take the current characteristics of today's uplink traffic into account and should mitigate inter-cell interference.

In this subsection, we focus on the uplink with non-saturated traffic conditions. The users are distributed currently in the scenario according to a homogeneous spatial random Poisson process. Poisson point processes are natural models for random configurations of points within a region. For the network scenario, this means that the users are homogeneously distributed over the entire network, but the distances between them are not uniformly distributed, but exponentially distributed. Consequently, the load in a cell may vary locally depending on the user concentration. In particular, there may be cells that have non-saturated traffic conditions.

The resource allocation strategy considered in this subsection is SFR with limitation strategy *PowerLimitation* and MCS optimization as defined in Section 2.4.5. The power limitation factor is set to -12 dB for the side bands.

In the following, *Frequency Reuse 1* and SFR are both evaluated in Figure 2.21 with and without the MCS optimization. The *reduce maximum MCS* strategy is used. Both approaches perform significantly better with the MCS optimization enabled, independent of the average sector load. For *Frequency Reuse 1*, the performance benefit is about 2 users per sector. There is still a significant gain with MCS optimization, even for a load of 26 users. With SFR, one additional user per sector can be accepted which results from the lower inter-cell interference due to the MCS reduction in the side partitions. If there are free resources in some sectors due to the varying user concentration, the MCS optimization lowers the MCS of the mobiles within these cells. Thus, the inter-cell interference generated by these users is reduced.

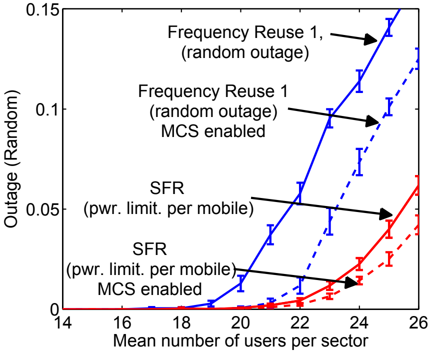
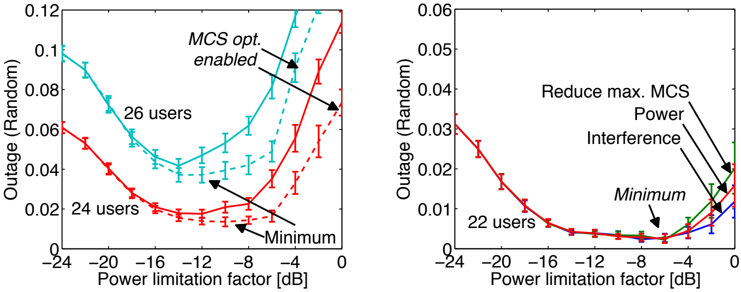


Figure 2.21: *Frequency Reuse 1 and SFR with and without MCS optimization enabled.*

Next, we investigate the impact of optimization on the power limitation factor of *SFR PowerLimitation*. Figure 2.22(a) shows the power factor at the x-axis and the random outage at the y-axis. Again, the *reduce maximum MCS* strategy is used. The upper curves reflect the parameter at a load of 26 users per sector on average. The other curve below shows the same for 24 users. The dashed curves show the behavior with MCS optimization enabled, the solid curves indicate the results with MCS optimization disabled respectively. The curves get smoother around the minimum with MCS optimization enabled. Thus, a less accurate adjustment of the factor can be tolerated. Furthermore, with increasing the average number of users, the minimum is shifting to higher power factors. Hence, the optimal setting for SFR, depends on the load in the sector. With MCS optimization enabled, the minimum is at -12 dB for 26 users and at -10 dB for 24 users. However, considering the confidence intervals, with MCS optimization it is sufficient to choose a parameter at -10 dB since there is no significant difference between the values.

In Figure 2.22(b), the different MCS optimization strategies *reduce maximum MCS*, *reduce maximum transmit power*, and *reduce maximum interference* are evaluated with the same configuration as in the figure before. There is no large impact on the performance of the MCS optimization if the power factor is chosen in a feasible manner. The same behavior can be found for 24 users and 26 users. Furthermore, the figure additionally shows the results related to Figure 2.22(a)



(a) Optimal parameter study at high load with and without MCS optimization.

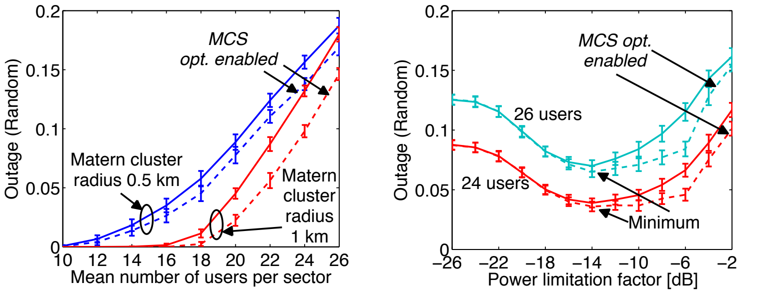
(b) Performance analysis of the three different implementations of the MCS optimization.

Figure 2.22: *Performance analysis of the MCS optimization algorithm.*

for 22 users. The minimum is at -6 dB. This again supports the outcome that the power factor is dependent on the user load. Similarly, the difference between the outage at -6 dB and -12 dB is only marginal.

In all previous evaluations, the users were scattered according to a spatial Poisson process in the network. In contrast, the performance evaluation is now investigated under more realistic conditions with a clustered spatial Matérn process. Clusters of fixed size t are generated. Each cluster is subsampled with a Poisson point distribution with a certain mean. The number of clusters is set to 20. The Matérn processes can be easily modified by changing the Matérn cluster radius.

The first investigation is the impact of the cluster radius on *Frequency Reuse 1*. Figure 2.23(a) shows the performance at either a cluster radius of 1 km or a radius of 0.5 km. The x-axis displays the overall average number of users per sector. The user concentration, however, now varies in each sector since the users are clustered around 20 specific cluster centers. The y-axis shows the outage as in the previous figures. As expected, with a decrease in the cluster radius, the outage probability increases since groups of users become spatially tighter. A sector may



(a) *Frequency Reuse 1* at cluster Matérn user location distribution.

(b) Parameter study at clustered Matérn user location distribution, Matérn radius 1 km.

Figure 2.23: *Performance analysis of the resource allocation schemes at different loads and different user location distributions.*

experience outage while other sectors are idle. Here again, the MCS optimization plotted in dashed lines improves the system performance.

Finally, SFR is investigated with a clustered user distribution in Figure 2.23(b). Depending on the radius of the cluster of users, the mean outage within the network changes. With a smaller cluster radius, i.e. with spatially tighter users concentrations, outage occurs much earlier than with a larger cluster radius. In both cases, the MCS optimization decreases the mean outage and allows for a more robust choice of the optimal power fraction parameter used in the power mask of SFR.

2.7 Lessons Learned

This section concludes the chapter and presents the most important findings. We have studied distributed resource allocation approaches under different settings. In particular, the focus was on the identification of approaches that are able to mitigate interference between cells.

First, research has been conducted to determine new opportunities to partition or restrict the different transmission resources. One of the objectives was to evaluate the different approaches with respect to the current traffic patterns, such as non-saturated traffic in the uplink. For this purpose, the uplink was examined which consists today typically of mainly TCP acknowledgements or short HTML request. Consequently, no full buffer scenario was investigated but a realistic uplink scenario with non-saturated traffic conditions. Furthermore, decentralized approaches were studied that do not require signaling and new combinations were compared with conventional schemes that do not use restriction or any resource partitioning such as *Frequency Reuse 1*.

It has been found that

1. an order metric (coordinated user allocation) is critical to the performance of the entire mobile communication system. The order metric specifies a sequence after that the users are assigned to the transmission resources.

This applies to both full frequency reuse as well as for FFR resource partitioning, *see Section 2.6.1 and Section 2.6.4.*

2. partial frequency reuse schemes such as SFR allow in conjunction with an order metric a performance gain for the resource allocation, *see Section 2.6.4.*
3. for SFR resource partitioning the kind of the limitation strategy for transmission resources significantly affects the performance of the resource allocation, *see Section 2.6.4.*
4. coordination with respect to user groups (aggregated coordination) is not more effective in comparison to the consideration of individual users. This is due to the fact that a limitation according to the aggregated transmit power of a group, allows a single user to use a significantly high transmit power, resulting in much interference, *see Section 2.6.4.*
5. for SFR applies: a coordinated user allocation with respect to inter-cell interference allows more performance gain than the differential treatment of users as specified in the different types of SFR, *see Section 2.6.4.*
6. even for the allocation at side-bands (frequency bands that in the first hand belong to other neighboring cells or sectors) the use of an order metric results in a performance gain. The effect is however less beneficial compared to the the use of a limitation strategy, *see Section 2.6.4.*
7. optimal limitation factor of SFR depends on the load in the network, *see Section 2.6.4.*

Second, the impact of the antenna configuration on *SFR*, *Frequency Reuse 1*, and *Frequency Reuse 3* was investigated. Advanced inter-cell interference mitigation techniques rely on a reasonable antenna configuration. The configuration has the purpose to provide, on the one hand, sufficient cell coverage and, on the other hand, good cell isolation to avoid inter-cell interference.

It has been found that

1. the antenna setting (downtilt) can significantly influence the performance of SFR and is dependent on the cell size and load, *see Section 2.6.3 and Section 2.6.5.*

2. due to path loss, multipath propagation, and antenna configuration, a differential treatment of users according to the geographic location within the cell is not useful for FFR. It is rather more beneficial to use the experienced interference of the users depending on the received signal strength at different positions in the cell, *see Section 2.6.3*.

The antenna configuration influences the propagation gain within the cell. Thus, with a clever configuration, the interference can be reduced. Consequently, the interference mitigation can be considered as less important in such a case. In Section 2.6.3, different antenna configurations were tested. For a fixed cell size, vertical downtilts are determined where the resource allocation works best.

The electrical downtilt reveals the best results for all resource allocation strategies and investigated downtilt angles. At full frequency reuse, the cell isolation becomes more important than the interference mitigation at a certain load. With SFR however better results are obtained with the electrical downtilt since the coverage is better utilized and interference is already mitigated by the SFR scheme and does not need to be done through the tilt of the antenna.

Third, a new resource allocation algorithm was proposed that is based on the intelligent choice of a modulation and coding scheme for a non-saturated uplink. Based on the previous results, a scheme was developed that a) is less load-dependent and consequently, does not need to be dynamically adapted, b) is interference reducing, and c) uses frequency partitioning.

It has been found that there is a non-linear relationship between transmission power and the selected modulation and coding scheme for the same amount of data. For example, a transmission with QPSK-1/2 modulation requires about 4.5 more REs compared to the transmission with QAM64-3/4. The accumulated transmission power with QPSK however is about 15 dB lower than with QAM64-3/4 at the same amount of transferred data. Choosing a low modulation scheme therefore represents a significant saving in transmission power, and as a result, a performance gain due to the smaller inter-cell interference.

The proposed algorithm takes advantage of this finding. If transmission resources are unused in a frame, it selects a more robust modulation and coding for the users to mitigate inter-cell interference. This is particularly interesting in the uplink with non-saturated traffic because, here, some resources remain unused.

In Section 2.6.6, it has been found that

1. lower modulation and coding schemes can be used to reduce the inter-cell interference.
2. if resources in a cell and the corresponding transmission frame are not used, the modulation and coding scheme can be lowered in order to fully utilize the resources. Although more resources are used, the inter-cell interference is reduced.
3. this approach is significantly less load dependent than pure SFR.

3 Application-Aware Resource Management

In the previous chapter, we have studied resource management at the physical layer for interference mitigation. In this chapter, we now deal with the application layer. We focus on *application-aware resource management* that integrates key performance indicators from the application layer within the network resource management.

3.1 Motivation and Objectives

In today's Internet, a large number of users consume an increasing variety of different applications and services. Each service has its own requirements on the communication network resulting from very different application capabilities. To address this heterogeneity in the network, one possibility is to consider the application layer within the resource management. The advantage of the consideration of the application layer is that the current applications, and thus the actual requirements of the applications on the network, can be taken into account. This enables efficient and tailor-made access networks, in which improvement is expected in the field of cost-efficiency of the network and perceived quality by the end user.

Compared to lower layer resource management, the goal is shifted away from the optimization of pure network parameters, towards the direct consideration of the application. This is based on the fundamental insight that users do not pay attention primarily on network parameters or settings, but on the application quality. The end user is typically not interested in technical parameters like the throughput of the network, as long as his video plays smoothly in acceptable resolution and image quality. The same applies to a user who surfs the Internet. A web browsing user is commonly not interested in the average throughput over time to his device, but in the first place, wants a fast website loading on his device.

The objective for the application-aware resource management is consequently to support the website loading or the smooth streaming in order to optimize the user-perceived quality through resource management actions in the network. Since this is not directly possible with only network resource management on lower layers, a different approach is utilized in this chapter.

Following cross-layer networking, corresponding ISO/OSI communication layers are linked together, which normally operate separately from each other. This means that for application-aware resource management QoE-relevant information about the applications (application layer) are available within the network (network layer) and can be involved within the resource management process. The challenge is to select the right information and define appropriate components to exploit it in an efficient manner. Many possibilities and concepts exist, exhibiting different degrees of complexity and benefits in terms of monitoring effort, usability of the information, and resulting improvement for the user.

The following example is intended to motivate the consideration of application information within the network resource management. In Figure 3.1, four home users are outlined, using a normal consumer Internet access. There is one shared link to the Internet. The users run different applications. The first user is watching a YouTube video. The second user talks via video chat with a friend. The third user is editing text documents with an online office thin-client application. The fourth user is downloading game updates from the Internet. All users are connected to a wireless router and are using the same link to the Internet.

Typically, the video chat user generates a constant bit rate traffic, because he transmits a consistent video image of himself and some constant voice data. The online office user generates only some minor data traffic if the document changes due to the thin-client application. YouTube on the other hand uses progressive

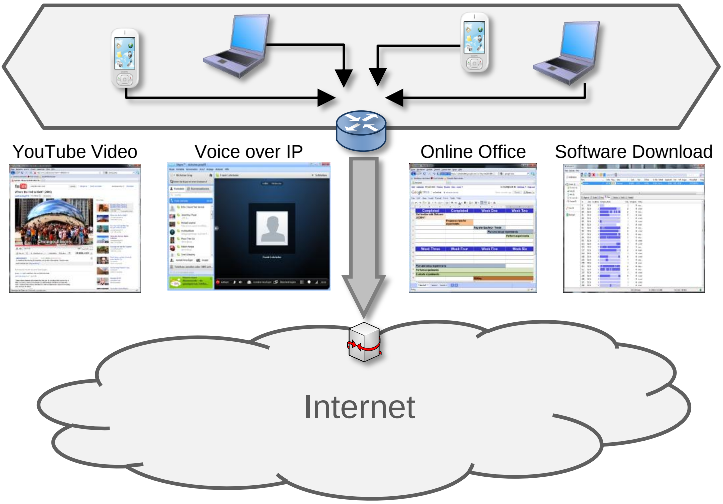


Figure 3.1: *Competing applications at a bottleneck link.*

streaming, which downloads the video data via HTTP. Similar to the download of game updates, this is a best-effort download. Consequently, we look closer at the situation for the YouTube user and the download user, who compete at this link against each other due to the best-effort nature of the generated traffic.

From a network point of view, both applications are using TCP transport layer protocol and equally share the link. In Figure 3.2, the bandwidth of the two observed users is shown. Depending on the link quality to the home access router, the users get a similar throughput to the Internet due to the TCP flow control. For the download user this may be fine. He expects a certain waiting time until the download is finished.

For YouTube, however, the situation is different. An important measure to quantify the perceived quality for a YouTube user is the buffer filling level of

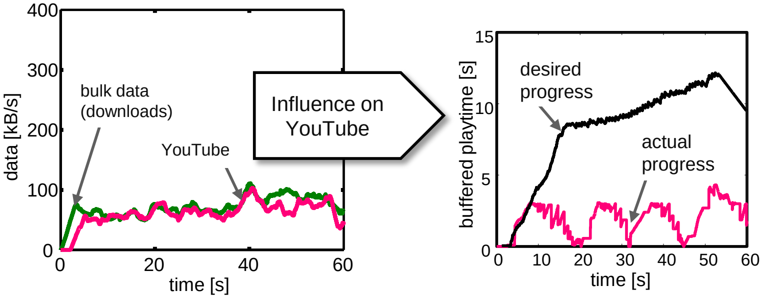


Figure 3.2: *Impact on application quality.*

the video player. As shown in Figure 3.2, the buffer level is low. It even decreases to zero seconds, so that the video playback may be interrupted if no lower quality of the video is available. It comes to the so-called *stalling*, the interruption of the video playback. In case of stalling, the user receives a small rotating circle at the video player. Stalling may have a severe impact on the user-perceived quality and results in abortion of the video playback by the user [87]. The desirable case from the user perspective is consequently a filled video playback buffer as shown in Figure 3.2, so that the video is smoothly played.

There are two possibilities of counteracting stalling. The application can be modified to take into account the network characteristics during playback, or the network properties may be varied to meet the requirements of the application. The former is commonly referred to as *network-aware application design* and includes all the adaptive approaches for applications such as adaptive and dynamic streaming. The second one is called *application-aware networking* which includes application-aware resource management as investigated in this chapter.

In the following, we discuss the concept of application-aware networking for access networks. We develop a resource management in order to allocate the resources according to the instantaneous application demands. Two major contributions to the research on resource management are given in this chapter. Firstly, an application-aware resource management framework is developed for access networks to directly address application needs within the network and optimize the performance of the network. Secondly, a monitoring approach for YouTube is designed that measures QoE relevant parameters.

The content of this chapter mainly refers to [3–6, 16–21, 24]. The remainder of this chapter is structured as follows. After this introduction, Section 3.2 presents background and related work about application- and user-centric resource management. The section is divided into three different subsections. First, mechanisms for traffic differentiation in networks are discussed in Section 3.2.1. Second, we describe in Section 3.2.2 corresponding metrics for the resource management. Third, this is followed by a brief introduction to the underlying theoretical concept *QoE modeling and management* in access networks in Section 3.2.3. In Section 3.3, we introduce the concept of application-aware networking. Subsequently, we develop application-aware monitoring approaches for YouTube video streaming in Section 3.4 and test them for videos in an access network. Finally, we conclude the chapter in Section 3.5.

3.2 Background and Related Work

This subsection provides a general overview on resource management for upper communication layers and summarizes the state of the art. The current state of the art consists of numerous approaches and concepts for traffic handling, which use information from different network layers and of different degrees of complexity in terms of their functionality, the information flow, and the number of involved entities. Thus, they achieve varying amounts of performance gains in comparison to the default best effort behavior in networks. This section gives an overview of related work and discusses the advantages and drawbacks of each approach.

A resource management for upper communication layers can usually be characterized by two different parts. First, the approaches differ according to their type of traffic differentiation mechanism in the network (Section 3.2.1). Secondly, there are differences in the use of information according to which the traffic differentiation changes (Section 3.2.2).

3.2.1 Traffic Differentiation in the Network

Each resource management approach performs a type of traffic differentiation in the network to enforce its goals. The traffic differentiation exploits appropriate information (see Section 3.2.2) and triggers management actions to improve the situation in the network. Therefore, commonly i) an *information/application-to-network interface*, and ii) a technology-dependent *enforcement mechanism* in the network is used as depicted in Figure 3.3. Each of the following approaches implements these two components in different ways.

QoS-based approaches [88–92] implement the application-network interface by defining classes according to network-level parameters such as minimum latency or throughput for the expected type of traffic. Involved networking devices enforce these QoS definitions by implementing queues within the forwarding plane. Applications with similar needs are assigned to similar queues. Typical representatives are implementations of IETF DiffServ [91] and IntServ [93] as well as mobile communication networks, in which so-called *QoS profiles* are defined for differentiated traffic handling [94–96]. The advantage of such approaches is the ease of implementation in a networking device due to the network-related specifications. The disadvantage is however twofold. First, the mapping of applications to QoS classes is often complex because of the heterogeneity of the applications. Applications often provide several different functions between which the user may select. Consequently, the requirements of an application can change according to the state of the application. The second disadvantage is that QoS classes normally meet only a set of fixed network parameters. For dynamic requirements of the applications, however, the QoS classes must be adjusted if at all possible. This is especially the case for applications with time-dynamic QoS requirements. For example, due to video encoding, download patterns, or user behavior, an application may not have a fixed demand for bandwidth.

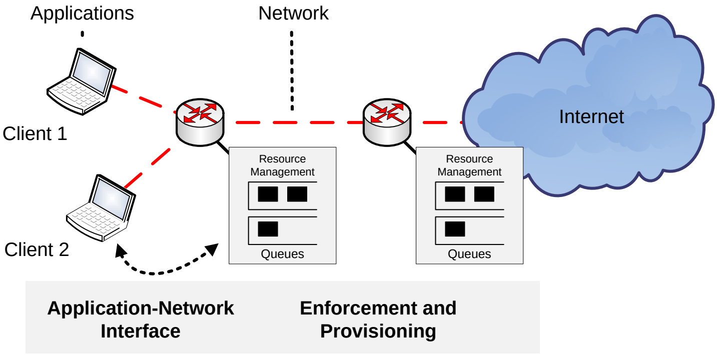


Figure 3.3: *Structure of a state-of-the-art traffic differentiation architecture with application-network interface and traffic enforcement mechanism.*

In [97], an application network-interface is proposed for OpenFlow-enabled networks. The interface allows application service providers to define application-specific requirements that are implemented by the network. The OpenFlow protocol is a standardized implementation of the principles of Software-Defined Networks (SDN) [98]. To achieve a network-wide enforcement, Application Label Switching (APLS) is introduced that adds additional protocol headers (*labels*) to IP flows that can be utilized by OpenFlow SDN switches in order to differentiate traffic. The advantage is the standardized implementation at network level. Furthermore, the approach delegates the specification of the required resources to the application provider. The issue with such an approach is that the application provider usually has no global view on the network and so the overall network performance may decrease compared to the optimal one since the application provider can only partly consider other parallel applications in the network. Without the global view of the network and the knowledge about utilization and application demands, it is difficult for individual applications to make effective and fair decisions.

Similar to this approach, a comprehensive additional protocol layer is proposed for networks in [99]. This Service Access Layer (SAL) facilitates a split between service level control and the application data plane, thus enabling service-aware network routing. For the realization the proposed protocol must be implemented of each network device.

In [13], the authors introduce a central entity that has knowledge about the network and application situation. It performs a dynamic resource allocation decision that utilizes SDN in order to enforce application demands in the network. An implementation featuring a video on demand application competing with a file download at a bottleneck link shows that monitoring the video on demand application's buffer state can help improving the end users' QoE. The work on participatory networking presented in [100] goes one step further and proposes a communication mechanism between applications and the network which is initiated by the applications. Key concepts include conflict resolution between the needs of different applications or users and decomposition of network control, i.e. limiting an application's authority. The authors of [101] introduce an OpenFlow based framework that aims at achieving fairness with respect to QoE among all users in an adaptive video streaming context. The expected QoE of a user is estimated on the quality information during the video streaming. This approach combines a network-aware application with a network based control entity. Based on monitoring results gathered via intercepting Media Presentation Description (MPD) files used in MPEG-DASH, the framework dictates video quality levels for each user in order to maximize the resulting QoE for every participant. Despite the achieved fairness, open issues here are how universal this approach is and how it can be applied to other applications. A general framework is usually required for including QoE information of all kinds of application for a multi-application networking scenario.

A system design featuring functional blocks that represent functions for packet handling in the network is presented in [102]. Feedback loops including a central decision unit allow for combining network-aware applications with an application-aware network. A central idea consists of translating user demands into application demands which in turn can be translated into network requirements. The system's capabilities are demonstrated and verified for various use cases including live video and video on demand scenarios.

Furthermore, there are other works that mainly address the application-network interface. In [103], the authors introduce Atlas, a machine learning based approach for classifying observed traffic. This technique not only allows reliably determining the application class but also the specific application. Combining this classification mechanism with application requirements derived from various models, this can result in efficient control solutions. In contrast to previously introduced architectures, the authors of [104] propose a system where users give explicit feedback on their QoE and assign priorities to running applications. This feedback is used in order to change the distribution of the available bandwidth among the competing applications.

A protocol realizing the exchange of information between network and application is standardized by the IETF *Application Layer Traffic Optimization* (ALTO) working group [105]. It aims at providing guidance to content delivery applications such as P2P or CDNs which have to select one or several hosts or endpoints out of a set of possible candidates. In such a way, appropriate candidates can be selected and the performance with respect to user-centric, network-centric, and application-centric metrics can be improved. Currently, several extensions of the ALTO protocol including data centers and cloud applications are discussed [106].

More work on application-aware resource management has been conducted in the context of data-center applications. Research on data center architectures already shows that combined solutions, e.g. for traditionally separated mechanisms such as routing and service migration, may increase data center efficiency. However, data center applications are no end-user applications as in access networks, but run in isolated and manageable server farms. In general, the constrained environment of data centers enables the combination of these traditionally separated mechanisms. In the work of [107], the authors propose an algorithm that integrates such a combined solution. In detail, the algorithm simultaneously combines the dynamic migration of virtual machines and the traffic routing in case of environments with changing demands. The authors demonstrate that their approach is effective, scalable, and cost-efficient. Data center application-aware solutions are also provided in [108–111].

In summary, there are many approaches in literature that propose user- or application-centric resource management. Known challenges are:

- **Flexible enforcement** Due to heterogeneous applications in a network with time-dynamic requirements, the enforcement must be able to dynamically respond to changes in the network and at applications, independent of the network technology.
- **Support for multiple applications** The monitoring or the application-network interface must be universally defined and applicable for multiple applications.
- **Resource Efficiency and Fairness** In order to increase the resource efficiency or for fairness reasons, a resource management mechanism requires a global view on the network and the application demands.

3.2.2 Information Metrics for Resource Management

The performance of the resource management is highly dependent on the utilized information. The more comprehensive the information is, the more complex is the monitoring of it. In the following, we briefly summarize proposed information metrics for resource management and traffic differentiation in the network.

The traditional solution is to perform traffic differentiation according to network-layer parameters [88–92]. However, a QoS-based provisioning alone is often not sufficient to provide an acceptable application quality due to download patterns, user behaviour, or time-dynamic requirements of applications. Instead, network resources depending on the current application state are required.

The works in [112–114] carry out a cross-layer optimization in order to integrate more appropriate information than QoS requirements within the network

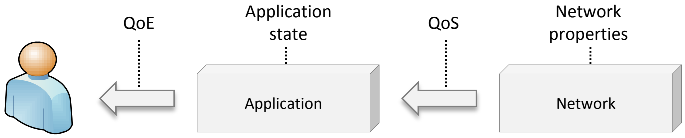


Figure 3.4: *Relationship between QoE, application state, and QoS.*

resource management. In [112,113], an application-specific solution is proposed. Important packets in MPEG videos, mainly key video frames (such as I-frames), are prioritized in case of limited transmission resources to optimize the perceived quality. In [115], a multi-layer video encoding with scalable video codec is used, which adapts according to the available network bandwidth. The goal is to control the different layers in different QoS classes with different priorities to specifically drop video layers with less importance if the network is congested.

Another, yet more extensive approach is to integrate QoE information, or QoE-related parameters. In [116], QoE-based scheduling for wireless mesh networks is proposed. The authors take into account not only video and audio streaming but also data traffic using simple MOS metrics which map QoE to simple QoS parameters. The work about QoE is commonly summarized under the term *QoE Management* and can be divided into two distinctive research fields. There are papers about understanding and modeling QoE for different applications [117, 118] and papers that make use of the QoE as resource management metric [101, 112–114,116,119]. More details are described in the next section. Commonly, the relationship between QoE and QoS network parameters is not of linear scale, i.e. altering the QoS parameters results in different QoE levels [118]. Consequently, there are resource management algorithms which define an acceptable end user quality at minimal resource utilization as control objective [117].

To extend this concept to general applications, other approaches have been developed. It was found during the modeling of QoE for applications that application parameters can reflect the QoE to a high degree [120–123]. Consequently, application-aware management approaches use cross-layer information from the running application to adjust the control decisions in the network [6, 124]. In [124], QoE metrics are used to differentiate between different services. A one-dimensional optimization function is generated for the radio resource management. This approach can be seen as a continuation of former work which used a so-called utility function with several different metrics for resource management [125–129].

In [114], a cross-layer method for web applications in an LTE mobile network is proposed which takes into account the QoE and service response times. The authors map user data rates and service response time to QoE MOS values for web traffic. The findings are used in a proposed LTE radio resource management which uses a utility function of the MOS values.

In summary, there are several approaches that use higher-layer information in network resource management. They differ in the type of information that is used for resource management. Figure 3.4 shows the relationship between the most commonly used information metrics in literature: *QoE*, *application state*, and *QoS*.

3.2.3 QoE Modeling and Management

QoE is one of the most important metrics for resource management [130]. In the following, we elaborate on the research steps necessary for the use of QoE in resource management. From a conceptual perspective, three steps are required: (1) modeling QoE, (2) monitoring QoE, (3) optimizing QoE.

Modeling QoE The fundamental step to integrate QoE within the resource management is understanding the heterogeneous application requirements and the impact of network disturbances on the user-perceived quality. This is done by modeling the user QoE for a specific application at different network conditions.

QoE modeling consists of two parts that are (1) the user perception quantifying the user-perceived quality and (2) application layer measurements providing application patterns depending on the actual network conditions [130]. The latter can be used to model application layer QoS, i.e., how often the service is interrupted and how long. However, in order to quantify the user satisfaction with a service, a user perception model has to be derived by means of subjective user studies. Thereby, the application layer measurements serve as input such as [131].

In [121–123, 132–134], models for the user QoE for specific applications are formulated. With respect to QoE management, these works specify possibilities for resource management by defining QoE relevant parameters for each application.

Monitoring QoE As a result of the QoE modeling process, QoE-relevant parameters are identified which have to be monitored accordingly. In general, monitoring includes the collection of information such as (i) the network environment (e.g., fixed or wireless); (ii) the network conditions (e.g., available bandwidth, packet loss); (iii) terminal capabilities (e.g., CPU power, display resolution); (iv) service and application-specific information (e.g., video bitrate, encoding, content genre).

The QoE monitoring can either be performed a) at the end user or terminal level, b) within the network, or c) a combination thereof. While the monitoring within the network can be done by the provider for fast reaction on degrading QoE, it requires mapping functions between network QoS and QoE. When taking into account application-specific parameters additional infrastructure like deep packet inspection (DPI) is required to derive and estimate these parameters within the network. Alternatively, monitoring at the end user level gives the best view on user-perceived quality. However, additional challenges arise, e.g., how to feed QoE information back to the network for resource management. In addition, trust and integrity issues are critical as users may cheat to get better performance.

In Section 3.4, the YoMo tool that monitors the buffer status of the YouTube video player is defined. This surveillance on the application performance allows for the prediction of the QoE and a timely notification for an QoE optimization mechanism. Both approaches, network-based and client-based, are proposed and evaluated in comparison to each other.

Optimizing QoE The final step is the actual resource management. The resource management should perform a dynamic adaptation and control of resources to deliver optimal QoE so that the user may not get dissatisfied or abandons the service. QoE control aims at reacting before the user encounters problems and uses monitoring information to adjust corresponding impact factors. QoE resource management addresses the following questions, (a) where to react, i.e., at the edge, within the network, or both; (b) when to react and how often; and (c) how to react and where which control knobs to adjust.

In this chapter, we follow the principles of QoE management for access networks. On the basis of QoE models in literature, we define hereinafter a complete application-aware resource management framework for access networks.

3.3 Application-Aware Resource Management

In this subsection, the concept of *application-aware resource management* is discussed. In particular, a corresponding framework is defined and explained in detail. We describe the general concepts, whereas the implementation is given in Chapter 4 for different types of access networks. This section demonstrates how the framework can be used to implement a traffic and resource management for an efficient and application-aware delivery of traffic in an access network.

The key idea is to collect information about network and application status at a central point in the network, in the following called *decision manager*. It is able to trigger a number of optimization actions that either change the traffic handling in the network or the traffic produced by the application. The first type of actions is summarized under the term *network control actions*, the second type of actions is referred to as *service control actions*.

The general goal of the framework is to improve the overall QoE in the network or rather to avoid QoE degradation. This framework follows a reactive approach for resource and traffic management. Resource management actions are only triggered if (a) a QoE degradation or an indication for an imminent QoE degradation has been detected and (b) a resource management action is available that will avoid or limit the QoE degradation without overly harming the QoE of other users.

The framework does not follow a proactive approach to optimize a QoE-based metric and it is also not targeting at an optimization of network parameters such as a balanced link utilization etc. As a consequence, it is intended to run in addition to a ”traditional“ traffic or resource management mechanism that does not take into account application layer performance but relies on typical network performance indicators such as load, available bandwidth, delay, packet loss, etc. and traffic classification. The idea is not to interfere with other mechanisms as long as the application layer performance and QoE is good enough. Only if a QoE degradation occurs in spite of these traffic and resource management mechanisms, the framework will trigger actions in order to avoid a QoE degradation.

Figure 3.5 shows the different resource management components of the framework. This is, first of all, the *decision manager* that receives information from the application and network monitors. It is additionally connected to the nodes in the network that actually enforce resource or traffic management decisions. When notified by an application monitor about a critical state of an application, the *decision manager* evaluates its set of resource management actions. This means that, based on the information on application and network status, it predicts how a resource management action changes the network status and estimates whether the QoE situation improves. From all resource management actions with potentially positive outcome the one to be executed is selected based on (a) the confidence in the prediction, (b) the degree of the QoE improvement, or (c) the effort for performing the resource management action.

Another key component is the *application monitor* that is running on the client or in the network. It sends the application status to the *decision manager*. This

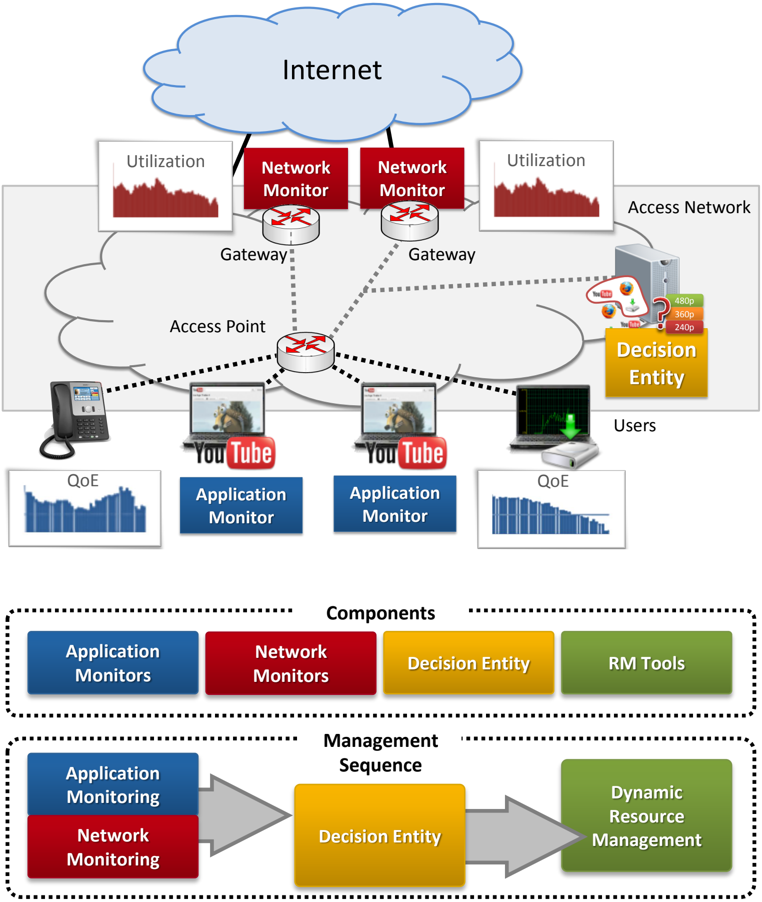


Figure 3.5: *Overview of the individual components and placement in the network.*

communication can be either event-driven or periodic, and either push- or pull-based. The task of the monitor is to keep track of the status of traffic intensive or QoE sensitive applications that are subject to the resource management decisions. The status of an application is a collection of QoE indicators that the customer will directly perceive as quality parameters. These indicators are application-specific and describe whether the current performance offered by the network leads to a QoE degradation.

The indicators cannot be directly mapped to a QoE value, as QoE describes the overall experience of a user with a certain service. Therefore, it also depends on many other factors such as non-measurable subjective user demands. However, the performance indicators indicate if a QoE degradation is imminent. Using these indicators, the framework is able to indirectly consider the QoE of applications within the resource management and avoid degradations.

To give an example, key performance indicators for RTP streaming are bandwidth on application layer, packet loss, or jitter that may be mapped to a QoE metric by using a QoE model. Key performance indicators for HTTP streaming services are the bandwidth on application layer or the buffered playtime in the client. When the buffered playtime is low and the bandwidth is below the video rate a period of stalling will probably occur if no measures are taken. According to [120, 135], stalling is the factor dominating the QoE for video clips clearly exceeding the significance of video resolution as a second impact factor. In the following, we focus on the buffered playtime and define a monitoring for YouTube video streaming in Section 3.4.

The third major component of the framework is the *network and flow monitor*. It monitors typical network parameters like load, packet loss, buffer status of the network interface, number of connections, etc. and sends these parameters to the *decision manager*. Additionally, it is able to monitor a certain flow in the network to observe its current throughput and state. Again, the communication may be pull- or push-based and periodic or event-driven.

All the mentioned components together form the *Application- and QoE-Aware Resource Management Framework*. Resource management algorithms and global decision managers have been studied sufficiently in the literature. In the following we define and evaluate an application monitoring for YouTube.

3.4 Implementation of an Application Monitoring for YouTube

The performance of application-aware resource management is highly dependent on the utilized information. Therefore, an essential component for the application-aware resource management is the application monitoring. In the following, we develop an approach for YouTube to estimate the client's buffer filling level within the network.

It holds according to [120, 135] that stalling is the dominating factor for a QoE degradation and clearly exceeds the significance of video resolution as a second impact factor. Accordingly, it must be monitored as part of a QoE estimation. Stalling occurs exactly if the video playback buffer is empty and if there is no more playback data available. Consequently, we propose in this section to monitor the YouTube video playback buffer in order to predict stalling. The buffer level corresponds exactly to the time until a stalling is going to happen if no more data arrive in the worst case.

To be able to monitor the YouTube buffer filling level, the YouTube Monitor (YoMo) has to fulfill several tasks: Firstly, it has to detect the YouTube flow. Secondly, it has to collect several additional information such as playback quality about the YouTube flow and thirdly, it has to monitor the YouTube buffer. To make our approach easier to understand, we first analyze the video streaming format of YouTube in Section 3.4.1 before we introduce the main ideas of our YoMo and their implementation in Section 3.4.2. The estimation of the amount of buffered playtime is a key concept of our approach described in detail in Section 3.4.3.

3.4.1 Description of the Streaming Technology of YouTube

For streaming video, YouTube either uses a proprietary Flash application or the built-in HTML5 video player of a browser is used. In case of the Flash player, it concurrently plays a Flash or MPEG-4 Part 14 (MP4) video file and downloads it via HTTP fragmented in multiple *video chunks*. The same applies to the HTML5-based version that can play various video formats depending on the browser version. Currently, the standard video container format is MP4.

At the beginning of such a so-called pseudo streaming, the client fills an internal buffer and starts the playback of the video as soon as a minimum buffer level, γ, is reached. During the time of simultaneous playback and downloading of chunks, the buffer level depends on download bandwidth and video rate. As long as the download bandwidth is higher than the video rate, the buffer increases, otherwise it shrinks. If the buffer runs empty, the video stalls. A stalling can be detected by a change of the YouTube player state from “playing” to “buffering”. The player state is hidden to the normal user, but can be retrieved from the YouTube API by JavaScript or ActionScript. Furthermore, the API allows to control the video playback and to get information about the currently displayed video.

A YouTube video is at present mostly encoded as an FLV or MP4 file which are both container formats for media files. FLV is developed by Adobe Systems. MP4 is standardized by the International Organization for Standardization (ISO) within ISO/IEC 14496-12:2004 which is based on the Adobe QuickTime file format. Both video container formats encapsulate synchronized audio and video streams, and are divided into a header and data part, in the following called *video fragments*. The header starts with a signature and contains information about the video fragments in the body of the file. The fragments encapsulate the data from the streams and contain information on their payload. This information includes the payload type, the length of the payload, and the time to which the fragment payload applies. FLV files may also contain metadata encapsulated in a fragment with a script data payload. The available properties depend on the software used for the encoding and may include the duration of the video, the audio and video rate, and the file size.

3.4.2 Main Functionality

The YouTube player establishes a new TCP connection at the beginning of the video playback or if the user jumps to another time in the video that is not prebuffered. YoMo runs at the client or at a gateway in the network and parses all uplink TCP flows in order to detect the video identifier which marks the beginning of a new YouTube video flow. Once a flow containing video data is recognized, the data is continuously parsed in order to retrieve the available meta information from the video file. The buffer monitoring task is more complex and will be explained in the following.

The YouTube buffer filling level is defined by the amount of playtime β, the player can continue playing if the connection to the server is interrupted. Figure 3.6 shows β as the difference between the currently available playtime T and the current time of the video t. YoMo constantly computes and visualizes β in a GUI and checks whether β falls below an alarm threshold β_a. In such a situation, like the one depicted in Figure 3.6, the instantaneous QoE is good, as the video is playing, but the application state is bad, as $\beta < \beta_a$ and the video is about to stall soon. Hence, YoMo has to notify the network resource management or decrease the video bandwidth.

3.4.3 Estimation of the Buffered Video Playtime

YoMo computes the buffered playtime as $\beta = T - t$. It decodes the video fragments in real time, and hence exactly knows the currently downloaded playtime T which is the time stamp of the last completely downloaded video fragment. This can be done in the network or at the client. Intuitively, t could easily be calculated as the time difference between the actual time and the time when the player starts to play the video. However, the playback of a YouTube video does not start immediately after the player has loaded, but only after an amount γ of bytes has been downloaded. Experiments with different videos and different connection speeds showed that γ is varying between 50 and 300 kB and is independent of the connection speed. In contrast, it is different for each of the considered videos.

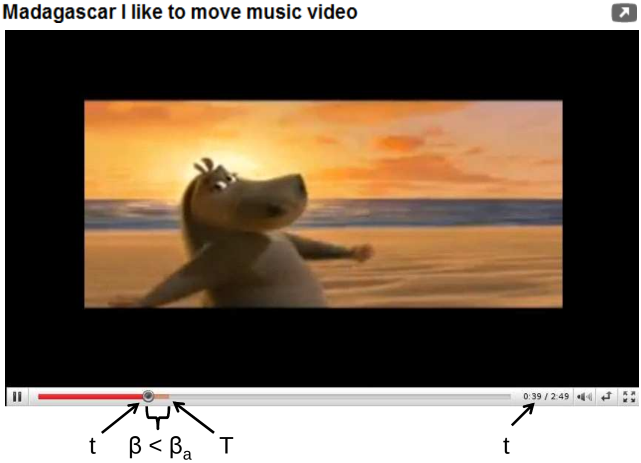


Figure 3.6: *The YoMo Parameters*

We analyzed the coefficient of correlation between γ and different video characteristics including information about the frame types of the original H.264 file embedded in the video fragments, but were not able to find a clear correlation which allows to derive γ from the properties of the displayed video.

It is hence not possible to calculate the amount of time which lies between the time when the user issues the request for the video and the time when the video actually starts to play. We therefore implemented two different methods for calculating t which we discuss in the following.

Network-Based Method This method uses the assumption that the video starts to play as soon as the first video fragment is completely downloaded. Clearly, this introduces a small error in the calculation of β which decreases however with an increasing connection speed and/or proximity to the client. This is the favorite method when using YoMo at networking elements since the download of the first video fragment can be detected within the network.

Client-Based Method The second method stands for the way of obtaining t from the YouTube player API which can be accessed at the client by scripting languages only. Hence, an additional monitoring software has to be installed at the client. YoMo uses a Mozilla Firefox extension which runs a JavaScript that retrieves t from the YouTube player. The extension additionally provides the actual value of t to YoMo.

3.4.4 Accuracy of the Monitoring Approach

In the remainder of this section, we investigate how exactly YoMo can indicate the prospective QoE. In order to find out how accurately YoMo predicts the time when a video stalls, we use a client running a measurement web page which allows to dump the YouTube player state. The client is connected to the Internet via a proxy which is able to modify the connection speed and to interrupt the connection and thereby cause the video to stall. In addition, YoMo is enabled either on the proxy or at the client dependent on the investigated method. The buffer estimation and the corresponding timestamps logged. During our experiments with 100 randomly chosen videos we observed that $\beta = 0\,\mathrm{s}$ is a sufficient but not a necessary condition for a stalling video: many videos already stall if $\beta \approx 0.5\,\mathrm{s}$. We therefore consider a video to stall as soon as $\beta \leq 0.5\,\mathrm{s}$.

In Figure 3.7 and Figure 3.8, we depict the estimation error Δt_s between the time when YoMo considers the video to stall and the video actually stalls, i.e. when the player state changes to buffering for the first and the second method respectively. For each considered bandwidth, the box depicts the inter quartile range of the estimation errors and whiskers which are 1.5 times longer than the interquartile range. Values beyond this range are shown by red crosses. Figure 3.7 presents the estimation accuracy of the *network-based method*, first. It shows that the estimation error decreases as the bandwidth increases. This is just a logical consequence of neglecting the time required for downloading γ, which gets smaller if the Internet connection is fast. While this method is thus sufficiently accurate for fast connections, it results in YoMo estimating the video to stall up

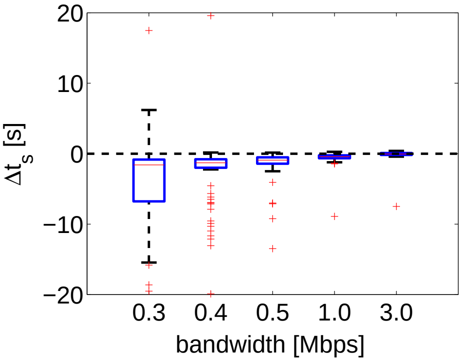


Figure 3.7: *Stall Time Estimation Error, Network-Based Method*

to 20 seconds earlier as it actually did if the connection is slow.

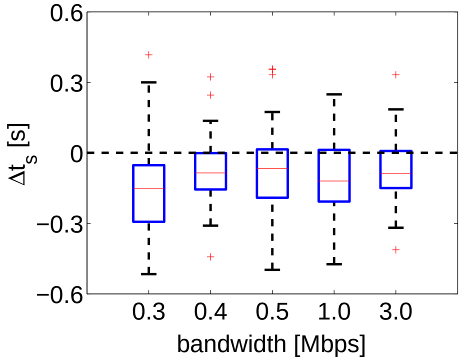


Figure 3.8: *Stall Time Estimation Error, Client-Based Method*

The results for the experiment with the *client-based method*, in contrast, show an error which is independent from the bandwidth. Moreover, YoMo estimates the stalling of the video on average roughly 0.1 s earlier than it actually happened, which is a significant improvement over the *network-based method*. In most cases, YoMo underestimates the remaining playtime, i.e. predicts the time of stalling earlier than it actually happened. The maximal estimation error in this direction is 0.5 s. In some cases, YoMo overestimated the remaining playtime with a maximal error below 0.5 s. Taking the inherent error of our assumption that a video stalls if $\beta < 0.5$ into account, these results demonstrate that YoMo, with the *client-based method* for the buffer estimation, is working as intended.

Based on these results, we use exclusively in the further part of this work YoMo with the *client-based method*, which gives us more accurate results.

3.5 Lessons Learned

This section concludes the chapter and summarizes the lessons learned. We introduced application-aware resource management to enable efficient and tailor-made access networks. The focus was on the description of the approach as well as on definition of the necessary components. Specific use cases are defined and evaluated in the following chapter for different types of access networks. Furthermore, the evaluation and implementation of an appropriate application monitoring was presented in this chapter. Application monitoring is not sufficiently studied in the current literature.

It has been shown that

1. network delivery problems may have a high negative impact on the QoE of Internet applications.
2. QoE modeling specifies and defines resource management options.

The proposed resource management approach in this chapter is *application-aware resource management*. It uses information about the status of an application (e.g. active/idle, displaying a video, application buffer level, chatting, displaying information, calculating, processing information, interaction with user), and integrates this information in the network resource management. Based on appropriate literature, application information reflect the QoE of the user to a high degree if properly selected.

It has been found that

1. the integration of application information may help to decide about efficient resource management actions
2. resource management with respect to application state requires dynamic monitoring of the application.
3. other necessary components for application-aware resource management are: resource management actions, a network monitoring, and a decision entity that receives all the information and triggers the management actions.

The performance of application-aware resource management is highly dependent on the utilized information. For this purpose, a dynamic monitoring was developed and evaluated on the example of YouTube video streaming. It consists of an optional browser plug-in that monitors the state, in particular the current playtime, of the Flash player and a packet sniffer that detects new Flash video transfers, extracts the videos metadata, and monitors the available playtime. There is a trade-off between complexity and accuracy. If the optional plugin is used, a program at the client-side (client-based monitoring) is necessary. If the plugin is omitted, the approach can be used within the access network on a networking element that forwards data to the Internet. However, if both components are used together, the approach allows for determining exactly the buffered playtime.

The buffered playtime directly reflects the occurrence of video stalling. According to [120, 135], stalling is the factor dominating the QoE for video clips clearly exceeding the significance of video resolution as a second impact factor.

It has been found that

1. the monitoring is able to accurately estimate the time when the YouTube player is stalling.

2. the monitoring is lightweight and easy to install while it provides valuable information to a network operator. If it is enabled, both parties may greatly benefit as the operator gets information for free which he can use for improving the user QoE.

4 Performance Assessment of Application-Aware Resource Management for Cellular and Wireless Mesh Networks

In this chapter of the thesis, use cases for application-aware resource management are carried out for different types of access networks. The objective is to evaluate the impact on user and network.

Internet access networks, such as cellular or wireless mesh access networks, are often the bottleneck of today's communication networks and consequently most strongly responsible for determining the user satisfaction. The limited bandwidth and the fact that access networks are most often used as mere bit pipes are unfavorable for the users' QoE. The lack of application-specific service guarantees is especially inadequate in the face of an increasing degree of heterogeneity of Internet applications and their individual service requirements. The implementations of application-aware resource management address this challenge by the interaction of monitoring tools, central decision manager, and resource management action.

In this chapter, wireless mesh and cellular access networks are investigated. Wireless mesh networks are based on a random channel access scheme for transmission. Giving strict QoS guarantees is thus difficult and complex, and does not facilitate the development of certain resource management approaches. It requires additional tools and mechanisms that work against the random access. We discuss in the following three new resource management mechanisms for mesh networks.

For cellular networks, in contrast, excellent opportunities exist to guarantee a certain bandwidth for a user. The scheduling at the base station allocates explicitly users the resources at certain times, on which they may send. However, the lack of studies and approaches lead to a disregard of such mechanisms for certain services and applications.

The research contribution of this chapter is a definition of different resource management approaches for these access networks. The benefits and costs are evaluated with respect to different applications.

The content of this chapter is taken from [3, 18, 22, 23].The remainder of this chapter is structured as follows. In Section 4.1, we give an overview of publications related to our work. In Section 4.2, application-aware resource management for YouTube video streaming in wireless mesh networks is investigated. We propose several approaches to improve the streaming of YouTube. In the second part, we propose application-aware resource management for cellular networks in Section 4.3. We influence the packet scheduling at the base station according to application needs. Finally, we conclude the chapter and summarize lessons learned in Section 4.4.

4.1 Related Work

In this section, a general overview is provided of the most relevant works from literature that propose application-aware resource management concepts for cellular and wireless mesh networks. The conceptual ideas, benefits, and drawbacks are outlined. This section is divided into two parts, considering approaches related to wireless mesh (Section 4.1.1) and cellular networks (Section 4.1.2).

4.1.1 Wireless Mesh Networks

Related work with respect to application-aware resource management for wireless mesh networks cover published radio resource management techniques and frameworks. Of particular interest are approaches for such networks deriving the QoE from network or application layer parameters and using them for resource management.

Standardization In the IEEE, standards are developed towards an improved usage of radio resources in heterogeneous wireless access network. A full-featured management system allowing the distributed optimization of radio resource usage and improving the QoS in heterogeneous wireless networks is defined by IEEE 1900.4 [136]. Further on, the IEEE 802.21 has the goal of specifying standards for media independent handover in heterogeneous radio access networks [137]. Both approaches have in common that they aim at optimizing resource usage in heterogeneous wireless networks by utilizing different kinds of information such as terminal capabilities, radio or network capabilities of devices, or device measurements. Therefore, a signaling framework between terminals and the network is established in order to enable context-aware resource management decisions.

QoS Resource Management Frameworks A similar idea is presented in [138] where the authors introduce a cooperative radio resource management framework for enabling seamless multimedia service delivery for wireless access networks. The framework enables a terminal to take a handover decision based on the QoS information broadcasted by the access networks. A simulation study demonstrates that the QoS broadcasting mechanism may be implemented within the IEEE 802.11 beacon frames. A comparable but more general framework has been presented in [139], which proposes an architecture for decentralized network management and control. Their main contribution is the idea of a piloting system which uses a control plane to measure different system parameters and provides them to control algorithms like routing or mobility management. In a simulation study of a mobile-initiated handover, this framework, taking into account the load of the APs, results in less rejected VoIP calls and a lower end-to-end delay than the traditional policy of considering the received signal strength only.

Wireless Mesh Networks Especially for wireless mesh networks many other possibilities for resource management exist. In [140], an overview on the existing alternatives is given, we therefore refrain from an exhaustive enumeration. In the work [141], the authors propose to dynamically constrain the bandwidth of best-effort traffic in order to ensure the quality of service requirements of multimedia applications. This is realized by the interaction of two tools, the traffic observer (TO) and the traffic controller (TC) running at each mesh node. The TO continuously monitors the QoE of the VoIP flows, which is calculated as Mean Opinion Score (MOS) from the measured packet loss. As soon as the MOS falls below a threshold, signaling messages are sent via the OLSR Hello message system to assure that the TC instances running on the same and on the neighboring nodes throttle the interfering best-effort traffic. The evaluation of this concept in a mesh testbed shows that TO and TC together allow to maintain a satisfying MOS score even in the presence of disturbing traffic.

QoE Resource Management The exponential relation between packet loss and MOS used in [141] has been discovered by Hoßfeld et al. [142] and is one example of mapping measurable QoS parameters to user experience. Many contributions similar to this work exist for wireless mesh network, see e.g. [141, 143, 144]. In [145], the QoE model for speech quality proposed by Raake [146] is modified in order to obtain a QoE estimation for an aggregate of VoIP calls. The resulting QoE based admission control scheme successfully avoids long periods of user dissatisfaction in a wireless IEEE 802.16 access network. Another example is [113] who use the peak signal to noise ratio for deriving the average QoE for video streams. The authors present a resource management scheme that optimizes video source rate, time slot allocation and modulation scheme in order to maximize the average video stream QoE.

4.1.2 Cellular Networks

There are different approaches that either perform application-related scheduling at the base station of a mobile network or that address the user and his applications directly. In the following, we discuss QoS scheduling, operating on QoS criteria at the network level, and QoE scheduling that tries to take into account the quality experienced by the user.

QoS Scheduling QoS requirements of an application are primarily hard network characteristics like maximum latency, minimum traffic rate, maximum delay, etc. These requirements are considered in so-called service classes as defined in the IEEE 802.16 standard or in QoS profiles of LTE bearers [147]. The packet scheduler tries to satisfy the criteria specified in the classes by defining an intelligent order for the packets. With this differentiation, groups of network flows with similar needs are prioritized equally and packet scheduling of different groups is done according to the QoS definitions. If the requirements are known and are met by the scheduler, a high user satisfaction can be guaranteed. These parameters however are very strict criteria and do not indicate how the quality, which is perceived of the user, changes if some or all of the requirements are violated. Hence, if there is congestion in the network and the requirements can no longer be guaranteed, the influence on the user cannot be predicted.

QoE-Aware Utility-Based Scheduling While it is important to consider network-level QoS parameters, it is even more important to take the perceived quality at the end user into account. The idea of utility-based scheduling [124, 148, 149] is based on the fact that different QoS parameters have a different effect on the QoE of an application. Thus, the QoS parameters are weighted and a utility function is defined based on them. The aim is to maximize the utility in order to optimize the QoE.

Cross-Layer Approaches and QoE Scheduling For cross-layer scheduling directly the QoE that the end user experiences is considered. Therefore, QoE models are commonly created of applications to identify quality indicators for the QoE [117, 118]. For video streaming, for example, the current resolution, frame rate, and the encoding can be seen as key quality indicators that affect the QoE. The scheduling is consequently performed according to the quality indicators. If different quality factors are weighted in a utility function, this can be regarded as an extension of the utility scheduling for different applications, in order to maximize the overall QoE. The problem in this case is the measurement of the quality indicators. Since they are not necessarily network parameters as in QoS scheduling, the measurement can be complicated and may require additional software as the proposed one in Section 3.4.

Knowledge about the QoE of an application within the scheduling process at a base station however provides more flexibility and also describes how a gradual violation of the QoS parameters affect the QoE. In [145, 150], the continuously monitored QoE of voice connections is considered for scheduling and admission control decisions for IEEE 802.16 networks. In [114], a mapping for web traffic is presented to allow for a direct consideration of web browsing within the scheduling.

4.2 Application-Aware Resource Management for Wireless Mesh Networks and YouTube

In this section, we investigate application-aware resource management for wireless mesh networks.

The motivation for the implementation of application-aware networking for mesh networks lies in the flexibility of such a network. In mesh networks, the performance of the network significantly depends on the individual channel qualities of the wireless links between the components and the optimization with resource management thereof [140]. Due to these characteristics, mesh networks maintain many different options and management possibilities to transmit data from one node to another. Consequently, there are many possibilities for optimization, which give excellent opportunities to quantify the benefits of application-aware resource management.

4.2.1 Problem Definition and Architecture

A mesh access network is a special form of a wireless ad-hoc network in which all nodes forward data, regardless of whether they are recipients of the data or not. Through this principle of cooperation, data can be passed across multiple network nodes until they reach the destination node or a gateway to the Internet. The structural design of a mesh network with clients, mesh access points, mesh nodes, and Internet gateways is depicted in Figure 4.1. The advantage of a wireless mesh network over other access networks is the large coverage, which is possible since the components are not wired, apart from the gateways to the Internet.

Problems in a mesh network however arise at different layers (e.g. multi antenna transmission at physical layer, multi channel section on data link layer, dynamic routing decisions at network layer). In the following, one issue is discussed in detail, which can be counteracted by application-aware resource management.

One problem is that a wireless link can have significantly varying capacity in the network[1]. This is mainly due to the fact that interference may occur due to the simultaneous transmission and reception of data while forwarding. Additionally, interference may occur due to the transmission on adjacent transmission paths as in any other wireless network. Especially, in a network with high meshing, it is likely that the transmissions interfere with each other, which reduces the overall performance. This issue is commonly counteracted by interference management or by a coordination of parallel transmissions [136].

Application-aware resource management evaluates information about applications and knows about the actual demand of the applications in the network. This knowledge can be exploited to deal with the varying link capacities. Con-

[1] An implication of this is that the theoretical capacity of a mesh network is still unknown and an open question in science [140].

sequently, a certain path with low capacity can be avoided in case of high application demands. Further on, a certain transmission can be delayed if the instantaneous application demand is low. Another example is to prefer critical transmissions to the disadvantage of other concurrent transmissions. This knowledge about the applications constitute in mesh networks a nominal additional *degree of freedom* for the resource management process. Table 4.1 provides a list of common degrees of freedom for resource management which defines resource management opportunities. The last line corresponds to the additional knowledge according to the application-aware resource management.

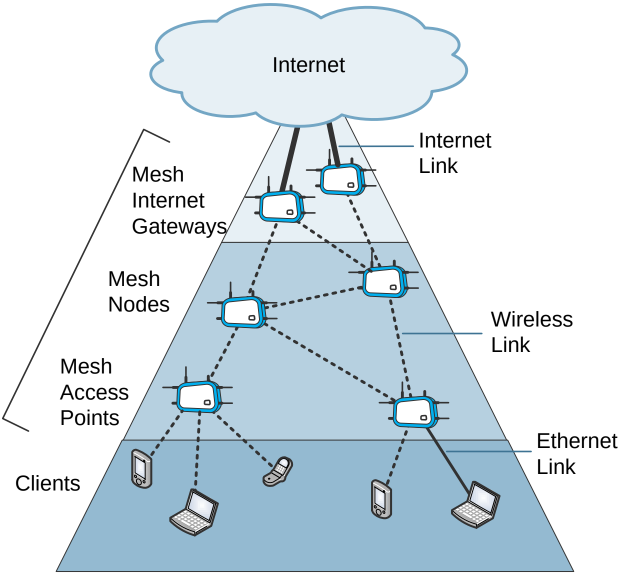


Figure 4.1: *Structure of an infrastructure mesh network [151].*

Table 4.1: *Degrees of freedom for the resource management of mesh networks.*

Networking layer	**Description of the degree of freedom**
Physical Layer	Selection of the physical settings for the transmission (modulation, rate adaptation, transmit and receive antenna, transmission channel)
Network Layer	Selection of the path due to the routing decision and network management
	Selection of transmission path due to current link capacity and load (might be influenced by other transmissions in parallel)
Application Layer	*Current application requirements on the network*

Following the definitions of Chapter 3, several components are required for application-aware resource management in mesh networks. In the following, we describe the architecture and give implementation details for the realization of the components within the mesh network. The framework is implemented by integrating YoMo as monitoring component within the resource management and other required components. The implementation follows the conceptual basics as described in Section 3.3.

In order to collect information about the network and application status, an *application monitor*, a *mesh network monitor*, and a central decision entity, the *mesh network advisor*, are required. The network advisor triggers the resource management actions that either change the traffic handling in the mesh network or the traffic produced by the application. We distinguish between *network control actions* and *service control actions.*

The resource management follows a reactive approach for resource and traffic management. As a consequence, it runs in addition to other traffic or resource management mechanisms that target on other degrees of freedom in the network. As long as the application layer performance and QoE is good enough, the resource management tries not to interfere with routing or other mechanisms. Only if a QoE degradation occurs, actions are triggered. Figure 4.2 shows the network structure with the different resource management components of the application-aware framework. Figure 4.3 provides an overview of the single components.

4.2.2 Implementation Details

Required components for the resource management are network and application monitors, a mesh network advisor (decision entity), and resource management actions. First, we describe the monitoring part of our implementation. The monitoring provides the necessary application information and helps to identify the overall network situation to enable a targeted resource management. Thereafter, the network advisor, the control algorithms, and actions are described. Since the focus is on YouTube in this subsection, all entities are described with respect to YouTube video streaming.

YouTube Application Monitor

Application monitoring is performed directly on the client side. Therefore, the client-based version of YoMo with the Mozilla Firefox extension is used that monitors the instances of Adobe Flash embedded on a website. On the one hand, we query the buffered playtime of the YouTube player. On the other hand, we additionally request general information for the resource management actions. This includes for example the possible video resolutions for YouTube which are both, offered by YouTube and currently supported by the end-user device.

In particular, the YouTube application monitor queries the following parameters and forwards them to the network advisor: (a) current buffered playtime in seconds, (b) available video resolutions as defined in [152], (c) current video resolution, and (d) flow information like transport layer ports and IP addresses.

The application information is continuously measured at the client but only reported in an event-based way. Information for resource management actions such as available video resolutions can be identified directly at the beginning of

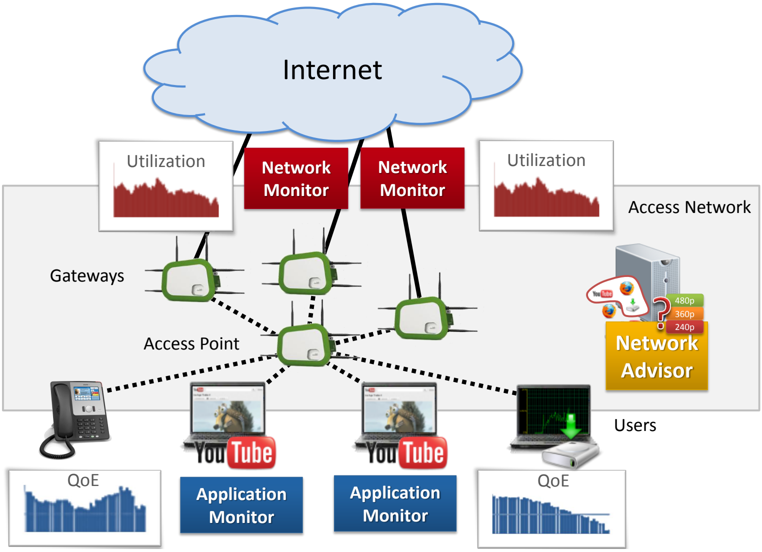


Figure 4.2: *Network structure with resource management components.*

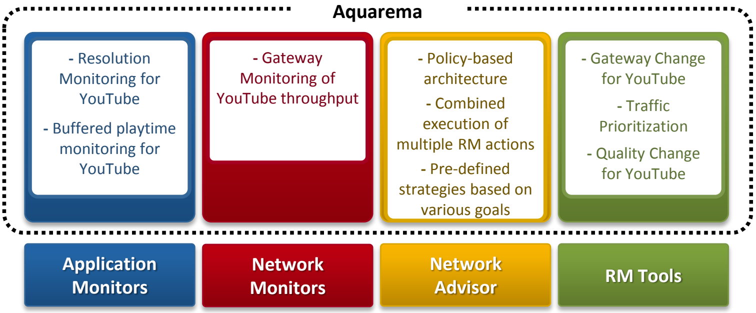

Aquarema
- Resolution Monitoring for YouTube
- Buffered playtime monitoring for YouTube
- Gateway Monitoring of YouTube throughput
- Policy-based architecture
- Combined execution of multiple RM actions
- Pre-defined strategies based on various goals
- Gateway Change for YouTube
- Traffic Prioritization
- Quality Change for YouTube
Application Monitors
Network Monitors
Network Advisor
RM Tools

Figure 4.3: *Used components and their task in the framework for mesh networks.*

the YouTube video playback. The data is sent once immediately after the video playback was detected. If the user changes the video resolution or a new video stream is requested by the YouTube player a similar message is sent to the network advisor right after the detection of the change. The buffered playtime is reported according to configurable thresholds. It may be necessary to adjust these thresholds based on the resource management action that uses the values, e.g. see Section 4.2.3 and Table 4.2.

Network and Flow Monitor

The network and flow monitor has two different functions. It measures the utilization of individual links and the current throughput of individual flows in the network. This information is used by the network advisor to estimate the benefit of possible resource management actions.

The load of the different links to the Internet is directly measured at the corresponding router or switch to the Internet. It is described by two values: the maximum capacity and the current throughput on the link. The current throughput is periodically polled every second by the network advisor from the network monitor at the router or switch. A moving average is calculated with a window size of 5 s to compensate short load peaks. We assume that the maximum capacity of the link is fixed and known at the network advisor.

To determine the throughput and state of individual flows, the network advisor sends the flow signature consisting of the IP address and transport layer port to the network monitor. The network monitor at the router uses a connection tracking module to gather the information. In the case of Linux OS, the kernel module *conntrack* is used. For Microsoft Windows based systems, the Event Tracking for Windows (ETW) is used in the network and flow monitor. If a flow is monitored, the router sends once in a second the current throughput of the flow to the network advisor which calculates the moving average for this flow.

Mesh Network Advisor

The mesh network advisor is the central entity triggering the resource management. It periodically collects information from the network and receives information from the event-based application monitors. All information is stored in a database so that a set of information about current applications in the network and the current network situation are known. Based on this information, the network advisor is able to trigger a number of resource management actions.

To be able to conduct the resource management actions, *strategies* are defined. Strategies map a certain application key performance indicator to a set of resource management actions. For example, there is a strategy for the buffered playtime of YouTube video streaming that is associated with the resource management action Gateway Change. This strategy is introduced in Section 4.2.3. In contrast, there is additionally a strategy that allows combined resource management. Here, in the resource management strategy, two actions are included, for instance, Gateway Change (network control) and Video Resolution Change (service control), see Section 4.2.3.

Within each strategy, for each application and key performance indicator, a critical threshold is defined. If this threshold is exceeded, the network advisor assumes that the application is in a critical condition. If this is the case, it runs the resource management actions of the set of actions defined in the strategy. Each resource management action returns a status information as return value that indicates (a) whether the action was successful, or (b) how long it should wait before the next resource management action is triggered. A waiting period after a transacted resource management action is necessary since it takes a short time until an action is enforced in the network. After each action the network advisor waits the time that the previous resource management action returned. If the action was not successful, it executes the next resource management actions in the list. If a resource management action was successful, the network advisor terminates resource management and evaluates again the key performance indicators whether the application is in a critical state or not.

YouTube Resource Management Actions

Resource management actions are undertaken at gateways and routers in the network. Depending on the capabilities such actions may range from additional entries in the routing table to the prioritization of application classes or flows, sophisticated resource allocation actions or even an active manipulation of the application layer content. After the rule is enforced a confirmation is sent to the network advisor. Figure 4.3 provides in addition to the components an overview of their functionality.

4.2.3 Definition of Different Resource Management Actions

We distinguish between two different types of resource management actions: network control and service control.

The concept of network control covers all measures which alter network properties or influence the packet flow in the network. The general goal of this concept is to improve the overall QoE of the users. To achieve this, the network has to react dynamically to changing network conditions and requirements of the users' applications.

In this work, two resource management actions that belong to network control are implemented: *Gateway Change* and *Buffer-based Prioritization*. The first management action allows a rerouting of packet flows to different gateways with less utilization. The second network management action implements traffic shaping to fairly distribute the available capacity according to the application needs. This is done by the prioritization of network flows in order to help applications if their QoE deteriorates. We further refer to this as Buffer-based Prioritization. A detailed description of the algorithms is given in the following subsections after the enumeration of the implemented service control mechanisms.

The second type of resource management action which is investigated in this work is service control. This includes mechanisms that control the users' applications such that the QoE of a single service is assured. Similar to network control actions, this implies that applications must accept resource management commands. As soon as a service level cannot be sustained, the service control mechanism notifies the application. If a degradation in the quality of the service is imminent, the application is adapted to the new conditions, if possible. Consequently, the application quality experienced by the user can be alleviated slowly and abrupt service failures which ruin users' QoE [118] can be avoided.

There are many different mechanisms for reaching this goal. In this work, video quality reduction of YouTube video streams is implemented by subsequently decreasing the video resolutions. In the following, it is called *Quality Change*. Other approaches include the adaptation of audio/video codecs as it is already implemented by Skype, or within the Annex G extension of H.264/MPEG-4 AVC video standard which is commonly referred to as Scalable Video Codec (SVC).

Both types of resource management actions improve the QoE in the network. Consequently, a combination of them is desirable for an efficient resource management. The combination, furthermore, can be done according to various objectives or provider preferences. For example, one provider policy can save network resources as long as an acceptable QoE can be maintained, or in contrast, the QoE can be maximized by using all available network resources. Eventually, a combined network and service control shall be provided, which utilizes network parameters, application parameters, and provider-dependent directives to maximize the perceived quality for a set of users in the network while, in each situation, minimizing network operator's costs.

Network Control: Gateway Change

In access networks such as wireless mesh networks multiple gateways to the Internet might exist. The resource management tool *Nigel* is responsible for dynamically assigning the clients to these different gateways [6]. Changing the Internet gateway of a client during run-time requires taking care of the active connections between a user and the Internet. To achieve this goal, Nigel follows the

Mobile IPv4 approach. It establishes an overlay network that ensures a seamless TCP handover. According to the Mobile IPv4 approach, an anchor - the ”gatekeeper” - is located in the Internet as so-called home agent which maintains the IP connection to the corresponding service. The overlay network between the access point and the home agent is established via IP tunnels. Thus, selecting another Internet gateway changes the routing of the IP tunnel. As a consequence, changing the Internet gateway of a client does not affect the actual connection between the home agent and the Internet service as only the virtual paths of the IP tunnels are changed. Based on the monitored information about the current gateway utilization and the needs of the hosted application streams, the network advisor decides which stream is assigned to which gateway. In the following, the algorithm and Nigel's gateway switching policy are described in detail.

Nigel is installed at the edges of the access network, namely on each access point and on the gatekeeper. The nigel instance at the access point manages the uplink direction while Nigel running at the gatekeeper is responsible for the downlink direction. To switch a stream to another gateway, a message is sent to Nigel running on the client's access point, naming the new gateway. It switches the uplink and sends a message to Nigel on the gatekeeper also naming the new gateway. Nigel on the gatekeeper switches the downlink and confirms the gateway switch.

In Figure 4.4 the resource management policy of the gateway change is depicted. The network controller checks at each status update of a YouTube flow if the condition for the resource management action is met, and if the video is already playing. The condition for Gateway Change for YouTube video streaming is that the buffered playtime is below 10 s and that the start time is at least 5 s ago.

We define for YouTube a threshold of 10 s buffered playtime and a start delay of at least 5 s to ensure that the web page and the video player is loaded as well as that the playing of the video has already begun. The download and initialization of embedded Flash objects within the browser can take up to a few seconds. Moreover, it may happen that a YouTube video request of the video player is redirected in some cases with HTTP error code 302 to a secondary YouTube server due to overload which costs some additional time.

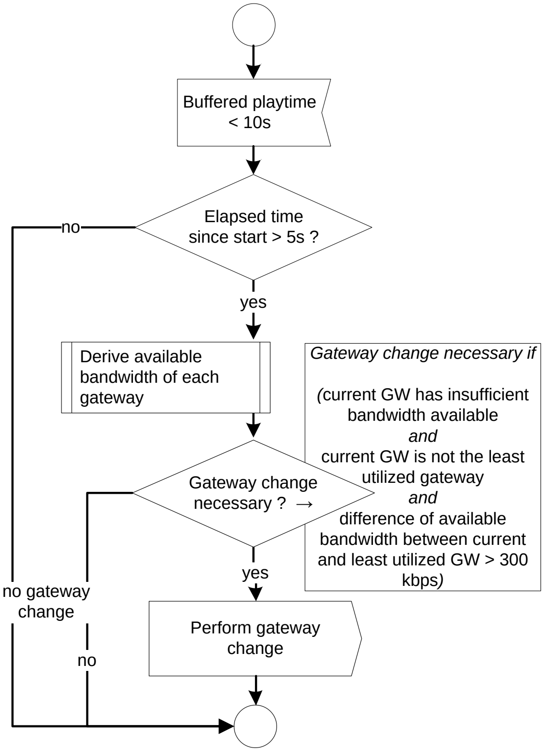


Figure 4.4: *Gateway Change algorithm.*

If all conditions are met, the resource management starts and the available capacity of each gateway is determined. As long as the current gateway has sufficient capacity, as long as the current gateway is the least utilized gateway, or as long as the capacity difference between the current and the least utilized gateway is negligibly small (less than 300 kbps), no gateway change is carried out. In all other cases, the flow is allocated to the least utilized gateway.

Network Control: Buffer-based Prioritization

When multiple YouTube streams compete for the available capacity of a gateway, the capacity assignment is handled arbitrarily by the TCP protocol control mechanisms. Therefore, it is possible that streams with similar needs get strongly different shares of the available capacity. Consequently, one video might struggle unnecessarily with low buffer sizes, increasing the probability of stalling (i.e. interruptions). To overcome this TCP-caused behavior, means to prioritize struggling streams and to distribute the available capacity more fairly according to the application state are required.

Table 4.2 shows the prioritization policy which is performed on each gateway. The video stream is assigned the respective priority 5 down to 1 with 1 being the highest priority. For this action not only one critical state threshold, i.e. buffered playtime is below one certain threshold, is considered. Instead, depending on the current buffered playtime of a YouTube stream, its priority is updated on every

Table 4.2: *Prioritization policy for YouTube.*

Buffered Playtime	Priority Class
$> 15\,s$	5
$> 10\,s$	4
$> 5\,s$	3
$> 2\,s$	2
$\leq 2\,s$	1

status update. The provided thresholds are critical for the resource management and were obtained empirically as the most adequate values. They must be far enough apart that the system does not tend to overreact, and close enough to allow sufficient priority changes, in order to avoid situations where one video is preferred for an excessively long period of time.

With this algorithm, the bandwidth can be allocated to the flows according to their buffered playtime. All flows of the highest priority class are processed first. The remaining capacity is now available for the flows of the second highest priority class. Again, they are served according to their needs and likewise the remaining capacity is available for next lower priority class. This distribution is continued until either no more streams or no more capacity is left. Thus, it is possible that flows of lower priority classes are not assigned any bandwidth at all. As their buffered playtime decreases, consequently, their priority increases and their needs are served again. Currently no actions are taken to distribute the available bandwidth equally within a priority class.

Service Control: Quality Change

In case of YouTube video streaming, this resource management action allows to dynamically change the video resolution on request. Depending on the uploaded video, YouTube currently offers 240p (i.e. 240 pixels vertical resolution), 360p, 480p, and even High-Definition (HD) videos with 720p, 1080p or ”Original”, which means a resolution of up to 4096x3072 pixels (4K). Each playback video quality requires different download bandwidths and consequently, a change in the video quality results in a change of the throughput of the YouTube video. This effect is exploited by the resource management action. If there is enough bandwidth available, the video quality is changed to the highest possible quality. However, if the network is congested and the application monitor measures a low buffer level of the YouTube video, a lower quality is suggested for the video to ensure a smooth video playback without stalling. The implementation of the quality change and the service control policy are described below.

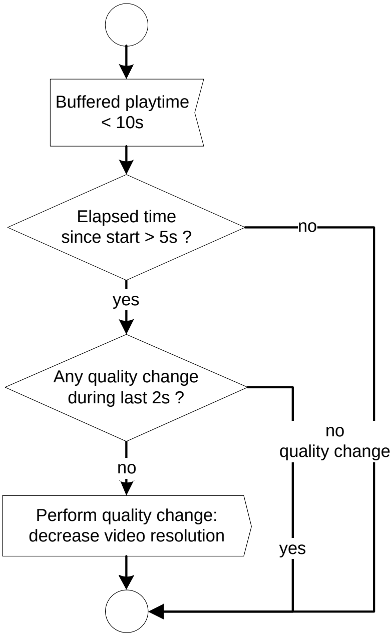


Figure 4.5: *Quality Change algorithm.*

To change the quality of a streamed video, the algorithm uses the YouTube player API which provides the possibility to set the playback quality of the video. The function causes the video to reload at its current position in the new quality just as if the user herself clicked the corresponding button at the video player. The old data is discarded and a new stream is requested from the YouTube servers. Beginning with a new FLV header, the servers start to stream the video in the new quality, i.e. the new resolution. Due to the new video stream and since the old data is discarded, every quality change causes a short stalling event but prevents the video from struggling with unsatisfiable needs which would result in even more and longer stallings.

To determine whether a quality change is necessary, the resource management algorithm performs checks on each critical application state which are depicted in Figure 4.5. The quality of the video is switched to the next lower level only if the video playback time is already larger than 5 s and there was no other quality change in the network during the last 2 s. These checks assure that only video streams which are not in the initial phase and are struggling (i.e. have a small buffer size) are changed by the algorithm.

Combined Control

While both network control and service control have proven their effectiveness in different test cases, for different purposes, combined control actions are required. As a start, two simple strategies are defined:

Network Control First As long as the problem can be solved by the network, only network control is used.

Service Control First As long as a sufficient QoE can be guaranteed, only service control is used.

For example, if the goal is to optimize the overall QoE, the first approach is useful. This means that all possible resources are utilized without considering the costs for transmission. In contrast, if the goal is to reduce the required transmission resources, the second strategy should be preferred. It can be used to a certain extent to rather provide a medium quality for all users than a high quality which may be, from the provider point of view, expensive compared to a lower quality.

4.2.4 Evaluation of Application-Aware Resource Management for Wireless Mesh Networks and YouTube

The resource management used for the measurements is based on the framework for mesh networks as described in the previous section.

For the evaluation, the framework is installed in a wireless mesh testbed which serves as access network for clients. The network consists of four mesh nodes which are connected by WiFi. One of the nodes is the access point (i.e. node to which all clients are connected) and the other three nodes are mesh gateways (i.e. mesh nodes having access to the Internet). The structure of the testbed is the same as depicted in Figure 4.2. Each gateway has a fixed capacity of 3 Mbps and, thus, forms the bottleneck of the network. Compared to the bandwidth of 3 Mbps at the gateway, we assume that the connection between testbed and the YouTube server provides enough capacity, so that it does not have any effect on the measurements. The network monitor tool is installed on each gateway node to report its utilization and available capacity to the Internet.

Up to four client PCs are connected to the access point node by WiFi. They give users the possibility to watch YouTube videos in a browser. On each client the YouTube application monitor is installed to signal the presence of video streams and to collect information. Additionally, one separate PC within the mesh network hosts the network advisor which receives all information from both, mesh monitors and application monitors, and decides about resource management actions. The network monitors at the gateways are connected directly to the advisor. The application monitors communicate with the advisor through the access point.

Reference Scenarios

The objective is to evaluate the resource management actions. Therefore, we compare the behavior without resource management with the behavior when resource management is enabled. We consider different video qualities and distinguish between synchronous, i.e. the videos start at the same time, and asynchronous start of YouTube videos. As metric for the evaluation, we focus on the buffered playtime since, according to [120, 135], stalling is the factor dominating the QoE of YouTube video streaming.

Table 4.3[2] lists the combinations that are used in the scenarios with a synchronous start of YouTube videos. The next section demonstrates how the startup

[2] A line is highlighted in this table since it is discussed later on in Section 4.2.4 or Section 4.2.4.

(synchronous or asynchronous start) has a big impact on the stalling behavior due to the pre-buffering pattern of YouTube. Therefore, in Table 4.4[2], for comparison, scenarios with a delayed start are defined. In this scenario, the videos start with an interval of 30 s. The first column in the tables indicates how many videos are used. If x is the number of videos in resolution 480p, y is the number of videos in 360p, and z is the number of videos in 240p, then $x/y/z$ denotes how many videos in 480p, 360p, or 240p are used for one test run. The other columns in the tables show the results within the testbed network as a reference without resource management and if only one gateway is used. The tables show the mean values of number of stalling events, stalling length, used bandwidth, and theoretical bandwidth, which were measured in the testbed. The discussion of the different results is done in the next section. The table headings contain abbreviations. The meaning of the abbreviations is explained in Table 4.5. For every combination at least 20 test runs are done. The columns in the tables show the average of all test runs.

Table 4.3: *Synchronous video start - no control*

Videos	n_s	t_s	$\overline{bw}$	bw_{tot}
0/0/1	0.00	0.00	0.92	0.69
0/1/0	0.00	0.00	1.61	1.20
0/0/2	0.14	0.61	1.84	1.38
0/1/1	0.14	0.79	2.53	1.90
1/0/0	0.00	0.00	2.55	1.90
0/2/0	0.55	13.19	2.98	2.41
1/0/1	0.92	47.45	2.96	2.59
0/0/4	3.50	20.71	2.98	2.76
1/1/0	1.57	144.00	2.97	3.11
2/0/0	1.57	263.43	2.98	3.80

Performance Investigations of the Reference Scenarios In the following, we determine situations in which YouTube encounters problems. In particular, this is the case if the network is overloaded. To allow a practical evalu-

Table 4.4: *Delayed video start - no control*

Videos	n_s	t_s	$\overline{bw}$	bw_{tot}
1/0/1	0.14	2.69	2.98	2.59
0/0/4	0.43	1.93	2.99	2.76
1/1/0	0.84	129.47	2.98	3.11
2/0/0	1.78	247.64	2.99	3.80

ation of our results, we restrict ourselves to the reference scenarios and our test network, and explain, based on estimations and practical measurements, when a critical situation may occur. Consequently, our results apply for the particular network only. However, the statement and the observations are also valid for other small to medium-sized access networks or other network structures.

In the reference scenarios, the same YouTube video with three different sizes of 240p, 360p, and 480p is used which have mean video rates of 0.69 Mbps, 1.20 Mbps, and 1.90 Mbps respectively. When considering only the mean video rate, videos with a total rate of up to 3 Mbps should be able to run smoothly in parallel on a single gateway (e.g. 4x240p with 2.76 Mbps, or 1x360p and 2x240p with 2.58 Mbps). The videos, however, are coded differently across the entire playing time using adaptive H.264/MPEG-4 AVC encoding. This may result in variable video bitrates. It means that even if on average a video may fit on a link, sometimes a higher temporal data rate is necessary to prevent the video from stalling. A video with high motion at the beginning and a slow 360-degree video pan over the scenery, for instance, at the end, is highly unequally encoded and requires more data at the beginning than at the end. In order to take this into account, YouTube generally transfers the video content within two transmission phases. At the beginning, the buffer is filled initially with a certain amount of data to compensate for variations in the video coding. This download pattern causes different download rates that have to be considered in the resource management.

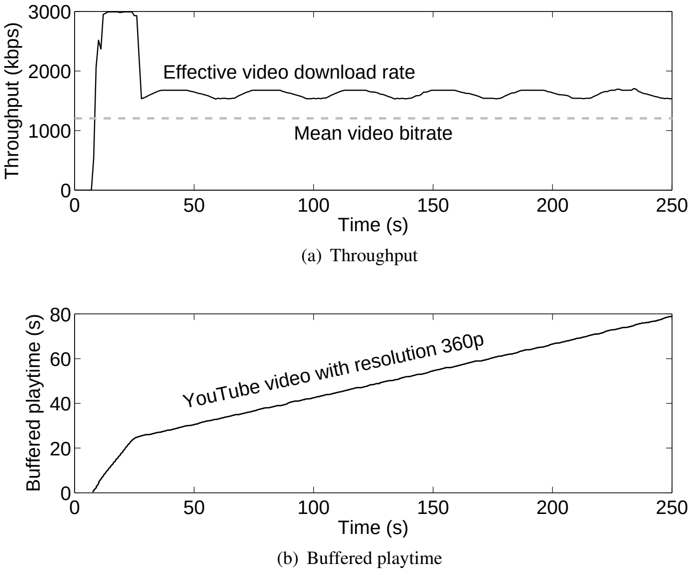
The specific download pattern of a YouTube video which is streamed in the testbed is shown in Figure 4.6. In the upper figure it can be seen that from

Table 4.5: *Used metrics and their abbreviations*

Abbreviation	Explanation
n_s	Average number of stallings during the video playback
t_s	Average stalling time during the video playback
$\overline{bw}$	Average bandwidth used on the gateway
bw_{tot}	Sum of average video bitrate of all videos
n_g	Number of conducted gateway changes
$\|GW\|$	Number of used gateways at the end of the run
n_p	Number of conducted prioritization changes
n_r	Number of conducted resolution changes
$\overline{buf}$	Average buffered time of the videos at the end of the run

the beginning the video uses the maximal available bandwidth of 3 Mbps and the buffered playtime increases rapidly (lower figure). After the initial burst, the stream is in periodic refill phase and the used bandwidth drops to a rate slightly above the mean video rate. As a consequence, the buffer occupancy increases more slowly.

Our measurements showed that the videos 240p, 360p, and 480p request a mean bandwidth of 0.92 Mbps, 1.61 Mbps, and 2.55 Mbps respectively in the first five minutes of the video. Compared to the mean total video rates, these values are about 25% higher. Considering these higher bandwidth values, the number of videos being able to run on a gateway in parallel need to be reconsidered. For instance, in case of 1x360p and 2x240p despite of the total mean video rate of 2.60 Mbps, the videos try to request a total bandwidth of 2x0.92 Mbps + 1x1.61 Mbps = 3.45 Mbps. Obviously, this data rate of 3.45 Mbps is too large for the gateway such that not all demands of the videos can be satisfied.



(a) Throughput

(b) Buffered playtime

Figure 4.6: *Reference measurement of YouTube streaming behavior for a 360p video in the testbed*

The reference scenarios can be divided into three categories depending on the mean video rate of the videos. For each category a different kind of resource management is performed later on.

Category 1: Video combinations having a total theoretical bandwidth of less than 2.1 Mbps. The average stalling length is around 0 s. They run smoothly on the gateway. No resource management is required.

Category 2: Video combinations having a total theoretical bandwidth between 2.1 Mbps and 3 Mbps. They use the maximal available bandwidth but stalling occurs occasionally. The performance of the individual videos depends strongly on the starting delay and order.

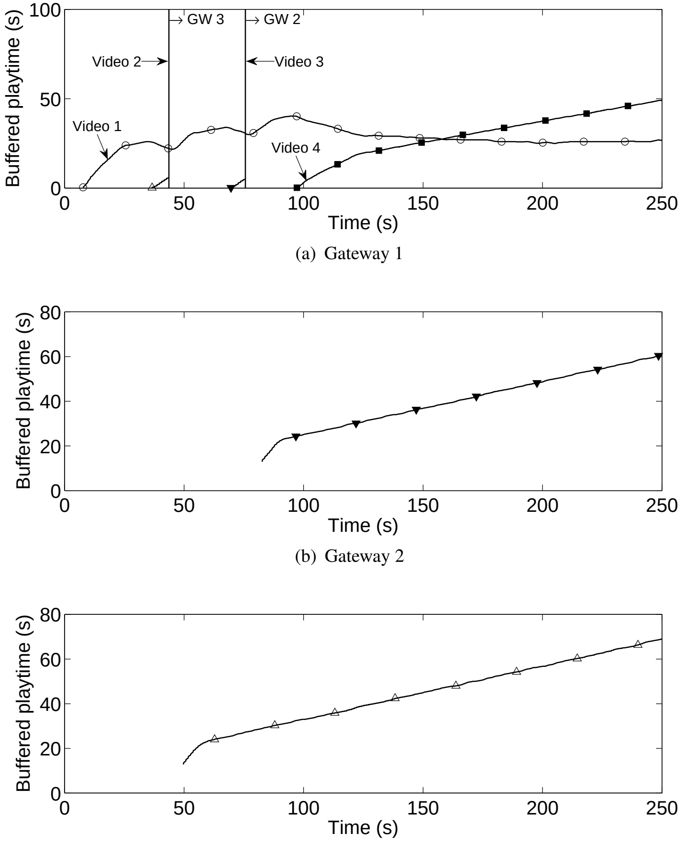
Category 3: Video combinations having a total theoretical bandwidth of more than 3 Mbps. They cannot run smoothly on the gateway and are almost permanently stalling. Therefore, a resource management has to be performed to reduce their required bandwidth. This is addressed later on in Section 4.2.4.

The second category is the most interesting one. If a video combination of this category is put on a gateway and the videos are started at the same time (synchronous start), there are two possible resulting effects. In the first case, one or two videos manage to fill their buffers as desired, resulting to the third video's inability to keep the buffer on a constant level. After a while, one of the videos will start stalling. In contrast, the others fill their buffer excessively. In particular, videos with higher resolutions suffer from this situation due to their higher bandwidth demands. The other possible effect is that all videos share the bandwidth equally. Especially in case of different resolutions, this is not the best choice. Instead, the videos should share the bandwidth proportional to their mean video rate since the throughput of videos with higher resolution should be higher than the throughput of low-quality videos, even if progressive download is used and a buffer is filled. Our measurements showed that basically all combinations without any resource management mechanisms end up in the first described situation.

Compared to Table 4.3 where the videos are started at the same time, in Table 4.4 the results for the asynchronous start are depicted. The videos started one after another and have their initial buffering phase in succession which leads to less stalling. Thus, a simple possibility for resource management is to delay the start of the videos. However, even if a video is started delayed, video combinations having a theoretical bandwidth of more than 3 Mbps cannot run smoothly without stalling.

Evaluating Different Control Approaches

Network Control: Gateway Change First, we consider the network control action Gateway Change. Four 360p videos are started sequentially with an interval of 30 s using the same gateway. Figure 4.7 shows the temporal progress



(a) Gateway 1

(b) Gateway 2

(c) Gateway 3

Figure 4.7: *Dynamic gateway change with 4x360p videos, sub-figures show individual gateways*

of the buffered playtime of the videos for the different gateways. At first, a single video is transmitted over *Gateway 1* and its playout buffer increases in the usual way. A second video is added to the gateway but its playout buffer cannot be filled properly. Thus, the stream is switched to *Gateway 3* where it gets enough capacity to fill its buffer. The same mechanism is applied to two more videos that use *Gateway 1* as the initial gateway. The first stream is switched to another gateway. The last stream is not switched since it would not improve the situation. In the end, *Gateway 2* and *3* host one stream each, and *Gateway 1* hosts two streams and all videos have sufficiently filled buffers.

This example shows that the gateway switching control mechanism helps struggling streams to increase their playout buffers which avoids stalling of the videos. According to [120, 135] stalling is the factor dominating the QoE for video clips. Consequently, the QoE of the users is increased. From a network perspective, this resource management leads to a balanced load on the available network resources. Compared to common load balancing on network-layer, however, we take the instantaneous application state into account, i.e. the current buffered playtime. This means that even situations where videos with different resolutions are used, or when users pause the video, or jump within the video can be addressed. For example, if a user is manually selecting a higher resolution, the resource management algorithm will recognize this due to a low buffer state and will relocate the flow to another gateway if capacity is available.

In Table 4.6 the other test runs and their aggregated statistics can be seen. Compared to the reference scenario, the average number of stallings and the average stalling lengths have diminished as up to three gateways are used. Especially when the videos are started delayed, the videos face almost no stalling. However, four 480p videos do not fit well on the three gateways of our testbed. Thus, with this combination stalling cannot be prevented.

In general, from a QoE perspective, stalling can be avoided if enough capacity is available to support all YouTube videos. However, situations where a video on one gateway buffers too much data and certain videos suffer from this cannot be avoided. This issue is addressed in the next section.

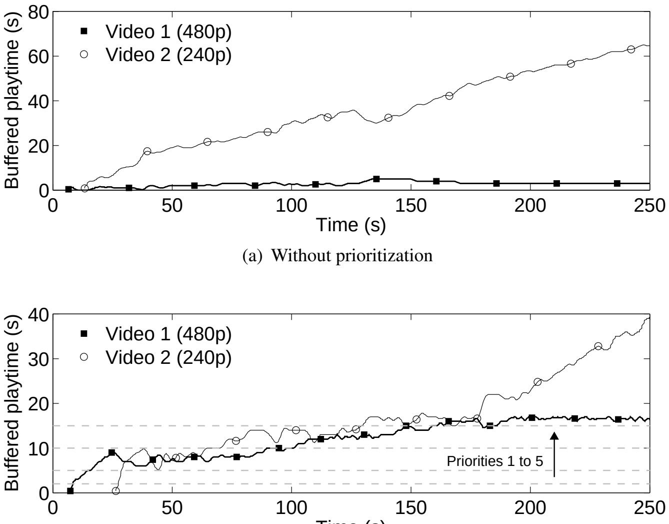
Table 4.6: *Delayed video start - dynamic gateway change*

Videos	n_s	t_s	n_g	$\|GW\|$	$\overline{bw}$	bw_{tot}
0/0/4	0.04	0.17	1.33	2.33	3.71	2.76
1/1/0	0.10	0.20	1.50	2.00	4.18	3.11
1/2/0	0.11	0.83	2.37	3.00	5.74	4.31
0/4/0	0.03	0.15	2.80	3.00	6.16	4.82

Network Control: Buffer-Based Prioritization In this section, we show that buffer-based prioritization of video streams helps to avoid stalling. In this example, a 480p and a 240p video stream compete for the bandwidth of the same gateway. Both videos do not fit at the same time on a single gateway and cause each other to stall as shown in the highlighted row in Table 4.3.

In Figure 4.8 the temporal progress of the buffered playtime is depicted. As described in the reference scenario and as can be seen in Figure 4.8(a), the 480p suffers most in this situation due to its higher bandwidth demand. The video cannot fill its buffer appropriately and is going to stall. In Figure 4.8(b), the situation with prioritization is depicted. The horizontal dashed lines represent the prioritization classes (cf. Table 4.2). If the buffered playtime of the stream is low, its priority is increased compared to the other stream. Then, the video is able to fills its buffer and the bandwidth requirements of the video are met until its priority becomes lower. Next, if the priority is lower, the other video can fill its buffer. This behavior continues until the end of the test run. Now, the buffer size of the 480p video oscillates around the last priority threshold and the 240p video continues to fill its buffer. Thus, with buffer-based prioritization it is possible that both streams coexist and none has a critically empty playout buffer.

In Table 4.7 it can be seen that in this example the average stalling length decreases from 47.45 s (cf. Table 4.3, highlighted line) down to 7.07 s. With the other combinations in the test scenario, the stalling decreases, too. This shows that buffer-based prioritization as a network control mechanism works well for our test network and is able to avoid TCP-caused problems with bandwidth shar-



(a) Without prioritization

(b) With prioritization

Figure 4.8: *Buffer-based prioritization with a 480p and a 240p video, sub-figures show the situation with and without prioritization*

ing. With respect to the QoE, this method allows an increase in QoE since a YouTube video which is almost stalling all the time can be supported without stalling, assuming the available capacity is enough for all YouTube videos.

Service Control: Video Resolution Change In the following, the performance of service control resource management actions is evaluated. The effects of video resolution change can be seen in Figure 4.9 in which two 480p videos are started sequentially with an interval of 30 s. The reference scenario (cf. Table 4.4, highlighted row) shows that in a normal situation the videos would stall

Table 4.7: *Synchronous video start - buffer-based prioritization*

Videos	n_s	t_s	n_p	$\overline{bw}$	bw_{tot}
0/2/0	0.18	0.42	36.00	2.99	2.41
0/1/2	0.33	1.64	95.00	2.99	2.59
1/0/1	0.63	7.07	116.73	2.99	2.59

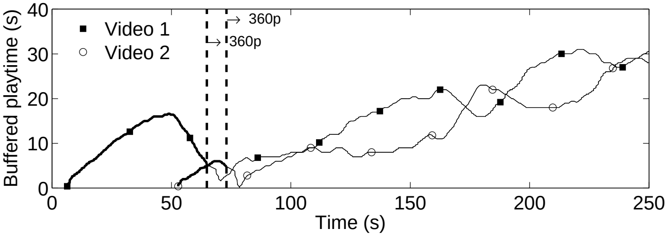


Figure 4.9: *Video resolution change of two YouTube videos*

permanently. In this scenario, service control is enabled which means that the video resolution is scaled down if the buffer occupancy drops below the control threshold. The figure shows that the 480p videos are changed to 360p one after another. Two 360p videos fit on a single gateway and each video is able to fill its playout buffer. Thus, in this case almost no stalling occurs and the video streams can coexist in the network. We have to point out that in our implementation, every resolution change (i.e. the start of a new video stream) causes a single stalling event which could be prevented by a smarter video player. Instead of discarding all data the smarter player could start the new stream early enough and switch the resolution seamlessly after playing out the whole buffer.

From a network point of view, a change to a lower resolution results in lower bandwidth requirements of the YouTube video. This has only a minor effect on the QoE (please refer to [120, 135] for a more refined analysis), but avoids

Table 4.8: *Synchronous video start - video resolution change*

Videos	n_s	t_s	n_r	$\overline{bw}$	bw_{tot}
1/1/0	0.70	15.98	2.10	2.45	3.11

Table 4.9: *Delayed video start - video resolution change*

Videos	n_s	t_s	n_r	$\overline{bw}$	bw_{tot}
0/2/0	0.08	0.89	0.27	2.86	2.41
1/1/0	0.43	6.54	1.35	2.83	3.11
2/0/0	0.76	17.83	2.86	2.55	3.80

stalling, which in turn avoids a severe QoE degradation. In fact this resource management action is particularly useful in overload situations. It works however only if the YouTube video is available in different resolutions.

In Tables 4.8 and 4.9 the aggregated statistics of the test runs are shown. It can be seen that the service control mechanism is useful as it helps the videos to fill their buffers to an adequate level. Stalling is short (especially with delayed video start) and could be fully prevented by a smarter player. Moreover, with service control, more video streams fit on the gateway than in the reference scenario. Thus, more users can be served at the same time in the test network.

This resource management action benefits, in our test network, in particular from the use of prioritization. Due to the YouTube streaming behavior with TCP, service control overreactions, i.e. too many and unnecessary resolution changes, might occur. To reduce the number of these overreactions, additionally applying buffer-based prioritization turned out to be helpful. In Figure 4.10, again, two 480p videos are started on the same gateway. This time, the buffer-based prioritization is activated as well. It turns out that both video resolutions are changed down to 360p and that the buffers of the videos are filled faster. The aggregated statistics of this example can be seen in Tables 4.10 and 4.11. In our test runs (delayed start) the average number of resolution changes dropped from 2.86 (cf.

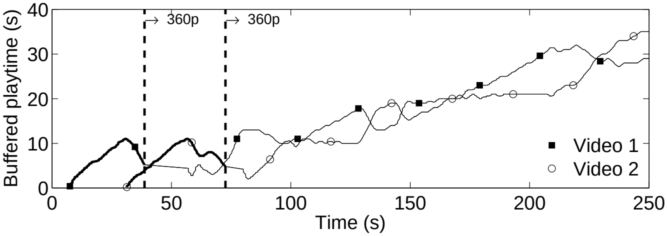


Figure 4.10: *Video resolution change with buffer-based prioritization for two YouTube videos*

Table 4.9) down to 2.36 by additional prioritization. Furthermore the average stalling length decreased from 17.83 s down to 4.25 s which further indicates the possible gain of a combined control strategy, which is covered in the next section.

Table 4.10: *Synchronous video start - video resolution change combined with buffer-based prioritization*

Videos	n_s	t_s	n_p	n_r	$\overline{bw}$	bw_{tot}
2/0/0	1.02	4.52	43.38	2.19	2.94	3.80

Table 4.11: *Delayed video start - video resolution change combined with buffer-based prioritization*

Videos	n_s	t_s	n_p	n_r	$\overline{bw}$	bw_{tot}
2/0/0	0.88	4.25	40.20	2.36	2.82	3.80

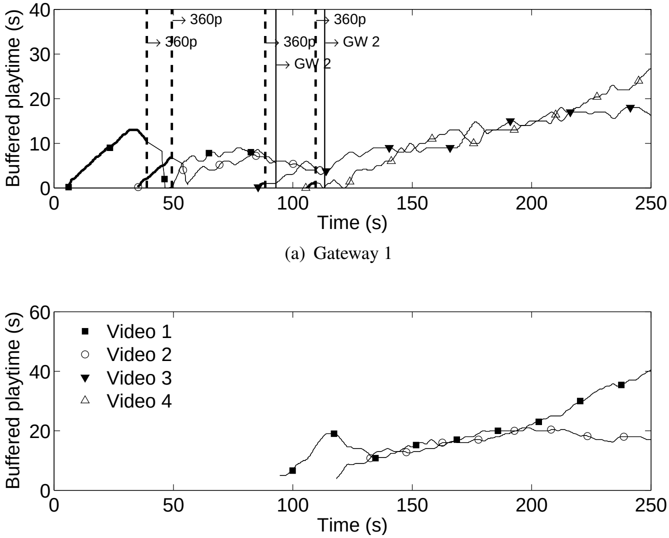
Combined Control To investigate the performance of a combination of the separately operating mechanisms, we consider the following example video combination. Four 480p videos are started in our testbed on the same gateway which would result in a heavy stalling according to our reference scenario without any control mechanisms. With combined control, the following three different strategies are examined.

Policy 1: Network Control First The four videos are distributed among the three available gateways. On one gateway two video streams remain which exceeds the capacity of the gateway. Thus, the resolution of the two videos are changed to 360p. Almost no stalling occurred and all videos could fill their buffers. This strategy reacts quite fast. However, many resources are needed.

Policy 2: Service Control First The videos are scaled down to a lower resolution first. As even four 360p videos do not fit on the gateway, the service control is applied again. Then, all videos have a resolution of 240p. As no more service control is possible, one stream is switched to another gateway. This stream could then be switched to a higher resolution again as the capacity of the gateway is sufficient. If prioritization is additionally enabled, it is even possible to use only a single gateway without much additional stalling. This ”strict” gateway minimization would be the most resource efficient strategy. However, there is a trade-off between resource utilization and buffer occupancy (i.e. risk of stalling).

Policy 3: Moderate Mix This example is depicted in Figure 4.11. All videos are scaled down to 360p first. Then, network control is enabled and two streams are switched to another gateway. Thus, again, two gateways are used but the videos can be kept on a higher resolution. This strategy addresses the trade-off between Network Control First and Service Control First.

The results of all strategies are summarized in Table 4.12. In case of our test scenario, the lowest number of stallings is achieved with the Network Control First strategy. Here, all three gateways are activated, which yields to a higher resource utilization compared to the other policies. Assuming that a quality of 360p is sufficient for an acceptable QoE for YouTube video streaming, the best trade-off between QoE and utilized network resources can be achieved with the Moderate Mix strategy.



(a) Gateway 1

(b) Gateway 2

Figure 4.11: *Combined control with policy Moderate Mix, sub-figures show the individual gateways (gateway 3 is not used and therefore, not displayed)*

To conclude this evaluation, we summarize work done so far. We evaluated the impact of resource management concepts for wireless mesh networks for YouTube video streaming in the previous subsection. The load can be balanced on different mesh gateways, if available, or the prioritization of the streams can be dynamically changed.

The results show that more YouTube users can be supported. The second option, service control, allows for changing the resolution of the YouTube video. Finally, the best trade-off between QoE and resource efficiency can be achieved

Table 4.12: *Combined control*

Policy	n_s	t_s	n_g	$\|GW\|$	n_p
Network control first	0.57	4.03	2.13	3.00	76.88
Strict service control first	1.04	10.47	1.00	2.00	77.92
Service control first	1.54	11.96	0.00	1.00	147.61
Moderate mix	0.86	6.93	2.32	2.00	87.53
	n_r	$\overline{buf}$	$\overline{bw}$	bw_{tot}	
Network control first	2.03	43.27	7.55	7.61	
Strict service control first	8.00	52.93	3.73	7.61	
Service control first	8.00	7.65	2.20	7.61	
Moderate mix 4.00	34.86	5.81	7.61		

using a combined control approach. Our findings here are that a strategy using a moderate mix of network and service control helps to keep the QoE on a high level without using too much resources in our test network and thus, reducing the energy consumption and the operational expenditure.

4.3 Application-Aware Packet Scheduling in Cellular Access Networks

Another important type of access network are cellular communication networks. The penetration rate is increasing worldwide [153] and the volume of data grows exponentially for years [154]. Mobile operators face the challenge that the costs must be reduced in order to increase profits, but the demand of the users steadily doubles every year. In the following, application-aware resource management is applied to the air interface of cellular networks.

Due to the ongoing liberalization of telecommunication markets, end users are in the position to freely choose between different mobile operators. The resulting intensive cost-driven competition among the different players has lead to the commoditization of Internet access services. Consequently, the mobile operators must reduce costs and increase the revenue per transmitted data. However, when price levels and pricing schemes become more and more low, another factor influencing a person's network choice comes into play: the quality of a service as perceived by the end user. Application-aware resource management may help in this case in both directions. It may improve the network efficiency, resulting in a potential cost reduction, as well as it might increase the user-perceived quality for a service. The numerical quantification of these benefits is subject to the following evaluation.

In the following, we restrict ourselves to the technical details for implementing application-aware resource management in cellular networks. In the remainder, the problem definition is given and the architecture is described in Section 4.3.1. Subsequently, two application-aware scheduling mechanisms are proposed in Section 4.3.2, which are evaluated and discussed in Section 4.3.4. In Section 4.3.3, the evaluation methodology is presented.

4.3.1 State of the Art Quality of Service Management in Cellular Networks

The challenges in mobile communication networks for operators are twofold: on the one hand, customers have high expectations on the delivered quality the services, which will be to a large extend based in the Internet and not under the control of the operator. On the other hand, the operator must minimize the ongoing costs of operating the network. These often contradictory objectives imply that for providing high QoE to the customers, novel solutions based on traffic differentiation according to the application or service requirements are needed.

The current mobile communication paradigm therefore is to differentiate on service level for provisioning of QoS to the end user. For this purpose, different QoS classes are standardized which are then mapped to different applications according to their approximate requirements.

From a technical perspective, in 3GPP LTE *bearers* are used to forward the data between user equipment and the Internet gateway. A bearer is a virtual connection between an end user device and the packet gateway of the mobile network to the Internet. Each bearer is set up with a bearer QoS profile that specifies guaranteed parameters on network level. With the classification into different QoS classes, groups of network flows with similar needs are prioritized equally and packet scheduling of different groups is done according to the QoS definitions.

The problem however is that, at present, almost every mobile network exclusively establishes only default bearers for data transmissions without guaranteed resources. The reason is that the mapping between IP traffic and QoS classes is not clear. Further information is required, e.g. more knowledge about the application, current requirements of the application, IP traffic-on-application mapping.

Application-aware resource management is able to provide more application-related information to the network. It can feedback quality indicators for application obtained by the monitoring. They can be used to determine the current demand of an application on the network. This allows to influence the bearer handling to address the quality of the applications of the users.

Similar to the discussions on mesh networks, this knowledge about the applications also provides another nominal degree of freedom for the network. A bearer can thus be regarded as critical, if applications are transported in it that are in a critical state. In contrast, even the traffic forwarding for bearers may be postponed if they contain non-critical traffic, releasing resources for other applications.

In the following, an application-aware traffic handling is defined in the air interface at the base station of a mobile network. At the base station, the order for the transmission of packets (*packet scheduling*) is affected based on the application-aware quality indicators. This is done in addition to other parameters such as signal strength or experienced inter-cell interference. In this work, it is assumed that the scheduling can be divided into two sub-functions. These are the *packet scheduling* and the *resource allocation*. The packet scheduling determines

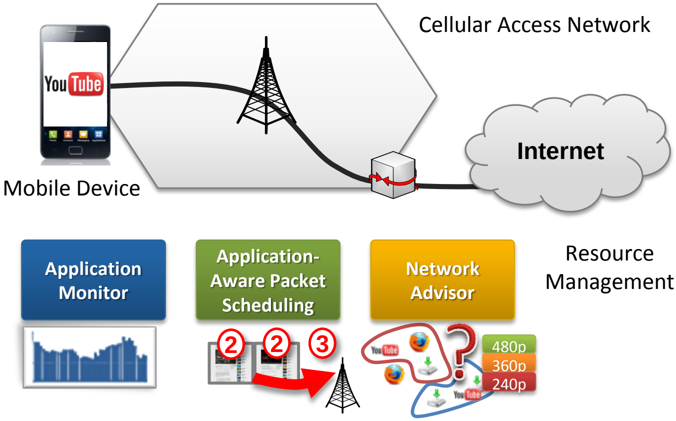


Figure 4.12: *Placement of components for application-aware packet scheduling in a cellular network.*

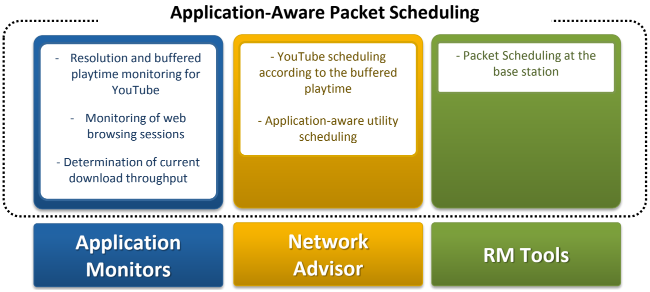

Application-Aware Packet Scheduling
- Resolution and buffered playtime monitoring for YouTube
- Monitoring of web browsing sessions
- Determination of current download throughput
- YouTube scheduling according to the buffered playtime
- Application-aware utility scheduling
- Packet Scheduling at the base station
Application Monitors
Network Advisor
RM Tools

Figure 4.13: *List of components and their functionality in application-aware packet scheduling.*

the order of processing of the packets. The resource allocation places the packets on the transmission frame according to this order and determines transmission parameters such as modulation and transmission strength.

Following the definitions for application-aware resource management in Chapter 3, several components are required. As indicated in Figure 4.12 and 4.13, two additional components are used within the network. The application monitor provides information to the network advisor. Hereupon, the network advisor influences the packet scheduling at the base station.

4.3.2 Definition of Application-Aware Packet Scheduling

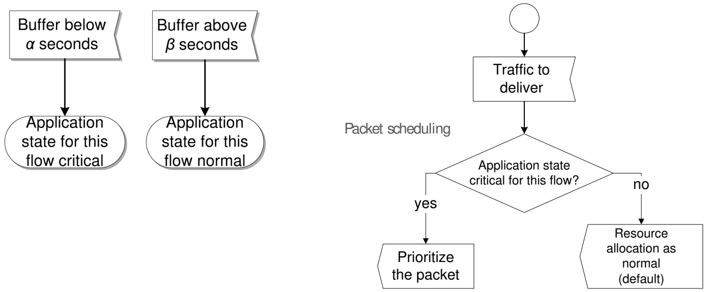
We define two distinct scheduling algorithms for application-aware packet scheduling: (1) buffered playtime scheduling for YouTube and (2) application-aware utility scheduling for multi-application scenarios. The first algorithm prioritizes YouTube on demand according to the buffer level. The second algorithm tries to schedule the applications according to their anticipated QoE value by weighting key quality indicators for applications. Here, the quality indicators are solely based on application information. Further on, we additionally define two reference scheduling algorithms for comparison.

YouTube Scheduling According to the Buffered Playtime

In this approach, the buffered playtime of YouTube is utilized to optimize the user perceived quality for YouTube. The scheduling is done as follows. As illustrated in Figure 4.14(a), as soon as the buffered playtime of one YouTube client falls below a threshold of α seconds, a signaling event is generated by the network advisor. Additionally, if the buffered playtime exceeds a second threshold of β seconds, again a signaling event is generated. In the scheduler, a bearer is tagged as being in a critical state if feedback is received indicating that the buffered playtime is below the threshold α. It is assumed that a user mainly runs a single application. Thus, a prioritization of a bearer only affects a single application as desired. A bearer is tagged as normal if the network advisor indicates that the threshold β is exceeded.

In Figure 4.14(b) the scheduling at the base station is depicted. If the scheduler receives a packet, it checks whether the packet belongs to a bearer in a critical state or not. If the state of the bearer is critical, then the client is prioritized by the scheduler. The scheduler prefers this packet over other users and allocates it to the transmission frame. In all other cases, the packet is passed to the resource allocator as in the normal case which means that the scheduling is done according to a certain fairness metric, which may consider channel quality or service-level QoS parameters.

The proposed scheduling does not follow a proactive approach to optimize the QoE. Only if a QoE degradation is imminent, in spite of the normal scheduling, this approach will prioritize a flow in order to avoid QoE degradation. The advantage of this approach is that the scheduling is done according to the state of the end user application to provide an acceptable quality, and not by the network by maintaining certain QoS levels for the application


Buffer below
α seconds
Application state for this flow critical
Buffer above
β seconds
Application state for this flow normal
Traffic to deliver
Packet scheduling
Application state critical for this flow?
yes
no
Prioritize the packet
Resource allocation as normal (default)

(a) Signaling events.

(b) Application-aware packet scheduling process at base station.

Figure 4.14: *Scheduling process and signaling events for the YouTube buffer-based scheduler.*

Application-Aware Utility Scheduling

The application-aware utility scheduling defines a utility function for scheduling. The utility function returns a value depending on the running applications which can be used to weight different bearers with each other. The approach is similar to a proportional fairness scheduler, as commonly used in mobile communication systems. However, instead of using a priority proportional to the possible throughput, a throughput inversely proportional to the current weighted application quality indicator is used. Hence, if the application is currently in a very good condition, it is assigned a very low priority. However, if the application is in a poor condition, it gets a higher priority. The quality indicator ranges from 1 (poor) to 5 (excellent). The network advisor maps certain application layer information onto this quality indicator by the metrics presented in Table 4.13. For YouTube, the buffered playtime is taken into account. For web browsing, the download time of a web page is monitored and for downloads, the current throughput is measured [155].

Table 4.13: *Mapping of quality indicator to application parameter.*

Quality indicator value	File download throughput [Mbps]	Web browsing page loading time [s]	YouTube buffer level [s]
1	< 0.25	> 5	< 2
2	0.25 - 0.5	3 - 5	2 - 4
3	0.5 - 1	2 - 3	4 - 8
4	1 - 2	1.5 - 2	8 - 16
5	> 2	< 1.5	> 16

Quality indicator value	Skype		
	resolution	frame rate F [fps]	image quality I
1	160x120	*any*	*any*
2	320x240	*any*	*any*
3	640x480	$3 + \frac{F}{35fps} + (2I - 1)$	
4			

Reference Schedulers

For comparison, two different reference schedulers are defined: the *round-robin scheduler* and the *proportional fair scheduler*. The round-robin scheduler assigns each user equal portions of transmission resources in circular order.

In contrast, the proportional fair scheduler addresses both fairness and throughput in the network. This is achieved by assigning each user i at transmission frame f a priority M which is based on the present achievable transmission rate r and the previously achieved overall throughput R.

$$M_i^R(f) = \frac{{r_i}^{\alpha}}{{R_i}^{\beta}}. \tag{4.1}$$

It should be noted that the resulting total bandwidth of the network depends on the scheduling since different assignments lead to different user throughput due to the adaptive modulation and coding of the users.

4.3.3 Evaluation Methodology

In the following, the simulation is described and the application models are defined.

Description of the Simulation

One mobile cell is simulated with a time-discrete event-based simulator for LTE mobile networks. The physical data transmission is performed on the basis of precalculated link-level curves for packet error and goodput from separate simulations with the LTE Downlink Link Level Simulator of the Vienna University of Technology[13]. The simulator implements a complete signal processing chain for the traffic channel. PHY and MAC functions are implemented according to LTE release 8 [156] as specified in [157, 158]. A carrier frequency of 2.5 GHz,

[31] http://www.nt.tuwien.ac.at/ltesimulator

a bandwidth of 5 MHz, and a cell diameter of 250 m have been chosen. The signaling and control channels are simulated as error-free. Based on this physical simulation, a complete system model is implemented with TCP transport protocol and application layer. TCP Cubic with congestion control, error detection and flow control is simulated for each user to obtain realistic scenarios even in overload situations. The propagation model for the data transmission consists of path loss, shadow fading, and multipath fading. Path loss is calculated according to the Winner II urban macro-cell model [159]. Furthermore, the shadow fading decorrelation distance is set to 50 m. The users move around randomly within the cell with a speed of 1 m/s. For this purpose, 200 SNR channel traces have been precalculated since on the fly computation is very time consuming. One SNR channel trace is assigned to each user with a random time offset. The users are able to watch YouTube videos, conduct Skype-like video calls, download files, or surf the Internet. The models are defined in the subsequent subsection.

Only the downlink is considered in this work, since it is assumed that this constitutes the bottleneck of the access network. The transmission is controlled by a packet scheduler. Each user has a packet buffer which is limited in size. The packet scheduler chooses the packets from the user queues according to the scheduling algorithm and passes them to the resource allocator. The resource allocation then selects the appropriate modulation and encoding based on the link-level curves depending on the users channel and places it in the frame.

Modeling of the Investigated Applications

For application-aware scheduling, it is important to model the application accurately. This section describes the modeling of the four investigated applications for the evaluation, namely file download, web browsing, YouTube, and Skype.

File Download The file download is the simplest application in the simulation. It represents the download of a big data file. Therefore, a best-effort transmission over TCP is simulated. The HTTP protocol is not simulated. The size of the downloaded data can be specified by the user. Hence, the download only depends on the simulated physical link and the behavior of the TCP congestion avoidance algorithm.

Web Browsing Web browsing of a user is modeled as follows. A web session consists of the download of a web page followed by an exponentially distributed reading time of a mean of 3 s. The web page itself consists of a main object and several embedded objects. Embedded objects are images, JavaScript code or CSS style sheet instructions. The number of embedded objects, the size of these objects and the size of the main object follow random variables whose distributions are listed in Table 4.14. TCP is used as transport protocol. The web server takes care about the TCP connection handling. The keep-alive timeout for HTTP/1.1 connections is set to 5s based on the values of the default configuration of the Apache web server. Furthermore, no speed or connection limit is set.

Table 4.14: *Web session simulation parameters.*

reading time	neg. exponential: Exp(3s)
volume main object	log-normal: $\ln\mathcal{N}(10\,\mathrm{kB}, 25\,\mathrm{kB})$ $\in [100\,\mathrm{B}, 2\,\mathrm{MB}]$
number of embedded objects	truncated Pareto(scale, shape, max): $Pr(1.1,2,55)$
volume embedded object	log-normal: $\ln\mathcal{N}(8\,\mathrm{kB}, 126\,\mathrm{kB})$ $\in [50\,\mathrm{B}, 2\,\mathrm{MB}]$

YouTube The YouTube Flash Player and a YouTube download server is simulated for YouTube users. The player processes HTTP data to display the YouTube video. In particular, it calculates the current buffered video playtime in seconds. The player may stall if the playtime buffer is empty. The play-out delay after stalling is set to 3 s buffered playtime which is the current value of the YouTube video player. Adaptive video streaming is not considered. The YouTube download server behavior follows [160] with refinements according to own measurements. The download speed is controlled by the server in two phases. The size

S_{ip} of the initial best-effort phase depends on the mean data rate x of the Adobe Flash video. It corresponds to a buffered playtime of 40 s, hence it is calculated as

$$S_{ip} = 40s \cdot x. \tag{4.2}$$

The periodic phase sends data in blocks of 64 kB with a fixed inter-arrival time. The inter-arrival time ΔT_{arr} depends on the target transmission rate which is 125 % of the mean data rate x of the Flash video, but has a maximum of 2.096 s. Therefore it is calculated as

$$\Delta T_{arr} = min(2.096\,s, \frac{64\,kB}{1.25x}). \tag{4.3}$$

Skype-like Video Conferencing The objective is to model a Skype-like application that dynamically adjusts the video parameters depending on the network quality. For this purpose, measurements of Skype from February 2012 serve as a basis for modeling. This section is separated into two different parts, describing the server model for the sending behavior on the one hand, and the self-adapting client behavior on the other hand.

Sending Behavior. We consider only video calls. A call can be started and finished. Hence, it cannot be degraded to a voice call or instant messaging. Additionally, no connection process and buddy list updates take place in background. Only the downlink direction is taken into account. Due to this, the application in the simulation only performs unidirectional transmitting of data directly from a so-called Skype server to a client with UDP transport protocol. Consequently, server and client must be started two times for a realistic video call, as in both directions usually video is transmitted. It is assumed that Skype can adjust three parameters in order to adapt to the current network performance. According to our measurements, the frame rate p_{mfr}, the resolution p_{res}, and the image quality p_{qua} can be adjusted during video calls. The maximum frame rate is set to 35 fps. The image quality is modeled by a factor between 0 and 1. An image quality of 1 corresponds to the best quality. Values below 1 indicate that some kind of lossy compression is used. There are three resolutions available. They are '640x480', '320x240', and '160x120' which result in a data rate ratio of 100 %, 25 %, and 6.25 %. The Skype server periodically sends data blocks to the client. The block inter-arrival time can be described by

$$\Delta T_{ar} = 1/p_{mfr}. \tag{4.4}$$

Next to p_{res} and p_{qua}, the block size additionally depends on a total data rate. It is set to $p_{tdr} = 1.2\,Mbps$. The maximum frame rate is set to $p_{mfr} = 35$. Consequently, the block size for our Skype-like application is described by

$$B_{size} = p_{tdr} \cdot p_{qua} \cdot p_{res}/p_{mfr}. \tag{4.5}$$

Self-Adapting Behavior. In order to instantly react to poor network conditions, the application in our model measures packet delay. A high packet delay is assumed to be caused by high network load. Therefore, the client signals the server that it should enter a *poor network* routine in order to decrease the quality of the video call. This is exactly done if the measured mean packet delay during the last second exceeds the threshold of 75 ms. In the next step, the application resets the parameters for the frame rate and the image quality. Afterwards, periodic requests of the current packet delay serve to estimate the current performance of the connection. In case the packet delay stays below the threshold of 75 ms, the frame rate is increased in our model up to 17 fps. In case the packet delay exceeds the threshold, the image quality is decreased in steps of 0.1 to a minimum of 0.5. Afterwards, the resolution is decreased, while the image quality is set to 1. The algorithm stops if either the minimum resolution and image quality or the targeted frame rate is reached.

In addition, the application in our model also performs a second routine in order to increase the video quality, if possible. This routine runs periodically every 10 s during the complete Skype video call. It is only activated if the previous described routine is not active. If the packet delay is lower than a threshold of 35 ms, the algorithm starts to increase the video encoding until the packet delay exceeds the threshold. The encoding is increased every 500 ms. Our model first tries to increase the resolution, afterwards the image quality is increased and finally the frame rate is increased. If the packet delay exceeds the threshold during this procedure, the encoding is set back to the last working encoding and the routine stops.

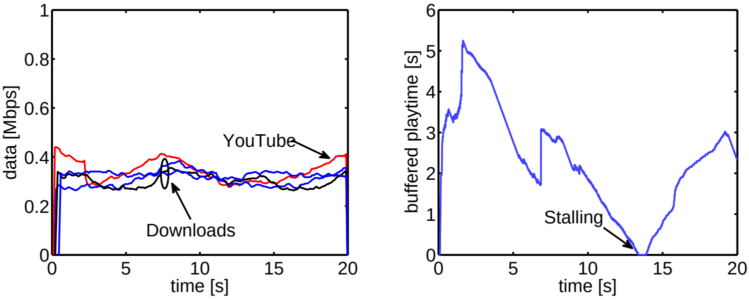
4.3.4 Evaluation of Application-Aware Packet Scheduling for HTTP Downloads and YouTube

In the following three subsections, the evaluation of the two application-aware scheduling algorithms is performed. The first scheduler is described in detail to illustrate the impact on the network performance. For this purpose, an evaluation on YouTube and dowloads is performed in this subsection (Section 4.3.4). In the following subsection, the impact of the scheduler on web browsing sessions and YouTube is presented (Section 4.3.5). For the second scheduler, three different scenarios with multiple users and random arrival requests are evaluated in the last subsection (Section 4.3.6). The aim is to conduct a comprehensive quantification of application-aware scheduling mechanisms.

At first, the YouTube scheduler according to the buffered playtime is investigated with a YouTube user and three HTTP downloads. The YouTube user starts at time instance zero, the downloads start randomly with a delay. The delay is determined according to an exponential distribution with a mean of two seconds. The main performance metric here is the buffered playtime of the YouTube video. The best-effort downloads are responsible for heavy load in the cell which affects the buffered playtime of the YouTube video. The results are compared to the round robin scheduling algorithm.

Reference Scenario

We begin with a brief presentation of the reference situation with round robin scheduling and the four users. Figure 4.15(a) depicts the throughput of the four users. The YouTube user is indicated by the red curve. Download users are shown



(a) Throughput

(b) Buffered playtime of YouTube video

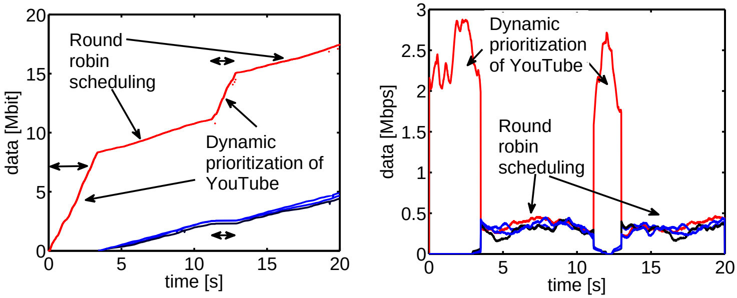
Figure 4.15: *Round robin scheduler with three download users and one YouTube user*

in blue color. On the x-axis the transmitted data in Mbps is shown. The y-axis shows the simulation time in seconds. The figure shows that the throughput is almost equal for all users and will only be influenced due to the different transmission channel conditions of the users since they are moving. Figure 4.15(b) shows the resulting buffered playtime of the YouTube video over the simulation time. The sharp increase of the buffer at 7 s is due to the video encoding since there is a small period with very low encoding rate from 5 to 6 s of video playtime. At 13 s the buffer is empty. The video begins a buffering period, and the user experiences video stalling.

Buffered Playtime Scheduler

In Figure 4.16 the same scenario is depicted as in the previous one but with the buffered playtime scheduler which dynamically prefers the YouTube video in the case of low YouTube buffer.

The first sub-figure of Figure 4.16 presents the cumulative downloaded data of the downloads and the YouTube video player. Together with Figure 4.16(b) the



(a) Cumulative data

(b) Throughput buffered playtime scheduler

Figure 4.16: *Buffered playtime scheduler with three download users and one YouTube user*

difference between the buffered playtime scheduler to the round robin scheduler is depicted. The YouTube flow is prioritized at the beginning and for 1.6 s at about 11 s due to the scheduling strategy. Almost no data is transferred at these time periods to the download users since YouTube is using nearly the whole bandwidth. Outside these time periods, the data is equally scheduled among the users as in round robin strategy. Figure 4.17 shows the corresponding buffered playtime. The buffer level is always greater than zero, and no stalling occurs. The thresholds for the prioritization are visible: if the buffer level is higher than 15 s, round robin scheduling is used. Since, in this scenario with four users, the throughput during the round robin phase is not sufficient, the buffer level decreases afterwards. If the buffer level falls below 10 s playback time, YouTube is prioritized again.

For quantifying the YouTube QoE, concrete mapping functions, depending on the length of stalling and the ratio of stalling, are proposed in literature. According to [161], one stalling already results in a QoE degradation from MOS 5 to 3.2 if the stalling length is 3 s until the flash player will restart the video playback. Another buffering period would further decrease the MOS value from 3.2 to 2.5. Contrary to YouTube, the QoE of file downloads is more robust. Especially for

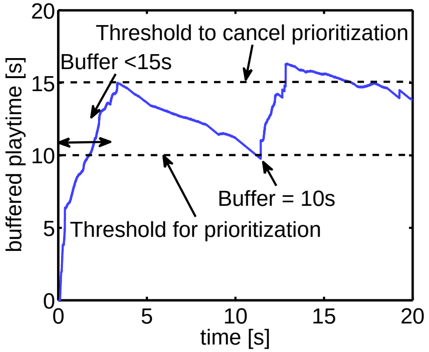


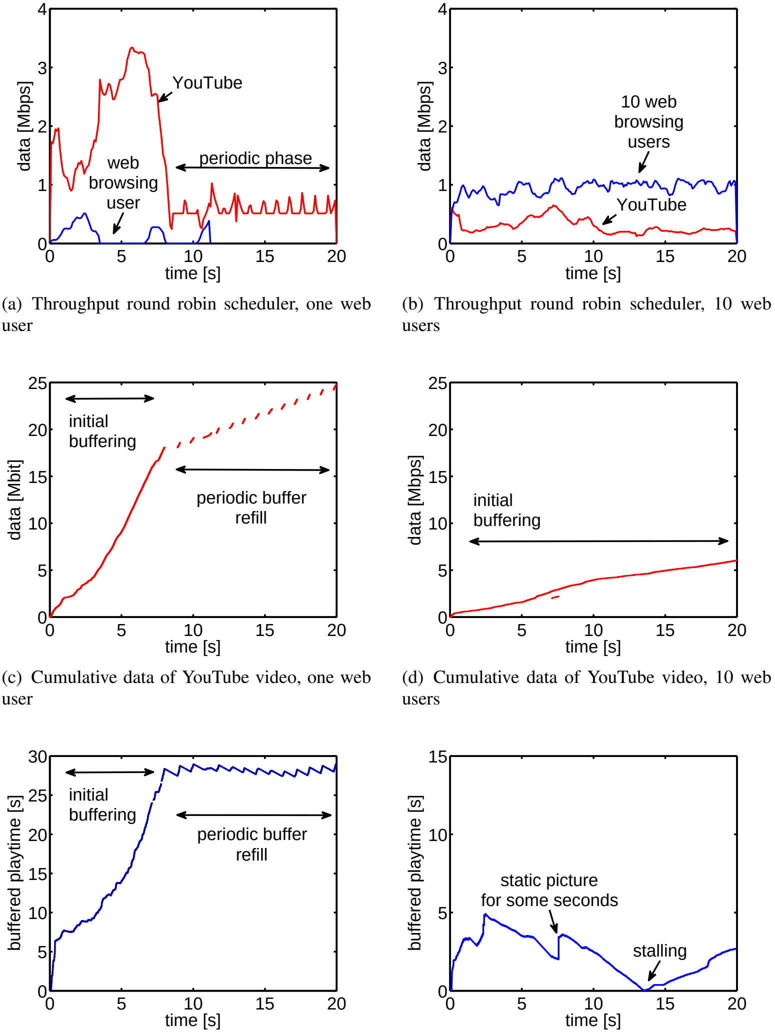
Figure 4.17: *Buffered playtime of YouTube video with buffered playtime scheduler.*

long downloads a small delay can be tolerated [155]. In this case, the download time of the downloads increase by 3.8 to 5.3 s per download depending on the channel conditions of the YouTube user for the two prioritization periods.

4.3.5 Evaluation of Application-Aware Packet Scheduling for Web Browsing Sessions and YouTube

In this subsection, web browsing users are simulated together with one YouTube user. A web session of a web user is defined as described in the simulation section, cf. Table 4.14.

Figure 4.18(a),(c),(e) show results for one web user and one YouTube user. Figure 4.18(a) shows in red color the throughput of the user who is watching the YouTube video and in blue color the web user throughput. The web user is watching three web pages at 7 s, and 10 s. The web traffic is influencing the YouTube throughput: the YouTube throughput is decreasing while the web traffic



(a) Throughput round robin scheduler, one web user

(b) Throughput round robin scheduler, 10 web users

(c) Cumulative data of YouTube video, one web user

(d) Cumulative data of YouTube video, 10 web users

(e) Buffered playtime of YouTube video, one web user

(f) Buffered playtime of YouTube video, 10 web users

Figure 4.18: *Web browsing with one YouTube video, one web user on the left, 10 web users at the right column.*

is increasing during the reading time of the web user. Figure 4.18(c) shows the corresponding accumulated data during the simulation time of the YouTube video only. The two download phases can be seen. At the beginning, YouTube is doing an initial buffering. Afterwards, there is a periodic buffer refill which is also reflected by the throughput in Figure 4.18(a). With one web user the YouTube video time buffer remains stable over the whole simulation time which is depicted in Figure 4.18(e). After the initial buffering, here, the buffer is kept at about 27 s.

Now, Figure 4.18(b),(d),(f) show the situation with 10 web users and round robin scheduler. The blue curve shows the throughput of all web users. The red curve shows the throughput of the YouTube user. The YouTube throughput decreases to about 300 kbit/s - 500 kbit/s due to the round robin scheduling which treats all TCP flows of the users equally. Figure 4.18(f) shows that the YouTube player is not even able to complete the initial best-effort buffering phase. With 10 users, the buffered playtime in Figure 4.18(f) remains below 5 s and stalls again at 13 s.

We now show the buffer progress with a buffered playtime scheduler which signals the current buffer level to the scheduler. Figure 4.19 contains three curves showing the buffer level over time for different scheduler settings. The curves are evaluated for 7 web users in parallel to the YouTube video. The round robin scheduler is included for comparison. For the top blue curve the scheduler is set to the same parameters as in the download scenario with buffered playtime scheduler. If the video time buffer is below 10 s the YouTube flow is strictly prioritized. At a threshold of 15 s round robin strategy is used until it falls below the critical 10 s. In this scenario, the buffer is not significantly decreasing after achieving the 15 s of buffered playtime. A smooth video playback is possible without stalling since the initial prioritization is enough for initially filling the buffer. The initial buffer level is able to compensate the variable encoding of the video for the whole simulation time. Note, this is video specific and depends on the encoding of the video. If the setting of the buffered playtime scheduler is changed to 9 and 10 s as thresholds, Figure. 4.19 shows that due to the variable encoding the critical threshold is reached very often at the beginning. The second scheduler setting has the advantage that the transmissions of web users are delayed for a shorter time period. However, users are more frequently delayed.

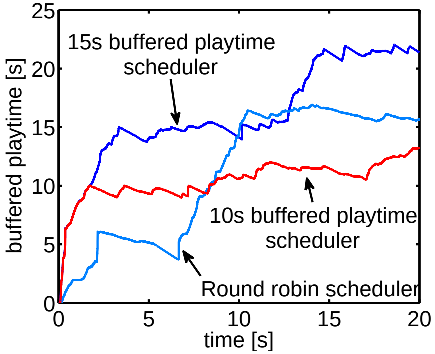


Figure 4.19: *Buffered playtime with different schedulers for 7 web users.*

4.3.6 Multi-Application Scenario: Application-Aware Utility Scheduling According to Quality Indicators

After the presentation of the YouTube scheduling, a comprehensive statistical study of the second application-aware scheduling approach follows. This evaluation considers three scenarios with a different number of users and multiple applications. In all three scenarios, users of YouTube, download users, web browsing users, and Skype users are simulated. As a reference scheduler, the proportional fair scheduler is used to obtain realistic results. In the following, the performance of the scheduler is statistically evaluated with the aim to compare the benefit and the impact of the approach on the applications.

Scenario Description

For this evaluation, 100 runs are conducted per scheduler and scenario. *Scenario I*, *II*, and *III* represent different situations in a mobile cell. In the first scenario, the cell is only slightly loaded. In the second scenario, the cell is crowded. In the third scenario, the system is overloaded. The Skype users, download users, and web browsing users start directly at the beginning of the simulation. The starting time of the YouTube users is calculated from a uniform distribution between second 5 and second 20. The reason therefore is due to the outcome of Section 4.2, synchronous vs. asynchronous start of YouTube videos. For each user one video out of 10 videos is randomly chosen. The system is simulated for 100 s. *Scenario I* simulates 19 users. There are 7 Skype users, two download users, four web browsing users, and 6 YouTube users. *Scenario II* consists of 25 users. There are 8 Skype users, three download users, 6 web browsing users, and 8 YouTube users. *Scenario III* simulates 34 users. There are 10 Skype users, four download users, 10 web browsing users, and 10 YouTube users.

Performance Metrics

The performance evaluation is based on carefully chosen application parameters as performance metrics. The application parameter provide a high correlation to the user-perceived quality:

- For **YouTube** the mean stalling probability is used as performance metric, since stalling is the main factor for a QoE degradation [161].
- For **web browsing** the download time of the content is chosen [155].
- For **file downloads** the amount of downloaded data is considered [155].

Skype is not considered in detail. As an adaptive application, it determines, in itself, the best video quality settings. Due to the adaptiveness the user QoE must be considered only to a limited extent in reality. Thus, Skype is considered as an additional background traffic within the network.

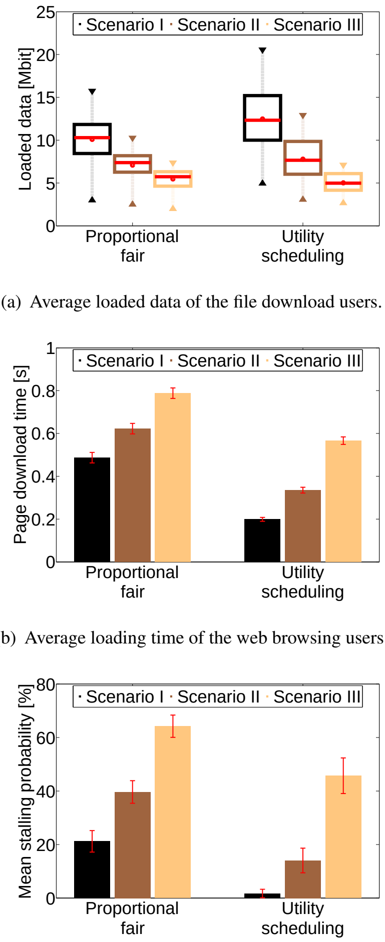
Impact on File Downloads

Figure 4.20(a) shows the amount of downloaded data of the users for the two different schedulers at all three scenarios. The red line indicates the median of the loaded data, the red dot indicates the average loaded data. The box shows the 40 % quantile of the results. The complete range is indicated by the dotted line and the triangles. The results show, on the one hand, a tendency with respect to the different scenarios. On the other hand, there are significant differences between the two scheduling strategies. With a higher number of users in the network, the amount of data that the users are able to download decreases. Accordingly, the 40 % quantiles and the entire range of the results become smaller indicating less variance. When comparing the two scheduler strategies in detail, a different result shows up depending on the employed scenario. In the slightly loaded *Scenario I*, the utility scheduler achieves an improvement in terms of downloaded data. On average, about 13 Mbit of data can be downloaded. These are about 3 Mbit more in comparison to the proportional fair scheduling with 10 Mbit. In the two other scenarios, however, the scheduler can not significantly improve the situation for the downloads. This is mainly due to the increased load in the network. As mentioned before the situation is deteriorating with increasing load, which results in fewer opportunities for the scheduling.

While an improvement for the downloads is desirable, it should be noted that a lower download throughput might be beneficial for other users. File downloads are flexible applications and a certain delay or a low download throughput might be tolerated. From a QoE perspective, assigning a lot of bandwidth is not necessary or even a waste of resources. Therefore, a moderate metric with respect to the file downloads was used in the utility function in order to gain resources for other applications as explained in the next section.

Impact on Web Browsing

Figure 4.20(b) shows the average page download time of the web browsing users between second 30 and 80 of the simulation. Only the transfer time within the mobile network is considered. Delays by the web server or the transmission over



(a) Average loaded data of the file download users.

(b) Average loading time of the web browsing users.

(c) Average stalling probability of the YouTube users.

Figure 4.20: *Evaluation of the amount of downloaded data, the page load time of web browsing users, and the stalling probability of YouTube.*

the Internet are not included here in this evaluation. The 95 % confidence intervals are indicated by the red lines on top of the bars. Again, if the number of active users inside the network increases, the page download time increases, too. However, the respective times of the scenarios differ with the two schedulers. The utility scheduler provides loading times of 0.2 s, 0.3 s and 0.6 s, which are about 0.3 s better than the times of the proportional fair scheduler for all scenarios. The result confirms that the utility scheduler takes the download time of web content into account and optimizes the network accordingly.

Impact on YouTube Video Streaming

In Figure 4.20(c), the average stalling probability for YouTube users is shown. Stalling is defined as at least one interruption between second 30 and 80 of the simulation. While the stalling probability increases with a larger number of users in the system, the proportional fair scheduler achieves the worst results in all three scenarios with stalling probabilities of 21 %, 40 %, and 64 %. The utility scheduler, in turn, is successful in improving the situation. In the case of YouTube, this is an improvement in *Scenario I* by a factor of seven.

Generally considered, the evaluation of the utility scheduler in comparison to the proportional fair scheduler demonstrated that application awareness can improve the overall situation of the applications in the network. However, there are different results for different applications that result from the definition of the utility function. There are applications that can tolerate a highly varying bandwidth according to such a utility scheduling. For example, the YouTube application has achieved useful results because the latter can tolerate dynamic changes in the bandwidth due to the video buffer. A constant UDP live video streaming in contrast does not favor such a scheduling.

4.4 Lessons Learned

This section concludes the chapter and presents its most important findings. The chapter provides implementations of application-aware resource management for different types of access networks. First, the impact of resource management concepts in wireless mesh networks is evaluated on the example of YouTube video streaming. Second, the impact of using application information for scheduling decisions is examined on the downlink within cellular networks on the user perceived quality. In particular, different applications such as web browsing, file downloads, progressive video streaming, and Skype video conferencing are considered in this case. The focus was on the quantification of benefits with respect to different applications.

The evaluation for wireless mesh networks was performed in a testbed. The necessary application-aware components have been implemented and integrated into the network. In the chapter, detailed implementation details are given for the realization.

It has been found that

1. an application-aware resource management can efficiently increase both, the resource utilization as well as the perceived quality. This applies for wireless mesh as well as for cellular access networks.
2. more YouTube users can be supported in an access network due to more efficient resource utilization with application-aware resource management.
3. the load can be balanced in a beneficial way on different Internet gateways for mesh networks, if multiple gateways are present in the access network.
4. a prioritization of IP flows can significantly improve QoE of users by dynamically changing the prioritization according to the buffered playtime.
5. on application side, the second proposed resource management option, service control, allows for changing the resolution of video streaming. A lower resolution results in a lower bandwidth with only a minor degradation of the QoE [120]. Thus, if no more resources are available or the

provider wants to reduce its operational expenditure, service control is the best choice.

6. the best trade-off between QoE and resource efficiency can be achieved using a combined control approach. Our findings here are that a strategy using a moderate mix of network and service control helps to keep the QoE on a high level without using too much resources in our test network and thus, reducing the energy consumption and the operational expenditure.

The investigations in the cellular access networks were carried out by simulation. A 3GPP LTE system level simulator has been designed in conjunction with a detailed model of different applications and TCP as well as wireless channel models. First, a scheduling algorithm is proposed that dynamically prioritizes users against other users if a QoE degradation is imminent. The prioritization is done in a proactive way according to the buffered playtime of the YouTube video player. Second, a comprehensive statistical study is conducted of an application-aware utility scheduling approach that directly integrates quality indicators for various different applications within the resource management.

For cellular networks applies that

1. it is necessary to adapt QoS mapping in the scheduler on the instantaneous requirements on the client side in order to guarantee good QoE at the end user.
2. for a YouTube video, a buffering period can be avoided at the expense of download time in cellular networks. Especially for long downloads, the overall QoE is improved since an increase of the download time can be tolerated for them and does not negatively influence the QoE.
3. a very flexible scheduling can be carried out due to the buffering of the video content. This can be exploited to obtain a multi-user diversity. However, the signaling of the buffer level to the scheduling entity is required.
4. an improvement can be expected of the overall user-perceived quality in case of application-aware scheduling for the investigated scenarios with multiple applications and services.

The results quantify the trade-off between the complexity of providing application information at network layer and the gain in terms of QoE.

5 Conclusion

Future access networks have to satisfy a large number of heterogeneous applications and services. This applies in addition to the challenges that they must be cost-effective, have to offer high quality Internet, and provide fast connections. A specialized resource management may help in many of these cases and can create a win-win situation for both users and the network.

In this monograph, we studied different new resource management approaches for performance optimization and network resource efficiency in access networks. The investigated approaches belong to different communication layers and meet different objectives. In the end, this work provides recommendations for network operators how a resource management for different types of networks and objectives needs to look like and what benefit can be expected in relation to the required complexity overhead.

Due to the increasing volume of traffic, the ambitious performance objectives of the network operators, and the cost pressure between providers, a management and efficient utilization of network resources is becoming increasingly important. For mobile communication networks, this means that the frequency resources must be carefully allocated according to various objectives. However, resources are usually aggressively reused throughout the network in order to exploit the limited frequency spectrum. This results in current mobile OFDMA-based systems in interference between neighboring cells if the same frequency is used. Consequently, frequency management and resource partitioning approaches are important means to enable efficient and high quality networks.

Current related works focus mainly on the downlink direction. The interference in the uplink and downlink however have considerably different characteristics. In the uplink, each user causes interference. Therefore, the uplink needs to be evaluated with suitable means. We investigated in this work distributed resource allocation approaches for the uplink of an OFDMA mobile communication network. The focus was on the identification of approaches that are able to mitigate inter-cell interference. We propose different coordination mechanisms for resource allocation which lead to a more efficient resource usage and consequently a higher number of supported users per cell. New combinations were compared with conventional schemes that do not use restriction or any resource partitioning such as universal frequency reuse.

Another challenge is that due to the heterogeneous applications, traffic characteristics significantly change in the networks. Many previous works consider the downlink with saturated traffic conditions, i.e., the connections are always busy with traffic. With respect to current traffic measurements however non-saturated links seem to be the more realistic choice for the uplink since the current traffic consists normally of constantly-recurring TCP acknowledgements, HTTP requests, or burst-wise video and voice traffic.

We proposed a new resource allocation algorithm that is based on the intelligent choice of a modulation and coding scheme for a non-saturated uplink. Based on the previous results, the scheme a) was developed to be less load-dependent, b) was developed to be interference reducing, and c) uses frequency partitioning. It is based on the findings that there is a non-linear relationship between transmission power and coding scheme for the same amount of data for the uplink. Thus, choosing a lower modulation and coding scheme yields to a significant saving in transmission power and generated interference. The approach enables a more efficient use of resources.

Not only on the physical layer, it is important to efficiently utilize resources. At the application layer, a dynamic and intelligent resource management is also necessary to meet the diverse requirements of today's applications. Compared to resource management on lower layers, the goal is moved from the optimization of network parameters towards the direct consideration of the user. This is based on the fundamental insight that users care about the resulting application quality and are not interested in technical parameters as long as the application runs smoothly and in high quality as expected. The objective for application-aware resource management is consequently to address directly application needs in order to optimize the user-perceived quality through resource management options.

Application-aware resource management is the approach to tailor access networks to have characteristics beneficial for the running applications and services. This is achieved through the integration of key performance indicators from the application layer within the resource management. In this work, an application-aware framework is defined consisting of several components such as application monitoring, network monitoring, decision entity, and resource management tool. Application monitoring is not sufficiently studied in the current literature. Thus, a dynamic monitoring was developed and evaluated on the example of YouTube video streaming. It has been found that the monitoring is able to accurately estimate the time when the YouTube player is stalling and consequently a quality degradation is imminent. Furthermore, the monitoring is lightweight and easy to install while it provides valuable information for the network operator. If users run it, both parties may greatly benefit as the operator gets information about the application status which can be used to allocate resources and improve user QoE.

We provided implementation examples for application-aware resource management for different types of access networks. First, the impact was studied in wireless mesh networks for YouTube video streaming. Second, the impact of using application information for scheduling decisions is examined in the downlink within cellular networks on the user-perceived quality. The focus was on the quantification of benefits with respect to different applications. The findings show that an application-aware resource management can efficiently increase both, the resource utilization as well as the perceived quality. This applies for wireless mesh as well as for cellular access networks. Further on, more YouTube users can be supported in an access network due to more efficient resource utilization.

In terms of cellular networks, this paper presents findings on YouTube streaming video. An unwanted video stalling, i.e. an interruption of the playback of YouTube videos, can be avoided by application-aware scheduling at the expense of download time. Especially for long downloads, the overall QoE is improved since an increase of the download time can be tolerated and does not negatively influence the QoE. Furthermore, an additional evaluation quantifies the impact of application-aware resource management in multi-application scenarios with varying load.

In the course of this monograph, we evaluated different resource management approaches for different types of access networks. The results fit into the broad field of resource management research dealing with technical, architectural, and performance issues in networks. The technical ones comprise for example the impact of other layer resource management mechanisms on the application performance. Further on, how dynamic network applications can adapt their demands to network capabilities. Second, architectural questions arise especially for other types of networks than access networks. The challenge here is in particular the component placement of decision unit and application monitoring. Third, performance questions have to be answered for all kinds of networks. Information exchange required for intelligent future network control implies additional overhead and therewith costs in terms of bandwidth and processing. Therefore, it is important to determine the amount of information that is necessary to improve user perceived quality without degrading network performance by excessive signaling.

The solution of these problems will result in implementation instructions, the specification of new communication interfaces, and operation guidelines, which will be of high importance for network and application interaction in the future Internet. This work provides a first step towards the understanding of the potential and the operation of this new network paradigm in access networks.

The thesis covers methodologies such as simulation and practical implementation. Use case-driven scenarios for wireless mesh networks are presented. The testbed evaluations illustrate the benefits and demonstrate the feasibility of the approach. The simulative analysis in contrast showed the potential of resource management approaches in different scenarios under different load for cellular networks. The investigated mechanisms enable a more efficient use of network resources as well as a possible improvement of the perceived quality for the users.

Bibliography and References

— Bibliography of the Author —

— Journals and Book Chapters —

[1] F. Wamser, R. Pries, D. Staehle, K. Heck, and P. Tran-Gia, “Traffic characterization of a residential wireless Internet access,” *Special Issue of the Telecommunication Systems (TS) Journal*, vol. 48: 1-2, May 2010.

[2] F. Wamser, D. Mittelstädt, D. Staehle, and P. Tran-Gia, “Impact of Electrical and Mechanical Antenna Downtilt on a WiMAX System with Fractional Frequency Reuse,” *FREQUENZ - Journal of RF-Engineering and Telecommunications*, vol. September/October, Sep. 2010.

[3] F. Wamser, D. Hock, M. Seufert, B. Staehle, R. Pries, and P. Tran-Gia, “Using Buffered Playtime for QoE-Oriented Resource Management of YouTube Video Streaming,” *Transactions on Emerging Telecommunications Technologies*, vol. 24, Apr. 2013.

[4] D. Hock, F. Wamser, M. Seufert, R. Pries, and P. Tran-Gia, “OC^{2}E^{2}AN: Optimized Control Center for Experience Enhancements in Access Networks,” *PIK - Praxis der Informationsverarbeitung und Kommunikation*, vol. 36, Feb. 2013.

[5] T. Hoßfeld, F. Liers, R. Schatz, B. Staehle, D. Staehle, T. Volkert, and F. Wamser, “Quality of Experience Management for YouTube: Clouds, FoG and the AquareYoum,” *PIK - Praxis der Informationverarbeitung und -kommunikation (PIK)*, Aug. 2012.

[6] B. Staehle, F. Wamser, M. Hirth, D. Stezenbach, and D. Staehle, "AquareYoum: Application and Quality of Experience-Aware Resource Management for YouTube in Wireless Mesh Networks," *PIK - Praxis der Informationsverarbeitung und Kommunikation*, 2011.

— Conference Papers —

[7] F. Wamser, L. Iffländer, T. Zinner, and P. Tran-Gia, "Implementing Application-Aware Resource Allocation on a Home Gateway for the Example of YouTube," in *Mobile Networks and Management, Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering*, Würzburg, Germany, Sep. 2014.

[8] F. Wamser, T. Zinner, J. Z. Zhu, and P. Tran-Gia, "Dynamic Bandwidth Allocation for Multiple Network Connections: Improving User QoE and Network Usage of YouTube in Mobile Broadband," in *ACM SIGCOMM Capacity Sharing Workshop (CSWS 2014)*, Chicago, IL, USA, Aug. 2014.

[9] F. Wamser, B. Staehle, R. Pries, D. Stezenbach, S. Deschner, and D. Staehle, "YouTube QoE-Aware Gateway Selection in Future Wireless Networks," in *EuroView2010*, Würzburg, Germany, Aug. 2010.

[10] R. Pries, F. Wamser, D. Staehle, K. Heck, and P. Tran-Gia, "On Traffic Characteristics of a Broadband Wireless Internet Access," in *Next Generation Internet Networks 2009 (NGI 2009)*, Aveiro, Portugal, Jul. 2009.

[11] ——, "Traffic Measurement and Analysis of a Broadband Wireless Internet Access," in *IEEE VTC Spring 09*, Barcelona, Spain, April 2009.

[12] M. Jarschel, F. Wamser, T. Höhn, T. Zinner, and P. Tran-Gia, "SDN-based Application-Aware Networking on the Example of YouTube Video Streaming," in *2nd European Workshop on Software Defined Networks (EWSDN 2013)*, Berlin, Germany, Oct. 2013.

[13] T. Zinner, M. Jarschel, A. Blenk, F. Wamser, and W. Kellerer, "Dynamic Application-Aware Resource Management Using Software-Defined Networking: Implementation Prospects and Challenges," in *IFIP/IEEE International Workshop on Quality of Experience Centric Management (QC-Man)*, Krakow, Poland, May 2014.

[14] F. Wamser, D. Mittelstädt, and D. Staehle, "Soft Frequency Reuse in the Uplink of an OFDMA Network," in *IEEE VTC Spring 10*, Taipei, Taiwan, May 2010.

[15] F. Wamser, D. Mittelstädt, D. Staehle, and P. Tran-Gia, "Advanced Interference Mitigation with Frequency Reuse Schemes in the IEEE 802.16m Uplink," in *ACM MSWIM 2010*, Bodrum, Turkey, Oct. 2010, pp. 132–139.

[16] F. Wamser, D. Hock, M. Seufert, R. Pries, and P. Tran-Gia, "Performance Optimization in Access Networks Using a Combined Control Strategy," in *Euroview 2012*, Würzburg, Germany, Jul. 2012.

[17] B. Staehle, F. Wamser, S. Deschner, A. Blenk, D. Staehle, O. Hahm, N. Schmittberger, and M. Günes, "Application-Aware Self-Optimization of Wireless Mesh Networks with AquareYoum and DES-SERT," in *Euroview 2011*, Würzburg, Germany, Aug. 2011.

[18] B. Staehle, M. Hirth, R. Pries, F. Wamser, and D. Staehle, "Aquarema in Action: Improving the YouTube QoE in Wireless Mesh Networks," in *Baltic Congress on Future Internet Communications (BCFIC)*, Riga, Latvia, Feb. 2011.

[19] B. Staehle, F. Wamser, R. Pries, D. Staehle, C. Mannweiler, A. Klein, J. Schneider, and H. D. Schotten, "Application- and Context-Aware Radio Resource Management for Future Wireless Networks," in *EuroView2010*, Würzburg, Germany, Aug. 2010, pp. 143–144.

[20] B. Staehle, M. Hirth, R. Pries, F. Wamser, and D. Staehle, “YoMo: A YouTube Application Comfort Monitoring Tool,” in *New Dimensions in the Assessment and Support of Quality of Experience for Multimedia Applications*, Tampere, Finland, Jun. 2010.

[21] M. Hirth, B. Staehle, F. Wamser, R. Pries, and D. Staehle, “QoE Prediction for Radio Resource Management,” in *The 6th International ICST Conference on Testbeds and Research Infrastructures for the Development of Networks & Communities (TridentCom 2010)*, Berlin, Germany, May 2010.

[22] F. Wamser, S. Deschner, T. Zinner, and P. Tran-Gia, “Investigation of Different Approaches for QoE-Oriented Scheduling in OFDMA Networks,” in *Mobile Networks and Management, Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering Volume 125*, Cork, Irland, Sep. 2013, pp. 172–187.

[23] F. Wamser, D. Staehle, J. Prokopec, A. Maeder, and P. Tran-Gia, “Utilizing Buffered YouTube Playtime for QoE-oriented Scheduling in OFDMA Networks,” in *International Teletraffic Congress (ITC)*, Krakow, Poland, Sep. 2012.

— Technical Reports —

[24] B. Staehle, M. Hirth, F. Wamser, R. Pries, and D. Staehle, “YoMo: A YouTube Application Comfort Monitoring Tool,” University of Würzburg, Tech. Rep. 467, Mar. 2010.

— Software Demonstrations —

[25] F. Wamser, D. Hock, M. Seufert, T. Zinner, and P. Tran-Gia, “Demonstrating the Benefit of Joint Application and Network Control Within a Wireless Access Network,” 32nd IEEE International Conference on Computer Communications (INFOCOM 2013), Turin, Italy, Apr. 2013.

[26] D. Hock, F. Wamser, M. Seufert, R. Pries, and P. Tran-Gia, “OC^{2}E^{2}AN: Optimized Control Center for Experience Enhancements in Access Networks,” Conference on Networked Systems (NetSys 2013), Stuttgart, Germany, Mar. 2013.

[27] F. Wamser, D. Hock, M. Seufert, R. Pries, and P. Tran-Gia, “Performance Optimization in Access Networks Using a Combined Control Strategy,” Euroview 2012, Würzburg, Germany, Jul. 2012.

[28] B. Staehle, F. Wamser, S. Deschner, A. Blenk, D. Staehle, O. Hahm, N. Schmittberger, and M. Günes, “Application-Aware Self-Optimization of Wireless Mesh Networks with AquareYoum and DES-SERT,” Euroview 2011, Würzburg, Germany, Aug. 2011.

[29] B. Staehle, F. Wamser, M. Hirth, D. Stezenbach, and S. D. und Dirk Staehle, “AquareYoum: Application and Quality of Experience-Aware Resource Management for YouTube in Wireless Mesh Networks,” winner of KuVS Communication Software Award, Kiel, Germany, Mar. 2011.

[30] F. Wamser, B. Staehle, R. Pries, D. Stezenbach, S. Deschner, and D. Staehle, “YouTube QoE-Aware Gateway Selection in Future Wireless Networks,” Euroview 2010, Würzburg, Germany, Aug. 2010.

— General References —

[31] M. Abaii, G. Auer, F. Bokhari, M. Bublin, E. Hardouin, O. Hrdlicka, G. Mange, M. Rahman, and P. Svac, “IST-4-027756 WINNER II; Interference avoidance concepts,” June 2007.

[32] M. Bublin, E. Hardouin, O. Hrdlicka, I. Kambourov, R. Legouable, M. Olsson, S. Plass, P. Skillermark, and P. Svac, “IST-4-027756 WINNER II; Interference averaging concepts,” June 2007.

[33] M. Bublin, T. Clessienne, E. Hardouin, O. Hrdlicka, B. Hunt, G. Mange, M. Olsson, S. Plass, K. Roberts, P. Skillermark, P. Svac, and X. Wei, "IST-4-027756 WINNER II; Smart antenna based interference mitigation," June 2007.

[34] WINNER II Deliverable, "IST-4-027756 WINNER II; WINNER II System Concept Description," Nov. 2007.

[35] A. S. Hamza, S. S. Khalifa, H. S. Hamza, and K. Elsayed, "A survey on inter-cell interference coordination techniques in OFDMA-based cellular networks," *Communications Surveys & Tutorials, IEEE*, vol. 15, no. 4, pp. 1642–1670, 2013.

[36] G. Boudreau, J. Panicker, N. Guo, R. Chang, N. Wang, and S. Vrzic, "Interference coordination and cancellation for 4g networks," *Communications Magazine, IEEE*, vol. 47, no. 4, pp. 74–81, 2009.

[37] Y. Xiang, J. Luo, and C. Hartmann, "Inter-cell interference mitigation through flexible resource reuse in OFDMA based communication networks," in *13th European Wireless Conference EW2007*, 2007.

[38] M. Necker, "A novel algorithm for distributed dynamic interference coordination in cellular OFDMA networks," Ph.D. dissertation, University of Stuttgart, 2009.

[39] WiMAX Forum, "WiMAX Forum Mobile System Profile; Release 1.0 Approved Specification," May 2007.

[40] IEEE, "IEEE Standard for local and metropolitan area networks; Part 16: Air Interface for Broadband Wireless Access Systems," May 2009.

[41] 3rd Generation Partnership Project., "3GPP TS 36.300 V9.10.0.; Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Overall description; Stage 2," January 2013, v9.10.0.

[42] WiMAX Forum, "Mobile WiMAX Part I: A Technical Overview and Performance Evaluation," February 2006.

[43] M. Sternad, T. Ottoson, A. Ahlén, and A. Svensson, "Attaining both Coverage and High Spectral Efficiency with Adaptive OFDM Downlinks," in *Proceedings of IEEE Vehicular Technology Conference, VTC2003-Fall*, Orlando, Florida, October 2003.

[44] Huawei and 3rd Generation Partnership Project, "3GPP R1-050507; Soft Frequency Reuse Scheme for UTRAN LTE," Athens, Greece, May 2005.

[45] C. Y. Wong, R. S. Cheng, K. B. Lataief, and R. D. Murch, "Multiuser ofdm with adaptive subcarrier, bit, and power allocation," *Selected Areas in Communications, IEEE Journal on*, vol. 17, no. 10, pp. 1747–1758, 1999.

[46] W. Rhee and J. M. Cioffi, "Increase in capacity of multiuser ofdm system using dynamic subchannel allocation," in *Vehicular Technology Conference Proceedings, 2000. VTC 2000-Spring Tokyo. 2000 IEEE 51st*, vol. 2. IEEE, 2000, pp. 1085–1089.

[47] J. Jang and K. B. Lee, "Transmit power adaptation for multiuser ofdm systems," *Selected Areas in Communications, IEEE Journal on*, vol. 21, no. 2, pp. 171–178, 2003.

[48] H. Yin and H. Liu, "An efficient multiuser loading algorithm for ofdm-based broadband wireless systems," in *Global Telecommunications Conference, 2000. GLOBECOM'00. IEEE*, vol. 1. IEEE, 2000, pp. 103–107.

[49] D. Kivanc, G. Li, and H. Liu, "Computationally efficient bandwidth allocation and power control for OFDMA," *Wireless Communications, IEEE Transactions on*, vol. 2, no. 6, pp. 1150–1158, 2003.

[50] J. Gross, H. Karl, F. H. Fitzek, and A. Wolisz, "Comparison of heuristic and optimal subcarrier assignment algorithms." in *International Conference on Wireless Networks*, 2003, pp. 249–255.

[51] Z. Shen, J. G. Andrews, and B. L. Evans, "Adaptive resource allocation in multiuser ofdm systems with proportional rate constraints," *Wireless Communications, IEEE Transactions on*, vol. 4, no. 6, pp. 2726–2737, 2005.

[52] G. Yu, Z. Zhang, Y. Chen, and P. Qiu, "An efficient resource allocation algorithm for OFDMA systems with multiple services," *IEEE Globecom '06*, 2006.

[53] T.-D. Nguyen and Y. Han, "A proportional fairness algorithm with qos provision in downlink OFDMA systems," *Communications Letters, IEEE*, vol. 10, no. 11, pp. 760–762, 2006.

[54] J. Cai, X. Shen, and J. W. Mark, "Downlink resource management for packet transmission in ofdm wireless communication systems," *Wireless Communications, IEEE Transactions on*, vol. 4, no. 4, pp. 1688–1703, 2005.

[55] A. Pokhariyal, K. I. Pedersen, G. Monghal, I. Z. Kovacs, C. Rosa, T. E. Kolding, and P. E. Mogensen, "Harq aware frequency domain packet scheduler with different degrees of fairness for the utran long term evolution," in *Vehicular Technology Conference, 2007. VTC2007-Spring. IEEE 65th.* IEEE, 2007, pp. 2761–2765.

[56] M. Andrews, K. Kumaran, K. Ramanan, A. Stolyar, P. Whiting, and R. Vijayakumar, "Providing quality of service over a shared wireless link," *Communications Magazine, IEEE*, vol. 39, no. 2, pp. 150–154, 2001.

[57] A. K. Khattab and K. M. Elsayed, "Opportunistic scheduling of delay sensitive traffic in OFDMA-basedwireless," in *proceedings of the 2006 International Symposium on World of Wireless, Mobile and Multimedia Networks.* IEEE Computer Society, 2006, pp. 279–288.

[58] H. Lei, C. Fan, X. Zhang, and D. Yang, "Qos aware packet scheduling algorithm for OFDMA systems," in *Vehicular Technology Conference, 2007. VTC-2007 Fall. 2007 IEEE 66th.* IEEE, 2007, pp. 1877–1881.

[59] IEEE, "IEEE Standard for Local and metropolitan area networks; Part 16: Air Interface for Fixed and Mobile Broadband Wireless Access Systems, Amendment 2: Physical and Medium Access Control Layers for Combined Fixed and Mobile Operation in Licensed Bands and Corrigendum 1," February 2006.

[60] R. Agarwal, V. Majjigi, R. Vannithamby, and J. M. Cioffi, "Efficient scheduling for heterogeneous services in OFDMA downlink," in *Global Telecommunications Conference, 2007. GLOBECOM'07. IEEE.* IEEE, 2007, pp. 3235–3239.

[61] P. Hosein and C. van Rensburg, "On the performance of downlink beamforming with synchronized beam cycles," in *Vehicular Technology Conference, 2009. VTC Spring 2009. IEEE 69th.* IEEE, 2009, pp. 1–5.

[62] C. Van Rensburg and P. Hosein, "Interference coordination through network-synchronized cyclic beamforming," in *Vehicular Technology Conference Fall (VTC 2009-Fall), 2009 IEEE 70th.* IEEE, 2009, pp. 1–5.

[63] J. Ellenbeck, M. Hammoud, B. Lazarov, and C. Hartmann, "Autonomous beam coordination for the downlink of an imt-advanced cellular system," in *Wireless Conference (EW), 2010 European.* IEEE, 2010, pp. 602–607.

[64] A. Ibing and V. Jungnickel, "Joint transmission and detection in hexagonal grid for 3gpp LTE," in *Information Networking, 2008. ICOIN 2008. International Conference on.* IEEE, 2008, pp. 1–5.

[65] J. Holfeld, V. Kotzsch, and G. Fettweis, "Order-recursive precoding for cooperative multi-point transmission," in *Smart Antennas (WSA), 2010 International ITG Workshop on.* IEEE, 2010, pp. 39–45.

[66] Z.-H. Liu and J.-C. Chen, "Design and analysis of the gateway relocation and admission control algorithm in mobile wimax networks," *Mobile Computing, IEEE Transactions on*, vol. 11, no. 1, pp. 5–18, 2012.

[67] J. Chen, W. Jiao, and Q. Guo, "An integrated qos control architecture for IEEE 802.16 broadband wireless access systems," in *Global Telecommunications Conference, 2005. GLOBECOM'05. IEEE*, vol. 6. IEEE, 2005, pp. 6–pp.

[68] B. Rong, Y. Qian, and H.-H. Chen, "Adaptive power allocation and call admission control in multiservice wimax access networks [radio resource management and protocol engineering for IEEE 802.16]," *Wireless Communications, IEEE*, vol. 14, no. 1, pp. 14–19, 2007.

[69] E. B. Rodrigues and F. R. P. Cavalcanti, "Qos-driven adaptive congestion control for voice over ip in multiservice wireless cellular networks," *Communications Magazine, IEEE*, vol. 46, no. 1, pp. 100–107, 2008.

[70] S. G. Kiani, G. E. Oien, and D. Gesbert, "Maximizing multicell capacity using distributed power allocation and scheduling," in *Wireless Communications and Networking Conference, 2007. WCNC 2007. IEEE.* IEEE, 2007, pp. 1690–1694.

[71] K. Doppler, C. Wijting, and K. Valkealahti, "Interference Aware Scheduling for Soft Frequency Reuse," in *Proc. of the IEEE Vehicular Technology Conference (VTC'09)*, 2009.

[72] M. Rahman and H. Yanikomeroglu, "Enhancing Cell-Edge Performance: A Downlink Dynamic Interference Avoidance Scheme with Inter-Cell Coordination," *IEEE Transactions on Wireless Communications*, p. 1, 2010.

[73] M. Bohge, J. Gross, and A. Wolisz, "Optimal power masking in soft frequency reuse based OFDMA networks," in *Proc. of the European Wireless Conference 2009 (EW'09)*, 2009, pp. 162–166.

[74] Y. Zhou and N. Zein, "Simulation Study of Fractional Frequency Reuse for Mobile WiMAX," in *Proc. IEEE Vehicular Technology Conference VTC Spring 2008*, 11–14 May 2008, pp. 2592–2595.

[75] A. Simonsson, “Frequency reuse and intercell interference co-ordination in E-UTRA,” *Proc. IEEE VTC 2007-Spring*, pp. 3091–3095, 2007.

[76] Ericsson and 3rd Generation Partnership Project, “3GPP R1-074444; On Inter-cell Interference Coordination Schemes without/with Traffic Load Indication,” Shanghai, China, October 2007.

[77] R. Chang, Z. Tao, J. Zhang, and C. Kuo, “A Graph Approach to Dynamic Fractional Frequency Reuse (FFR) in Multi-Cell OFDMA Networks,” in *Proc. of the IEEE ICC 2009*, June 2009.

[78] M. Necker, “A Graph-Based Scheme for Distributed Interference Coordination in Cellular OFDMA Networks,” in *Proceedings of the 67th IEEE Vehicular Technology Conference (VTC2008 - Spring)*, May 2008.

[79] IEEE, “IEEE Standard for local and metropolitan area networks; Part 16: Air Interface for Broadband Wireless Access Systems; Amendment 3: Advanced Air Interface,” May 2011.

[80] D. Tse and P. Viswanath, *Fundamentals of wireless communication*. Cambridge university press, 2005.

[81] A. M. Law, W. D. Kelton, and W. D. Kelton, *Simulation modeling and analysis*. McGraw-Hill New York, 2006, vol. 4.

[82] IEEE, “IEEE 802.16m Evaluation Methodology Document (EMD),” January 2009.

[83] F. Gunnarsson, M. N. Johansson, A. Furuskar, M. Lundevall, A. Simonsson, C. Tidestav, and M. Blomgren, “Downtilted base station antennas-a simulation model proposal and impact on HSPA and LTE performance,” in *Vehicular Technology Conference, 2008. VTC 2008-Fall. IEEE 68th*. IEEE, 2008, pp. 1–5.

[84] 3rd Generation Partnership Project., “3GPP TR 25.996 Spatial Channel Model for Multiple Input Multiple Output (MIMO).”

[85] S. N. Chiu, D. Stoyan, W. S. Kendall, and J. Mecke, *Stochastic geometry and its applications.* John Wiley & Sons, 2013.

[86] R. L. Streit, *Poisson Point Processes.* Springer, 2010, vol. 1.

[87] T. Hoßfeld, R. Schatz, M. Seufert, M. Hirth, T. Zinner, and P. Tran-Gia, "Quantification of YouTube QoE via Crowdsourcing," in *IEEE International Workshop on Multimedia Quality of Experience - Modeling, Evaluation, and Directions (MQoE 2011)*, Dana Point, CA, USA, Dec. 2011.

[88] Z. Wang and J. Crowcroft, "Quality-of-service routing for supporting multimedia applications," *Selected Areas in Communications, IEEE Journal on*, vol. 14, no. 7, pp. 1228–1234, 1996.

[89] I. Foster, A. Roy, and V. Sander, "A quality of service architecture that combines resource reservation and application adaptation," in *International Workshop on Quality of Service (IWQOS).* IEEE, 2000, pp. 181–188.

[90] E. Crawley, R. Nair, B. Rajagopalan, and H. Sandick, *A Framework for QoS-based Routing in the Internet*, Internet Engineering Task Force (IETF) Std., 1998.

[91] K. Nichols, S. Blake, F. Baker, and D. Black, *RFC 2474: Definition of the differentiated services field (DS field) in the IPv4 and IPv6 headers*, Internet Engineering Task Force (IETF) Std., 1998.

[92] J. Klotz, F. Knabe, and C. Huppert, "Resource allocation algorithms for minimum rates scheduling in mimo-ofdm systems," *European Transactions on Telecommunications*, vol. 21, no. 5, pp. 449–457, 2010.

[93] J. Wroclawski, *RFC 2211: Specification of the Controlled-Load Network Element Service*, Internet Engineering Task Force (IETF) Std., 1997.

[94] 3GPP, *TS 23.107 V12.0.0; Quality of Service (QoS) concept and architecture*, 3GPP Std., Sep. 2014.

[95] P. Szilágyi and C. Vulkán, "Application aware mechanisms in hspa systems," in *ICWMC 2012, The Eighth International Conference on Wireless and Mobile Communications*, 2012, pp. 212–217.

[96] Nokia Siemens Networks, "Cell load and application-aware traffic management," whitepaper, Tech. Rep., 2012.

[97] S. Paul and R. Jain, "Openadn: Mobile apps on global clouds using openflow and software defined networking," in *Globecom Workshops (GC Wkshps), 2012 IEEE*. IEEE, 2012, pp. 719–723.

[98] N. McKeown, "Software-defined networking," *INFOCOM keynote talk*, 2009.

[99] E. Nordström, D. Shue, P. Gopalan, R. Kiefer, M. Arye, S. Ko, J. Rexford, and M. J. Freedman, "Serval: An end-host stack for service-centric networking." in *NSDI*, 2012, pp. 85–98.

[100] A. D. Ferguson, A. Guha, C. Liang, R. Fonseca, and S. Krishnamurthi, "Participatory networking: An api for application control of sdns," in *Proceedings of the ACM SIGCOMM 2013 conference on SIGCOMM*. ACM, 2013, pp. 327–338.

[101] P. Georgopoulos, Y. Elkhatib, M. Broadbent, M. Mu, and N. Race, "Towards network-wide qoe fairness using openflow-assisted adaptive video streaming," in *Proceedings of the 2013 ACM SIGCOMM workshop on Future human-centric multimedia networking*. ACM, 2013, pp. 15–20.

[102] T. Zinner, T. Hoßfeld, M. Fiedler, F. Liers, T. Volkert, R. Khondoker, and R. Schatz, "Requirement driven prospects for realizing user-centric network orchestration," *Multimedia Tools and Applications*, pp. 1–25, 2014.

[103] Z. A. Qazi, J. Lee, T. Jin, G. Bellala, M. Arndt, and G. Noubir, "Application-awareness in sdn," in *Proceedings of the ACM SIGCOMM 2013 conference on SIGCOMM*. ACM, 2013, pp. 487–488.

[104] K. Li, W. Guo, W. Zhang, Y. Wen, C. Li, and W. Hu, “Qoe-based bandwidth allocation with sdn in ftth networks,” in *Network Operations and Management Symposium (NOMS), 2014 IEEE.* IEEE, 2014, pp. 1–8.

[105] *IETF Working Group on Application-Layer Traffic Optimization (ALTO)*, Internet Engineering Task Force (IETF) Std.

[106] R. Alimi, Y. Yang, and R. Penno, *Application-Layer Traffic Optimization (ALTO) Protocol*, Internet Engineering Task Force (IETF) Std. RFC 7285, September 2014.

[107] J. W. Jiang, T. Lan, S. Ha, M. Chen, and M. Chiang, “Joint vm placement and routing for data center traffic engineering,” in *INFOCOM, 2012 Proceedings IEEE.* IEEE, 2012, pp. 2876–2880.

[108] D. Xie, N. Ding, Y. C. Hu, and R. Kompella, “The only constant is change: incorporating time-varying network reservations in data centers,” *ACM SIGCOMM Computer Communication Review*, vol. 42, no. 4, pp. 199–210, 2012.

[109] L. Popa, G. Kumar, M. Chowdhury, A. Krishnamurthy, S. Ratnasamy, and I. Stoica, “Faircloud: sharing the network in cloud computing,” in *Proceedings of the ACM SIGCOMM 2012 conference on Applications, technologies, architectures, and protocols for computer communication.* ACM, 2012, pp. 187–198.

[110] B. Heller, S. Seetharaman, P. Mahadevan, Y. Yiakoumis, P. Sharma, S. Banerjee, and N. McKeown, “Elastictree: Saving energy in data center networks.” in *NSDI*, vol. 10, 2010, pp. 249–264.

[111] M. Jarschel and R. Pries, “An openflow-based energy-efficient data center approach,” in *Proceedings of the ACM SIGCOMM 2012 conference on Applications, technologies, architectures, and protocols for computer communication.* ACM, 2012, pp. 87–88.

[112] J. Gross, J. Klaue, H. Karl, and A. Wolisz, “Cross-layer optimization of OFDM transmission systems for MPEG-4 video streaming,” *Computer Communications*, vol. 27, no. 11, pp. 1044–1055, 2004.

[113] S. Khan, Y. Peng, E. Steinbach, M. Sgroi, and W. Kellerer, “Application-driven cross-layer optimization for video streaming over wireless networks,” *Communications Magazine, IEEE*, vol. 44, no. 1, pp. 122–130, 2006.

[114] P. Ameigeiras, J. J. Ramos-Munoz, J. Navarro-Ortiz, P. Mogensen, and J. M. Lopez-Soler, “QoE oriented cross-layer design of a resource allocation algorithm in beyond 3G systems,” *Computer Communications*, vol. 33, no. 5, 2010.

[115] C. Huang, H. Juan, M. Lin, and C. Chang, “Radio resource management of heterogeneous services in mobile WiMAX systems [Radio Resource Management and Protocol Engineering for IEEE 802.16],” *Wireless Communications, IEEE*, vol. 14, no. 1, pp. 20–26, 2007.

[116] A. Reis, J. Chakareski, A. Kassler, and S. Sargento, “Quality of experience optimized scheduling in multi-service wireless mesh networks,” in *IEEE Conference on Image Processing (ICIP).* IEEE, 2010, pp. 3233–3236.

[117] F. Agboma and A. Liotta, “QoE-aware QoS management,” in *6th International Conference on Advances in Mobile Computing and Multimedia.* ACM, 2008, pp. 111–116.

[118] M. Fiedler, T. Hoßfeld, and P. Tran-Gia, “A Generic Quantitative Relationship between Quality of Experience and Quality of Service,” *IEEE Network, Special Issue on Improving QoE for Network Services*, Jun. 2010.

[119] R. Pries, D. Hock, and D. Staehle, “QoE based Bandwidth Management Supporting Real Time Flows in IEEE 802.11 Mesh Networks,” *Praxis der Informationsverarbeitung und Kommunikation*, vol. 32, no. 4, pp. 235–241, 2010.

[120] F. Dobrian, A. Awan, D. Joseph, A. Ganjam, J. Zhan, V. Sekar, I. Stoica, and H. Zhang, "Understanding the impact of video quality on user engagement," in *ACM SIGCOMM conference.* ACM, 2011, pp. 362–373.

[121] M. Grafl and C. Timmerer, "Representation switch smoothing for adaptive http streaming," in *Proceedings of the 4th International Workshop on Perceptual Quality of Systems (PQS 2013),(Vienna, Austria)*, 2013.

[122] T. Hoßfeld, R. Schatz, M. Varela, and C. Timmerer, "Challenges of qoe management for cloud applications," *Communications Magazine, IEEE*, vol. 50, no. 4, pp. 28–36, 2012.

[123] M. Jarschel, D. Schlosser, S. Scheuring, and T. Hoßfeld, "An evaluation of qoe in cloud gaming based on subjective tests," in *Innovative Mobile and Internet Services in Ubiquitous Computing (IMIS), 2011 Fifth International Conference on.* IEEE, 2011, pp. 330–335.

[124] S. Thakolsri, S. Khan, E. Steinbach, and W. Kellerer, "QoE-Driven Cross-Layer Optimization for High Speed Downlink Packet Access," *Journal of Communications*, vol. 4, no. 9, pp. 669–680, Oct. 2009.

[125] M. Xiao, N. Shroff, and E. Chong, "A utility-based power-control scheme in wireless cellular systems," *Networking, IEEE/ACM Transactions on*, vol. 11, no. 2, pp. 210–221, 2003.

[126] M. Andrews, L. Qian, and A. Stolyar, "Optimal utility based multi-user throughput allocation subject to throughput constraints," in *IEEE INFOCOM*, vol. 4. IEEE, 2005, pp. 2415–2424.

[127] G. Song and Y. Li, "Utility-based resource allocation and scheduling in ofdm-based wireless broadband networks," *Communications Magazine, IEEE*, vol. 43, no. 12, pp. 127–134, 2005.

[128] A. Saul, "Simple optimization algorithm for mos-based resource assignment," in *VTC Spring 2008. IEEE.* IEEE, 2008, pp. 1766–1770.

[129] X. Pei, G. Zhu, Q. Wang, D. Qu, and J. Liu, “Economic model-based radio resource management with qos guarantees in the cdma uplink,” *European Transactions on Telecommunications*, vol. 21, no. 2, pp. 178–186, 2010.

[130] T. Hoßfeld, R. Schatz, M. Varela, and C. Timmerer, “Challenges of QoE Management for Cloud Applications,” *IEEE Communications Magazine*, Apr. 2012.

[131] T. Hoßfeld, T. Zinner, R. Schatz, M. Seufert, and P. Tran-Gia, “Transport Protocol Influences on YouTube QoE ,” University of Würzburg, Tech. Rep. 482, Jul. 2011.

[132] D. Strohmeier, S. Egger, A. Raake, R. Schatz, and T. Hoßfeld, “Web Browsing,” in *Quality of Experience: Advanced Concepts, Applications and Methods*, S. R. Sebastian Möller, Ed. Springer: T-Labs Series in Telecommunication Services, ISBN 978-3-319-02680-0,, Mar. 2014.

[133] T. Hoßfeld and A. Binzenhöfer, “Analysis of Skype VoIP Traffic in UMTS: End-to-End QoS and QoE Measurements,” *Computer Networks*, vol. 52, Feb. 2008.

[134] T. Hoßfeld, M. Seufert, C. Sieber, and T. Zinner, “Assessing Effect Sizes of Influence Factors Towards a QoE Model for HTTP Adaptive Streaming,” in *6th International Workshop on Quality of Multimedia Experience (QoMEX)*, Singapore, Sep. 2014.

[135] T. Hoßfeld, R. Schatz, M. Seufert, M. Hirth, T. Zinner, and P. Tran-Gia, “Quantification of YouTube QoE via Crowdsourcing,” in *IEEE International Workshop on Multimedia Quality of Experience*, Dana Point, CA, USA, Dec. 2011.

[136] *IEEE 1900.4, Architectural Building Blocks Enabling Network-Device Distributed Decision Making for Optimized Radio Resource Usage in Heterogeneous Wireless Access Networks*, Std., Feb. 2009.

[137] G. Lampropoulos, A. K. Salkintzis, and N. Passas, "Media-independent handover for seamless service provision in heterogeneous networks," *Communications Magazine, IEEE*, vol. 46, no. 1, pp. 64–71, 2008.

[138] E. H. Ong and J. Y. Khan, "Cooperative radio resource management framework for future ip-based multiple radio access technologies environment," *Computer Networks*, vol. 54, no. 7, pp. 1083–1107, 2010.

[139] T. Bullot, D. Gaïti, G. Pujolle, and H. Zimmermann, "A piloting plane for controlling wireless devices," *Telecommunication Systems*, vol. 39, no. 3-4, pp. 195–203, 2008.

[140] I. F. Akyildiz and X. Wang, "A survey on wireless mesh networks," *Communications Magazine, IEEE*, vol. 43, no. 9, pp. S23–S30, 2005.

[141] R. Pries, D. Hock, N. Bayer, M. Siebert, D. Staehle, V. Rakocevic, B. Xu, and P. Tran-Gia, "Dynamic bandwidth control in wireless mesh networks: A quality of experience based approach," *18th ITCSS on QoE*, 2008.

[142] T. Hoßfeld, P. Tran-Gia, and M. Fiedler, "Quantification of quality of experience for edge-based applications," in *Managing Traffic Performance in Converged Networks.* Springer, 2007, pp. 361–373.

[143] R. Lopes Gomes, W. Moreira Junior, E. Cerqueira, and A. Jorge Abelém, "Using fuzzy link cost and dynamic choice of link quality metrics to achieve qos and qoe in wireless mesh networks," *Journal of Network and Computer Applications*, vol. 34, no. 2, pp. 506–516, 2011.

[144] V. Borges, M. Curado, and E. Monteiro, "Cross-layer routing metrics for mesh networks: Current status and research directions," *Computer Communications*, vol. 34, no. 6, pp. 681–703, 2011.

[145] T. M. Bohnert, D. Staehle, G.-S. Kuo, E. Monteiro, and Y. Koucheryavy, "Speech quality aware admission control for fixed ieee 802.16 wireless man," in *Communications, 2008. ICC'08. IEEE International Conference on.* IEEE, 2008, pp. 2690–2695.

[146] A. Raake, "Short-and long-term packet loss behavior: towards speech quality prediction for arbitrary loss distributions," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, no. 6, pp. 1957–1968, 2006.

[147] 3GPP, *TS 23.401 V10.1.0 ; General Packet Radio Service (GPRS) enhancements for Evolved Universal Terrestrial Radio Access Network (E-UTRAN) access*, 3GPP Std., Sep. 2010.

[148] G. Song and Y. G. Li, "Utility-Based Resource Allocation and Scheduling in OFDM-Based Wireless Broadband Networks," *IEEE Communications Magazine*, vol. 43, no. 12, pp. 127–143, Dec. 2005.

[149] S. Khan, S. Duhovnikov, E. Steinbach, and W. Kellerer, "MOS-Based Multiuser Multiapplication Cross-Layer Optimization for Mobile Multimedia Communication," *Advances in Multimedia*, vol. 2007, 2007.

[150] T. M. Bohnert, D. Staehle, and E. Monteiro, "Speech Quality Aware Resource Control for Fixed and Mobile WiMAX," in *WiMAX Evolution*, F. F. Marcos Katz, Ed. John Wiley & Sons, Jan. 2009, p. 227.

[151] B. Staehle, "Modeling and Optimization Methods for Wireless Sensor and Mesh Networks," Ph.D. dissertation, University of Würzburg, Jul. 2011.

[152] Google, "YouTube Player API Reference," Mar. 2012.

[153] Bundesnetzagentur Germany, "Anzahl der Mobilfunkanschlüsse pro 100 Einwohner in Deutschland von 1990 bis 2013," 2014.

[154] Cisco, "Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2013-2018," February 2014.

[155] S. Egger, T. Hoßfeld, R. Schatz, and M. Fiedler, "Waiting Times in Quality of Experience for Web Based Services," in *QoMEX 2012*, Yarra Valley, Australia, Jul. 2012.

[156] 3GPP Technical Specification Group RAN, "E-UTRA; LTE physical layer – general description," 3GPP, Tech. Rep. TS 36.201 Version 8.3.0, March 2009.

[157] ——, "E-UTRA; physical channels and modulation," 3GPP, Tech. Rep. TS 36.211 Version 8.7.0, May 2009.

[158] ——, "E-UTRA; multiplexing and channel coding," 3GPP, Tech. Rep. TS 36.212, March 2009.

[159] Winner II consortium, "Channel Models Part II: Radio Channel Measurements and Analysis Results, Deliverable 1.1.2," IST-4-027756 WINNER II, September 2007, Tech. Rep., 2007.

[160] S. Alcock and R. Nelson, "Application flow control in YouTube video streams," *ACM SIGCOMM Computer Communication Review*, vol. 41, no. 2, pp. 24–30, 2011.

[161] T. Hoßfeld, R. Schatz, M. Seufert, M. Hirth, T. Zinner, and P. Tran-Gia, "Quantification of YouTube QoE via Crowdsourcing," in *IEEE International Workshop on Multimedia Quality of Experience*, Dana Point, CA, USA, Dec. 2011.

ISSN 1432-8801