Dynamic Resource Allocation for Low-Delay Video Streaming in the Uplink of OFDMA Mobile Networks

vorgelegt von
M.Sc. Minh Hieu Le

an der Fakultät IV - Elektrotechnik und Informatik
der Technischen Universität Berlin
zur Erlangung des akademischen Grades
Doktor der Ingenieurwissenschaften
- Dr.-Ing -

genehmigte Dissertation

Promotionsausschuss:

Vorsitzender:	Prof. Dr.-Ing. Thomas Sikora
Gutachter:	Prof. Dr.-Ing. Adam Wolisz
Gutachter:	Prof. Dr.-Ing. Anja Klein
Gutachter:	Prof. Dr. Holger Karl

Tag der wissenschaftlichen Aussprache: 08. September 2021

Berlin 2022

Acknowledgments

This thesis would not be possible without the support of many people. First of all, I would send my deepest thank to Professor Adam Wolisz for his guidance. It is an honor to pursue my doctoral degree under his supervision at Telecommunication Networks Group (TKN).

I would acknowledge the Vietnam government and the German Academic Exchange Service (DAAD) for giving me a chance to study in Germany.

I am also thankful to all my friends at TKN. In particular, I would thank Arash Behboodi and Konstantin Miller for their fruitful cooperation. I am grateful to Manoj Rege, Onur Ergin, and Jan Hauer for sharing my hard times and happy times. Special thanks go to Mathias Bohge for his warm heart.

Finally, I am forever indebted to my wife Hanh Vuong, our daughters Ban-Mai and Thu-Duong, and my big family in Vietnam. Without them, I could not finish this long and challenging journey.

Abstract

In recent years, video has been inducing an exponential growth in mobile traffic. Today, mobile operators pay several billions of euros for radio frequency bands to cope with enormous traffic demands. At the same time, the enduring demand growth increasingly forces mobile operators and video service providers to look for efficient solutions to make the best possible service quality out of the limited radio spectrum.

This thesis focuses on improving the Quality of Experience (QoE) of multiple low-delay video streams in the uplink of mobile networks using Orthogonal Frequency Division Multiple Access (OFDMA). Several mobile networks like LTE and WiMAX have been using OFDMA. In that context, this thesis provides multiple contributions. The first contribution is a novel dynamic resource allocation approach, which exploits wireless channels' random variations to improve user throughput while suppressing Multiple Access Interference (MAI). MAI presents when user signals are not perfectly synchronous in the uplink of OFDMA networks. Next, the thesis presents two cross-layer video adaptation approaches, which adopt the introduced dynamic resource allocation approach. Those two approaches target low-delay video streaming services using non-layered and layered video coding. While the video industry has been broadly using non-layered video coding, layered video coding might be more prevalent in the future. Those two technologies have distinct adaptation principles, so they require different solutions tailored particularly for them. In both cases, via efficient mathematical transformations, the large-timescale problem of video adaptation (in a few seconds) is pursued via a series of small-timescale resource allocation problems (each in a few milliseconds). By doing that, video adaptation algorithms can quickly react to wireless channel variations and meet low latency requirements. As for non-layered video coding, another contribution is to consider potential throughput gains via efficient resource allocation algorithms as selecting video quality. Throughout the thesis, we develop several mathematical optimization problems and adaptation algorithms to determine the performance gain of proposed approaches.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

More than 100 years ago, physicist and inventor Nikola Tesla predicted

> *"It will soon be possible to transmit wireless messages all over the world so simply that any individual can carry and operate his own apparatus,"*

Nowadays, we can only admire the accuracy of his incredible imagination. Thanks to wireless technology, talking to someone thousands of kilometers away, in some remote areas, and even on the move is no longer a problem. Furthermore, mobile systems have been dramatically evolving over the last two decades, from voice service and short messages to broadband Internet connections. Modern mobile networks can enable a large variety of applications such as multimedia services (e.g., video streaming, video gaming), cloud services (e.g., storage, computing), Internet of Things (IoT) (e.g., smart home, smart city, industry 4.0), autonomous driving and web-based services.

Along with the evolution of mobile networks, global mobile data traffic has been continuously growing. That trend is forecasted to continue in at least the next few years [1]. The primary driving force of the traffic growth has been the enormous number of intelligent devices and the popularity of video streaming applications like YouTube [1] and Netflix [2]. Unlike in traditional services, where playback can only start when the receiver finishes downloading the entire video file, playback (of some buffered content pieces) and downloading (the rest) can simultaneously occur in streaming services.

Offloading is one crucial technique to alleviate the traffic demand in mobile networks. The basic idea is to relieve congested mobile networks by routing mobile traffic to available small-cell networks like WiFi that use unlicensed spectrum. More than half of total mobile traffic was offloaded through WiFi or femtocell in 2019 [1]. However, the traffic load on mobile networks is still tremendous.

To cope with this problem, on the one hand, mobile operators might pay multiple billion euros to utilize additional frequency bands. On the other hand, the spectrum crunch forces mobile operators to improve the performance of all communication

[1] `http://youtube.com`
[2] `http://netflix.com`

processes. Intelligent utilization of precious radio resources is especially crucial to provide the best possible service quality to consumers with the least cost.

Regarding radio technologies, several modern communication systems like WiFi, Long Term Evolution (LTE) and Worldwide Interoperability for Microwave Access (WiMAX) (a less favorite alternative to LTE) selected Orthogonal Frequency Division Multiplexing (OFDM) and its multiple user version Orthogonal Frequency Division Multiple Access (OFDMA). OFDMA has several unique features. An important one is its robustness against interference. OFDMA achieves that by splitting the overall spectrum into several small subcarriers. Those subcarriers can be overlapping but differentiable (i.e., being able to be demodulated) at the receiver. Another advantage of OFDMA lies in the possibility for mobile operators to dynamically adapt resource allocation to, among other things, channel diversities to improve user throughput. Resource allocation can also be adapted to assure fairness among users. One critical disadvantage of OFDMA is the strict synchronization requirement between multiple devices in time and frequency domains.

In the meantime, the video industry has widely utilized advanced streaming solutions to provide high user experience over dynamic transmission environments like the Internet. At the heart of such solutions are advanced video encoders. Modern encoders can achieve high compression ratios while maintaining good video quality. Additionally, they also offer flexible video adaptation for service providers and users to select suitable video quality subject to, for instance, user experience, available bandwidth, and computation power. In the video world, Quality of Experience (QoE) is commonly used to measure the subjective perception judged by users [2]. Note that QoE is required since Quality of Service (QoS) centers on network parameters (e.g., throughput and delay) and does not guarantee good viewer experience.

HTTP Based Adaptive Streaming (HAS) is today the most dominant video streaming approach on the Internet. Several companies have developed their proprietary solutions based on HAS, for example, Adobe HTTP Dynamic Streaming, Apple HTTP Live Streaming, Microsoft Smooth Streaming. Dynamic Adaptive Streaming over HTTP (DASH) is the only international standardized solution that can enable cooperation between vendors. Several service providers like YouTube and Netflix have adopted DASH as the *de-facto* standard.

Most streamed content nowadays is Video on Demand (VoD). In VoD services, the entire video content is encoded and stored on remote media servers. Users can download and buffer a mass number of video segments, which can later absorb link rate fluctuations and reduce the possibility of video stalls during the video playback. Fewer video stalls result in better QoE. That is why some users might have to wait a few tens of seconds before the YouTube application starts playing their requested videos.

In recent years, the amount of low-delay streaming applications has been increasing [3]. The primary difference of such applications compared to VoD lies in their tight latency constraints required to assure good user QoE (e.g., a few hundreds of milliseconds versus a few tens of seconds for VoD services). Examples of low-delay streaming applications are video conferencing, live news streaming, video gaming, vehicle-to-vehicle communication, robot, and telesurgery. In addition, apart from

video content generated by organizations like service providers and media companies, more and more User Generated Content (UGC) traffic (e.g., Facebook [3], TikTok [4]) is streamed on the Internet, creating more video streams in the uplink (i.e., from users toward networks).

However, the development of low-delay streaming applications over mobile networks poses several challenges to academia and also industry. The following paragraphs shortly describe some key challenges.

- First, most studies on adaptive streaming in the literature focused on VoD, while only a little attention looked at low-delay streaming. In most targeted scenarios, buffer sizes are on the order of tens of seconds and thus not suitable for low-delay services, in which the delay needed for building up a big buffer level is not tolerable.

- Second, typical streaming applications operate entirely on the application layer while treating underneath networks as black boxes. This separation between the application layer and the link layer follows the traditional Open System Interconnection (OSI) architecture. The adoption of OSI can facilitate deployment and reduce complexity. However, since video adaptation algorithms in such applications work based on video segments and each segment is typically a few seconds long, such adaptive algorithms tend to be too slow to cope with vast and rapid variability of link rate.

- Third, modern mobile networks like LTE and WiMAX feature flexible frameworks to deliver different QoS levels for different applications. For instance, the QoS framework in LTE is implemented based on the QoS Class Identifier and the Guaranteed Bit Rate (GBR) [4]. However, networks do not incorporate efficient frameworks to address QoE. Usually, modern resource allocation strategies adopt proportional fairness regarding QoS metrics (e.g., throughput, error probability, packet loss) as allocating resources to users [5]. Unfortunately, such approaches result in sub-optimal resource allocation schemes for adaptive video streaming [6].

Motivated by those challenges, this thesis investigates the problem of equally improving QoE of multiple low-delay streams, which compete for precious resources in the uplink of OFDMA mobile networks. To that aim, we strive to develop cross-layer approaches that jointly consider video adaptation and resource allocation.

Cross-layer approaches have been recently emerged to become an attractive means to serve video traffic. In the literature, at least two concrete aspects of cross-layer approaches can boost the performance of adaptive video streaming services. First, the physical layer's information (e.g., channel information and available radio resource), which takes effect in a few milliseconds, can be beneficial. One example is to improve the estimation of available link rate, which is crucial for selecting proper video bitrates

[3] `http://facebook.com`
[4] `http://tiktok.com`

[7]. Second, in the opposite direction, information about video content characteristics can help the Base Station (BS) to efficiently allocate precious radio resources to users that can benefit most from them. Furthermore, the joint consideration of multiple streams can assure a certain level of fairness while avoiding overloading Radio Access Network (RAN) [8]. Especially, since the resource allocation can perform at the scale of a few milliseconds (compared to a few seconds of video adaptation), such cross-layer adaptive algorithms can react much snappier to fast channel variations and serve low-delay streaming better [9].

This thesis mainly focuses on three main questions considering cross-layer approaches enhancing QoE of multiple low-delay streams in the uplink.

- Challenge 1: **How to deal with imperfect synchronization?**

  In the uplink, the signal arriving at BS is the superposition of multiple components sent by multiple users simultaneously. Due to several reasons like mobility and imperfect oscillators, typical component signals exhibit different synchronization errors (aka offsets) and are not perfectly synchronized. Lack of perfect synchronization then damages the orthogonality between subcarriers and causes Multiple Access Interference (MAI) that can severely degrade user throughput [10].

  The most common approach to deal with imperfect synchronization is to (i) first estimate synchronization errors and then (ii) counteract them [11]. Despite the simple principle, implementation of this approach deeply involves very complex signal processing techniques. Typically, sizable overhead is adopted for the estimation purpose to ease complexity, but that leads to a reduction of spectral efficiency. Besides, due to several reasons, such as inadequate implementations, there is always a chance that residual offsets still exist and significantly deteriorate user signals [12].

  One question that emerges in this context is how to efficiently deal with imperfect synchronization and MAI with less overhead so that more resources can convey video traffic.

- Challenge 2: **How to develop efficient cross-layer approaches?**

  One challenge as developing cross-layer approaches stems from the difference in timescales of adaptation mechanisms. In particular, while resource allocation operates based on resource units in milliseconds' timescale, video adaptation performs on video segments whose lengths are in the range of a few seconds.

  Then the question is how to efficiently couple the large-timescale video adaptation with the small-timescale resource allocation. Efficient resource allocation algorithms should, on the one hand, quickly react to fast wireless variations and, on the other hand, achieve long-term goals instead of instantaneous performance gains (e.g., short-term throughput). Besides, video adaptation should fairly maximize user QoE while not exceeding the achieved link rate. Specifically, cross-layer algorithms also need to consider tight latency requirements to avoid video stalls.

- Challenge 3: **How to derive efficient resource allocation schemes and video adaptation decisions?**

  In general, Optimization Problems (OPs) are typically required to find efficient resource allocation schemes and video adaptation decisions. Formulating and solving those OPs are, however, not trivial [13]. One primary challenge lies in the discreet nature of resource allocation, where each resource unit is assigned uniquely to one user. Another challenge is to balance the system's spectral efficiency against the fairness between users, who compete for the shared resources. Finding optimal solutions to the trade-off between those two potentially conflicting criteria can lead to highly complex mathematical problems and, thus, is generally challenging. Therefore, the achievement of proper cross-layer approaches, which can be efficiently solved to achieve optimized performance, is critical.

This thesis manages to provide multiple contributions regarding the above questions. First, we present a novel Dynamic Resource Allocation (DRA) approach that uses less overhead and achieves valuable throughput gains compared to conventional approaches that deal with imperfect synchronization and MAI. We then extend the proposed approach to leverage valuable throughput gains for enhancing the QoE of multiple low-delay video streams. In particular, the thesis addresses the second question above by presenting new cross-layer approaches for two types of video streaming services, which adopt (i) Non-Layered Video Coding (NLVC) and (ii) Layered Video Coding (LVC). While the former has broadly functioned in the industry, the latter is considered an essential solution for future applications. Those two technologies have distinct adaptation principles, and, thus, require different solutions tailored particularly for them. As for NLVC, another contribution lies in the explicit consideration of the potential throughput gain achieved via efficient resource allocation algorithms as selecting video bitrate. Within this thesis, we develop several OPs and their less-complex versions to determine the potential of the proposed solutions.

The rest of the thesis is organized into six chapters. First, Chapter 2 provides short summaries about wireless channels, OFDM and OFDMA systems, and adaptive streaming technologies. Chapter 3 then gives an overview of the current result in the literature regarding challenges under consideration. The last section of that chapter presents the scope of the thesis. Chapter 4 presents the proposed approach that uses less overhead when dealing with imperfect synchronization and MAI in the uplink of OFDMA networks. OPs of resource allocation are formulated and solved to achieve user throughput gains while suppressing MAI. We introduce several heuristic approaches to reduce complexity while achieving relatively good performance. Chapter 5 and Chapter 6 propose cross-layer approaches for enhancing QoE of low-delay video streaming services that adopt NLVC and LVC, respectively. Finally, in Chapter 7, conclusions are drawn, and issues for future work are presented.

Chapter 2

Background

This chapter provides some background required to discuss key challenges and main contributions of this thesis. The first section covers the basics of wireless channels. Consequently, we summarize key aspects of OFDM and OFDMA systems. Finally, this chapter presents the principles of video encoding and adaptive video streaming.

2.1 Wireless Channel

Understanding the main characteristics of wireless channels is crucial to design any efficient communication system. This section briefly describes the attenuation effects of wireless channels and their models.

The behavior of wireless channels can be generally explained by three physical propagation mechanisms: reflection, diffraction, and scattering. Figure 2.1 illustrates those mechanisms. For instance, received signals can yield severe losses due to the diffraction around the edges of surrounding buildings and the scattering by uneven surfaces like trees. Consequently, multiple copies of the transmitted signal following different paths can arrive and overlap at the receiver. This phenomenon is referred to as multipath propagation.

Apart from the distortion caused by multipath propagation, the received signal is also distorted by path loss and shadowing. Besides, received signals also suffer from additional thermal noise. The following sub-sections discuss those factors in more detail.

Path loss

As traversing from the transmitter to the receiver through space, the electromagnetic wave's power decreases along the way. This degradation is known as path loss. In the simplest case, when there is only one ray following the Line of Sight (LOS) path, path loss can be derived analytically from the theory of the electromagnetic field. As a result, path loss, defined as the ratio of the received power $¶_{\mathrm{RX}}$ to the transmitted power P_{TX}, usually takes the following form [14]:

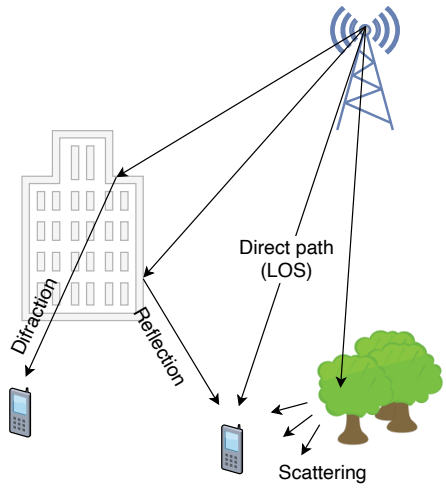

Figure 2.1: Illustration of main propagation mechanisms

$$h_p = \frac{P_{\mathrm{RX}}}{P_{\mathrm{TX}}} = \left(\frac{\lambda}{4\pi d}\right)^2, \tag{2.1}$$

where λ is the wavelength of radio signals, and d is the distance between the transmitter and the receiver.

However, the assumption of the free space environment described above is valid, perhaps, only for aeronautical communication. For the propagation in large open areas like in the countryside, the two-ray model can effectively predict path loss [15]. In this model, the second ray reflected on the ground is also considered apart from the direct one. When more than two rays exist (like in urban areas), one can use the ray-tracing method to develop efficient path loss models. However, this method demands exact information about objects' locations and tedious geometrical calculations.

To practically predict path loss, extensive measurement campaigns have been conducted and used to develop empirical models. Measurement results can also be used to adjust analytical models, resulting in semi-empirical models. In such models, the averaged path loss over time is modeled as a function of distance and typically formulated as shown in [16]. Particularly, we have:

$$h_p = \frac{P_{\mathrm{RX}}}{P_{\mathrm{TX}}} = K\left(\frac{d_0}{d}\right)^{\gamma}, \tag{2.2}$$

where d_0 is the reference distance. K and γ are coefficients representing environment characteristics (e.g., urban or rural areas, with or without LOS). Those coefficients are derived by fitting analytical models to measurement results. Two well-adopted empirical models for urban environments are the Okumura-Hata model [17] and the Lee model [18]. In the scope of the research project European Cooperative for Scientific and Technical 231 (COST-231), the Okumura-Hata urban model was extended to cover a more elaborated frequency range [19]. COST-231 also proposed another model for microcells and small macrocells by combining models proposed by Walfisch

and Ikegami. Particularly, assuming the LOS case and a distance more than 20 m, path loss as the function of distance d and frequency f_c yields:

$$h_p \text{ [dB]} = 42.6 + 26\log(d) + 20\log(f_c) \tag{2.3}$$

Shadowing

Path loss models predict the average attenuation over time for a given distance between the transmitter and the receiver. However, practical path loss measurements for the same distance at different places differ from predicted values. The difference is explained by the existence of large objects in the nearby environment (like surrounding buildings and mountains). The statistical variation observed in path loss measurements is presented by shadowing. Shadowing is efficiently modeled as a zero-mean Gaussian process [20]. The stochastic model of shadowing loss, denoted by h_s, is given by:

$$p(h_s) = \frac{1}{\sqrt{2\pi\sigma_s^2}} e^{-\frac{h_s^2}{2\sigma_s^2}} \tag{2.4}$$

where σ_s^2 is the variation expressed in decibels. Typical σ_s takes a value in the range from 5 to 12 dB depending on communication systems [18].

Furthermore, since shadowing is caused by large obstacles, it exhibits a correlation in space [21]. A model of the auto-correlation of shadowing values can be found in [22] and shown as:

$$\rho(r) = \frac{1}{\sqrt{\sigma_s^2}} e^{\frac{d}{d_0}} \tag{2.5}$$

where d_0 is the reference distance, which, according to measurements, varies between 25 m and 100 m at 1900 MHz or between a few and few tens meters for 900 MHz.

Multipath fading

Due to multipath propagation, each received signal is constituted by multiple copies of the transmitted one. Each copy undergoes a different environment, thus yielding different power, phase, and delay variations. Due to phase differences, those copies can interact constructively or destructively, resulting in fast and vigorous variations of received signal power. This effect is called multipath fading. In general, multipath fading occurs much faster compared to the varying pace of shadowing.

The time difference between the first and the last arriving copies is called delay spread Δt_d. Delay spread relates to the coherence bandwidth of channel [23], which is the frequency range over which channel behavior is approximately unchanged. However, the relation between delay spread and coherence bandwidth is subjective. One reason lies in the definition of the similarity of channel responses. An approximation of coherence bandwidth W_c can be found in [14] and formulated as:

$$W_c \approx \frac{1}{2\pi\Delta t_d} \tag{2.6}$$

Based on the comparison between coherence bandwidth and signal bandwidth, a fading channel can be classified as flat or frequency selective. The difference is that frequency components of a transmitted signal are treated equally in the flat fading channel and differently in the frequency selective channel.

In addition, due to delay spread, one signal symbol can spread and interfere with adjacent ones, causing Inter-Symbol Interference (ISI). One efficient way to avoid ISI is to insert GI between adjacent signal symbols. ISI can be eliminated if GI's length exceeds delay spread.

When the transmitter, the receiver, and surrounding objects are not stationary, the received signal suffers from frequency shifts caused by the Doppler effect. The range between the minimal and the maximal frequency shift is called Doppler spread Δf_d. Importantly, there is a reciprocal relationship between Doppler spread and coherence time T_c [23]. In essence, the channel's coherence time is the period over which the channel behavior does not change significantly. Generally, the larger the Doppler spread, the shorter the coherence time. The relation is again subjective. It can be explained by the fact that a path with a substantial Doppler shift may have a too weak amplitude that gives no strong distortion. One example model is shown in [14] and cited below.

$$T_c \approx \sqrt{\frac{9}{16\pi(\Delta f_d)^2}} = \frac{0.423}{\Delta f_d} \tag{2.7}$$

Depending on how signal symbol duration is compared to coherence time, wireless channels can be classified as either fast fading or slow fading. In the first case, the channel's responses changes rapidly within one symbol. In the meantime, slow fading channel can be considered static over one or several signal symbols.

Due to Doppler spread, signals generated on one frequency band can interfere with adjacent bands. When those bands come from the same transmitter, the interference is referred to as Inter-Carrier Interference (ICI). One means to cope with ICI is to insert a Guard Band (GB) between two adjacent frequency bands.

In this thesis, fading channel is assumed to be flat. Consequently, the following discussion focuses only on modeling flat fading channels.

Mathematically, flat fading can be modeled via stochastic processes. Assuming there are a large number of statistically independent paths and there is no direct path between the transmitter and the receiver (i.e., Non Line of Sight (NLOS)), the Probability Density Function (PDF) of amplitudes of complex received signals can be efficiently modeled by the Rayleigh distribution [14]. It means

$$p(h_f) = \frac{h_f}{\sigma^2} \times e^{\frac{-(h_f)^2}{2\sigma^2}} \tag{2.8}$$

where $\sigma = \frac{1}{2}\overline{P}_{\mathrm{RX}}$ and $\overline{P}_{\mathrm{RX}}$ is the average received signal power. When there is a dominant LOS path, the Rayleigh distribution is replaced by the Rice distribution.

Apart from the fading amplitude h_f, another important property of fading channel is the Doppler power spectral density, which describes how much spectral spread the Doppler effect causes. The corresponding effect of Doppler spread is the auto-correlation of channel responses in time. That property is efficiently modeled by the Jakes model, which is based on an assumption of equal-strength rays and uniformly distributed arrival angles [24]. The Jakes fading model is a deterministic method for simulating time-correlated Rayleigh fading channels. This model is widely used and also adopted in this thesis.

Thermal noise

Noise is generally not a useful signal caused by arbitrary sources like human-made devices like microwave ovens. In this thesis, only thermal noise is relevant and, thus, considered. Generally, thermal noise originates from the heat caused by random movements of charged particles (like electrons) in the circuitry. That heat then distorts useful signals.

Thermal noise can be effectively modeled as a zero-mean Gaussian stochastic process [22]. It means the power spectral density of thermal noise is constant over all frequency ranges. Due to that, it is said to be white, like the white light that contains all frequency. The Power Spectral Density (PSD) of thermal noise can be computed (in Watt per Hertz) as the product of the Boltzmann constant and temperature.

2.2 OFDM Basics

OFDM has been adopted in several modern communication standards. For instance, OFDM presents itself in several WiFi standards in the IEEE 802.11 family. In addition, OFDM also takes place in several broadcast standards like Digital Audio Broadcasting (DAB) and Digital Video Broadcasting (DVB). Besides, Asymmetric Digital Subscriber Line (ADSL) also uses OFDM to enable fast data transmissions over copper telephone lines.

Essentially, OFDM is a form of Frequency Division Multiplexing (FDM). In FDM systems, the total channel bandwidth is divided into several non-overlapping sub-bands. The transmitter can then modulate those sub-bands simultaneously to send its data. Typical FDM implementations require small gaps or GBs inserted between sub-bands to avoid ICI. However, the adoption of GB reduces spectral efficiency, since GBs convey no user data.

OFDM distinguishes itself from conventional FDM systems by, first of all, allowing sub-bands to be overlapping. In OFDM systems, the smallest sub-band is one subcarrier. Typically, a group of adjacent subcarriers forms one subchannel. The problem of ICI is avoided by assuring mutual orthogonality between subcarriers. Roughly speaking, orthogonality means the PSD of each subcarrier has its maxima exactly where those of all other subcarriers equal zero. Figure 2.2 illustrates the subcarrier orthogonality in the frequency domain. As a result of orthogonality, OFDM can improve spectral efficiency compared to FDM. Another advantage of OFDM lies in its ro-

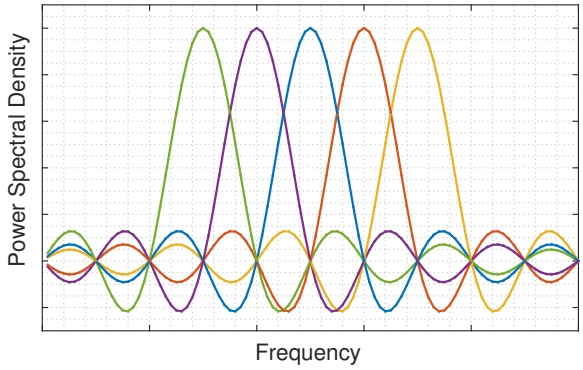

Figure 2.2: Illustration of subcarrier orthogonality

bustness against multipath fading. Due to the division of bandwidth to narrow-band subcarriers, the common bandwidth of subcarrier is normally smaller than the channel's coherence bandwidth. As the result, OFDM is resistant to frequency selective fading.

Analytically, let B be the total bandwidth, which is divided into N orthogonal subcarriers. Consequently, subcarriers are equally spaced on the frequency axis by a distance of:

$$f_0 = \frac{B}{N} = \frac{1}{T_{\text{sym}}}, \tag{2.9}$$

where T_{sym} denotes the OFDM symbol duration.

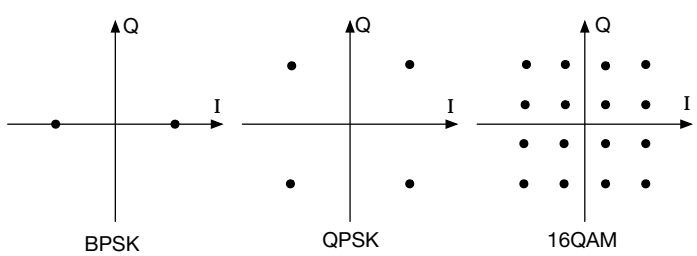

Figure 2.3: Example modulation schemes

At the transmitter, first, channel coding is performed on user data, so that transmission errors can be detected and corrected at the receiver [25]. Next, the coded data is mapped on a complex constellation following a modulation scheme like Quadrature Amplitude Modulation (QAM) or Phase Shift Keying (PSK) into complex symbols d_k. Figure 2.3 illustrates some modulation schemes, where I_k and Q_k denote in-phrase and quadrature components of complex symbols, respectively. As depicted in Figure 2.4, the stream of those symbols is then fed to a Serial to Parallel Converter, so that N symbols can be modulated on N subcarriers simultaneously.

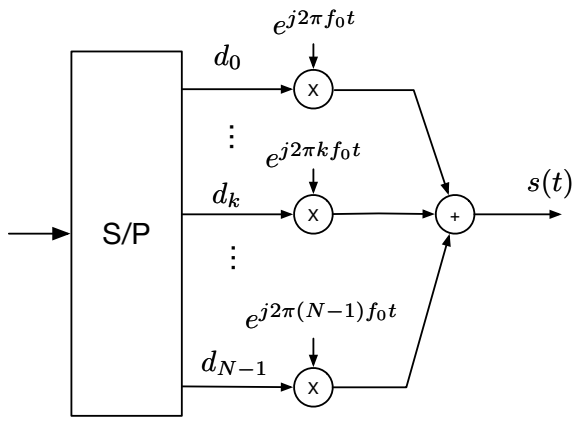

Figure 2.4: Block diagrams of OFDM modulation

The modulation of d_k with the k-th subcarrier, becoming s_k, is presented in the following equation.

$$s_k(t) = Re\left\{ d_k e^{j2\pi k f_0 t} \right\}, \quad 0 \leq t \leq T_{\text{sym}} \tag{2.10}$$

Note that function $Re()$ in (2.10) returns the real part of a complex symbol. Consequently, signals on N subcarriers are then multiplexed into one OFDM signal $s(t)$. For the scaling purpose, $s(t)$ is normalized to N. Then the final form of OFDM signal $s(t)$ yields:

$$s(t) = \frac{1}{N} \sum_{k=0}^{N-1} s_k(t), \qquad 0 \leq t \leq T_{\text{sym}} \tag{2.11}$$

Next, $s(t)$ is modulated with the nominal carrier frequency f_c before being fed to the Digital to Analog Converter (DAC) and transmitted over the air. In practice, frequency f_c specifies the frequency band licensed to a network operator. $s(t)$ is called the baseband signal, and the transmitted signal is the bandpass signal. Without loss of generality, the baseband signal is used to discuss the channel's impact and the OFDM demodulation.

From equations (2.10) and (2.11), it can be seen that modulated OFDM signals can be derived by using Inverse Discrete Fourier Transform (IDFT) instead of N oscillators. IDFT can be efficiently implemented by the low complex Inverse Fast Fourier Transform (IFFT) algorithm.

In this thesis, the frequency spacing is selected much smaller than the channel coherence bandwidth, and all OFDM symbols are also much smaller than the coherence time. In that case, the wireless channel is slow and flat fading, and the channel response on subcarrier k in time t is then characterized by a complex-valued symbol $h_k(t)$. Consequently, the received signal takes the following form:

$$r(t) = \sum_{k=0}^{N-1} r_k(t) + n(t) = \sum_{k=0}^{N-1} h_k(t)s_k(t) + n(t), \tag{2.12}$$

where $n(t)$ denotes the thermal noise. $n(t)$ is assumed to be zero-mean Gaussian and flat across all subcarriers. Now we investigate the demodulation. The basic idea is to exploit the orthogonality condition. The mathematical description of orthogonality (between two subcarriers m and n) is shown below:

$$\frac{1}{T_{\text{sym}}} \int_0^{T_{\text{sym}}} e^{j2\pi m f_0 t} e^{-j2\pi n f_0 t} = \frac{1}{T_{\text{sym}}} \int_0^{T_{\text{sym}}} e^{j2\pi (m-n) f_0 t} = \delta(m-n) \tag{2.13}$$

where $\delta(m-n)$ is the delta function, which equals one when $m = n$ and zero otherwise. Based on this property, the signal on subcarrier k can be demodulated using oscillator k as following:

$$r_k(t) = r(t)e^{-j2\pi k f_0 n t} = h_k(t)d_k + \theta(t) \tag{2.14}$$

where $\theta(t)$ is a term of thermal noise:

$$\theta(t) = \frac{1}{T_{\text{sym}}} \int_{t=0}^{T_{\text{sym}}} n(t)e^{-j2\pi k f_0 t} \tag{2.15}$$

Similar to the transmitter, Discrete Fourier Transform (DFT) can replace the bank of oscillators to efficiently demodulate data symbols d_k. In practice, DFT is implemented by the Fast Fourier Transform (FFT) algorithm in order to reduce computational complexity.

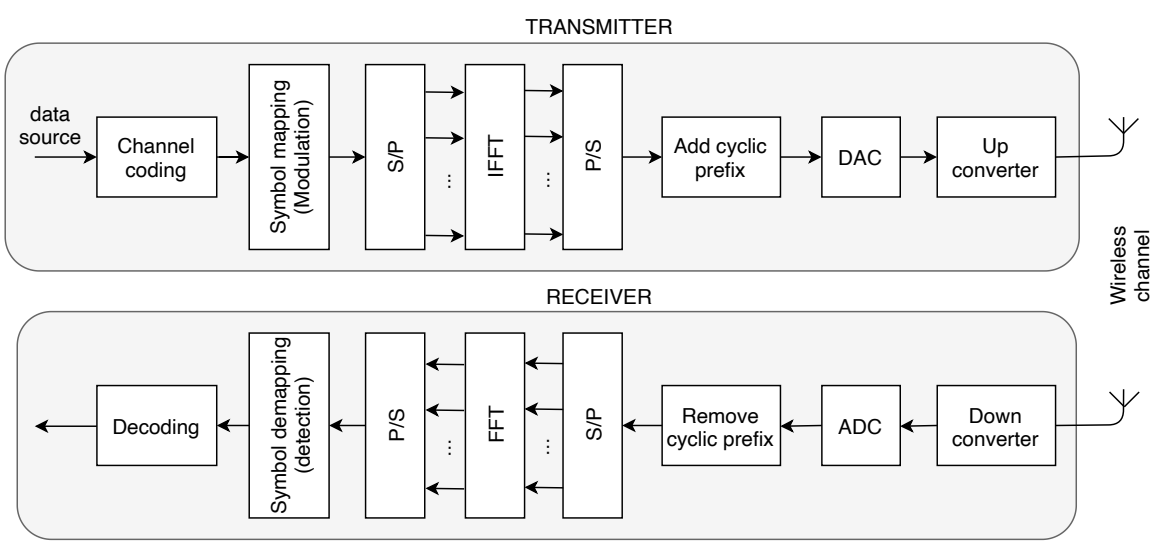

Figure 2.5: Block diagrams of OFDM transmitter and receiver

Finally, block diagrams of the transmitter and the receiver are illustrated in Figure 2.5. Note that S/P and P/S are the Serial to Parallel and Parallel to Serial converters.

To eliminate ISI, a Guard Interval (GI), whose duration is denoted by T_g, is prepended to each OFDM symbol. GIs will be deleted at the receiver before received signals are fed to the FFT module. GIs' lengths are chosen to be bigger than the expected delay spread so that multipath components from one symbol cannot interfere with the following ones.

Moreover, instead of just using an empty guard (i.e., the transmitter generates no power during this period), some samples in the tail of the OFDM symbol are copied and used as GI. For that reason, GI is also called Cyclic Prefix (CP). Figure 2.6 illustrates the problem when using empty GI. As it is shown on the left of Figure 2.6, when GI is zero and a time offset exits, the sum of signals on two subcarriers over time is not zero. The highlighted half of the sine symbol is the culprit. In this case, ICI arises and deteriorates received signals. In contrast, using CP can avoid this issue. Further analysis can be found in [26].

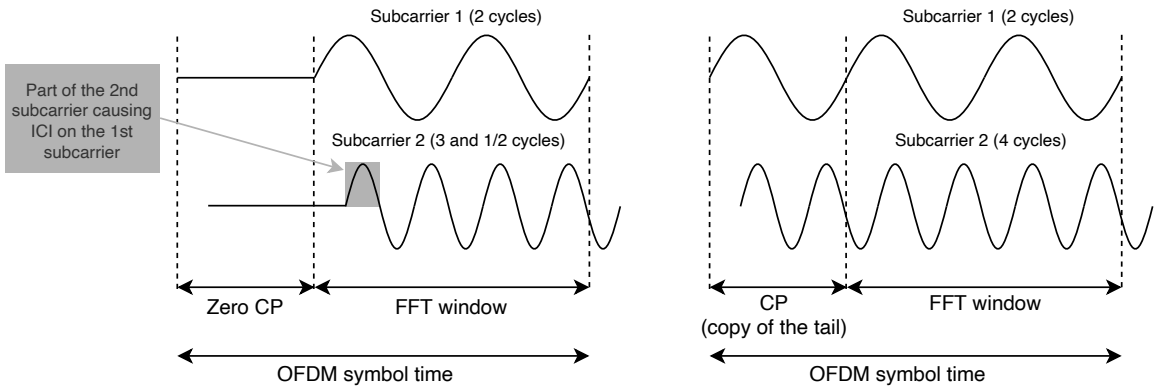

Figure 2.6: Empty GI causing ICI

However, adding CP results in overhead in time and reduces spectral efficiency by a factor of $T_g/(T_{\text{sym}} + T_g)$. The efficiency reduction can be negligible if the OFDM symbol duration is much larger than the delay spread, i.e., $T_{\text{sym}} >> T_g$. The sum of symbol duration and CP is denoted by T and $T = T_{\text{sym}} + T_g$.

2.3 OFDMA Basics

OFDMA is the multiple-access variation of OFDM, where unique sets of sub-bands are assigned to different users and users' data are transmitted simultaneously. One advantage of OFDMA compared to OFDM is that OFDMA can exploit multiuser diversity to improve the system performance. Multiuser diversity basically means users experience wireless channels of the same frequency band differently, and the probability that all users suffer from deep fading attenuation at the same time is typically low. Thus, intelligent algorithms allocating sub-bands to suitable users can increase spectral efficiency. Resource allocation typically operates on subchannels

(instead of subcarriers). That is to balance signaling overhead and frequency diversity. As mentioned above, each subchannel consists of some adjacent subcarriers. Figure 2.7 illustrates three assignment strategies. In this example, three users are sharing 12 subchannels. Each user takes a block of consecutive subchannels in the blocking assignment or a set of interleaved subchannels in the interleaving method. Finally, subchannels are dynamically assigned to users in the general assignment.

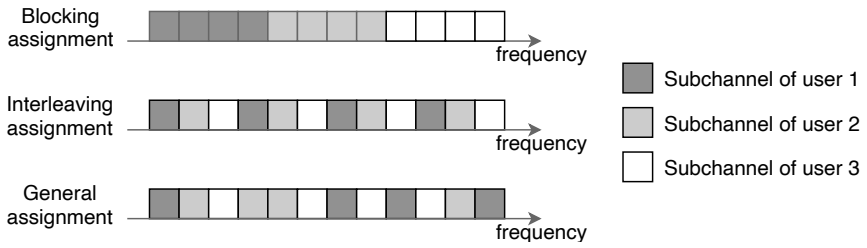

Figure 2.7: Illustration of frequency assignment strategies

The block diagram for the downlink, i.e., from BS to M users, is illustrated in Figure 2.8. At first, considering Channel State Information (CSI), a subset of subchannels is assigned to each user. Then, similar to OFDM, user bitstream is grouped and mapped on constellations of modulation schemes like QAM and PSK to obtain complex symbols d_m.

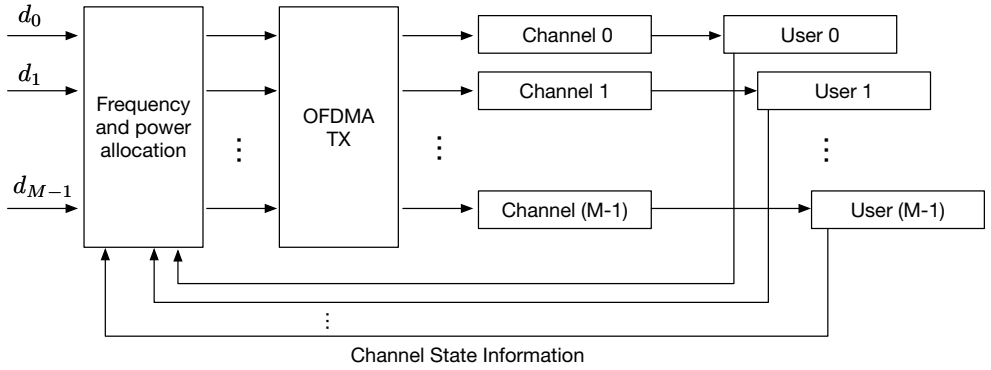

Figure 2.8: Block diagram of the downlink of a typical OFDMA system

Usually, the mapping process is realized by performing an algorithm to select a proper Modulation and Coding Scheme (MCS) subject to the channel state. Essentially, higher MCS order means more bits can be loaded into one symbol but higher sensitivity to noise. Figure 2.9 shows the relation of sensitivity (in term of Bit Error Rate (BER)) of different MCSs and Signal to Noise Ratio (SNR). As the name implies, SNR is the ratio of received power to noise. Consider the same tolerable BER, higher schemes should be used only when the channel condition is good.

To achieve proper resource allocation schemes, CSI is crucial so that the transmitter can adapt resource allocation schemes to time-varying channel states in the most efficient way. In this thesis, the transmitter in the uplink, i.e., the user device, equally distributes maximal power budget among assigned subcarriers.

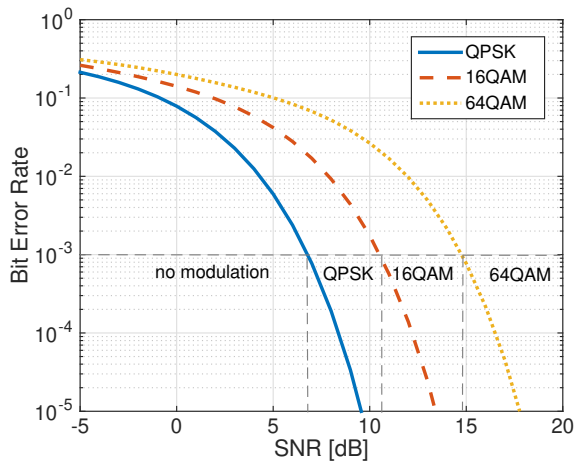

Figure 2.9: Illustration of MCS selection

2.4 Adaptive Video Streaming

This section summarizes the basis of video coding and video streaming techniques.

2.4.1 Video coding

Video coding is essentially the process of compressing and changing the format of video content while maintaining a reasonable perceptual quality. The main purpose of video coding is efficiency and compatibility regarding storage, transmission, and interoperability. Video decoding is the reverse process. The term *"codec"* indicates a pair of encoding and decoding algorithms to compress and decompress video data. There are three main compression approaches in the literature: hybrid, wavelet, and parametric compressions. Among them, hybrid compression has been proved to be most efficient and, thus, broadly adopted [27]. Hybrid compression combines three coding algorithms underneath, which are entropy, prediction, and transform coding.

- Entropy coding is a type of lossless data compression achieved by representing frequently occurring input patterns by short codewords and rare patterns by long ones. However, the compression efficiency is insignificant by applying only entropy coding.

- Prediction coding aims to predict the unknown content based on known ones. The idea is to exploit similarities in the spatial and temporal dimensions of video content. For instance, pixels belonging to the blue sky in a video frame (i.e., in spatial dimensions) tend to have similar values of luminance (i.e., pixel intensity) and chrominance (i.e., pixel color). Similarly, the change of picture pixels in consecutive frames can exhibit a temporal correlation to the object's movement. The exploitation of similarities within a frame is referred to as INTRA prediction, while the one between frames is INTER prediction. Using prediction

algorithms, the sender can send only a small amount of data representing the difference between the predicted signal and the actual one.

- Transform coding, which is based on Discrete Cosine Transform (DCT) [28], is used to compress data further. Roughly speaking, DCT can be seen as a cut-down version of FFT, where only the real part of FFT is returned. Using this method, video content is separated into parts of different importance concerning visual quality. Consequently, compression is achieved by discarding less critical information.

Video coding standards are required to ensure compatibility. Two main standardization bodies are International Telecommunication Union - Telecommunication Sector (ITU-T) and International Standard Organization (ISO)-International Electrotechnical Commission (IEC). While ITU-T develops the Moving Picture Experts Group (MPEG) standard family (e.g., MPEG-1, MPEG-2, MPEG-4), ISO works on the H.26x series (e.g., H.264 and H.265). The most common standard used by 91% of video industry developers by September 2019 is H.264/MPEG-4 Advanced Video Coding (shortly H.264/AVC) [29], which is jointly developed by ITU-T and ISO/IEC. A key advantage of H.264/AVC is its compression efficiency (i.e., only half or less the bitrate of MPEG-4) for good video quality while not increasing the implementation complexity significantly.

Besides, H.264/AVC provides enough flexibility so that it can be adopted for a wide variety of applications, including broadcast, DVD storage, RTP/IP packet networks, and multimedia telephony systems. In practice, H.264/AVC defines a set of supported profiles and levels that target different application areas [30]. While a profile defines the set of coding algorithms that can be used, a level defines certain limits for key parameters (e.g., maximal resolution, maximal output bitrate) of coding parameters. For instance, the baseline profile targets applications that require a low computational complexity and a high error decoding resilience, while the main profile aims at high coding efficiency and low error robustness.

To boost the scalability, ITU-T and ISO/IEC JTC jointly developed the Scalable Video Coding (SVC) extension to H.264/AVC [31]. SVC allows the same encoded video can be decoded at different quality levels with, for instance, different resolutions and frame rates while avoiding transcoding or re-encoding. SVC achieves its scalability via the concept of LVC, which will be the main subject of the following sub-section. LVC is also supported in H.265, i.e., High Efficiency Video Coding (HEVC), which is the direct successor of H.264/AVC. HEVC targets very high resolutions (e.g., 4K and 8K) and double compression efficiency while maintaining similar or the same video quality compared to H.264/AVC. Some key features of HEVC are flexibility and more efficient compression.

2.4.2 Layered and Non-Layered Video Coding

In a system using LVC, video content is encoded into a hierarchical structure of layers, including one base layer and some enhancement ones. During the decoding

process, the base layer is decoded first to construct basic images, which provides the lowest quality regarding resolution, frame rate, and fidelity. The quality is then enhanced by decoding higher layers in the hierarchical structure. In other words, the same encoded content can be decoded at different bitrates. It is important to note that higher layers can only be successfully decoded when lower layers are available; missing the base layer leads to a decoding failure despite the presence of all enhancement ones.

LVC exploits three scalability dimensions of video content: time, space, and fidelity. They are also referred to as temporal, spatial, and quality scalability, respectively. Figure 2.10 illustrates three scalabilities types.

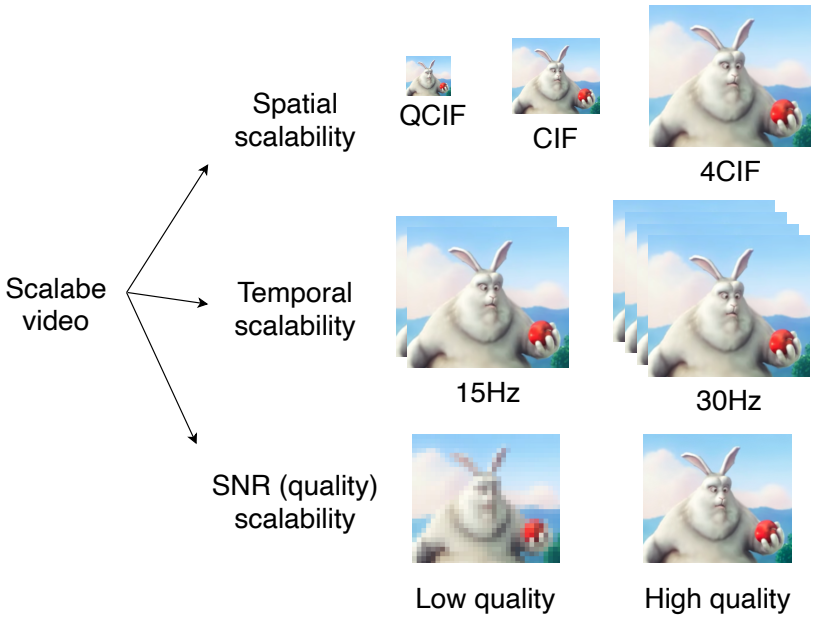

Figure 2.10: Illustration of three scalability types

On the contrary, NLVC streams do not contain any subset, which can be decoded partially. It means, encoded data using NLVC is not scalable. In other words, if a large proportion of the stream is missing, it cannot be decodable properly.

2.4.3 Streaming applications and latency requirement

Regarding latency requirements, streaming applications can be categorized into three types, which are interactive, low-delay, and VoD streaming. Some examples of interactive services are video conferencing and video gaming. To assure a good user experience for such multimedia streams, ITU-T recommends limiting the latency under 100 ms [32]. Live sport streaming is one example of low-delay streaming services. Typically, the tolerable latency for this type spans from one to few seconds.

One fundamental characteristic of interactive and low-delay streaming is that the encoding, downloading, and playback processes happen simultaneously. Thus there is a specific limit for available contents clients can download. On the contrary, in the case of VoD, the entire video content is available on the server. Therefore, users can download as much as possible.

In this thesis, Chapter 5 approaches low-delay streaming, and Chapter 6 targets even lower latency requirements toward interactive streaming.

2.4.4 Adaptive Streaming

Essentially, adaptive streaming is the technology that can adapt on the fly the bitrate of video content to, for instance, available link rate. The *de-facto* approach to develop adaptive streaming applications is to adopt HAS. In a HAS system, video content is divided into a series of segments, each of which has a constant playback duration ranging from one to tens seconds. Each segment can be decoded independently of other segments. The encoded content is stored on Hypertext Transport Protocol (HTTP) servers. A client downloads video segments from servers using the HTTP protocol. By adopting HTTP, HAS can leverage the ubiquitous delivery infrastructure developed for web traffic, including Content Delivery Network (CDN) and cache servers. Also, HTTP can ease the deployment since it is typically allowed to pass middleboxes, such as firewalls and Network Address Translation (NAT) devices.

A client starts to play the video when a sufficient amount of data is buffered. An initial delay is the waiting time between the first request and the start of playback. During the playback, new segments are downloaded, while buffered ones are played out. If the download of a segment is not finished when the player needs it, the playback goes into a stall. Such an event is also referred to as buffer underrun. Different strategies dealing with video stalls are available for different types of applications. For the real-time and low-delay applications, the player can ignore missing segments, strive to download and play next available segments to minimize the latency. In contrast, with VoD, the player normally halts, re-buffers, and resumes the playback only when the buffer level exceeds a pre-defined threshold. In general, buffer underrun can strongly degrade the perceived quality. Therefore, video adaptation algorithms are needed to adapt video bitrate to network conditions and ensure continuous playback.

To enable the adaptation of video streaming, a HAS server can provide multiple sub-streams of the same video content, which are suitable for different conditions like available link rate. Service providers typically select the set of supported sub-streams during the planning phase [33].

NLVS and LVS

Either LVC or NLVC can be used to derive sub-streams. In this thesis, Layered Video Streaming (LVS) and Non-Layered Video Streaming (NLVS) are used to refer to streaming services using LVC and NLVC, respectively. Importantly, HAS by design is codec-agnostic. Therefore, the integration of different coding standards should require only minor changes in the implementation [34]. In either case, the metadata describing available sub-streams is required for video adaptation process. Some vital information is segment indices, representation indices, links to individual representations, and bitrates of representations. Figure 2.11 depicts the adaptation principle of NLVS and LVS.

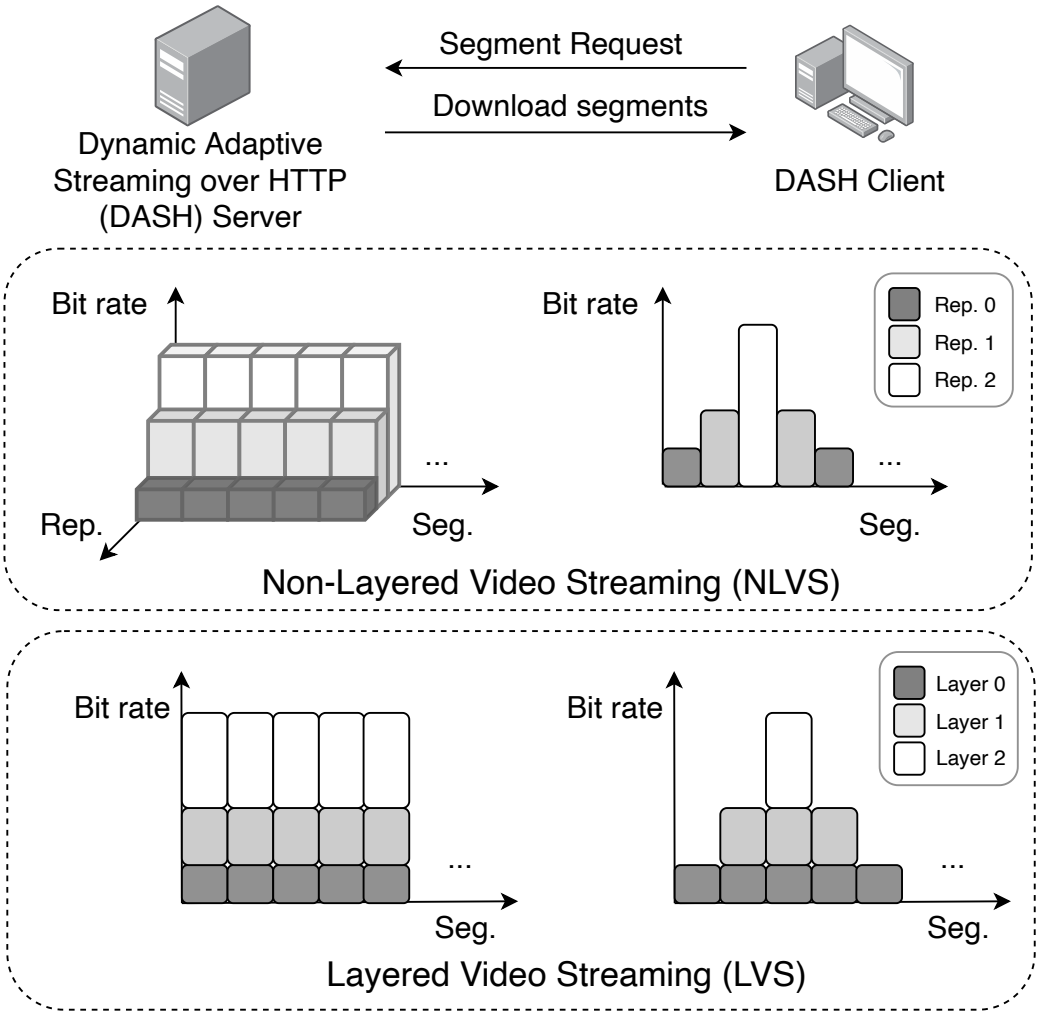

Figure 2.11: Illustration of HAS using NLVS and LVS

In NLVS, each input video segment is encoded multiple times with different coding parameters (e.g., resolution and frame frequency) into multiple representations. The encoding process of different representations is independent of each other. In literature, this approach is also known as Multiple Description Coding (MDC) [35]. In the example shown in Figure 2.11, there are three independent representations for each segment. Note that, this example assumes Constant Bitrate (CBR), therefore segments belonging to one representation have the same bitrate. Alternatively, bitrates can vary among segments when Variable Bitrate (VBR) encoding is enabled. Nevertheless, the key point is that the bitrate of a low-quality representation is expected to be smaller than those of higher ones.

Most services nowadays adopt NLVS. One reason is the low implementation complexity of video adaptation. However, NLVS has several disadvantages, such as large additional content storage (due to several representations) and non-optimal quality selection under varying network conditions [36].

Recently, LVS has drawn more interest from both academics as well as industry. In LVS, a sub-stream can be seen as a combination of encoded layers. In that way, LVS offers multiple adaptation points while streaming a segment. For instance, after successfully downloading a few layers, a client can decide either to download the next enhancement layer or the base layer of the next segment. This advantage of LVS facilitates great responsiveness to bandwidth fluctuation. Another advantage of LVS over NLVS is the requirement for storage. Concerning NLVS, since multiple independent representations of the same content need to be stored, storage of roughly 200% to 300% of the highest video quality is required [36]. In the case of LVS, the server stores only the metadata describing the hierarchical structure together with the only-encoded-once video content. The overhead of metadata is estimated at roughly 10% for each enhancement layer [37]. Thus, the overhead needed for a typical configuration of one base layer and seven enhancement layers is 70%. Compared to a few hundred percent in the NLVS case, the reduction in storage capacity is significant. Another advantage of LVS is the robustness to the large bandwidth fluctuation. Notably, since a whole NLVS segment is required for decoding, a deep fade of wireless channels can extend download time and cause video stalls. When adopting LVS, the player can cancel the unfinished downloads of enhancement layers and play out buffered layers to avoid buffer underrun.

One disadvantage of LVS lies in the additional signaling overhead as many HTTP requests are needed for each segment [36]. Fortunately, the HTTP/2 standard has recently introduced the Server Push feature that allows the server to respond with a pre-configured sequence of segments to a single request [38]. This feature has the potential to reduce the protocol overhead.

2.4.5 Quality of Experience

One primary factor in developing any video streaming system is to measure user satisfaction or QoE. However, several reasons, like the complex human visual and neural systems, make the assessment of QoE become a highly complex task.

Evaluation methods

There are three main approaches to evaluating QoE: subjective tests, objective models, and data-driven models [39].

Concerning subjective tests, video is evaluated by human viewers. Mean Opinion Score (MOS) is commonly used for subjective tests. MOS is a measure for the arithmetic average over individual human-judged quality values. Typically, MOS has a scale from 1 (bad) to 5 (excellent).

Subjective tests obviously can bypass the lack of complete knowledge about human perception, but this method is very costly due to human involvement and, thus, not applicable for a large number of videos. For that reason, objective models are the most of interests. Many studies have been conducted to study the relationship between influencing factors and human perception. One common metric is Peak Signal-to-Noise Ratio (PSNR), which is computed by averaging the squared intensity

difference (i.e., Mean Square Error (MSE)) of distorted and reference image pixels. Although PSNR does not reflect very well perceived quality, it is appealing because it is simple to calculate, has clear physical meanings, and is mathematically convenient in the context of optimization.

In order to calculate PSNR, full knowledge of the original content is required. For that reason, this method belongs to the group of full reference methods. Alternatively, Structural Similarity (SSIM) can provide a better assessment. In this method, the high sensitivity of the human visual system to structural distortion is exploited [40]. Importantly, PSNR and SSIM are measured for each video segment, thus considered as short-term quality measurements.

Finally, data-driven models have emerged from big data technology, where a large amount of information about viewers' opinions and content characteristics are available for analysis.

Key factors affecting QoE

This section describes key factors that can strongly affect user QoE. Those factors are initial delay, video stalling, and quality fluctuation.

At the beginning of each streaming session, an initial delay is required to build up the playback buffer. Roughly speaking, the impact of initial waiting time on QoE strongly depends on the streaming application. But, in general, several studies have pointed out that the initial delay is expected by the users from the everyday usage of video applications and considered less important for QoE [41].

While the initial delay is expected, video stalling is unexpected. The impact of video stalling is thus much worse than initial delay. Furthermore, other researches suggest that video stalling is even more important than many other factors like frame rates and quantization parameters [33]. Importantly, it is shown that both the stalling duration and the number of stalling events have an exponential impact on QoE [42].

For those reasons, it is crucial to mitigate video stalling. In addition, the fluctuation of selected representations, or the adaptation trajectory, can also affect QoE. Although the impact is less severe than video stalling, its impact on QoE must not be neglected [33].

Chapter 3

Related Work and Scope of the Thesis

The first three sections of this chapter discuss main problems considered in this thesis and the gap in the literature. In particular, the first one concerns previous approaches dealing with imperfect synchronization in the uplink of OFDMA systems. The second section then addresses DRA in OFDMA systems. The state of the art regarding cross-layer video adaptation is then presented. The last section summarizes the scope of this thesis and the main contributions.

3.1 OFDMA Synchronization

One fundamental disadvantage of OFDMA lies in the stringent requirement of synchronization in time and frequency. Lack of synchronization results in interference that severely degrade the system performance. In the downlink, time errors can lead to incorrect placements of FFT windows and, thus, give rise to ISI. Besides, frequency errors can result in overlaps of frequency bands or demodulation at incorrect frequencies, causing ICI. In the uplink, misalignment in the time and frequency domains between users' signals arriving at BS can additionally introduce MAI.

However, acquiring sufficient synchronization is exceptionally challenging due to several reasons like mobility and imperfect oscillator clocks. In the literature, there are plenty of studies on this topic. For instance, a general search with the keyword *"OFDM synchronization"* on IEEExplore [1] gives more than 4000 results, and about a half of them are published in the last ten years (from 2011 to 2021).

This section gives an overview of synchronization challenges in OFDMA systems.

A typical approach dealing with synchronization errors is first to estimate and then counteract time and frequency offsets. For that reason, this approach is generally referred to, in this thesis, as the Estimation-Correction Based Approach (ECBA).

Commonly the estimation task is accomplished in two phases. The first phase, which typically happens at the beginning of OFDMA frames, is to achieve a coarse estimate. Consequently, the long-term deviation caused by the Doppler effect and the oscillator's clock drifts is addressed by a fine-tracking process in the second phase [11].

[1] https://ieeexplore.ieee.org/

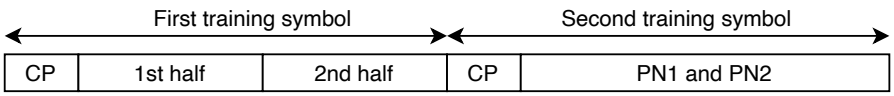

Figure 3.1: Illustration of training symbols in [44]

3.1.1 Synchronization in the downlink

Coarse estimations

One common approach for acquiring synchronization errors in the downlink is to use pilot signals dedicated for the estimation purpose [11]. Many works following this approach first utilize a training sequence of a few OFDM symbols in time. Some important works are introduced in [43]–[49]. Those solutions then distinguish themselves from others by proposing different training signals and corresponding estimation algorithms.

Schmidt and Cox introduced an exemplary work in [44]. This work is adopted as the base for many other studies. Therefore, we first summarize the proposed method in more detail. The proposed solution utilizes two training symbols. While the first one consists of two identical halves, the second one is generated by modulating a Pseudo Noise (PN) sequence on the even subcarriers and another PN sequence on the odd subcarriers.

The acquisition of time offsets then means finding the beginning of the first training symbol. Due to the special repetitive structure, the receiver can detect that symbol by searching for the delay where the autocorrelation function of time samples yields the maximum value. Once the first symbol is located, time offset can be calculated, and the placement of FFT windows can be adjusted.

A large frequency offset (exceeding subcarrier spacing) is possible by decomposing it into a fraction and an integer part of frequency spacing. The two parts are then addressed sequentially. First, the fraction part is determined by computing the phase difference of samples in two identical halves. Second, the integer part is exposed by the modification of PN numbers at the FFT output. Finally, time error estimates are then achieved by finding the delay at which the normalized autocorrelation function of received signals is maximal. However, the autocorrelation function in [44] exhibits a large plateau, which greatly reduces the estimation accuracy.

Other works like [45], [46] then aim to enhance the accuracy by developing more efficient training patterns and corresponding estimation algorithms. In general, proposed algorithms become less effective when time errors are small, e.g., less than 2% of the FFT window in [46].

Another approach utilizes so-called blind estimations. Those algorithms exploit inherent properties of OFDMA signals to estimate offsets, and no dedicated resource is required for the estimation purpose. The first advantage of this approach is the improvement of spectral efficiency since more resources convey user data. Second, estimation algorithms can take place entirely on the receiver without signaling with the transmitter. However, those advantages come with the cost of higher complexity and lower accuracy compared to those utilizing training sequences.

One main class of blind estimations achieves time errors by exploiting the correlation between the CP and the OFDMA symbol's tail, or the cyclostationary of the OFDM transmission. Some examples are [50], [51]. However, the performance of such solutions degrades strongly when significant multipath fading destroys the similarity between the CP and the tail. Besides, another class reserves some un-modulated subcarriers (i.e., null subcarriers) to estimate time and frequency offsets. Important works in this class are [52], [53].

Fine tracking estimations

The presence of non-negligible sampling frequency errors and Doppler shifts gives rise to long-term variations of time and frequency errors. Those deviations need to be tracked periodically throughout OFDM frames to avoid ICI and ISI.

One important work that explicitly deals with sampling frequency errors is introduced by Kim et al. in [54]. In this work, some pilot subcarriers are required. The proposed algorithm then tracks residual time offsets by using a Phrase Locked Loop (PLL) to compute the difference of phase changes between all pilot subcarriers at the IFFT output to obtain accurate results.

Fine frequency tracking is commonly achieved via closed control loops. In such systems, the FFT output is fed back to compute errors, which are then fed to a Voltage Controlled Oscillator (VCO). Consequently, VCO generates an exponential term to compensate frequency offsets of received signals [55]. Next, pilot subcarriers or blind estimations can be used to compute errors. Key researches using such approaches include [50], [55]–[58]. In those works, errors can be computed from samples either in the time or the frequency domain (i.e., at the FFT output). One example of methods operating in the frequency domain is [59]. In this work, a Maximum Likelihood (ML)-based approach is used to approximate frequency offsets. That work is then slightly improved in [55]. In general, using training signals can improve estimate accuracy and reduce complexity compared to blind estimations but requires some resources dedicated for the estimation task.

3.1.2 Synchronization in the uplink

Generally, achieving good synchronization in the uplink is much more challenging than in the downlink. Particularly, since all component signals (on sub-bands) in the downlink come from BS, component signals yield the same synchronization offsets. Consequently, the receiver has to track time and frequency offsets from one entity. In the uplink, signals arriving at BS are constituted of multiple components sent from multiple users, thus possessing different time and frequency offsets. The BS thus needs to simultaneously estimate and track plenty of offset values from all component signals. Furthermore, the next task after achieving offset estimates is to counteract synchronization errors. However, this task is also challenging. One fundamental reason is that adjusting BS's clock to synchronize with one user clock can increase the errors of others.

One common approach for simplifying those challenges is to adopt long CP so that there is no overlap between OFDMA symbols in the time domain. As a result, component signals are effectively quasi-synchronous in the time domain, and only frequency offsets remain. To that aim, CPs accommodate not only the maximal delay spread but also the maximal time offset (caused by, e.g., different propagation delay).

In general, many solutions have been developed for acquiring and correcting frequency offsets in the uplink. Similar to the aforementioned classification in the downlink, proposed approaches can roughly be classified into two groups depending on if training sequences are required or not. In case of no training sequence, blind estimations are required.

Suppose no training sequence is available. The estimation task generally becomes much more challenging, since an exhaustive multiuser multi-dimensional ML search is required for blind estimations [60]. Importantly, blind estimations are only possible when a special structure of signals, like in the interleaving assignment, can be exploited for the estimation purpose.

This section focuses on frequency estimations in the uplink. And, since proposed algorithms in the literature are tailored for different frequency assignment strategies, this section is divided into three sub-sections addressing blocking, interleaving, and general assignment.

Estimation for blocking assignment

In the case of blocking assignment, each user takes a unique set of continuous subchannels. Blocking assignment simplifies the synchronization task to a large extent. In the presence of frequency errors, only a few subcarriers located at the borders may experience significant ICI. To mitigate this problem, GB are typically inserted between sub-bands to avoid ICI as proposed in [53], [61]. Suppose frequency offsets are adequately smaller than GB. In that case, BS can easily separate signals from different users by feeding received signals through a bank of band-pass filters in the frequency domain. Each filter targets one sub-band of one user. Consequently, conventional offset estimation methods can be adopted.

Estimation for interleaving assignment

Adopting interleaving assignment, subchannels assigned to each user are equally spread over the system bandwidth. The motivation is to exploit frequency diversity. Despite that advantage, interleaving assignment is very sensitive to frequency offsets since subchannels can overlap much more frequently than in block assignment. That makes the synchronization task for interleaving assignment very challenging.

Some important works in this context are [60], [62]. The common idea of those works is to adopt blind estimations that exploit the periodic structure of interleaving assignment. Roughly speaking, that structure can be seen as equally-spaced subchannels, which belong to one user, have an equal frequency offset. Based on this idea, introduced solutions focus on designing efficient estimators with reasonable complex-

ity. For instance, the work in [62] proposes an iterative estimation scheme using Space Alternating Generalized Expectation Maximization (SAGE) for performing ML parameter estimation. In addition, a series expansion when evaluating the ML function is used in [60].

Frequency estimation for general assignment

Adopting general assignment, BS can exploit knowledge of CSI to assign suitable subchannels to users. This assignment strategy is more flexible than two previous assignment strategies and able to exploit multiuser diversity [55]. However, the absence of a predefined structure of sub-band assignments makes the synchronization task extremely challenging.

The most common approach in the literature typically requires some dedicated training symbols for the estimation. Some important works that use training symbols are shown in [63]–[66]. For instance, it is assumed in [64] that each user transmits a training block at the beginning of each OFDMA frame. The estimation of frequency offsets is then persuaded via a ML algorithm. However, the introduced solution is prohibitively complex since they demand an enormous searching space over multiple dimensions. To reduce complexity, the work resorts to the alternating projection, which uses a sequence of projections to convert the joint multi-dimensional search to a sequence of one-dimensional searches. However, the proposed system is still highly complex, while the estimation result is not optimal.

Alternatively, the work in [63] leverages a mathematical solution of certain OP to replace alternating projection algorithms. The simulation result shows a clear outperformance compared to [64]. Another approach for reducing complexity is shown in [65]. In this work, the alternating projection is substituted by an interactive scheme. In each step, users are divided into groups and handled separately to transform a large mathematical matrix into smaller ones.

The common feature of the proposed algorithms in [63]–[65] is that estimation errors can be significant when SNR is weak. For instance, MSE is around 10^{-2} when SNR is 0dB. The higher the SNR, the smaller the MSE value.

In the meantime, a family of sub-optimal estimators was provided in [67] to achieve lower complexity. In essence, this work aims to replace the exact ML criterion with approximated criteria, which are easier to compute and fairly efficient. This is done by approximating the inverse of a frequency offset matrix with the inverse of a predetermined matrix. Proposed estimators are shown to be asymptotically efficient while requiring reasonable complexity. However, the complexity reduction comes at the cost of estimate accuracy.

Interestingly, the work in [68] exploits the tile structure used in IEEE 802.16 networks and proposes to embed training signals on a few subcarriers (instead of all subcarriers of a subchannel). However, this work again requires an exhaustive iterative process for estimating frequency offset.

Synchronization Error Correction

Once time and frequency offsets in the uplink are estimated, BS can counteract them to restore orthogonality among users' signals. Unfortunately, as mentioned above, this problem is not trivial, because the alignment of one specific user would cause misalignment to all the others. Suppose long CP can facilitate the quasi-synchronization, the following paragraph focuses only on the correction of frequency offsets.

One early work in this context is introduced in [69], where signals from each user are treated separately by one detector. As a result, multiple FFT blocks are needed. An alternative design is introduced in [61], referred to as CLJL, where only one FFT unit is used. Consequently, a post-FFT processing technique, which is based on circular convolution, is employed to correct frequency offsets. However, CLJL performs well only for the blocking assignment. For the case of interleaving or a general assignment, the solution in [70] proposes an iterative interference cancellation scheme to enhance the performance of CLJL. Alternatively, the method of linear multiuser detection can be used instead of interference cancellation as shown in [71]. Roughly speaking, proposed solutions aim at restoring orthogonality among users by applying a linear transformation to the FFT output.

In general, the performance of correction algorithms depends on the assignment strategy, where they perform better for blocking assignments than interleaving and general ones. Apart from that, the performance degrades when SNR decreases.

3.1.3 Summary

In general, the ECBA approach has some main disadvantages. First, estimation and correction algorithms are generally implemented using complex signaling processing techniques.

Second, large overheads are expected to reduce complexity. For instance, corresponding to the Partially Used Sub-Channelization (PUSC) method in the IEEE 802.16m standard, each uplink tile consists of 4 adjacent subcarriers in frequency and 3 symbols in time, and 4 out of the 12 subcarrier-symbol combinations are for pilot signals, i.e., approximately 33% of system resources is for the overhead [72].

Third, proposed methods following ECBA are strongly coupled with selected resource assignment strategies, and not every scheme can be efficiently handled. In general, due to the implementation complexity, there is always a chance that residual offsets still exist [12]. For instance, a well-known study in [73] proposes a frequency offset tracking algorithm for the IEEE 802.11e OFDMA uplink. In this paper, a fluctuation of frequency offsets over OFDMA symbols is explicitly considered. Evaluation results in this work show that certain estimation errors exist by the presence of thermal noise plus interference induced by frequency offsets. Especially, it shows that when the variation of frequency offsets is significant, like 10% of the frequency spacing between subcarriers, residual frequency offsets can be significant. In practice, the standard IEEE 802.16 requires a precision of less than 2% of frequency spacing and 25% of symbol duration should be maintained [74].

3.2 Dynamic Resource Allocation

The idea of adapting transmission parameters to channel states to improve the communication performance can be traced back to the study in [75] introduced by Hayes in 1968. Later, that idea has been widely adopted for various wireless communication systems as an essential means to cope with unreliable channels [76], [77]. For multiuser systems like OFDMA, another motivation for adaptive approaches is to exploit multiuser diversity [78]. The basic idea is that channel states of different users are unlikely to be bad or good at the same time. Consequently, frequency resources are dynamically assigned to users with good channel states, aiming to improve spectral efficiency.

In the literature, the term DRA generally encompasses adaptive schemes of transmission power, bandwidth, and MCS. Note that the term Adaptive Coding and Modulation (ACM) indicates the adaptation of MCSs. In the literature, adaptive modulation is also referred to as adaptive bit loading.

This section discusses the main results in the literature regarding DRA in OFDMA systems. First, important studies in the downlink from a BS to multiple users (i.e., point-to-multipoint) are briefly summarized. Second, significant works on DRA in the uplink are reviewed. In this section, an explicit discussion about the relation between DRA and MAI is included. Although the thesis mainly concerns the uplink, a short review for the downlink is still necessary, since most of DRA algorithms for the uplink are extended from those for the downlink.

Researches on DRA algorithms for OFDMA systems can be divided into two main groups, which are margin-adaptive and rate-adaptive. On the one hand, margin-adaptive algorithms aim to minimize the transmission power while providing users with a minimum QoS support (regarding, for instance, data rate or BER). On the other hand, rate-adaptive algorithms strive to maximize throughput with constraints on the maximal transmission power. In this thesis, only rate-adaptive algorithms in the single-cell context are relevant and considered.

3.2.1 DRA in the downlink

Regarding the downlink of OFDMA systems, a straightforward objective of rate-adaptive algorithms is to maximize the total cell throughput. In other words, the goal is to maximize the system's spectral efficiency, which is defined as the average number of bits that can be sent per one Hertz. This problem has been well-studied in the literature. That challenge is normally formulated as multiuser sum-rate maximization problems. For instance, Kim et al. considered adaptive subcarrier allocation jointly with adaptive modulation in [79]. The authors managed to convert nonlinear OPs to linear ones. Jang et al. then considered the optimization of transmission power and subcarrier allocation [80]. Importantly, that work proved that, in theory, the total cell throughput of multiuser systems is maximized if each subcarrier is assigned to the user with the best channel gain on it. Consequently, the total transmission power is distributed among subcarriers following the water-filling theorem. Roughly speaking,

it means more transmission power is applied to subcarriers experiencing good channel gains.

To derive optimal resource allocation schemes for practical system states, formulated OP need to be solved numerically. However, that task is extremely challenging. The first reason lies in the discreet nature of subcarrier assignment, where each subcarrier is typically assigned to one user. As a result, integer programming problems are expected. Those problems in basic forms are non-linear and generally hard to solve. Particularly, the sum-rate maximization problem is proved to be NP-complete [13]. It means no known algorithms that can provide optimal solutions for such a problem in polynomial time. As a result, an extensive computation load can be expected to solve this problem numerically. The work in [81] further considers several related problems of adaptive allocation in the OFDMA downlink and proves that they are generally NP-hard. Thus, sub-optimal and less complex approaches are required to alleviate the problem. To that goal, three different methods are commonly adopted [82]. The first method is to relax integer constraints so that integer problems can be converted into linear forms, which can be solved more efficiently [83]. The second method is to split the DRA problem into two separate and less complex problems of power allocation and frequency allocation. The third method is to develop heuristic algorithms. As an example, the sum-rate maximization problem can be efficiently solved by the greedy algorithm in [84].

The sum-rate maximization, however, leads to unfairness among users. For instance, more frequency resources will be allocated to users located closer to BS since their channel gains are better than those in the further locations. It is important that spectral efficiency can be harmonized with fairness among users. Note that while spectral efficiency is a technical parameter, fairness is subjective. In general, fairness and spectral efficiency tend to be conflicting, and fairness among users comes at the cost of a sub-optimal spectral efficiency [82].

Suppose fairness is simply defined by user throughput. One way to tackle fairness is to ensure that each user can acquire a minimum rate as proposed in [85]. To that aim, new constraints to enforce minimum throughput are added to the sum-rate maximization problem. In this work, a heuristic two-step resource allocation process is adopted. First, the number of subcarriers assigned to users and the transmission power allocation is derived by a greedy algorithm. Second, subcarriers are then assigned to users afterward using the Hungarian algorithm. Simulation results show that the total throughput in the cell can be improved by 90%, and the potential gain loss due to sub-optimal approaches is 10%.

Alternatively, fairness is pursued in [86] by maximizing the minimum user throughput, or shortly max-min user throughput. In that work, it is first assumed that each subcarrier can be shared among multiple users. By doing that, the authors can derive a convex problem and greatly reduce the complexity of the original max-min problem. Based on that, a sub-optimal greedy algorithm is then used to derive allocation schemes. Throughput gains achieved by the proposed heuristic approach are shown to be very close to one of the optimal solutions. This work, though, assumes that users have the same QoS requirements, which is not the case for practical systems.

Another approach for enforcing fairness is to maximize the weighted-sum rate like in [87]. The main difference compared to the traditional sum-rate maximization is that user throughput is weighted to assure that users with high priority can receive more resources, and vice versa. By enforcing a constant transmission power allocation, a further reduction in computational complexity can be achieved. The work, however, neglected the discussion on how to select proper weights in practical systems.

In [88], proportional fairness is assured by imposing a set of non-linear constraints into the sum-rate maximization problem. In this work, proportional fairness based on fairness indices is formulated to provide an efficient way to prioritize users, instead of merely assigning arbitrary weights as in [87]. However, the proposed power allocation algorithm requires solving iterative non-linear methods, which are generally complex. The authors in [89] relaxed proportional constraints to propose a non-iterative method with significantly reduced complexity. Various other heuristics have been proposed to reduce the complexity (e.g., [90]–[93].

Suppose the relation between spectral efficiency and user fairness is defined in a higher layer of the networking stack. In that case, a better notion of fairness than a mere throughput can be derived in the form of utility functions. Such a DRA algorithm that utilizes information from other layers in the stack is said to adopt the cross-layer design. In [94], [95], the authors establish a theoretical framework and general algorithms for cross-layer optimizations. Various utility functions are adopted to bridge the QoS requirement in the Media Access Control (MAC) layer and the physical layer's transmission schemes. Consequently, corresponding utility-based OP are introduced. To solve those OPs, the subcarrier assignment is first derived by using a sorting search algorithm. Then, the power adaptation is achieved by either a sequential linear approximation of the water-filling algorithm for the continuous rate formulation (i.e., Shannon equation) or a greedy power algorithm for discrete rate formulation (i.e., a limited set of available MCS). Numerical results show significant performance gains for cross-layer optimization.

In the meantime, the potential gain achieved by different optimal rate-adaptive algorithms is compared in [96]. It is shown that a dynamic scheme of transmission power, subcarriers, and MCS can improve the average throughput up to 100% compared to the static approach. Other combinations of dynamic subcarrier allocation or power allocation with adaptive MCS can also significantly improve the average user throughput.

3.2.2 DRA in the uplink

A large number of studies focus on exploiting the DRA approach in the downlink of OFDMA systems, a few others consider the uplink.

The sum-rate maximization problem for the uplink is considered in [97]. In this work, adaptive mechanisms of subcarrier assignment and power allocation are jointly considered. First, the adequate OP to find the optimal resource allocation is formulated. Second, a sub-optimal greedy algorithm based on the Karush-Kuhn-Tucker condition then allocates subcarriers and power. Simulation results show that the proposed algorithm produces almost near-optimal solutions. However, the continuous

Shannon's capacity instead of realistic models considering the discreet set of MCSs is adopted in this work. Besides, the proposed greedy approach requires an iterative algorithm, which is potentially computationally demanding, to realize the water-filling power allocation.

Similar to the work in [85] proposed for the downlink, an approach maximizing total cell throughput with constraints for minimum user throughput is introduced in [98]. The main difference between the two studies lies in multiple constraints for different users' maximal power allocation. To solve that problem, subcarrier assignment is achieved via a two-step algorithm. An initial subcarrier allocation considering per-user fairness is followed by the second allocation of residual subcarriers to increase the sum rate. However, it is not clear in this work how to select minimum requirements of user throughput.

A general framework based on utility functions to balance spectral efficiency and fairness is adopted in [99]. Via proposed utility functions, a notion of max-min fairness can be pursued. The authors manage to propose low-complexity greedy algorithms to achieve near-optimal results.

In general, a common drawback of proposed DRA algorithms in the uplink lies in the strong assumption of perfect synchronization between user signals. In literature, the study in [100] proposes an alternative approach to deal with imperfect synchronization in the uplink. In that work, the authors first propose to use short CP to cope with only delay spread. Shortened CP aims to reduce the overhead but leads to residual time offsets. Offsets in time and frequency domains are then mitigated by using appropriate resource allocation schemes. Especially, resource allocation includes the usage of GB in frequency. This approach is driven by the strong dependency of MAI on resource allocation as analytically formulated in [10]. An optimization model of resource allocation for the uplink is then introduced to maximize the minimum user throughput. The impact of MAI is, however, not included in the formulated OP. The proposed OP instead enforces a GB between two frequency regions of two users irrespectively to synchronization conditions. Consequently, the derived resource allocation scheme after solving the OP is fed to an algorithm, which re-assigns GBs to users if MAI is trivial. It is shown that the proposed heuristic approach can significantly improve the average throughput in the cell as well as the minimum user throughput. Not considering the impact of MAI in the OP is the main drawback of this work. In addition, the impact of different CP lengths is neglected.

To the best of our knowledge, no other studies consider imperfect synchronization or the mitigation of MAI as optimizing resource allocation. Thus, the work in [100] is the main baseline approach for our studies.

3.3 Cross-Layer Video Adaptation

Many studies in the literature target adaptive streaming from the client's perspective and model the end-to-end link as a black box (e.g., [101], [102]). That approach can perform generally well in wired networks, where channel states are relatively sta-

ble. However, those solutions tend to underperform in wireless networks [103], where the channel can fluctuate strongly and rapidly.

Pioneering researches in [6], [104], [105] demonstrate the benefit of cross-layer approaches for adaptive video streaming. In such systems, several adaptation strategies in different layers of the OSI architecture are jointly considered. However, potential performance gains comes at the cost of, among others, implementation complexity and communication overhead. Recently, due to the explosive video traffic growth, mobile network operators and service providers are increasingly forced to look for new ways to serve video traffic more efficiently, creating a strong incentive to consider cross-layer approaches [106].

Cross-layer approaches for single-user video streaming typically focus on packet scheduling, error protection, and video adaptation as maximizing the perceived quality (e.g., [107]). However, as multiple users compete for the bandwidth in the bottleneck link of the RAN, several performance problems concerning fairness, stability and the efficiency of resource utilization have been observed [8], [108]. Several studies in the literature have been conducted to resolve those problems.

A systematic framework for optimizing multiple streams over generic wireless networks is introduced in [109]. Importantly, the proposed cross-layer adaptation strategy allows users to aim at long-term video quality instead of immediate throughput. Video adaptation is formulated as a multiuser Markov decision process, and the objective is to minimize the total distortion of all streams. A learning algorithm is then introduced to solve the Markov decision problem. However, the model of resource allocation and channel throughput is highly abstracted. In particular, it is assumed that user throughput is a convex increasing function of the number of assigned frequency resources. In addition, resource allocation is abstracted as portions of the total bandwidth. So the time, frequency, and multiuser diversity of wireless channels and the potential gain of DRA are not considered. It is interesting to point out that the work, in the meantime, considers an extremely detailed video model with consideration of image frame types, the dependency between image frames, and latency deadlines.

A joint optimization of network resource allocation and video adaptation for HAS-based applications is considered in [110]. In this work, a VoD service using NLVC is assumed. Similar to the work in [109], a generic wireless link model based on a convex feasible rate region is adopted to model a fairly general class of network-related constraints. The highly abstracted model of resource allocation centering on resource share is again adopted.

Regarding OFDMA networks, the work in [8] introduces an in-network resource management framework, named AVIS, that targets HTTP-based adaptive video streams. NLVC is adopted in this work. The proposed framework consists of two main units, which are an allocator and an enforcer. First, the allocator selects, for each user, a properly targeted video bitrate and a resource share to accommodate the selected video bitrate. The desired selection aims to maximize the total utility of all users. In the second step, the enforcer schedules video packets of streams following allocated resource shares. However, the discrepancy between the estimated

throughput as the input for the allocator and the actual throughput of the wireless channel is not explicitly discussed.

A joint consideration of resource allocation and video adaptation for LTE networks is considered in [111]. Similar to the aforementioned studies, adopted resource allocation models are limited to the concept of resource share, and the achieved throughput is the product of the share and the maximal achievable bitrate when a user takes all resources. Based on such models, the selection of video bitrates and resource allocation is derived by solving an OP that maximizes the total network utility. This approach is commonly known in the literature as Network Utility Maximization (NUM). Knowing the selected video representations, a proxy between BS and the remote media server can intercept clients' requests and rewrite them according to video adaptation decisions. Perceived quality is modeled by a simple linear function that maps transmission rate to MOS. However, the work does not take account of the resource allocation process to enforce the required transmission rate needed for selected presentations.

A more detailed model of resource allocation is considered in [112]. In this paper, a joint optimization of resource allocation and packet scheduling is the focus. Similar to [111], the NUM approach is adopted to capture some notion of fair maximization of user quality. Particularly, the network utility is formulated as the importance of scheduled video packets, defined as the distortion when packets are missing. Importantly, by exploiting the Lagrange dual decomposition method, the optimal solution can be derived. However, the problem of video quality selection is not explicitly considered. In addition, the model of perceived quality is based on distortion. It does not consider some important factors that strongly affect QoE like video stalls.

Some researches have been conducted to consider video streaming applications using LVC. Interestingly, several of them strive to exploit the nature of LVC to support low-delay streaming services. For instance, the authors in [113] propose a cross-layer design to optimize video quality by jointly adapting source rate by dropping enhancement layers. In the lower layer, MCSs are selected with the consideration of video content. Via the adaptation of MCS, the unequal error protection per layer is included. Roughly speaking, lower MCS schemes, which are more robust to interference and noise, are used for more important layers and vice versa. The importance of layers is determined based on perceptual loss estimation. The problem of resource allocation is, however, neglected in this work.

In [114], the problem in DASH based systems is studied. In this work, end-to-end distortion and buffer level are considered in the considered video adaptation algorithm. Besides, channel state is considered as allocating resources to users. A utility function based on the average downloading time is used. Based on that utility function, an OP following NUM approach is formulated to maximize the sum weighted utility over all users. The problem is then transformed using the Lagrangian dual decomposition method and, consequently, solved by a sub-gradient algorithm. A serious weakness of this work, however, is to assume that video adaptation and resource allocation operate on the same timescale.

Another work in [115] proposes a cross-layer solution for LVS in OFDMA networks. The work aims to maximize the aggregate ergodic (average) throughput subject to

the tolerable distortion difference between streams. By adopting the ergodic throughput, the design is significantly simplified, but, the potential throughput gain by the exploitation of channel diversity is not considered.

Most studies on cross-layer designs for adaptive video focus on the downlink. One specific aspect of the uplink that requires additional study stems from constraints on transmission power. Unlike in the downlink, each user in the uplink has a discreet budget of the maximal power it can emit signals. For instance, the work in [116] considers multiuser streaming service in the uplink. The proposed algorithm strives to allocate transmission power and frequency resources to users accordingly to instantaneous channel states and the rate-distortion information of the video stream. A cross-layer OP is proposed to minimize the sum of distortion rates over all users. The main problem of the work however lies in the strong assumption that the channel slowly varies, specifically, at the same interval of video. In addition, the problem of potential residual offsets is completely neglected.

3.4 Scope of the Thesis

Dealing with imperfect synchronization is generally a critical task in the uplink of OFDMA networks. The typical approach is to estimate and then counteract synchronization offsets, thus the name ECBA. However, adopting ECBA leads to a significant overhead required for estimation algorithms. In addition, there is always a chance that residual offsets exist, which can significantly deteriorate user signals.

One question that emerges here is whether one can cope with synchronization offsets more effectively than ECBA. One possible answer is to exploit DRA for mitigating MAI and improving user throughput simultaneously. The motivation for this approach is twofold. On the one hand, MAI depends strongly on resource allocation, including the usage of CP and GB as well as the assignment of resources to users. Thus, MAI can be suppressed when using suitable resource allocation schemes. On the other hand, the exploitation of channel diversity can provide valuable throughput gain, which can help improve video quality.

There has been so far very little understanding about integrating the mitigation of MAI in resource allocation optimization models. This thesis aims to address two important issues in this context. First, far too little attention has been paid to studying the effectiveness of different combinations of CP and GB regarding MAI and user throughput. Unlike GB, which can be dynamically assigned, the selected length of CP is fixed. Second, it is not clear how to derive optimal resource allocation schemes that can maximize the system performance while effectively suppressing MAI.

In Chapter 4, the impact of different combinations of GB and CP on MAI in static conditions of channel gains, offsets, and frequency assignments is first studied. The result of this investigation allows us to select a suitable CP length for dynamic scenarios. Consequently, an optimization model that explicitly includes the usage of GBS and MAI on user throughput is formulated. The objective is to fairly improve the throughput of multiple users. Via efficient mathematical transformations, the formulated OP can be numerically solved by common optimization software solvers.

Several sub-optimal heuristics algorithms are also provided to reduce computational complexity.

It is well understood that the potential throughput gain and the fast-paced adaptation (in the timescale of milliseconds) of DRA algorithms can be exploited in cross-layer solutions to improve the performance of low-delay streaming services. Existing cross-layer video adaptation algorithms that exploit DRA mostly assume perfect synchronization in the uplink achieved by, implicitly, adopting ECBA. As a result, proposed OP do not consider residual synchronization offsets. In addition, again, a large overhead needed for ECBA causes the inefficiency of resource usage. Another issue is that the problem of resource allocation is typically limited to deriving users' resource shares; thus, channel diversity is not exploited to achieve throughput gains.

This thesis addresses those issues by proposing novel cross-layer video adaptation approaches based on the optimization model of resource allocation in Chapter 4. Two adaptive streaming paradigms are considered in this thesis, which bases on NLVS and LVS. The two paradigms distinguish from each other in their video adaptation principles that require separate studies. These two paradigms are both supported by the open standard DASH. While the former has been widely deployed in streaming services like YouTube and Netflix, the former is expected to be more popular in the near future.

Non-Layered Video Streaming: A specific feature of NLVS is: a suitable quality of each video segment must be selected before the transmitter delivers the corresponding data of that segment. Due to that, an important task of any video adaptation algorithms is to foresightedly optimize video quality selections. The selected bitrates, on the one hand, does not exceed the future link rate to avoid video stalls but, on the other hand, maximizes user QoE. So far, no research has been found that takes into account potential throughput gains achieved by DRA as deriving the optimal quality selection.

Once the quality of a video segment is determined, the mobile network strives to deliver the selected video representation before its deadline. In low-delay services, the tight deadline gives a thrust of efficient resource allocation strategies that can confront fading channels and achieve the requested throughput. Note that one video segment needs to be conveyed over a large number of OFDMA resource units due to its large amount of data. Therefore, the long-term QoE generally needs to be pursued over a sequence of resource allocation instances. It is unclear how to drive the short-term OP of resource allocation to meet the long-term video adaptation goal.

In Chapter 5, a novel video adaptation algorithm consisting of two components is proposed. The first component, a video quality selector, takes into account (i) a proper estimation model of future throughput and (ii) the potential throughput gain via DRA as selecting appropriate video qualities. This approach is expected to enable the selection of higher bitrates.

Furthermore, two realistic use cases are considered, which are live streaming with Hard Latency Constraint (LSH) and with Soft Latency Constraint (LSS). In the first use case, clients skip segments that miss their playback deadlines so that a given strict upper bound on the live latency can be met. In the second use case, clients prefer to play out the content without gaps. To that aim, whenever a segment cannot be

delivered by its playback deadline, the playback is halted until the playback buffer is raised above a certain threshold, effectively increasing the latency. In addition, the proposed approach takes into account not only dynamically changing network conditions but also, in the LSS use case, individual buffer levels.

Given quality selections as the output of the first component, the second component, a DRA algorithm, strives to deliver the requested video qualities. To that aim, a sequential process of adapting resource allocation schemes to instantaneous wireless channel states is performed to gradually match users' demands with achieved link rates. By optimizing DRA, valuable throughput gains by exploiting the channel diversities to efficiently combat the wireless channel's fluctuations. Especially, we introduce solutions for both the downlink and the uplink. For the uplink, imperfect synchronization among users is explicitly considered by incorporating the mitigation of the MAI as deriving DRA schemes.

Layered Video Streaming: Unlike NLVS, where video adaptation is available only at the borders of video segments, LVS provides many more adaptation points. In other words, a receiver can decode a segment after dropping some enhanced video layers of that segment. This feature gives rise to a possibility to tightly couple video adaptation with resource allocation. However, to the best of our knowledge, no research has considered the practical problem of imperfect synchronization as developing cross-layer video adaptation algorithms for low-delay LVS.

In Chapter 6, a cross-layer approach can tightly integrate long-term QoE objectives into a series of quality-driven DRA. At each DRA step, resources are allocated to users according to their utility determined by QoE constraints and quality fairness. By doing that, the proposed approach can react to the channel's fluctuations at the pace of few milliseconds and thus better support low-delay streaming. To push the potential of this approach to the limit, the main focus is on an extreme use case, where the playback buffer is enough only for one video segment.

Chapter 4

MAI aware Dynamic Resource Allocation

In the uplink of OFDMA networks, imperfect synchronization between users' signals arriving at BS causes MAI. MAI can severely degrade user throughput.

This chapter presents a novel DRA approach that mitigates MAI and improves user throughput simultaneously. The idea is to derive proper resource allocation schemes, which can, on the one hand, assign resources to users that experience good channel gains. On the other hand, the mutual MAI, which strongly depends on resource allocation, is suppressed. The proposed approach explicitly considers the joint usage of GB and CP to mitigate MAI. The performance evaluation shows that, in comparison to ECBA, the proposed approach efficiently utilizes available wireless resources and provides valuable throughput gains. The main results of this chapter have been published in [117], [118].

4.1 System model

We consider one single cell as depicted in Figure 4.1. OFDMA presents itself in the downlink as well as in the uplink. Time Division Duplexing (TDD) is the duplex method. This selection is inspired by the wide adoption of TDD in commercial LTE and WiMAX systems. One important technical advantage of TDD is the flexibility to cope with asymmetric traffic. Within the cell, one BS locates at the center of the cell and communicates with multiple users. Among all associated users, M users are currently active, which means they exchange data with BS. Besides, the full buffer model is assumed in this chapter. It means users always have some data to send.

The notation is summarized in Table 4.1.

Radio Resource

The total bandwidth B at the center frequency f_c is available within the cell under consideration. In addition, we assume inter-cell interference can be reasonably neglected through proper frequency reuse patterns. Figure 4.2 illustrates an example

Table 4.1: Notation overview

M	Number of active users
B	Total available bandwidth
f_c	Center frequency
f_0	Frequency spacing between subcarriers
N_{sca}	Number of subcarriers, indexed by r and r'
N_{sch}	Number of subchannels, indexed by i and j
$\mathbb{S}_i$	Set of subcarriers belonging to subchannel i
S	Number of subcarriers per subchannel
$N = N_{\text{sca}}$	FFT and IFFT window size
$T_{\text{sym}} = 1/f_0$	OFDM symbol duration without CP
$T_{\text{sam}} = T_{\text{sym}}/N$	Sampling interval
$T_g = (v_1 + v_2)T_{\text{sam}}$	Length of CP
v_1	Proportion of CP to cope with synchronization errors
v_2	Proportion of CP to cope with channel delay spread
$P_{\text{max}}^{\text{TX}}$	Maximal transmission power generated by each user
$P_{r,m}^{\text{TX}}$	Transmission power of subcarrier r of user m
$P_{r,m}^{\text{RX}}$	Received power on subcarrier r sent by user m
$P_{i,m}^{\text{TX}}$	Average transmission power on subchannel i of user m
$P_{i,m}^{\text{RX}}$	Average received power on subchannel i of user m
$H_{i,m}$	Average channel gain of subchannel i of user m
$\text{MAI}_{r,m}^{r',m'}$	MAI caused by subcarrier r' of m' on subcarrier r of m
$\overline{\text{MAI}}_{i,m}^{j,m'}$	Average MAI caused by subchannel j of m' on subchannel i of m
$\bar{\gamma}_{i,m}$	Average SINR of subchannel i of user m
$b_{i,m}$	Number of bits sent on subchannel i by user m
τ_m	Time offset of user m with respect to BS's clock
θ_m	Frequency offset of user m with respect to BS's clock

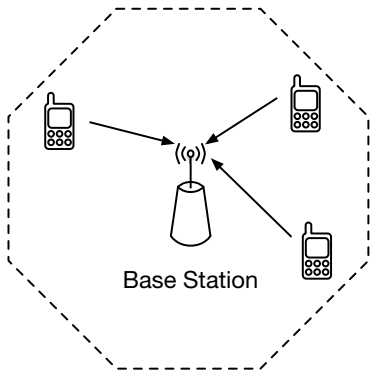

Figure 4.1: A single cell under consideration

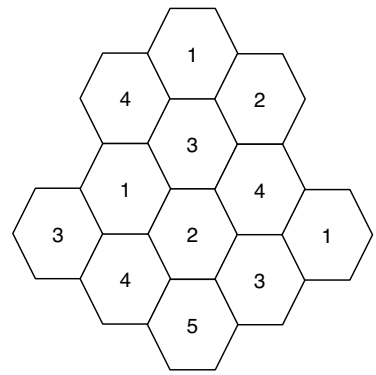

Figure 4.2: An example of frequency reuse patterns

of such frequency reuse patterns. The numbers in that figure denote the frequency partition indices.

The cell's available bandwidth is divided into N_{sca} subcarriers. Thus, the subcarrier spacing is $f_0 = B/N_{\text{sca}}$. By dividing bandwidth into narrow-band subcarriers, the coherence bandwidth of wireless channels typically spans over multiple subcarriers in frequency [26]. Consequently, a few adjacent subcarriers are grouped into one subchannel, and the size of each subchannel is selected to be equal the coherence time. By doing that, frequency diversity can be efficiently exploited while limiting the signalling overhead for addressing resource units [23]. Let $\mathbb{S}_i$ be the set of subcarriers in the subchannel i, and S be the number of subcarriers in each subchannel. The number of subchannels is an integer denoted by N_{sch}, and $N_{\text{sch}} = N_{\text{sca}}/S$. Subcarriers in a subchannel are indexed by r and r', while the indices of subchannels are i and j.

To guarantee the orthogonality between subcarriers, all symbol durations on all subcarriers have the same size, denoted by T_{sym}, and is the multiplicative inverse of subcarrier spacing, i.e., $T_{\text{sym}} = 1/f_0$. A CP, whose size is denoted by T_g, is prepended to each OFDMA symbol. The total symbol length, including CP, then yields: $T = T_{\text{sym}} + T_g$. The selection of the CP's length is discussed in the next section.

Without loss of generality, the size of FFT and IFFT modules, denoted by N, is selected to be equal to the number of subcarriers, i.e., $N = N_{\text{sca}}$. Consequently, the sampling interval is computed as $T_{\text{sam}} = T_{\text{sym}}/N_{\text{sca}}$.

Consequently, time is divided into OFDMA frames, each of which is consisted of one uplink and one downlink frame. The proportion between uplink and downlink frames is fixed. The number of OFDM symbols in each uplink frame is denoted by N_{sym}. Importantly, the duration of OFDMA frames is assumed to be smaller than the coherence time of wireless channels. This assumption can be justified by selecting adequate system parameters. For instance, assuming a vehicle moves at the velocity, denoted by v, of 30 km/h (i.e., 8.34 m/s) and a frequency band centers around f_c of 1900 MHz, then as it is shown in Chapter 2, the coherence time can be approximated by:

$$\frac{0.423}{f_D} = \frac{0.423c}{f_c v} = 8 \times 10^{-3}\ [s], \tag{4.1}$$

where f_D is the Doppler shift and c is the speed of light: $c = 3 \times 10^8$ [m/s]. Accordingly, it is recommended by, for instance, the WiMAX forum to choose the frame length of 5 ms for urban scenarios [72], which is smaller than the coherence time of 8 ms.

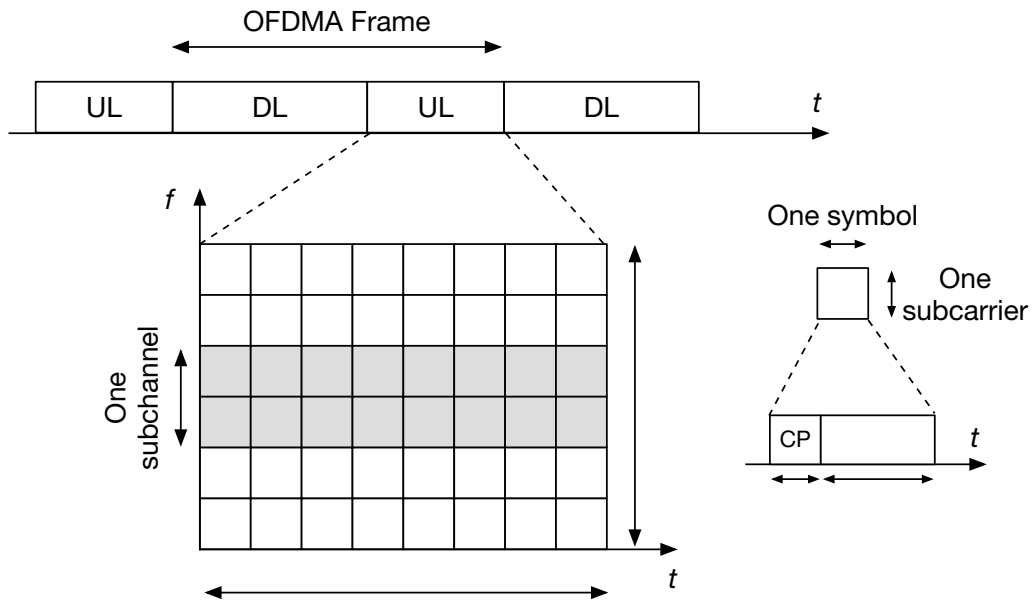

Figure 4.3: Illustration of resource structure

Finally, resource allocation is performed based on resource units, each of which consists of one subchannel in frequency and one uplink frame (i.e., N_{sym} symbols) in time. In this thesis, resource unit and resource block are used interchangeably.

4.1.1 Wireless Channel

Signals traversing wireless channels are attenuated, distorted, and shifted in time and frequency due to path loss, shadowing, and multipath fading. Since path loss and shadowing slowly fluctuate in time and frequency, as mentioned in Chapter 2, it is reasonable to assume that path loss h_p and shadowing h_s are constant for all subchannels and symbols in one OFDMA uplink frame.

On the contrary, multipath fading causes rapid and significant fluctuations of channel gains over small ranges of time and frequency. By selecting proper OFDMA parameters as discussed in the previous section, typically, multipath fading coefficients are constant over one OFDMA frame in time and subcarriers in the same subchannel.

Consequently, let $H_{i,m}(t)$ be the channel gain on subchannel i assigned to user m during the OFDMA uplink frame in time t, i.e., $0 \leq t \leq T_{\text{sym}} N_{\text{sym}}$. As discussed

in the previous paragraph, we can write $H_{i,m}(t) = H_{r,m}(t)$ for all subcarriers r in the subchannel i (i.e., $r \in \mathbb{S}_i$). For simplicity, the index of OFDMA frame t is omitted unless specifically required. Consequently, the received power can be written as $P_{i,m}^{\mathrm{RX}} = P_{i,m}^{\mathrm{TX}} H_{i,m}$.

In this thesis, we make two important assumptions about channel knowledge for the uplink. First, CSI of each resource unit in the uplink is available at BS for DRA algorithms. In practice, during the downlink, to correctly receive signals from BS, users need to estimate channel quality and offsets anyway. CSI of the downlink can be then sent to BS [72]. Moreover, it is assumed that the overhead required to carry channel knowledge to BS is negligible by using advanced compression techniques such as one proposed in [119] and [120]. In addition to the knowledge in the downlink, BS can perform channel estimation on few reference resource units at the beginning of the uplink frame [121]. The resource amount for signaling is normally negligible compared to user data.

Second, we also assume that CSI is not subject to delay or error. We justify this assumption by arguing that, as mentioned above, the coherence time of wireless channels can span over more than one OFDMA frame. In addition, it is also shown in [122] that even a completely stale CSI can still be very useful to derive efficient resource allocation. Evermore, the assumption of ideal channel knowledge can also be achieved by using promising channel prediction techniques (e.g., [123]). Consequently, the impact caused by outdated or erroneous CSI on resource blocks is not considered in this model.

4.1.2 Multiple Access Interference

Unlike in the downlink, signals arriving at BS in the uplink are the aggregate of elements sent from several users. Due to several reasons (including propagation delay, Doppler shift, and oscillator errors), user signals are shifted differently in time and frequency. Let τ_m $[samples]$ and θ_m $[Hz]$ be the offset in time and frequency, respectively. Note that here only the integer part of time offset is taken into account; the fractional part is included in channel coefficients.

The misalignment of user signals plagues users' orthogonality and causes MAI. An analytical model of MAI is formulated in [10]. We adopt that model in this thesis and summarize the final result in the following. Let $\Delta\tau$ and $\Delta\theta$ be the relative differences between user m and m' in the time and frequency domains, respectively. Then, MAI caused by subcarrier r' of user m' on subcarrier r of user m takes the following form.

$$\mathrm{MAI}_{r,m}^{r',m'} = \frac{P_{r',m'}^{\mathrm{RX}}}{N^2} \times \frac{A(r'-r, \Delta\theta, \Delta\tau)}{sin^2[\frac{\pi}{N}(r'-r+\Delta\theta)]} \tag{4.2}$$

where $A(.)$ is a complex function of relative offsets as well as the distance between subcarriers r' and r. In addition, $A()$ also depends on the CP's length. To show the closed form of the function $A(.)$, let N_p be the number of fading paths for all users, and let $T_g = (v_1 + v_2)T_{\mathrm{sam}}$, where v_2 is used to deal with delay spread, and v_1 aims

to protect signals from two-way time offset caused by propagation delay and clock errors. Consequently, function $A(.)$ has the following form.

- Case 1: $(-v_1/2 - v_2) > \Delta\tau \geqslant (-N + v_1/2)$:

$$A = \sum_{p=0}^{N_p-1} \alpha_p^m \big\{ sin^2\big[\frac{\pi}{N}(p - \Delta\tau - \frac{v_1}{2} - v_2)(\Delta\theta + r' - r)\big] + sin^2\big[\frac{\pi}{N}(p - \Delta\tau - \frac{v_1}{2} - v_2 - N)(\Delta\theta + r' - r)\big]\big\}$$

- Case 2: $(N_p - 1 - v_1/2 - v_2) > \Delta\tau \geqslant (-v_1/2 - v_2)$:

$$A = sin^2(\pi\Delta\theta) \sum_{p=0}^{v-|\Delta\tau|} \alpha_p^m + \sum_{p=v_1+v_2-\Delta\tau+1}^{N_p-1} \alpha_p^m \big\{ sin^2\big[\frac{\pi}{N}(p - \Delta\tau - \frac{v_1}{2} - v_2)(\Delta\theta + r' - r)\big] + sin^2\big[\frac{\pi}{N}(p - \Delta\tau - \frac{v_1}{2} - v_2 - N)(\Delta\theta + r' - r)\big]\big\} \tag{4.3}$$

- Case 3: $(-v_1/2) \geqslant \Delta\tau \geqslant (N_p - 1 - v_1/2 - v_2)$:

$$A = sin^2(\pi\Delta\theta) \sum_{p=0}^{N_p-1} \alpha_p^m$$

- Case 4: $(N_p - 1 + v_1/2) \geqslant (\Delta\tau > v_1/2)$ then:

$$A = sin^2(\pi\Delta\theta) \sum_{p=|\Delta\tau|+1}^{N_p-1} \alpha_p^m + \sum_{p=0}^{|\Delta\tau|} \alpha_p^m \big\{ sin^2\big[\frac{\pi}{N}(p - \Delta\tau + \frac{v_1}{2})(\Delta\theta + r' - r)\big] + sin^2\big[\frac{\pi}{N}(p - \Delta\tau + \frac{v_1}{2} + N)(\Delta\theta + r' - r)\big]\big\} \tag{4.4}$$

- Case 5: $(N + v_1/2) \geqslant \Delta\tau > (-N_p - 1 + v_1/2)$:

$$A = \sum_{p=0}^{N_p-1} \alpha_p^m \big\{ sin^2\big[\frac{\pi}{N}(p - \Delta\tau + \frac{v_1}{2})(\Delta\theta + r' - r)\big] + sin^2\big[\frac{\pi}{N}(p - \Delta\tau + \frac{v_1}{2} + N)(\Delta\theta + r' - r)\big]\big\}$$

where α_p^m be the average power received on path p of user m. For simplicity, it is assumed that $\alpha_p^m = 1/N_p$ for all users and all paths.

Since one subchannel is the smallest addressable unit in DRA algorithms, it is useful to define the average MAI per subchannel. Let $\overline{\text{MAI}}_{i,u}^{j,u'}$ be the average MAI caused by subchannel j of user u' on subchannel i of user u as follows:

$$\overline{\text{MAI}}_{i,m}^{j,m'} = \frac{1}{S} \sum_{r \in \mathbb{S}_i} \sum_{r' \in \mathbb{S}_j} \text{MAI}_{r,m}^{r',m'} \tag{4.5}$$

4.1.3 Signal to Noise plus Interference Ratio

We define the average Signal to Noise plus Interference Ratio (SINR) $\bar{\gamma}_{i,m}$ of subchannel i, which is uniquely assigned to user m, is given by

$$\begin{aligned} \bar{\gamma}_{i,m} &= \frac{P_{i,m}^{\text{RX}}}{\sigma^2 + \sum_{\forall m' \neq m} \sum_{\forall j \neq i} \overline{\text{MAI}}_{i,m}^{j,m'}} \\ &= \frac{P_{i,m}^{\text{TX}} H_{i,m}}{\sigma^2 + \sum_{\forall m' \neq m} \sum_{\forall j \neq i} \overline{\text{MAI}}_{i,m}^{j,m'}} \end{aligned} \tag{4.6}$$

where σ^2 is the thermal noise power.

4.1.4 Adaptive Coding and Modulation

Given a SINR value of subchannel, the capacity of that subchannel can be derived using the Shannon equation. Then, the number of bits that can be conveyed on subchannel i in one OFDMA symbol by user m, denoted by $a_{i,m}$, yields the following form.

$$a_{i,m} = T_{\text{sym}} \times S \times f_0 \times \log_2(1 + \bar{\gamma}_{i,m}) \tag{4.7}$$

In practice, however, only a fixed amount of MCS, denoted by K, is available. In this thesis, ACM is adopted. Thus the highest MCS scheme is selected subject to the tolerable BER.

Analytically, let Th_k [dB] be the minimum SINR required to use MCS k. Accordingly, let B_k [*bits*] be the number of bits can be sent per one subcarrier in one symbol when MCS k is chosen. Then we use function $F(.)$ to reflect the selected MCS, which can satisfy the predetermined target error probability P_{err}. For example, modulation scheme 64-QAM with coding rate 2/3 allows the transmitter to send a total of 6 bits and, thus, 4 payload bits per one OFDMA symbol. In general, due to the discreet set of available MCS, function $F(.)$ is a piece-wise constant function over SINR.

Finally, we have:

$$b_{i,m} = S \times F(\bar{\gamma}_{i,m}, P_{\text{err}}) \tag{4.8}$$

4.1.5 Medium Access Control

Before users can use radio resources, they need to be informed about the resource allocation scheme. This task is carried out by a resource allocator, typically located at BS. The resource assignment can be derived following static methods such as blocking or interleaving assignments. Alternatively, the allocator can adapt resource allocation schemes to instantaneous system conditions. As mentioned above, perfect knowledge of CSI and user offsets are assumed to be available for DRA algorithms.

After deriving resource allocation schemes, BS sends that information to users in the downlink. Transmission of that info is assumed to be done on a separate control channel and not considered in this thesis. In addition, the signaling channel is assumed to be always error-free and is never delayed.

4.2 Problem Statement and Proposed Approach

It has been well known that user throughput in the uplink of the OFDMA systems can be strongly degraded by MAI [10]. As discussed in Chapter 3, the most common way to deal with MAI is collectively referred to as ECBA. However, ECBA has several disadvantages including large overhead and residual offsets.

The question that arises here is whether one can reduce the overhead required for MAI mitigation and then dedicate more resources for user data. To address this question, we resort to efficient resource allocation algorithms. This approach is motivated by the fact that MAI strongly depends on the assignment of frequency subchannels to users, which can be seen in Section 4.1.2. Especially, we make two important remarks as follows.

- MAI caused by subchannel j of user m' on subchannels i of other user m (i.e., $\overline{\mathrm{MAI}}_{i,m}^{j,m'}$) is proportional to its received power at BS (i.e., $P_{j,m'}^{\mathrm{RX}}$). Therefore, the better the channel quality on the interfering subchannel j, the stronger MAI is caused on other subchannels. However, from the perspective of user m', worse received power means a reduction of throughput. On the other way, MAI is reduced when the received power of the interfering subchannel i is weaker.

- The extreme case is when the interfering user generates zero power on the assigned subchannel j, then obviously MAI on other subchannels becomes zero. The subchannel, in that case, is used as GB.

Consequently, the development of efficient resource allocation strategies concerning MAI mitigation and user throughput in the OFDMA uplink is the main focus of this chapter. In our approach, less resource is used for estimation and correction algorithms. As a result, residual offsets, however, exit. The impact of MAI is then minimized via proper resource allocation schemes.

To that aim, we first focus on a simple system with static conditions of wireless channels and synchronization errors. It means channel is assumed to be flat over all subchannels and symbols, and time and frequency offsets are constant. Then a

static resource allocation scheme using static usage of GB and CP is of the most interest. By doing that, the impact of random channel variations can be excluded from the problem of MAI mitigation. The goal of this work is to investigate the impact of different combinations of GB and CP on user throughput under the same synchronization error condition. This investigation is crucial since, while the GB can be dynamically assigned, the length of CP is static. The selection of CP length thus needs to be elaborated.

Next, resource allocation algorithms that can deal with both MAI and varying wireless channels are investigated. The goal of such algorithms is to find proper resource allocation schemes that can maximize user throughput, which is a function of not only wireless channel states but also synchronization offsets. Intuitively, it is expected that the size of GB should be flexible such that GBs protect each user signal individually (from others). Therefore GBs are assigned dynamically with respect to instantaneous channel conditions.

Two main challenges, which are two trade-offs, then emerge.

- **Assignment trade-off**: Assigning a good/bad subchannel i to user m obviously increases/decreases the m-th user's throughput (see (4.6)), but increases/decreases $\overline{\text{MAI}}_{j,m'}^{i,m}$ causing on all other subchannel j of all other users $\forall m' \neq m$ (see (4.2)), thus decreases/increases the m'-th user's throughput.

- **GB trade-off**: Taking some subchannels from a badly-synchronized user m' and setting them as GBs reduces MAI on other subchannels of all other users $m \neq m'$. However, that mitigation of MAI comes at the cost of frequency wastage and, thus, throughput losses. Therefore, considering only channel quality without consideration of MAI might increase MAI and thus lead to insufficient resource allocation schemes.

The problem of the two aforementioned trade-offs can be formulated as mathematical OPs. However, due to the discrete nature of the assignment problem, where each resource unit can be assigned to only one user, and only one MCS is selected, desired OPs tend to have highly complex formulations; the basic problem of DRA is proved to be NP-hard in [13]. Consequently, we face the problem of solving such OPs to find optimal resource allocation schemes. We aim to develop efficient mathematical transformations to solve those OPs and derive efficient resource allocation schemes. Furthermore, to reduce computational complexity, we also consider sub-optimal heuristic approaches.

4.3 MAI Mitigation via Static Resource Allocation

In this section, we consider the impact of static usage of guards in frequency and time domains, i.e., GB and CP, on MAI and user throughput. Based on the analytical model shown in Section 4.1.2, we make the following remarks about the usage of CP.

- CP protects user signals against only time offsets and does not have any effects on frequency offsets. Particularly, with only time offsets, the negative impact of

MAI is fully mitigated when CP's length is greater than two times the maximal time offset. However, with the existence of frequency offsets, increasing the length of CP further does not help mitigate MAI but only increases resource wastage.

- Due to the nature of OFDMA, CP is static, and its length is fixed in advance. Hence one cannot dynamically adapt CP's length to, for instance, synchronization errors during the transmission. Consequently, CP might be either too long (i.e., $v_1 > 2\Delta\tau$), which leads to throughput losses, or not long enough (i.e., $v_1 < 2\Delta\tau$), which results in insufficient protection.

- Another problem is that CP cannot be parameterized differently for different subchannels or users. In other words, the duration is the same for all users, although users might be differently asynchronous in time; hence, CP cannot cope with users' offsets in time individually.

Unlike CP, a subchannel can be dynamically set as GB only when needed. Also, assigning some subchannels as GBs can deal with MAI, no matter if MAI is caused by time offsets or frequency offsets, or both. That is because, as seen in (4.2), $P^{\mathrm{RX}}_{r',m'} = 0$ results in $\mathrm{MAI}^{r',m'}_{r,m} = 0$ regardless of time and frequency offsets (i.e., $\Delta\tau$ and $\Delta\theta$).

The question is then how to select efficient lengths to counteract the impact of time offsets, and then using GB to deal with time and the frequency offset residuals. To answer that question, we consider multiple static combinations of GB and CP. We then compute MAI and user throughput based on the adopted analytical model.

To focus on the impact of guards on MAI, it is useful to neglect the fluctuation of channel conditions and the variation of time and frequency offsets from the model in Section 4.1. More specifically, it is assumed that the wireless channel is flat over frequency and time for all users, and time and frequency offsets are static. In this case, the model of MAI and user throughput become deterministic.

In addition, it is assumed there are only two users in the cell denoted by m_1 and m_2. The user m_1 takes some subcarriers located in the center, i.e., around f_c. Other subcarriers (on two sides) are assigned to user m_2 that cause MAI on neighbors. Two users are not perfectly synchronized in both time and frequency. As a result mutual MAI exists and leads to losses in SINR and, consequently, in user throughput.

Consequently, given a specific selection of CP and GB, MAI and SINR can be computed directly from (4.2) and (4.6), respectively. For simplicity, the continuous rate model, i.e., the Shannon capacity as shown in Section 4.1, is used for the performance evaluation in this section. In addition, since the resource allocation scheme is static, the overhead for signaling the assignment scheme is not relevant, thus the investigation in this section can be performed at the granular level of subcarrier instead of subchannel. The main parameters are shown in Table 4.2. Note that, for CP, the length of v_2 is selected to equal the maximal delay spread; v_1 takes different lengths for different scenarios.

First, the impact of using only CP on SINR are shown in Figures 4.4, 4.5, and 4.6. When frequency offset is zero (i.e., $\Delta\theta = 0$), as shown in Figure 4.4, the length

Table 4.2: System parameters.

Parameters	Values
Number of users	$M = 2$
Number of subcarriers	$N_{\text{sca}} = 128$ (indices from 1 to 128)
Subcarriers assigned to user m_1	with index in $[48, 80]$
Subcarriers assigned to user m_2	with index in $[1, 47]$ and $[81, 128]$
Transmission power on subcarrier	0 dBm (i.e., 1mW)
Number of propagation paths	$N_p = 16$
Proportion of CP to deal with delay spread	$v_2 = N_p$
Channel gain of the flat wireless channel	-90 dB
Thermal noise power	-133 dBm

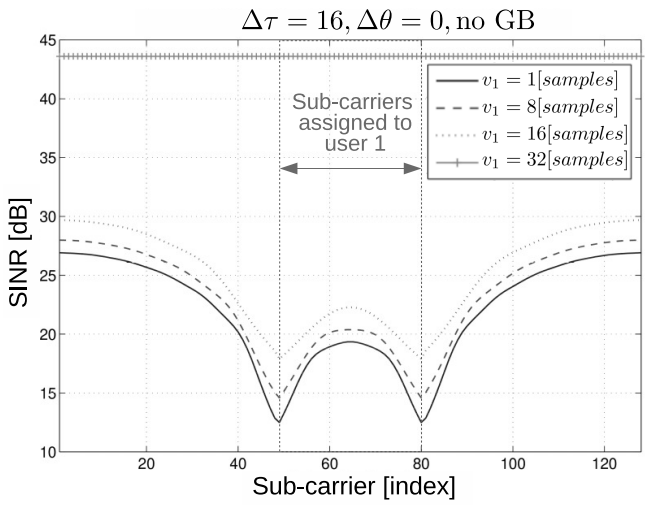

Figure 4.4: Impact of CP on SINR in case of time offsets only

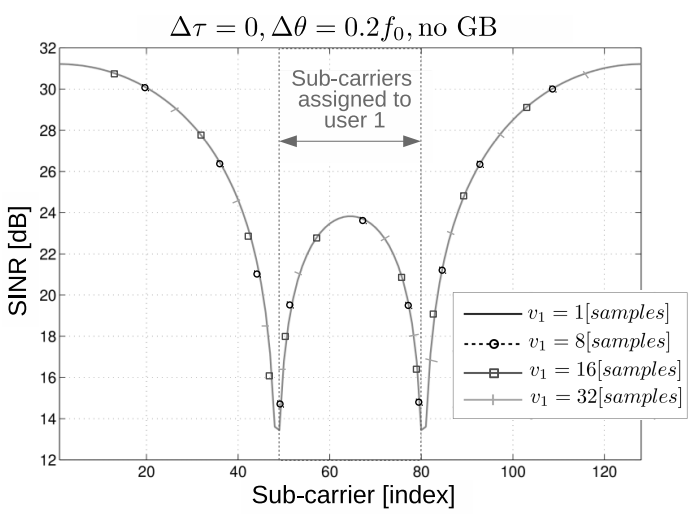

Figure 4.5: Impact of CP on SINR in case of frequency offsets only

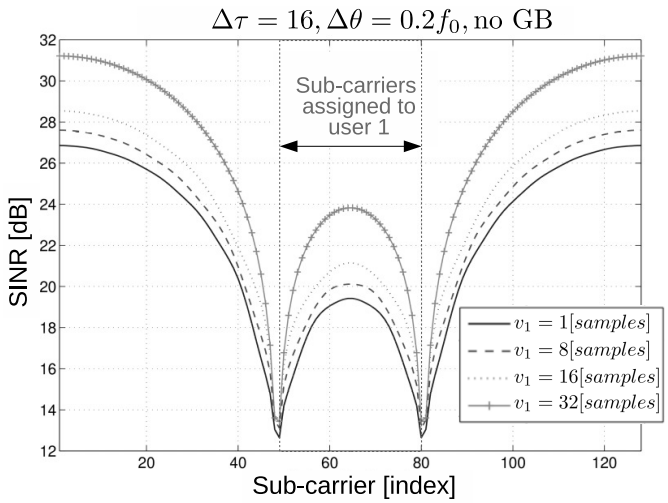

Figure 4.6: Impact of CP on SINR in case of time and frequency offsets

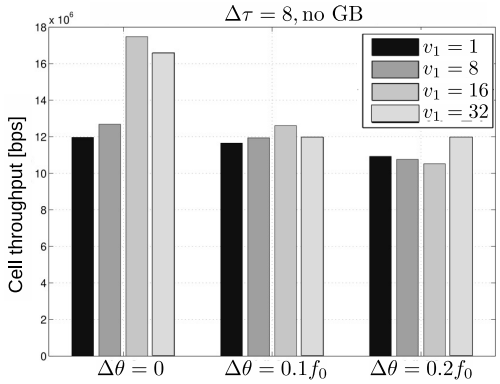

Figure 4.7: Impact of CP on cell throughput

of CP has strong impact on SINR. Roughly speaking, the longer the CP, the less MAI and the higher SINR. Especially, when v_1 equals two times of time offset (i.e., $v_1 = 2\Delta\tau$), the negative impact on SINR caused by time offset is fully mitigated. Therefore choosing CP equal to two times of time offset would be the best choice when no frequency offset is present. It is however important to note that when v_1 exceeds $2\Delta\tau$, SINR is not improved (since MAI cannot take a negative value), thus that extra proportion leads to resource wastage.

On the other side, in case only frequency offset $\Delta\theta$ is not zero, CP has no impact on SINR as shown in Figure 4.5. Thus the longer v_1, the more wastage CP causes. Similarly, with the existence of both time and frequency offsets, as shown in Figure 4.6, long CP cannot fully cope with MAI.

Regarding the overall cell throughput, results in Figure 4.7 demonstrate that as frequency offset increases, the optimal length of CP changes, and eventually, one has to choose the minimum length CP.

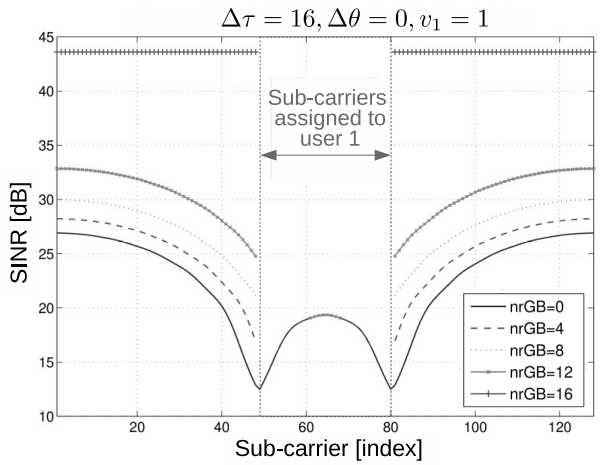

Figure 4.8: Impact of GB on SINR in case of time offsets only

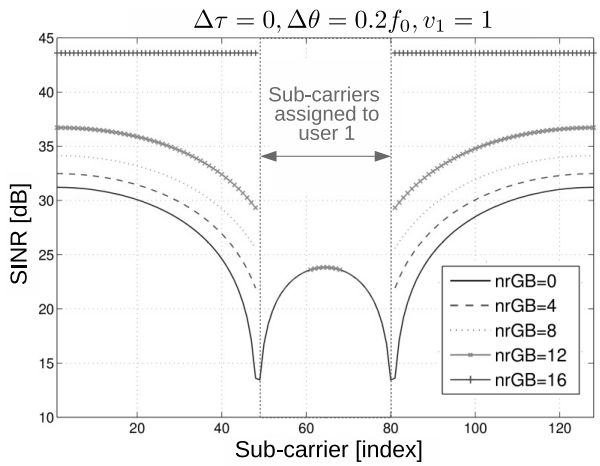

Figure 4.9: Impact of GB on SINR in case of frequency offsets only

The impact of using only GBs (i.e., $v_1 = 1$) on SINR is presented in Figures 4.8, 4.9 and 4.10. It can be easily seen that, unlike CP, GB can fight against both time and frequency offsets. In the extreme case, assigning all subcarriers of the first user as GBs (i.e., number of GB is 16 subcarriers on each side) means no transmission from user 1, and thus no MAI on the second user's signals (i.e., SINR values of all subcarriers of user 2 reach maximal). In general, GBs can protect user signals from frequency offset and time offsets. However, the mitigation of MAI and the improvement of SINR are obtained at the cost of resource wastage. Note that as received power is zero, SINR in logarithm scale reaches negative infinity and thus not shown in the figures.

More specifically, as no frequency offset is present, Figure 4.11 shows it is still better to use long CPs rather than GBs. However, one can still combine short CPs and long GBs to achieve a better cell throughput. But as soon as frequency offsets occur, using GBs can also improve the cell throughput. For instance, when time offsets are non zero, using GBs with short CPs is the best choice as it can be seen in

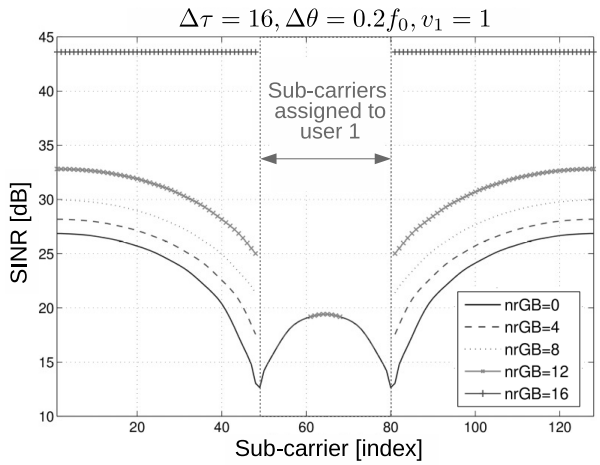

Figure 4.10: Impact of GB on SINR in case of time and frequency offsets

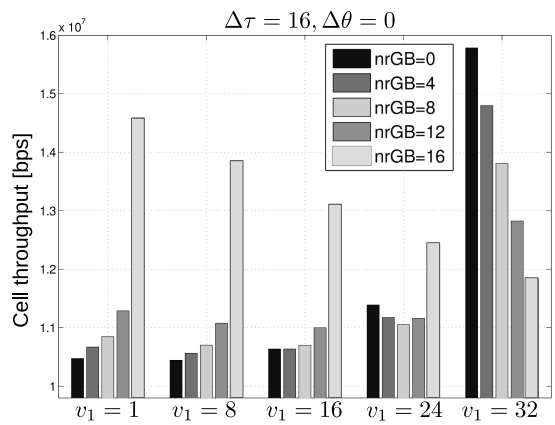

Figure 4.11: Cell throughput in case of time offsets only

4.12. In the presence of both frequency and time offset, one can see in Figure 4.13 that using GBs with short CPs is still the best choice concerning cell throughput.

It is important to mention that using cell throughput as the criteria for performance evaluation neglects fairness between users. For instance, in Figure 4.14, it can be seen that the maximum cell throughput is achieved when the throughput of another user is zero.

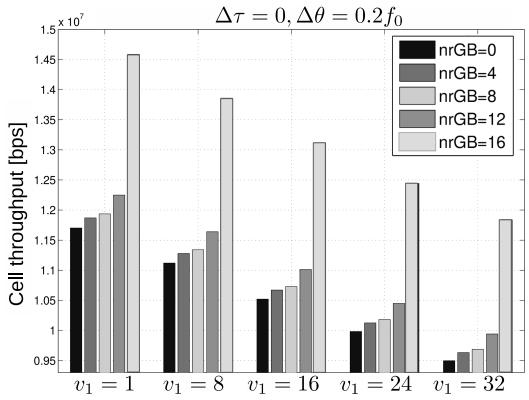

Figure 4.12: Cell throughput in case of frequency offsets only

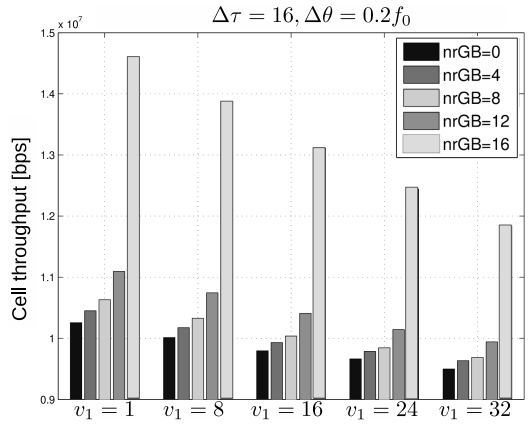

Figure 4.13: Cell throughput in case of time and frequency offsets

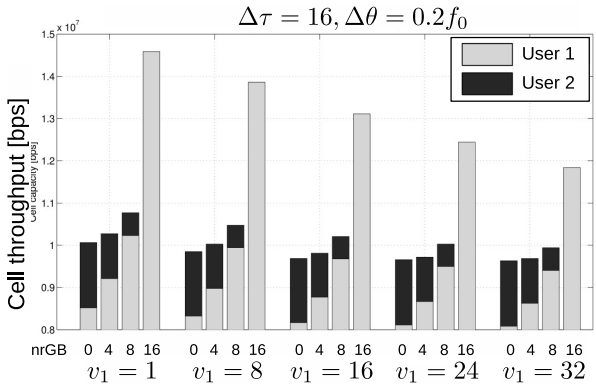

Figure 4.14: User and cell throughput in case of time and frequency offsets

4.4 MAI Aware Dynamic Resource Allocation

The investigation in the previous section has shown that proper static usage of GBs and short GIs can outperform a mere adoption of long GIs concerning user throughput. In addition, the analysis exposes the need for dynamic algorithms to efficiently assign resources to users and deal with MAI. Especially GBs should be used dynamically to protect users' signals only when needed. Efficient resource allocation schemes should also exploit random variations of fading channels in time, frequency, and multi-user domains to improve user throughput.

To that aim, we face two trade-offs as mentioned in Section 4.2, which are in resource assignment and GB usage. To resolve those challenges, proper OPs are formulated with the goal to (i) fairly maximize throughput of multiple users and, at the same time, (ii) suppress mutual MAI. Mathematical transformations are also proposed to derive equivalent OPs, which can be solved numerically more efficiently. Besides, sub-optimal OPs are also introduced to reduce computational complexity.

4.4.1 Basic optimization problem

This section is dedicated to formulating the general mathematical problem that aims to maximize the minimum (shortly max-min) user throughput.

First, let $x_{i,m}$ be a binary variable representing the assignment of subchannel i to user m, $x_{i,m}$ takes 1 if user m takes subchannel i and 0 if not. Consequently, the throughput gain user m achieves on subchannel i can be written as:

$$x_{i,m} F(\bar{\gamma}_{i,m}, P_{\text{err}}) \tag{4.9}$$

where the average SINR of subchannel in decibels takes the form

$$\bar{\gamma}_{i,m} = 10 \log_{10}\left(\frac{P_{i,m}^{\text{RX}}}{\sigma^2 + \sum_{\forall m' \neq m} \sum_{\forall j \neq i} \overline{\text{MAI}}_{i,m}^{j,m'} x_{j,m'}}\right) \tag{4.10}$$

To derive the analytical model of max-min user throughput problems, an auxiliary variable, denoted by ϵ, is defined as the lower bound of user throughput. Then the general formulation of the max-min problem is described as follows:

$$\begin{aligned} &\max \quad \epsilon \\ \text{subject to (s.t.)} \quad a) \quad &\sum_{i=0}^{N_{\text{sch}}-1} x_{i,m} F(\bar{\gamma}_{i,m}, P_{\text{err}}) \geq \epsilon, \quad \forall m \\ b) \quad &\sum_{m=0}^{M-1} x_{i,m} \leq 1, \quad \forall i \end{aligned} \tag{4.11}$$

The first constraint in (4.11) assures that each user can achieve a throughput above the lower threshold ϵ. Therefore the objective function, which is to maximize

ϵ, in coupling with the first constraint essentially reflects the max-min policy. In addition, the second constraint ensures that each subchannel is assigned to at most one user or left un-modulated and thus set as a GB.

Essentially, the basic OP in (4.11) is non-continuous and non-convex due to binary variables $x_{i,m}$ and the non-linear function $F(.)$ mapping SINR to throughput. This problem belongs to the Mixed-Integer Programming (MIP) class. It is mixed-integer since the formulation contains integer variables $x_{i,m}$ as well as continuous variables ϵ.

In general, it is challenging to solve MIP problems. Several works like [13], [79], [124] have proven that such OPs are **NP-hard**. The term NP-hard is from the computational complexity theory and stands for non-deterministic polynomial. Roughly speaking, no known algorithm can solve this problem in polynomial time. Therefore, a major research focus in the literature is to develop algorithms that can effectively solve the problem concerning computational complexity. One way to measure the complexity of algorithms is based on solving time.

4.4.2 Equivalent optimization problem

Solving OP (4.11) is challenging. Next, we strive to transform that problem into an equivalent problem, which can be solved more efficiently.

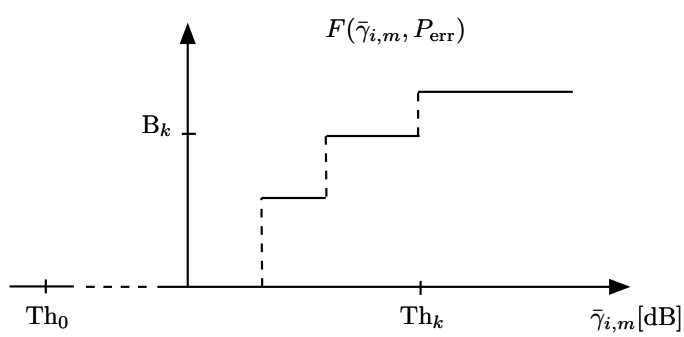

Figure 4.15: Adaptive Coding and Modulation function $F(.)$

First, we address the discreet and non-linear function $F(.)$. In this context, we adopt the well-known mathematical transformation method using piece-wise linear functions to transform function $F(.)$ to an equivalent and linear form.

Let K be the number of available MCSs, each of which requires a minimum SINR threshold Th_k to satisfy tolerable error rate P_{err}, and allows the transmitter to send B_k payload bits per symbol per subcarrier.

Adopting ACM, for a given SINR value, the highest MCS scheme is selected subject to P_{err}. Figure 4.15 illustrates the adaptive selection of MCS. Essentially, The dynamic loading algorithm of ACM then can be written as follows:

$$
\begin{aligned}
&\text{for } (\forall i, k, m) \\
&\quad \text{if } (\mathrm{Th}_k < \bar{\gamma}_{i,m}) \text{ then } F(\bar{\gamma}_{i,m}, P_{\mathrm{err}}) = B_k \\
&\text{end}
\end{aligned}
\tag{4.12}
$$

Further, we have:

$$
\begin{aligned}
& \mathrm{Th}_k \leq \bar{\gamma}_{i,m} \\
\Leftrightarrow \quad & \mathrm{Th}_k \leq 10 \log_{10} \left(\frac{P_{i,m}^{\mathrm{RX}}}{\sigma^2 + \sum_{j,m'} \overline{\mathrm{MAI}}_{i,m}^{j,m'}} \right) \\
\Leftrightarrow \quad & \Lambda_{i,m,k} \geq \sum_j \sum_{m'} \frac{\overline{\mathrm{MAI}}_{i,m}^{j,m'}}{\sigma^2}
\end{aligned}
\tag{4.13}
$$

where

$$
\Lambda_{i,m,k} = 10^{(\tilde{\gamma}_{i,m} - \mathrm{Th}_k)/10} - 1 \tag{4.14}
$$

and $\tilde{\gamma}_{i,m}$ is the SNR of subchannel i of user m and takes the form:

$$
\tilde{\gamma}_{i,m} = \frac{P_{i,m}^{\mathrm{RX}}}{\sigma^2} \tag{4.15}
$$

To make the mathematical transformation valid, it is necessary to add an auxiliary MCS scheme representing the case when $\bar{\gamma}_{i,m}$ is too small to use any MCSs; it means $B_0 = 0$ when $\Lambda_{i,m,0} \leq \gamma_{i,m} < \Lambda_{i,m,1}$, $\forall i, m$. Th_0 is a sufficiently small constant; so that $\Lambda_{i,m,0}$ has to be larger than all possible values of $\sum_j \sum_{m'} \overline{\mathrm{MAI}}_{i,m}^{j,m'} x_{j,m'}$.

Let $z_{i,m,k}$ be an integer optimization variable, which represent the selection of MCSs. $z_{i,m,k}$ takes 1 if MCS k is chosen for subchannel i of user m and 0 if not. Then (4.12) can be written as:

$$
b_{i,m} = \sum_{k=0}^{K} z_{i,m,k} B_k \tag{4.16}
$$

together with

$$
\sum_{k=0}^{K} z_{i,m,k} \Lambda_{i,m,k} \geq \sum_j \sum_{m'} \frac{\overline{\mathrm{MAI}}_{i,m}^{j,m'}}{\sigma^2} \tag{4.17}
$$

we have the formulation in the following form:

$$
\begin{aligned}
& \max \quad \epsilon \\
s.t. \quad & a) \sum_{i=0}^{N_{\text{sch}}-1} x_{i,m} b_{i,m} \geq \epsilon \quad , \forall m \\
& b) b_{i,m} \leq \sum_{k=0}^{K} z_{i,m,k} B_k \quad , \forall i, m \\
& c) \sum_{k=0}^{K} z_{i,m,k} \Lambda_{i,m,k} \geq \big(\sum_{\forall j \neq i} \sum_{\forall m' \neq m} \overline{\text{MAI}}_{i,m}^{j,m'} x_{j,m'} \big) \quad , \forall i, m \\
& d) \sum_{k=0}^{K} z_{i,m,k} \leq 1 \quad , \forall i, m \\
& e) \sum_{m=0}^{M-1} x_{i,m} \leq 1 \quad , \forall i
\end{aligned}
\tag{4.18}
$$

OP in (4.18) is referred to as OP1. Essentially, OP1 is now continuous, but not linear due to the quadratic term in the first constraint (i.e., $x_{i,m} b_{i,m}$). This formulation belongs to the Mixed Integer Quadratically Constrained Problem (MIQCP) class. In general, solving MIQCP is challenging but, this class of problems, fortunately, features some useful structures that can be exploited and solved in more efficient ways [125]. In the literature, some studies like [126], [127] enhance the branch-and-cut algorithm to divide searching spaces and find the optimal solutions more efficiently. In practice, there are a few software optimizers on the market, such as IBM ILOG CPlex and Gurobi, that can solve this problem.

4.4.3 Sub-optimal optimization problems

In the previous section, we manage to transform the basic OP to an equivalent problem OP1, which can be solved more efficiently. However, solving such a problem is still a very challenging task. To alleviate that challenge, we consider other sub-optimal solutions, which can balance performance and complexity.

Variations of OP1

We propose three variations of OP1, which are less complex than the original problem.

- Problem OP11

  The first sub-optimal algorithm is based on a chunking scheme, where N_{ck} adjacent subchannels are grouped into a chunk. On the one hand, the bigger size of chunks, the smaller impact of frequency diversity, thus resulting in sub-optimal results. On the other hand, the searching space is greatly reduced from $(M+1)^{N_{\text{sch}}}$ to $(M+1)^{N_{\text{sch}}/N_{\text{ck}}}$.

Analytically, the formulation of OP11 is similar to (4.18), except that number of frequency subchannels becomes the number of chunks (i.e., $N_{\text{sch}}/N_{\text{ck}}$), and subchannel indices i and j are now indices for chunks.

- Problem OP12

  The second way is to relax some variables in OP1. Particularly, we introduce OP12, in which $z_{i,m,k}$ is now defined as a continuous variable that takes a value in the range of $[0, 1]$ (instead of either zero or one in (4.18)). Continuous $z_{i,m,k}$ can greatly relax the complexity of integer programming problems, although this clearly violates the principle of ACM, in which only one MCS is selected. Via OP12, the impact of pure mathematical relaxation (without a sufficient consideration from the engineering point of view) on the system performance is considered.

- Problem OP13

  Finally, we introduce a concurrent optimization model, denoted by OP13, where a timeout value is defined as the maximum solving time. Optimization software solvers stop the solving process when that period is over and return, most likely, sub-optimal results. In this way, we can target a certain level of complexity and work around extremely hard instances of OP1.

Minimization of maximal normalized MAI

In this section, we consider the possibility of improving the system performance by merely enforcing MAI mitigation without consideration of user throughput. This approach is inspired by the concept of zero-forcing equalizers [23], which is intensively studied to deploy multiple antennas (i.e., Multiple Input Multiple Output (MIMO) systems) in wireless communications. We develop a heuristic OP to minimize the mutual damage caused by MAI. We formulate a new problem in a min-max form, in which the objective function is to minimize the sum of MAI plagued on subchannels assigned to each user.

Analytically, the value of MAI caused by subchannel j of m' on subchannel i of m now can be reflected by the term of $x_{i,m}x_{j,m'}\overline{\text{MAI}}_{i,m}^{j,m'}$. Consequently, the total MAI over all subchannels that are assigned to user m yields the following form.

$$\sum_{\forall i}\sum_{\forall j\neq i}\sum_{\forall m'\neq m} x_{i,m}x_{j,m'}\overline{\text{MAI}}_{i,m}^{j,m'} \tag{4.19}$$

At this point, one can continue to complete the mathematical formulation, but it is important to note that: if received power is not considered jointly with MAI, then although MAI might be reduced, SINR is still small if received power is small. As a result, throughput gain is low.

Therefore, we propose to normalize MAI to the received power. By doing that, it is expected that the sum of MAI caused by other users on subchannel i is decreased, and at the same time, users tend to take subchannels whose channel gains are relatively good. This supposedly leads to an improvement in throughput. The goal now is

to minimize the maximal sum of MAI normalized to the received power as shown following:

$$
\begin{aligned}
& \min \quad \epsilon \\
s.t. \quad a) \quad & \sum_{i=0} \sum_{m' \neq m} \sum_{j \neq i} x_{i,m} x_{j,m'} \frac{\overline{\mathrm{MAI}}_{i,m}^{j,m'}}{P_{i,m}^{\mathrm{RX}}} < \epsilon \quad , \forall m \\
b) \quad & \sum_{m} x_{i,m} \leq 1 \quad , \forall i
\end{aligned}
\tag{4.20}
$$

Here, an auxiliary variables ϵ is used together with the first constraint in (4.20) to impose the min-max strategy, and the second constraint is to assure one subchannel can be assigned to one user or set as GB.

As the formulation does not directly consider user throughput, it is possible that no subchannels are assigned to users and all subchannels are set as GBs, then MAI reaches zero for all users. Therefore, we need a constraint assuring a minimum number of subchannels assigned to each user. Let Γ be the minimum number of subchannels that each user takes. Then we derive the complete formulation, referred to as OP2, in the following.

$$
\begin{aligned}
& \min \quad \epsilon \\
s.t. \quad a) \quad & \sum_{i} \sum_{m' \neq m} \sum_{j \neq i} x_{i,m} x_{j,m'} \frac{\overline{\mathrm{MAI}}_{i,m}^{j,m'}}{P_{i,m}^{\mathrm{RX}}} < \epsilon \quad , \forall m \\
b) \quad & \sum_{m} x_{i,m} \leq 1 \quad , \forall i \\
c) \quad & \sum_{i} x_{i,m} \geq \Gamma \quad , \forall u
\end{aligned}
\tag{4.21}
$$

As it can be seen from (4.21), OP2 is also a MIQCP problem. Thus this approach is supposed to be equally complex as OP1. The OP needs to be solved in an iterative process to derive the resulting resource allocation, each loop with an increasing input value of Γ.

Optimization problem with fixed-width GBs

In this thesis, for the sake of completeness, we describe in this section another sub-optimal approach proposed by Bohge et al. in [100]. In that work, an OP is formulated to maximize user throughput without consideration of MAI. It means instead of SINR, only SNR is adopted. By doing that, the mathematical challenge can be avoided; the problem can be formulated in a linear form, which can then be easily solved by linear optimization techniques (e.g., simplex method) [83]. Besides, a constraint is added to force the allocation of GB to mitigate MAI.

To formulate the problem, optimization binary variable $x_{i,m,d}$ has now 3 dimensions: subchannel i, user m and transmission mode d. d takes 1 for user data and 0 for GB. The formulation of OP3 is shown in (4.22).

$$
\begin{aligned}
& \max \quad \epsilon \\
s.t. \quad a) \quad & \sum_{i=0}^{N_{\text{sch}}-1} x_{i,m,1} F(\bar{\gamma}_{i,m}, P_{\text{err}}) \geq \epsilon, \quad \forall m \\
b) \quad & \sum_{m=0}^{M-1} \sum_{d=0}^{1} x_{i,m,d} = 1, \quad \forall i \\
c) \quad & \text{if } ((x_{i,m,1} == 1) \text{ AND } (x_{i-1,m,1} == 0)) \\
& \rightarrow x_{i-1,m,0} = 1, \quad \forall i, m
\end{aligned}
\tag{4.22}
$$

where $\bar{\gamma}_{i,m}$ [dB] denotes the SNR and has the following form

$$
\bar{\gamma}_{i,m} \text{ [dB]} = 10 \log_{10}(\frac{P_{i,m}^{\text{RX}}}{\sigma^2}) \tag{4.23}
$$

In (4.22) constraint c) forces one subchannel to be a GB between 2 adjacent resource blocks assigned to 2 different users. As it can be seen from (4.23), $\tilde{\gamma}_{i,m}$ is independent from optimization variables. As a result, OP3 is linear.

4.4.4 Evaluation

To numerically evaluate the performance of the proposed approaches, we simulate the system under consideration on the network simulation framework OMNeT++ using the MiXiM library. Parameters are selected based on the IEEE 802.16m standard [128]. Values of parameters are shown in Table 4.3.

First, we select parameter values for fading channels as recommended in [129] and [72]. Particularly, the root mean square of delay spread is set to equal $\sigma_\tau = 0.251\ \mu s$. Consequently, coherence bandwidth of channels is defined as the bandwidth over which the frequency correlation function is over 0.9. Then corresponding to [14], we have $B_c \approx \frac{1}{50\sigma_\tau}$, which equals 7.968 kHz. Thus each subchannel consists of 8 adjacent subcarriers.

The maximal time offset is chosen to equal 25% of the symbol duration (i.e., T_{sym}). And the clock accuracy of user oscillators is assumed to be 1 ppm (i.e., $\pm 1 \times 10^{-6}$ Hz). Given the center frequency f_c of 2.5 GHz, the maximal frequency offset is set to equal 2.5×10^3 Hz (which approximately equals 20% of frequency spacing).

We simulate in total 20 runs, and a total of 100 uplink frames per run. And for each uplink frame, users are dropped uniformly in the cell. Apart from our proposed OPs, we also simulate the solution introduced in [100] (referred to as OP3) as reference. During each uplink frame, at the same system status, instances of OP1, OP11, OP12, OP13, and OP3 are formulated and sent to the software optimizer, which is the Gurobi software solver. Since OP2 contains logical expressions, it could

Table 4.3: System parameters.

Parameters	Values
Number of users	4
Number of subcarriers	128
Subcarrier spacing	10940 Hz
Number of subcarriers per subchannel	8
Number of subchannels	16
Number of subchannels per chunk	2
Transmission power on subcarrier	1 mW
User maximal time offset	25 μs
User maximal frequency offset	2000 Hz
Cell radius	250 m
Path loss model	COST231 Walfish-Ikegami
Log normal shadowing std. dev.	10 dB
Multipath fading model	Jakes's/Clarke's model
Penetration and other losses	10 dB
Delay spread (rms)	0.251 μs
Receive Antenna Gain	14 dB
v_2	equal the maximal delay spread
v_1	0 μs

not be solved by Gurobi, the ILOG CPlex software optimizer is used to solve OP2 instances. ILOG CPLex is also used in the original work in [100]. Resource allocation schemes achieved at the output of the optimization solver are used to calculate user throughput.

To compare the performance achieved by the proposed OPs with the one of ECBA, a simple system, based on the IEEE 802.16m standard [128], with long CPs together with pilot subcarriers to deal with MAI is simulated. In this system, first, CP's length is assumed to be sufficient to deal with delay spread, propagation delay, and clock errors. ISI is then fully mitigated. Pilot subcarriers are inserted following the PUSC method in the standard IEEE 802.16m [128]. It means 33% of overall resources are pilots. Finally, the blocking assignment strategy is adopted to statically assign subchannels to users. We evaluate two cases, referred to as ECBA1 and ECBA2. First, the estimation and correction are assumed to be perfect in ECBA1. Thus no time and frequency offsets exist. Consequently, we consider a more realistic scenario, in which the ECBA implementation is not perfect, leading to a residual frequency offset. In ECBA2, we assume the residual frequency offset is 10 Hz. Note that the Doppler shift for the velocity of 30 km/h is about 70 Hz. It is important to mention that results for the simple ECBA system and the proposed OPs are derived from the same system status (e.g., wireless states and synchronization offsets).

Figure 4.16 shows the minimum user throughput achieved by different algorithms as well as by the ECBA approach. The two most left columns illustrate the comparison of OP1 and ECBA1 in the idealized scenario. As it can be seen, OP1 provides

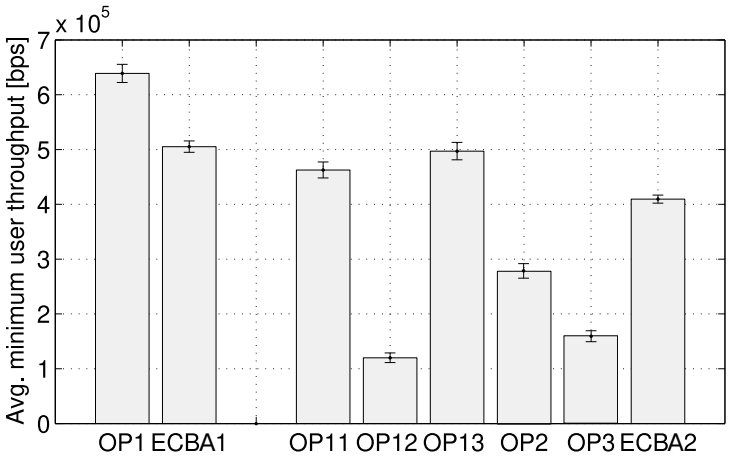

Figure 4.16: Avg. of minimum user throughput

a significant gain (about 26%) compare to ECBA1. On the right, with the existence of residual frequency offsets, applying DRA offers an improvement of minimum user throughput by 25% compared to ECBA2. Furthermore, among OPs, OP1 provides a significant improvement compared to OP2 and OP3, equivalent to about 129% and 300%, respectively. Therefore by considering channel quality and MAI, the system performance can be greatly improved subject to the desired goal. The chunking approach OP11 and the concurrent OP13 also provide improvement (190% and 115% compared to OP2 and OP3, respectively), but the mathematical relaxation of in OP12 leads to poorer performances.

Concerning the solving time, Figure 4.17 shows the quartiles of solving time for different OPs. Since distributions of solving time have long tails, mean values with confidence intervals are replaced by quartiles. As it shows, although OP11 provides sub-optimal resource allocation, it could be solved much faster than OP1. The median value of OP11 is less than 0.1 s equal 10% of the one of OP1.

To have an insight, we quantize the fragmentation of resource allocation by investigating the number of GBs and the number of Heterogeneous Junctions (HJ), which is defined as the border between 2 user data blocks assigned to 2 different users. First, Figure 4.18 shows the average number of GBs inserted when different OPs are used. Furthermore, the histogram of the number of GBs of OP1 is shown in Figure 4.19. As it can be seen from Figure 4.18, 4.19 and 4.20, OP1 improves the minimum user throughput as expected by exploiting better the frequency and multi-user diversities.

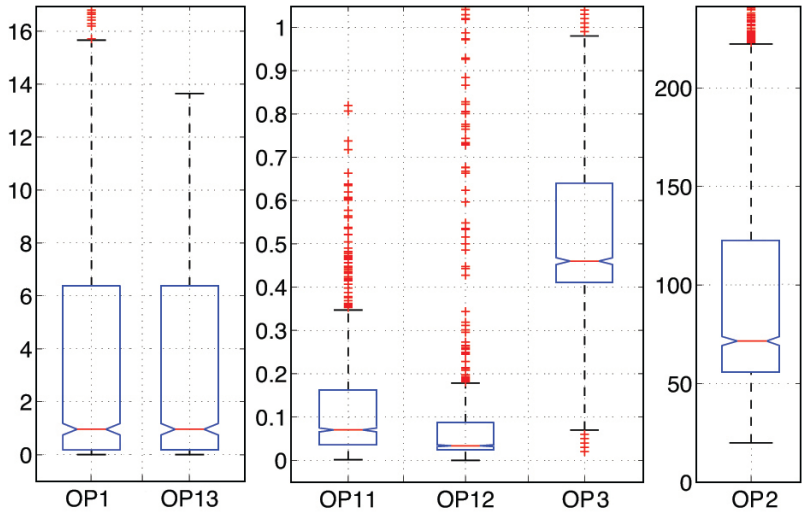

Figure 4.17: Quartiles of solving time in second

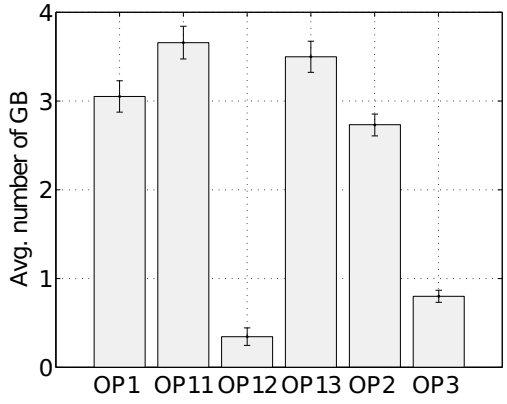

Figure 4.18: Avg. number of GBs

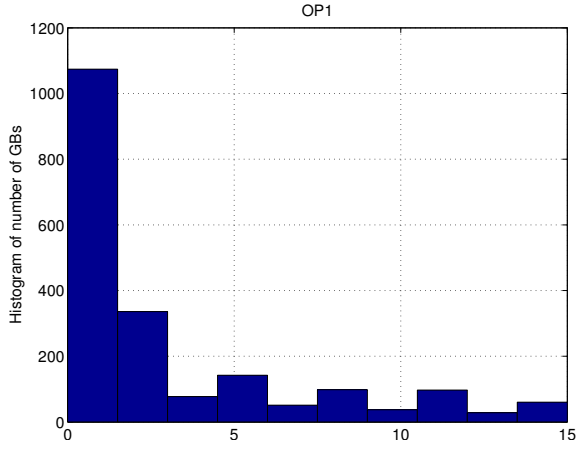

Figure 4.19: Histogram of number of GBs of OP1

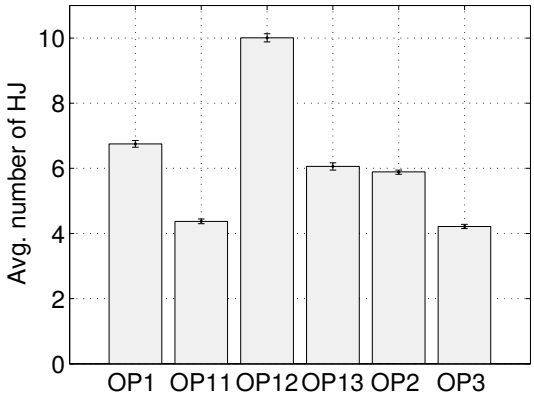

Figure 4.20: Avg. number of HJs

4.5 Conclusion

The received signal in the uplink of OFDMA systems is highly sensitive to the imperfect synchronization, which stems from the fact that received signal is the aggregate of several elements sent by multiple users. Essentially, the time and frequency offsets damage the orthogonality among subcarriers, and cause MAI. Managing the synchronization problem and dealing with MAI is one of the primary challenge to deploy OFDMA in the uplink.

In this chapter, the possibility to deal with MAI via efficient resource allocation schemes is investigated. First, the impact of static resource allocation on MAI in a simple and static scenario without fading channel is studied. It is shown that, instead of using long CP and employing estimation techniques as most of previous researches in the literature, one can use time and frequency guards together to mitigate the effect of MAI and avoid large overhead of long CP. When no frequency offset is present, it is better to use long CP rather than GBs, however one can instead choose short CP and long GB to achieve the same performance. In case of frequency offset in the system, using GBs can also improve the overall cell throughput. In presence of both frequency and time offset, using GBs with short CP achieves the best performance. Using cell rate as the criteria for throughput performance neglects the fairness between users and another formulation is needed to deal with this case.

Second, the investigation is then extended to cope with more realistic scenarios. In this context, the resource allocation scheme needs to adapt to the channel conditions and user offset profiles with the goal to fairly improve user throughput. We formulated a max-min OP and provided simplifications of the OP to reduce the computational complexity and, thus, the solving time. Numerical results demonstrate that the system throughput, defined as maximum of minimum user throughput, can be significantly improved by proper resource allocation.

Chapter 5

Cross-Layer Algorithm for Non-Layered Video Streaming

The video industry has broadly adopted NLVS for VoD services. It is, however, challenging to develop efficient adaptive algorithms using NLVC for low-delay services. Demanding latency requirements are the main reason. The challenge is even more noteworthy when considering mobile networks, where network condition varies enormously and swiftly.

This chapter presents our novel algorithms that aim to improve the QoE of multiple low-delay video streams over OFDMA networks. Those cross-layer algorithms jointly consider video adaptation and DRA. We propose different algorithms targeting scenarios in the uplink as well as in the downlink. In addition, we consider two re-buffering policies, which are LSH and LSS. In those algorithms, the video qualities of segments are first determined. The selection balance several factors. On the one hand, the selected segment qualities aim to improve the long-term QoE assessed over multiple segments. Our video adaptation algorithms also enforce QoE fairness among users. On the other hand, the total bitrates demand of selected qualities should not exceed the system's available bandwidth. Given the selection of segments' qualities, a series of short-term DRA problems, each of which occurs in the millisecond range, adapt resource allocation schemes to instantaneous channel states. The goal is to proactively match effective throughput with bitrate demand. In that way, the system strives to deliver requested video segments before their deadlines. Simulation results demonstrate the benefit of the proposed algorithms. The main result of this chapter has been published in [130].

5.1 Non-Layered Video Streaming over OFDMA Networks

We consider a single OFDMA cell, in which M users stream low-delay video contents in the uplink simultaneously.

5.1.1 Streaming model

In this chapter, NLVS is adopted. It means each video content is available in several representations. Each representation targets a perceived quality and has a specific Mean Media Bit Rate (MMBR). Essentially, the MMBR of a segment is computed by dividing the segment's payload size (in bits) by its length (in seconds). In this thesis, MMBR and segment bitrate have the same meaning. Roughly speaking, the better the quality, the higher the video bitrate. The available representations are split into equally long video segments, or, shortly, segments, such that switching between qualities is feasible on segment boundaries. It means the pace of video adaptation increases when the segment length decreases, and vice versa.

Let L_m denote the number of available video representations for user m, indexed by l. And let $Q_{l,s,m}$ and $R_{l,s,m}$ be the video quality and its corresponding bitrate of segment s of representation l in stream m, respectively. We denote the selected qualities and bitrates by $Q_m(s)$ and $R_m(s)$, respectively. Notably, the perceived quality of segments is measured using the PSNR metric. PSNR gives a moderately accurate measurement of user satisfaction [131]. Nevertheless, PSNR is still widely adopted in the literature due to its conceptual and computational simplicity [132].

Recall that QoE involves human complex visual and neural systems, and is very subjective. It is well known that QoE depends on several factors, e.g. initial delay, quality fluctuation, and video stalls. It is challenging to address all those aspects in a single analytical model. In this work, we address key factors in different aspects of our algorithms.

We adopt the quality model, called QoE Index (QID), in [110] to consider two important factors affecting QoE. Those factors are the average quality and the quality fluctuation over several segments. (For that reason, segments' qualities $Q_{l,s,m}$ are short-term measurements, and QID is a long-term evaluation.) Particularly, we compute QID over first s segments (i.e. $0 \leq s' \leq s$) via function $U_m(s)$ with the following equation:

$$U_m(s) = \text{mean}(\mathcal{A}_m(s)) - \text{std}(\mathcal{A}_m(s)) \tag{5.1}$$

where:

$$\mathcal{A}_m(s) = \{Q_m(s'), 0 \leq s' \leq s\} \tag{5.2}$$

Here $\mathcal{A}_m(s)$ is the set of selected qualities for segments $s' \in [0, s]$; mean() and std() are two functions returning average and standard deviation.

In principle, different streams can have different segment lengths. (Note that all segments in a stream have the same duration.) Selecting an appropriate segment length is, however, not a trivial task. Selected values should harmonize several factors as discussed in [133]. On the one hand, using short segments increases the pace of video adaptation, thus quickening the system's response to, among other things, channel variations. Choosing long segments, n the other hand, results in better encoding efficiency. Besides, segment lengths also affect segments' data sizes and downloading time. Typical values are in the range of 1 to 10 seconds.

In the context of low-delay streaming in mobile networks, short lengths are preferred. That is because quick reactions to channel variations and small downloading durations are critical to meet stringent latency requirements. For the sake of simplicity, it is assumed that all streams have the same segment length, denoted by T_{seg}. This assumption was also adopted in several studies in the literature (e.g., [111], [134]).

Conventionally, we adopt a slotted time axis corresponding to video segments as illustrated in Figure 5.1.

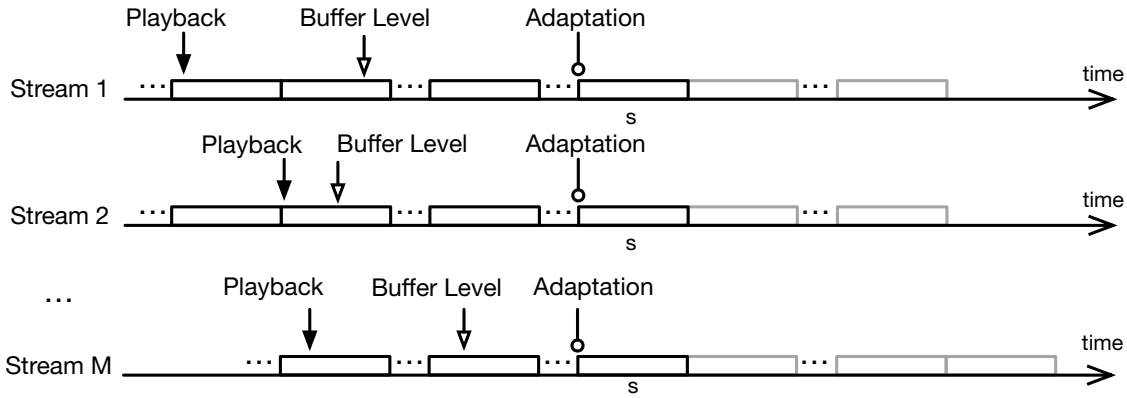

Figure 5.1: Illustration of three concurrent processes: playing back, downloading and adapting

Before starting the playback of all streams, each video player fetches some segments to build an initial buffer up to a targeted pre-buffer level. The goal of pre-buffering is to mitigate the negative impact of throughput fluctuations during the playback afterward.

After the pre-buffering phase, three processes take place simultaneously. They are video playback, video downloading and video adaptation. Particularly, at the beginning of each slot, one new segment of each stream is available for downloading. Also at this point, a proper quality is selected for each newly available segment. In the meantime, video players play buffered segments, while segments are downloaded and queued in the playback buffer.

Regarding video adaptation, the quality selection takes into account long-term QID measurements, a notion of QoE fairness among users, and projected user throughput. Information about available representations and their characteristics is assumably available for video adaptation algorithms.

Let $B_m(s)$ denote the buffer level, measured in playback time, of stream m at the beginning of time slot s. To simplify the notation, it is assumed that after the pre-buffering phase (i.e., $s = 0$), each client buffers the same number of segments N_{pre}, i.e., $B_m(0) = N_{\text{pre}} T_{\text{seg}}$.

Furthermore, each stream has the same number of segments, denoted by N_{seg}.

In this work, a theoretical transport protocol is assumed for lossless transmission. Also, the round trip delay of the communication between clients and servers is not considered.

Table 5.1: Notation overview

$Q_m(s)$	Selected video quality for user m in time slot s
$R_m(s)$	Selected video bitrate for user m in time slot s
$B_m(s)$	Buffer level at the start of time slot s
$Q_{l,s,m}$	Quality of representation l for segment s in stream m
$R_{l,s,m}$	Video bitrates of representation l associated with $Q_{l,s,m}$
$\hat{C}_m(s)$	Estimated link rate of stream m in time slot s
$\hat{G}_m(s)$	Targeted resource share for stream m in time slot s
$\hat{A}_m(s)$	Expected throughput for m in time slot s
$C_m(s)$	Empirical link rate of user m in time slot s
$G_m(s)$	Actual resource share for user m in time slot s
$A_m(s)$	Actual throughput for user m in time slot s

Finally, Table 4.1 summarizes the main notation used in this chapter.

5.1.2 OFDMA model

The same OFDMA model shown in 4.1 is inherited in this chapter. The rest of this section focuses on mapping low-level OFDMA resources to high-level video adaptation.

It is assumed that each video time slot s contains N_{fis} OFDMA frames, and we use t to index OFDMA frames in slot s, $0 \leq t \leq N_{\text{fis}}$. In the frequency domain, M users share in total N_{sch} subchannels for streaming video.

During each OFDMA frame, let $\bar{\gamma}_{i,m}(t)$ be the SINR of subchannel i of user m during frame t. The calculation of $\gamma_{i,m}(t)$ yields the formulation shown in (4.6). Similarly, let $x_{i,m}(t)$ indicates whether resource unit (i,t) is allocated to user m or not. Consequently, we denote the number of bits user m can send during frame t by $a_m(t)$, which can be computed as follows:

$$a_m(t) = \sum_{i=0}^{N_{\text{sch}}-1} x_{i,m}(t) F(\bar{\gamma}_{i,m}(t)) \tag{5.3}$$

For convenience, the resource share assigned to user m in video time slot s is given by $G_m(s)$, which yields

$$G_m(s) = \frac{1}{N_{\text{sch}} N_{\text{fis}}} \sum_{i=0}^{N_{\text{sch}}-1} \sum_{t=0}^{N_{\text{fis}}-1} x_{i,m}(t) \tag{5.4}$$

And the total throughput achieved by user m during time slot s is given by $A_m(s)$. We have $A_m(s) = 1/T_{\text{seg}} \sum_t a_m(t)$.

5.2 Joint Resource Allocation and Video Adaptation Scheme

In this section, we first describe the general idea of our cross-layer algorithms. Next, we formulate adaptive algorithms of video adaptation and DRA. Finally, the last sub-section introduces system architectures that integrate proposed solutions into existing mobile networks.

5.2.1 A novel cross-layer approach for low-delay streaming

Generally, one key challenge as considering video adaptation jointly with resource allocation lies in the difference between their timescales [135]. In particular, on the one hand, video adaptation takes place at intervals of a few seconds (at segments' boundaries). Resource allocation, on the other hand, typically occurs every a few milliseconds (at the level of OFDMA frames) so that adaptive algorithms can exploit channel diversities and achieve throughput gains.

In the context of NLVS, we address the challenge of different timescales by exploiting a feature of streaming systems. Particularly, a client must select a suitable quality for each segment before downloading it. Once the quality is selected, the client strives to download it before its playback deadline. Changing the quality of a video segment during its download causes overheads, wastage of resources, and additional delays. Due to that, a natural approach to deal with the timescale difference is to consider a two-phases approach. In the first phase, video adaptation algorithms derive appropriate quality selections. Given desired quality selections, radio networks need to deliver selected data on time to avoid video stalls in the second phase. A diagram of the overall system is depicted in Figure 5.2.

Quality selection

The problem of selecting appropriate video representations is challenging. On the one hand, too optimistic selections can overload radio networks and cause video stalls. Too conservative decisions, on the other hand, can avoid video stalls but underutilize radio resources and result in sub-optimal QoE. In the context of low-delay services, playback buffer is typically small due to demanding latency requirements. As a result, QoE is exposed strongly to channel fluctuations, making video adaptation is a great challenge.

To address that challenge, we develop efficient video adaptation algorithms that balance several factors (including buffer levels and video fluctuations) and provide appropriate quality selections. To do that, mathematical OPs are developed. The goal is to improve QoE by increasing QID. In that way, we aim to increase average video quality and, at the same time, reduce quality fluctuation. Concretely, we maximize

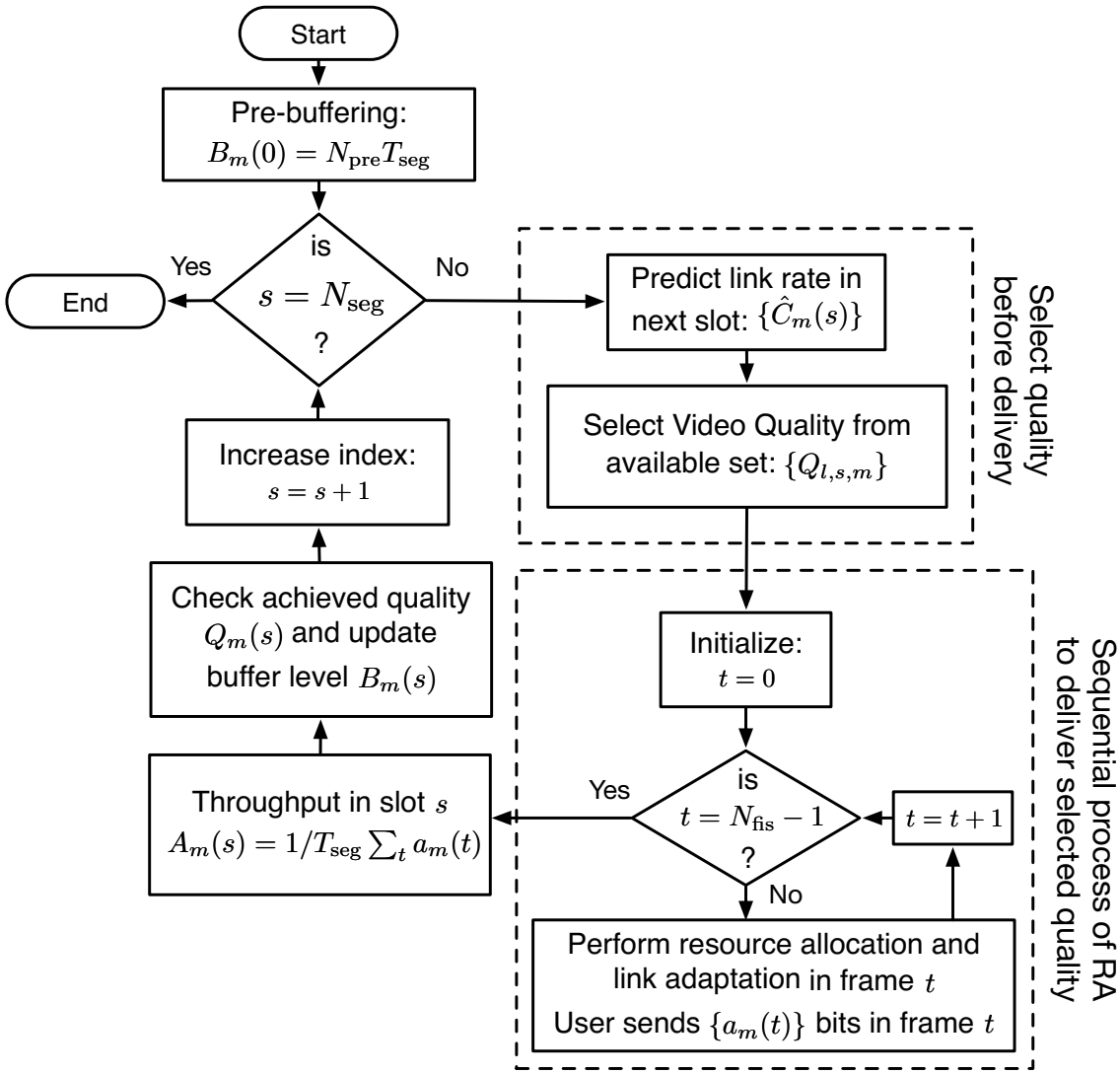

Figure 5.2: Illustration of the proposed cross-layer adaptive streaming approach

the minimum QID among users. The max-min formulation is adopted to enforce fairness among users concerning QID. Finally, our video adaptation algorithms strive to avoid playback stalls by considering buffer levels as well as forecasted link rates.

In particular, at the beginning of each time slot, we select appropriate qualities of the next segments for all users. The quality selection is derived by solving an instance of the max-min QID problem. Selected qualities are constrained from above by estimated link rates. Let $\{\hat{C}_m(s), 0 \leq m < M\}$ be the set of estimated link rates of all users.

Especially, we propose a novel approach to estimate future link rates. Unlike other studies, our approach considers, not only, channel statistics, but also, potential throughput gains via DRA. Consequently, higher segment qualities can be selected. In this section, we assume that link rate estimations are given. We will study the problem of estimating link rates in Section 5.3.

Video delivery

Given the set of selected qualities, denoted by $\{Q_m(s), 0 \leq m < M\}$, radio resources are assigned to users so that selected representations are transmitted to viewers. Particularly, during each OFDMA frame, which is a few milliseconds long, each user takes some OFDMA subchannels and sends some pieces of video segment. For each stream, the accumulated amount of transmitted data is increased gradually over a large number of OFDMA frames. Eventually, the accumulated amount should match segments' payload sizes before their deadline. Otherwise, video stalling is encountered.

In this context, we adopt DRA algorithms presented in Chapter 4. Via that adoption, we can address more realistic assumptions, where residual synchronization offsets exist in the uplink. Besides, those DRA algorithms can achieve valuable throughput gains by exploiting channel diversities. Those valuable gains are then used to assure the video delivery on time. The objective of DRA algorithms is to increase users' intermediate transmitted data amounts towards expected video bitrates. Particularly, we formulate max-min OPs that maximize the minimum accumulated amount. Notably, weights are added to assure users with lower buffer levels receive more resources.

By the end of each time slot, achieved link rate in that slot is computed and used to estimate the link rate of the next slot.

Video stalling

When a segments cannot be delivered completely by its deadline, video stalling happens. It means $B_m(s) < T_{\text{seg}}$. We consider two realistic policies dealing with such events. They are LSH and LSS.

In the case of LSH, when the playback deadline of a segment reaches before its download finishes, that segment is dropped. The video player strives then to fetch the next segment. This policy might be preferred for video applications, where viewers prioritize low delays over missing frames. Examples of such applications are video conferencing and live sport streaming. Missing segments clearly causes QoE degradation. For that reason, we consider the number of missing frames together with QID to achieve a better QoE evaluation of applications using LSH.

LSS might be preferred when viewers want to play out the content without content gaps. Consequently, whenever a segment cannot be delivered before its deadline, the playback is halted. Video player strive to raise its buffer level above a reasonable threshold before resuming the playback. For this case, in addition to QID, we evaluate the number of video stalling events and the total re-buffering times to assess QoE.

5.2.2 Video Adaptation

In this section, we consider video adaptation algorithms for low-delay streaming adopting LSH and LSS in detail.

Low delay streaming with hard latency constraints

First, we formulate an adaptive algorithm for the LSH case. In this context, we target a particularly low delay of two times the segment duration (neglecting the processing time at the server, the client, and the round trip time). At the beginning of a streaming session, a video player loads one segment into its playback buffer and immediately starts the playback. That is, $B_m(0) = T_{\text{seg}}$. During each slot, a complete segment must be delivered to viewers. If a segment cannot be delivered by its deadline, which is at the end of time slot, it is discarded. Such an extreme latency requirement demonstrates the direct impact of the performance of adaptive algorithms on QoE. Figure 5.3 illustrates our proposed approach for LSH case.

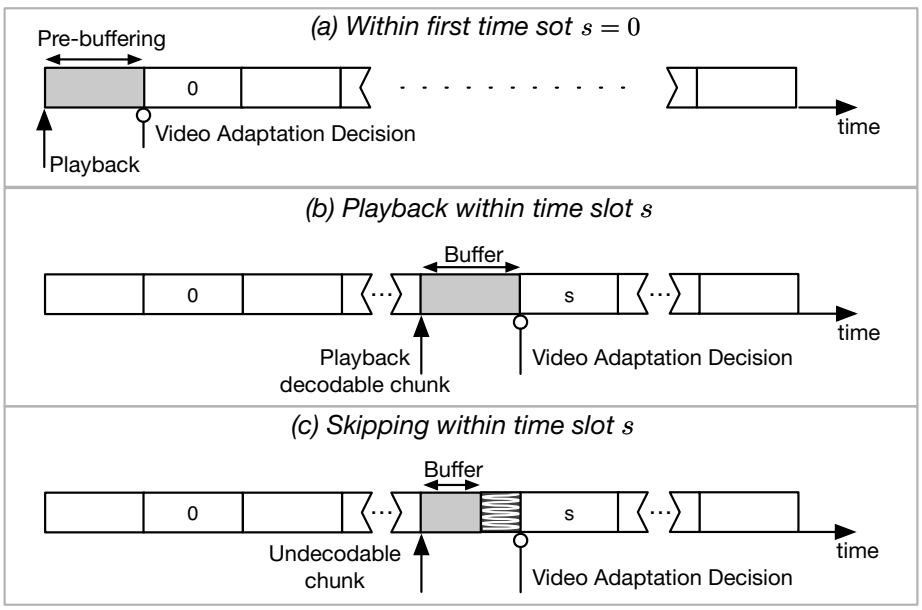

Figure 5.3: Illustration of LSH strategy

We introduce binary optimization variables $z_{l,s,m}$ reflecting quality selections. Those variables take 1 if user m selects representation l ($0 \leq l < L_m$) in slot s, and 0 otherwise. Then the max-min OP for quality selections in slot s yields the form in (OP-LSH).

Here, constraints (C1) and (C2) express the quality and the MMBR of segment s in the selected representation. Constraint (C3) ensures that each user m selects exactly one representation for segment s. Finally, constraints (C4.H) and (C5) represent upper limits on link rate and resource share. More precisely, constraint (C4.H) ensures that the MMBR of segment s in the selected representation does not exceed the throughput expected in time slot s, given by the link rate estimate $\hat{C}_m(s)$ multiplies by the user's link share $\hat{G}_m(s)$. Constraint (C5) ensures that the total allocated resource blocks does not exceed the total available limit.

$$\max_{m} \min \quad U_m(s) \tag{OP-LSH}$$

$$\text{s.t.} \quad Q_m(s) = \sum_{l=0}^{L_m-1} z_{l,s,m} Q_{l,s,m}\,, \qquad \forall m \tag{C1}$$

$$R_m(s) = \sum_{l=0}^{L_m-1} z_{l,s,m} R_{l,s,m}\,, \qquad \forall m \tag{C2}$$

$$\sum_{l=0}^{L_m-1} z_{l,s,m} \leq 1\,, \qquad \forall m \tag{C3}$$

$$R_m(s) \leq \hat{G}_m(s)\hat{C}_m(s)\,, \qquad \forall m \tag{C4.H}$$

$$\sum_{m=0}^{M-1} \hat{G}_m(s) \leq 1 \tag{C5}$$

It is important to note that, due to the discrete set of available qualities and MMBRs of available representations, the problem of video adaptation, in general, is non-linear, non-convex, and, hence, NP-hard [136]). Thus there are no efficient approaches available to derive the global optimum of such problems. The formulation in (OP-LSH) alleviates the challenge by exploiting the piece-wise linearization method to provide a linear programming problem for obtaining an approximated global optimal solution.

Low delay streaming with soft latency constraints

Figure 5.4 illustrates the proposed approach for LSS case. As it shows, more than one segments are pre-buffered. Accordingly, the playback deadline of segments can be in a few slots.

In this case, a playback buffer can be exploited to absorb short-term throughput degradation. Concretely, we allow the selected MMBR to exceed the estimated throughput when the playback buffer level is high enough so that the segment can still be downloaded before its playback deadline. Consequently, constraint (C4.H) in (OP-LSH) transforms to

$$R_m(s) \leq \hat{G}_m(s)\hat{C}_m(s)\Big[1 + \lambda \frac{B_m(s-1)}{T_{\text{seg}}}\Big]\,, \quad \forall m \tag{5.5}$$

where a real-valued coefficient $\lambda \in (0, 1)$ represents a safety margin to reduce the risk of buffer underrun. The complete formulation then takes the formulation in (OP-LSS).

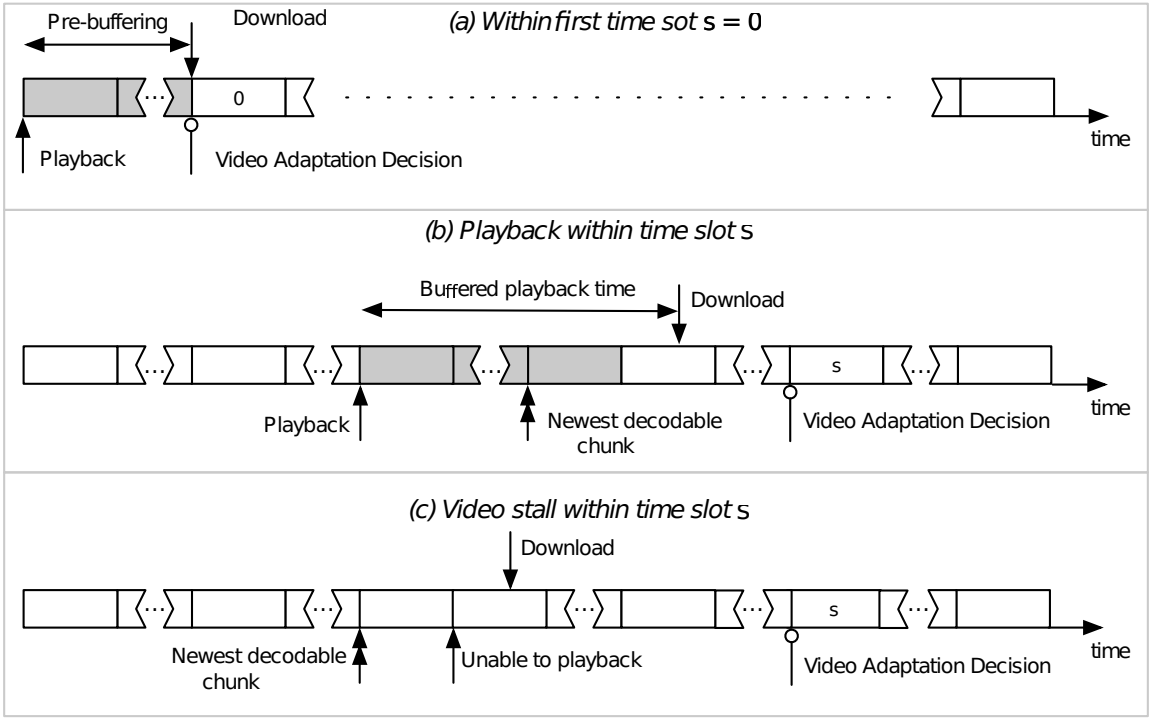

Figure 5.4: Illustration of LSS strategy

$$\max_{m} \min \quad U_m(s) \tag{OP-LSS}$$

$$\text{s.t.} \quad Q_m(s) = \sum_{l=0}^{L_m-1} z_{l,s,m} Q_{l,s,m} \,, \qquad \forall m \tag{C1}$$

$$R_m(s) = \sum_{l=0}^{L_m-1} z_{l,s,m} R_{l,s,m} \,, \qquad \forall m \tag{C2}$$

$$\sum_{l=0}^{L_m-1} z_{l,s,m} \leq 1 \,, \qquad \forall m \tag{C3}$$

$$R_m(s) \leq \hat{G}_m(s)\hat{C}_m(s)\Big[1 + \lambda \frac{B_m(s-1)}{T_{\text{seg}}}\Big] \,, \qquad \forall m \tag{C4.S}$$

$$\sum_{m=0}^{M-1} \hat{G}_m(s) \leq 1 \tag{C5}$$

5.2.3 Dynamic Resource Allocation

In this section, we exploit DRA to proactively match users' long-term user throughput with selected video bitrates. Particularly, we consider a series of resource allocation decisions, each of which takes place in one OFDMA frame. We formu-

late an OP, which takes into account instantaneous SINR, intermediate achieved throughput (accumulated over all frames in the time slot under consideration), and the selected representation. At the output, optimal resource allocation schemes and associated MCSs are given. Primarily, we separately consider the problems in the downlink and in the uplink. The ultimate difference between them relies in the mitigation of MAI in the uplink. We adopt the max-min formulation to balance efficiency and fairness.

DRA for the downlink

We consider the problem of DRA for OFDMA frame t ($0 \leq t \leq N_{\text{fis}} - 1$) in video time slot s. Let $\tilde{a}_m(t)$ be the sum of video data (in bits) that user m has sent over all previous OFDMA frames in time slot s. We have:

$$\tilde{a}_m(t) = \sum_{t'=0}^{t-1} a_m(t') \tag{5.6}$$

Then the amount of data delivered in frame t (given in (5.3)) is a direct result of the DRA decision in that frame.

To pursue fairness concerning selected video bitrates, the achieved throughput sum is further normalized by the amount of video data buffered at the transmitter and not yet delivered to the receiver. The normalization weight is given by $W_m(s) = \sum_{u \in \psi_m(s)} R_m(u) T_{\text{seg}}$, where $\psi_m(s)$ is the set of the segments that have qualities selected.

Essentially, $W_m(s)$ ensures more resources are assigned to users with higher chances of video stalling (i.e. higher values of $W_m(s)$). The intermediate amount data that has been delivered $\tilde{a}_m(t)$ forces long-term user throughput to match selected video bitrate. The final OP at time t is given as:

$$\max_m \min \quad \frac{1}{W_m(s)} \Big[\tilde{a}_m(t) + \sum_{i=0}^{N_{\text{sch}}-1} x_{i,m}(t) b_{i,m} \Big] \tag{OP-DL}$$

$$\text{s.t.} \quad \sum_{m=0}^{M-1} x_{i,m}(t) \leq 1 \,, \quad \forall i \,, \tag{C6}$$

Note that $b_{i,m}$ is the coefficient derived from function $F(.)$ representing the ACM feature. Since MAI does not occur in the downlink, $F(.)$ does not include any optimization variable. It basically takes instantaneous values of channel gains and computes a coefficient reflecting the number of bits user m can send on subchannel i in frame t.

DRA for the uplink

We extend the OP1 in Section 4.4 to efficiently assign resources to users in the uplink. The difference is the existence of two new coefficients, which are $W_m(s)$ and

$\tilde{a}_m(t)$. Similar to the OP for the downlink, they are added to enforce long-term goals. The problem of DRA in the uplink frame t can be written as follows:

$$\max_m \min \quad \frac{1}{W_m(s)}\Big[\tilde{a}_m(t) + \sum_{i=0}^{N_{\text{sch}}-1} x_{i,m}(t) b_{i,m}\Big] \tag{OP-UL}$$

$$\text{s.t.} \quad \sum_{m=0}^{M-1} x_{i,m}(t) \leq 1\,, \qquad \forall i \tag{C6}$$

$$\sum_{k=0}^{K} z_{i,m,k}\Lambda_{i,m,k} \geq \sum_{j\neq i}\sum_{m'\neq m} \frac{\overline{\text{MAI}}_{i,m}^{j,m'}}{\sigma^2} x_{j,m'}(t)\,, \qquad \forall i, m \tag{C7}$$

$$\sum_{k=0}^{K} z_{i,m,k} \leq 1\,, \qquad \forall i, m \tag{C8}$$

$$b_{i,m} \leq \sum_{k=0}^{K} z_{i,m,k} B_k\,, \qquad \forall i, m \tag{C9}$$

Similar to problem (OP-LSH), the problem (UL) is a linear OP thanks to the piece wise linear method. Recall that constraints (C7)-(C9) essentially construct a linear formulation of the discreet ACM process. Here B_k is the number of bits user can be send if MCS scheme k is selected. $\Lambda_{i,m,k}$ is the SINR lower limit required to use MCS scheme k (subject to a tolerable BER). Finally, binary optimization variables $z_{i,m,k}$ present MCS selections.

5.2.4 Proposed system architecture

In this section, we propose two system architectures that can enable our proposed cross-layer algorithms in the downlink and in the uplink of OFDMA networks. We envision a center controller, which can be allocated at BS, collect inputs, make the decisions about video adaptation and resource allocation for multiple low-delay streams in one cell. One critical requirement is that important information like buffer levels and content characteristics is available for the cross-layer optimization.

Streaming in Downlink

The proposed architecture for the downlink is illustrated in Figure 5.5. To derive appropriate video quality selections, the controller requires some inputs. First one is about CSI, which measured directly at BS or fed back from users. The controller also collects information about buffer levels and users' requests. Based on those inputs, instances of proposed OPs for video adaptation can be formulated and solved by optimization software solvers (e.g. Gurobi). Results are optimized video quality selections. Consequently, selected qualities are then used to override the selections in users' requests and sent to remote servers on behalf of users. The selected bitrates are also sent to the scheduler that assigns radio resources to users. The task of the

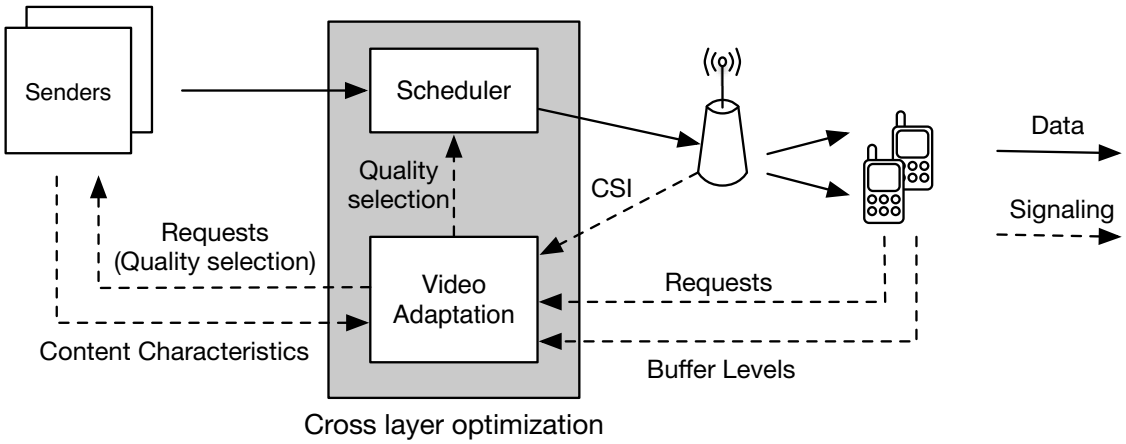

Figure 5.5: System architecture for the adaptive streaming in the downlink

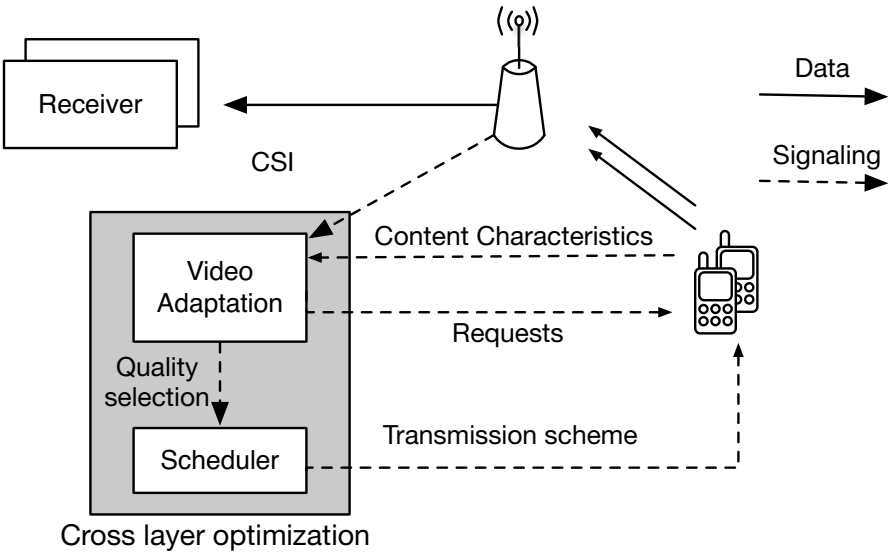

Figure 5.6: System architecture for adaptive streaming in the uplink

scheduler is then to adapt resource allocation and MCS to channel states to gradually match users' long-term throughput with selected bitrates.

Streaming in Uplink

The proposed architecture for the uplink is shown in 5.6. As it shows, the controller in this case sends requests (dictating quality selections) to users on behalf of remote receivers. Unlike in the downlink, the controller have to send resource allocation schemes to users so that they can send video data in the uplink.

A realization of this architecture, which is still compatible with the DASH streaming paradigm, may use the ServerPush feature of HTML5 [38] to actively push the video segments selecting for each of them an appropriate representation. Buffer levels may be reported using the MPEG-DASH reporting functionality.

5.3 Link Rate Estimation

Accurate link rate estimation is crucial to choose appropriate video qualities. However, deriving good estimates is very challenging due to, among others, wireless channels' variations. While under-estimated link rates result in lower video qualities, overestimations cause recurrent video re-buffering or skipping video segments. One way to significantly enhance prediction quality is to exploit information about wireless links and to estimate link rates based on the ergodic capacity of fading channels as used in, for instance, [111], [134], [137]. However, adopting the ergodic capacity implicitly means that resources are not adapted to particular channel states, and therefore achieved throughput is averaged out over all possible channel realizations. This assumption does not hold for DRA. In this work, we show that an ergodic capacity expression can be modified slightly to provide predictions in this case.

5.3.1 Throughput Estimation using Ergodic Capacity

In OFDMA systems, it is reasonable to assume that fading processes on radio resource blocks are independent. Recall that, one resource block is made of one OFDMA frame in time and one subchannel in frequency. As it shows in Chapter 2, that assumption can be easily realized via selecting a proper frame duration and a subchannel width. And when transmission is carried out over a sufficiently large number of resource blocks, the average throughput can be approximated well by the ergodic capacity [23].

Suppose user m takes all available radio resources, which are N_{sch} subchannels in frequency and N_{fis} OFDMA frames within slot s. SNR can be considered as a random variable, independently realized over resource blocks. Let $\Gamma_m(s)$ denote the random variable of SNR. Function $\mathbb{E}()$ returns the mean value computed over $\Gamma_m(s)$ in time slot s. The ergodic capacity of user m in time slot s, denoted by $C_m^{\text{erg}}(s)$, yields:

$$\frac{1}{N_{\text{fis}}N_{\text{sch}}}\left\{\sum_{t=0}^{N_{\text{fis}}-1}\sum_{i=0}^{N_{\text{sch}}-1}\log_2(1+\bar{\gamma}_{i,m}(t)\right\} \longrightarrow \mathbb{E}\Big[\log_2(1+\Gamma_m(s))\Big] = C_m^{\text{erg}}(s) \tag{5.7}$$

One can use different methods to estimate the channel's ergodic capacity. In this work, we consider there estimation methods. The first one is Statistical Generation (SG) method. This method assumes PDF of fading processes is available by analyzing channel gains in the previous time slot [138]. Consequently, one can generate a sufficient number of channel coefficients following the given PDF. Ergodic capacity can be computed offline by averaging over those channel coefficients. The output of this method is denoted by $C_m^{\text{SG}}(s)$.

The second method, called Low Bound Prediction (LBP) method, is adopted in [134]. This method provides a lower bound on ergodic capacity. If the average SNR in the last time slot is given by $\bar{\Gamma}_m(s-1)$, the LBP method approximates the ergodic capacity by:

$$C_m^{\text{LBP}}(s) = e^{1/\bar{\Gamma}_m(s-1)} E_i\left(1, \frac{1}{\bar{\Gamma}_m(s-1)}\right), \tag{5.8}$$

where $E_i(1,x)$ is the exponential integral function defined as $E_i(1,x) = \int_x^\infty \frac{e^{-\tau}}{t} d\tau$.

Tight Bound Prediction (TBP), introduced in [139], aims to provide a closer approximate of the ergodic capacity. The output of this method is given by:

$$C_m^{\text{TBP}}(s) = \log_2(1 + \bar{\Gamma}_m(s-1)(e^{-\rho})). \tag{5.9}$$

where ρ is the Euler's constant, i.e., $\rho \approx 0.57721566$.

Normally, it is expected that the number of OFDMA resource blocks within a time slot is large. For instance, in a typical OFDMA system, there are roughly 10^5 blocks in a 500 ms time slot with a total bandwidth of 18 MHz [72]. As a result, the convergence to ergodic capacity occurs even when users do not take all resource blocks. Suppose user m takes a resource share of $\hat{G}_m(s)$, the expected throughput $\hat{A}_m(s)$ can be approximated as:

$$\hat{A}_m(s) = \hat{G}_m(s)\hat{C}_m(s) \tag{5.10}$$

5.3.2 Estimation with Dynamic Resource Allocation

Prediction methods based on the channel's ergodic capacity implicitly assume Static Resource Allocation (SRA). Regarding DRA, resource blocks are typically allocated to users, who experience good channel gains. As a result, it is expected that achieved link rates can exceed predicted values. In addition, the intricate impact of MAI in the uplink is not considered in those prediction methods. To the best of our knowledge, no other work in the literature has proposed prediction methods that project link rates for DRA and consider MAI in the uplink.

In this work, we bridge that gap in the literature by extending prediction methods targeting SRA to consider DRA. In general, we utilize throughput gains achieved by DRA in the last time slot to adjust the throughput approximation in the next time slot. Figure 5.7 illustrates the proposed method. The prediction procedure of the next slot (for instance $s+1$) consists of the following steps:

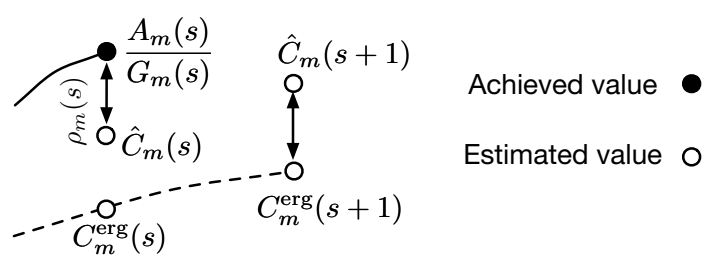

Figure 5.7: Illustration of the proposed estimation

1. The throughput gain $\rho_m(s)$ achieved by DRA is computed from the achieved throughput and the predicted value in the last time slot s. Particularly, we have:

$$\rho_m(s) = \frac{A_m(s)}{G_m(s)} - \hat{C}_m(s) \tag{5.11}$$

2. The ergodic capacity without potential gains by DRA is computed via one of three methods shown above. The approximation for the next slot $s+1$ utilizes known channel gains in the previous slot, i.e. $\{H_{i,m}(t)\}$.

3. The capacity with potential gains by DRA is predicted as:

$$\hat{C}_m(s+1) = C_m^{\text{erg}}(s+1) + \beta\rho_m(s) \tag{5.12}$$

   where β is a coefficient to scale the impact of previous gain $\rho_m(s)$ on the expectation.

Based on the predicted $\hat{C}_m(s+1)$, the throughput is obtained as:

$$\hat{A}_m(s+1) = \hat{C}_m(s+1)\hat{G}_m(s+1) \tag{5.13}$$

Note that $\hat{G}_m(s+1)$ is found by solving optimization problems (OP-LSH) and (OP-LSS).

5.4 Evaluation

We evaluate the proposed algorithms through Monte Carlo simulations based on the network simulator OMNeT++ with realistic models of wireless channels and video trace files.

5.4.1 Simulation setup

Wireless channels

We consider a cell where 4 users share 16 subchannels. In the time domain, one OFDMA frame spans over 47 OFDM symbols. Users move in the cell at the speed of 50 km/h following the Manhattan mobility grid model. The propagation channel consists of either LOS or NLOS. The probability of NLOS (as a function of distance d) is computed as: $p_{\text{NLOS}}(d) = 0.9\{1 - [1.24 - 0.6\log(d)^3]^{1/3}\}$. Path loss models for LOS and NLOS cases are recommended models in COST-231 [19].

We model shadowing loss that includes spatial correlation based on the MOSAIC model in [140].

To address realistic scenarios, we explicitly consider the fact that the fading process within a long duration of one video segment is most likely non-stationary. Particularly, we adopt the model of piece-wise stationary fading processes, each of which

is reasonably assumed to last for 100 ms [141]. The time-varying statistics of those stationary channels are derived from the highly detailed QuaDRIGA simulator [142].

Imperfect synchronization between users' signals is considered in the uplink. Specifically, users' signals yield small residual frequency offsets, which are equally distributed from 0 to 30 Hz (due to, e.g., oscillator inaccuracy and the Doppler effect), while perfectly synchronized in time. In addition, the specification of MCSs in LTE is considered for the ACM feature.

Link Rate Prediction

We evaluate prediction errors of three prediction methods: SG, LBP, and TBP. Especially, the evaluation is done in the downlink and in the uplink, with realistic assumptions like non-stationary channels and imperfect synchronization. Essentially, at the beginning of each time slot, we predict users' throughput using three methods. During the slot, resource blocks are assigned to users following either SRA or DRA schemes. By the end of that slot, we compute the actually achieved throughput and compare that value with predicted ones.

Regarding the prediction, a simple scenario is assumed: each user requires a link rate proportional to its capacity, i.e., $\hat{C}_m(s)/M$. To that aim, each users takes an equal share, i.e., $\hat{G}_m(s) = 1/M$. The expected throughput yields: $\hat{A}_m(s) = \hat{C}_m(s)/M$.

During slot s, blocking assignment is adopted as for the SRA approach. It means user takes a continuous chunk of N_{sch}/M subchannels. Alternatively, proposed DRA algorithms are exploit. Particularly, we set the expected throughput as targeted rates, i.e., $W_m(s) = \hat{C}_m(s)$, in (OP-DL) for the downlink and in (OP-UL) for the uplink.

We simulate and compute users' throughput $A_m(s)$ achieved using either SRA or DRA at the end of slot s. The prediction quality is assessed via relative prediction errors defined as $100(A_m(s) - \hat{A}_m(s))/\hat{A}_m(s)$ (in percentage). We perform a total of 30 simulation runs of 4 users in 60 seconds.

Video Traffic

We use the trace file of the movie *"The silence of the lambs"* provided by Arizona State University [132] to feed the simulation. In the chosen trace, video content is encoded nine times separately; the encoding scheme is MPEG-4 single layer (non-scalable) with the format G16B15. For convenience, the size of the video segment is set to equal one Group of Picture (GOP) of one second. It means there are 200 OFDMA frames within each video segment. Users' video sequences are extracted from the common trace but from different starting points to take into account the heterogeneity among users' video contents. In the case of LSH, video streams have the same size of 300 s in the downlink and 90 s in the uplink, and for LSS simulation terminates only when a minimum number of video segments are played back, which are 300 s in the downlink and 90 s in the uplink.

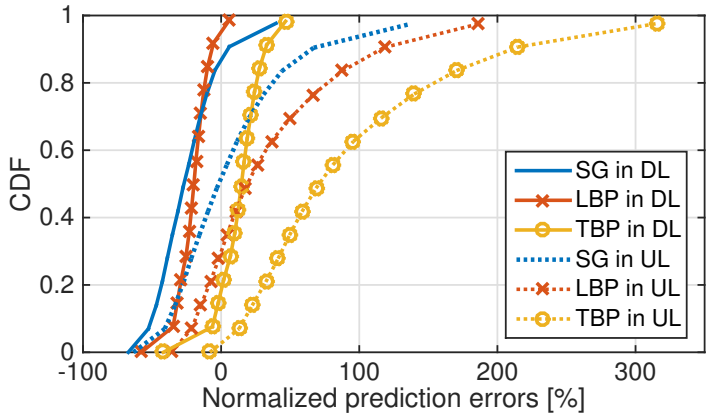

Figure 5.8: Performance of prediction methods

5.4.2 Evaluation of Link Rate Prediction

We show numerical results for SRA cases in Figure 5.8, where solid lines correspond to the downlink and dash lines to the uplink. Note that the continuous Shannon rate is used here instead of discrete ACM rates as computing users' throughput. The achieved throughput is then compared to the predicted ones. As it is shown, in general, TBP provides mostly overestimated values. LBP is the most conservative one since the expected values rarely overshoot the channel capacity. Finally, SG tends to give both under- and over-estimation. In addition, all methods lead to overestimation in the uplink due to the lack of consideration of MAI. Based on this result, the LBP seems to be the most appropriate choice for the SRA approaches in order to avoid video stalls.

Next, the results for DRA approaches are shown in comparison with the predicted throughput using the LBP method and achieved throughput through SRA. Note that the actual user throughput is calculated using a realistic model of link adaptation for both SRA and DRA, as opposed to the prediction step where the continuous Shannon rate is considered. The results shown in Figure 5.9 illustrate the significant improvement of DRA compared to SRA. From this figure, one can also see that throughput achievement by DRA tends to fluctuate around the predicted value in the downlink and exceed 50% of that in the uplink, respectively.

5.4.3 Evaluation of Video Performance

We then evaluate the video performance of the proposed algorithms and the baseline, which uses SRA and no exploitation of DRA and buffer levels. First, for the LSH, consistent with the results from the previous section, link rates of incoming time slot are estimated to equal 60% and 40% of the lower bound of ergodic capacity (i.e., $\hat{C}_m^{\text{LBP}}$) for the downlink and the uplink, respectively. In addition, after some preliminary investigation, the coefficient β in (5.12), used to scale the impact of DRA gain, is chosen as $\beta = 0.2$. We present QID, which is computed by function $U_m()$ (defined in (5.1)) for all segments of all streams in Figure 5.10 and number of skipped

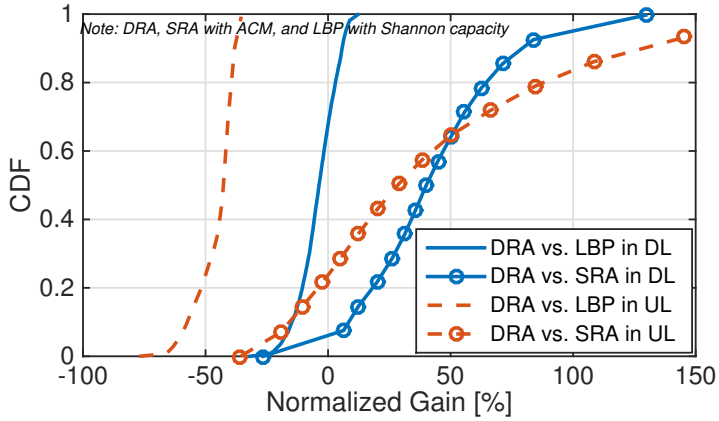

Figure 5.9: Throughput performance gain by DRA

segments in Figure 5.11. As it can be seen, DRA can greatly improve QoE by not only increasing the median of QID by 10dB in the downlink and more than 20dB in the uplink but also strongly mitigates the number of video stalls for both the downlink and the uplink cases.

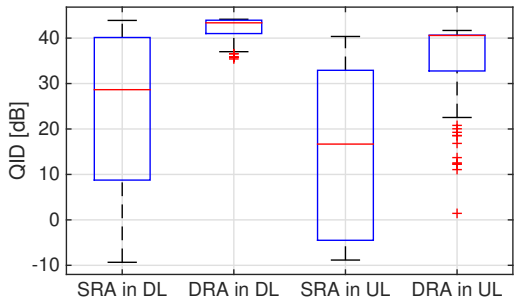

Figure 5.10: QID in case of LSH

Regarding the LSS case, predicted link rates are computed equal to 60% of $\hat{C}_m^{\text{LBP}}$ for SRA in the downlink and the uplink; these ratios are 80% and 60% used for DRA in the downlink and uplink. The parameters of proposed algorithms are selected as follows $\beta = 0.2$, $\lambda = 0.1$ for downlink, and $\beta = 0.2$, $\lambda = 0.1$ for the uplink. We show the overall number and duration of video stalls in cell in addition to QID in Figure 5.12. It can be seen that the performance gains by DRA are less obvious due to the deployment of buffer to absorb the throughput fluctuation. Particularly, DRA increases the median QID by more than 2 dB and 0.5 dB in the downlink and the uplink, respectively. Meanwhile, the advantage of DRA can be easily noticed as the proposed approaches can efficiently avoid the video stalls with respect to both number of events as well as duration.

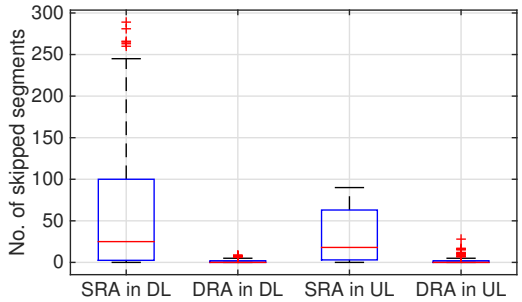

Figure 5.11: Number of skipped segments in case of LSH

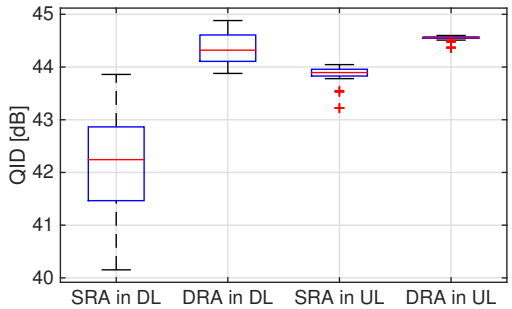

Figure 5.12: QID in case of LSS

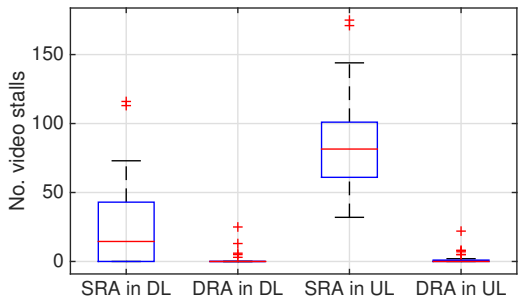

Figure 5.13: Number of skipped segments in case of LSS

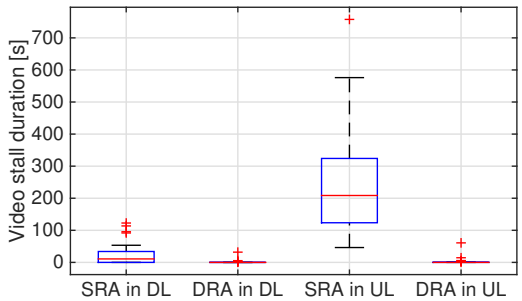

Figure 5.14: Interruption Duration in case of LSS

5.5 Conclusion

This chapter introduces novel cross-layer algorithms to enhance the QoE of multiple low-delay video streams in an OFDMA mobile cell. The proposed solutions consist of two key components: video quality selector and dynamic resource allocator. We address two use cases: low-delay streaming with hard latency constraints and with soft latency constraints. We separately consider downlink and uplink transmissions, where imperfect synchronization in the uplink distinguishes these two cases. Through DRA, the multiuser and frequency diversities can be efficiently exploited. Simultaneously, MAI in the uplink can be mitigated to provide valuable throughput gains, and resources can be allocated more efficiently to users who benefit most from them. As a separate contribution, we evaluate several link rate estimation methods revealing significant throughput gains due to DRA. Simulation results demonstrate significant QoE gains for all considered use cases.

Chapter 6

Cross-Layer Algorithm for Layered Video Coding

LVC has been supported by several modern standards like MPEG H.264-AVC/SVC and HEVC. By adopting LVC, video quality can be incrementally improved by streaming enhanced layers when the channel's capacity allows. This feature gives rise to a new paradigm of video adaptation.

This chapter presents a novel cross-layer approach, which jointly exploits the unique adaptation paradigm of LVC and DRA of OFDMA systems to improve the QoE of multiple low-delay streams. The main result of this chapter has been published in [143].

6.1 Layered Video Streaming over OFDMA networks

We consider one single cell using OFDMA, where M users compete for N_{sch} sub-channels to stream their videos simultaneously in the uplink. The OFDMA system model shown in Section 5.1.2 in the previous chapter is also adopted in this chapter. Therefore, the description of the adopted OFDMA system model is omitted in this section. The rest of this section concentrates on the video streaming system.

We aim to exploit the adaptation mechanism of LVS and target extremely low-delay applications, in which downloaded video content is not buffered but decoded and played immediately at the receiver. Such a solution is potentially desired by interactive streaming applications like video conferencing and autonomous driving.

We adopt the slotted time axis aligned with video segments' boundaries. At the beginning of a time slot, video data of the previous slot is completely encoded and queued in the transmission queue. When BS assigns radio resources to users, they dequeue and send video data from the base layer to the upper enhancement layers. Users stop sending video data in either two cases. In the first case, video data in the transmission buffer cannot be fully transmitted when the time slot is over. In the second case, users manage to transmit all segments' available video layers. In either case, queued video data is removed and replaced by the data of the next segment.

A central optimizer located at BS jointly considers video adaptation and resource allocation. Based on several factors (including instantaneous channel states, synchronization offsets between users, and video characteristics), appropriate resource allocation schemes are derived and sent to users. Resource allocation schemes are supposed to be sufficient so that all users in the cell can deliver good QoE.

It is assumed that transmitted data arrives at the receiver following the transmitted order so that they can be appropriately decoded. To realize that, video packages can be prioritized based on their significance in the hierarchical structure to maintain the correct decoding order. Besides, video packages can assumably be segmented arbitrarily small and fed to MAC frames.

We adopt the continuous rate-distortion model introduced in [144] to characterize the dependency of perceived quality on MMBR. That model was commonly adopted in several studies in the literature (e.g., [115], [145], [146]).

In this chapter, PSNR is also used to measure the perceived quality of video segments. Consequently, let $Q_m(s)$ be the PSNR value of segment s of stream m. It takes the following form.

$$Q_m(s) = 10\log_{10}\big[\frac{255^2}{\mathrm{MSE}(s)}\big] \tag{6.1}$$

where MSE presents the quality loss computed between the source and the reconstructed video. MSE can be efficiently approximated as below:

$$\mathrm{MSE}(s) \equiv D_e\big[R_e(s)\big] = \frac{\theta(s)}{R_e(s) - R_0(s)} + D_0(s) \tag{6.2}$$

The distortion D_e is a function of video bitrate R_e. Characteristics of each video segment is modeled by three coefficients denoted by θ, R_0 and D_0. It is assumed that those coefficients of all encoded segments are available for video adaptation algorithms.

Figure 6.1 illustrates the continuous rate-distortion curves of two segments of two streams. In this example, for the same perceived quality, the segment of the second user requires a lower bitrate than the one of the first user. One example factor causing that difference is the motion in video segments. The variety of video characteristics is referred to as video content diversity.

For simplicity, we write:

$$Q_m(s) = G[R_e(s)] \tag{6.3}$$

Since PSNR is a short-term quality measurement per segment, it cannot reflect the dependency of QoE on quality fluctuations between successive segments. To enforce the smoothness of the video adaptation trajectory, we utilize the average PSNR over multiple segments. In particular, let $\mu_m(s)$ be the mean perceived quality of all transmitted segments until the beginning of slot s. It means

$$\mu_m(s) = \frac{1}{s}\sum_{s'=0}^{s-1} Q_m(s') \tag{6.4}$$

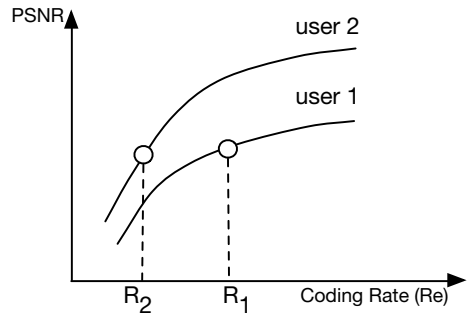

Figure 6.1: Illustration of the variety of video characteristics

By the end of video streams, we compute the overall QoE index, denoted by QID, to evaluate the overall performance in the cell. In particular, QID has the following form.

$$\text{QID} = \sum_{m=0}^{M-1} \left(\mu_m - \psi_m\right) \quad , \tag{6.5}$$

where μ_m and ψ_m are the mean and the variance of PSNR values over all segments of video contents, respectively.

Motivated by the same reason as discussed in Section 5.1, it is assumed that all video segments have an equal and static duration T_{seg}. Besides, we assume that each user streams in total N_{seg} segments in the uplink.

A theoretical transport protocol is adopted to provide lossless transmission. The amount of signaling data is also assumed to be negligible in comparison with the volume of video data.

6.2 Joint Adaptation Algorithm

This section presents our novel cross-layer algorithm, which aims to fairly increase the QoE of multiple low-delay video streams that compete for precious radio resources in the OFDMA uplink. The cross-layer algorithm jointly considers video adaptation and DRA.

To develop an efficient cross-layer algorithm, we aim to exploit the special adaptation principle of LVC. As mentioned in Section 2.4, while quality adaptation of NLVC happens only on segments' boundaries, LVC provides many adaptation points within each video segment. In particular, the quality of each layered video segment can be adapted on the fly by sending and decoding different numbers of enhancement layers. As a result, it is not necessary to select a specific quality for each segment before streaming that segment. For each video segment, the transmitter always starts streaming from the base layer to provide the minimum quality. It continues with enhancement layers in the hierarchical structure of layered video to incrementally increase the segment's quality if possible (e.g., the channel capacity allows). At this point, we make a remark: video adaptation can be indirectly controlled via the num-

ber of resources assigned to users. In addition, that streaming approach also eases the need for efficient estimations of future link rates.

Motivated by the above discussion, we envision a novel cross-layer approach, where the problem of optimizing video quality in the timescale of video segments (in seconds) can be split into a sequence of sub-problems, each of which is tightly coupled with the problem of optimizing DRA in the timescale of milliseconds. By doing that, adaptation algorithms can react quickly to channel fluctuations and avoid video stalls. Furthermore, we adopt the proposed DRA approach presented in Chapter 4 to explicitly cope with imperfect synchronization in the uplink and also achieve valuable throughput gains (by exploiting the channel diversity) to boost QoE.

To develop the desired cross-layer algorithm, we first consider the video adaptation problem. In this work, the max-min formulation is adopted to fairly improve all streams' qualities. Particularly, we formulate an OP to maximize the minimum weighted PSNR value among users. The weight is the average of achieved PSNR values, i.e., $\mu_m(s)$, to reduce quality fluctuations. We refer to this max-min weighted problem as **OP-VID**. The mathematical formulation of **OP-VID** (for slot s) is shown in (6.6).

$$\begin{aligned} &\max \quad \epsilon \\ s.t. \quad &\frac{Q_m(s)}{\mu_m(s)} \geq \epsilon \qquad ,\forall m \end{aligned} \tag{6.6}$$

where ϵ is an auxiliary variable to present the lower bound of weighted PSNR values, which is maximized.

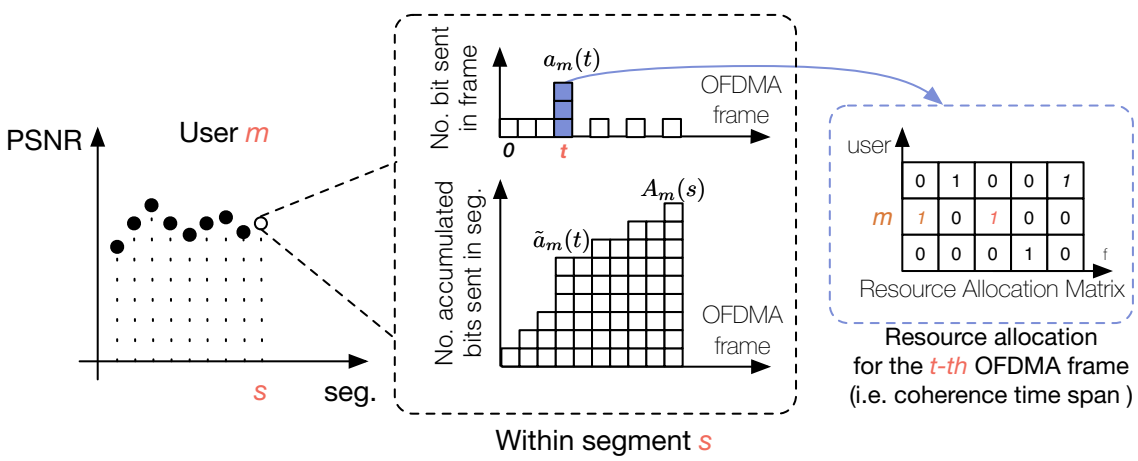

Figure 6.2: Illustration of the timescale difference

Figure 6.2 illustrates the timescale difference and the sequential DRA process in each video segment. As it is shown, one segment is transmitted in a large number N_{fis} of OFDMA frames. During each OFDMA frame t with $(0 < t < N_{\text{fis}})$ in segment s, user m takes some resources and sends $a_m(t)$ bits of video data.

The total amount of data user m sends in slot s, denoted by $A_m(s)$, is accumulated over all uplink frames in that slot.

$$A_m(s) = \sum_{t=0}^{N_{\text{fis}}-1} a_m(t) \tag{6.7}$$

The computation of perceived quality of segment s of the stream m is then given by:

$$Q_m(s) = G[A_m(s)] = G\big[\sum_{t=0}^{N_{\text{fis}}-1} a_m(t)\big] \tag{6.8}$$

As it shows the quality $Q_m(s)$ is achieved through a sequential process of transmission of N_{fis} OFDMA frames.

6.2.1 Sequential Process of Quality Driven Resource Allocation

Let us consider resource allocation of, for instance, uplink frame t in slot s. During this frame, the instantaneous channel condition within frame t is available for optimizing resource allocation. Before frame t, user m has sent an accumulated data amount of $\tilde{a}_m(t-1)$.

Next, BS derive an efficient resource allocation scheme. Following that scheme, resource blocks are assigned to users. Using the assigned resources, user m sends a data amount of $a_m(t)$. Then, we can write:

$$\tilde{a}_m(t) = \tilde{a}_m(t-1) + a_m(t) \tag{6.9}$$

Corresponding to the accumulated amount of transmitted data $\tilde{a}_m(t)$, the intermediate quality $q_m(t)$ is estimated by the following formulation:

$$q_m(t) = G[\tilde{a}_m(t)] = q_m(t-1) + g(a_m(t)) \tag{6.10}$$

Essentially, function $g(.)$ estimates the intermediate quality achieved for the currently-being-sent segment. Based on $g(.)$, one can drive the sequence of resource adaptation (each happens for one uplink frame) so that the intermediate quality is incrementally improved toward the desired quality $Q_m(s)$ by the end of the slot.

Next, we present an efficient formulation of the coupling function $g(.)$. First, based on the definition, we have

$$Q_m(s) = 10\log_{10}(255^2) - 10\log_{10}\Big[\frac{\theta(s)}{R_e(s) - R_0(s)} + D_0(s)\Big] \tag{6.11}$$

For simplicity, we omit the notion of segment s. Then we take the derivative of PSNR $Q_m(s)$ to R_e

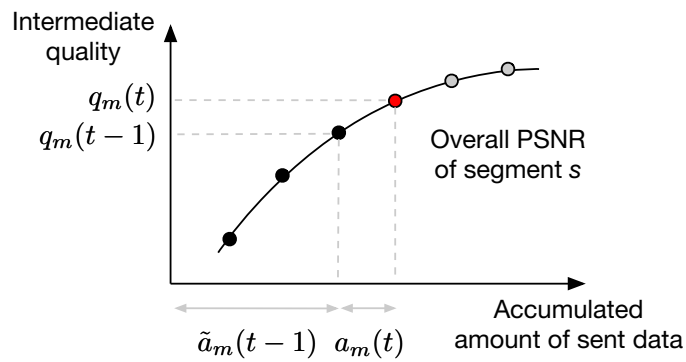

Figure 6.3: Illustration of the sequential process

$$\frac{\delta Q_m(s)}{\delta R_e} = \frac{10\theta}{\ln(10)} \times \frac{1}{(\theta + D_0(R_e - R_0))(R_e - R_0)} \tag{6.12}$$

Furthermore, since one segment typically consists of a large number of uplink frames, the amount of transmitted data during each uplink frame is much smaller than the amount of multimedia data. Hence, it is reasonable to approximate $\delta R_e \approx a_m(t)/T_{\text{seg}}$.

Then, we can write

$$\delta Q_m(s) \approx q_m(t) - q_m(t-1) = \Delta q_m(t) = g(a_m(t)) \tag{6.13}$$

where

$$g(a_m(t)) = \frac{10\theta}{\ln(10)} \frac{a_m(t)/T_{\text{seg}}}{(\theta + D_0(\tilde{a}_m(t-1)/T_{\text{seg}} - R_0))(\tilde{a}_m(t-1)/T_{\text{seg}} - R_0)} \tag{6.14}$$

Note that parameters θ, D_0, R_0 and $\tilde{a}_m(t)$ are known in the frame t. Then, for simplicity, we can write:

$$q_m(t) = q_m(t-1) + \alpha_m(t)\frac{a_m(t)}{T_{\text{seg}}} \tag{6.15}$$

where the coefficient $\alpha_m(t)$ takes the following form

$$\alpha_m(t) = \frac{10\theta}{\ln(10)} \frac{1}{(\theta + D_0(\tilde{a}_m(t-1)/T_{\text{seg}} - R_0))(\tilde{a}_m(t-1)/T_{\text{seg}} - R_0)} \tag{6.16}$$

Function $g(.)$ becomes a very simple linear function of the intermediate throughput $a_m(t)$. The simplicity of $g(.)$ provides a huge computational benefit as we design the cross-layer algorithm of interest.

6.2.2 Dynamic Resource Allocation

In this section, we consider the problem of optimizing resource allocation. We extend the approach introduced in Section 4.4 to include the QoE goal by using the novel function $g(.)$. In particular, we pursue the goal of **OP-VID** in a series of quality-driven DRA OPs, referred to as **OP-DRA**. At the beginning of each uplink frame, one instance of OP-DRA is formulated and solved to find the appropriate resource allocation scheme, which can improve the intermediate quality toward the desired quality value. The general formulation of OP-DRA is described as follows.

$$\begin{aligned} &\max \quad \epsilon \\ s.t. \quad &\frac{q_m(t)}{\mu_m(s)} \geq \epsilon \ , \quad \forall m \end{aligned} \tag{6.17}$$

We combine the introduced coupling function $g(.)$ and the equivalent transformation of function $F(.)$ above to formulate the exact formulation of **OP-DRA** as shown in (6.18).

$$\begin{aligned} &\max \quad \epsilon \\ s.t. \quad &a) \ \frac{1}{\mu_m(s)}\Big[q_m(t-1) + \frac{\alpha_m(t)}{T_{\text{seg}}} \sum_{i=0}^{N_{\text{sch}}-1} x_{i,m}(t) b_{i,m}(t)\Big] \geq \epsilon \ , \qquad \forall m \\ &b) \ \sum_m x_{i,m}(t) \leq 1 \ , \qquad \forall i \\ &c) \ \sum_{k=0}^{K} z_{i,m,k}\Lambda_{i,m,k} \geq \Big[\sum_{j\neq i}\sum_{m'\neq m} \frac{\overline{\text{MAI}}_{i,m}^{j,m'}}{\sigma^2} x_{j,m'}(t)\Big] \ , \quad \forall i, m \\ &d) \ \sum_{k=0}^{K} z_{i,m,k} = 1 \ , \quad \forall i, m \\ &e) \ b_{i,m}(t) \leq \sum_{k=0}^{K} z_{i,m,k} B_k \ , \quad \forall i, m \end{aligned} \tag{6.18}$$

Note that the last three constraints in (6.18) represents the selection of MCS scheme as discussed in (4.18).

6.3 Evaluation

We evaluate the performance of the proposed cross-layer algorithm and baselines via simulation. Particularly, we develop a simulation based on the network simulation framework OMNeT++ and using the optimization solver Gurobi. In total, we simulate three other solutions as baselines.

- Static resource allocation + Rate Adapting:

  Resource allocation and video adaptation run independently. Resource allocation schemes are static. Particularly, the block-wise assignment is adopted, where each user takes an equal and continuous block from the available bandwidth. Segments' video bitrates are equal to achieved link rates (thus named Rate Adapting).

- Adaptive resource allocation + Rate Adapting:

  Resource adaptation and video adaptation do not collaborate. This solution deploys the proposed DRA in Chapter 4, which maximizes the minimum user throughput without considering how much users benefit from the given throughput (i.e., throughput-driven DRA). Video bitrate is then matched to the available throughput. By considering this setup, we can investigate the benefit of cross-layer algorithms.

- Proportional-Fairness resource allocation + Rate Adapting:

  For a comparison with other works in literature, the solution in [137] might be the most appropriate. However, since the main focus of that work associates with several complex aspects of video delivery, while resource allocation is generous, we cannot directly compare our work with it. Therefore, we choose to compare with a baseline used in [137]. It is essentially a DRA algorithm based on proportional fairness. Particularly, the corresponding optimization is described as follows.

$$\max \Big[\sum_{m} \frac{1}{\rho_m} x_{i,m} F(\bar{\gamma}_{i,m}, P_{\text{err}}) \Big] \tag{6.19}$$

  where ρ_m is the peak throughput achieved when user m takes all resources. The video data transmission rate is independently matched to the available resource.

For each uplink frame, the same profile of the system condition is fed to all simulated solutions. Based on the resource allocation schemes after solving DRA OPs, the simulation computes SINR, selects appropriate MCS schemes, computes the total amount of data users send (i.e., a_m^t). After the last frame of the segment, the video quality of the segment s is computed: $Q_m(s) = G[(] \sum_t a_m^t]$.

We assume there are four users are streaming videos in the uplink. Each video stream consists of 30 segments, and each segment in its turn includes 25 OFDMA uplink frames, i.e., $N_{\text{fis}} = 25$ for al users. To simulate segments' characteristics, we adopt the distortion rate curves introduced in [147]. In total, we use four different video profiles to simulate video contents. We illustrate these data profiles in Figure 6.4.

Furthermore, to simulate video content diversity, we adopt the following policy. In the beginning, four segments sent by four users have the characteristics following four different data profiles, mainly user m takes profile m. After an interval of 5 segments, each user takes the profile with the index increased by 1, i.e., user m now

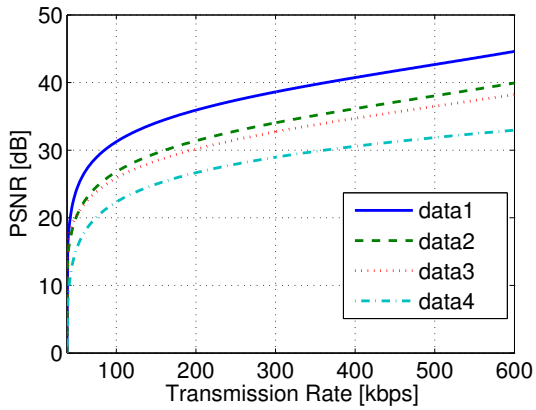

Figure 6.4: Video Rate-Distortion Curves

takes profile $m+1$ and, especially user four now takes the first profile. The process continues similarly until the end of the video sequences. The main parameters for the OFDMA system are listed in Table 6.1.

Table 6.1: Simulation parameters for channel.

Parameters	Values
Number of subcarriers	512
Number of subchannels	8
Number of subcarriers per subchannel	64
Power per subcarrier	$1\,mW$
MS maximal time offset	$2\,\mu s$
MS maximal frequency offset	$200\,Hz$
Cell radius	$100\,m$

A total of 30 runs are conducted for each solution's simulation. For each run, in the beginning, users' locations are defined in the way that their distances to BS are uniformly distributed in the cell. During the process of video streaming, mobility is modeled based on the Manhattan Mobility Grid. In each run, 750 optimization instances (for 30 segments times 25 uplink frames) times four solutions are formulated and solved.

We compute the overall QID in the cell. Figure 6.5 shows the average of QID over 120 simulated video streams. As it shows, the novel cross-layer algorithm significantly improves QID, particularly by about 75% and 40% in comparison to those of the **(MMT+RA)** and **(PF+RA)**.

Next, we analyze achieved PSNR values to find out what is the cause of the QoE improvement. To demonstrate all results, we show in Figure 6.6 the cumulative distribution function (CDF) of all PSNR values collected for 3600 simulated segments (30 runs times four users times 30 segments) of each solution. Through the CDF, we show the quality variability. As it shows, the novel approach outperforms others concerning the quality fluctuation.

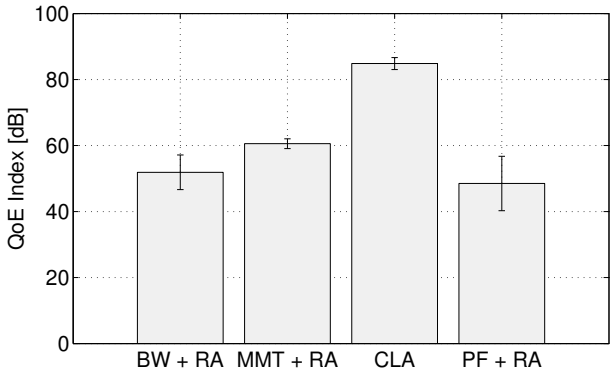

Figure 6.5: QoE Index (QID)

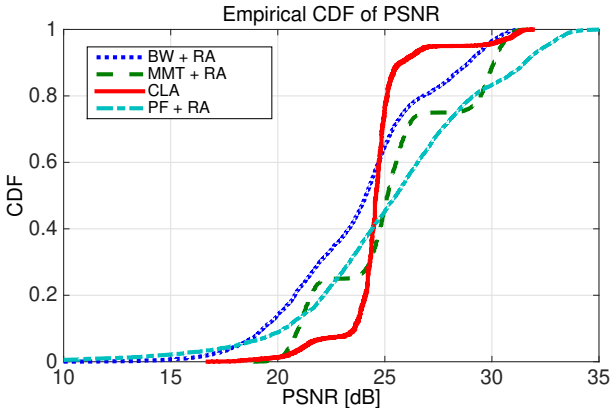

Figure 6.6: CDF of PSNR values

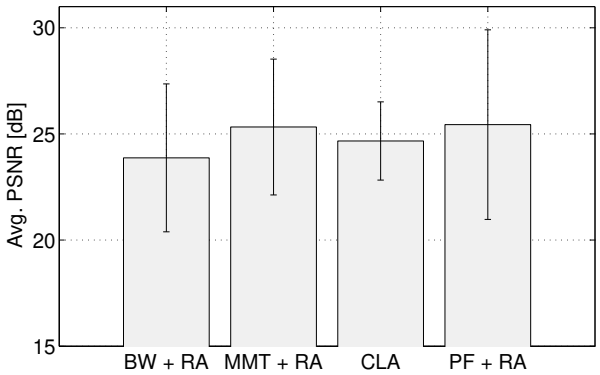

Figure 6.7: Average of PSNR values

We show in Figure 6.7 the average of PSNR achieved in different solutions. As it shows, the cross-layer algorithm cannot provide a higher PSNR average. The proportional fairness and max-min user throughput, which run independently from the video adaptation, can provide a better average PSNR. This can be explained by

the nature of the max-min fairness, where the worse user essentially limits the whole system performance.

Therefore, the main cause for the QoE improvement is the efficient reduction of the PSNR fluctuation between segments. In other words, the novel cross-layer algorithm can match the achieved throughput to the bitrate demand.

6.4 Conclusion

In this chapter, we introduce a novel cross-layer approach, which jointly considers video adaptation using LVC and resource adaptation in OFDMA networks, to improve the QoE of multiple video streams in the uplink. Specifically, our objective is to increase QoE in three aspects, which are (1) increasing short-term quality per video segment, (2) reducing the quality variation between adjacent segments, and (3) taking into account quality fairness among users. To deal with the unavoidable problem of timescale difference, we pursue the QoE improvement through a sequential series of short-scale quality-driven DRA problems. The proposed quality-driven DRA exploits the frequency and multi-user diversities to improve spectral efficiency and at the same time assign resources to users subject to the long-term QoE goal.

Chapter 7

Conclusions and Outlook

Video streaming has been the main traffic generator in mobile networks. The video traffic growth increasingly forces mobile operators and video service providers to look for efficient solutions to make the best possible service quality out of the limited radio spectrum. One of the open issues is how to develop robust video adaptation algorithms for low-delay streaming applications over mobile networks. Most studies in the literature consider VoD (which targets a buffering delay of several seconds). Low-delay streaming receives only a little attention.

This thesis introduces novel cross-layer approaches to improve the QoE of multiple low-delay video streams in the uplink of mobile networks using OFDMA. Those approaches jointly consider the slow-paced video adaptation and the fast-paced dynamic resource allocation. Consequently, adaptive algorithms are developed to quickly react to wireless channel variations and assure the delivery of video segments before their deadlines. Regarding resource allocation, a novel dynamic resource allocation approach is proposed to efficiently suppress MAI and enhance user throughput.

Several issues remain as future work. First, other objective functions that balance fairness between users and spectral efficiency can be developed to simplify introduced cross-layer algorithms. Second, the complexity of the proposed OP can be further reduced by adopting more efficient mathematical transformations (e.g., Lagrangian relaxation). Third, while in principle all presented algorithms can apply to the downlink, the evaluation remains a future work issue. Finally, an implementation of the introduced algorithms on a real-world testbed would strongly justify the performance gain in according systems.

Appendix A

Acronym

3GPP Third Generation Partnership Project

ACM Adaptive Coding and Modulation

ADC Analog to Digital Converter

ADSL Asymmetric Digital Subscriber Line

AVC Advanced Video Coding

AWGN Additive White Gaussian Noise

BER Bit Error Rate

BS Base Station

CBR Constant Bitrate

CDN Content Delivery Network

CIR Channel Impulse Response

CP Cyclic Prefix

CSI Channel State Information

DAB Digital Audio Broadcasting

DAC Digital to Analog Converter

DASH Dynamic Adaptive Streaming over HTTP

DCT Discrete Cosine Transform

DFT Discrete Fourier Transform

DRA Dynamic Resource Allocation

DVB Digital Video Broadcasting

DVD Digital Versatile Disk

ECBA Estimation-Correction Based Approach

EM Expectation Maximization

FDD Frequency Division Duplexing

FFT Fast Fourier Transform

GB Guard Band

GI Guard Interval

GOP Group of Picture

HAS HTTP Based Adaptive Streaming

HEVC High Efficiency Video Coding

HTTP Hypertext Transport Protocol

ICI Inter-Carrier Interference

IDFT Inverse Discrete Fourier Transform

IEC International Electrotechnical Commission

IFFT Inverse Fast Fourier Transform

ISI Inter-Symbol Interference

ISO International Standard Organization

ITU-T International Telecommunication Union - Telecommunication Sector

LOS Line of Sight

LSH Hard Latency Constraint

LSS Soft Latency Constraint

LTE Long Term Evolution

LVC Layered Video Coding

LVS Layered Video Streaming

MAC Media Access Control

MAI Multiple Access Interference

MCS Modulation and Coding Scheme

MDC Multiple Description Coding

MIMO Multiple Input Multiple Output

MIP Mixed-Integer Programming

MIQCP Mixed Integer Quadratically Constrained Problem

ML Maximum Likelihood

MMBR Mean Media Bit Rate

MMSE Minimum Mean Square Error

MOS Mean Opinion Score

MPEG Moving Picture Experts Group

MS Mobile Station

MSE Mean Square Error

NAT Network Address Translation

NLOS Non Line of Sight

NLVC Non-Layered Video Coding

NLVS Non-Layered Video Streaming

NUM Network Utility Maximization

OFDM Orthogonal Frequency Division Multiplexing

OFDMA Orthogonal Frequency Division Multiple Access

OP Optimization Problem

OSI Open System Interconnection

PDF Probability Density Function

PER Packet Error Rate

PLL Phrase Locked Loop

PN Pseudo Noise

PSD Power Spectral Density

PSK Phase Shift Keying

PSNR Peak Signal-to-Noise Ratio

PSS Packet-switched Streaming Service

PUSC Partially Used Sub-Channelization

QAM Quadrature Amplitude Modulation

QID QoE Index

QoE Quality of Experience

QoS Quality of Service

RA Resource Allocation

RAN Radio Access Network

SAGE Space Alternating Generalized Expectation Maximization

SINR Signal to Noise plus Interference Ratio

SNR Signal to Noise Ratio

SRA Static Resource Allocation

SSIM Structural Similarity

SVC Scalable Video Coding

TCP Transport Control Protocol

TDD Time Division Duplexing

TDMA Time Division Multiple Access

TTI Transmission Time Interval

UDP User Datagram Protocol

UGC User Generated Content

VBR Variable Bitrate

VCO Voltage Controlled Oscillator

VoD Video on Demand

WiMAX Worldwide Interoperability for Microwave Access

Appendix B

Publication

Conference Proceedings

- Hieu Le, Daniel Willkomm, Adam Wolisz. Optimizing User Throughput with Consideration of Multiple Access Interference in the OFDMA Uplink. In Proceedings of the International Wireless Communications and Mobile Computing Conference (IWCMC), Jul. 2013.
- Hieu Le, Arash Behboodi, and Adam Wolisz. Dynamic Resource Allocation in OFDMA Uplink for MAI Mitigation and Throughput Improvement. In Proceedings of the IEEE 80th Vehicular Technology Conference (VTC), Sep. 2014.
- Hieu Le, Arash Behboodi, and Adam Wolisz. Quality driven resource allocation for adaptive video streaming in OFDMA uplink. In Proceedings of the IEEE 26th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC), Aug. 2015.
- Hieu Le, Konstantin Miller, Arash Behboodi, and Adam Wolisz. Cross-layer approach for HTTP-based low-delay adaptive streaming in mobile networks. In Proceedings of the IEEE 18th International Symposium on A World of Wireless, Mobile and Multimedia Networks (WoWMoM), Jun. 2017.

Technical Reports

- Hieu Le, Arash Behboodi, Adam Wolisz, "Multiple Access Interference Mitigation in OFDMA Uplink via static assignment of Guard Bands and Guard Intervals", TKN Technical Report Series TKN-13-004, Technical University Berlin, Oct. 2013.
- Hieu Le, Arash Behboodi, Adam Wolisz, "Multiple Access Interference Mitigation in OFDMA Uplink using Dynamic Resource Allocation and Guard Bands", TKN Technical Report Series TKN-13-005, Technical University Berlin, Oct. 2013.

Bibliography

[1] Cisco, "Cisco Visual Networking Index: Forecast and Trends, 2017–2022 (White Paper)," Tech. Rep., 2019.

[2] K. Brunnström, S. A. Beker, K. De Moor, A. Dooms, S. Egger, M.-N. Garcia, T. Hossfeld, S. Jumisko-Pyykkö, C. Keimel, M.-C. Larabi, B. Lawlor, P. Le Callet, S. Möller, F. Pereira, M. Pereira, A. Perkis, J. Pibernik, A. Pinheiro, A. Raake, P. Reichl, U. Reiter, R. Schatz, P. Schelkens, L. Skorin-Kapov, D. Strohmeier, C. Timmerer, M. Varela, I. Wechsung, J. You, and A. Zgank, *Qualinet White Paper on Definitions of Quality of Experience*. Mar. 2013.

[3] P. Sweeting, "Video in 2014: Going live and over the top (Analysis Report)," GigaOM Media, San Francisco, CA., Tech. Rep., Jul. 2014.

[4] F. Capozzi, G. Piro, L. Grieco, G. Boggia, and P. Camarda, "Downlink Packet Scheduling in LTE Cellular Networks: Key Design Issues and a Survey," *IEEE Commun. Surv. Tutor.*, vol. 15, no. 2, pp. 678–700, Second 2013, ISSN: 1553-877X. DOI: 10.1109/SURV.2012.060912.00100.

[5] M. Andrews, "A Survey of Scheduling Theory in Wireless Data Networks," in *Wireless Communications*, vol. 143, May 2010, pp. 1–17. DOI: 10.1007/978-0-387-48945-2_1.

[6] M. v. D. Schaar and S. S. N, "Cross-layer wireless multimedia transmission: Challenges, principles, and new paradigms," *IEEE Wirel. Commun.*, vol. 12, no. 4, pp. 50–58, Aug. 2005, ISSN: 1536-1284. DOI: 10/dgvz48.

[7] X. Xie, X. Zhang, S. Kumar, and L. E. Li, "piStream: Physical Layer Informed Adaptive Video Streaming over LTE," in *Proceedings of the 21st Annual International Conference on Mobile Computing and Networking*, ser. MobiCom '15, New York, NY, USA: ACM, 2015, pp. 413–425, ISBN: 978-1-4503-3619-2. DOI: 10/gcpx6d.

[8] J. Chen, R. Mahindra, M. A. Khojastepour, S. Rangarajan, and M. Chiang, "A Scheduling Framework for Adaptive Video Delivery over Cellular Networks," in *Proceedings of the 19th Annual International Conference on Mobile Computing & Networking*, ser. MobiCom '13, New York, NY, USA: ACM, 2013, pp. 389–400, ISBN: 978-1-4503-1999-7. DOI: 10.1145/2500423.2500433.

[9] Y. Sanchez, E. Grinshpun, D. Faucher, T. Schieri, and S. Sharma, "Low latency DASH based streaming over LTE," in *2014 IEEE Visual Communications and Image Processing Conference*, Dec. 2014, pp. 1–4. DOI: 10.1109/VCIP.2014.7051489.

[10] A. M. Tonello, N. Laurenti, and S. Pupolin, "Analysis of the uplink of an asynchronous multi-user DMT OFDMA system impaired by time offsets, frequency offsets, and multi-path fading," in *Vehicular Technology Conference Fall 2000. IEEE VTS Fall VTC2000. 52nd Vehicular Technology Conference (Cat. No.00CH37152)*, vol. 3, 2000, 1094–1099 vol.3. DOI: 10.1109/VETECF.2000.886275.

[11] M. Morelli, C. C. J. Kuo, and M. O. Pun, "Synchronization Techniques for Orthogonal Frequency Division Multiple Access (OFDMA): A Tutorial Review," *Proc. IEEE*, vol. 95, no. 7, pp. 1394–1427, Jul. 2007, ISSN: 0018-9219. DOI: 10.1109/JPROC.2007.897979.

[12] X. Wang, T. T. Tjhung, Y. Wu, and B. Caron, "SER performance evaluation and optimization of OFDM system with residual frequency and timing offsets from imperfect synchronization," *IEEE Trans. Broadcast.*, vol. 49, no. 2, pp. 170–177, Jun. 2003, ISSN: 0018-9316. DOI: 10/fgvshm.

[13] J. Gross and M. Bohge, "Dynamic Mechanisms in OFDM Wireless Systems: A Survey on Mathematical and System Engineering Contributions," TKN, Technical Report TKN-06-001, Jan. 2006.

[14] T. Rappaport, *Wireless Communications: Principles and Practice*, Second. USA: Prentice Hall PTR, 2001, ISBN: 978-0-13-042232-3.

[15] M. Feuerstein, K. Blackard, T. Rappaport, S. Seidel, and H. Xia, "Path loss, delay spread, and outage models as functions of antenna height for microcellular system design," *IEEE Trans. Veh. Technol.*, vol. 43, no. 3, pp. 487–498, Aug. 1994, ISSN: 1939-9359. DOI: 10.1109/25.312809.

[16] A. Neskovic, N. Neskovic, and G. Paunovic, "Modern approaches in modeling of mobile radio systems propagation environment," *IEEE Commun. Surv. Tutor.*, vol. 3, no. 3, pp. 2–12, Third 2000, ISSN: 1553-877X. DOI: 10.1109/COMST.2000.5340727.

[17] Y. Oda, R. Tsuchihashi, K. Tsunekawa, and M. Hata, "Measured path loss and multipath propagation characteristics in UHF and microwave frequency bands for urban mobile communications," in *IEEE VTS 53rd Vehicular Technology Conference, Spring 2001. Proceedings (Cat. No.01CH37202)*, vol. 1, May 2001, 337–341 vol.1. DOI: 10.1109/VETECS.2001.944860.

[18] W. C. Y. Lee, *Mobile Cellular Telecommunications: Analog and Digital Systems*. McGraw-Hill, 1995, ISBN: 978-0-07-038089-9.

[19] COST231, "Urban transmission loss models for mobile radio in the 900- and 1,800 MHz bands (Revision 2)," The Hague, The Netherlands, Tech. Rep., Sep. 1991.

[20] J. K. Cavers, *Mobile Channel Characteristics*. USA: Kluwer Academic Publishers, 2000, ISBN: 0-7923-7926-8.

[21] M. Gudmundson, "Correlation model for shadow fading in mobile radio systems," *Electron. Lett.*, vol. 27, no. 23, pp. 2145–2146, Nov. 1991, ISSN: 0013-5194. DOI: 10.1049/el:19911328.

[22] K. Wehrle, M. Günes, and J. Gross, *Modeling and Tools for Network Simulation*. Springer Science & Business Media, Sep. 2010, ISBN: 978-3-642-12331-3.

[23] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication*. May 2005. DOI: 10.1017/CBO9780511807213.

[24] W. C. Jakes, *Microwave Mobile Communications*. IEEE Press, 1974, ISBN: 978-0-7803-1069-8.

[25] S. Faruque, "Introduction to Channel Coding," in *Radio Frequency Channel Coding Made Easy*, ser. SpringerBriefs in Electrical and Computer Engineering, S. Faruque, Ed., Cham: Springer International Publishing, 2016, pp. 1–16, ISBN: 978-3-319-21170-1. DOI: 10.1007/978-3-319-21170-1_1.

[26] R. Prasad, *OFDM for Wireless Communications Systems*. Artech House, 2004, ISBN: 978-1-58053-799-5.

[27] D. Karwowski, T. Grajek, K. Klimaszewski, O. Stankiewicz, J. Stankowski, and K. Wegner, "20 Years of Progress in Video Compression – from MPEG-1 to MPEG-H HEVC. General View on the Path of Video Coding Development," in *Image Processing and Communications Challenges 8*, R. S. Choraś, Ed., vol. 525, Cham: Springer International Publishing, 2017, pp. 3–15, ISBN: 978-3-319-47273-7 978-3-319-47274-4. DOI: 10.1007/978-3-319-47274-4_1.

[28] N. Ahmed, T. Natarajan, and K. R. Rao, "Discrete Cosine Transform," *IEEE Trans. Comput.*, vol. C-23, no. 1, pp. 90–93, Jan. 1974, ISSN: 0018-9340. DOI: 10/c4kqx4.

[29] Bitmovin, "2019 Video Developer Report - The Future of Video: AV1 Codec, AI & Machine Learning, and Low Latency," Tech. Rep., Sep. 2019, ch. Blog Post.

[30] T. Wiegand, G. Sullivan, G. Bjontegaard, and A. Luthra, "Overview of the H.264/AVC video coding standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 13, no. 7, pp. 560–576, Jul. 2003, ISSN: 1051-8215. DOI: 10/cr56b4.

[31] H. Schwarz, D. Marpe, and T. Wiegand, "Overview of the Scalable Video Coding Extension of the H.264/AVC Standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17, no. 9, pp. 1103–1120, Sep. 2007, ISSN: 1051-8215. DOI: 10/bg25pj.

[32] ITU-T, "Requirements for low-latency interactive multimedia streaming. Recommendation F.746.1," International Telecommunication Union (ITU), Recommendation, 2014.

[33] M. Seufert, S. Egger, M. Slanina, T. Zinner, T. Hoßfeld, and P. Tran-Gia, "A Survey on Quality of Experience of HTTP Adaptive Streaming," *IEEE Commun. Surv. Tutor.*, vol. 17, no. 1, pp. 469–492, Firstquarter 2015, ISSN: 1553-877X. DOI: 10.1109/COMST.2014.2360940.

[34] Y. Sánchez, T. Schierl, C. Hellge, T. Wiegand, D. Hong, D. D. Vleeschauwer, W. V. Leekwijck, and Y. Lelouedec, "Improved caching for HTTP-based Video on Demand using Scalable Video Coding," in *2011 IEEE Consumer Communications and Networking Conference (CCNC)*, Jan. 2011, pp. 595–599. DOI: 10/cghmsr.

[35] V. Goyal, "Multiple description coding: Compression meets the network," *IEEE Signal Process. Mag.*, vol. 18, no. 5, pp. 74–93, Sept./2001, ISSN: 10535888. DOI: 10/bbsr55.

[36] R. Huysegems, B. De Vleeschauwer, T. Wu, and W. Van Leekwijck, "SVC-Based HTTP Adaptive Streaming," *Bell Labs Tech. J.*, vol. 16, no. 4, pp. 25–41, Mar. 2012, ISSN: 1538-7305. DOI: 10/gcpvcz.

[37] T. Oelbaum, H. Schwarz, M. Wien, and T. Wiegand, "Subjective performance evaluation of the SVC extension of H.264/AVC," in *2008 15th IEEE International Conference on Image Processing*, Oct. 2008, pp. 2772–2775. DOI: 10.1109/ICIP.2008.4712369.

[38] M. Belshe, M. Thomson, and R. Peon, "Hypertext Transfer Protocol Version 2 (HTTP/2)," IEFT, Tech. Rep., 2015.

[39] Y. Chen, K. Wu, and Q. Zhang, "From QoS to QoE: A Tutorial on Video Quality Assessment," *IEEE Commun. Surv. Tutor.*, vol. 17, no. 2, pp. 1126–1165, Secondquarter 2015, ISSN: 1553-877X. DOI: 10/gcpx8r.

[40] Z. Wang, A. C. Bovik, and H. R. Sheikh, "Image Quality Assessment: From Error Visibility to Structural Similarity," *IEEE Trans. IMAGE Process.*, vol. 13, no. 4, p. 14, 2004. DOI: 10/c7sr27.

[41] T. De Pessemier, K. De Moor, W. Joseph, L. De Marez, and L. Martens, "Quantifying the Influence of Rebuffering Interruptions on the User's Quality of Experience During Mobile Video Watching," *IEEE Trans. Broadcast.*, vol. 59, no. 1, pp. 47–61, Mar. 2013, ISSN: 1557-9611. DOI: 10.1109/TBC.2012.2220231.

[42] T. Hoßfeld, M. Seufert, M. Hirth, T. Zinner, P. Tran-Gia, and R. Schatz, "Quantification of YouTube QoE via Crowdsourcing," in *2011 IEEE International Symposium on Multimedia*, Dec. 2011, pp. 494–499. DOI: 10.1109/ISM.2011.87.

[43] H. Nogami and T. Nagashima, "A frequency and timing period acquisition technique for OFDM systems," in *Proceedings of 6th International Symposium on Personal, Indoor and Mobile Radio Communications*, vol. 3, Sep. 1995, pp. 1010–. DOI: 10/b2g5sm.

[44] M. Schmidl and D. C. Cox, "Blind synchronisation for OFDM," *Electron. Lett.*, vol. 33, no. 2, pp. 113–114, Jan. 1997, ISSN: 0013-5194. DOI: 10/cn4vmr.

[45] H. Minn, M. Zeng, and V. K. Bhargava, "On timing offset estimation for OFDM systems," *IEEE Commun. Lett.*, vol. 4, no. 7, pp. 242–244, Jul. 2000, ISSN: 1089-7798. DOI: 10/b3b8h6.

[46] K. Shi and E. Serpedin, "Coarse frame and carrier synchronization of OFDM systems: A new metric and comparison," *IEEE Trans. Wirel. Commun.*, vol. 3, no. 4, pp. 1271–1284, Jul. 2004, ISSN: 1536-1276. DOI: 10/dxxgjp.

[47] P. H. Moose, "A technique for orthogonal frequency division multiplexing frequency offset correction," *IEEE Trans. Commun.*, vol. 42, no. 10, pp. 2908–2914, Oct. 1994, ISSN: 0090-6778. DOI: 10/cpxczs.

[48] A. J. Coulson, "Maximum likelihood synchronization for OFDM using a pilot symbol: Analysis," *IEEE J. Sel. Areas Commun.*, vol. 19, no. 12, pp. 2495–2503, Dec. 2001, ISSN: 0733-8716. DOI: 10/b2565w.

[49] ——, "Maximum likelihood synchronization for OFDM using a pilot symbol: Algorithms," *IEEE J. Sel. Areas Commun.*, vol. 19, no. 12, pp. 2486–2494, Dec. 2001, ISSN: 0733-8716. DOI: 10.1109/49.974613.

[50] J. J. van de Beek, M. Sandell, and P. O. Borjesson, "ML estimation of time and frequency offset in OFDM systems," *IEEE Trans. Signal Process.*, vol. 45, no. 7, pp. 1800–1805, Jul. 1997, ISSN: 1053-587X. DOI: 10/b33rc5.

[51] H. Bolcskei, "Blind estimation of symbol timing and carrier frequency offset in wireless OFDM systems," *IEEE Trans. Commun.*, vol. 49, no. 6, pp. 988–999, Jun. 2001, ISSN: 0090-6778. DOI: 10/dp96bq.

[52] H. Liu and U. Tureli, "A high-efficiency carrier estimator for OFDM communications," *IEEE Commun. Lett.*, vol. 2, no. 4, pp. 104–106, Apr. 1998, ISSN: 1089-7798. DOI: 10/bv74rk.

[53] S. Barbarossa, M. Pompili, and G. B. Giannakis, "Channel-independent synchronization of orthogonal frequency division multiple access systems," *IEEE J. Sel. Areas Commun.*, vol. 20, no. 2, pp. 474–486, Feb. 2002, ISSN: 0733-8716. DOI: 10/dq4z7x.

[54] D. K. Kim, S. H. Do, H. B. Cho, H. J. Chol, and K. B. Kim, "A new joint algorithm of symbol timing recovery and sampling clock adjustment for OFDM systems," *IEEE Trans. Consum. Electron.*, vol. 44, no. 3, pp. 1142–1149, Aug. 1998, ISSN: 0098-3063. DOI: 10/d5tnvj.

[55] M. Morelli, A. N. D'Andrea, and U. Mengali, "Feedback frequency synchronization for OFDM applications," *IEEE Commun. Lett.*, vol. 5, no. 1, pp. 28–30, Jan. 2001, ISSN: 1089-7798. DOI: 10.1109/4234.901817.

[56] F. Daffara and O. Adami, "A novel carrier recovery technique for orthogonal multicarrier systems," *Eur. Trans. Telecomm.*, vol. 7, no. 4, pp. 323–334, Jul. 1996, ISSN: 1541-8251. DOI: 10/fkvww4.

[57] N. Lashkarian and S. Kiaei, "Class of cyclic-based estimators for frequency-offset estimation of OFDM systems," *IEEE Trans. Commun.*, vol. 48, no. 12, pp. 2139–2149, Dec. 2000, ISSN: 0090-6778. DOI: 10/cctk6q.

[58] J. Lei and Tung-Sang Ng, "A consistent OFDM carrier frequency offset estimator based on distinctively spaced pilot tones," *IEEE Trans. Wirel. Commun.*, vol. 3, no. 2, pp. 588–599, Mar. 2004, ISSN: 1558-2248. DOI: 10.1109/TWC.2004.825350.

[59] F. Daffara and A. Chouly, "Maximum likelihood frequency detectors for orthogonal multicarrier systems," in *Technical Program, Conference Record, IEEE International Conference on Communications, 1993. ICC '93 Geneva*, vol. 2, May 1993, 766–771 vol.2. DOI: 10/bkq3hm.

[60] H. T. Hsieh and W. R. Wu, "Blind Maximum-Likelihood Carrier-Frequency-Offset Estimation for Interleaved OFDMA Uplink Systems," *IEEE Trans. Veh. Technol.*, vol. 60, no. 1, pp. 160–173, Jan. 2011, ISSN: 0018-9545. DOI: 10/fpzscq.

[61] J. Choi, C. Lee, H. W. Jung, and Y. H. Lee, "Carrier frequency offset compensation for uplink of OFDM-FDMA systems," *IEEE Commun. Lett.*, vol. 4, no. 12, pp. 414–416, Dec. 2000, ISSN: 1089-7798. DOI: 10/c9w58v.

[62] J.-H. Lee and S.-C. Kim, "Detection of Interleaved OFDMA Uplink Signals in the Presence of Residual Frequency Offset Using the SAGE Algorithm," *IEEE Trans. Veh. Technol.*, vol. 56, no. 3, pp. 1455–1460, May 2007, ISSN: 1939-9359. DOI: 10.1109/TVT.2007.895574.

[63] J. Chen, Y.-C. Wu, S. C. Chan, and T.-S. Ng, "Joint Maximum-Likelihood CFO and Channel Estimation for OFDMA Uplink Using Importance Sampling," *IEEE Trans. Veh. Technol.*, vol. 57, no. 6, pp. 3462–3470, Nov. 2008, ISSN: 1939-9359. DOI: 10.1109/TVT.2008.920473.

[64] M. O. Pun, M. Morelli, and C. C. J. Kuo, "Maximum-likelihood synchronization and channel estimation for OFDMA uplink transmissions," *IEEE Trans. Commun.*, vol. 54, no. 4, pp. 726–736, Apr. 2006, ISSN: 0090-6778. DOI: 10/dx3t4g.

[65] Z. Wang, Y. Xin, and G. Mathew, "Iterative carrier-frequency offset estimation for generalized OFDMA uplink transmission," *IEEE Trans. Wirel. Commun.*, vol. 8, no. 3, pp. 1373–1383, Mar. 2009, ISSN: 1558-2248. DOI: 10.1109/TWC.2009.080028.

[66] S. Sezginer and P. Bianchi, "Asymptotically Efficient Reduced-Complexity Frequency Offset Estimation for Uplink MIMO-OFDMA Systems," in *2007 IEEE International Conference on Communications*, Jun. 2007, pp. 2877–2882. DOI: 10.1109/ICC.2007.478.

[67] ——, "Asymptotically Efficient Reduced Complexity Frequency Offset and Channel Estimators for Uplink MIMO-OFDMA Systems," *IEEE Trans. Signal Process.*, vol. 56, no. 3, pp. 964–979, Mar. 2008, ISSN: 1053-587X. DOI: 10/bcfmh2.

[68] Y. Zeng and A. R. Leyman, "Pilot-Based Simplified ML and Fast Algorithm for Frequency Offset Estimation in OFDMA Uplink," *IEEE Trans. Veh. Technol.*, vol. 57, no. 3, pp. 1723–1732, May 2008, ISSN: 0018-9545. DOI: 10/bcj23g.

[69] A. Tonello and S. Pupolin, "Performance of single user detectors in multitone multiple access asynchronous communications," in *Vehicular Technology Conference. IEEE 55th Vehicular Technology Conference. VTC Spring 2002 (Cat. No.02CH37367)*, vol. 1, May 2002, 199–203 vol.1. DOI: 10.1109/VTC.2002.1002692.

[70] D. Huang and K. B. Letaief, "An interference-cancellation scheme for carrier frequency offsets correction in OFDMA systems," *IEEE Trans. Commun.*, vol. 53, no. 7, pp. 1155–1165, Jul. 2005, ISSN: 0090-6778. DOI: 10.1109/TCOMM.2005.851558.

[71] Z. Cao, U. Tureli, Yu-Dong Yao, and P. Honan, "Frequency synchronization for generalized OFDMA uplink," in *IEEE Global Telecommunications Conference, 2004. GLOBECOM '04.*, vol. 2, Nov. 2004, 1071–1075 Vol.2. DOI: 10.1109/GLOCOM.2004.1378122.

[72] S. Ahmadi, *Mobile WiMAX: A Systems Approach to Understanding IEEE 802.16m Radio Access Technology*. Academic Press, Dec. 2010, ISBN: 978-0-08-096097-5.

[73] P. Sun, M. Morelli, and L. Zhang, "Carrier Frequency Offset Tracking in the IEEE 802.16e OFDMA Uplink," *IEEE Trans. Wirel. Commun.*, vol. 9, no. 12, pp. 3613–3619, Dec. 2010, ISSN: 1536-1276. DOI: 10/c264px.

[74] T. Jiang, L. Song, and Y. Zhang, *Orthogonal Frequency Division Multiple Access Fundamentals and Applications*. CRC Press, Apr. 2010, ISBN: 978-1-4200-8825-0.

[75] J. Hayes, "Adaptive Feedback Communications," *IEEE Trans. Commun. Technol.*, vol. 16, no. 1, pp. 29–34, Feb. 1968, ISSN: 0018-9332. DOI: 10/dcnftg.

[76] L. Li and A. J. Goldsmith, "Capacity and optimal resource allocation for fading broadcast channels .II. Outage capacity," *IEEE Trans. Inf. Theory*, vol. 47, no. 3, pp. 1103–1127, Mar. 2001, ISSN: 0018-9448. DOI: 10/cwdcm6.

[77] ——, "Capacity and optimal resource allocation for fading broadcast channels .I. Ergodic capacity," *IEEE Trans. Inf. Theory*, vol. 47, no. 3, pp. 1083–1102, Mar. 2001, ISSN: 0018-9448. DOI: 10.1109/18.915665.

[78] R. Knopp and P. A. Humblet, "Information capacity and power control in single-cell multiuser communications," in *ICC '95 Seattle, 'Gateway to Globalization', 1995 IEEE International Conference on Communications, 1995*, vol. 1, Jun. 1995, 331–335 vol.1. DOI: 10/fk8z58.

[79] I. Kim, H. L. Lee, B. Kim, and Y. H. Lee, "On the use of linear programming for dynamic subchannel and bit allocation in multiuser OFDM," in *IEEE Global Telecommunications Conference, 2001. GLOBECOM '01*, vol. 6, 2001, 3648–3652 vol.6. DOI: 10/dhpn5m.

[80] J. Jang and K. B. Lee, "Transmit power adaptation for multiuser OFDM systems," *IEEE J. Sel. Areas Commun.*, vol. 21, no. 2, pp. 171–178, Feb. 2003, ISSN: 1558-0008. DOI: 10.1109/JSAC.2002.807348.

[81] P.-H. Huang, Y. Gai, B. Krishnamachari, and A. Sridharan, "Subcarrier Allocation in Multiuser OFDM Systems: Complexity and Approximability," May 2010, pp. 1–6. DOI: 10.1109/WCNC.2010.5506244.

[82] M. Bohge, J. Gross, A. Wolisz, and M. Meyer, "Dynamic resource allocation in OFDM systems: An overview of cross-layer optimization principles and techniques," *IEEE Netw.*, vol. 21, no. 1, pp. 53–59, Jan. 2007, ISSN: 1558-156X. DOI: 10.1109/MNET.2007.314539.

[83] S. P. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, UK ; New York: Cambridge University Press, 2004, ISBN: 978-0-521-83378-3.

[84] Jiho Jang and Kwang Bok Lee, "Transmit power adaptation for multiuser OFDM systems," *IEEE J. Sel. Areas Commun.*, vol. 21, no. 2, pp. 171–178, Feb. 2003, ISSN: 0733-8716. DOI: 10/fr44bw.

[85] H. Yin and H. Liu, "An efficient multiuser loading algorithm for OFDM-based broadband wireless systems," in *IEEE Global Telecommunications Conference, 2000. GLOBECOM '00*, vol. 1, 2000, 103–107 vol.1. DOI: 10/c94sj8.

[86] W. Rhee and J. M. Cioffi, "Increase in capacity of multiuser OFDM system using dynamic subchannel allocation," in *VTC2000-Spring. 2000 IEEE 51st Vehicular Technology Conference Proceedings (Cat. No.00CH37026)*, vol. 2, 2000, 1085–1089 vol.2. DOI: 10/b24fh3.

[87] L. M. C. Hoo, B. Halder, J. Tellado, and J. M. Cioffi, "Multiuser transmit optimization for multicarrier broadcast channels: Asymptotic FDMA capacity region and algorithms," *IEEE Trans. Commun.*, vol. 52, no. 6, pp. 922–930, Jun. 2004, ISSN: 0090-6778. DOI: 10/ftjgh4.

[88] Z. Shen, J. G. Andrews, and B. L. Evans, "Adaptive resource allocation in multiuser OFDM systems with proportional rate constraints," *IEEE Trans. Wirel. Commun.*, vol. 4, no. 6, pp. 2726–2737, Nov. 2005, ISSN: 1536-1276. DOI: 10/bfkhbs.

[89] I. C. Wong, Z. Shen, B. L. Evans, and J. G. Andrews, "A low complexity algorithm for proportional resource allocation in OFDMA systems," in *IEEE Workshop onSignal Processing Systems, 2004. SIPS 2004.*, Oct. 2004, pp. 1–6. DOI: 10/bqrk45.

[90] T. C. H. Alen, A. S. Madhukumar, and F. Chin, "Capacity enhancement of a multi-user OFDM system using dynamic frequency allocation," *IEEE Trans. Broadcast.*, vol. 49, no. 4, pp. 344–353, Dec. 2003, ISSN: 0018-9316. DOI: 10.1109/TBC.2003.819525.

[91] H. Zhu and J. Wang, "Chunk-Based Resource Allocation in OFDMA Systems - Part II: Joint Chunk, Power and Bit Allocation," *IEEE Trans. Commun.*, vol. 60, no. 2, pp. 499–509, Feb. 2012, ISSN: 0090-6778. DOI: 10/dp9tnn.

[92] H. Zhu and J. Wang, "Chunk-Based Resource Allocation in OFDMA Systems—Part II: Joint Chunk, Power and Bit Allocation," *IEEE Trans. Commun.*, vol. 60, no. 2, pp. 499–509, Feb. 2012, ISSN: 1558-0857. DOI: 10.1109/TCOMM.2011.112811.110036.

[93] Z. Han, Z. Ji, and K. J. R. Liu, "Fair multiuser channel allocation for OFDMA networks using Nash bargaining solutions and coalitions," *IEEE Trans. Commun.*, vol. 53, no. 8, pp. 1366–1376, Aug. 2005, ISSN: 0090-6778. DOI: 10/b36dsw.

[94] G. Song and Y. Li, "Cross-layer optimization for OFDM wireless networks-part I: Theoretical framework," *IEEE Trans. Wirel. Commun.*, vol. 4, no. 2, pp. 614–624, Mar. 2005, ISSN: 1536-1276. DOI: 10/b6d5r9.

[95] ——, "Cross-layer optimization for OFDM wireless networks-part II: Algorithm development," *IEEE Trans. Wirel. Commun.*, vol. 4, no. 2, pp. 625–634, Mar. 2005, ISSN: 1536-1276. DOI: 10/ccsb58.

[96] M. Bohge, J. Gross, and A. Wolisz, "The potential of dynamic power and sub-carrier assignments in multi-user OFDM-FDMAa cells," in *GLOBECOM '05. IEEE Global Telecommunications Conference, 2005.*, vol. 5, Nov. 2005, pp. 2932–2936. DOI: 10.1109/GLOCOM.2005.1578295.

[97] K. Kim, Y. Han, and S.-L. Kim, "Joint subcarrier and power allocation in uplink OFDMA systems," *IEEE Commun. Lett.*, vol. 9, no. 6, pp. 526–528, Jun. 2005, ISSN: 1089-7798. DOI: 10/fchs3n.

[98] L. Gao and S. Cui, "Efficient subcarrier, power, and rate allocation with fairness consideration for OFDMA uplink," *IEEE Trans. Wirel. Commun.*, vol. 7, no. 5, pp. 1507–1511, May 2008, ISSN: 1536-1276. DOI: 10/bvgqvg.

[99] C. Y. Ng and C. W. Sung, "Low complexity subcarrier and power allocation for utility maximization in uplink OFDMA systems," *IEEE Trans. Wirel. Commun.*, vol. 7, no. 5, pp. 1667–1675, May 2008, ISSN: 1536-1276. DOI: 10.1109/TWC.2008.060723..

[100] M. Bohge, F. Naghibi, and A. Wolisz, "The use of guard bands to mitigate multiple access interference in the OFDMA uplink," in *International OFDM-Workshop 2008 (InoWo'08)*, Hamburg, Germany, 2008, p. 5.

[101] J. Jiang, V. Sekar, and H. Zhang, "Improving Fairness, Efficiency, and Stability in HTTP-Based Adaptive Video Streaming With Festive," *IEEEACM Trans. Netw.*, vol. 22, no. 1, pp. 326–340, Feb. 2014, ISSN: 1063-6692. DOI: 10/f5s47b.

[102] K. Miller, A.-K. Al-Tamimi, and A. Wolisz, "QoE-Based Low-Delay Live Streaming Using Throughput Predictions," *ACM Trans Multimed. Comput Commun Appl*, vol. 13, no. 1, 4:1–4:24, Oct. 2016, ISSN: 1551-6857. DOI: 10/gcpvcc.

[103] X. Zhu and B. Girod, "Video streaming over wireless networks," in *2007 15th European Signal Processing Conference*, Sep. 2007, pp. 1462–1466.

[104] S. Shakkottai, T. S. Rappaport, and P. C. Karlsson, "Cross-layer design for wireless networks," *IEEE Commun. Mag.*, vol. 41, no. 10, pp. 74–80, Oct. 2003, ISSN: 0163-6804. DOI: 10/d9s5hw.

[105] S. Khan, Y. Peng, E. Steinbach, M. Sgroi, and W. Kellerer, "Application-driven cross-layer optimization for video streaming over wireless networks," *IEEE Commun. Mag.*, vol. 44, no. 1, pp. 122–130, Jan. 2006, ISSN: 1558-1896. DOI: 10.1109/MCOM.2006.1580942.

[106] O. Oyman, J. Foerster, Y. j Tcha, and S. c Lee, "Toward enhanced mobile video services over WiMAX and LTE [WiMAX/LTE Update]," *IEEE Commun. Mag.*, vol. 48, no. 8, pp. 68–76, Aug. 2010, ISSN: 0163-6804. DOI: 10/fp4hmw.

[107] X. Yin, V. Sekar, and B. Sinopoli, "Toward a Principled Framework to Design Dynamic Adaptive Streaming Algorithms over HTTP," in *Proceedings of the 13th ACM Workshop on Hot Topics in Networks*, ser. HotNets-XIII, New York, NY, USA: ACM, 2014, 9:1–9:7, ISBN: 978-1-4503-3256-9. DOI: 10/gcpvb8.

[108] S. Akhshabi, L. Anantakrishnan, A. C. Begen, and C. Dovrolis, "What Happens when HTTP Adaptive Streaming Players Compete for Bandwidth?" In *Proceedings of the 22Nd International Workshop on Network and Operating System Support for Digital Audio and Video*, ser. NOSSDAV '12, New York, NY, USA: ACM, 2012, pp. 9–14, ISBN: 978-1-4503-1430-5. DOI: 10/gcpxhn.

[109] F. Fu and M. V. D. Schaar, "A systematic framework for dynamically optimizing multi-user wireless video transmission," *IEEE J. Sel. Areas Commun.*, vol. 28, no. 3, pp. 308–320, Apr. 2010, ISSN: 0733-8716. DOI: 10/fnnzdw.

[110] V. Joseph, S. Borst, and M. I. Reiman, "Optimal rate allocation for adaptive wireless video streaming in networks with user dynamics," in *IEEE INFOCOM 2014 - IEEE Conference on Computer Communications*, Apr. 2014, pp. 406–414. DOI: 10.1109/INFOCOM.2014.6847963.

[111] A. E. Essaili, D. Schroeder, E. Steinbach, D. Staehle, and M. Shehada, "QoE-Based Traffic and Resource Management for Adaptive HTTP Video Delivery in LTE," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 25, no. 6, pp. 988–1001, Jun. 2015, ISSN: 1051-8215. DOI: 10/gcpvcj.

[112] L. He and G. Liu, "Optimal cross layer design for video transmission over OFDMA system," in *2012 IEEE International Conference on Communications (ICC)*, Jun. 2012, pp. 1154–1159. DOI: 10/gcpvb4.

[113] A. A. Khalek, C. Caramanis, and R. W. Heath, "A Cross-Layer Design for Perceptual Optimization Of H.264/SVC with Unequal Error Protection," *IEEE J. Sel. Areas Commun.*, vol. 30, no. 7, pp. 1157–1171, Aug. 2012, ISSN: 0733-8716. DOI: 10/gcp2jm.

[114] M. Zhao, X. Gong, J. Liang, W. Wang, X. Que, and S. Cheng, "QoE-Driven Cross-Layer Optimization for Wireless Dynamic Adaptive Streaming of Scalable Videos Over HTTP," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 25, no. 3, pp. 451–465, Mar. 2015, ISSN: 1051-8215. DOI: 10/gcpz7z.

[115] S. Cicalò and V. Tralli, "Distortion-Fair Cross-Layer Resource Allocation for Scalable Video Transmission in OFDMA Wireless Networks," *IEEE Trans. Multimed.*, vol. 16, no. 3, pp. 848–863, Apr. 2014, ISSN: 1520-9210. DOI: 10/f5vv7f.

[116] D. Wang, L. Toni, P. C. Cosman, and L. B. Milstein, "Uplink Resource Management for Multiuser OFDM Video Transmission Systems: Analysis and Algorithm Design," *IEEE Trans. Commun.*, vol. 61, no. 5, pp. 2060–2073, May 2013, ISSN: 0090-6778. DOI: 10/gcpvb7.

[117] H. Le, D. Willkomm, and A. Wolisz, "Optimizing user throughput with the consideration of multiple access interference in the OFDMA uplink," in *2013 9th International Wireless Communications and Mobile Computing Conference (IWCMC)*, Jul. 2013, pp. 889–894. DOI: 10.1109/IWCMC.2013.6583675.

[118] H. Le, A. Behboodi, and A. Wolisz, "Dynamic Resource Allocation in OFDMA Uplink for MAI Mitigation and Throughput Improvement," in *2014 IEEE 80th Vehicular Technology Conference (VTC2014-Fall)*, Sep. 2014, pp. 1–5. DOI: 10.1109/VTCFall.2014.6965949.

[119] D. J. Love and R. W. Heath, "OFDM power loading using limited feedback," *IEEE Trans. Veh. Technol.*, vol. 54, no. 5, pp. 1773–1780, Sep. 2005, ISSN: 0018-9545. DOI: 10/dbc4xf.

[120] H. Nguyen, J. Brouet, V. Kumar, and T. Lestable, "Compression of associated signaling for adaptive multi-carrier systems," in *2004 IEEE 59th Vehicular Technology Conference. VTC 2004-Spring (IEEE Cat. No.04CH37514)*, vol. 4, May 2004, 1916–1919 Vol.4. DOI: 10.1109/VETECS.2004.1390607.

[121] P. Sure and C. M. Bhuma, "A survey on OFDM channel estimation techniques based on denoising strategies," *Engineering Science and Technology, an International Journal*, vol. 20, no. 2, pp. 629–636, Apr. 2017, ISSN: 2215-0986. DOI: 10.1016/j.jestch.2016.09.011.

[122] M. A. Maddah-Ali and D. Tse, "Completely Stale Transmitter Channel State Information is Still Very Useful," *IEEE Trans. Inf. Theory*, vol. 58, no. 7, pp. 4418–4431, Jul. 2012, ISSN: 0018-9448. DOI: 10/f328d6.

[123] S. Gifford, C. Bergstrom, and S. Chuprun, "Adaptive and linear prediction channel tracking algorithms for mobile OFDM-MIMO applications," in *MILCOM 2005 - 2005 IEEE Military Communications Conference*, Oct. 2005, 1298–1302 Vol. 2. DOI: 10.1109/MILCOM.2005.1605857.

[124] M. Ergen, S. Coleri, and P. Varaiya, "QoS aware adaptive resource allocation techniques for fair scheduling in OFDMA based broadband wireless access systems," *IEEE Trans. Broadcast.*, vol. 49, no. 4, pp. 362–370, Dec. 2003, ISSN: 0018-9316. DOI: 10/fr35z4.

[125] J. Lee and S. Leyffer, Eds., *Mixed Integer Nonlinear Programming*, ser. The IMA Volumes in Mathematics and Its Applications. New York: Springer-Verlag, 2012, ISBN: 978-1-4614-1926-6.

[126] A. Saxena, P. Bonami, and J. Lee, "Convex relaxations of non-convex mixed integer quadratically constrained programs: Extended formulations," *Math. Program.*, vol. 124, no. 1-2, pp. 383–411, Jul. 2010, ISSN: 0025-5610, 1436-4646. DOI: 10.1007/s10107-010-0371-9.

[127] ——, "Convex relaxations of non-convex mixed integer quadratically constrained programs: Projected formulations," *Math. Program.*, vol. 130, no. 2, pp. 359–413, Dec. 2011, ISSN: 0025-5610, 1436-4646. DOI: 10.1007/s10107-010-0340-3.

[128] IEEE, *802.16m-2011 - IEEE Standard for Local and metropolitan area networks Part 16: Air Interface for Broadband Wireless Access Systems Amendment 3: Advanced Air Interface*, May 2011.

[129] W. Forum, "WiMAX System Evaluation Methodology," Tech. Rep., Jul. 2008.

[130] H. Le, K. Miller, A. Behboodi, and A. Wolisz, "Cross layer approach for HTTP-based low-delay adaptive streaming in mobile networks," in *2017 IEEE 18th International Symposium on A World of Wireless, Mobile and Multimedia Networks (WoWMoM)*, Jun. 2017, pp. 1–9. DOI: 10.1109/WoWMoM.2017.7974322.

[131] K. Seshadrinathan and A. C. Bovik, "Temporal hysteresis model of time varying subjective video quality," IEEE, May 2011, pp. 1153–1156, ISBN: 978-1-4577-0538-0. DOI: 10.1109/ICASSP.2011.5946613.

[132] P. Seeling and M. Reisslein, "Video Transport Evaluation With H.264 Video Traces," *IEEE Commun. Surv. Tutor.*, vol. 14, no. 4, pp. 1142–1165, Fourth 2012, ISSN: 1553-877X. DOI: 10.1109/SURV.2011.082911.00067.

[133] D. I. Forum, "Guidelines for Implementation: DASH-IF Interoperability Points (Version 4.3)," Tech. Rep., Nov. 2018.

[134] D. Bethanabhotla, G. Caire, and M. J. Neely, "Adaptive Video Streaming for Wireless Networks With Multiple Users and Helpers," *IEEE Trans. Commun.*, vol. 63, no. 1, pp. 268–285, Jan. 2015, ISSN: 0090-6778. DOI: 10/gcpvcf.

[135] J.-N. Hwang, "Multimedia Networking: From Theory to Practice," 2009. DOI: 10.1017/CBO9780511626654.

[136] A. Seetharam, P. Dutta, V. Arya, J. Kurose, M. Chetlur, and S. Kalyanaraman, "On Managing Quality of Experience of Multiple Video Streams in Wireless Networks," *IEEE Trans. Mob. Comput.*, vol. 14, no. 3, pp. 619–631, Mar. 2015, ISSN: 1536-1233. DOI: 10/gcpvch.

[137] V. Joseph and G. de Veciana, "NOVA: QoE-driven optimization of DASH-based video delivery in networks," in *IEEE INFOCOM 2014 - IEEE Conference on Computer Communications*, Apr. 2014, pp. 82–90. DOI: 10/gcpvb5.

[138] E. Yaacoub and Z. Dawy, "A Survey on Uplink Resource Allocation in OFDMA Wireless Networks," *IEEE Commun. Surv. Tutor.*, vol. 14, no. 2, pp. 322–337, Second 2012, ISSN: 1553-877X. DOI: 10.1109/SURV.2011.051111.00121.

[139] O. Oyman, R. Nabar, H. Bolcskei, and A. Paulraj, "Tight lower bounds on the ergodic capacity of Rayleigh fading MIMO channels," in *Global Telecommunications Conference*, vol. 2, Dec. 2002, 1172–1176 vol.2, ISBN: 978-0-7803-7632-8. DOI: 10.1109/GLOCOM.2002.1188380.

[140] D. Kitchener, W. Tong, M. Naden, and Z. Peiying, "Correlated Lognormal Shadowing Model (Technical Report)," IEEE, Tech. Rep. IEEE C802.16j-06/059, 2006.

[141] A. Duel-Hallen, S. Hu, and H. Hallen, "Long Range Prediction of Fading Signals: Enabling Adaptive Transmission for Mobile Radio Channels," *IEEE Signal Process. Mag.*, vol. 17, no. 3, pp. 62–75, May 2000.

[142] K. Börner, J. Dommel, S. Jaeckel, and L. Thiele, "On the requirements for quasi-deterministic radio channel models for heterogeneous networks," in *2012 International Symposium on Signals, Systems, and Electronics (ISSSE)*, Oct. 2012, pp. 1–5. DOI: 10.1109/ISSSE.2012.6374332.

[143] H. Le, A. Behboodi, and A. Wolisz, "Quality driven resource allocation for adaptive video streaming in OFDMA uplink," in *2015 IEEE 26th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, Aug. 2015, pp. 1277–1282. DOI: 10.1109/PIMRC.2015.7343495.

[144] K. Stuhlmuller, N. Farber, M. Link, and B. Girod, "Analysis of video transmission over lossy channels," *IEEE J. Sel. Areas Commun.*, vol. 18, no. 6, pp. 1012–1032, Jun. 2000, ISSN: 1558-0008. DOI: 10.1109/49.848253.

[145] H. Mansour, V. Krishnamurthy, and P. Nasiopoulos, "Channel Aware Multiuser Scalable Video Streaming Over Lossy Under-Provisioned Channels: Modeling and Analysis," *IEEE Trans. Multimed.*, vol. 10, no. 7, pp. 1366–1381, Nov. 2008, ISSN: 1941-0077. DOI: 10.1109/TMM.2008.2004915.

[146] R. Deng and G. Liu, "QoE driven cross-layer scheme for DASH-based scalable video transmission over LTE," *Multimed Tools Appl*, pp. 1–25, Apr. 2017, ISSN: 1380-7501, 1573-7721. DOI: 10/gcpvcs.

[147] D. Wang, P. C. Cosman, and L. B. Milstein, "Cross Layer Resource Allocation Design for Uplink Video OFDMA Wireless Systems," in *2011 IEEE Global Telecommunications Conference - GLOBECOM 2011*, Dec. 2011, pp. 1–6. DOI: 10.1109/GLOCOM.2011.6134147.