Dynamic Algorithms in Multi-User OFDM Wireless Cells

vorgelegt von
Diplom-Ingenieur
James Richard Gross
aus Berlin

von 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. Bernd Mahr
Berichter: Prof. Dr.-Ing. Adam Wolisz
Berichter: Prof. Dr. Martin Grötschel
Berichter: Prof. Dr.-Ing. Karl Dirk Kammeyer
Tag der wissenschaftlichen Aussprache: 15. Juni 2006

Berlin 2006
D83

Alles Wissen und alle Vermehrung unseres Wissens
endet nicht mit einem Schlußpunkt, sondern mit Fragezeichen.

Hermann Hesse

Abstract

This thesis addresses optimization issues in dynamic multi-user OFDM systems. The investigations are focused on the down-link of a single wireless cell. Dynamic sub-carrier assignments to terminals are periodically generated by the access point based on channel predictions. Several issues are considered in this context.

First, the thesis investigates optimal and suboptimal assignment strategies for a particular optimization problem in multi-user OFDM, referred to as rate adaptive problem. Jointly assigning power and sub-carriers to terminals yields the optimal performance. However, this strategy is found to have a high computational complexity. It is proposed in the literature to only assign the sub-carriers dynamically while keeping the power per sub-carrier fixed. As the thesis shows, this setup performs almost equally well in small cells, while the joint optimal assignment pays off especially in large cells. However, this second approach does not reduce the computational complexity of the problem: As the thesis shows, even the (optimal) dynamic assignment of sub-carriers only is NP-complete. Hence, in practical systems suboptimal assignment algorithms will have to be applied. The thesis presents a suboptimal scheme for the joint rate adaptive assignment of sub-carriers and transmit power, based on linear programming. This scheme substantially outperforms previously proposed suboptimal algorithms and provides a performance close to the optimal one. In contrast to the optimal solution, it can always be computed in polynomial time. Finally, a modification is presented which enables the application of the suboptimal schemes also to variable rate packet streams. This modification is shown to perform well for an example scenario where terminals in the cell receive MPEG-4 encoded video streams.

Second, the thesis addresses the issue of signaling in dynamic multi-user OFDM systems. As the assignments are generated at the access point and the terminals have to be informed of their new assignments prior to the payload transmission, a signaling scheme is required. This is found to decrease the performance of the system, in particular as a function of the number of sub-carriers and the frame duration. Furthermore, the thesis presents and evaluates two approaches which can reduce the signaling overhead for systems whose performance would suffer from the high cost of signaling: Compression and optimization. Both approaches exploit the correlation of assignments in time and frequency. It is shown that both schemes decrease the overhead due to signaling significantly, extending the parameter range for which the signaling overhead is still acceptable. However, the compression approach is recommended as the complexity of the optimization scheme is high.

Finally, the thesis outlines a new system concept of the IEEE 802.11a standard which employs a dynamic multi-user OFDM algorithm. The proposal is compliant with legacy IEEE 802.11a systems and includes a protocol scheme for retrieving the channel knowledge from all terminals as well as distributing the signaling information. The new system concept is found to provide a significant performance gain, especially if multiple stations can be served in parallel by the multi-user approach.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Future wireless communication systems are widely expected to provide higher data rates to users. This can be observed from the evolution of cellular communication systems as well as from the evolution of wireless local area networks. In both cases, the recent standard is considerably "faster" regarding its transmission rate than previous versions. However, the wireless channel is quite a different medium compared to wired links such as copper wires or optical communication links. While the attenuation of wired links is more or less constant in time, it varies rapidly (in time) for wireless channels, known as *fading*. In addition, the wireless communication channel is well known to suffer from an effect similar to an "echo": Several copies of the same signal, representing a certain bit combination, interfere at the receiver as they travel along different paths. This "echo" distorts the received signal, as some copies arrive slightly earlier than other ones. The higher the data rate of such a wireless channel is, the more severe is the distortion due to the "echo". This limits the amount of data that can be transmitted reliably over a wireless channel by traditional transmission schemes. Hence, it is difficult to increase the data rate of wireless communication systems.

This has lead to the development of several alternative transmission schemes, which are not similarly affected by the "echo"of the channel. One of these transmission schemes is **O**rthogonal **F**requency **D**ivision **M**ultiplexing (*OFDM*) . In OFDM, the given system bandwidth is split into many small sub-channels. Then, data is communicated in parallel on these sub-channels, rather than conveying it sequentially. The key point is that the transmission rate per single sub-channel is low. Thus, the "echo" of the channel does not distort the signal that much. However, the overall transmission rate is still high. Due to this characteristic, the OFDM transmission scheme has become quite popular recently: It is applied already in digital broadcasting systems and wireless local area networks. In addition, several upcoming standards consider using OFDM as the underlying transmission scheme.

In particular, *dynamic schemes* for wireless OFDM systems have been discussed recently. In most OFDM systems, the attenuation of different sub-channels varies (when considering a point-to-point connection between two OFDM transceivers). Out of the set of sub-channels, some will always be in a good state (having a relatively low attenuation) while others might be in a bad state (having a relatively high attenuation). Therefore, a superior transmission policy consists of grouping the transmit power and the bits to be sent on the sub-channels in a relatively

good state while avoiding sub-channels in a bad state. It has been shown that such dynamic schemes significantly increase the performance of OFDM point-to-point connections. However, such a dynamic policy requires the transmitter to have knowledge of the channel states which calls for some additional mechanisms.

If multiple terminals are present within the communication range, another degree of diversity regarding the attenuation values of the sub-channels appears. It is well known that the attenuation component due to fading of a single wireless channel is statistically independent at different locations. If multiple terminals are located at different places, this leads to a multiuser diversity of the sub-channel attenuations. In such a point-to-multi-point communication setting, a dynamic scheme can assign different sets of sub-carriers to different terminals, in addition to distributing the power and modulation types dynamically. Such dynamic multi-user OFDM systems have been proposed for the first time about five years ago. It has been shown, that – theoretically – the system performance (in terms of the bit rate) can be doubled, compared to static approaches.

Despite the theoretical advantages of such dynamic schemes for OFDM systems, no such dynamic multi-user system has been implemented so far. Two issues in particular have impeded the realization of such systems: First of all, the assignments have to be generated at some node or terminal in the network, which needs to have sufficient computational resources for this task. For example, this could be the access point. The problem is that the assignments have to be generated *fast* compared to the time variations of the wireless channel. Otherwise, the dynamic assignments are of no use and might even lead to a significant performance loss.

Furthermore, after the assignments have been generated at the transmitter, the receivers have to be informed of their current assignments in order to decode the payload data. This requires the transmission of management data on some control channel, referred to as signaling information. It is not known how large this signaling overhead is and whether this overhead is larger than the performance gain which dynamic schemes can achieve.

In this thesis, multiple contributions to the field of dynamic multi-user OFDM systems are presented. Chapter 4 and Chapter 5 address the two issues (fast assignment algorithms and signaling overhead) outlined above; they form the main part of the thesis. Chapter 6 then draws on the results of these chapters to present a proposal for a new protocol standard within the family of IEEE 802.11 wireless local area networks.

The rest of the thesis is structured as follows: After providing an overview of related background issues (Chapter 2), the state of the art regarding dynamic schemes for OFDM systems is presented (Chapter 3). The scope of the thesis is discussed in detail in Section 3.4.

Then the issue of the complexity and performance of assignment algorithms is investigated in Chapter 4. Theoretically, the optimal performance is achieved by dynamically distributing sub-carriers and transmit power to different terminals. However, the computational complexity of this optimization approach is not known. This issue is investigated in detail in Chapter 4. As it turns out, the jointly optimal assignment of sub-carriers and power is too complex for application in practical systems. Hence, suboptimal algorithms are of interest. Chapter 4 presents a very good suboptimal scheme for the joint dynamic assignment of sub-carriers and power. However, due to various constraints, the dynamic power assignment might not be possible for future dynamic OFDM systems. In this case, only sub-carriers are assigned dynamically (in

an optimal way). Interestingly, this optimization problem is found to be too complex as well. Therefore, suboptimal algorithms have to be applied in this special case, too. Such suboptimal algorithms are also evaluated in Chapter 4. Finally, a modification is introduced which enables the scheduling of variable rate packet streams to sub-carriers, using the suboptimal schemes presented before. Considering an example scenario where terminals receive MPEG-4 encoded video streams, this modfication is shown to work well with an suboptimal scheme in a centralized system approach.

After dealing with algorithms for the dynamic multi-user OFDM approach, the issue of signaling is investigated in depth in Chapter 5. Regarding the issue of signaling, nothing is known so far. The chapter first discusses a rather simple model, allowing a quantification of the impact due to signaling. It turns out that the signaling information reduces the performance of dynamic multi-user OFDM systems significantly. In particular, a few parameters are identified which have a strong impact on the overhead. Thus, if these parameters cause a high overhead of a specific OFDM system, the question arises if there exist reduction schemes for the signaling overhead. Two such schemes (optimization as well as compression) are proposed and finally evaluated in Chapter 5.

Finally, Chapter 6 proposes an outline for a new extension within the family of IEEE 802.11 wireless local area networks. Based on the results of Chapter 4 and Chapter 5, this proposal enables the access point to apply dynamic multi-user OFDM burst transmissions in order to increase the throughput of the system. As the chapter shows, the system performance is always increased if several stations can be served by the access point in parallel, even if the MPDU sizes of the stations are small. Also, the new concept features a mode for the up-link direction, i.e. from a station to the access point. The proposal is compliant with legacy IEEE 802.11a systems. Protocol mechanisms are developed which allow the acquisition of the channel knowledge as well as the distribution of the signaling information.

Finally, in Chapter 7 conclusions are drawn as well as issues for future work are discussed.

Chapter 2

Background Issues

In this chapter multiple background issues are discussed, providing theoretical foundations of issues raised afterwards in this thesis. First, the two issues of wireless channel behavior (Section 2.1) and **O**rthogonal **F**requency **D**ivision **M**ultiplexing (*OFDM*) transmission (Section 2.2) are covered. It is assumed that the reader has some knowledge about random processes in general (specifically, about the first-order and second-order moments of random processes, refer to [1]) and that the reader is familiar with the basic concepts of digital data transmission (specifically, about baseband signal representation and modulation schemes, to some extent covered by [2]). Thereafter, an introduction to certain issues regarding information theory are given in Section 2.3. Finally, in Section 2.4 some basic information on mathematical optimization is summarized.

2.1 Communications over Wireless Channels

For any communication system, a metric referred to as **S**ignal-to-**N**oise-and-**I**nterference **R**atio (*SNIR*) is of fundamental importance. The SNIR is defined as the ratio between the received power of the data signal of interest divided by the sum of all "disturbing" sources of power, i.e. *interference* and *noise*. From information theory, it is known that a higher SNIR increases the maximum possible error-free transmission rate (referred to as Shannon capacity [3]) of any communication system and vice versa[1]. Correspondingly, the higher the SNIR, the more decreased the bit error rate in practical systems.

While one aspect of the SNIR is the sum of all distracting power sources, another issue is the received power. This depends on the transmit power and the attenuation between transmitter and receiver. Hence, given a fixed transmit power, three factors influence the SNIR: The attenuation between transmitter and receiver, the received interference power and the noise power. All three factors are in general time-varying and are denoted in the following by $h(t)$ for the attenuation, $j(t)^2$ for the interference power, and $n(t)^2$ for the noise power. Given a transmit power of P_{tx}, the SNIR is defined as:

[1] Originally, this result is based on the **S**ignal–to–**N**oise **R**atio (*SNR*) and not on the SNIR.

$$\text{SNIR} := \frac{P_{\text{tx}} \cdot h(t)^2}{j(t)^2 + n(t)^2} \tag{2.1}$$

The attenuation depends strongly on the medium utilized. This medium, for example an optical fiber or an infrared link, is referred to as *communication channel*. There exists a large variety of channels, which may be divided into two groups, *wired* and *wireless* ones. Wireless channels differ a lot from wired channels due to their highly random attenuation. In contrast to wired channels, the attenuation between two transceivers may change randomly by magnitudes within a couple of milliseconds, for example. Therefore, reliable communications over wireless channels is quite challenging.

Within the group of wireless channels, different categories can be found. For example, channels can be classified by the structural environment surrounding the transceivers. Many different such *propagation environments* have been identified such as urban, suburban, indoor, underwater or orbital propagation environments, which differ in various ways.

In this section, common models are presented which capture essential aspects of the attenuation, the noise and the interference encountered in wireless communications. Most attention is spent on the attenuation, as a proper model of its statistical characteristics is quite important afterwards in this thesis. However, a compact discussion of noise and interference models is also provided.

2.1.1 Attenuation Variations

Measuring the attenuation of a wireless transmit signal at a certain frequency over some time reveals a highly random behavior [4]. However, the results of extensive measurement campaigns [5] allow to distinguish three different components contributing to the overall attenuation. First of all, if the *distance* between transmitter and receiver is kept constant and multiple samples of the attenuation are recorded for different placements, these samples vary significantly. However, the average attenuation over all placements and over time depends on the distance alone. Models have been developed which reliably predict this average attenuation depending on the particular environment of transmitter and receiver [6]. This component of the attenuation is termed *path loss*.

For a given placement of transmitter and receiver, these path loss models predict a certain attenuation. However, the measured attenuation varies constantly over time on a time-scale of about milliseconds. If the long-term (over multiple seconds) average of the attenuation is taken, it is often found to differ from the predicted attenuation[2] [5]. These deviations have been found to be of stochastic nature and are termed *shadowing*. As with the path loss, shadowing models have to be parameterized according to the particular environment of transmitter and receiver.

Finally, the short-term variations (in the range of milliseconds) of the attenuation are found to follow certain first- and second-order statistics, depending on various environment parameters such as a possible line-of-sight between transmitter and receiver, the center frequency etc. These

[2]Recall that the path loss model only predicts an average attenuation for all placements with a certain distance, not for a particular placement.

short-term fluctuations are termed *fading* and they are modeled by an appropriately parameterized stochastic process in general.

Combining all three effects yields the overall attenuation of the wireless channel. The corresponding model, which is used commonly [4], is given in Equation 2.2.

$$h = h_{\mathrm{pl}} \cdot h_{\mathrm{sh}} \cdot h_{\mathrm{fad}} \tag{2.2}$$

Especially for wireless channels the attenuation usually fluctuates strongly over time. These variations over time result basically from changes within the propagation environment of transmitter and receiver. First of all, the propagation environment of the transmitter and receiver might change due to objects moving around, doors or windows being opened and closed etc. This might affect the shadowing component and/or the fading component, such that the resulting attenuation changes over time. Alternatively (or in combination with moving objects) the propagation environment might change due to mobility of the transmitter and/or the receiver. This usually results in a change of the path loss, the shadowing and the fading component. However, it is important to note that the fading component is affected on a much smaller time scale than shadowing and path loss. Note that in addition to the time-selective nature of the wireless channel, there might also be a frequency-selective behavior of the attenuation, mainly due to the fading component, as discussed below in detail.

Path Loss

It is well known that the average power of an electromagnetic wave decreases quadratically for an increasing distance between transmitter and receiver in free space [7]. However, many different studies have found that the received average power decreases by an exponent significantly larger than 2 for an increasing distance [7, 5]. This difference can be explained for example by the *two-ray model*. It provides a prediction of the attenuation due to path loss, assuming that two electromagnetic rays interfere constructively and destructively with each other as the distance increases. One ray reaches the receiver by a **L**ine-**o**f-**S**ight (*LOS*) while the second ray is reflected by the ground. Thus, this model only matches transmission scenarios with a LOS and no other objects surrounding the path. These assumptions are not valid in almost all practical environments, where a **n**on **L**ine-**o**f-**S**ight (*NLOS*) is quite common and the number of paths is much higher than two. Alternatively much more paths can be taken into consideration by ray-tracing methods. These enable the calculation of the attenuation for every possible path between transmitter and receiver and finally add up all these path components to a resulting attenuation. Unfortunately, this method requires exact data about the terrain, the buildings and the vegetation and is very demanding in terms of computational capacity.

For these reasons, empirical models [5] have been developed to calculate the path loss between a transmitter and a receiver, separated by a certain distance from each other in specific environments for different frequencies. These models are usually of the form:

$$h_{\mathrm{pl}} = \frac{P_0}{P_{\mathrm{tx}}} = K \cdot \frac{1}{d^{\alpha}} \tag{2.3}$$

or, in dB,

$$h_{\mathrm{pl}}[\mathrm{dB}] = \frac{P_0}{P_t}[\mathrm{dB}] = 10\,\log(K) - 10\,\cdot\alpha\cdot\log(d) \quad ,$$

where the constants K and α are fitted to measurement results. The factor K is the loss over one unit of a reference distance. The distance d between transmitter and receiver is in units of the reference distance, and has to be defined along with the exponent. Note that there is no difference between LOS and NLOS settings anymore, since the models are obtained by averaging measurements under both conditions, and the different propagation effects are approximatively condensed in a single parameter - the path loss exponent α.

Popular models of this kind are the Okumura-Hata model [8] and the Lee model [9]. These models are wide spread over the research community and are also used in this thesis. An overview of other path loss models can be found in [10].

Shadowing

The path loss models presented above provide an average attenuation which can be expected at a certain distance between transmitter and receiver. However, it is known from measurements [7, 5] that for a fixed distance the average attenuation (over time) varies significantly for several different positions of the receiver, as objects such as buildings or trees might obstruct the transmission paths. These stochastic, location-dependent variations are referred to as *shadowing*. Shadowing is modeled as a stochastic process with an average of 1, as the average over many different location dependent variations yields exactly the predicted value of the path loss model. Note that shadowing is an abstraction which represents the result of a sum of several propagation phenomena which occur when an electromagnetic wave propagates in an environment. As shadowing is due to the obstruction by larger objects, it has a high correlation in space. For example, if the distance between transmitter and receiver is kept constant and the receiver moves along a circle around the transmitter, the attenuation due to shadowing stays constant over several meters, as a large building "blocks" the area around a certain position. Therefore, shadowing varies rather over "large" time scales such as seconds if the transmitter and receiver are mobile.

From measurements of the attenuation for a variety of environments and distances, the variation of the measured signal level relative to the average predicted path loss can be calculated (see Figures 2.37 to 2.41 from [5]). Its distribution is normal with 0 mean (in dB), which implies a log-normal distribution of the received power around the mean value corresponding to the path loss. This hypothesis has been verified with the χ^2 and Kolmogrov-Smirnov test with a high confidence interval. Thus, the shadowing process is modeled by a first-order distribution given by

$$p\left(h_{\mathrm{sh}}\right) = \frac{1}{\sqrt{2\cdot\pi\cdot\sigma_{\mathrm{sh}}^2}}\cdot e^{-\frac{h_{\mathrm{sh}}^2}{2\sigma_{\mathrm{sh}}^2}} \quad , \tag{2.4}$$

where σ_{sh}^2 is the variation and all variables are expressed in dB.

The standard deviation $\sqrt{\sigma_{\mathrm{sh}}^2}$ of the shadowing process has been found to take values between 5 dB and 12 dB [11, 9, 12], depending on the considered environment and system.

Fading

Fading is the interference of many scattered signals arriving at an antenna [4]. It is responsible for the most rapid and violent changes of the signal strength itself as well as its phase. These signal variations are experienced on a small time scale, mostly a fraction of a second or shorter, depending on the velocity of the receiver. The following discussion is based on [4].

The physical basis of fading is given by the reception of multiple copies of the transmitted signal, each one stemming from a different propagation path. Depending on the environment of transmitter and receiver, there can be many or only few objects reflecting the transmitted radio signal. In general these objects lead to a situation shown in Figure 2.1, which is called a *multi-path signal propagation environment*.

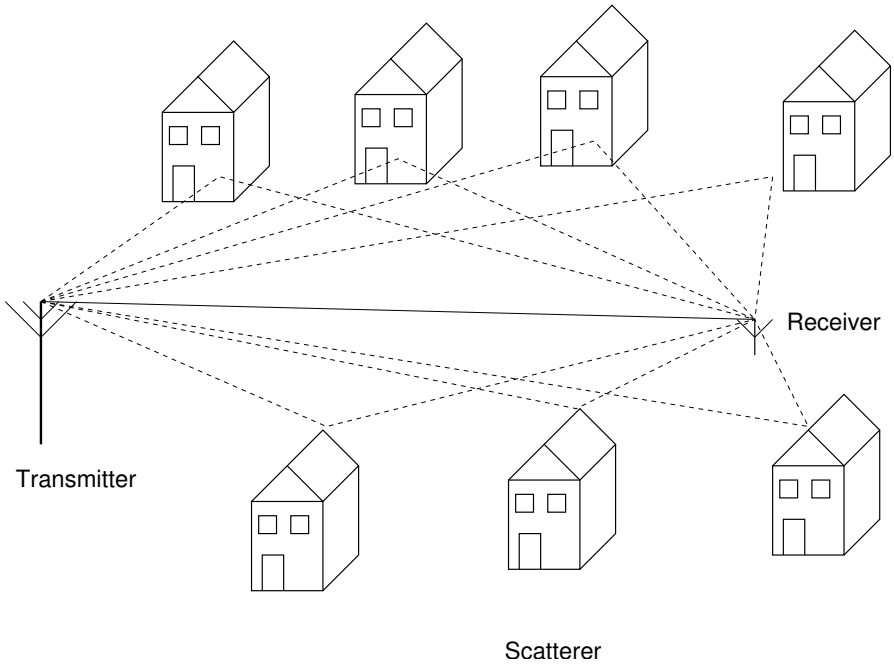


Figure 2.1: Multi-path propagation scenario.

In such a typical environment, each path i has a different length l_i. Due to this difference, each signal traveling along a path arrives with a different *delay* $\tau_i = \frac{l_i}{c}$, where c is the speed of light. Some signal copies traveling along short paths will arrive faster than other copies traveling along longer paths. Physically, this situation is comparable to an acoustic echo. In communications, the channel is said to have a memory, since it is able to store signal copies for a certain time span. The difference between "earliest" and "latest" received signal copy is often referred to as *delay spread* $\Delta\sigma$. Apart from the delay spread, each signal copy is

attenuated differently, since the signal paths have to pass different obstacles like windows, walls of different materials, trees of different sizes and so on. Also, each signal traveling along its path might reach the receiver by a different angle. If the receiver or the transmitter is moving, this leads to a Doppler shift of all signals, according to their angle of arrival. This results in a *Doppler spread* Δf_d.

Taking all this into account, the multi-path propagation of a transmitted radio wave results in a specific interference pattern for each propagation environment, where at certain places the waves interfere constructively while at other places they interfere destructively. If all elements within the propagation environment (transmitter, receiver etc.) do not move, the receiving signal will only by distorted by the *delay spread* and the varying attenuation per path. In this case, the interference situation of the channel stays constant over time and therefore the channel is said to be *time invariant*. In contrast, if any kind of movement is encountered in the propagation environment, all or some paths change in time. As a consequence the wireless channel becomes *time variant* (see Figure 2.2). Correspondingly, if the delay spread $\Delta\sigma$ of the channel is zero, the channel is said to be *frequency invariant*. Otherwise, the channel's attenuation varies for different frequencies and therefore the channel is said to be *frequency variant* (see Figure 2.2). In contrast to the Doppler spread, the delay spread is almost always non zero. Thus, almost always a wireless channel is frequency variant.

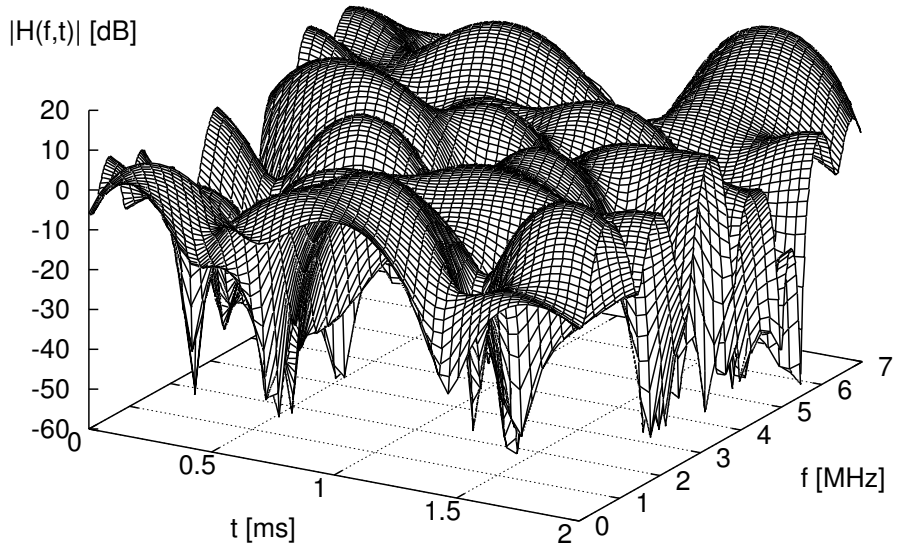


Figure 2.2: Time and frequency variant attenuation due to fading of a broadband wireless channel. Depending on the observed bandwidth and duration, the channel might be time and frequency selective or not.

The frequency or time variant behavior of a wireless channel do not have to be harmful as such. Depending on the severity of the channel's variance compared to fundamental transmission parameters, channels are classified. The severity of the time variant behavior caused by the Doppler spread depends on the time span the receiver processes the incoming communication

signal. If coherent detection[3] is assumed, this processing time is the symbol length T_s. In general, $n \cdot T_\mathrm{s}$ represents the processing time span (as differential detection or an equalization process cause $n > 1$). If the fade rate of the time selective process, given by the Doppler spread Δf_d, is larger than the processing rate (given by $\frac{1}{n \cdot T_\mathrm{s}}$), then the *fading* is called *time selective* [4]. In contrast, if the fade rate is much lower than the processing rate, therefore if $\Delta f_d \cdot n \cdot T_\mathrm{s} \ll 1$, then the *fading* is called *not time selective*.

Correspondingly, the severity of the frequency variant behavior can be estimated by the product of the required baseband bandwidth of the signal (denoted by B) and the delay spread. If the delay spread is very small compared to the reciprocal of the bandwidth[4], then it has almost no impact on the reception of the signal ($\Delta\sigma \cdot B \ll 1$). In this case the transfer function (attenuation function) of the channel has no variations within the signal's bandwidth. The *fading* is called to be *flat or frequency non selective*. On the other hand, if the delay spread is significant compared to the reciprocal of the bandwidth, then the channel has a *frequency selective* behavior. That is, at certain frequency ranges of the baseband signal the received signal is significantly more attenuated than at other ranges. In this case the receiver observes **I**nter**s**ymbol **I**nterference (*ISI*) (see Figure 2.3), as time-domain manifestation of the frequency selective behavior. If the delay spread is for example half of the symbol time, then signal copies of two consecutively sent symbols interfere at the receiver, such that the 'fast' signal copy of the latter sent symbol interferes with the 'slow' signal copy of the previous sent symbol.

In Figure 2.3 the effect of ISI is illustrated. A wireless channel with three paths is assumed and a delay spread of $\Delta\sigma = 90$ time units (channel impulse response with three major propagation paths at the top of the picture). On the left side, the symbol time is $T_\mathrm{s} = 400$ time units, which is much bigger than the delay spread. Therefore, symbols (shown here as 'high' or 'low' values) transmitted are only marginally influenced by ISI. The received signal $eN(t)$ is almost not corrupted. In contrast, on the right side of the picture, the resulting signal is shown for a symbol duration of $T_\mathrm{s} = 100$ time units. Thus, the symbol rate is four times higher, but the delay spread of the channel is almost identical to the symbol duration. As a consequence, the transmitted symbols are severely corrupted.

In practical situations most of the time both, Doppler and delay spread, are present. However, both effects can be of harm or not, depending on the ratio between the symbol time (baseband bandwidth) and the characteristic values of the effects Δf_d and $\Delta\sigma$. Therefore, a channel might be categorized as one of four different types. The four categories are listed in Table 2.1.

Mathematically, fading can be modeled as a stochastic process in time and frequency. It is common to characterize this process by its first- and second-order statistics. Regarding the first-order statistics is has been shown that the attenuation of the channel can be assumed to be *Rayleigh* distributed, as given in Equation 2.5[5].

$$p\left(h_\mathrm{fad}\right) = h_\mathrm{fad} \cdot e^{\frac{-{h_\mathrm{fad}}^2}{2}} \tag{2.5}$$

[3] This refers to perfectly synchronized transmitters and receivers

[4] The baseband bandwidth requirement of any communication system is strongly related to the rate of digital symbols transmitted. Thus, the delay spread might also be compared to the symbol duration T_s

[5] Modeling the fading process requires a complex random process. In the following, h_fad denotes the magnitude of this complex process, as it determines the attenuation experienced at the receiver. For a detailed discussion refer to [13, 4].

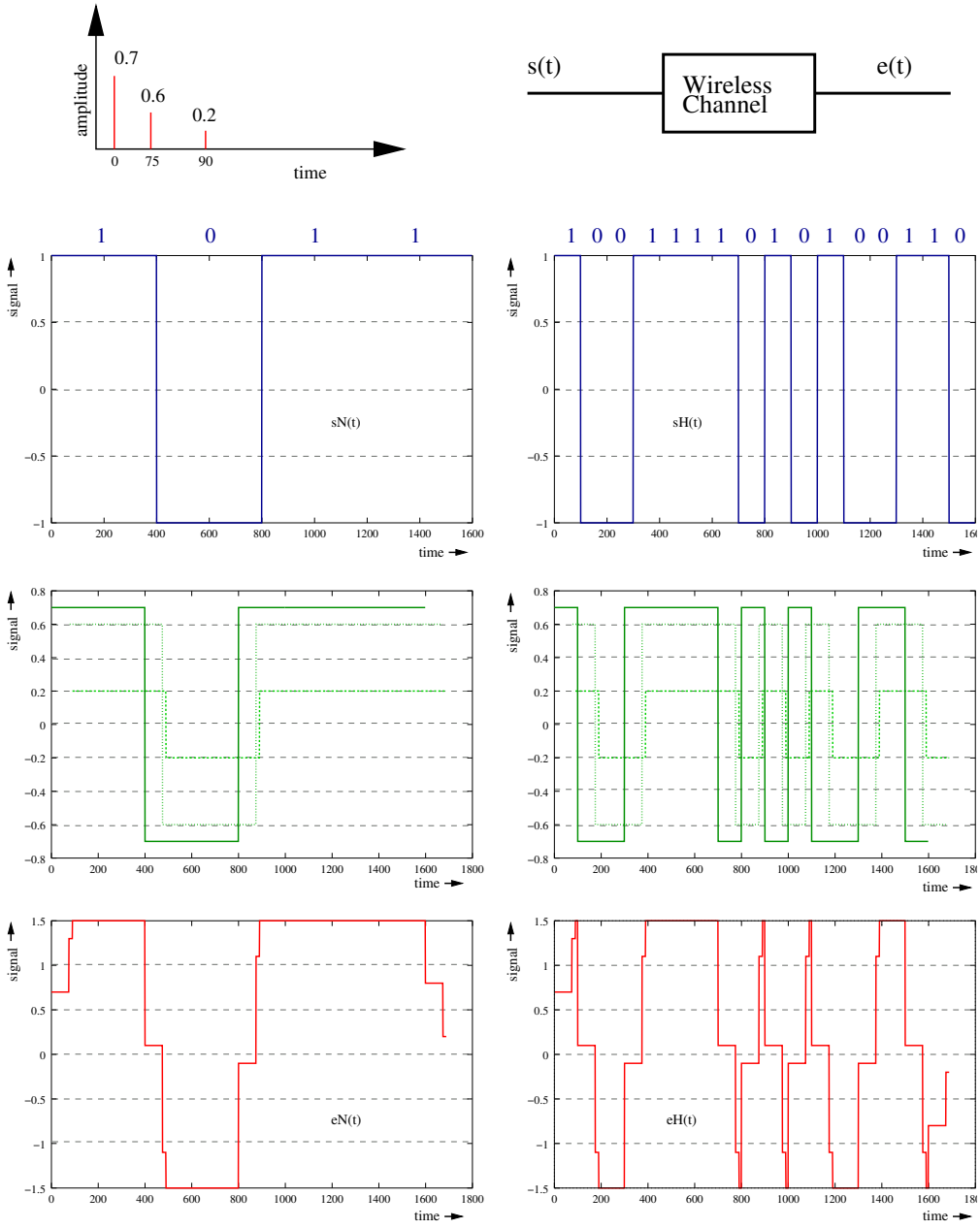


Figure 2.3: Example illustration of the impact of ISI. Top: Channel impulse response with a delay spread of $\Delta\sigma = 90$ time units. Left three pictures below: Transmit signal, propagation path copies of the transmit signal and resulting interference signal for a symbol duration of $T_s = 400$ time units. The received signal is almost not effected. Right three pictures: Corresponding conditions for a symbol duration of $T_s = 100$ time units. The data rate is four times larger, but the received signal is distorted significantly.

Criteria	Category
$\Delta\sigma \cdot B \ll 1, \Delta f_d \cdot n \cdot T_\mathrm{s} \ll 1$	not frequency selective (flat), not time selective (slow)
$\neg\,(\Delta\sigma \cdot B \ll 1), \Delta f_d \cdot n \cdot T_\mathrm{s} \ll 1$	frequency selective, not time selective (slow)
$\Delta\sigma \cdot B \ll 1, \neg\,(\Delta f_d \cdot n \cdot T_\mathrm{s} \ll 1)$	not frequency selective (flat), time selective (fast)
$\neg\,(\Delta\sigma \cdot B \ll 1), \neg\,(\Delta f_d \cdot n \cdot T_\mathrm{s} \ll 1)$	frequency selective, time selective (fast)

Table 2.1: Categories in order to characterize the fading of a wireless channel depending on the Doppler and delay spread

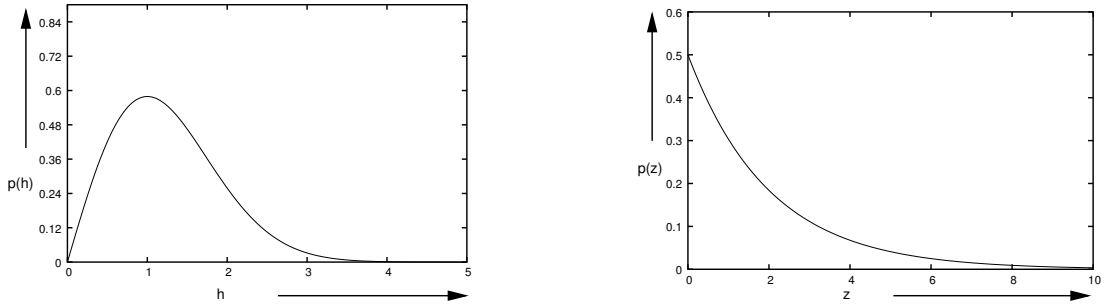


Figure 2.4: Probability density functions related to the magnitude of the fading attenuation: Rayleigh probability density (left) and χ^2 probability density (right).

Assuming the magnitude of the attenuation to be Rayleigh-distributed is a rather pessimistic model, as this assumes no dominating path to be present among all paths of the propagation environment between transmitter and receiver. A plot of the distribution is given in Figure 2.4.

For determining the actual SNR at the receiver, the instantaneous power has to be obtained rather than the instantaneous amplitude. This is given by the squared amplitude $z = h_\mathrm{fad}^2$ and the distribution of z is given in Equation 2.6. In fact, z has a χ^2 distribution with two degrees of freedom. A plot of the distribution is given in Figure 2.4.

$$p\,(z) = \frac{1}{2} \cdot e^{\frac{-z}{2}} \tag{2.6}$$

As mentioned, the Rayleigh distribution corresponds to a propagation environment with a NLOS setting, which is encountered for example in indoor scenarios as well as in macrocells of urban areas. Other scenarios require a different distribution (for example the Rice distribution for LOS settings [4]).

As with the first-order statistics of the fading process, models have also been developed for the second-order statistics. A very common model for these statistics is the *wide-sense stationary uncorrelated scatterer* (WSSUS) model [4, 2]. This model assumes the fading process to be stationary in time as well as in frequency. Hence, the autocorrelation in time as well as in frequency depends only on the time or frequency "shift", not on the absolute time or frequency. Therefore, in an WSSUS model the autocorrelation can be easily expressed given the power

delay profile and the power spectrum of the process. One popular setting for a WSSUS model is to assume the time correlation to be characterized by a *Jakes power spectral density*, while the frequency correlation is characterized by an *exponential power delay profile*. Both densities are parameterized solely by their corresponding spreads – the Jakes density by the Doppler spread, the exponential profile by the delay spread.

In order to roughly characterize the correlation (strength) in time as well as in frequency, two metrics have become quite accepted. They are the *coherence time* and the *coherence frequency*. The coherence time indicates the time span the wireless channel roughly stays constant. One mathematical definition of the coherence time is given in Equation 2.7, which equals the time shift during which the autocorrelation function drops to a value of 0.98 [4]. However, this definition is somewhat subjective and other definitions can be found in [2, 6, 14].

$$T_{\mathrm{c}} = \frac{1}{2\pi\Delta f_d} \tag{2.7}$$

The coherence bandwidth measures roughly the frequency spacing for which the channel does not change significantly. Again the exact mathematical definition is to some extend subjective. One definition of the coherence bandwidth is given by Equation 2.8, following [4]. Other definition might be found in [2, 6, 14].

$$W_{\mathrm{c}} = \frac{1}{2\pi\Delta\sigma} \tag{2.8}$$

2.1.2 Noise and Interference

Two major sources of *additive effects* are considered in general which potentially distort the signal. The first one is *noise*. Noise is always of stochastic nature and varies with time. It is denoted by $n(t)$.

The second effect corrupting the received signal is *interference*. Interference is caused by other RF transmitting electronic devices. As with the noise, interference has a stochastic nature and varies with time. It is denoted by $j(t)$. Interference is either caused by other systems operating in the same frequency band in the case of unlicensed bands or caused by co- and adjacent-channel interference in licensed bands. *Co-channel interference* happens due to frequency re-utilization, for example in a cellular environment. *Adjacent channel interference* is due to bandpass filters which produce a small power interference in neighboring bands.

The quality of the demodulation and decoding of a signal depends on the difference between the power of the received signal and the power of other signals in the same frequency band - interference - in addition to the noise. As this thesis does not consider interference to be present at the receiver, the effect of interference is neglected below.

Noise is always present and comes from several sources, for example atmospheric disturbances, electronic circuitry, human-made machinery etc. The first two belong to the group of thermal noise sources, described in depth below. Noise produced by human-made sources is described in more detail in [15].

Thermal noise is due to the movement of charged particles inside electronic components existent in every receiver system and is therefore unavoidable. The characteristics of thermal

noise were firstly studied for a resistor. It may be modeled as a zero-mean, wide sense stationary Gaussian stochastic process. The power spectral density of thermal noise can be obtained using the maximum power transfer theorem (Equation 2.9).

$$S = n_0/2 = k \cdot \mathcal{T}[\text{W/Hz}] \tag{2.9}$$

Here, $k = 1.37 \cdot 10^{-23}\, J/deg$ is the Boltzman constant and $\mathcal{T}$ refers to the temperature..

This kind of process is called white, i.e. thermal noise sources are modeled as having a flat power spectral density. It contains all frequencies, in analogy to white light, which contains all light frequencies. Accordingly, white noise is uncorrelated. However, in practical systems of limited bandwidth, the noise is filtered and is at the output no longer white, taking the shape of the filter's transfer function. This means that the noise becomes correlated when it is low-pass filtered. The noise power at the output of the filter depends on the filter's bandwidth (B). This average noise power at the output of the filter can be expressed as given in Equation 2.10.

$$\sigma^2 = g \cdot n_0 \cdot B \qquad , \tag{2.10}$$

where g is the power gain of the filter at the center frequency.

2.1.3 Simulation of Wireless Channels

As discussed above, a wireless channel can be considered on several different layers as well as over various different time spans. Thus, different metrics can be of interest for a simulation. It is quite important to state first what exactly is to be simulated regarding these aspects. In this section the generation of the instantaneous SNR is considered according to the composition following Equation 2.11.

$$v(t) = \frac{P_0\,(t)}{n(t)^2} = \frac{h(t)^2 \cdot P_{\text{tx}}\,(t)}{n(t)^2} = \frac{h_{\text{pl}}(t)^2 \cdot h_{\text{sh}}(t)^2 \cdot h_{\text{fad}}(t)^2 \cdot P_{\text{tx}}\,(t)}{n(t)^2} \tag{2.11}$$

Regarding the time span, channel samples separated by milliseconds are of interest, assuming that the channel is constant in between[6]. As the models discussed above distinguish three different processes influencing the SNR (apart from the transmit power), the simulation of the attenuation, the noise and the interference is discussed below.

Noise is often simulated by just considering a fixed power threshold which is constant over time. In reality it varies strongly over time. However, these variations are very fast and the simulation of these variations is only required if a system behavior on this small time scale is considered (typically in the range of μs). For example, if the attenuation, transmit power and interference stay constant over a time span of 1 ms, the symbol error rate obtained from assuming a constant noise threshold compared to the symbol error rate obtained from modeling a varying noise level are negligible, if the constant impact of the noise is determined correctly (taking the average noise power). As these rather large time scales are of interest in this thesis, the impact due to noise is modeled constant below.

[6]This requires a certain coherence time of the channel, which is not discussed in depth here. However, it is well justified in the simulation scenarios considered later in this thesis.

In contrast, interference might have a time varying behavior even on larger time scales, depending on the source of the interference. In order to simulate the influence due to interference, it is necessary to specify the interference source in detail (co-channel or adjacent channel interference source etc.). However, in this work no interference influence has been assumed, therefore no simulation technique for interference is discussed.

The impact due to attenuation can be decomposed into three different elements, as described by Equation 2.2. For the impact of the path loss, primarily the distance between transmitter and receiver has to be determined. This is done by a mobility model if terminals are assumed to roam. Otherwise, a static distance value has to be generated. In this work, transmitters and receivers were assumed to be static, while other objects in the propagation environment are assumed to be mobile (with important consequences for the simulation of fading, see below). The terminals are always assumed to be uniformly distributed over the area of the cell. Thus, given a certain cell radius, terminals are positioned in the cell according to a uniform distribution. This leads to the situation that the majority of the terminals has a rather large distance to the access point (relative to the cells radius). Once each terminal is associated with a certain distance to the transmitter, a homogeneous path loss model according to Equation 2.3 is assumed, yielding the path loss for each terminal. As the terminals do not roam, the path loss is constant.

The impact due to shadowing is stochastic, in contrast to the impact due to path loss. Shadowing varies as objects within the propagation environment move and circumstances change. The time scale on which shadowing varies is large (in the range of seconds), compared to the time scales of interest in this thesis (samples spaced by a few ms). For the shadowing, independent samples were generated according to a log-normal distribution (Equation 2.4), as discussed in Section 2.1.1.

The last factor regarding the attenuation is the fading. Fading is a stochastic element of the attenuation and varies on rather short time scales. Apart from being time selective, it is also frequency selective. This is in contrast to path loss and shadowing, which are both not frequency selective (unless rather large bandwidths > 100 MHz are considered). In this work, not only the first-order statistics are of interest but also the second-order statistics. A common assumption for the first-order statistics is the Rayleigh distribution of the attenuation (as the SNR is of interest, the square of the attenuation has a χ^2 distribution). The second order statistics can be characterized by a Jakes-distributed power spectrum and an exponential power delay profile (the power spectrum yields the correlation in time while the power delay profile yields the correlation in frequency). Simulating the fading with given first-and second order statistics can be done by using the method of Rice [16]. First, inphase and quadrature components are generated, which are colored Gaussian processes. Colored Gaussian processes can be generated by the superposition of harmonic functions with random amplitude, frequency and phase. Thus, the magnitude of two such processes (if the one is assumed to be the inphase component and the other one is the quadrature component) is already Rayleigh distributed. In order to give this resulting process the right correlation in time and frequency, each Gaussian process is colored appropriately, with random frequency values drawn from a Jakes-distribution and random delays drawn from an exponential distribution. Note that the resulting superposition of harmonic functions is a deterministic process which is sampled according to the needs of the simulation.

2.2 Orthogonal Frequency Division Multiplexing

In order to increase the data rate of a **S**ingle**c**arrier **M**odulation (*SCM*) system[7], one option is to increase the symbol rate. Increasing the symbol rate leads to a larger bandwidth requirement. However, once the attenuation of the channel is not flat anymore for the required bandwidth, the conveyed information is distorted by ISI (see Section 2.1.1). Thus, in broadband channels the performance of traditional SCM schemes is limited.

Unless the delay spread of a channel is much bigger than the symbol duration, equalizers can reduce the impact of ISI. For example, the **G**obal **S**ystem for **M**obile Communications (*GSM*) system is designed for a maximum delay spread of $\Delta\sigma = 20\mu s$. The symbol duration of the used **G**aussian **M**inimum **S**hift **K**eying (*GMSK*) modulation is $T_\mathrm{s} = 4\ \mu\mathrm{s}$ (200 kHz channel bandwidth). Therefore, ISI might occurs over five adjacent symbols. Although the maximum delay spread is bigger than the symbol duration, the complexity of an equalizer for such a scenario (with a ratio of 5) is still acceptable. A powerful equalizer can also cope with ISI over a longer symbol sequence but there is a limit. For example, if ISI effects 100 adjacent symbols, the calculation of the corresponding number of filter coefficients inside an equalizer would be too complex for a realistic system [2]. As a consequence, the equalizer becomes too expensive.

In such situations, the usage of a **M**ulti-**c**arrier **M**odulation (*MCM*) scheme is an alternative. The principle of such a scheme is this: A broadband channel is divided into many narrow-band sub-channels. Accordingly, the high rate data stream is divided into several parallel low rate data streams. The lower data rate streams are transmitted now via the narrow-band sub-channels. If a broadband channel is divided into N narrow-band sub-channels, the data rate per sub-channel is reduced by the factor of N. This way, the symbol duration for each sub-channel is increased by the factor of N and subsequently the impact of ISI is reduced.

MCM systems were introduced already in the late 1950's for **H**igh **F**requency (*HF*) radio links of the U.S. military [17]. As a special form of MCM systems, **O**rthogonal **F**requency **D**ivision **M**ultiplexing (*OFDM*) was patented in the U.S. in 1970 [18]. Despite the early introduction of MCM systems, the practical relevance was quite low. Until about 10 years ago, these systems suffered from their implementation complexity of real-time Fourier transformations as well as in the stability of oscillators used in transmitters and receivers [17]. Due to technological progress mainly in the field of digital signal processing meanwhile these issues are no longer demanding.

OFDM was first described in the 1960's [19, 20, 21] and was patented in 1970 [18]. If a high rate data stream is to be transmitted over a frequency-selective wireless channel, MCM schemes are an alternative as discussed above. In addition to the concept of MCM, the sub-channel (also referred to as sub-carrier) spectra overlap each other but do not interfere at the transmit frequencies (carriers). In communications, such channels are referred to as *orthogonal* [2]. Given a certain bandwidth B for communications, orthogonality is achieved by using the same symbol duration T_s for all sub-carriers. Thus, the sub-carriers are equidistantly spaced on the frequency axis by a distance of [22]

$$\frac{B}{N} = \Delta f = \frac{1}{T_\mathrm{s}} \quad . \tag{2.12}$$

[7] This refers to an equalized QPSK modulation system, for example.

This leads to the orthogonality of the sub-carrier signals. Although the same symbol duration is applied, different modulation types can be used per sub-carrier. Analytically, the n-th sub-carrier signal is given by a function $\tilde{s}_n(t)$ (refer to Equation 2.13).

$$\tilde{s}_n(t) = S_{n,t} \cdot e^{j \cdot 2\pi \cdot n \cdot \Delta f \cdot t} \tag{2.13}$$

$S_{n,t}$ refers to the modulation symbol at time t on sub-carrier n. The superposition of all N sub-carrier signals yields the OFDM block, the superposition of all N modulated sub-carriers (Equation 2.14).

$$\tilde{s}(t) = \frac{1}{\sqrt{N}} \cdot \sum_{n=1}^{N} s_n(t) = \frac{1}{\sqrt{N}} \cdot \sum_{n=1}^{N} S_{n,t} \cdot e^{j \cdot 2\pi \cdot n \cdot \Delta f \cdot t} \tag{2.14}$$

Compared to an equivalent[8] SCM system, the symbol duration is N times longer per sub-carrier. When designing an OFDM system for a certain wireless channel, the symbol time T_s is chosen such that it is bigger than the delay spread $\Delta\sigma$ of the channel. This requires an appropriate choice of N, according to Equation 2.12. However, ISI might still occur. In order to prevent this completely, a fixed signal pattern is transmitted in addition. This pattern extends each OFDM block. It is a cyclic extension of the time domain representation of the OFDM block and is referred to as the *guard interval*. It has a duration of T_g. The combination of the guard interval and the OFDM block represent an *OFDM symbol*. This symbol has a time length of $T_\mathrm{s} + T_\mathrm{g}$ (see Figure 2.5).

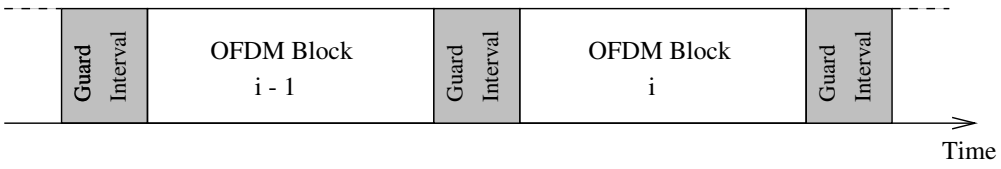


Figure 2.5: OFDM symbols, each consisting of an OFDM block and the guard interval.

The function of the guard interval is to lower the impact due to ISI. At the receiver, the guard interval is removed and only the signal within the time span $[0, T_\mathrm{s}]$ is considered. If the guard interval is appropriately chosen, ISI occurs only within the guard intervals. Therefore, ISI does not affect the symbol following the guard interval. However, such a system design leads to a loss of throughput, since the time duration of the guard interval (of length T_g) is lost for data transmission.

As digital symbols are modulated to each sub-carrier at a rate of $\frac{1}{T_\mathrm{s}}$, the frequency domain representation of the modulated signal of sub-carrier n is a sinc-function (Equation 2.15). The frequency-domain representation of the complete OFDM signal is therefore the sum of all N frequency-domain sub-carrier representations (Figure 2.6). As mentioned previously, orthogonality is given by the fact that at the carrier frequencies of each sub-carrier the frequency-domain amplitude of all adjacent sub-carriers is zero.

[8] Equivalent refers here to the symbol rate.

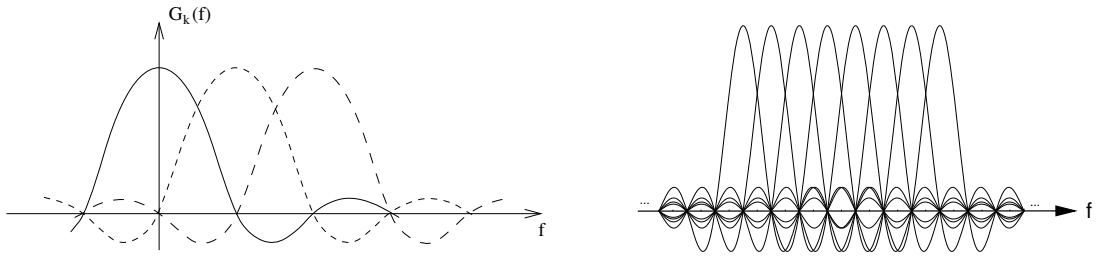


Figure 2.6: OFDM spectrum (Left: Detailed principle – Right: Complete spectrum).

$$G_n(f) = T_\mathrm{s} \cdot \mathrm{sinc}\left(\pi T_\mathrm{s}\left(f - n\Delta f\right)\right) \tag{2.15}$$

In order to generate an OFDM signal, the inverse discrete Fourier transform (IDFT) is applied [23, 24]. First, all N modulation symbols $S_{n,t}$ are determined[9] for the next OFDM block. Then, the overall OFDM signal is generated by applying an inverse discrete Fourier transform where the modulation symbols per sub-carrier are the frequency domain weights (Equation 2.14).

In practical systems, this IDFT is implemented by an IFFT in order to reduce the computational complexity. Before the discrete-time transmission signals are D/A converted, the guard interval is added in front of the symbol. At the receiver, after removing the guard interval, the obtained signal is applied to a discrete Fourier transform (the FFT in practice) in order to obtain the modulation symbols. The complete transmission system is shown in Figure 2.7.

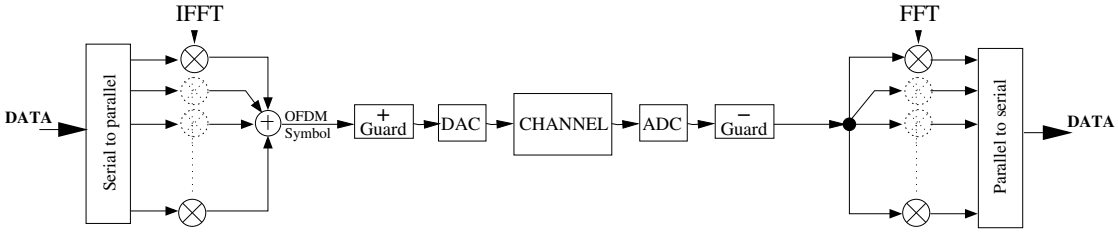


Figure 2.7: Complete OFDM transmission sketch.

Apart from the robustness against frequency selective fading, OFDM has further advantages. These are mainly [25]:

- Transmitter and receiver can easily be implemented using the FFT.
- The bandwidth efficiency (ratio between used bandwidth and conveyed information) of OFDM is very high.

[9]The modulation symbol per sub-carrier is chosen according to the current bit sequence to be represented.

However, there are several disadvantages that should be mentioned. These are (a discussion is given below):

- The bit error rate decreases only slowly in Rayleigh-fading channel environments (flat fading behavior) if no channel coding is applied.
- OFDM systems are sensitive to non-linear effects.
- The frequency synchronization requirements are high.

Flat fading behavior is observed for many modulation schemes. An example illustration of the impact of flat fading is given in Figure 2.8 where the bit error rate is given versus the SNR for BPSK modulation with radio channel error statistics according to a Rayleigh-distributed attenuation and additive white Gaussian noise. A countermeasure is error-correction coding in combination with interleaving introduced to the communication system. However, this reduces the bandwidth efficiency by accepting redundancy in the system. Diversity techniques such as space diversity may be applied to the system in order to yield better transmission results for temporarily attenuated sub-carriers. Finally, sub-carrier selection schemes can be applied, which simply "switch off" sub-carriers that are attenuated too much [26].

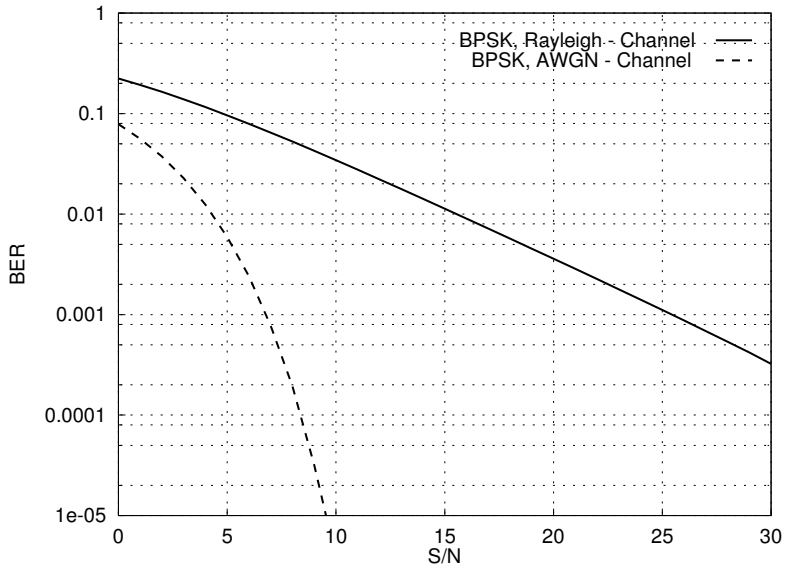


Figure 2.8: Bit error rate of a BPSK modulated wireless channel with additive, white Gaussian noise and Rayleigh-distributed attenuation. [27].

Non-linear effects are mainly introduced to the system by the equipment used. A particularly important part of the system is the amplifier of the transmitter. After transforming the frequency-domain signal to the time domain sample sequence, the amplitudes of these samples vary quite drastically. This leads to the widely known **P**eak-to-**A**verage **R**atio **P**roblem (*PARP*)

[22]. Compared to the average power of the sample sequence, certain parts have an instantaneous power which is magnitudes higher. Thus, the amplifier has to provide a large range of linear behavior. Otherwise, the time domain signal is distorted, which leads to modulation detection errors (and ultimately to bit errors). It has been claimed that this issue can be resolved by introducing clipping [22] to avoid sub-carrier interference or by the introduction of interleaving in combination with channel coding [22]. A further approach is based on a sub-carrier selection scheme. Several "badly" attenuated sub-carriers are not employed for data transmission. Instead, their coefficients are modified, leading to a signifcant reduction of the peaks of the transmit signal [28].

Finally, imperfect frequency synchronization of oscillators as well as temporary phase jitter due to the Doppler effect lead to a performance degradation in OFDM systems. Again, this results in an error floor, as shown in Figure 2.9. Using accurate oscillators in the receiver as well as channel coding can improve this behavior [26].

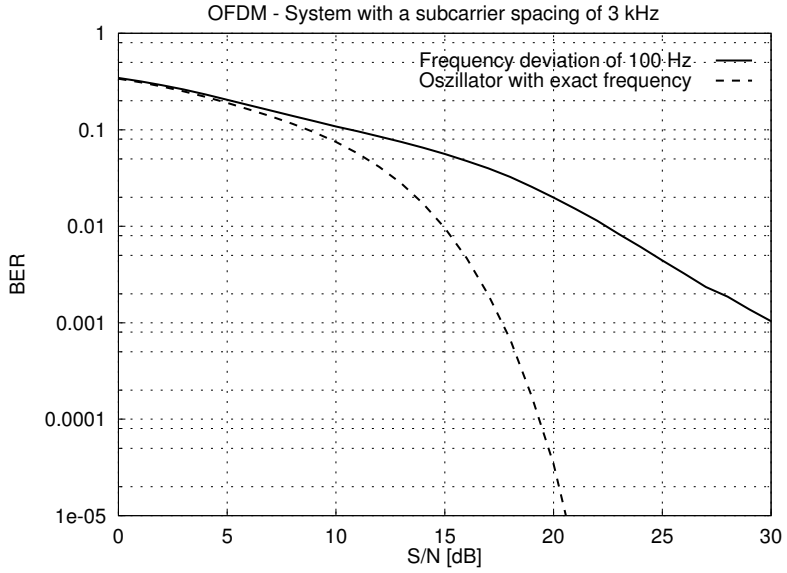


Figure 2.9: Bit error rate of an OFDM system with and without correct frequency synchronization [27].

Today, OFDM is applied as transmission scheme in many communication systems, wired and wireless ones. It is applied in **A**symmetric **D**igital **S**ubscriber **L**ine (*ADSL*) systems as the twisted-pair telephone cable is highly frequency selective for larger bandwidths [29]. In the context of ADSL the transmission scheme is also often referred to as **D**iscrete **M**ulti-**t**one (*DMT*) (a synonym for OFDM). ADSL systems provide asymmetric high-speed data connections on traditional twisted-pair telephone cables and have been developed a the end of the 1980's and beginning 1990's (first mass market introduction started in 1997). In ADSL systems, approximately a bandwidth of 1 MHz is split into $N = 256$ sub-carriers and is used for down-link data communications (the first 4 kHz are required for the traditional telephone system while the

next 60 kHz are reserved for the ADSL up-link). Note that exist varies further variants of DSL systems with different system parameters.

One of the first standards for wireless mass market communication systems based on OFDM have been the broadcasting standards for digital radio and television, **D**igital **A**udio **B**roadcasting (*DAB*) [30] and **D**igital **V**ideo **B**roadcasting (*DVB*) [31][10]. These standards have been developed during the 1990's and were released in 1995 (DAB) and 1997 (DVB). DAB systems utilize a bandwidth of about 1.5 MHz, in Germany at a transmit frequency of about 1.47 GHz. The given bandwidth is divided into different numbers of sub-carriers, ranging from $N = 192$ to $N = 1536$. In the case of DVB systems, a system bandwidth of 8 MHz is split for into $N = 2048$ or $N = 8192$ sub-carriers (there exist different modes). The transmit frequency is about 200 MHz in Germany, as it is with the analog television systems.

Meanwhile, OFDM is part of several **W**ireless **L**ocal **A**rea **N**etwork (*WLAN*) standards. It is the physical layer transmission scheme for IEEE 802.11a [32] and HIPERLAN/II [33]. Both systems rely on the same physical layer specifications and utilize a system bandwidth of 16.25 MHz, divided into $N = 52$ sub-carriers. As transmit frequencies, the ISM bands in the 5 GHz band are used. While both standards have been released at the end of the 1990's, today only IEEE 802.11a has become a successful commercial product. It even has been extended for use in the 2.4 GHz ISM bands, leading to the standard IEEE 802.11g (which basically combines the 802.11a standard with the 802.11b standard). In addition to these WLAN standards, a relatively new **W**ireless **M**etropolitan **A**rea **N**etwork (*WMAN*) standard IEEE 802.16 has been released [34], which is also based on OFDM. In IEEE 802.16 systems, a (varying) bandwidth between 1.75 MHz and 20 MHz is split into $N = 256$ sub-carriers. The transmit frequencies vary between 2 GHz and 11 GHz. In contrast to WLAN, IEEE 802.16 is designed to provide high-speed wireless access at a much larger range (about a few kilometers).

Apart from these existing standards, OFDM is proposed for several upcoming systems. One of them is a new standard for mobile broadband wireless access, probably released under the name IEEE 802.20. As 802.16 has no capabilities to support true roaming of wireless terminals, 802.20 is going to have these functionalities. It will definitely be based on OFDM as transmission scheme. In addition, OFDM is likely to be selected as transmission scheme for future broadband cellular communication systems, such as 3G evolutions (OFDM extensions, referred to as long term evolution (LTE) of 3G) and 4G systems.

System	Bandwidth	Sub-carrier Number	Center Frequency
DSL	1 MHz	256	Baseband
DAB	1.5 MHz	1536	1.47 GHz
DVB	8 MHz	2048	200 MHz
IEEE 802.11a	16.25 MHz	52	5.4 GHz
IEEE 802.16	1.75 MHz	256	2 GHz

Table 2.2: System parameters (part I) of selected OFDM transmission systems

[10]There exist different DVB standards also for satellite broadcasting (DVB-S) and cable broadcasting (DVB-C). Here only the terrestrial broadcasting standard (DVB-T) is discussed.

System	Application
DSL	Wired, physical layer protocol
DAB	Digital broadcasting of audio
DVB	Digital broadcasting of video
IEEE 802.11a	Wireless local area network
IEEE 802.16	Wireless metropolitan area network

Table 2.3: System parameters (part II) of selected OFDM transmission systems

2.3 Entropy of an Information Source

The main goal of lossless data compression is to transform a certain amount of data into a new representation, which contains exactly the same information, but whose length is as small as possible. The proliferation of communication networks results in a massive transfer of data over communication links, i.e. wireless channels. Compressing transmit data can reduce the costs of communications significantly. It can be said that, for example, when the amount of data to be transmitted is reduced up to 50% by means of lossless compression, the effect is that of doubling the capacity of the communication channel.

This section gives an introduction to parts of Shannon's information theory which are related to data compression. In [3], Shannon formulated the theory of data compression, where a fundamental limit to lossless data compression was established. Obtaining this limit, called the entropy rate H, will help to obtain a theoretical limit to the compression rate that can be achieved for any information source. Most issues discussed here are based on [35, 36].

2.3.1 Source Modeling

In 1948, Shannon defined the theoretical limit to lossless data compression [3]. This limit is known as entropy rate, and is denoted by H. It is possible, in theory, to compress an information source in a lossless manner (i.e. so that the original data can be recovered completely once it has been compressed) with a compression rate close to H. It is mathematically impossible to reach a better compression than H. Hence, this theory sets the fundamental limits on the performance of all data compression algorithms.

The exact value of H depends on the statistical nature of the information source. Therefore, it is important to have a good model of the source. Throughout this section, English text is used as an example of information source to be compressed, as it was done in [3]. Let the source be a book selected randomly from a library. Mathematically, we can represent this book by

$$\mathcal{X} = (X_1, X_2, X_3, X_4, \ldots)$$

where $\mathcal{X}$ represents the whole book, X_1 represents the first character in the book, X_2 the second character, and so on. Assume that this book has an infinite length, as this simplifies the study from a mathematical point of view. The characters in this book are either a lower-case letter or a *space*. Then, the source alphabet $\mathcal{A}$ is defined to be the set of all 27 possible values:

$$\mathcal{A} = \{a, b, c, d, e, f, g, h, i, j, k, l, m, n, o, p, q, r, s, t, u, v, w, x, y, z, \textit{space}\}$$

Now consider to compress this book. As it is not known what book has been selected, the characters in the book ($X_i, i = 1, 2, \ldots$) are *random variables* which take values on the alphabet $\mathcal{A}$. The whole book $\mathcal{X}$ is an infinite sequence of random variables, a *random process*. At this point, there are several ways the statistical properties of this book can be modeled.

- Zero-Order Model: In this model, it is assumed that each character is statistically independent of all other characters, and the 27 possible values in the alphabet $\mathcal{A}$ are equally likely (uniformly distributed) to occur. If this example were accurate, then a typical opening of the book could look as follows (examples taken from [3]):

  xfoml rxkhrjffjuj zlpwcfwkcyj ffjeyvkcqsgxyd qpaamkbzaacibzlhjqd

- First-Order Model: It is known that in English some letters occur more frequently than others. For example, a and e are a lot more common than q and z. In this model, the characters are still supposed to be statistically independent, but now a non-uniform probability distribution exists. A typical text according to this model could look like this:

  ocro hli rgwr nmielwis eu ll nbnesebya th eei alhenhttpa oobttva nah

- Second-Order Model: The previous two models assumed statistical independence between the characters. However, that does not reflect the nature of English language. This can be shown with the following example:

  so#e let#ers in th#s sent#nce are missi#g

  Inspecting the sentence, even though some letters are missing, it still can be understood what is being said from looking at the context. This implies that there is some dependency between the characters. For example, the letter u occurs with a rather low probability ($p_u = 0.022$). However, given that the previous character is q, the probability of u being the character that follows is much higher ($p_{u|q} = 0.995$). Also, it can be said that characters close to each other in a sentence are more dependent than those that are far from each other. In this model, the present character X_i depends on the previous character X_{i-1}, but is statistically independent of all the other characters $(X_1, X_2, \ldots, X_{i-2})$. According to this model, the probability distribution of X_i varies, depending on what the previous character X_{i-1} was. An example of a sequence generated according to this model is given below:

  on ie antsoutinys are t inctore st be s deamy achin d ilonasive tucoowe

- Third-Order Model: This model is an extension of the previous one, where the present character X_i is dependent on the previous two characters (X_{i-2}, X_{i-1}), but is independent of all previous characters. An example would be:

  in no ist lat whey cratict froure bers grocid pondenome of

As it can be seen, the resemblance to English text increases with each of the steps taken above. In fact, a Fourth-Order Model approximation would already be surprisingly close to real language, as is shown in the following example [37]:

> the generated job providual better trand the displayed code abovery upondults well the coderst in thestical it do hock bothe merg instates cons eration never any puble and to theory evential callegand to elast benerated in with pies as is with the

However, the complexity associated with this models grows exponentially with n, being n the number of variables the current character depends on.

2.3.2 Entropy Rate of a Source

The value of the entropy rate of a source depends only on the statistical properties of the source. If the model chosen for the source is a simple one, then it is easy to calculate the entropy rate. Hence, the value of H varies depending on the model used. Here, an arbitrary source is considered, however staying with the example of English text:

$$\mathcal{X} = (X_1, X_2, X_3, X_4, \ldots)$$

- Zero-Order Model: For this model, the value of H is obtained by the following equation:

$$H(X_i) = \log_2 m \quad \text{bit/character} \ , \tag{2.16}$$

  where m is the size of the source alphabet. As said before, this size is $m = 27$ for English text. Thus, the entropy rate of English text is $H = 4.75$ bit/character, according to this model.

- First-Order Model: While the characters are still statistically independent, now p_i is the probability of the i-th letter in the alphabet. The entropy rate using this model is:

$$H(X_i) = \sum_{i=1}^{m} p_i \cdot \log_2 (1/p_i) \ \text{bit/character} \ , \tag{2.17}$$

  Using this model, the entropy rate of English text would be 4.07 bit/character.

- Second-Order Model: Let $p_{j|i}$ be the conditional probability that the current character is the j-th letter in the alphabet, given that the previous character was the i-th. Then, the entropy rate for this model is calculated as follows:

$$H(X_i|X_{i-1}) = \sum_{i=1}^{m} p_i \sum_{j=1}^{m} p_{j|i} \cdot \log_2 \left(1/p_{j|i}\right) \ \text{bit/character} \ , \tag{2.18}$$

  Again, back to the text example, the entropy rate given by this model would be 3.36 bit/character.

- Third-Order Model: Now define $p_{k|j,i}$ as the conditional probability of the present character being the k-th letter in the alphabet $\mathcal{A}$, given that the previous two characters where the j-th one (for X_{i-1}) and the i-th one (for X_{i-2}). According to this model, the entropy is:

$$H(X_i|X_{i-1}, X_{i-2}) = \sum_{i=1}^{m} p_i \sum_{j=1}^{m} p_{j|i} \sum_{k=1}^{m} p_{k|j,i} \cdot \log_2 \left(1/p_{k|j,i}\right) \text{ bit/character} \ , \quad (2.19)$$

  With this model, the amount of bits theoretically needed to code a character in English text is reduced to 2.77.

It is clearly seen that the more knowledge of the source is given (meaning by knowledge the number of previous characters known) the lower is the number of bits needed to code the current character. This basic fact of compression is captured in the following equation:

$$H(X|Y) \leq H(X) \quad . \quad (2.20)$$

2.3.3 Lossless Coding Schemes

The concept of *block coding* is presented in the following. As example, a source with alphabet $\mathcal{A} = \{a, b\}$ is considered. Suppose that both letters 'a' and 'b' are equally likely to occur. However, given that 'a' occurred in the previous character X_{i-1}, the probability of the current character X_i being 'a' again is 0.9. The same happens to the letter 'b', i.e. $p(X_i = b \,|\, X_{i-1} = b)$ = 0.9. An n-th order block code is a mapping which assigns to each block of n consecutive characters a sequence of bits of varying length. If a *First-Order Block Code* is used, the following mapping could be used:

Block	$p(Block)$	Codeword
a	0.5	0
b	0.5	1
	$H(X_i)$ = 1	

Consider the following sequence as original data:

aaaaaaabbbbbbbbbbbbbaaaa

In this case it would require exactly 24 bits to code the 24 characters, i.e. an average of 1 bit/character. However, using a code with a higher order could reduce the amount of bits needed per character. A *Second-Order Block Code* would look like this:

Block	$p(Block)$	Codeword
aa	0.45	0
bb	0.45	10
ab	0.05	110
ba	0.05	111
	$H(X_i, X_{i-1})$ = 0.825	

In the example, 20 bits would be needed to represent the 24 characters, an average of 0.83 bits/character. It can be seen that using block coding reduces the amount of bits needed to code the data. This can be explained with the concepts defined previously in this section. The idea is to take advantage of the mutual information that exists between the two random variables X_i and X_{i-1}. If both variables are coded independently, a first-order block code would be used as the one shown above. This requires $H(X_i) + H(X_{i-1})$ bits to code the information. However, with the second-order code, a lower number is needed, given by $H(X_i, X_{i-1}) = H(X_i) + H(X_{i-1}) - I(X_i; X_{i-1})$ bits. If the two variables are not independent, as in the example, the compression of the source can be achieved. Naturally, using more complex codes will bring more compression (if the source offers this degree of correlation), as it is shown in the last example, using a *Third-Order Block Code*:

Block	$p(Block)$	Codeword
aaa	0.405	0
bbb	0.405	10
aab	0.045	1100
abb	0.045	1101
bba	0.045	1110
baa	0.045	11110
aba	0.005	111110
bab	0.005	111111
	$H(X_i, X_{i-1}, X_{i-2})$ = 0.68	

In this case, only 17 bits are used to code the 24 characters (0.71 bits/character). Note that the higher the order of the code, the lower the average codeword length. This means, in other words, that better compression can be achieved. The codes that have been used here are *Huffman codes* [38]. The here presented example of block coding gives a brief introduction to the issue how the entropy rate of a sources is achieved by practical algorithms. Today, there exist many different such compression algorithms, which employ quite different mechanisms.

2.4 Mathematical Programming

The term "mathematical programming" is an expression from operations research and refers to *static* optimization problems [39]. Such optimization problems are characterized by the fact that a single decision-maker has complete information, i.e. all possible decisions are known as well as the cost of each decision [40]. In contrast, an optimization problem might either involve multiple decision makers and/or only a stochastic description of the cost. In the first case results of game theory [41, 42] have to be applied while in the second case a stochastic optimization approach has to be taken [43, 44]. However, in this section only optimization problems with a single decision-maker and complete information are discussed.

An optimization problem is defined by the pair (S, F), given as follows [45]:

$$\begin{aligned} \min \quad & F(\mathbf{x}) \\ \text{subject to (s.t.)} \quad & \mathbf{x} \in S \end{aligned} \tag{2.21}$$

where $F : S \rightarrow \mathbf{R}$ is the *objective function* (the cost of a decision), and S is the *feasible solution set* (also referred to as optimization space). Any point $\mathbf{x} \in S$ is called a *feasible solution*, where $\mathbf{x}$ is (usually) a n-vector $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ of the n-dimensional Euclidean space $\mathbf{R}^n$. Obviously, $S \subseteq \mathbf{R}^n$.

In general, one speaks of an instance of an optimization problem if the complete "input data" (i.e. all parameters defining S and F) are given. For example, an instance of the famous traveling salesman problem [45] is given if the distance matrix is specified. One speaks of an optimization problem as of the collection of all instances.

An optimization problem "asks" for a specific feasible solution $\mathbf{x}^*$, referred to as *optimal* solution, for which the following holds:

$$F(\mathbf{x}^*) \leq F(\mathbf{x}) \qquad \forall \mathbf{x} \in S \; . \tag{2.22}$$

However, such an optimal solution $\mathbf{x}^*$ might not exist at all. For example, S might be empty. Then, the optimization problem is called *infeasible*. On the other hand, it may happen that $F(\mathbf{x})$ is not bounded on S. Then the optimization problem is called *unbounded* and again no optimum exists with respect to Equation 2.22. Finally, S can be open which also leads to the situation that the optimization problem might not be solvable.

Optimization theory basically considers procedures which yield such optimal solutions to optimization problems. It is important to note that these procedures can be arbitrarily complex and in fact for some classes of optimization problems today there is no better way known of finding the optimal solution than enumerating all feasible solutions and comparing their objective values. Static optimization problems are classified by the structure of the feasible set S and the composition of the cost function F. In this section, two particularly important classes of optimization problems are covered furthermore, namely linear programs and linear integer programs.

2.4.1 Linear Programming

Linear programming (LP) is an exact, effective, and easily accessible optimization approach, which is applicable to optimization problems whenever they can be formulated as linear programming problem [39]. Today, LP solvers are available on both the commercial and freeware basis, and some of them are capable of solving large linear programs with many millions of variables and constraints. Thus, many other optimization problems are casted into linear programming formulations (whenever possible) or are modified (i.e. approximated) such that they become linear programming problems in order to obtain optimal or approximate solutions.

The general form of an linear programming problem is as follows [40]:

$$\begin{aligned} \max \quad & \sum_j c_j \cdot x_j \\ \text{s. t.} \quad & \sum_j a_{i,j} \cdot x_j \leq b_i \qquad \forall i = 1, 2, \ldots, m \; . \end{aligned} \tag{2.23}$$

Denote by $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ the vector of the variables, by $\mathbf{A} = \{a_{i,j}\}$ the $m \times n$ matrix of coefficients, by $\mathbf{c} = (c_1, c_2, \ldots, c_n)$ the cost coefficient vector and by $\mathbf{b} = (b_1, b_2, \ldots, b_m)$

the *right-hand side* vector. Then, the linear program of Equation 2.23 can be formulated in the *matrix form* as follows[11]:

$$\begin{array}{ll} \max & \mathbf{c}^T \cdot \mathbf{x} \\ \text{s. t.} & \mathbf{A} \cdot \mathbf{x} \leq \mathbf{b} \ . \end{array} \tag{2.24}$$

A linear program is given in *standard form* if, in addition to (2.24), all variables are constrained to be non-negative, i.e.:

$$\begin{array}{ll} \max & \mathbf{c}^T \cdot \mathbf{x} \\ \text{s. t.} & \mathbf{A} \cdot \mathbf{x} \leq \mathbf{b} \\ & \mathbf{x} \geq 0 \ . \end{array} \tag{2.25}$$

Thus, a linear program consists of a feasible set S which is defined by a set of linear inequalities. In addition, the objective function is also linear. Note that any equality constraint of the form $\sum_j a_{i,j} \cdot x_j = b_i$ can be reformulated by stating two inequalities $\sum_j a_{i,j} \cdot x_j \leq b_i$ and $-\sum_j a_{i,j} \cdot x_j \leq -b_i$. Therefore, even if the set S is originally defined by a set of linear equalities or a mixture of linear equalities and inequalities, one can always reformulate the definition of S such that only inequalities are obtained. Also, any general linear program in the form of (2.24) can be reformulated in the standard form of (2.25).

The set S of feasible solutions $\{\mathbf{x} \ : \ \mathbf{A} \cdot \mathbf{x} \leq \mathbf{b}\}$ is called a *polyhedron* [46]. A polyhedron is the intersection of finitely many closed half-spaces of $\mathbf{R}^n$. If the polyhedron is bounded, the set is called a *polytope*. For example, consider the following set of points:

$$\left\{(x_1, x_2) \in \mathbf{R}^2 \ : \ -x_1 + x_2 \leq 1 \ \wedge \ x_1 + x_2 \leq 2\right\} \qquad .$$

This set is an unbounded polyhedron, shown in the left graph of Figure 2.10. In contrast, if the set is restricted to contain only positive points (thus, if the conditions $x_1 \geq 0$ and $x_2 \geq 0$ are added) the polyhedron is bounded and a polytope is obtained (shown in the right graph of Figure 2.10).

A feasible solution $\hat{\mathbf{x}}$ is called an extreme point of the polytope, if it cannot be expressed by a linear combination of other feasible solutions $\mathbf{x}^1, \mathbf{x}^2, \ldots$. Such an extreme point of a polytope is called a *vertex*. An important property of LP problems is that if an LP problem is bounded and feasible, then the optimal solution set contains at least one vertex (all optimal solutions of a problem form the optimal solution set).

As example, consider the following LP problem (based on the polytope shown in Figure 2.10):

$$\begin{array}{ll} \max & x_1 + 3x_2 \\ \text{s. t.} & -x_1 + x_2 \leq 1 \\ & x_1 + x_2 \leq 2 \\ & x_1, x_2 \geq 0 \quad . \end{array} \tag{2.26}$$

The polytope and the vertices of this example are shown in Figure 2.11. The optimal solution to Problem (2.26) is given at the vertex $(0.5, 1.5)$.

[11] In the context of vectors and matrices $\leq$,$=$, and $\geq$ are understood in the following componentwise.

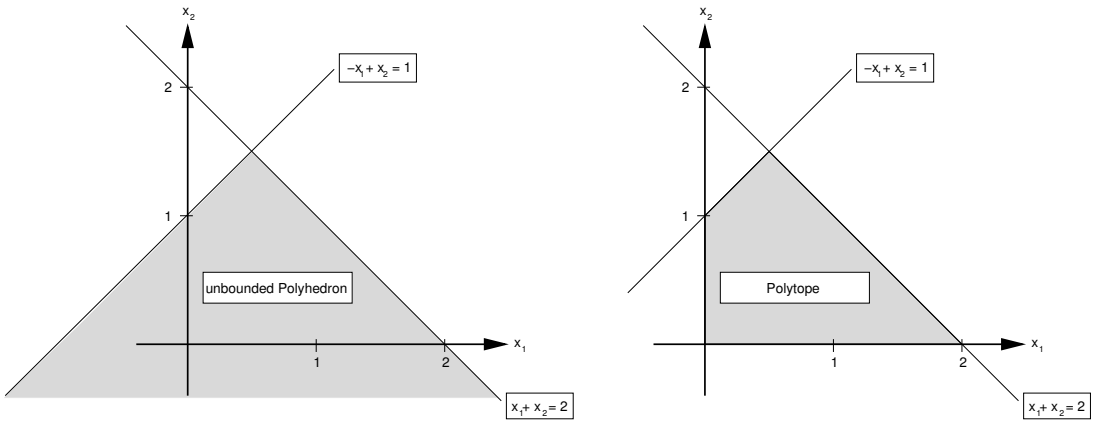


Figure 2.10: Left: Example of an unbounded polyhedron – Right: Example of a bounded polyhedron (polytope).

Figure 2.11 motivates a simple graphical solution method. One can obtain an optimal vertex by "sliding" the plane orthogonal to the objective function vector into the direction of the objective function vector. The last point met by the plane is an optimal solution (if the problem is feasible). However, for problems of a higher dimension of $\mathbf{R}^3$ this is rather difficult. Thus, a precise, efficient solution algorithm is of interest. From the discussion above it is clear that if an LP is feasible, there exists at least one vertex which is an optimal solution. Thus, one way to obtain an optimal solution would be to enumerate all vertices of the LP and compare their objective values. However, the set of all vertices of some LP is usually not known as such and has to be computed. Also, there might exist exponentially many vertices as the number of variables and constraints increases for an LP. Hence, the enumeration method is computationally not viable.

Instead, a different solution method has become the standard algorithm for linear programming problems, known as the simplex[12] method [46]. The idea behind the simplex method is quite simple: Initially, obtain some vertex of the feasible solution space of the linear program. If this initial vertex is not optimal, select one of the edges leaving the vertex in a direction along the objective function improves and traverse to the next vertex. Eventually, an optimal vertex is obtained by this method. In the following a simplified version of the simplex method is discussed for the maximization case.

Given a linear program in standard form, such as the one in (2.25), the first step is to reformulate the problem in what is known as canonical form, given as:

$$\begin{aligned} \max \quad & \mathbf{c}^T \cdot \mathbf{x}' \\ \text{s. t.} \quad & \mathbf{A} \cdot \mathbf{x}' = \mathbf{b} \\ & \mathbf{x}' \geq 0 \ . \end{aligned} \tag{2.27}$$

This might require the introduction of additional variables, known as *slack variables*. Consider the linear constraint $-x_1 + x_2 \leq 1$. This can only be reformulated as an equality constraint by

[12]The name simplex refers to a special kind of polytope.

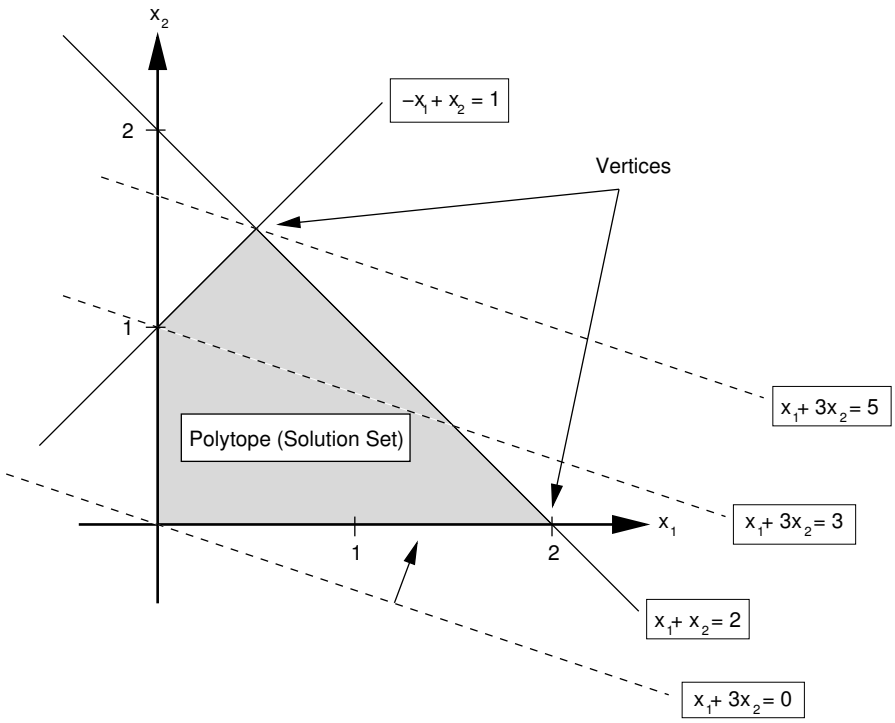


Figure 2.11: Illustration of the linear program (2.26).

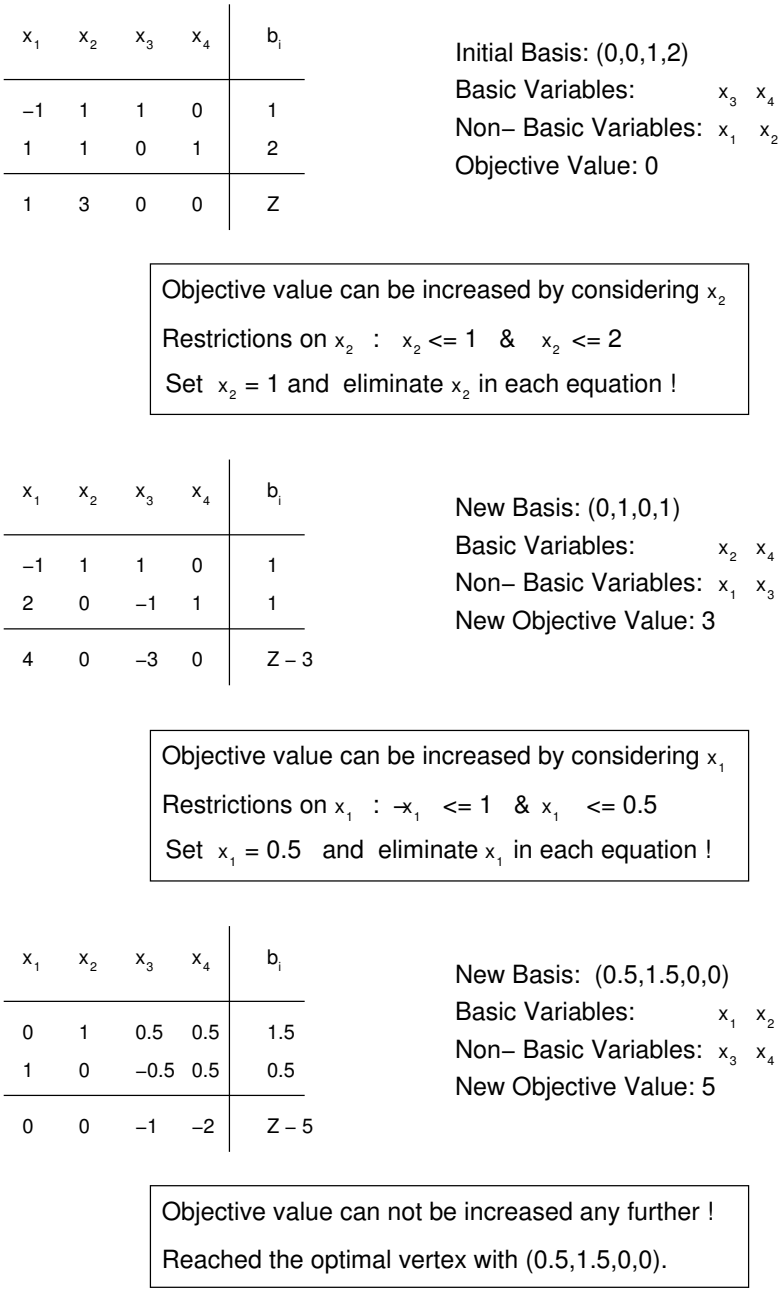


Figure 2.12: The iteration steps of the simplex method for the example linear program of (2.26).

introducing the slack variable x_3, which yields then $-x_1 + x_2 + x_3 = 1$. For each inequality constraint, one such slack variable has to be introduced. For the example given in (2.26), the corresponding linear program in canonical form is given by:

$$\begin{aligned} \max \quad & x_1 + 3x_2 \\ \text{s. t.} \quad & -x_1 + x_2 + x_3 = 1 \\ & x_1 + x_2 + x_4 = 2 \\ & x_1, x_2, x_3, x_4 \geq 0 \quad . \end{aligned} \tag{2.28}$$

In general, for an $n \times m$ matrix $\mathbf{A}$, the canonical formulation yields $n + m$ variables ($\mathbf{x}$ has now $n+m$ components). The first n variables are called *nonbasic variables*, the other m variables are referred to as *basic variables*. The *basic solution* is obtained by setting $(x_1, \ldots, x_n) = 0$, which yields for the remaining m variables $(x_{n+1}, \ldots, x_m) = \mathbf{b}$. If the non-negativity constraints of (2.28) are fulfilled (if $\mathbf{b} \geq 0$), this is a basic *feasible* solution and therefore an initial vertex has been obtained. Next, the algorithm computes the objective value for this vertex and checks whether this objective value can be improved (with respect to the optimization goal) by increasing some non-basic variable. If this is not the case, the current vertex is optimal. Otherwise, the increase from the non-basic variable is determined by considering the maximum value of the variable with respect to the constraints. Then, the considered variable is set to its maximum and becomes a basic variable in exchange for some previous basic variable. This iteration step is called a *pivot step* and corresponds to traversing an edge of the polytope to the next vertex. After each pivot step the new vertex is obtained by setting all non-basic variables to zero. These pivot steps are repeated until no further improvement in the objective function can be achieved by increasing a non-basic variable. The result is an optimal vertex. Figure 2.12 illustrates this algorithm for the example given in (2.28), representing the LP as *simplex tableau* [45].

A crucial property of the simplex method is that although, in general, it is an exponential-time algorithm, in practice it is a very efficient approach, in most cases with the computation time proportional to $n + m$. Although more recent polynomial-time algorithms for linear programs exist (the ellipsoidal method [47]), it still appears that in practice the simplex method is one of the easiest and fastest approaches [39].

The vertex based approach gives way to several versions of the simplex algorithm, including the *dual* simplex algorithm and the *primal dual* simplex algorithm [45]. All of these algorithms are based on the common technique of determining the necessary quantities during the calculation process without explicit use of the whole coefficient matrix A.

2.4.2 Integer Programming

In *integer programming* the variables are required to be integer values. These optimization problems often result from modeling the reality, as the variables might be actual products such as cars or light bulbs. Apart from optimization problems where all variables are required to be integer, it might also be the case that only some variables are required to be integer. In these cases the optimization problem is referred to as *mixed integer programming* problem. In the following only linear objective functions and constraints are considered.

The general form of an integer linear programming problem is as follows:

$$\begin{aligned} \max \quad & \sum_j c_j \cdot x_j \\ \text{s. t.} \quad & \sum_j a_{i,j} \cdot x_j \leq b_i \qquad \forall i = 1, 2, \ldots, m \\ & x_j \text{ integer} \qquad \forall j \ . \end{aligned} \tag{2.29}$$

Alternatively, the problem of (2.29) can also be formulated in matrix form or in the canonical form and the transformation of any form to the other one is straightforward.

Solving an integer linear program turns out to be rather difficult. An intuitive approach is to first solve the corresponding linear program (*relaxing* the integer constraints) by the use of simplex algorithm and then further process the (possibly) non-integer solution, for example by rounding the result to the next integer value. While in certain cases this might yield an optimal solution, in general there are serious limitations to this approach. It turns out that the procedure of rounding to the next *feasible* solution becomes quite difficult, as illustrated in Figure 2.13.

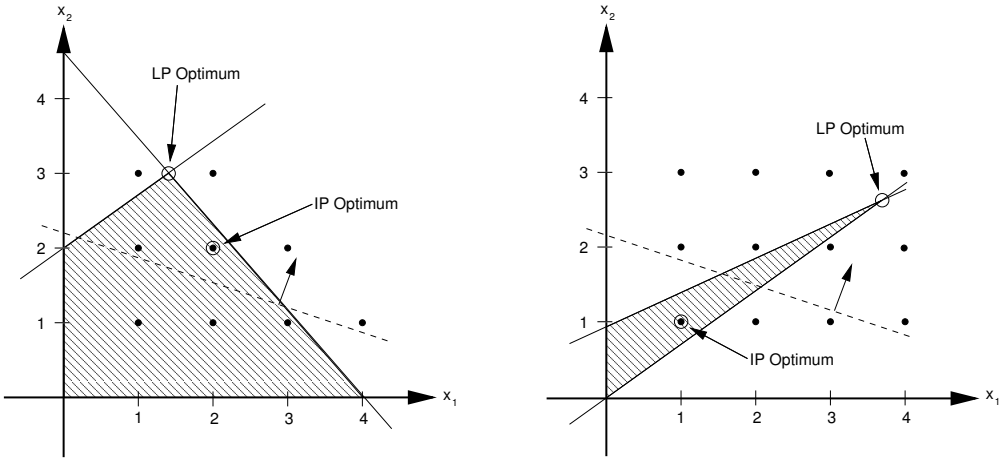


Figure 2.13: Two examples of an integer programming problem. Left: IP for which rounding is rather easy. – Right: IP for which rounding is difficult.

Instead of rounding, a method referred to as *branch and cut* (B&C) is a widely implemented algorithm to solve an integer programming problem. In B&C, initially the optimal relaxed solution is generated. In case that one of the variables, which are required to be integer, are fractional in the relaxed solution, this variable is "branched on" by constructing two polytopes out of the initial single one. All fractional solutions of this chosen variable are excluded, as shown in Figure 2.14. Then, the algorithm continues by evaluating the relaxed solutions for the two new polytopes and may have to perform the branching operation again. The resulting data is stored in a tree where the original polytope is the root node (together with the relaxed solution). Each node in the tree might have two children.

Clearly, by the branching operation the number of polytopes increases quite fast to a high number. However, if an integer solution has been found for some "child-polytope" which has

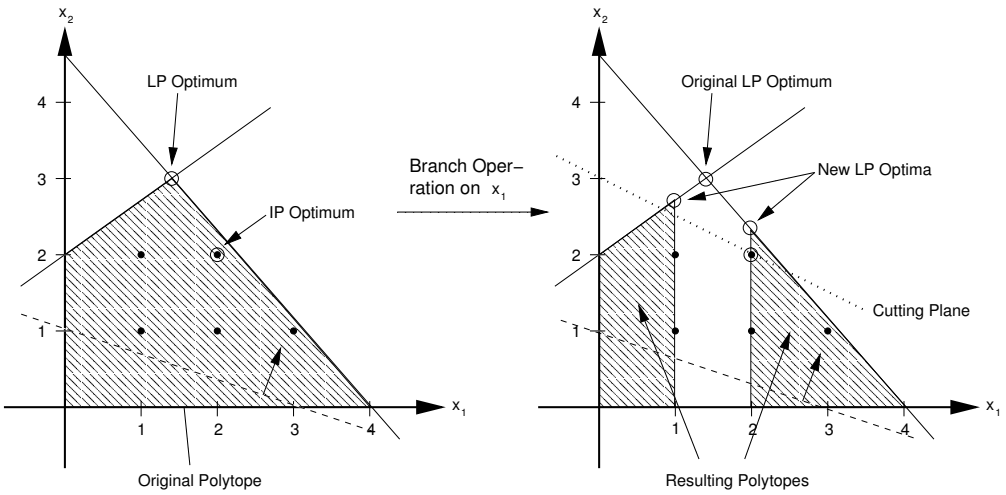


Figure 2.14: Example of the branching operation of the B&C algorithm.

a higher objective value than the relaxed solution for other "child-polytopes", all feasible solutions of the corresponding "child-polytopes" can safely be excluded. This reduces the number of "child-polytopes" again, once such feasible solutions are found. Essentially, this method increases the lower bound on the maximum objective function.

In contrast, the "cut" part of the B&C algorithm decreases the upper bound on the optimum. This is done by obtaining additional constraints on the original polytope which "chop" off certain parts of the relaxed polytope without changing the integer hull of the original polytope. These additional constrains are referred to as *cutting planes* (see Figure 2.14). They reduce the upper bound on the integer optimal solution and therefore help closing the gap between the best integer objective value and the relaxed upper bound.

Despite this and several other sophisticated approaches, integer linear programming problems remain to be rather difficult in general. In the worst case, enumeration of *all* feasible integer solution remains the only way to obtain the optimum. Sometimes the integer programming problem has a special structure which can be exploited to construct a special solution algorithm, for example, based on graph theory. These special cases are covered by what is called *combinatorial optimization* [46]. For example, the (weighted) matching problem in bipartite graphs is such a special integer programming problem.

Chapter 3

Related Work and Scope of the Thesis

The field of dynamic OFDM systems has been under investigation for about ten years so far. During this time span, two main directions have been considered: Dynamic OFDM for point-to-point communications (i.e. communications between a single transmitter and receiver) and dynamic OFDM for point-to-multi-point communications (between one transmitter and several receivers, as in the down-link for example), such as the down-link of a wireless cell. In the first case the dynamic approach is often referred to as (single user) *bit-* and/or *power loading* for OFDM systems [48] while the latter approach is often referred to as (dynamic) **M**ulti**u**ser-**OFDM** (*MU-OFDM*) system [49]. In this section, the state of the art regarding these two directions is presented.

3.1 Dynamic Schemes for Point-to-Point Communications: State of the Art

Consider the following scenario (see Figure 3.1): Data is to be conveyed from a transmitter to a receiver. A certain bandwidth of B [Hz] is available for this task. OFDM is applied as transmission scheme: The bandwidth B is divided into N sub-carriers, each with a bandwidth of $\frac{B}{N}$. Assume that at the transmitter the power for each sub-carrier can be varied separately. These power *assignments* at time t are denoted by $p_n^{(t)}$. At time t, each sub-carrier n experiences an attenuation of $h_n^{(t)}$. These sub-carrier attenuations are known to the transmitter.

Loading algorithms for such point-to-point communications adapt the transmit power (power loading) and/or the modulation type (bit loading) per sub-carrier. As input, the loading algorithm takes the sub-carrier attenuations and generates the power and/or bit assignments from them. Loading algorithms are always driven by improving some system metric, for example, decreasing the overall transmit power (with the constraint of providing some fixed data rate) or increasing the overall data rate (with the constraint of a fixed transmit power budget). In addition, the loading algorithm is based on a (functional) relationship between spent transmit power, obtained data rate and current attenuation of the sub-carrier (referred to in the following as power-rate function). It is obvious that the underlying assignment problems solved by loading algorithms are optimization problems. In the next section, the several modeling approaches are

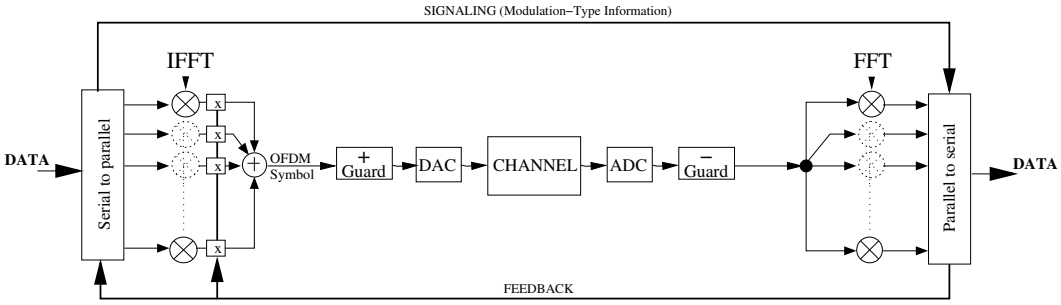


Figure 3.1: Transmission sketch of a point-to-point dynamic OFDM communication scheme: While the transmitter requires a feed-back regarding the channel attenuations, the receiver requires additional information regarding the applied modulation types per sub-carrier.

presented.

3.1.1 Optimization Models and Algorithms for Point-to-Point Communications

Initially, consider the Shannon capacity of an OFDM system (defined further below). To start with, consider the receiver SNR of sub-carrier n at time t. It is defined as the following:

$$v_n^{(t)} = \frac{p_n^{(t)} \cdot \left(h_n^{(t)}\right)^2}{\sigma^2} \quad . \tag{3.1}$$

Clearly, the SNR depends on the transmit power $p_n^{(t)}$, the current attenuation $h_n^{(t)}$ and the noise power level. From the transmitter point of view, the SNR per sub-carrier can only be influenced by the transmit power. Assume in particular that the transmitter knows the current attenuation $h_n^{(t)}$. Then, the transmitter can easily calculate how much transmit power to put on sub-carrier n such that the received SNR is equal to $v_n^{(t)}$:

$$p_n^{(t)} = \frac{v_n^{(t)} \cdot \sigma^2}{\left(h_n^{(t)}\right)^2} \quad . \tag{3.2}$$

Next, consider the Shannon capacity (see Figure 3.2) of this sub-carrier n, assuming (for a moment) that the transmit power and the attenuation are fixed. It is defined as [35]

$$C_n = \Delta f \cdot \log_2\left(1 + v_n^{(t)}\right) = \Delta f \cdot \log_2\left(1 + \frac{p_n^{(t)} \cdot \left(h_n^{(t)}\right)^2}{\sigma^2}\right) \quad . \tag{3.3}$$

The Shannon capacity of a channel denotes the highest data rate, for which the bit error probability can be reduced arbitrarily (by means of error correction coding). If a higher data rate than the capacity rate is conveyed, an irreducible amount of bit errors will remain in the information

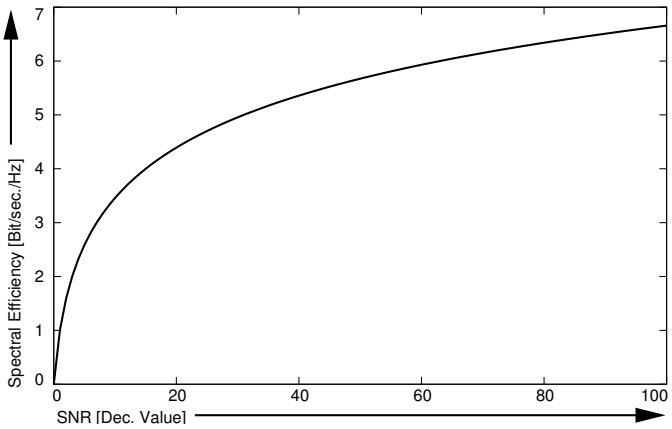


Figure 3.2: Shannon capacity curve for an increasing SNR: Depicted is the spectral efficiency, which is the capacity (according to Equation 3.3) divided by the channel's bandwidth.

conveyed, even if the most sophisticated coding schemes are applied. The capacity formula defines a continuous functional relationship between the transmit power and the achieved Shannon capacity (i.e. the maximum error free bit rate). Thus, for any additional Δp of transmit power an corresponding additional ΔC of Shannon capacity is obtained. If the transmitter knows the current attenuation, it can enforce theoretically any capacity on the sub-carrier if the transmit power is not restricted.

Considering now multiple sub-carriers with different attenuations (but still fixed in time), a natural question arises regarding the Shannon capacity of the whole OFDM system: Consider first the simple case where the transmit power is equally distributed per sub-carrier. Then, the Shannon capacity of the system is simply the sum of the capacities of the single sub-carriers:

$$C = \Delta f \cdot \sum_n \log_2 \left(1 + v_n^{(t)}\right) \quad . \tag{3.4}$$

However, if the transmit power can be varied, the maximum error-free rate is not obtained from an equal power distribution, unless the attenuations of all sub-carriers are equal. If the attenuations differ, the Shannon capacity is obtained by solving a maximization problem [2], where the transmit power allocations per sub-carrier are the variables. Note that the overall power budget is limited to the value $P_{\max}$, otherwise the Shannon capacity of the system would not be bounded.

$$\begin{aligned} C = \quad & \max \quad \Delta f \cdot \sum_n \log_2 \left(1 + \frac{p_n^{(t)} \cdot \left(h_n^{(t)}\right)^2}{\sigma^2}\right) \\ & \text{s. t.} \quad \sum_n p_n^{(t)} \leq P_{\max} \quad . \end{aligned} \tag{3.5}$$

Problem (3.5) is a non-linear, continuous optimization problem where the objective function is concave and the feasible solution set is convex. Therefore, it can be solved analytically by

applying the technique of Lagrangian multipliers [50]. The Lagrangian function, obtained from Equation 3.5, is:

$$L\left(\vec{P}^{(t)},\lambda\right) = \Delta f \cdot \sum_n \log_2\left(1+\frac{p_n^{(t)}\cdot\left(h_n^{(t)}\right)^2}{\sigma^2}\right) + \lambda\cdot\left(\sum_n p_n^{(t)} - P_{\max}\right) \quad . \tag{3.6}$$

According to the technique of Lagrangian multipliers, the optimal transmit power distribution is obtained by solving the following equation system:

$$\nabla L\left(\vec{P}_{\text{opt}}^{(t)},\lambda_{\text{opt}}\right) = 0 \Leftrightarrow \begin{pmatrix} \frac{\partial L}{\partial p_1^{(t)}} \\ \vdots \\ \frac{\partial L}{\partial p_N^{(t)}} \end{pmatrix} = \vec{0} \tag{3.7}$$

$$\lambda_{\text{opt}}\left(\sum_n p_n^{(t)} - P_{\max}\right) = 0 \quad . \tag{3.8}$$

After some transformations this equation system yields the (optimal) power distribution given by Equation 3.9.

$$p_n^{(t)} = \frac{1}{N}\left(\sum_i \frac{\sigma^2}{\left(h_i^{(t)}\right)^2} + P_{\max}\right) - \frac{\sigma^2}{\left(h_n^{(t)}\right)^2} \tag{3.9}$$

The lower the relative attenuation of some sub-carrier is compared to all other attenuations, the more transmit power this sub-carrier will receive. This analytical expression is in accordance to the well known water-filling solution [35] (see Figure 3.3). Water-filling applies originally to a continuous[1] attenuation function of the channel and provides the Shannon capacity of such a channel. It is assumed that the transmit power can be adapted on an infinitely small granularity. Thus, a continuous power function is obtained from the water-filling solution. The Shannon capacity is achieved by *adapting* the transmit power to the transfer function. Roughly speaking, given a fixed power budget, more power is applied to frequency areas with a lower attenuation compared to the other frequencies. Correspondingly, less power is applied if the attenuation is higher. This optimal power distribution is similar to inverting the transfer function and pouring a liquid (water), i.e. power, into the shape. Hence, the scheme is termed "water-filling" (see Figure 3.3).

Computationally, the result of Equation 3.9 might lead to one significant problem. The analytical solution for the optimal power distribution may yield *negative power allocations* for some sub-carriers. This is obviously not applicable in reality. In order to generate a realizable optimal power allocation for each sub-carrier, the power allocation according to Equation 3.9 is then regenerated, excluding the sub-carriers for which the negative transmit power was obtained.

[1] Continuous refers here to the frequency domain.

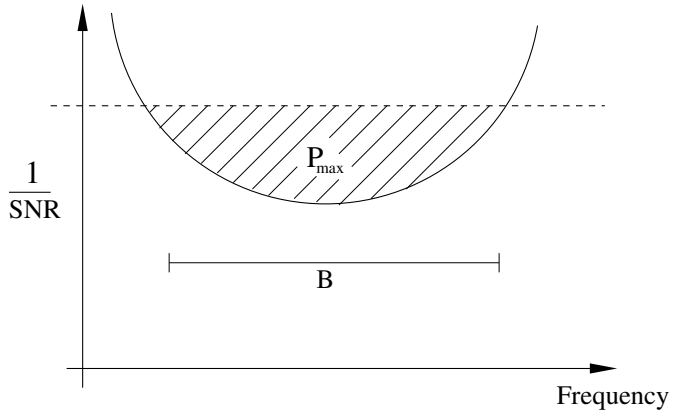


Figure 3.3: Principle of information theory's "water-filling" approach.

However, in the worst case the complete power allocation has to be redone $N - 1$ times until an optimal allocation is obtained (as always only one sub-carrier is obtained with a negative transmit power, which is then excluded).

In practical OFDM transmission systems, there is no technological counterpart to a continuous relationship between the transmit power and the rate, as given by the Shannon capacity formula. In addition, Shannon capacity is a theoretical limit on the rate, assuming infinitely long coding words. In practice, such coding schemes simply cannot be applied. Thus, a different relationship between transmit power and rate is required.

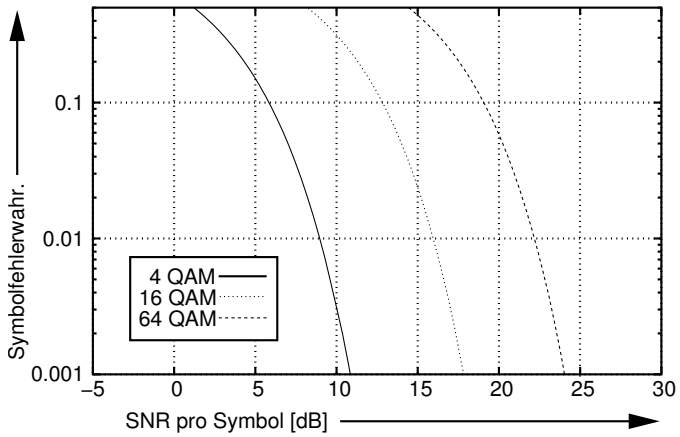


Figure 3.4: Symbol error probability vs. increasing SNR for three different QAM types (4-, 16-, 64-QAM).

In [51], one of the first studies on dynamic power allocation for discrete sub-carrier amounts, an analytical relationship for a certain transmission scheme, referred to as quadrature amplitude modulation [2], was used. This relationship is derived from the symbol error probability function (see Figure 3.4), as defined below:

$$p_{\mathrm{sym}} = 1 - \left(1 - 2 \cdot \left(1 - \frac{1}{\sqrt{M}}\right) \cdot Q\left(\sqrt{\frac{3}{M-1} \cdot v(t)}\right)\right)^2 \tag{3.10}$$

where p_{sym} is the symbol error probability, $\log_2(M)$ is the number of bits transmitted with each symbol, $v(t)$ is the SNR at time t and Q is a function defined as:

$$Q(x) = \frac{1}{\sqrt{(2\pi)}} \cdot \int_x^\infty e^{\frac{-y^2}{2}} dy \quad . \tag{3.11}$$

Equation 3.10 can be transformed such that the resulting bit rate (M) is obtained, depending on the transmit power. However, it is important to note the difference between the functional relationship in Equation 3.10 and the relationship given by the Shannon capacity formula of Equation 3.3. For QAM systems, as for any practical modulation system, the (symbol or bit) error probability enters the equation. Thus, the relationship between the SNR and the achieved bit rate also depends on the target error rate of the system. If a low error probability is of interest, the transmitter has to spend more power to achieve a certain bit rate. In the following, this functional relationship between achieved bit rate, transmit power (or SNR respectively) and the error probability is referred to as the *power-rate function* $F(v(t), p_{\mathrm{err}})$. An example power-rate function is shown in Figure 3.5 for three different QAM types and a target symbol error rate of 10^{-1} and 10^{-2}.

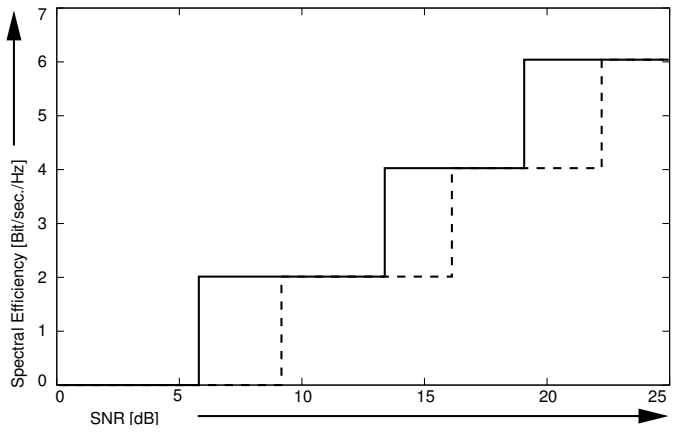


Figure 3.5: Power-rate function resulting if three different QAM types are available (4-, 16-, 64-QAM) for two different target symbol error probabilities 10^{-1} (solid line) and 10^{-2} (dashed line).

From this it follows that three different metrics can be considered for the optimization of practical systems (see Figure 3.6): Maximizing the data rate for a given power budget and a target bit error probability (called the bit rate maximization problem) [51], minimizing the transmit power for a certain given rate and a target bit error probability (called the margin maximization problem) [52] and minimizing the bit error rate for a given bit rate and power budget [48].

The principle challenges regarding loading algorithms are twofold: How to generate the optimal power allocation or at least a considerably "good" solution, and how to implement such a scheme in a real system.

Consider the following, general formulation of the bit rate maximization problem:

$$\begin{aligned} \max_{\vec{P}^{(t)}} \quad & \sum_n F\left(\frac{p_n^{(t)} \cdot \left(h_n^{(t)}\right)^2}{\sigma^2}, p_{\text{err}}\right) \\ \text{s. t.} \quad & \sum_n p_n^{(t)} \leq P_{\text{max}} \end{aligned} \tag{3.12}$$

A practical OFDM transmission system might have M different modulation types per sub-carrier. By specifying a target bit error rate of the system, a certain required SNR can be determined from Equation 3.10 for each modulation type (see Figure 3.5). This can be precomputed prior to the application of the system. If the transmitter knows the current attenuations of the sub-carriers, a required power level is obtained from the precomputed SNR values. Ultimately, the transmitter has to decide which of the $M + 1$ power levels[2] it assigns to each sub-carrier in order to obtain the maximum rate at the target bit error rate. Essentially, this becomes an integer programming problem [53] (Section 2.4.2). In general, integer programming problems are difficult to solve. Although the amount of possible solutions is finite (each sub-carrier might be allocated one of $M + 1$ power levels, therefore there are $(M + 1)^N$ possible power allocations), finding the optimal solution might remain a difficult task, often requiring a brute force enumeration of all possible solutions.

An iterative approach to obtain the optimal power allocation to each sub-carrier has been patented by Hughes-Hartogs [54]. The principle of this algorithm is quite simple, in fact it is based on a greedy strategy (see Algorithm 1): For each sub-carrier, the amount of power required to transmit data with the lowest modulation type is calculated. Then, the sub-carrier which requires the least amount of power for transmitting an additional bit at the target error probability is selected, the amount of power is allocated to it and the required additional power for applying the next higher modulation type is calculated for this sub-carrier (while the total amount of available power is decreased by the allocated amount). This algorithm continues in this manner until no more transmit power is available. It determines for a discrete amount of modulation types the optimal power allocation with respect to the target bit error probability while maximizing the data rate. The exact same scheme can also be used to determine the optimal power allocation in order to minimize the transmit power subject to a bit rate constraint. In this case the algorithm simply runs until the target data rate is reached. Also, this algorithm can be used to minimize the bit error rate of the system.

Although the Hughes-Hartogs algorithm does not require an enumeration of all possible power allocations, the required amount of iterations might be very high. For example, assume the M modulation steps to differ by one bit. Then, for transmitting a total of 1000 bits the algorithm will have to perform 1000 iterations. In each step it compares the required additional transmit power to convey one more bit for each sub-carrier and chooses the one with the lowest

[2] A sub-carrier may also be allocated no transmit power at all.

Algorithm: Hughes-Hartogs

Result : Optimal Sub-Carrier Power Allocation

Given an overall maximum power value $P_{\max}$ *and the set of the current attenuation values for each sub-carrier. Then, a transmit power matrix of values* $p_{m,n}^{(t)}$ *is calculated in advance, holding the required transmit power to convey* m *bits on sub-carrier* n*. The optimal power allocation can be obtained using the following procedure:*

Calculate the "incremental power matrix", consisting of values $\Delta p_{m,n}^{(t)} = p_{m,n}^{(t)} - p_{m-1,n}^{(t)}$
Set $P_{tot} = 0$
while $P_{tot} \leq P_{max}$ **do**
    Search row 1 of the incremental power matrix for the smallest $\Delta p_{1,n}^{(t)}$
    if *(*$\Delta p_{1,n}^{(t)}$ *is the smallest)* **then**
        Assign one additional bit to sub-carrier n.
        Increment $P_{tot} = P_{tot} + \Delta p_{1,n}^{(t)}$
        Update column n of the incremental power matrix by $\Delta p_{i,n}^{(t)} = \Delta p_{i+1,n}^{(t)}$ (move all terms of this column up by one position)
    end
end

Algorithm 1: The Hughes-Hartogs Loading Algorithm.

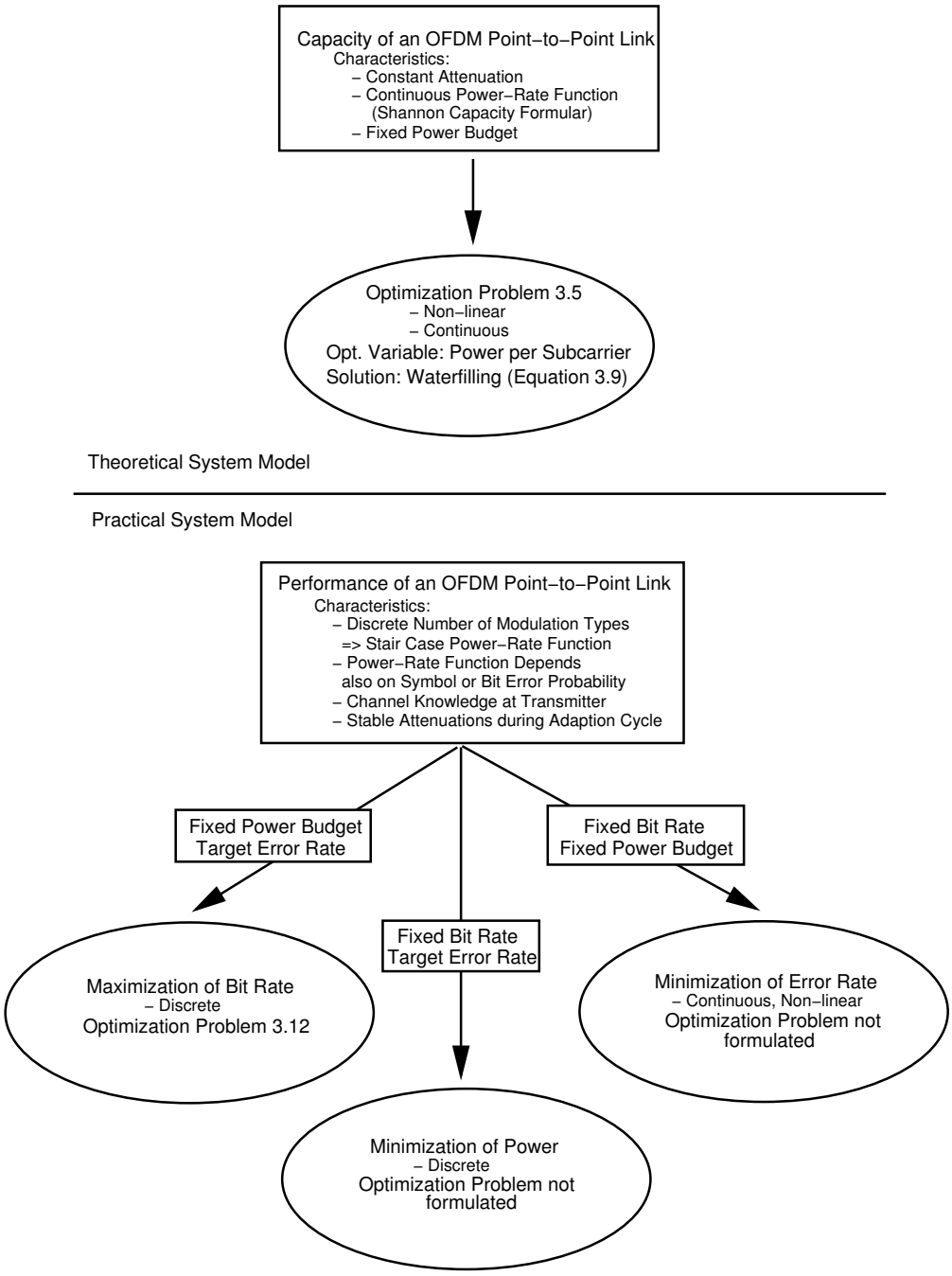


Figure 3.6: Relationship between the different optimization approaches for point-to-point OFDM systems.

required power.

However, the Hughes-Hartogs algorithm provides a good insight to the problem of allocating power to sub-carriers in order to optimize a certain metric. In particular, note that power is only allocated to the sub-carrier which requires the least amount of it in order to transmit one bit. This observation serves as basis for an optimality condition on the power allocations: Assume a certain power allocation $\vec{P}^{(t)}$ is optimal with respect to the bit rate maximization problem. Then, it is not possible to find any two sub-carriers x, y for which the "exchange" of a transmit bit pays off. Denote by $\Delta p^{(t)}_{m+1,x}$ the power required to transmit one more bit on sub-carrier x (as formulated in the chart of Algorithm 1). Thus, if the current power allocation is optimal, there is no pair of sub-carrier x, y for which the following condition holds: $\Delta p^{(t)}_{m+1,x} < \Delta p^{(t)}_{m-1,y}$. Taking one bit from sub-carrier y does not "free" enough transmit power to put the bit on sub-carrier y (and vice versa). An equivalent statement can be constructed for the minimization of power or for the minimization of the bit error rate. A power allocation which satisfies this constraint has been termed *efficient* in [55, 56].

In the literature, the major focus has been on obtaining a power allocation which can be computed faster than by the Hughes-Hartogs approach while providing a performance as good as possible. Using the optimality condition a given allocation can be checked quite fast for optimality. Note that although most algorithms have been constructed for a specific objective function, they can be "reprogrammed" for any objective function as discussed above (maximization of the bit rate, minimization of the power, minimization of the bit error rate).

Next, two specific approaches of the literature are presented. The first scheme has been one of the first loading algorithms published at all. It does not provide the optimum solution but has been constructed to reach a good suboptimal solution much faster than the Hughes-Hartogs algorithm. The second approach provides the optimal solution, given an allocation of a previous transmission phase.

Chow et al. [52] presented one of the first faster algorithms than the Hughes-Hartogs approach, initially considering minimizing the transmit power. They propose to start with an *equal* power distribution and then alter this distribution in order to reach the required rate. After obtaining the bit rate per sub-carrier with the equal power distribution, they iteratively increase or decrease the transmit power margin per sub-carrier, depending on the difference between currently achieved data rate and target rate (for example, if the total rate is larger than the target, the system can accept a higher noise margin while still providing the target rate, therefore the transmit power for all sub-carriers can be lowered by a certain *factor* for all sub-carriers). After this refinement has been performed for a finite number of times, the bit-and power allocation is finalized with a last adjustment. Compared to the Hughes-Hartogs algorithm, this proposal determines an allocation much faster. However, it provides a suboptimal result (the loss of performance is rather small, as claimed by the authors).

In [57], a different approach is taken by Lai et al. Typically, wireless channels are time variant. Therefore, the power allocation has to be updated periodically in order to keep up with the fading process in time (Section 2.1.1). The coherence time can be viewed as an appropriate update cycle for the power and bit allocations in OFDM systems. In such a case, some sub-carrier attenuations will not have changed much from one update cycle to the next one. Therefore, also the corresponding allocations will be almost the same. This motivates the idea of the authors

to use the power allocation of the previous update cycle for the power allocation of the current one. For each reallocation cycle, the proposed algorithm first sorts the sub-carriers according to their current attenuation. Prior to this, the previous bit allocations are sorted according to their rate. Next, the algorithm allocates the best current sub-carrier (in terms of the attenuation) the highest number of bits which had been allocated in the previous cycle. The second best sub-carrier receives the second highest number and so on. Then, the required power allocation is computed. The algorithm refines its allocation by finding suitable sub-carrier pairs for switching a bit allocation (according to the optimality condition mentioned above). Finally, the optimal bit allocation is found.

There are much more dynamic loading schemes in the literature. For further reading, refer to [55, 58, 48, 59, 60, 56, 61] and the references given therein.

Apart from these *fully* adaptive schemes, there exist also two special cases for the power distribution in OFDM systems. In the first case, the transmit power is simply distributed statically. As the attenuation values might differ strongly, the resulting SNR per sub-carrier varies, too. These varying SNR's motivate the idea to adapt the modulation types solely, referred to as *adaptive modulation* (see Figure 3.7). Given a certain target bit error probability, for each modulation type a SNR range can be obtained for which it is applied (as discussed above). Then for each sub-carrier the modulation type is simply adapted according to the SNR ranges. An excellent, in depth discussion of adaptive modulation for multi-carrier systems is given in [62].

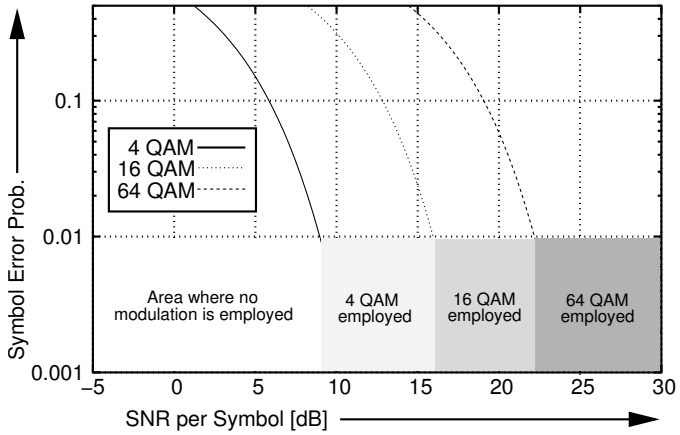


Figure 3.7: Adaptive modulation scheme with three different QAM types (4-, 16-, 64-QAM) and a symbol error probability of 10^{-2}.

The second specific approach to power loading is to vary the transmit power solely while considering an OFDM system with a *single* modulation type. This has been suggested by Hunziker et al. in [63]. As the throughput is fixed in such a case (for each OFDM symbol the same number of bits is transmitted), the objective is to minimize the bit error probability while subject to a total transmit power budget. Since the transmit power is varied whereas only one modulation type is available, the respective optimization problem becomes a non-linear, continuous problem. The authors obtain an analytical expression for the bit error probability of each bit in an OFDM symbol depending on the sub-carrier attenuations and the transmit power per sub-

carrier. Then, using the Lagrange multiplier technique, they obtain an expression for the optimal power allocation, assuming perfect channel knowledge at the transmitter.

3.1.2 System Studies Regarding Dynamic Loading Schemes

In Section 3.1.1, the issue of modeling a dynamic point-to-point OFDM connection as optimization problem has been discussed. Depending on the model, it was shown that the resulting optimization problem has a varying complexity. In addition, some (suboptimal and optimal) loading algorithms were discussed. However, regarding the application in practical OFDM systems, various additional aspects arise, such as the performance gain that is achieved by suboptimal or optimal loading schemes (compared to static schemes). Does it pay off to apply dynamic schemes at all? How close do suboptimal schemes get to the optimal performance? How strong is the impact due to erroneous channel state information? These issues, as well further ones regarding the application of dynamic loading schemes in OFDM systems, are addressed in this section.

Performance Results The most important question relates to the performance gain that can be achieved by dynamic loading schemes. Comparing the optimal data rate in case of the bit rate maximization problem to the rate of an equalized SCM system[3], Willink et al. find in [64] a $5\ \mathrm{dB}$ gain (target bit error probability of 10^{-5}) for a wide range of total transmit power. Hence, according to this study, a SCM system with the same bandwidth requires a three times higher transmit power to achieve the same data rate as an OFDM system which applies optimal dynamic loading (according to the bit rate maximization problem).

In [65], Czylwik compares the performance of an equalized SCM system to the performance of an OFDM system with fixed and adaptive modulation. For four different wireless channels, Czylwik finds an improvement (in terms of bit error probability) of around $2\mathrm{dB}$ for a line-of-sight scenario when switching from the SCM system to the fixed OFDM system (both applying the same modulation type), and a further improvement of around $4\ \mathrm{dB}$ when switching from the fixed OFDM system to the adaptive OFDM system. In case of no line-of-sight, the performance gains are much larger, around $8\ \mathrm{dB}$ for switching from the SCM system to fixed OFDM and around $5\ \mathrm{dB}$ when switching from the fixed to the adaptive OFDM. Similar results advocating adaptive modulation were found by Rohling et al. in [66]. In a further study [67], Czylwik investigates the performance difference (in terms of bit error probabilities) between an OFDM system with fixed modulation type (equal power distribution), adaptive modulation (equal power distribution) and a fully dynamic loading scheme (which is optimal with respect to the considered objective function). For all considered channels (two different ones, relying on measurements) fixed OFDM is significantly outperformed (by about $5\ \mathrm{dB}$ at an bit error probability of 10^{-3}) by adaptive OFDM. However, the difference between adaptive modulation and the loading scheme is rather small, around $1\ \mathrm{dB}$. This indicates that the computational expense related to adapting the bit *and* power allocation according to the loading scheme is not worth the performance

[3] Considering a perfectly equalized SCM system takes the influence of ISI on the bit error rate out of the investigations. However, the channel is still frequency selective, which is exploited by loading algorithm of the OFDM system

gain achieved by it, at least for these channel characteristics studied. Similar results were found by Barreto et al. in [68] for an IEEE 802.11a-like system applying adaptive modulation and a dynamic optimal loading scheme, instead of performing link adaptation (same power and modulation allocation for all sub-carriers) which is implemented in 802.11a systems today. In [60], the authors find that the biggest advantage for a coded, interleaved system is obtained by performing adaptive modulation and not by also adapting the transmit power in a suboptimal way (adaptive modulation achieves at a target bit error probability of 10^{-4} a 3 dB gain versus a static system, while additional power adaption only adds 1 dB more). In [63], the same authors show that soft decision decoding yields a much larger performance gain than adaptive power allocation.

However, it should be mentioned that in the context of DSL systems fully dynamic loading schemes are employed, despite the computational burden. The reason for this might be that in the context of DSL the sub-carrier attenuations are much more diverse (in frequency domain) than in typical wireless transmission systems. In [61], it is mentioned that the attenuation over twisted pair loops might vary by 60 dB and more, which appears to be more than usually observed in wireless systems (as a rule of thumb a 10 dB fade has a probability of 10^{-1}, a 20 dB fade has a probability of 10^{-2} and so on [4]). This indicates that the diversity of the transfer function plays a major role. For the frequency selective fading the diversity depends on the delay spread of the propagation environment. A large delay spread might lead to a situation where power adaption does pay off, however this has never been studied so far.

Furthermore, note that in [56] run time experiments have been conducted for several loading algorithms. These run times are found to be well below 1 ms for a wide variety of sub-carrier numbers.

Channel Knowledge A second issue is the channel knowledge. In practical systems, any channel knowledge is erroneous to some extent. Therefore, especially in wireless channels the question is by how much the gained performance is decreased if realistic channel knowledge is assumed instead of perfect knowledge. In [65], a comparison is presented for different mobile velocities[4] ($0 - 10$ m/s) for uncoded OFDM systems with adaptive modulation at a carrier frequency of 1.8 GHz. The bit error probability curves are compared to the performance with ideal channel knowledge. The results indicate that as long as the mobile velocity is rather small (about 2 m/s) the performance loss is small. However, for a velocity of 10 m/s the performance loss is quite significant. Note that the performance at high velocities is still considerably better than the one achieved by SCM systems with equalization (not considering spread spectrum techniques) at these velocities. In [68], the authors compare the results of uncoded, adaptive modulation based on perfect channel knowledge in IEEE 802.11a systems to the performance of the same adaptive modulation scheme based on channel estimates (assuming a linear minimum mean-square error estimator). No fading variations in time are assumed, the estimation errors stem from estimating the frequency selective transfer function upon training symbol data. In that study, the authors find the performance difference to be around 2 dB at a bit error probability of 10^{-3}. In general this difference is larger for systems operating with a small power budget (low receiver SNR). It is worth noting that the performance loss due to the estimation error is smaller

[4]The velocity of a transmitter or receiver has a strong influence on the selectivity of the channel and therefore on the quality of channel estimates at the transmitter.

in case of link adaption schemes (currently performed in IEEE 802.11a systems). Hunziker et al. develop in [63] also a power adaption scheme assuming outdated channel knowledge at the receiver. With this improved allocation scheme they demonstrate a small performance gain (with respect to the bit error probability at a fixed data rate) compared to equal power allocation, while the power allocation scheme for perfect channel knowledge (introduced above) is outperformed by the equal power allocation in the case of realistic channel knowledge. In [69], Love et al. take the feed-back overhead for loading schemes in FDD systems into consideration and present a scheme to reduce this overhead by applying compression.

Signaling Overhead A third issue discussed in the context of implementing dynamic power and modulation schemes for point-to-point communications is the impact due to signaling. As the modulation type is varied per sub-carrier at the transmitter, the receiver has to be informed of the chosen modulation types for the next allocation cycle in order to decode the sub-carrier symbols correctly. Thus, the transmitter has to inform the receiver. Obviously, this requires system resources and will decrease the performance gain achieved by adaptive power and modulation algorithms. However, it has not been studied so far by how much this affects the performance of such a transmission system.

Some investigations though interpret the potential signaling overhead as motivation to consider countermeasures. Hunziker et al. propose in [63] to only vary the transmit power while applying a single modulation type. Nguyen et al. [70] study the compression gain that can be obtained when applying loss-less compression schemes (run length codes in combination with universal variable length codes) to the signaling information. They find quite good compression gains of about 70%, significantly more than may be achieved by using the Lempel-Ziv compression algorithm [71]. Alternatively, one can consider to exploit the correlation in frequency of sub-carrier attenuations by grouping various sub-carriers together to a sub-band and then allocate power and modulation types adaptively to these sub-bands. This reduces the signaling overhead since only for each group of sub-bands the new modulation type will have to be indicated (instead of for each sub-carrier). This was suggested by Lei et al. in [72]. However, their results indicate that the performance loss due to this method is only small, if the groups are relatively small. For a 20 MHz system with 512 sub-carries and considering indoor (delay spread at about 0.8 μs) and outdoor (delay spread at about 5 μs) propagation environments, a grouping of already eight sub-carriers leads to a severe performance decrease such that the application of a fixed OFDM system performs at the same level. For a grouping of four sub-carriers to one sub-band the performance is decreased by 1 dB compared to an individual sub-carrier loading.

3.2 Dynamic Schemes for Point-to-Multi-Point Communications: State of the Art

Recall the system model that was introduced in Section 3.1. However, now J terminals are present (see Figure 3.8). In particular, consider only the down-link transmission direction of a single cell with no interference. The wireless cell is served by an access point. The bandwidth is split into N sub-carriers and M different modulation types are available. Again, $F\left(v(t), p_{\mathrm{err}}\right)$

denotes the power-rate function (which is the relationship between the transmit power per sub-carrier and the conveyable amount of bits for the next down-link phase at a certain target error probability). As each terminal j might experience a different attenuation for each sub-carrier n, a suitable system description is given by the attenuation matrix $\mathbf{H}^{(t)}$. Each matrix element $h_{j,n}^{(t)}$ represents the attenuation that terminal j experiences on sub-carrier n at time t.

$$\mathbf{H}^{(t)} = \left(h_{j,n}^{(t)} | \forall\, j, n \right) \tag{3.13}$$

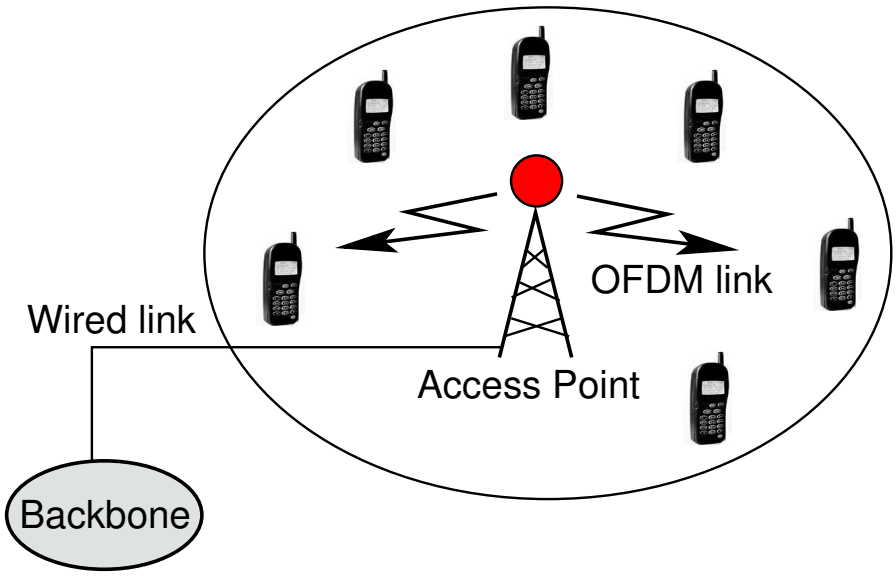


Figure 3.8: A cellular point-to-multi-point OFDM scenario, consisting of one transmitter (base station) and several receivers (terminals).

Obviously, in such a transmission scenario the given system resources (power, bandwidth, time etc.) have to be shared by the terminals. For example, in today's IEEE 802.11a [32] WLAN system (based on OFDM) the resource sharing is implemented as **T**ime **D**ivision **M**ultiplexing (*TDM*) . Time is slotted and, after a specific medium access procedure, the access point exclusively uses all sub-carriers in order to transmit data to some terminal. Thus, the connection becomes a point-to-point connection during each time slot (unless a broadcast frame is transmitted). Dynamic schemes adapting the power and modulation type per sub-carrier could be applied (but are not in todays IEEE 802.11a, instead only link adaption is applied, where *all* sub-carriers are adapted equally).

Another opportunity arises from the diversity of the wireless sub-carriers. As several terminals are located in the cell, sub-carriers which have quite a high attenuation regarding some terminal j are quite likely to have a low attenuation regarding some different terminal i. Thus, the multi-user communication scenario is characterized by a spatial selectivity of the sub-carriers (also referred to as multi-user diversity). The reason for the spatial selectivity is the fact that the fading processes are statistically independent for different terminals, as long as their receive

antennas are separated considerably [5]. Thus, if always all sub-carriers are assigned for a time slot to one terminal (as in TDM systems), sub-carriers with a high attenuation regarding some terminal could probably be better utilized if there is a terminal for which these sub-carriers have lower attenuation. The more terminals there are, the higher is the probability that there is some other terminal with a lower attenuation.

*Dynamic **M**ulti**u**ser-**OFDM** (MU-OFDM) systems* exploit this spatial diversity of sub-carriers. The set of sub-carriers is shared in a **F**requency **D**ivision **M**ultiplexing (*FDM*) fashion. Disjoint subsets of sub-carriers are dynamically assigned to terminals for the next down-link phase. The assignment of sub-carrier subsets to terminals depends on the attenuation matrix $\mathbf{H}^{(t)}$ (or an estimate of this matrix). Periodically, these assignments are recalculated at the access point. While all terminals receive the complete OFDM symbol (with all sub-carrier information in it), each sub-carrier only demodulates the specific sub-carriers it was assigned. Therefore, some form of control channel is required to inform the terminals about their assignments.

The essential property of a dynamic MU-OFDM system is the dynamic assignment of disjoint sub-carrier sets to terminals. A dynamic algorithm performing these assignments is driven by improving some metric. As with the loading algorithms of the previous section, different metrics can be of interest, depending on the given scenario. Apart from the dynamic assignment of sub-carriers, the distribution of the transmit power can be performed by two ways. Either it is also dynamically assigned. Then a joint sub-carrier and power assignment algorithm generates these allocations. Or it is distributed statically (and only the sub-carriers are assigned dynamically).

3.2.1 Optimization Models for Dynamic MU-OFDM Systems

Assume for the beginning that the transmit power is assigned dynamically as well as the sub-carriers. The specific assignment of sub-carrier n to terminal j at time t is a variable of the system. Denote each assignment by the binary variable $x_{j,n}^{(t)}$, where

$$x_{j,n}^{(t)} = \begin{cases} 1 & \text{if } n \text{ is assigned to } j \text{ at } t \\ 0 & \text{if } n \text{ is not assigned to } j \text{ at } t. \end{cases} \tag{3.14}$$

The set of all assignment variables $x_{j,n}^{(t)}$ forms the binary assignment matrix $\mathbf{X}^{(t)}$. The transmit power $p_n^{(t)}$ for each sub-carrier is a variable of the system, as well. Depending on this power assignment for each sub-carrier n, one of the M modulation types is applied during down-link phase t. In such a system, a straightforward optimization approach is to maximize the overall bit rate of the cell per down-link phase, as given in Equation (3.15).

[5] There is no definition for the minimum spacing. However, a spacing of one wavelength is assumed to be sufficient.

$$
\begin{aligned}
\max_{\vec{P}^{(t)},\mathbf{X}^{(t)}} \quad & \sum_{j,n} F\left(\frac{p_n^{(t)} \cdot \left(h_{j,n}^{(t)}\right)^2}{\sigma^2}, p_{\text{err}}\right) \cdot x_{j,n}^{(t)} \\
\text{s. t.} \quad & \sum_{n} p_n^{(t)} \leq P_{\text{max}} \\
& \sum_{j} x_{j,n}^{(t)} \leq 1 \qquad \forall\, n
\end{aligned}
\tag{3.15}
$$

Two general constraints exist for dynamic MU-OFDM systems: While the first constraint limits the overall transmit power (as in the case of problem (3.12) for the power loading), the second constraint is new and is not present in Equation (3.12). It formulates the requirement that each sub-carrier can be assigned to at most one terminal at a time. It has been shown that this requirement yields a significantly superior performance [73][6].

The optimization problem (3.15) is easy to solve by a greedy strategy: Assign each sub-carrier to the terminal with the lowest attenuation [73]. Afterwards a loading algorithm is applied in order to distribute the transmit power with respect to the objective function (for example, applying the Hughes-Hartogs algorithm). However, this approach is not suitable in many cases. Unless all terminals have the same average attenuation, some terminals will have a significantly lower overall attenuation than others (for example, because some terminals are closer to the access point than others). As a consequence, terminals with a high overall attenuation experience high delays for the transmission of packets (as they are seldom assigned a sub-carrier). Therefore, two different problems have been considered in the literature, known as *margin adaptive* and *rate adaptive* optimization problems [74]. These two approaches take the maximum delay within the cell into account.

In the case of the rate adaptive problem, the minimum throughput ϵ of any terminal in the cell is maximized. Thus, the maximum delay in the cell is decreased as much as possible. This problem is given in Equation (3.16).

$$
\begin{aligned}
\max_{\vec{P}^{(t)},\mathbf{X}^{(t)}} \quad & \epsilon \\
\text{s. t.} \quad & \sum_{j} x_{j,n}^{(t)} \leq 1 \quad \forall\, n \\
& \sum_{n} p_n^{(t)} \leq P_{\text{max}} \\
& \sum_{n} F\left(\frac{p_n^{(t)} \cdot \left(h_{j,n}^{(t)}\right)^2}{\sigma^2}, p_{\text{err}}\right) \cdot x_{j,n}^{(t)} \geq \epsilon \quad \forall\, j
\end{aligned}
\tag{3.16}
$$

[6] In this study the authors investigate the assignment of joint sets of sub-carriers to terminals. However, it turns out that exclusive assignments are superior.

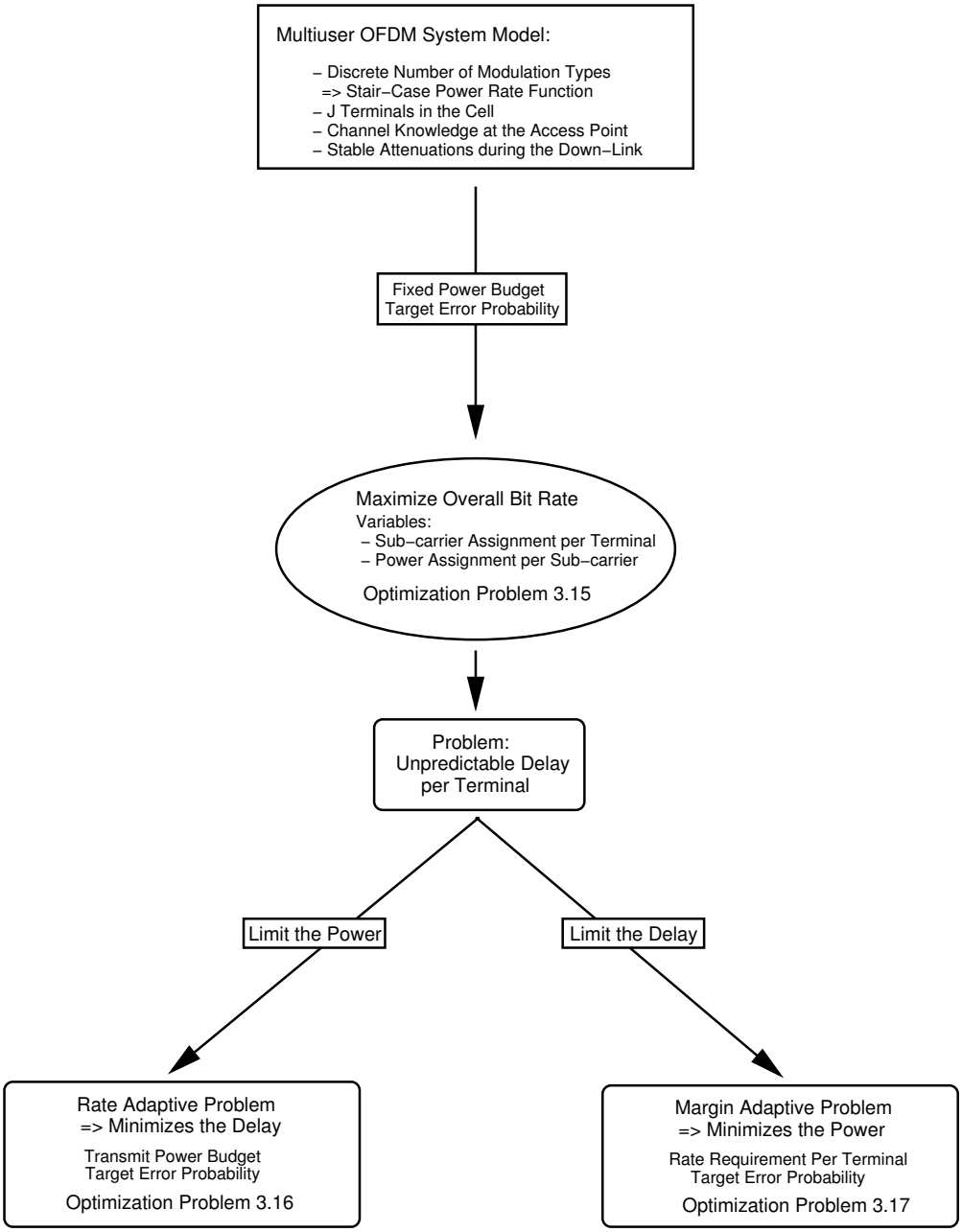


Figure 3.9: Relationship between the different optimization approaches for point-to-multi-point OFDM systems.

In the case of the margin adaptive problem, it is assumed that each terminal j has to be provided a certain throughput, denoted by $r_j^{(t)}$. The objective of this approach is to minimize the transmit power, while achieving the specific rate per terminal per frame. Hence, the delay within the cell is fixed. The resulting optimization problem is given by Equation (3.17).

$$\begin{aligned} \min_{\vec{P}^{(t)},\mathbf{X}^{(t)}} \quad & \sum_{j,n} p_n^{(t)} \cdot x_{j,n}^{(t)} \\ \text{s. t.} \quad & \sum_{j} x_{j,n}^{(t)} \leq 1 \qquad \forall\, n \\ & \sum_{n} F\left(\frac{p_n^{(t)} \cdot \left(h_{j,n}^{(t)}\right)^2}{\sigma^2}, p_{\text{err}}\right) \cdot x_{j,n}^{(t)} \geq r_j^{(t)} \qquad \forall\, j \end{aligned} \tag{3.17}$$

The nature of these two optimization problems depends on the power-rate function (the function $F\left(v(t), p_{\text{err}}\right)$). If the Shannon capacity formula is assumed, the optimization problems are non-linear optimization problems. However, as practical systems usually only feature a limited number of modulation types (with a discrete number of bits transmitted per symbol), $F\left(v(t), p_{\text{err}}\right)$ becomes a stair-case function (as in Figure 3.5). Then, the resulting optimization problem can be formulated as integer programming problem. In the literature, it is stated that solving a non-linear optimization problem (as obtained if the Shannon capacity formula is assumed) is computationally more demanding than solving an integer programming problem [75].

Still, integer programming problems are difficult to solve. Multiple publications [74, 76] claim the rate or margin adaptive optimization problem to be *NP*-hard, however, this has not been proven rigorously. Despite this uncertainty regarding the complexity, it has been shown (see the discussion below) that the performance gain stemming from dynamic sub-carrier assignments is large compared to OFDM systems which assign sub-carriers statically (either in TDM or FDM, not adapting to the sub-carrier attenuations)[7]. Therefore, a major scientific focus within the last years has been on finding efficient algorithms which solve one of the two optimization problems as good as possible.

Apart from jointly assigning sub-carriers and transmit power, certain variants are also considered in the literature. For example, it is also possible to fix the power per sub-carrier and only assign the sub-carriers dynamically to different terminals [77]. Obviously, in this case only a rate maximization, as described formally in problem (3.15) or (3.16), can be pursued. The transmit power cannot be minimized. In a similar manner, it is possible to fix the modulation type per sub-carrier and consider only a dynamic sub-carrier assignment minimizing the transmit power [78], as described in Equation (3.17).

[7] Above, the notion of static systems was defined as systems not reacting to channel variations. In the multi-user case a static system rather refers to a system which does not adapt the sub-carrier assignments to the current channel states. Thus, fixed blocks or time slots are assigned periodically to terminals. On top of these fixed bandwidth resources, an adaptive modulation scheme might be applied. Still, such a scheme is referred to as static system.

3.2.2 Algorithms for Dynamic MU-OFDM Systems

In this section, first a general overview is provided, then various approaches for the rate and margin adaptive approach are presented in detail.

As it is widely believed that generating the optimal solution for the rate or margin adaptive approach is computationally difficult, suboptimal algorithms have been of interest.

One often considered approach is to divide the optimization problem into two parts: First, each terminal is allocated a certain number of sub-carriers. Then, based on this allocation, the sub-carriers are assigned to the terminals, i.e. the specific sub-carrier/terminal pairs are generated. The allocation process can be based on the average attenuation per terminal, for example. In general, it is quite easy to generate the sub-carrier allocations. Once the allocations are available, a second algorithm starts to generate the assignments. If the allocations are available (i.e. if the number of sub-carriers each terminal will receive during the next down-link phase), the resulting optimization problem, known as *assignment problem*, can be solved efficiently. For example, as objective for this problem the overall attenuation of the assigned sub-carriers can be minimized. Because each terminal has to receive a certain number of sub-carriers, the delay of each terminal is to some extent limited. The resulting assignments are optimal regarding the chosen objective function. For illustration purposes, the assignment problem, as discussed above, is given in Equation (3.18). Note that $m_j^{(t)}$ denotes the sub-carrier allocations.

$$\begin{aligned} \max_{\mathbf{X}^{(t)}} \quad & \sum_{j,n} h_{j,n}^{(t)} \cdot x_{j,n}^{(t)} \\ \text{s. t.} \quad & \sum_{j} x_{j,n}^{(t)} \leq 1 \qquad \forall\, n \\ & \sum_{n} x_{j,n}^{(t)} \leq m_j^{(t)} \qquad \forall\, j \end{aligned} \tag{3.18}$$

The problem of Equation (3.18) has a graph-theoretic counterpart: The bipartite weighted matching problem [46]. This optimization problem can be solved by the well-known *Hungarian algorithm* [79], which has a complexity of $O\left(N^3\right)$ [46](note that N equals here the total number of sub-carriers in the system). Often, this algorithm is referred to in the context of dynamic MU-OFDM systems. Other optimal algorithms are discussed in [53], however, the worst-case complexity of $O\left(N^3\right)$ has not been reduced so far.

As second suboptimal scheme, some authors also try to solve the margin- or rate adaptive problem by developing heuristics, mostly based on sorting procedures. Also, some contributions apply the technique of local search to the considered problem. In local search, an initial (hopefully "good") solution is generated which is then iteratively improved by local steps. Therefore, this technique can also be termed a "two-step approach". However, it differs from the two-step approach of allocation and assignment, mentioned above.

Finally, two publications also suggest to relax the integer constraint on the sub-carrier assignments. Thus, each sub-carrier can now be assigned to multiple terminals during one down-link phase. From a system point of view, this might be realized by allowing time-sharing of one sub-carrier during one down-link phase.

In the following, some of the related work is reviewed, considering the problems of (3.16) or (3.17).

Margin Adaptive Approach

In 1999, Wong et al. [49] were the first to face the joint sub-carrier and power optimization of the margin adaptive problem. The authors considered a continuous function $F\left(v(t), p_{\mathrm{err}}\right)$ as relationship between the bit rate achieved and the transmit power spent. Thus, they obtained a continuous optimization problem (as they initially relax the integer constraint on the sub-carrier assignments), which is more tractable but not realizable in practice. Relaxing the integer constraint yields a lower bound to the minimum power needed to transmit the required amount of data. In order to solve the continuous optimization problem, they used standard Lagrange optimization techniques (as mentioned above for problem (3.5)). Applying this method yields solutions that depend on the Lagrange multipliers λ_k, where λ_k can be interpreted as power coefficient of terminal k ($k's$ need for power). In order to find a set of Lagrange multipliers that satisfy the optimization problem constraints, while at the same time delivering optimal solutions, an iterative search algorithm is introduced. This algorithm is the paper's major contribution. As the obtained optimal solution belongs to the non-realistic real value, the authors suggest a suboptimal multi-user adaptive OFDM scheme (MAO), where the sub-carrier *assignment* follows the proposed algorithm (each sub-carrier is assigned to the terminal that holds the greatest share of it), and then a single-user bit allocation (presented in [80]) is applied to each terminal's sub-carrier assignment set. The MAO approach is then compared to various static OFDM approaches and the minimum, relaxed solution. It is shown that – in terms of average bit SNR– the MAO scheme is never worse by more than $0.6\ \mathrm{dB}$ from the optimal solution, whereas the static schemes are 3 to $5\ \mathrm{dB}$.

In a later work [81], the same authors refer to [49] as a solution which is not suitable for time varying channel, as the required iterative computation of the Lagrange multipliers λ_k makes a real-time implementation for time-varying channels impractical. Hence, they present a simpler approach to the margin adaptive problem which is based on predetermining the number of sub-carriers each terminal receives. This results in a bipartite weighted matching problem, for which good algorithms are known to provide the optimal solution (as discussed above). However, the authors claim even this algorithm to be too slow for fast time-varying fading channels and propose a suboptimal solution. This algorithm consists of two phases (the *Constructive Initial Phase* and the *Iterative Improvement Phase*) and belongs to the group of local search algorithms; perfect channel knowledge as well as predetermined numbers of sub-carriers to be assigned per terminal are assumed. In the first phase, the channel conditions for each terminal are evaluated. The sub-carriers are then ordered with respect to their attenuations in descending order (for each terminal). Starting at some terminal, each one gets assigned the individual predetermined number of those sub-carriers that are best for it. If a sub-carrier is already assigned to another terminal, the next best sub-carrier is chosen. Once each user is assigned its individual predetermined number of sub-carriers, the iterative improvement phase begins. Since these assignments result from a greedy strategy, the assignment might be trapped in a local minimum. Hence, pairs of terminals are selected to exchange sub-carriers in a way that the overall required transmit power decreases. To find appropriate pairs, all possible pairs of terminals are examined

regarding possible power savings that result from switching sub-carriers between them. Once all possible sub-carrier switching values are calculated for each terminal-pair, the greatest value is chosen and the appropriate switching is performed. Then all power saving values are updated and the process is repeated. If the greatest value turns out to be negative, switching is stopped. The authors present simulation results which show that this suboptimal algorithm performs almost as good as the optimal solution to the margin adaptive problem with predetermined sub-carrier amounts per terminal, while at the same time delivering assignment values up to 70 times faster than the optimal one based on the Hungarian algorithm. Compared to static OFDM schemes, the greedy dynamic technique is shown to outperform them by at least $10\ \mathrm{dB}$ in terms of the necessary bit SNR for a certain target bit error probability.

Algorithm: Resource Allocation Algorithm BABS

Result : Optimal Resource Allocation (number of sub-carriers) among Users

Given the individual terminal rate requirements $r_j^{(t)}$, the maximum rate per sub-carrier R_{max}, the attenuation matrix $\mathbf{H}^{(t)}$ and a power-rate function $F^{-1}(\cdot)$ (delivering the transmission power, according to the number of bits to send, the minimum BER requirements and available coding/modulation schemes), let terminal j be allocated $m_j^{(t)}$ sub-carriers at time t as follows:

Calculate the initial allocation number per terminal j: $\quad m_j^{(t)} = \lceil \frac{r_j^{(t)}}{R_{max}} \rceil \quad \forall\, j$

while $\sum_j m_j^{(t)} > N$ **do**

  Search for $j^* = arg\min_j m_j^{(t)}$

  Set $m_{j^*}^{(t)} = 0$

end

Calculate the average sub-carrier attenuation per terminal j: $h_{j,avg}^{(t)}$.

while $\sum_j m_j^{(t)} < N$ **do**

  Calculate the difference in transmission power needed when an additional sub-carrier is allocated to terminal j: $\Delta p_j^{(t)} = \frac{m_j^{(t)}+1}{h_{j,avg}^{(t)}} F^{-1}\left(\frac{r_j^{(t)}}{m_j^{(t)}+1}\right) - \frac{m_j^{(t)}}{h_{j,avg}^{(t)}} F^{-1}\left(\frac{r_j^{(t)}}{m_j^{(t)}}\right) \quad \forall\, j$

  Search for $j^* = arg\min_j \Delta p_j^{(t)}$

  Set $m_{j^*}^{(t)} = m_{j^*}^{(t)} + 1$

end

Algorithm 2: The BABS Algorithm.

Wong's greedy algorithm is often referred to and serves as comparison basis for many follow-up algorithms. One such approach was presented in 2000 by Kivanc et al. in [82]. In this case the problem is solved by dividing it into two steps: *Resource Allocation* and *Sub-Carrier*

Algorithm: Sub-carrier Assignment Algorithm ACG

Result : Near Optimal Distribution of Sub-Carriers among Users

Given the number of sub-carriers allocated $m_j^{(t)}$ and the set of sub-carriers allocated A_j for each terminal j (where $\#A_j$ denotes the cardinality of the set A_j), determine the sub-carriers assignments:

1 Initialize: $A_j = \{\} \quad \forall\, j$
for n=*1:*N **do**
2 Search for $j^* = arg\min_j h_{j,n}^{(t)}$
while $\#A_j = m_j^{(t)}$ **do**
3 Gate the for this terminal: $h_{j,n}^{(t)} = 0$
4 and search again: $j^* = arg\min_j h_{j,n}^{(t)}$
end
5 Set $A_{j^*} = A_{j^*} \cup \{n\}$
end

Algorithm 3: The ACG Algorithm.

Assignment. Resource Allocation (determining the number of sub-carriers each terminal should receive) is done using the greedy descent BABS algorithm (cf. Algorithm 2): Based on each terminals average SNR over all sub-carriers, a terminal j gets assigned the number of sub-carriers it needs to reach its minimum transmission rate. Once all rate requirements are fulfilled, the remaining sub-carriers are given to terminals with lower SNR-values. As long as the overall number of sub-carriers is too small to fulfill all data rate requirements, the terminals with the smallest initial sub-carrier allocation values are omitted. Once the number of sub-carriers to be assigned is determined for each terminal, the specific assignment of the sub-carriers is done by the Amplitude Craving Greedy (ACG) algorithm (cf. Algorithm 3): Each sub-carrier is given to the terminal with the highest gain on that sub-carrier, unless the number of sub-carriers so far given to that terminal equals the number of sub-carriers to be assigned (previously determined by the BABS algorithm). Simulations show that the power requirements of the combination BABS/ACG are only marginally higher than the power requirements of the optimal Lagrange approach, while at the same time both greedy algorithm combinations outperform the Lagrange approach in terms of CPU runtime.

In a later journal-paper version [83], the authors deploy algorithm-simulations to two different channel models (highly correlated vs. uncorrelated channel profiles, i.e. adjacent sub-carriers do or do not face quite similar channel attenuations). The results are compared to a single-step frequency allocation[84] that was introduced by Rohling et al. as a simple centralized frequency assignment algorithm. In this scheme, basically for each terminal a subset of qualitatively "good" sub-carriers is selected. No detailed specification of the algorithm is provided, though. It is shown that the algorithms presented yield lower outage probability and lower power requirements at reasonable complexity in both channel scenarios.

Based on that work, Kim et al. propose another adaptive sub-carrier allocation (ASA) algorithm in [85]. The algorithm is a local search algorithm, consisting of two phases, the *Initial Sub-Carrier Allocation (ISA)* and the *Residual Sub-Carrier Allocation (RSA)*, where the ISA phase is based on the BABS algorithm. However, instead of permanently excluding terminals whose requirements are fulfilled – as it is done in ACG – all terminals are reconsidered, once the minimum number of sub-carriers is assigned to each terminal and the RSA phase starts. The combination of both parts – ISA and RSA – has a complexity that is similar to the complexity of the ACG algorithm and is lower than most of the other ASA algorithms such as MAO in [49], while at the same time being almost as good as the MAO algorithm and clearly better than the ACG algorithm in terms of average BER vs SNR. Additionally, a block-wise processing of the ASA is proposed, in case the sub-carrier spacing is smaller than the system's coherence frequency. Given that the coherence bandwidth is m times greater than the sub-carrier spacing, blocks of m adjacent sub-carriers can be grouped and allocated as a whole, which decreases the required computational effort.

A different approach can be found in [86] which follows the idea of separating the problem of sub-carrier assignments into two stages. Pfletschinger et al. present a framework that is based a former approach in [87]. They introduce a low-complexity two-step algorithm. In addition to the rate requirements always considered in margin adaptive problems, individual maximum power constraints per terminal are considered. These enable the algorithm to be also applied in the up-link direction, where different terminals have different maximum transmit power values

due to the resulting CCI. As in [82], the initial number of sub-carriers per terminal is computed by dividing the required data-rate by the maximum number of bits per sub-carrier. The authors claim that usually there are more sub-carriers available than initially assigned. Therefore, in the next step, the remaining sub-carriers can be distributed among the terminals in order to meet their individual power constraints. If there are still sub-carriers left when all terminals meet their power requirements, the maximum power values are lowered and the algorithm is re-applied. For the computation of the power levels that match a certain constellation, the authors assume linear and time-invariant channels. In the second step the sub-carriers are assigned according to the individual **C**hannel–to–**N**oise **R**atio (*CNR*) values: Each terminal chooses the available sub-carrier with the best CNR. The terminal that is allowed to choose next is determined by terminal specific priority values. A terminal's priority is computed by dividing the number of sub-carriers still to assign to that terminal by the overall number of sub-carriers that is still to assign to all terminals. In their simulations, the authors compare their algorithm to the results of the iterative algorithm based on Lagrange optimization by Wong et al. in [49]. The power requirements were chosen to be large in order to be able to compare both approaches, where only the two-step algorithm considers individual power constraints. It is shown that for this scenario the two-step algorithm assigns only little additional power to the terminals, while at the same time being up to a thousand times faster regarding the computation speed.

In [88], Zhang et al. discuss the optimal performance of the margin adaptive problem applied to a more realistic system model. In this work, the authors assume a single modulation type to be available. The link layer at the access point comprises of two parts: A packet scheduler and a resource allocation "layer". Each terminal in the cell receives a data stream consisting of packets. The packets are queued separately per terminal at the access point and are passed to the resource allocation "layer" depending on a certain scheduling policy. The objective of the resource allocation "layer" is to minimize the transmit power while conveying all packets during one down-link phase. In order to achieve this, different sub-carriers can be allocated to different terminals at *different times during one down-link phase* (apart from allocating a different transmit power to different sub-carriers in order to meet a target bit error probability). Thus, each OFDM symbol is interpreted as sub-slot of a down-link frame such that each sub-carrier can be assigned to multiple terminals during one down-link phase. For each down-link phase a control slot is reserved for transmitting the assignment informations for the next phase. The authors formulate the problem as integer programming problem and consider in their simulations only the optimal performance. This performs quite well compared to a non-adaptive scheme (not specified further) in terms of system throughput and average packet delay at a varying average transmit power.

Further schemes for the margin adaptive problem can be found in [75, 89].

Rate Adaptive Approach

The first ones to consider efficient algorithms solving (3.16) were Yin et al. in [76]. The authors formulate the rate adaptive problem as non-linear optimization problem. The power-rate function is not specified directly, but modeled via a general monotonic function which can be inverted. Next, the authors claim the problem to be a *NP*-hard problem without proof. As they point out, the pure rate maximization (without requiring each terminals rate to be above a certain

limit) equals a single-user water-filling problem which is easy to solve. In order to solve the rate adaptive problem, the authors suggest a two-step approach, which is often referred to in later work. In the first step they propose to perform the *resource allocation* (see Algorithm 4), meaning the allocation of power and sub-carrier amounts to all active terminals in the current system. They propose to base the allocation on the average channel to noise ratio of each terminal across all sub-carriers. The exact algorithm is based on the assumption that the amount of sub-carriers allocated should be proportional to the amount of assigned power (relatively to the total amount of sub-carriers and transmit power). After allocating each terminal one sub-carrier, the corresponding power assignment per terminal is calculated. Next, if the total amount of power does not correspond to the allocated sub-carrier ratio, sub-carriers are continued to be assigned to the terminals until either all sub-carriers are assigned and the mentioned power relationship cannot be fulfilled or the power relationship is fulfilled. In the first case the system is then assumed to be overloaded and some terminals have to be dropped. Otherwise, the second step of the two-step approach is performed. This is called the sub-carrier assignment. In this step, depending on each sub-carrier's attenuation towards each terminal, an assignment is performed such that the chosen sub-carriers are optimal for each terminal (i.e. sub-carriers with a lower attenuation – thus in a better state – cannot be assigned to any terminal without significantly "hurting" a different terminal). This assignment problem is known to be efficiently solvable by the Hungarian algorithm [79], which is proposed in this paper for this problem. Once the assignment has been performed, each sub-carrier set is bit and power loaded with the transmit power determined in the first step for each terminal. After discussing this algorithmic solution, the authors compare their system approach with the optimal solution to the rate adaptive problem and also to a static approach. They find that the optimal performance is about twice as large as in the case of the static approach while their suboptimal solution achieves 90% of the optimal performance and therefore outperforms the static scheme significantly. Interestingly, they do not find any increase of the cell's capacity due to multi-user diversity. No run time analysis is presented. In order to perform the allocation process of the first step, the terminals have to specify for each data stream the expected data rate. However, as average and instant rates differ for many known application types significantly (for example, for VBR streams) this allocation scheme might lead to rate over- or underprovisioning. In addition, if the terminals cannot specify any rate requirement, this allocation scheme cannot be performed. The allocation process basically allocates power and sub-carrier resources according to the margin adaptive approach and not according to a rate adaptive approach.

Rhee et al. came up with another approach to increase the throughput of multi-user OFDM systems using dynamic sub-carrier allocation [77]. The authors consider the single cell rate adaptive problem. Initially, they introduce the optimization problem, considering Shannon capacity as power-rate function. Next, the optimization problem is relaxed (the integer constraint on the sub-carrier assignments is dropped) which leads to a convex optimization problem. The authors claim that in most cases only a small number of sub-carriers are shared, despite the relaxation on the assignment of sub-carriers (sub-carriers might be assigned now to multiple different terminals simultaneously). However, the authors claim the optimization problem to be too complex for consideration in practical systems. Therefore, they present a quite simple heuristic based on a *constant* power assignment for all sub-carriers (see Algorithm 5). The au-

thors reason that a constant power distribution does not reduce the systems performance too much due to the fact that most assigned sub-carriers will be in quite a good state. For single user water-filling solutions it is known that the achieved performance (i.e. capacity) of assigning each sub-carrier equal power (as long as only quite good sub-carriers are considered at all) is almost identical to the one achieved if an optimal (variable) power distribution is applied (as discussed in Section 3.1). It remains open if this also applies to multi-user scenarios. Finally, the authors present simulation results for an increasing transmit power, comparing a static OFDM approach to the optimal performance and the presented suboptimal algorithm. They find that the dynamic MU-OFDM performance is around $2\ \mathrm{bits/s/Hz}$ better than the performance of the static scheme, while the optimal and suboptimal dynamic schemes do not differ significantly. This performance gain of the dynamic scheme versus the static one increases as the number of terminals in the cell increases (about 100% gain for 10 terminals in the cell). The authors explain this behavior with the increasing multi-user diversity in the system.

Kim et al. demonstrate in [74] how the rate adaptive (and also margin adaptive) optimization problem (assuming a non-linear, continuous power-rate function) can be converted into a integer programming problem. The *general* rate adaptive optimization problem is nonlinear due to the nonlinear power-rate function. For this function the Shannon capacity formula can be used, which is obviously nonlinear. Alternatively, there are formulas for the family of M-QAM (see Section 3.1.1) which relate transmit power, error probability and bit rate to each other. However, these formulas are also nonlinear. In contrast, practical systems will consist of a finite number of possible modulation types. Instead of dealing with a two-dimensional assignment matrix (one dimension for the sub-carriers, one for the terminals), it is possible to deal with a three-dimensional assignment matrix, the third dimension representing the applied modulation type. Therefore, the resulting expression for the rate achievable per terminal with any assignment choice becomes a simple summation of binary values multiplied by a matrix holding the bit values for each possible assignment. This is a (linear) integer combination of discrete variables. Thus, as the authors claim, the nonlinear optimization problem becomes a linear integer optimization problem. However, it remains difficult due to the integer constraint. Next, the authors propose a suboptimal solution method for this problem. They propose to allocate each terminal an equal amount of sub-carriers and then perform an assignment. For this process the power per sub-carrier is fixed. After the assignment has been performed (which is basically an assignment as problem (3.18), however, the authors propose to use a slightly different objective function), a loading scheme is applied (in fact, they propose one scheme for the margin adaptive problem and one for the rate adaptive problem). Then, the authors show simulation results of their suboptimal algorithm both for the margin adaptive and rate adaptive problem, compared to the optimal performance and an iterative method proposed in [49]. They find that their suboptimal solution performs quite well compared to the optimal solution while the iterative solution has a significantly lower performance. However, no complexity analysis is presented. In addition, the simulation model does not reflect a realistic set up of a wireless cell as the path loss is not taken into consideration.

Shen et al. [90] investigate a power adaption scheme for dynamic MU-OFDM assignments. The basic approach studied is the rate adaptive approach. However, they address a modified version where not all terminals in the cell are assigned the same bit rate (and thus the minimal

Algorithm: Yin's combined Resource Allocation Algorithm

Result : Optimal Resource Allocation (number of sub-carriers and power) among Terminals

Given the individual terminal rate requirements $r_j^{(t)}$, the maximum transmittable power $P_{\max}$, the average channel-gain values $h_j(t)$ per terminal and a power-rate function $F^{-1}(\cdot)$ (delivering the transmission power, according to the number of bits to send, the minimum BER requirements and available coding/modulation schemes).

1 Let each terminal j initially be allocated $m_j^{(t)} = 1$ sub-carrier.

2 Compute the overall number of allocated sub-carriers: $N_a = \sum_{j=1}^{J} m_j^{(t)} = J$.

3 Determine the power needed for each terminal to fulfill its rate requirement using only that sub-carrier: $P_j = m_j^{(t)} \cdot f^{-1}\left(\frac{r_j^{(t)}}{m_j^{(t)}}\right) / g_j(t)$.

4 Determine the total amount of power needed to fulfill all data-rate requirements: $P_a = \sum_{j=1}^{J} P_j$.

while $P_a > \frac{N_a}{N} P_{\max}$ **do**

5 Compute $\Delta P_j = m_j^{(t)} \cdot f^{-1}\left(\frac{r_j^{(t)}}{m_j^{(t)}}\right) / g_j(t) - \left(m_j^{(t)} + 1\right) \cdot f^{-1}\left(\frac{r_j^{(t)}}{\left(m_j^{(t)}+1\right)}\right) / g_j(t)$.

6 Select $j^* = \arg\min_{j \in J} \Delta P_j$.

7 Update $m_{j^*}^{(t)} = m_{j^*}^{(t)} + 1$ and $P_{j^*} = P_{j^*} - \Delta P_{j^*}$.

8 Update N_a and P_a.

end

9 Allocate the remaining amount of sub-carriers $N - N_a$ to the terminal j^* with the highest average channel gain value g_{j^*}

Algorithm 4: Combined power and sub-carrier allocation algorithm.

Algorithm: Rhee's Heuristic Assignment Algorithm

Result : Suboptimal Terminal/Sub-carrier Assignment Sets.

Given a scenario of J terminals sharing N sub-carriers in one cell, the individual sub-channel gain values $h_{j,n}^{(t)}$ per terminal, a fixed transmit power per sub-carrier of $\frac{P_{\max}}{N}$ and the Shannon capacity function $F(\cdot)$ (delivering the capacity of a channel, depending on the channel gain):

1 Set the momentary data-rate per terminal $r_j^{(t)} = 0$ for all $j \in J$.
2 Let A be the set of available sub-carriers $A = \{1, 2, \ldots, N\}$.

for $j = 1$ *to* J **do**
3 Find n^* such that $|h_{j,n^*}^{(t)}| \geq |h_{j,n}^{(t)}|$ for all $n \in A$ with $n \neq n^*$.
4 Update $r_j^{(t)} = F\left(h_{j,n^*}^{(t)}\right)$ and $A = A \backslash \{n^*\}$.
end
while $A \neq \emptyset$ **do**
5 Find j^* such that $r_{j^*}^{(t)} \leq r_j^{(t)}$ for all j with $j \neq j^*$ and $0 \leq j \leq J$.
6 For j^* find n^* such that $|h_{j^*,n^*}^{(t)}| \geq |h_{j^*,n}^{(t)}|$ for all $n \in A$ with $n \neq n^*$.
7 Update $r_{j^*}^{(t)} = r_{j^*}^{(t)} + F\left(h_{j^*,n^*}^{(t)}\right)$ and $A = A \backslash \{n^*\}$.
end

Algorithm 5: Combined allocation and assignment algorithm.

terminal rate is maximized). Instead, a rate ratio between the different terminals is assumed. As a result the achieved down-link capacity is split up according to the ratios. The authors assume that a certain suboptimal assignment algorithm is applied, introduced in [77]. Based on this algorithm, for each down-link phase a certain sub-carrier assignment is generated. Next, instead of applying equal power distribution, as suggested by the authors of the suboptimal algorithm, here adaptive power distribution is proposed and an optimal power distribution method is discussed. The authors use the Shannon capacity formula as power-rate function. Therefore, solving for the optimal power distribution is a non-linear optimization problem. The authors propose to apply the method of Lagrange for this problem. In order to find the power distribution, the authors divide the allocation process into two steps: First allocate the power between different terminals, next find for each sub-carrier set of each terminal the power distribution according to water-filling. In the simulation part, the authors compare the minimum achieved throughput for the case of equal and adaptive power distribution. Three cases are compared: Fixed OFDM-TDM with equal power distribution and the two dynamic variants. The authors find that the fixed OFDM-TDM is outperformed on average by 100%, while the adaptive power distribution provides a substantial capacity increase compared to the equal power distribution. Only results for an increasing number of terminals are presented. Note that the authors assumed correlated channels with fading and path loss, restricting the maximum path loss difference to 40 dB.

Further schemes regarding the rate adaptive problem can be found in [75, 91].

3.2.3 System Studies Regarding Dynamic MU-OFDM Systems

While the previous sections presented the modeling and optimization of dynamic MU-OFDM systems, in this section various system-related investigations are presented. These studies provide insight to issues such as the performance gain which is achieved by dynamic MU-OFDM systems (compared to static schemes), the impact due to erroneous channel state information, the performance behavior of dynamic MU-OFDM systems depending on several system parameters and the impact due to signaling.

As with dynamic loading algorithms, the most important question relates to the performance gain that can be achieved by an optimal dynamic MU-OFDM algorithm. Interestingly, the potential gain of the rate- or margin adaptive problem has never been investigated systematically. Wong et al. indicate a potential gain of 100% for the optimal solution in certain propagation environments [49], compared to a static scheme. It is not discussed, if this performance advantage varies (for example, for a varying number of terminals in the cell). It is also not investigated which optimal dynamic schemes provide how much potential gain. Specifically, it is not clear if an optimal dynamic power distribution in combination with dynamic sub-carrier assignments potentially pays off in comparison to adaptive modulation (fixed power distribution) with optimal dynamic sub-carrier assignments.

How much of this performance gain can be reached by suboptimal algorithms? This question is to some extent open, too. For example, following the two-step approach of Yin et al. [76] achieves about 90% of the optimal performance. Rhee et al. [77] find for a fairly simple heuristic a much smaller "optimality gap" of about 2% (assuming a different parameter set). However, there are some inconsistencies with the results presented in these studies. In addition, it is not

clear how the performance behaves in different scenarios, i.e. for a varying transmit power, for a varying number of terminals, for a varying number of sub-carriers etc. Also, despite the fact that the different suggested algorithms are faster than the optimization, no run-time results for the algorithms have been published so far. These would indicate how fast these algorithms really are.

Apart from considering the optimal performance of a dynamic MU-OFDM system, other studies consider implementation aspects of dynamic MU-OFDM systems. In [92], Rohling et al. consider different multiplexing schemes for the down-link of broadband OFDM systems. They find that dynamic MU-OFDM clearly outperforms TDM and CDM variants. However, FDM is also found to require the largest overhead due to signaling (as is discussed below).

In [93], Wang et al. study various effects in combination with a certain system concept of a multi-user OFDM system in FDD. The authors propose a system where the time and frequency radio resources are divided into bins (chunks) of certain size (for example, 120 symbols obtained from combining 10 sub-carriers of bandwidth 10 kHz each and 12 symbols of length 111 μs each). Given perfect channel knowledge at the access point (of all terminal bins during the next three bins), the access point assigns bins to the terminal with the best channel attenuation (basically, this is an assignment strategy according to Equation 3.15). Then, adaptive modulation is applied to the specific bin. The authors show in the simulation part how the system benefits from multi-user diversity and an increasing transmit power. Analytically, they find a method to obtain the spectral efficiency of the system and validate their results by means of simulation. Also, they present results on a varying size of the bin (indicating that there are significant consequences for the systems performance if the bin size is chosen too large or too small due to the caused overhead) and investigate the performance of the system at different average receiver power levels. However, as neither the rate- or margin adaptive approach is considered, no delay or rate requirements for each terminal are assumed and evaluated.

As dynamic MU-OFDM schemes provide a significant performance gain, they are also subject to a certain overhead due to signaling. The signaling information contains the full assignment set (which terminal is assigned which sub-carrier with which modulation type). This might be quite costly. However, qualitative or quantitative studies regarding the signaling overhead are not known at all. Some studies consider the potential signaling overhead as motivation to investigate the building of blocks of sub-carriers which are assigned to different terminals instead of assigning individual sub-carriers [94, 85]. These studies show that block building is quite limited as soon as the sub-carrier blocks are larger than the coherence bandwidth. In these cases the performance loss (in bit error rate) is quite large. It is important to stress that these studies do not quantify the signaling overhead and therefore cannot quantify the improvement in terms of signaling overhead by grouping of sub-carriers.

3.3 Summary on Dynamic Schemes in MU-OFDM Systems

In this section, an overview of the related work regarding dynamic schemes for point-to-point OFDM communications and dynamic schemes for point-to-multi-point OFDM communications is presented. This summary provides an overview of the two previous Sections 3.1 and 3.2. Most statements are not verified by citations, though, as the corresponding statements with citations

can be found in the previous sections.

Point-to-Point Communications Algorithmically, the problem of power and bit loading in the presence of frequency selective fading for OFDM systems is well understood. There exist algorithms achieving the optimal performance while requiring only a small amount of computational steps. The incorporation of different system features into the optimization model bear some difficulties, though. Although the effect of interleaving has been included into the optimization model, optimal allocation strategies for coded systems have not been developed so far. Unfortunately, the impact of convolutional coders on the bit error rate of a modulated system has not been expressed analytically, therefore the incorporation of such schemes into the optimization model is difficult.

Furthermore, it is an open question whether the dynamic adaption of transmit power and modulation types should be performed. Only adapting the modulation types and keeping the transmit power fixed for all sub-carriers is outperformed only slightly by the adaption of both. Although the allocation problem can be solved optimally, the required computational power does not seem to pay off. However, this might be dependent on the considered propagation scenario as joint power and bit loading is performed in commercial DSL systems.

Compared to static systems, dynamic schemes achieve a significant performance gain which calls for the application of such schemes in wireless systems. Although they rely on receiver feed-back (regarding the sub-carrier attenuations) and on a signaling scheme in order to indicate the new modulation allocations per sub-carrier (see Figure 3.1), the performance gain is still substantial. In addition, it should be mentioned that with such dynamic schemes most of the complexity required to achieve the performance gain has to be spent on the transmitter side. This is in contrast to schemes like soft decision decoding, where a high additional complexity has to be spent at the receiver as well.

Point-to-Multi-Point Communications Dynamically assigning power, sub-carriers and modulation types to different terminals simultaneously provides a substantial performance gain. It is open in which situations such schemes particularly pay off and in which situations not. Certain studies indicate for special scenarios that this optimal performance gain can be realized closely by (suboptimal) approximation schemes. It is known that dynamic MU-OFDM systems benefit from an increasing number of terminals in the system due to the increasing multi-user diversity. However, it is open which transmit situations else qualify in particular for dynamic MU-OFDM schemes. Also, it is not known how much additional performance is provided potentially by distributing power and sub-carriers dynamically in contrast to only distributing sub-carriers dynamically. Finally, the complexity of the most studied optimization problems, the rate and margin adaptive problems, is not known.

In addition, most considered system investigations treat several aspects too superficially. For example, in practical access systems the data is going to be transmitted in packets. This leads to the transmission of only a few packets compared to the continuous transmission of some data to each terminal for each down-link phase. Also, due to the discrete size of packets, assigned resources per down-link phase may not be completely required, such that random data has to padded to fill the assigned resource. One system model following this packet approach has

been discussed in [88]. However, it only considers a single modulation type to be available. Introducing adaptive modulation increases the complexity of the approach quite a bit.

Furthermore, the impact of realistic channel knowledge has not been considered at all. Although it can be expected that the impact of erroneous channel knowledge is similar as with loading approaches, estimating the channel becomes more difficult as there are multiple terminals in the cell. For all of these the channel estimation has to be performed resulting in a higher requirement regarding the available computational power at the access point. Therefore, this might lead to a higher (negative) impact on the performance as a reliable estimation over a time span of one down-link phase cannot be provided for more than a couple of terminals in the cell.

In this context, the required feed-back in the up-link has also not been considered. Assuming a TDD system, the terminals have to report the current attenuations through the up-link. These values are then processed at the access point (i.e. input to an estimation algorithm). Obviously, this requires some system resources in the up-link, depending on the granularity of the attenuations reported back to the access point. This has not been investigated so far at all. Finally, also the overhead in the down-link due to signaling has not been investigated at all.

3.4 Scope of the Thesis

From the above discussion it is clear that the area of dynamic adaptions for point-to-point OFDM communications is well understood. This has lead to the application of such schemes, for example in DSL systems. In contrast, dynamic schemes for point-to-multi-point OFDM communications have not been implemented so far. The reason for this is a lack of understanding regarding theoretical and practical issues. In this thesis, I basically consider dynamic schemes for point-to-multi-point OFDM communications. In particular, the down-link of a single (no interference) wireless cell is assumed, where several terminals are associated to one access point. The access point coordinates all data transmissions.

First, regarding such systems it is not clear how sub-carriers should be assigned to terminals. Clearly, some performance metric is to be improved by the dynamic assignment. This leads to the consideration of appropriate optimization models. However, for meaningful optimization models the theoretical complexity involved in the assignments is not understood. Although multiple authors claim the rate and/or margin adaptive problem to be difficult optimization problems, these problems have not been classified regarding their computational complexity. Thus, although many authors propose suboptimal methods, it is open if there exist efficient algorithms which *optimally* solve these problems, even as the number of terminals or the number of sub-carriers increases. In contrast, these problems might be *NP*-hard, such that any algorithm solving them optimally will have unacceptable high run times as system parameters like the number of sub-carriers increases. In this case, suboptimal algorithms have to be used. Multiple of them have been introduced in the literature (see Section 3.2.2). Unfortunately, all these investigations do not provide comparable results such that certain algorithms qualify more than others. Very different scenarios are considered, sometimes even the optimal upper bound is not provided in the published results.

Furthermore, it is open which dynamic adaptions (dynamic sub-carrier assignments, dynamic power distribution) exactly provide how much performance, for the optimally solved

cases and in the case of using appropriate suboptimal algorithms. These performance differences might vary as system parameters change, for example, in the case of a large number of terminals in the cell compared to a small one. Nothing is known about such cases. Nonetheless, these issues are important for system engineers, as different approaches require a different amount of computational power at the access point. In addition, it is open how to apply most of the discussed approaches in realistic, packet-oriented communication systems. The optimization models for the rate and margin adaptive problems do not consider the fact that packets are present at the access point instead of an infinitely long stream of data to be transmitted to each terminal. Very few optimization models have been investigated considering packet queues at the access point.

Summarizing all this, dynamic MU-OFDM systems are not well understood, both from the theoretical point of view (complexity, potential performance of different approaches) and the practical point of view (comparable performance results, run times, application to packet-oriented communication systems). Almost all publications regarding dynamic MU-OFDM systems only suggest a new (suboptimal) solution method and study its performance regarding some scenario.

Further issues remain completely open. Dynamically assigning sub-carriers to terminals requires more than only some additional computational power at the access point. At least, two further practical aspects have to be considered: How to obtain the current sub-carrier attenuations of all active terminals and how to indicate to terminals which sub-carriers they have been assigned. Thus, the system performance might be decreased in practice by:

- Potentially inadequate channel knowledge (channel estimation, outdated or erroneous feed-back).
- Overhead in order to update the channel knowledge at the access point (which requires some time span in order to be transmitted from all terminals to the access point).
- Overhead in order to update the terminals of their current assignments.
- Transmit errors in the stream of the signaling data, i.e. bit errors.

The former issues are related to the channel knowledge, the latter issues are related to the signaling. In both cases, overhead as well as new sources for erroneous behavior are introduced. It is possible that the required overhead annihilates the potential performance increase provided by the dynamic MU-OFDM approach (compared to static approaches). In addition, the transmitted overhead might be received erroneous (in case of the signaling information by the terminals, in case of the channel feed-back by the access point). Such erroneous information can reduce the system performance further.

The issue of channel knowledge is closely related to the reciprocal character of the wireless channel[95]. Theoretically, any wireless channel is reciprocal, i.e. the attenuation is equivalent in both transmission directions. Thus, no feed-back channel is required for the channel state information if the transmission scheme is identical in the down-link and up-link (in contrast to a system where OFDM is used in the down-link and CDMA is used in the up-link, for example). Then, each terminal simply transmits a short training symbol sequence to the access point. The

access point uses these sequences to estimate the attenuation of each sub-carrier for each terminal, yielding a quite accurate system description. It is important to note that a feed-back channel for the channel knowledge would decrease the channel knowledge accuracy at the access point, compared to the method of directly transmitting training symbols to the access point. Using a feed-back channel delays the channel knowledge. In addition, it might lead to transmit errors in the feed-back channel itself. If a channel estimation is required, the terminals would require additional computational power. Thus, as long as the wireless channel can be assumed to be reciprocal, using a feed-back channel is rather complicated compared to the method of directly measuring the channel at the access point. The time requirements can be easily calculated for this method while providing the current sub-carrier attenuations. If the wireless channel can be assumed to be constant during the coherence time, working with these "plain" attenuations is a promising approach.

This situation changes if the transmission scheme for the up-link is not OFDM (as it is for the down-link) but some form of SCM scheme, for example. An asymmetric system design can be motivated by the PARP, as discussed in Section 2.2. Due to the PARP, terminals using OFDM in the up-link direction would be required to have much better amplifiers (with a larger range of linearity), increasing the price for the devices. Hence, other transmission schemes might be applied for the up-link direction. Thus, a feed-back channel is required to transmit the sub-carrier attenuations to the access point. Depending on the duplex scheme, reporting the channel attenuations back by this feed-back channel might be delayed significantly (in a **T**ime **D**ivision **D**uplex (*TDD*) scheme), while in a **F**requency **D**ivision **D**uplex (*FDD*) scheme the feed-back would not be delayed significantly.

Thus, the issue of channel knowledge is closely related to the specific design of the transmission system in the up-link as well as to the issue of the duplex scheme. It appears as if the issue of channel knowledge is rather an engineering problem when implementing the system as long as the up-link transmission scheme is not OFDM in a TDD system. In the three other system layouts (TDD system with an OFDM up-link, FDD system with an OFDM up-link, FDD system with no OFDM up-link) the acquisition of the channel knowledge is rather straightforward.

Throughout this thesis, a TDD scheme is assumed while the up-link transmission scheme is assumed to be OFDM. Thus, the issue of channel knowledge is not further considered as the access point retrieves the channel state information during the up-link phase. However, the design of an appropriate feed-back channel for other systems (as for example an TDD system with a different transmission scheme in the up-link than OFDM) is an open issue, but considered as future work.

In contrast, the signaling problem is important regardless of the system design. The need for signaling is obvious: As the access point dynamically assigns sub-carriers to terminals, prior to the payload transmission the terminals have to be informed of their new assignments. Otherwise, they cannot decode their data. In addition to this, per sub-carrier also the used modulation/coding combination has to be signaled. This information has to be received by the terminals prior to the payload data transmission. In addition to the pure signaling data, the signaling overhead also has to be error protected. This is a quite important issue. If bit errors occur in the signaling information, certain terminals will decode the wrong data. Most probably, these terminals will identify afterwards these received packets as erroneous ones and invoke the **A**utomatic **R**epeat

Request (*ARQ*) procedure, if lossless transmission is required. Errors in the signaling part cannot be corrected by error coding in the payload data, as terminals receive large packet chunks on each sub-carrier. Major parts of the packet are corrupted and direct error coding cannot compensate for the loss. Therefore, the signaling information itself has to be protected by error coding, leading to a larger overhead. A very important question is if the required signaling information has a significant qualitative and quantitative effect on the performance of dynamic MU-OFDM systems. In the worst case, the signaling information consumes most of the performance increase provided by a dynamic MU-OFDM system. Thus, the resulting performance equals more or less the one of a static system. Nothing is known regarding this issue so far from the literature.

In this thesis, I consider the issues of the algorithmic problems regarding dynamic MU-OFDM systems and the issue regarding the signaling information as the focus of the work. The thesis is restricted to the down-link of TDD-OFDM wireless systems. In particular, the rate adaptive optimization problem is considered.

In Chapter 4, the complexity of the rate (and margin) adaptive optimization problem is first considered. Next, the optimal performance regarding the rate adaptive approach is studied while comparing different degrees of dynamic approaches (pure power adaption, pure adaptive modulation, adaptive modulation with sub-carrier assignments and power adaption with sub-carrier assignments). Then different suboptimal schemes are studied which can be applied in practical systems. Finally, it is discussed how such suboptimal schemes can be applied to packet-oriented transmission systems.

In Chapter 5 the cost due to signaling is studied. Based on a selected dynamic MU-OFDM approach a basic model for the representation of the signaling overhead is introduced. It is used to obtain initial results on the impact due to the signaling overhead. Next, various schemes are investigated which can reduce the signaling overhead. Entropy models are used to find upper limits for the application of compression schemes. In addition, approaches are studied which incorporate the cost of signaling into the optimization model. This results in new problems for which the system performance is studied. Finally, compression algorithms are applied to the signaling overhead as comparison schemes.

In Chapter 6 an outline of a concept is presented which defines an IEEE 802.11a compliant system based on dynamic MU-OFDM. So far, no such system concept has been proposed. The new proposal is based on changing the IEEE 802.11a standard slightly to enable a dynamic MU-OFDM approach. The advantage of the new proposal is to boost the throughput for most data transmissions by dynamically assigning sub-carriers and transmit power to stations.

Finally, in Chapter 7 the conclusions of the thesis are presented. In addition, a few open issues remain, which are presented as well.

Chapter 4

Assignment Algorithms for Dynamic MU-OFDM Systems

In this section, several issues regarding the rate adaptive optimization problem are investigated. First, in Section 4.1, a system model is presented. In this context, the rate adaptive approach is stated as an integer programming problem. Then, in Section 4.1.1, the complexity of this approach is discussed. Afterwards, in Section 4.1.2, the optimal performance of the rate adaptive approach is studied. Finally, in Section 4.2 and 4.3, a new contribution regarding suboptimal algorithms for the rate adaptive optimization problem is discussed.

4.1 Modeling the Rate Adaptive Approach in Dynamic MU-OFDM Systems

Throughout this section, the following system model is used: Assume a single cell of a wireless system with radius r_{cell}. Within this cell, one access point coordinates all data transmissions. $\hat{J}$ terminals are associated with this access point. The terminals are uniformly distributed over the area of the cell. Out of the number of associated terminals, J terminals are currently active, which means that each of these $numTerminals$ terminal is receiving a data stream. It is assumed that the access point always has some data to be transmitted to each terminal. Additional control information which would be requires for the indication of data fragments is not taken into consideration.

Physical Layer System Assumptions For data transmission, a total bandwidth of B [Hz] can be utilized at the center frequency f_{c}. For this bandwidth, a maximum transmit power of P_{max} is allowed to be radiated. OFDM is applied as transmission scheme in the up-link as well as in the down-link direction. The given bandwidth is split into N sub-carriers with a bandwidth of $\frac{B}{N}$ each. In order to guarantee orthogonality between the sub-carriers, the symbol length for all sub-carriers is identical and is related to the bandwidth of each sub-carrier by $\frac{N}{B} = T_{\text{s}}$. Although each sub-carrier employs the same symbol rate, per sub-carrier a different amount of bits might be represented by a symbol. M modulation types are available. Based on the

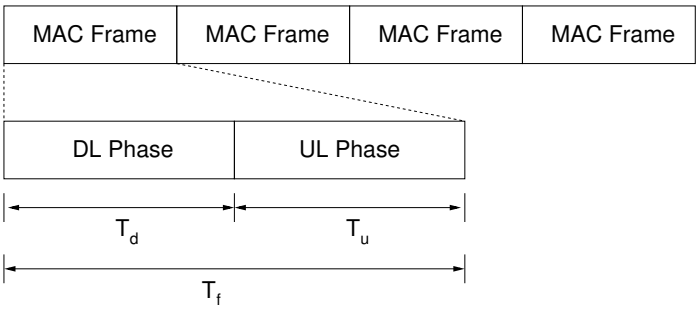


Figure 4.1: MAC layer timing structure

modulation symbols per sub-carrier, one OFDM symbol is generated by applying the inverse fast Fourier transformation to the N sub-carriers (see Section 2.2). Prior to the transmission of this discrete time sequence, a cyclic extension of length T_{g} is added, referred to as guard interval. Therefore, the duration of the complete OFDM symbol is $T_{\mathrm{s}} + T_{\mathrm{g}}$.

Wireless Channel Model As objects within the cell move constantly, the perceived signal quality per sub-carrier and terminal varies permanently. The instant SNR of sub-carrier n for terminal j at time t is given by $v_{j,n}^{(t)} = \frac{p_n^{(t)} \cdot \left(h_{j,n}^{(t)}\right)^2}{\sigma^2}$, where $p_n^{(t)}$ denotes the transmission power, $h_{j,n}^{(t)}$ denotes the attenuation of sub-carrier n versus terminal j, and σ^2 denotes the noise power. The attenuation causes the variation of the perceived SNR; it varies due to path loss, shadowing and fading (see Section 2.1.1). The attenuation of each sub-carrier is assumed to be constant over the time unit T_{f}. This time unit is considered to be smaller than the coherence time of the wireless channel (see Section 2.1.1).

Medium Access Control Layer Time is divided into frames of length T_{f} (see Figure 4.1). During each frame a fixed time span is reserved for down-link and up-link transmissions. Thus, a TDD scheme is considered. During the reserved down-link transmission, dynamic MU-OFDM is applied. Apart from reserving a time span within every frame, the up-link is not considered any further. The time duration of one down-link phase is denoted by T_{d}. As one OFDM symbol has a length of T_{s}, the total number of symbols to be transmitted during one down-link phase equals

$$S = \frac{T_{\mathrm{d}}}{T_{\mathrm{s}} + T_{\mathrm{g}}} \quad . \tag{4.1}$$

Prior to each down-link phase, the access point generates new assignments of sub-carrier subsets and power for each terminal in the cell. Delay effects owing to an inadequate processing power at the access point are not considered. Since the terminals do not know the new assignments a priori, the access point has to signal these assignments to the terminals first. Assume a separate control channel to be available. Transmissions on this control channel are always error-free and are never delayed. Thus, prior to the payload transmission, all terminals are informed of their

new assignments.

The generation of the new assignments per frame is based on the knowledge of the sub-carrier attenuations towards each terminal, denoted by the matrix $\mathbf{H}^{(t)} = \left(h_{j,n}^{(t)} | \forall\, j,n\right)$. It is assumed that this knowledge is available. As the terminals apply OFDM also in the up-link, a simple scheme as explained in Section 3.4 can be applied to acquire very accurate channel knowledge.

Optimization Models In the literature, two different objective functions have been essentially investigated so far regarding a dynamic MU-OFDM system model: Minimizing the transmit power (while conveying a fixed amount of data to each terminal in the cell, known as margin adaptive approach); and maximizing the minimal throughput per terminal (while the transmit power is upper bounded to a certain limit, known as rate adaptive approach).

For both objective functions reasonable scenarios exist. Clearly, minimizing the transmit power is useful since the interference caused in other systems is reduced. Thus, one might argue that the margin adaptive approach is chosen whenever the interference is the major concern. This might apply to unlicensed frequency bands where each transmission system tries to reduce the radiated transmit power as much as possible, for example. However, it might also apply to cellular systems, if the frequency reuse pattern is rather low. Then, the interference impact of a particular cell to neighbor cells with the same frequency band is to be reduced, which leads to the margin adaptive problem.

In contrast, if the interference impact to neighboring systems can be neglected as long as a certain maximum transmit power limit P_{max} is preserved, the rate adaptive approach is of much more interest. This might also apply to both, the cellular setting as well as the unlicensed frequency band setting. In this thesis, the rate adaptive approach is chosen to be further investigated. The focus is on maximizing the rate for the terminals currently associated to the access point.

A fairly simple optimization problem results from purely maximizing the total throughput of the system.

$$\begin{aligned} \max_{\vec{P}^{(t)},\mathbf{X}^{(t)}} \quad & S \cdot \sum_{j,n} F\left(\frac{p_n^{(t)} \cdot \left(h_{j,n}^{(t)}\right)^2}{\sigma^2}, p_{\text{err}}\right) \cdot x_{j,n}^{(t)} \\ \text{s. t.} \quad & \sum_{j} x_{j,n}^{(t)} \leq 1 \qquad \forall\, n \\ & \sum_{n} p_n^{(t)} \leq P_{\text{max}} \end{aligned} \tag{4.2}$$

Below, the parameters and variables involved in formulating this optimization problem are summarized (from the system model text above):

- t – *Constant, integer*: System time, i.e. index of the link layer frames or down-link phases.
- J – *Constant, integer*: Number of active wireless terminals in the cell.

- N – *Constant, integer*: Number of sub-carriers the bandwidth is split into.
- S – *Constant, integer*: Number of OFDM symbols transmitted during a down-link phase, as given in Equation 4.1.
- $h_{j,n}^{(t)}$ – *Constant, real*: Attenuation of sub-carrier n towards terminal j during frame t, all attenuation values during frame t are contained in matrix $\mathbf{H}^{(t)}$.
- σ^2 – *Constant, real*: Average noise power per sub-carrier.
- $P_{\max}$ – *Constant,real*: Transmit power budget.
- $F\left(v(t), p_{\mathrm{err}}\right)$ – *Continuous Function, real*: Functional mapping between the SNR of a sub-carrier given by $p_n^{(t)} \cdot \left(h_{j,n}^{(t)}\right)^2 / \sigma^2$ and the amount of bits conveyable per symbol, referred to as power-rate function (see Section 3.1.1). This mapping can describe the switching policy of the adaptive modulation. For example, an approximation of the QAM error function might be used in combination with a target bit error- or symbol error probability. Also, the Shannon capacity function might be used as a functional relationship. If the transmit power per sub-carrier is fixed, the value $b_{j,n}^{(t)}$ can be directly obtained (see below).
- $b_{j,n}^{(t)}$ – *Constant, real*: Number of payload bits to be conveyed by one symbol on sub-carrier n to terminal j during down-link phase t. This value is obtained from the sub-carrier attenuation if the transmit power per sub-carrier is fixed and $F\left(v(t), p_{\mathrm{err}}\right)$ is specified. Then, this bit number can be obtained directly considering the target error probability. The bit values of all possible assignment pairs j, n during down-link phase t form the bit matrix $\mathbf{B}^{(t)}$.
- $x_{j,n}^{(t)}$ – *Variable, binary*: Assignment of sub-carrier n to terminal j during down-link phase t. All assignments during down-link phase t are denoted by the assignment matrix $\mathbf{X}^{(t)}$.
- $p_n^{(t)}$ – *Variable, real*: Power assignment of sub-carrier n during down-link phase t. All power assignments form the power allocation vector $\vec{P}^{(t)}$.

In practical communication systems only a limited number of M modulation/coding combinations is available. Thus, given the attenuation $h_{j,n}^{(t)}$ of a certain sub-carrier n regarding a certain terminal j, the power assignment $p_n^{(t)}$ only takes one of M discrete values. Therefore, Equation (4.2) can be reformulated and becomes an integer programming problem. The following additional notations are introduced:

- M – *Constant, integer*: Number of modulation types available.
- $p_{j,n,m}^{(t)}$ – *Constant, real*: Power amount required to transmit data with modulation type m to terminal j on sub-carrier n. This value results directly from the attenuation $h_{j,n}^{(t)}$,

the noise power σ^2, the power-rate function $F\left(v(t), p_{\text{err}}\right)$ of the corresponding modulation types and the target error probability $p_{\max}$. For example, the corresponding SNR per modulation type might take the form of a stair-case function (as in Figure 3.5).

- $r\left(m\right)$ – *Constant, integer*: Number of bits transmitted per symbol by modulation type m.
- $y_{j,n,m}^{(t)}$ – *Variable, binary*: Binary decision variable, ruling the assignment of sub-carrier n to terminal j with modulation type m at time t. All these assignments form the matrix $\mathbf{Y}^{(t)}$.

Then, optimization problem (4.2) can be rewritten as:

$$\begin{aligned} \max_{\mathbf{Y}^{(t)}} \quad & S \cdot \sum_{j,n,m} r\left(m\right) \cdot y_{j,n,m}^{(t)} \\ \text{s. t.} \quad & \sum_{j,m} y_{j,n,m}^{(t)} \leq 1 \qquad \forall\, n \\ & \sum_{j,n,m} p_{j,n,m}^{(t)} \cdot y_{j,n,m}^{(t)} \leq P_{\max} \end{aligned} \tag{4.3}$$

The major difference (of the formulations) between problem (4.2) and problem (4.3) is the new, first constraint in Equation (4.3). As the decision variable $y_{j,n,m}^{(t)}$ is binary, for each sub-carrier only one terminal and modulation type can be chosen. The second constraint limits the transmit power. In contrast to problem (4.2), formulating the problem as in Equation (4.3) immediately allows to estimate the number of feasible solutions (as the problem is now an integer programming problem). For each sub-carrier one of $J \cdot M$ different, feasible assignments is chosen. Therefore, the total number of feasible (integer) solutions is $N^{J \cdot M}$.

Solving this problem is quite simple: Each sub-carrier is assigned to the terminal with the lowest attenuation. Then, an optimal loading algorithm distributes the power among the chosen sub-carrier assignments and determines therefore the amount of bits to be transmitted per sub-carrier. This approach has a low complexity and achieves the best possible capacity[1] of the cell. If the average attenuation per terminal over all sub-carriers is more or less equal for all terminals, this approach can be applied. The assignment of sub-carriers depends then only on the current stochastic variation due to fading. In such a case, even if a terminal receives only quite a small bandwidth for a certain down-link phase, it will receive probably more during some future down-link phase, in which the fading is less destructive. For example, this might apply to rather small cells with a line-of-sight for all terminals.

However, if the path loss and shadowing components cause a significant difference between the average attenuation of different terminals, the greedy solution to Equation 4.3 will cause a severe imbalance between different terminals. Several terminals will experience a large delay

[1] Capacity denotes in this discussion and thereafter the total average throughput of the cell, not the information theoretic measure.

in receiving packets. This might occur, if the the cell size is rather large, for example. This difference of the average attenuation is more likely to occur in general. In such a case, the primary interest is to maximize the throughput for all terminals equally, such that the delay is upper bounded for all terminals in the cell. Therefore, solving the problem stated in Equation 3.16, known as *rate adaptive* problem, is more suitable. For example, if the cell size is large, the rate adaptive approach does not cause the problem of starvation. Using the above mentioned notation, problem (3.16) takes the following form:

$$\begin{aligned} \max_{\mathbf{Y}^{(t)}} \quad & \epsilon \\ \text{s. t.} \quad & \sum_{j,m} y_{j,n,m}^{(t)} \leq 1 \quad \forall\, n \\ & \sum_{j,n,m} p_{j,n,m}^{(t)} \cdot y_{j,n,m}^{(t)} \leq P_{\text{max}} \\ & S \cdot \sum_{n,m} r\,(m) \cdot y_{j,n,m}^{(t)} \geq \epsilon \quad \forall\, j \end{aligned} \tag{4.4}$$

Regarding the optimization approach of Equation 4.4, dynamic power assignments, as well as dynamic sub-carrier assignments are involved, which yields the optimal performance. However, complexity issues might lead to a simpler assignment approach, where the transmit power is fixed per sub-carrier. A straightforward way is to distribute the transmit power equally among all sub-carriers. If the transmit power per sub-carrier is fixed, a modulation type is directly obtained for each sub-carrier/terminal pair[2].

$$\begin{aligned} \max_{\mathbf{X}^{(t)}} \quad & \epsilon \\ \text{s. t.} \quad & \sum_{j} x_{j,n}^{(t)} \leq 1 \quad \forall\, n \\ & S \cdot \sum_{n} b_{j,n}^{(t)} \cdot x_{j,n}^{(t)} \geq \epsilon \quad \forall\, j \end{aligned} \tag{4.5}$$

Note that optimization problem (4.5) has a lower number of feasible solutions. For each sub-carrier only one out of J assignments is possible, thus an overall number of N^J feasible solutions exist.

4.1.1 Complexity of the Rate Adaptive Optimization Problem

In this section, it is shown that the rate adaptive optimization problem in dynamic MU-OFDM systems is *NP*-complete. Initially, concentrate on optimization problem (4.5), where the dynamic power assignment has been replaced by a fixed distribution. In addition, a system with

[2]If the transmit power is fixed, an SNR is immediately obtained for each terminal/sub-carrier pair. For this SNR value one can always find a modulation type which provides the best throughput while fulfilling the target error probability.

a finite amount of modulation/coding schemes is considered, such that all elements of matrix $\mathbf{B}^{(t)}$ are integer values.

Consider first the *recognition version* [45] of the rate adaptive optimization problem (4.5). The recognition version is a decision problem formulation of the optimization problem, asking if there is a feasible solution of a combinatorial instance *greater or equal* (*lower or equal*) to some integer value L. Thus, the recognition version is a question that can be answered either by yes or no. Any instance of an optimization problem can be solved by considering the recognition version of the problem's instance polynomially often [45].

The recognition version of problem (4.5) is introduced as:

HEAVY DISJOINT SUBSETS :
Given J sets of N integers $B_j = \{b_{j,1}, \dots b_{j,N}\}$ and an integer L, are there J disjoint subsets $A_j \subseteq \{1, \dots N\}$ such that $\forall\ j : \sum_{n \in A_j} b_{j,n} \geq L$?

This is the recognition version of problem (4.5), as for each terminal j the N different sub-carriers of the system are represented by the integer sets B_j. The assignments per terminal are recorded in the integer sets A_j, which have to be disjoint such that each sub-carrier is assigned to one terminal only.

In order to prove that this recognition problem is *NP*-complete, first it has to be shown that all problems in *NP* polynomially reduce to HEAVY DISJOINT SUBSETS. In general, a problem P_1 reduces in polynomial time to some other problem P_2, if and only if there exists a polynomial-time algorithm $\mathfrak{a}_1$ for P_1 that uses several times an algorithm $\mathfrak{a}_2$ for P_2 as a subroutine. Thus, the problem P_1 can be said to be solved by translating it to a problem P_2 for which an algorithm $\mathfrak{a}_2$ is known (or at least assumed to be known). The output of the algorithm $\mathfrak{a}_2$ is then translated back and provides a solution for P_1. By picking some *NP*-complete problem P_1 and showing that this reduces to the above recognition problem P_2, it is shown that P_2 is *NP*-hard. Performing this reduction proves that HEAVY DISJOINT SUBSETS is at least as difficult as any *NP*-complete problem.

However, a second step is required proving that HEAVY DISJOINT SUBSETS is not harder than any *NP* complete problem. This is done by verifying in addition that the decision problem belongs to the class of *NP*. A decision problem P belongs to the class *NP* if there exists a *concise certificate* for any potential solution of P, which can be checked in polynomial time for validity. A concise certificate of a potential solution refers to a representation of it, polynomially bounded in length.

First consider the reduction. For this, the following *NP*-complete problem PARTITION is chosen:

PARTITION :
Given N integers $c_1, \dots, c_N$, is there a subset $S \subseteq \{1, \dots, N\}$ such that $\sum_{n \in S} c_n = \sum_{n \notin S} c_n$?

Suppose now there is an algorithm $\mathfrak{A}$ solving the recognition problem HEAVY DISJOINT SUBSETS. In addition, there is a problem instance of PARTITION given (the set of integers $\{c_1, \dots, c_N\}$). Can this instance of PARTITION be solved by calling $\mathfrak{A}$ with a specific param-

eterization?

The answer is yes. The reduction works as this: Denote by $C = \{c_1, \ldots, c_N\}$ the set of given integers. As input, $\mathfrak{A}$ receives $J = 2$ sets of N identical integers $B_1 = C$ and $B_2 = C$. Prior to invoking $\mathfrak{A}$, the sum $l = \sum_{n \in \{1,\ldots,N\}} c_n$ is checked if it is even or odd. In the case that l is odd, we stop immediately since this instance of PARTITION cannot be solved. Else, set $L = \frac{l}{2}$, where L is the lower limit used in HEAVY DISJOINT SUBSETS.

Algorithm $\mathfrak{A}$ will now try to find two disjoint subsets such that the "weight" of each subset is greater or equal half of the sum l of the PARTITION integers. Using this parameterization of $\mathfrak{A}$, PARTITION is solved because there cannot be two disjoint subsets both with *different* "weights" greater or equal than l. Generating the special parameterization for algorithm $\mathfrak{A}$, given the instance of PARTITION, can obviously be done in polynomial time. Therefore, a polynomial time reduction of PARTITION to HEAVY DISJOINT SUBSETS is found, the recognition version of problem (4.5). Hence, HEAVY DISJOINT SUBSETS is *NP*-hard and therefore the optimization problem (4.5) is *NP*-hard, too.

It remains to be shown that HEAVY DISJOINT SUBSETS belongs to the class of *NP*. Consider a potential solution of an instance of this problem. It can be represented by denoting N pairs (j, n), referring to the integer $b_{j,n}$ that was chosen. This certificate is polynomially bounded in space as it grows linearly with N. This certificate can also be validated quite fast. Iterating through the set of pairs, for each j the integer $b_{j,n}$ is summed separately. Next, each of these summations are checked if they are greater or equal L. If so, then it remains to validate that all n of the pairs are disjoint. Therefore, while skipping through the pairs also each n is denoted in a sorted list. If for some n this list already contains an entry, the potential solution is discarded. All these operations are polynomially bounded if varying the parameters N or J. Therefore, it can be concluded that HEAVY DISJOINT SUBSETS is *NP*-complete.

Next consider the optimization problem (4.4), where the power can be distributed dynamically (with the constraint on the total amount of transmit power in the used band) besides distributing the sub-carriers dynamically. As stated, assuming a common wireless system, a certain number of modulation types M will be available. Given a target error probability for the data transmission, for every sub-carrier/terminal combination M transmit power values can be calculated, which are required to reach the (error probability dependent) SNR for each modulation type. The power assignment then refers to choosing one of M different power values.

Based on the proof that problem (4.5) is *NP*-complete, it is straightforward to prove that problem (4.4) is *NP*-complete, too. First, consider the reduction. Define the following recognition version of problem (4.4):

CONSTRAINED HEAVY DISJOINT SUBSETS :
Given J sets of $N \cdot M$ integers $B_j = \{b_{j,1,1}, \ldots, b_{j,N,M}\}$, $J \cdot N \cdot M$ real numbers $p_{j,n,m}$ and two numbers $L \in \mathbf{N}$ and $P_{\text{max}} \in \mathbf{R}$, are there J disjoint subsets $A_j \subseteq \{1, \ldots N\}$ and a set of N integers $m_n \in \{1, \ldots, M\}$ such that $\forall\ j : \sum_{n \in A_j} b_{j,n,m_n} \geq L$ and $\sum_j \sum_{n \in A_j} p_{j,n,m_n} \leq P_{\text{max}}$?

In this recognition problem, the integer sets B_j correspond to all possible sub-carrier/modulation type combinations which are possible. The real numbers $p_{j,n,m}$ correspond to the required

power. As in the case of fixing the transmit power, the disjoint sets A_j contain then the sub-carrier assignments for each terminal j. The N integer values m_n denote the modulation assignments per sub-carrier n. For both of these assignments then the limit L must hold, which ensures that each terminal receives an amount of bits greater or equal to L. In addition, for the modulation assignments a limit on the total transmit power $P_{\max}$ must hold.

Now, suppose an algorithm $\mathfrak{Z}$ solves the recognition problem CONSTRAINED HEAVY DISJOINT SUBSETS. Obviously, this algorithm $\mathfrak{Z}$ will also solve a problem instance of HEAVY DISJOINT SUBSETS. In this case, each combination of j and n has only one single value for $b_{j,n,m}$ instead of M ones. Therefore, all other values for $b_{j,n,m}$ are set to zero. In addition, set all values of $p_{j,n,m}$ to zero, since this constraint is of no interest in problem HEAVY DISJOINT SUBSETS. Now, pass the lower limit L together with the integer sets B_j, the numbers $p_{j,n,m}$ and some value for $P_{\max}$ to the algorithm $\mathfrak{Z}$. If there is a solution it will be a valid solution for any problem instance of HEAVY DISJOINT SUBSETS. Therefore, reasoning as above, it has been proven that optimization problem (4.4) is *NP*-hard.

In fact, CONSTRAINED HEAVY DISJOINT SUBSETS is also *NP*-complete. To prove this, consider the certificate for a possible solution of an instance of CONSTRAINED HEAVY DISJOINT SUBSETS. The solution can be denoted by N triples $\langle j, n, m\rangle$. Compared to the certificate of HEAVY DISJOINT SUBSETS a third element is added, indicating which real number $p_{j,n,m}$ is chosen (therefore, which power level is chosen for the sub-carrier/terminal pair). This certificate scales again linearly with N. Also, it can be validated in polynomial time. In addition to the validation in case of problem HEAVY DISJOINT SUBSETS, the transmit power limit has to be checked, which is one further summation. But certainly the complete process is polynomially bounded in time. Thus, problem (4.4) is *NP*-complete.

Interestingly, it is not difficult to see from the recognition problem CONSTRAINED HEAVY DISJOINT SUBSETS that also the margin adaptive optimization problem is *NP*-complete. To see this, consider the discrete margin adaptive optimization problem below (where the requested rate per terminal is represented by ϵ_j.

$$\begin{aligned} \min_{\mathbf{Y}^{(t)}} \quad & \sum_{j,n,m} p_{j,n,m}^{(t)} \cdot y_{j,n,m}^{(t)} \\ \text{s. t.} \quad & \sum_{j,m} y_{j,n,m}^{(t)} \leq 1 \quad \forall\, n \\ & \sum_{n,m} r\,(m) \cdot y_{j,n,m}^{(t)} \geq \epsilon_j \qquad \forall\, j \end{aligned}$$

This optimization problem has exactly the same recognition problem CONSTRAINED HEAVY DISJOINT SUBSETS. However, the optimization problem is solved using the recognition problem by successively reducing the limit $P_{\max}$ for a fixed L. In contrast, the optimization problem (4.4) is solved using recognition problem CONSTRAINED HEAVY DISJOINT SUBSETS by successively increasing L while fixing $P_{\max}$. As a consequence, it follows that the margin adaptive approach is *NP*-complete as well.

4.1.2 Performance of the Optimally Solved Rate Adaptive Problem

In the previous section, it has been shown that the rate adaptive optimization problem (as well as the margin adaptive approach) is *NP*-complete. Even if the transmit power is fixed (which could be believed to reduce the complexity), the rate adaptive problem remains *NP*-complete. The consequence of this result is that suboptimal schemes are to be used for the assignment, as one cannot expect to find a fast optimal algorithm for the rate adaptive problem (with or without dynamic power distribution). Despite the theoretical complexity of the optimal problems, suboptimal schemes will most likely differ substantially in their average run-times if the power is to be distributed dynamically compared to the case were the power per sub-carrier is fixed. If dynamic power assignments were known to provide not much additional performance (even in the optimal case), system engineers could stick to developing suboptimal schemes with a static power distribution. However, nothing is known about the performance difference between these two approaches. Thus, it is investigated in this section regarding the *optimal* performance. Primarily, the optimal solution of the rate adaptive problem (4.4) is compared to the optimal solution of the (still) rate adaptive problem (4.5). In addition, three other schemes are considered for illustration purposes. The (five) different schemes considered are the following (see also Figure 4.2):

- **Rate Adaptive with Loading**: The optimal solution to the fully dynamic rate adaptive problem, stated in Equation (4.4). As pointed out, power and sub-carriers are distributed dynamically per down-link phase. The access point has to solve an integer programming optimization problem for each down-link phase. As for each sub-carrier one of J terminals is assigned one of the M modulation types, there are $N^{J \cdot M}$ different feasible solutions (regarding the integer constraint on the assignments only; taking the other constraints into account, reduces the amount of feasible solutions).

- **Rate Adaptive with Adaptive Modulation**: The optimal solution to the rate adaptive problem (4.5), which only assigns sub-carriers dynamically. The transmit power is fixed and yields for each sub-carrier/terminal pair a certain modulation type depending on the attenuation value. Thus, the modulation type applied per sub-carrier and terminal pair is not variable but is fixed. This simplification leads to a much smaller space of feasible solutions, N^J, regarding the integer constraint of the problem. However, as stated in the previous section, the computational complexity of the problem remains the same as if the transmit power could be distributed dynamically. Still, in practice it can be expected that the generation of the optimal solution is performed faster than of the fully dynamic approach.

- **Raw Rate Maximization**: The first comparison scheme for the two rate adaptive problems is the optimal solution to the raw bit rate maximization problem, formulated as integer programming optimization problem in equation (4.3). Primarily, this optimal solution is taken into consideration as it provides the maximum throughput of the cell. Comparing it to the throughput of the rate adaptive solutions characterizes the "cost" for taking the delay within the cell also into consideration (as the rate adaptive problem optimizes

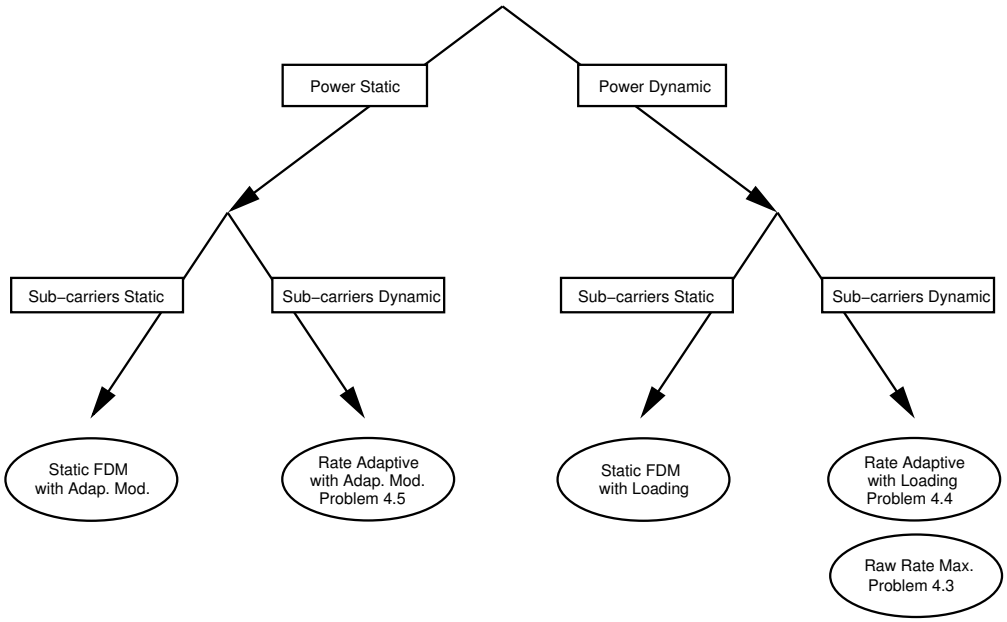


Figure 4.2: The relationship between the five different schemes investigated in this section.

the maximum delay within the cell). The solution to problem (4.3) can be achieved by a simple greedy algorithm.

- **Static FDM with Loading**: The second comparison scheme applies bit and power loading, but a static sub-carrier distribution. Each terminal in the cell is assigned the same sub-carrier set for each down-link phase. All sets have the same cardinality. Also, each terminal receives the same total transmit power, which is then dynamically loaded to the sub-carriers.

- **Static FDM with Adaptive Modulation**: The last comparison scheme is a fully static approach. Each terminal receives the same sub-carrier set for each down-link phase. Also, the transmit power is fixed per sub-carrier. This yields a certain modulation type to be applied for each sub-carrier for each down-link phase, as a fixed transmit power results in a certain SNR for the sub-carrier/terminal pair, which determines the modulation type. Therefore, from a computational point of view, this approach requires the least resources at the access point.

All five approaches can be further characterized. If neither dynamic sub-carrier assignments are performed nor dynamic power distribution, one yields the purely static approach (in FDM). If no dynamic sub-carrier assignments are performed but dynamic power distribution is enabled, the static FDM approach with bit and power loading results. Contrarily, if dynamic sub-carrier assignments are performed but no dynamic power distribution, this matches optimization problem (4.5). If in addition also dynamic power distribution is enabled, problem (4.4) is obtained (or the raw bit rate maximization of problem (4.3)) (see Figure 4.2).

The goal of this investigation is to quantify the optimal performance of the five schemes, assuming an example parameterization of a wireless cell. Two different metrics are considered, as the five approaches do not have the same objective functions. The first metric is the average throughput per terminal in the cell (which is the objective function of problem 4.3 and of the second comparison scheme, the static sub-carrier assignment with loading). The second metric is the largest average transmission delay for conveying a data packet (which is the objective function of the two considered rate adaptive optimization approaches). It is obtained from the minimum throughput per cell. It is straightforward to obtain from this average minimum throughput per cell the maximum transmission delay for conveying a data packet of a certain size. In the following, the size of a typical IP packet – $1500\ \mathrm{byte}$ – is assumed.

For the given cell setting (parameterization is given below), the five schemes are compared for an increasing cell radius (while keeping all other parameters fixed). As the cell radius increases, the distance between the terminals and the access point increases. Terminals are assumed to be equally distributed over the area of the cell. If the cell radius increases, the average attenuation per terminal increases, too. However, some terminals remain quite close to the access point (due to the assumed distribution) while the distance between the access point and (most) other terminals becomes larger. Thus, the spread of the average attenuation between close terminals and terminals far away increases. As a result, it becomes more difficult to equally maximize the throughput of all terminals in the cell. This applies only to the optimal solution of the rate adaptive problems ((4.4) and (4.5)).

Methodology

For each down-link phase, instances of the optimization problems have to be solved (where the instance is a parameterized optimization problem). The solution for such an instance of one of the discussed optimization problems is a certain number of bits that can be transmitted per down-link phase to each terminal. However, the introduced metrics are averages over a longer time period. In order to obtain these system level results, the following procedure is conducted. Initially, channel trace files of the attenuations of each sub-carrier are generated from a computer program (refer to Section 2.1). Each attenuation sample consists of path loss, shadowing and fading components, as parameterized below. One sample is generated for every down-link phase and sub-carrier per terminal.

Once the trace file of the attenuations is generated, for each down-link phase the attenuation matrix is obtained. Based on this matrix, the instances for the corresponding optimization problems can be built. A solver for linear (integer) programming problems called CPLEX [96] is used. As input, CPLEX requires a special file format which can be generated by a second tool, called ZIMPL [97][3]. After solving the linear integer programs, the resulting assignments are collected by a script. The script keeps record of each frame throughput. Typically, each trace file represents a few seconds (equaling a few thousands of down-link phases). Finally, the average throughput per terminal is obtained by dividing the total amount of bits transmitted to each terminal by the simulated time.

[3] I gratefully thank Torsten Koch from ZIB for enabling the usage of ZIMPL

Scenario Parameterization

The following scenario is chosen: The center frequency is assumed to be $f_c = 5.2$ GHz while the system bandwidth is set to $B = 16.25$ MHz. This bandwidth is split into $N = 48$ sub-carriers. Each OFDM symbol has a duration of 4 μs including a guard interval of $T_g = 0.8$ μs. These parameters correspond to the WLAN standard IEEE 802.11a [32].

The sub-carrier attenuations $h_{j,n}^{(t)}$ are generated based on the three components path loss, shadowing and fading. For the path loss a standard model, as discussed in Section 2.1.1, is used. The parameters are chosen as $K = -46.7$ dB and $\alpha = 2.4$, according to a large open space propagation environment. For the shadowing independent stochastic samples from a log-normal distribution are assumed, characterized by a zero mean and a variance of $\sigma_{\text{sh}}^2 = 5.8$ dB. The samples are regenerated every second. While the path loss and shadowing affect all sub-carriers of a terminal equally, each sub-carrier experiences an individual fading component. Each sample of the fading process is assumed to be Rayleigh-distributed. The frequency and time correlation of the samples are characterized by a Jakes-like power spectrum and an exponential power delay profile. The Jakes-like power spectrum is parameterized by the maximum speed within the propagation environment $v_{\max}$ and the center frequency f_c. A maximum speed of $v_{\max} = 1\text{m/s}$ is assumed. The exponential power delay profile is characterized by the delay spread $\Delta\sigma$. It is set to $\Delta\sigma = 0.15$ μs. The noise power σ^2 is determined by considering the noise in a receiver at an average temperature of 20° C over the bandwidth of a sub-carrier $\frac{B}{N}$ (see Section 2.1.2).

For the adaptive modulation scheme, a target symbol error probability of $p_{\text{sym,max}} = 10^{-2}$ is chosen for the payload transmission. $M = 4$ (BPSK, QPSK, 16-QAM and 64-QAM) modulation types are available. Given a specific SNR value, the adaptive modulation always chooses the modulation type with the highest data rate with respect to the target symbol error probability.

Performance Evaluation

In Figure 4.3 the average throughput results are given for a cell with $J = 4$ active terminals, where the cell radius increases from 10 m to 200 m. As the radius increases, the average throughput decreases in all considered schemes. The lowest performance is provided by the static scheme with adaptive modulation. The best performance is provided by the raw capacity maximization approach (Problem (4.3)). Comparing the two different dynamic versions of the rate adaptive maximization, (obviously) applying dynamic power and sub-carrier assignments provides a higher throughput than applying only dynamic sub-carrier assignments with adaptive modulation. Also, the static approach with adaptive modulation is outperformed by the static approach with bit and power loading.

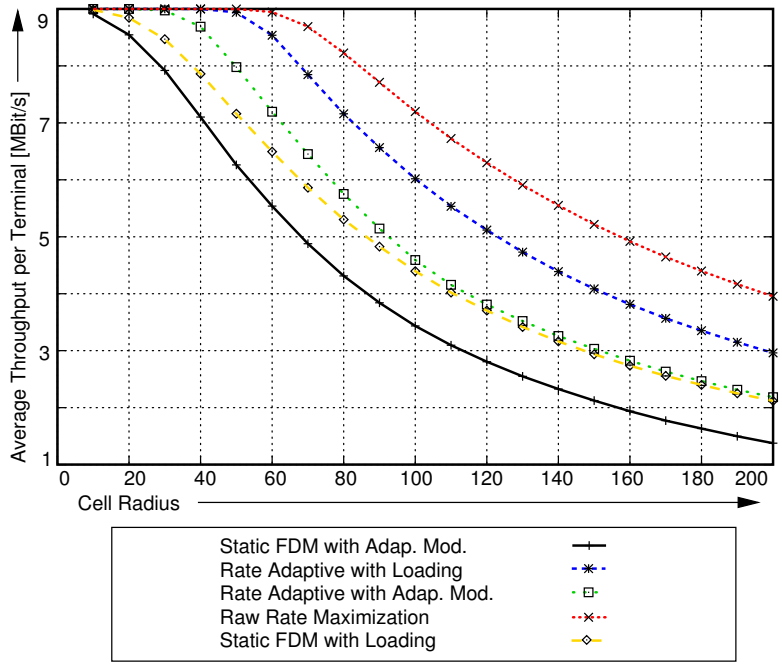
In particular, all dynamic schemes provide the same average throughput up to a radius of 30 m. Up to a radius of 50 m both dynamic schemes applying dynamic power and sub-carrier assignments provide identical performance (regarding the average throughput). In all these cases the available transmit power in combination with the dynamic adaption schemes enables the usage of the highest modulation type (in this study 64-QAM) on all sub-carriers. Therefore, the dynamic schemes would benefit from the usage of even higher modulation types (256-QAM, for example). This would lead to even higher throughput values. Once all schemes do not pro-

vide the highest possible throughput any more (at about $80\ \mathrm{m}$), a constant throughput difference exists between the schemes. Switching from the static scheme with adaptive modulation to the static scheme with bit loading yields a throughput increase of about $0.5\ \mathrm{MBit/s}$ per terminal. Switching from the static scheme with bit loading to the optimal performance provided by the solution to Problem (4.5) yields a throughput increase of again about $0.5\ \mathrm{MBit/s}$ per terminal. This difference is larger for smaller cell sizes (at $50\ \mathrm{m}$, for example), while it is smaller for larger cell sizes (at $150\ \mathrm{m}$, for example). Switching from the rate adaptive approach with adaptive modulation to the one with dynamic power allocation provides a throughput gain of about $1.2\ \mathrm{MBit/s}$. Finally, the raw capacity maximization approach yields an additional performance gain of about $0.9\ \mathrm{MBit/s}$.

In addition, a huge difference of these schemes exist considering the upper limit for the transmission delay. The corresponding graph is also shown in Figure 4.3. Here, the maximum delay for the transmission of one IP packet of size $1500\ \mathrm{byte}$ is given. Apart from the raw capacity maximization approach, the static scheme with adaptive modulation provides also the worst performance - the delay increases drastically with an increasing cell radius. If bit loading is added to the static scheme, the delay decreases significantly, especially for larger cells. Compared to this, the delay limit for the two rate adaptive approaches is much lower, but still increasing for an increasing cell radius. The rate adaptive approach with dynamic power assignments outperforms the rate adaptive approach with adaptive modulation. However, the performance difference is only marginal.

The delay results indicate how the rate adaptive approach with dynamic power assignments obtains its throughput advantage, compared to the rate adaptive approach with adaptive modulation: The throughput of the terminals farer away does not significantly differ, since the delay limits would differ more in that case. Instead, the throughput of the terminals close to the access point differs significantly.

In general, it is observed that the static scheme with adaptive modulation provides the worst performance. Adding dynamic power distribution (bit loading) improves this performance for the static scheme, especially for large cells. Using the rate adaptive approach with adaptive modulation outperforms both static variants, while the rate adaptive approach with dynamic power distribution even outperforms the one without bit loading. In particular, note that adding bit loading to the system approach yields a higher performance increase in combination with dynamic sub-carrier assignments, compared to static sub-carrier assignments. In Figure 4.4 the throughput and delay results are shown for a larger number of terminals in the cell, here with $J = 8$ terminals. Consider the throughput results first. In general, the qualitative behavior is the same as with the case of $J = 4$ terminals in the cell. The performance of the static approach with adaptive modulation is the worst, maximizing the raw capacity achieves the best average throughput per terminal. In between, the rate adaptive approach with dynamic power assignment outperforms the rate adaptive approach with adaptive modulation. Again, the rate adaptive approach with adaptive modulation outperforms the static approach with bit loading. As the radius increases, the average throughput drops for all five schemes. However, the throughput differences between the schemes changes. First, the rate adaptive approaches perform better compared to the static scheme with adaptive modulation and bit loading. Second, the rate adaptive approach with adaptive modulation performs better in general compared to the rate adaptive



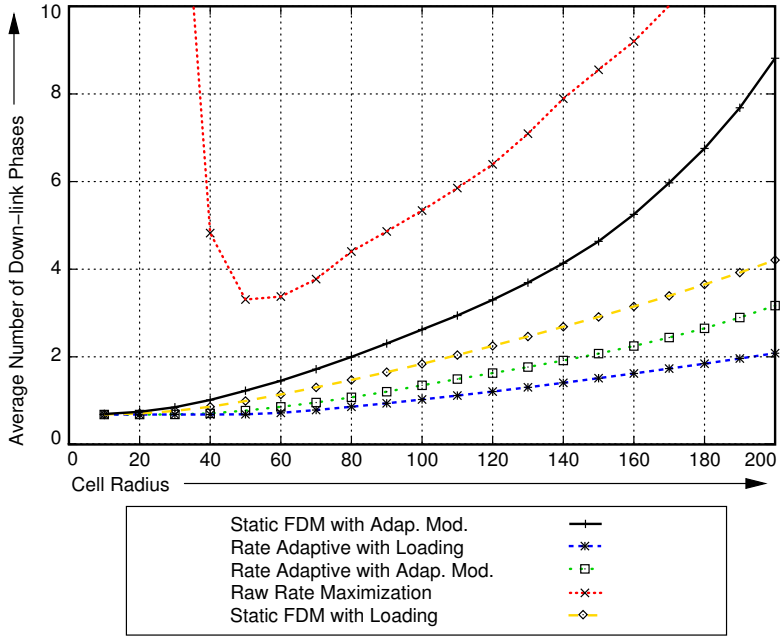
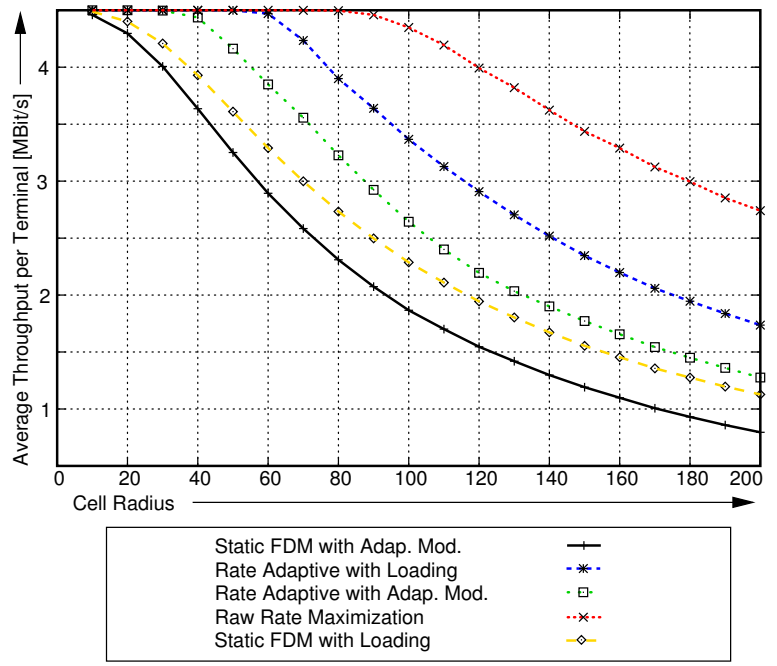


Figure 4.3: Average throughput and maximal delay results for five different system approaches for an increasing cell radius and $J = 4$ terminals in the cell.

approach with dynamic power assignments. Third, the raw capacity maximization outperforms all other schemes much more as in the case with a smaller number of terminals. These results become even more evident if one considers the upper transmission delay limits in the cell. Here, the two rate adaptive approaches are quite close to each other while the two static approaches provide a much larger delay. Note that the transmission delay limits of the raw capacity maximization are well beyond the ranges of this delay plot[4]. For small and medium cell sizes, the performance difference between the two rate adaptive approaches is rather small (delay limits are almost identical, throughput results are closer in terms of percent than in the case of a lower number of terminals present). However, for larger cell sizes the performance differences increases. As the number of terminals increases even more, this qualitative and quantitative behavior continues. In Figure 4.5 the delay and throughput results are given for $J = 16$ terminals in the cell. The two rate adaptive approaches achieve for smaller cell sizes performance results, which are quite close to each other (the percentage gap becomes smaller compared to the cases with less terminals in the cell). The static approach with adaptive modulation is significantly outperformed by all dynamic approaches while the raw capacity maximization approach outperforms all other schemes even more. The static approach with bit loading improves the performance of the purely static approach, however, the increase is much smaller than in the cases with the less terminals in the cell. In contrast, the performance difference between the two rate adaptive approaches increases. Thus, bit loading pays off even more for dynamic scheme with sub-carrier assignments, if many terminals are located in the cell. This is demonstrated more accessible in Figure 4.6, where the spectral efficiency[5] is shown for the five different schemes while the number of terminals increases. The graph on top of this figure shows the results for a small cell radius of $50\ \mathrm{m}$, while the graph below shows the values for a larger cell size (radius of $100\ \mathrm{m}$). In a small cell, the spectral efficiency increases moderately for the static approach with adaptive modulation while it increases stronger for the rate adaptive approach with adaptive modulation. For the other rate adaptive approach and the raw capacity maximization the spectral efficiency increases only slightly, as almost all assigned sub-carriers have the highest possible modulation type. In this case, these schemes would benefit from the availability of a higher modulation type. For the static approach with bit loading, the spectral efficiency increases first, but then decreases. However, the gap between all considered schemes is rather low. As the radius increases (radius of $100\ \mathrm{m}$ shown in the lower graph of Figure 4.6) the gap between the raw capacity maximization approach and the other approaches increases, while in general the gap between the two rate adaptive approaches decreases with an increasing number of terminals. In Figure 4.7 the spectral efficiency is shown for a relatively large cell radius of $150\ \mathrm{m}$. Note in particular the large gap between the raw capacity maximization approach and all other ones (this gap increases for an increasing number of terminals in the cell). Also, note that the rate adaptive approach with dynamic power assignment achieves almost twice the spectral efficiency of the static scheme with adaptive modulation, whereas the other rate adaptive approach achieves a spectral efficiency, which is higher by about 50%, compared to the static approach with adaptive

[4] In several cases the minimal average throughput is zero, as the terminals far away never receive a sub-carrier. Thus, computationally the delay limit is infinity.

[5] In this study, the spectral efficiency is measured by the amount of bits transmitted on average per second, divided by the bandwidth. As the down-link has a duration of only half a frame, the instantaneous spectral efficiency *per down-link phase* is higher by a factor of 2.



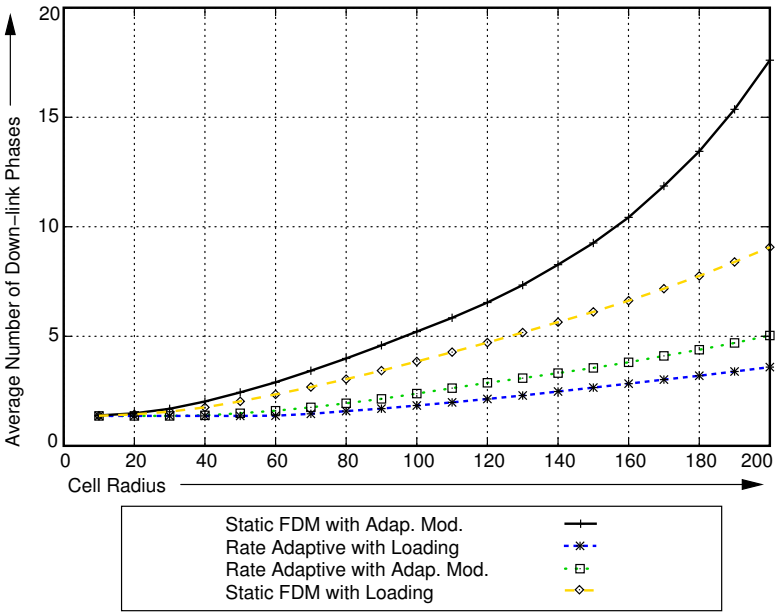
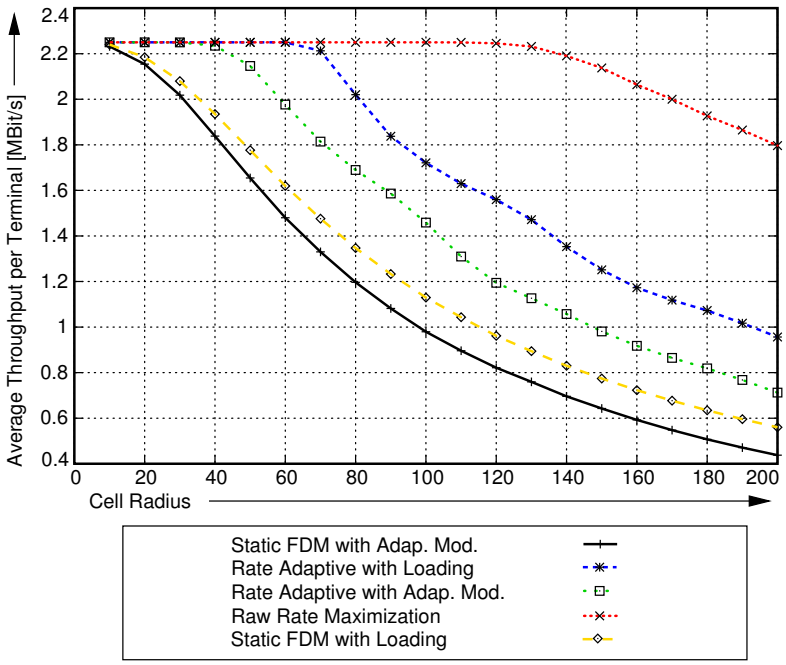


Figure 4.4: Average throughput and maximal delay results for five different system approaches for an increasing cell radius and $J = 8$ terminals in the cell.



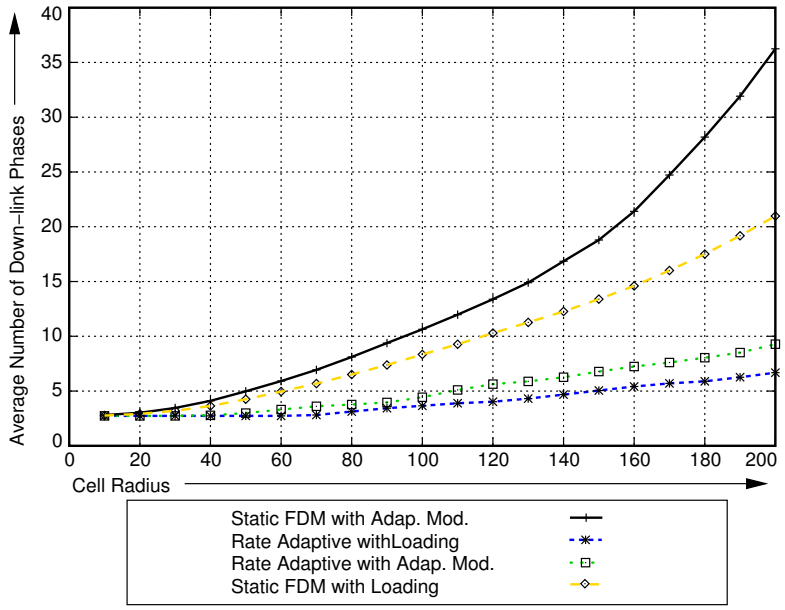
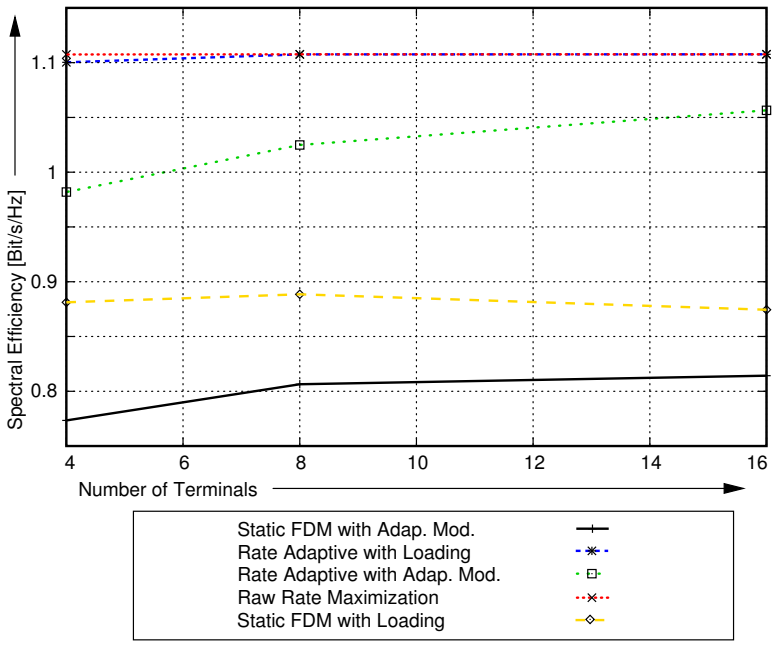


Figure 4.5: Average throughput and maximal delay results for five different system approaches for an increasing cell radius and $J = 16$ terminals in the cell.



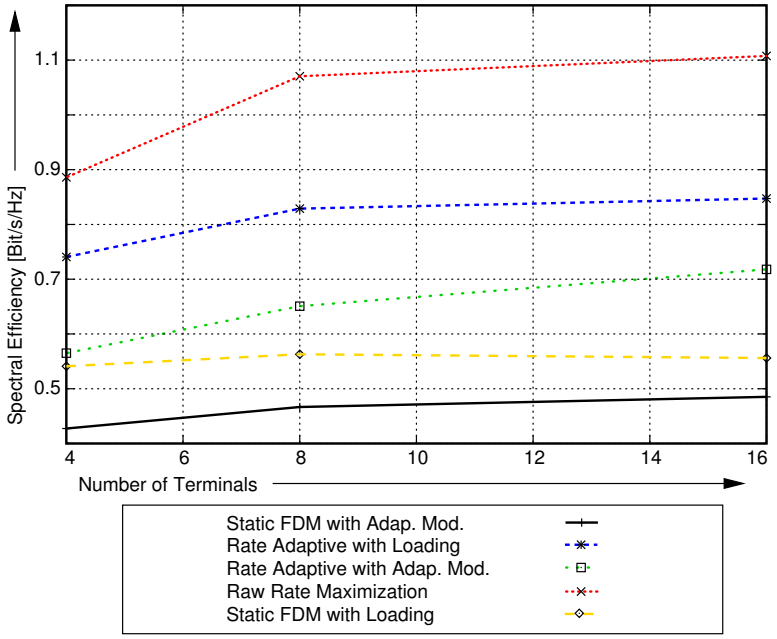


Figure 4.6: Spectral efficiency of the five different approaches for an increasing number of terminals in the cell at a cell radius of 50 m (top) and 100 m (below).

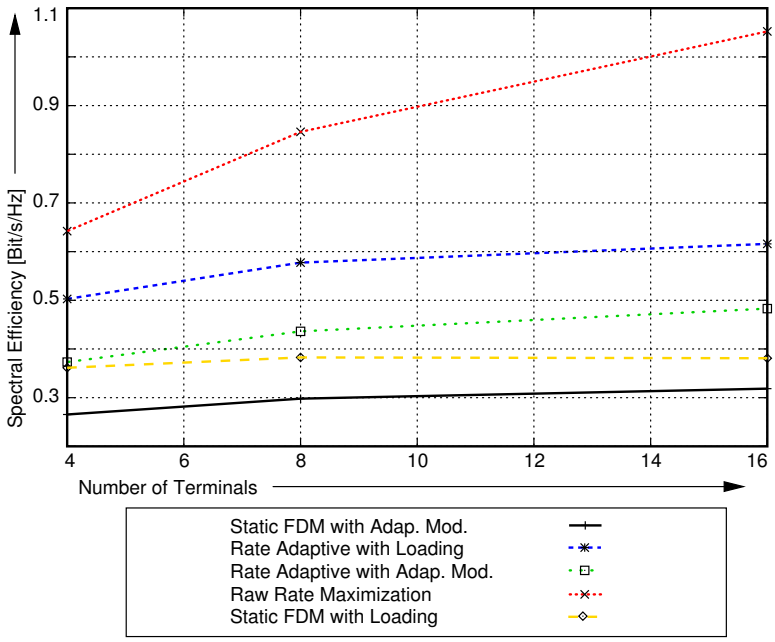


Figure 4.7: Spectral efficiency of the five different approaches for an increasing number of terminals in the cell at a cell radius of 150 m.

modulation.

Summary

Optimally solving the so far discussed dynamic schemes achieves an impressive performance advantage in terms of throughput, compared to static approaches. Dynamic schemes benefit in particular from a higher number of terminals in the cell due to multi-user diversity. Maximizing the capacity without respect to the delay of packet transmissions works well in terms of throughput. However, in almost all cases, terminals that are far away do not obtain much throughput, leading to severe problems regarding the delay. Comparing the performance of the two rate adaptive approaches, the one with dynamic power assignments always outperforms the one with adaptive modulation. If the cell size is rather small and a large number of terminals is located in the cell, this performance difference is rather small, for larger cells (> 100 m), or for a small number of terminals in the cell, the dynamic power adaption pays off well in terms of average throughput per terminal. Regarding the delay, both schemes perform quite the same. Thus, especially for larger cell sizes, the rate adaptive approach with dynamic power assignments should be considered as dynamic MU-OFDM approach. However, this also requires a higher computational power at the access point (despite the fact that both rate adaptive approaches are *NP*-complete). Considering dynamic power distribution for static schemes, it can be stated that it pays off, especially in larger cells with a low number of terminals. For a high number of terminals, using bit and power loading in combination with a static sub-carrier assign-

ment provides only a slight performance increase. In general, bit and power loading performs much better with dynamic sub-carrier assignments than without.

4.2 Suboptimal Schemes for the Rate Adaptive Optimization Problem

In Section 4.1.2 it was shown that the optimal performance (with respect to the rate adaptive approach) is achieved by a jointly dynamic assignment of sub-carriers and power. However, the optimization problem (4.4) has been shown to be *NP*-complete. Thus, suboptimal schemes have to be applied in practical systems. In this section, several suboptimal algorithms are evaluated. Two of them have been proposed recently in the literature while the third scheme is a novel contribution.

In the literature it is proposed to simplify the rate adaptive optimization problem by splitting the whole process into sub-carrier and power assignment phases [76, 90, 74]. First, the sub-carrier assignments are performed, then the transmit power is allocated. For both steps, two issues have to be determined: First, how much sub-carriers and/or transmit power is assigned to each terminal. Second, what is the specific assignment, i.e. in case of the sub-carriers, which sub-carriers obtains each terminal specifically. Note that once the power budget and the sub-carrier assignments per terminal are determined, the power assignment per sub-carrier is similar to the point-to-point connection loading problem (3.12). This problem can be solved by the Hughes-Hartogs algorithm (refer to Section 3.1), for example.

The following two schemes, proposed in the literature, are considered in this section (see also Figure 4.8):

- **Kim et al.**: Kim et al. suggest in [74] to first perform the dynamic sub-carrier assignment according to the maximum weighted bipartite matching approach (optimization problem (4.6)). However, they modify the objective function such that the sum of the assigned reciprocal attenuation values is minimized (in contrast to maximizing some bit rate function):

$$\begin{aligned} \min_{\mathbf{X}^{(t)}} \quad & S \cdot \sum_{j,n} \frac{x_{j,n}^{(t)}}{\left(h_{j,n}^{(t)}\right)^2}. \\ \text{s. t.} \quad & \sum_{j} x_{j,n}^{(t)} \leq 1 \qquad \forall\, n \\ & \sum_{n} x_{j,n}^{(t)} \leq \frac{N}{J} \qquad \forall\, j \end{aligned} \tag{4.6}$$

  This different objective function penalizes the assignment of attenuated sub-carriers (with an attenuation coefficient below 1) extraordinarily, as the reciprocal values are considered. In addition, they propose to allocate each terminal the same number of sub-carriers prior to the assignment (second constraint in optimization problem (4.6)).

  After the sub-carrier assignments have been performed, the transmit power is loaded to the assigned sub-carriers iteratively. For each iteration step, one additional bit is allocated to

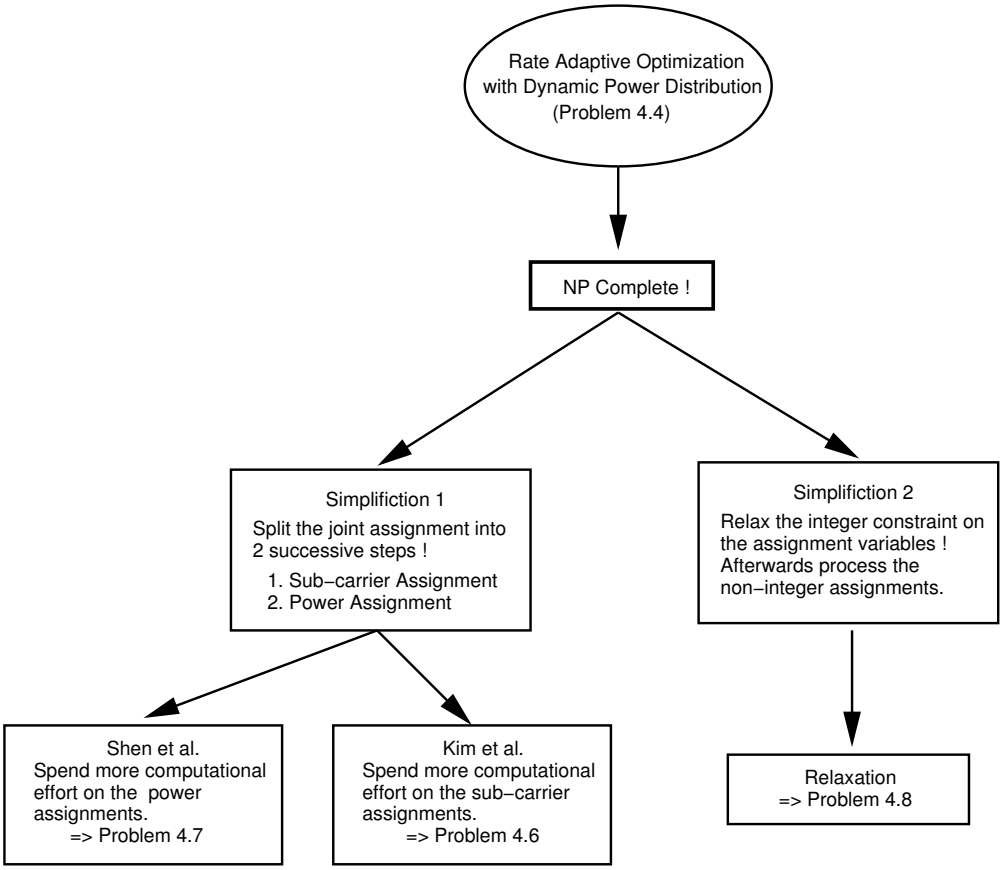


Figure 4.8: Three different suboptimal algorithms considered for the rate adaptive problem with dynamic power distribution.

each terminal. For each terminal this bit is always loaded to the sub-carrier which requires the least additional transmit power. This loading process is terminated once all transmit power is allocated.

Note that the sub-carrier assignment problem has a complexity of $O\left(N^3\right)$, while the loading process has a complexity of $O\left(N \cdot M\right)$.

- **Shen et al.**: Here the authors propose to first perform a sub-carrier assignment by some heuristic method proposed by Rhee et al. in [77] (refer to Section 3.2.2, Algorithm 5). For this algorithm a static power distribution is assumed.

  After generating the sub-carrier assignments, the transmit power is reassigned [90]. Assigning the transmit power is split into two steps: First, determine a power budget per terminal, given the sub-carrier assignments. Then, the allocated power per terminal is loaded to the sub-carriers. Originally, Shen et al. propose a non-linear optimization problem to be solved for determining the power budget per terminal. The solution strategy suggested is based on Lagrange multipliers. Instead of considering this rather complex method, a different approach for power loading is considered in this section. Once the sub-carrier assignments have been fixed, a rate adaptive optimization can be performed which *only* distributes the transmit power. This can be formulated as integer programming problem:

$$\begin{aligned} \max_{\mathbf{Y}^{(t)}} \quad & \epsilon \\ \text{s. t.} \quad & \sum_{j(n),m} y_{j,n,m}^{(t)} \leq 1 \quad \forall\, n \\ & \sum_{j,n,m} p_{j,n,m}^{(t)} \cdot y_{j,n,m}^{(t)} \leq P_{\text{max}} \\ & S \cdot \sum_{n \in N_j} \sum_{m} r\left(m\right) \cdot y_{j,n,m}^{(t)} \geq \epsilon \quad \forall\, j \quad , \end{aligned} \tag{4.7}$$

  where $j\left(n\right)$ denotes the terminal which has been assigned to each sub-carrier n, and N_j denotes the set of sub-carriers assigned to terminal j. Note that this combination of sub-carrier and power assignments is an upper bound to the published scheme in [90] regarding this system model.

  In this case, more complexity is put to the power assignment (the exact computational complexity of Equation 4.7 is not known) while the sub-carrier assignment has a complexity of $O\left(N^2\right)$.

Both suboptimal approaches first perform the dynamic assignment of sub-carriers and then perform power and bit loading. In contrast, it is proposed here to perform a jointly dynamic (suboptimal) assignment of power and sub-carriers, in the following referred to as **Relaxation**. It is based on the relaxation of the integer constraint of variable $y_{j,n,m}^{(t)}$ in problem (4.4). The resulting optimization problem is given below:

$$
\begin{aligned}
\max_{\mathbf{Y}^{(t)}} \quad & \epsilon \\
\text{s. t.} \quad & \sum_{j,m} y_{j,n,m}^{(t)} \leq 1 \quad \forall\, n \\
& \sum_{j,n,m} p_{j,n,m}^{(t)} \cdot y_{j,n,m}^{(t)} \leq P_{\text{max}} \\
& S \cdot \sum_{n,m} r\,(m) \cdot y_{j,n,m}^{(t)} \geq \epsilon \quad \forall\, j \\
& 0 \leq y_{j,n,m}^{(t)} \leq 1 \quad \forall\, j,n,m
\end{aligned}
\tag{4.8}
$$

This is a linear optimization problem, which can be solved by efficient algorithms. However, the solution to the relaxed optimization problem (4.8) will most likely be not applicable to a practical OFDM system, as several terminals are assigned shares of the same sub-carriers. Also, each terminal might be assigned multiple modulation shares to be applied during the next downlink phase, which cannot be realized. Thus, the relaxed solution has to be processed further.

A fairly simple rounding scheme is suggested initially. Each sub-carrier is assigned to the terminal/modulation combination with the highest share. It is important to note that usually this approach will not lead to a feasible solution either, as the transmit power constraint of problem (4.4) is possibly violated. The rounding might lead to a too high overall transmit power. Thus, the assignments have to be further improved. If the transmit power is too high, the next step is to reduce the modulation type of some sub-carrier for the terminal which has the highest throughput (under the currently infeasible rounded assignment). Then the transmit power is recalculated. If it is still too high, the terminal with the currently highest throughput is chosen and the modulation type of its sub-carriers is decreased by one constellation (for example, reducing from QPSK to BPSK). After a few such modifications, a feasible solution is obtained.

In the context of the rate adaptive optimization problem (regarding problem (4.5) as well as regarding problem (4.4)) the relaxation approach has not been previously proposed. In general, relaxing an integer programming problem and subsequently rounding the obtained result is known as one (somewhat reasonable) approach to solve integer optimization problems. In the context of other optimization problems in OFDM systems (i.e. the margin adaptive optimization approach), relaxation has been studied in [49, 75]. However, for the corresponding optimization problems other suboptimal schemes have been proposed which achieve a better performance. Thus, so far relaxation is not known as a particular suitable suboptimal assignment scheme for dynamic multiuser OFDM systems.

Apart from these three suboptimal schemes, the optimal solution to problem 4.4, referred to as **Rate Adaptive with Loading**, is considered as comparison scheme. Also, the static solution discussed in Section 4.1.2 as **Static FDM with Loading** is considered in the evaluation part.

Performance Evaluation

The same system model and scenario parameterization is used as in Section 4.1.2. As performance metric the average throughput and the maximum delay is considered (see also Section 4.1.2). All five approaches are studied for a varying number of terminals and a varying cell radius.

Initially, consider the average throughput and transmission delay results of the five different schemes for an increasing cell radius with $J = 4$ terminals (Figure 4.9). For all five schemes, the average throughput per terminal decreases as the radius increases. For example, the average throughput of the optimal solution to problem (4.4) decreases from about $9\ \mathrm{MBit/s}$ down to $3\ \mathrm{MBit/s}$ for the largest considered radius. This is clearly due to the increasing attenuation as the radius increases. The performance gap between the static approach and the optimal rate adaptive approach is about $2\ \mathrm{MBit/s}$ for small cells and $1\ \mathrm{MBit/s}$ for large cells. The graphs of the suboptimal schemes are in between. The scheme by Shen et al. provides the lowest average throughput of all dynamic schemes evaluated in this section. It outperforms the static scheme but is lower by about $0.5 - 1\ \mathrm{MBit/s}$ compared to the optimal performance. In contrast, the suboptimal scheme by Kim et al. performs quite well in terms of the average throughput per terminal. Compared to the optimum, the performance is lower by at most $0.3\ \mathrm{MBit/s}$. However, the best suboptimal scheme is the relaxation scheme. For a cell radius larger than $40\ \mathrm{m}$ it provides an average throughput which is at most $0.1\ \mathrm{MBit/s}$ below the optimum one. Next, consider the maximum delay of the cell for this setting with $J = 4$ terminals (lower graph of Figure 4.9). Clearly, the maximum transmission delay of the cell increases as the cell radius increases. This is simply due to the fact that the distance between the access point and terminals is increasing, which increases the average attenuation. The delay of the static scheme is outperformed by all dynamic schemes. Basically, the static scheme provides a delay which is twice as large as the delay of the dynamic schemes. Out of the group of the suboptimal schemes, the relaxation approach provides the lowest maximum delay. It is only slightly larger than the minimum delay. The approach by Shen et al. provides the worst dynamic delay. The approach by Kim et al. is in between. However, the difference of all four dynamic schemes is at most half of a down-link phase.

If more terminals are present in the cell, these qualitative results stay more or less the same, but the quantitative differences change. Consider first the case with $J = 8$ terminals in the cell (Figure 4.10). Regarding the average throughput and maximum transmission delay, the relaxation scheme remains the best suboptimal scheme. The second best is the approach by Kim et al. while the approach by Shen et al. provides the worst performance of all dynamic schemes. However, the performance difference between the relaxation scheme and the approach by Kim et al. increases in terms of the average throughput and in terms of the maximum delay metric. This is due to the fact that the relaxation scheme now provides a performance which is closer to the optimal one. In contrast, the approach by Kim et al. and the approach by Shen et al. provide worse performance (in terms of percentage of the optimum) as the number of terminals increase. As the number of terminals increases even further ($J = 16$ in Figure 4.11), the performance difference between the three suboptimal schemes increases, too. Especially for larger cell sizes, the relaxation approach now provides a substantially larger performance than the other two suboptimal schemes. In addition, the performance of the relaxation approach

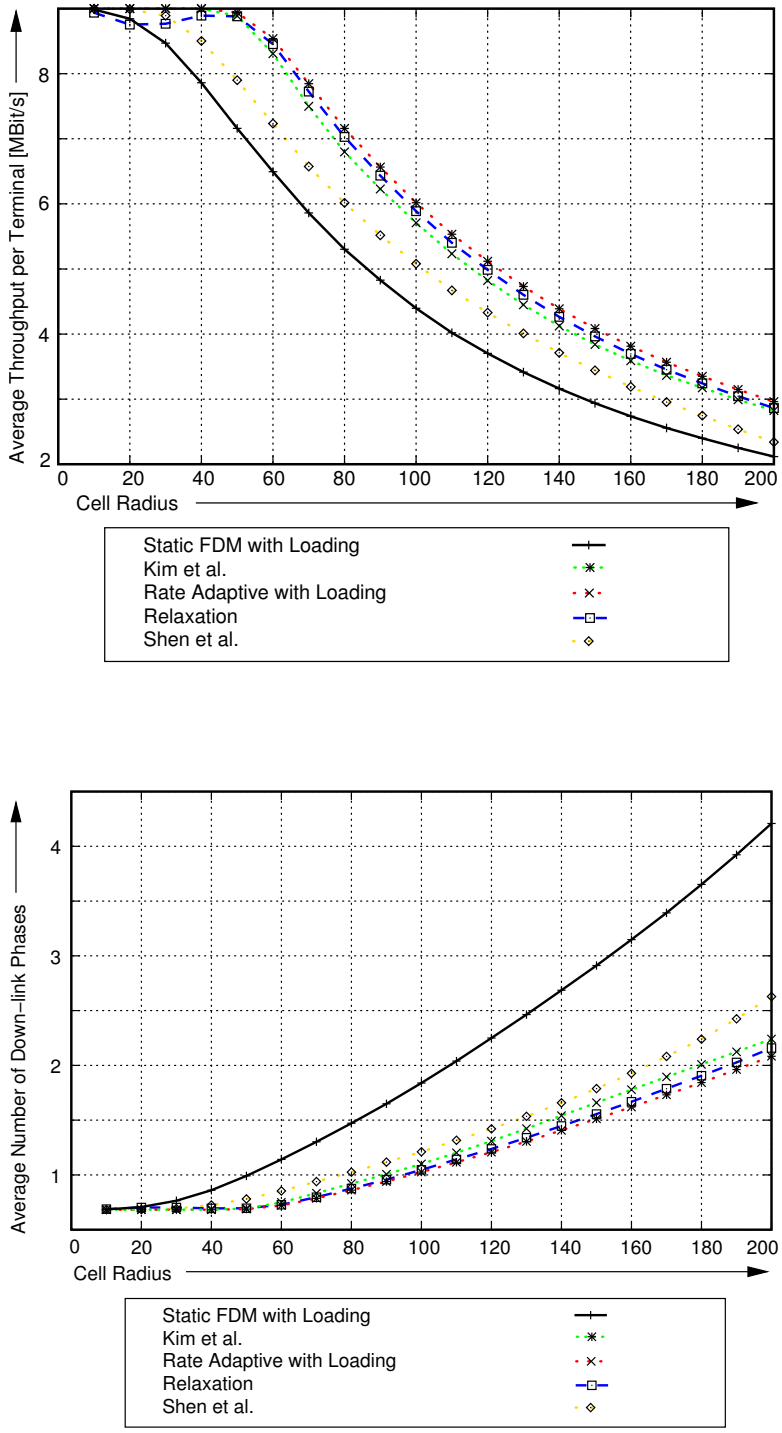
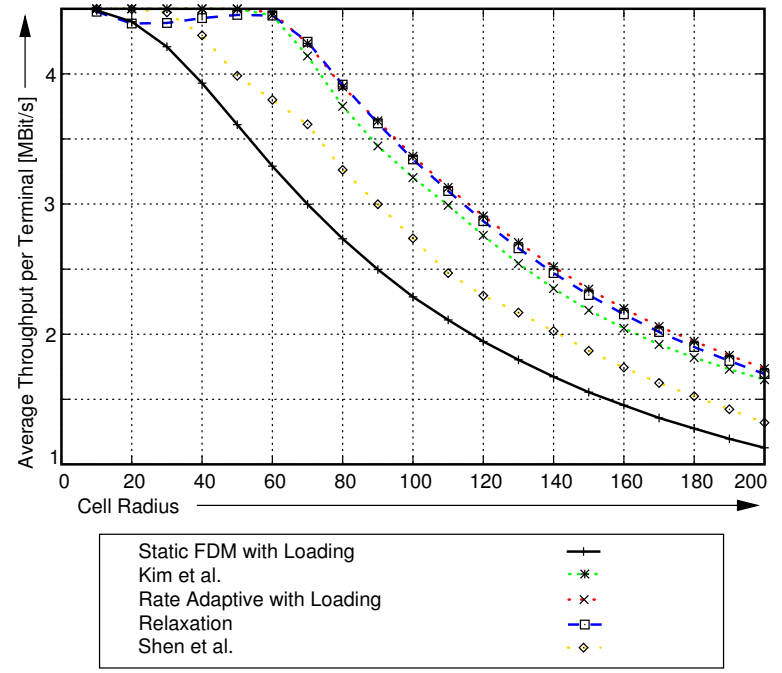


Figure 4.9: Average throughput and maximal delay results for five different system approaches with dynamic power distribution for an increasing cell radius and $J = 4$ terminals in the cell.



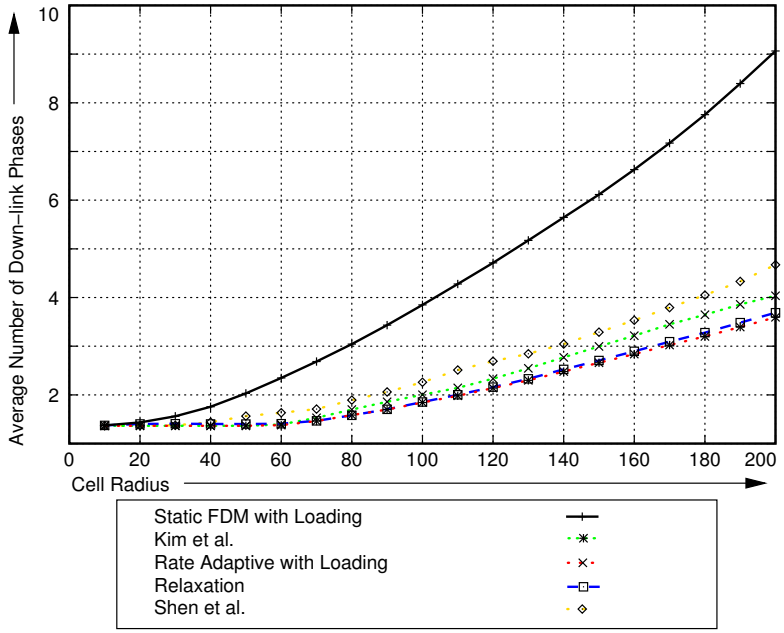


Figure 4.10: Average throughput and maximal delay results for five different system approaches with dynamic power distribution for an increasing cell radius and $J = 8$ terminals in the cell.

is always close to the optimum. Regarding the average throughput, it even outperforms the optimum solution to the rate adaptive problem for cell sizes of $70\ \mathrm{m}$ up to $120\ \mathrm{m}$[6]. The approach by Shen et al. provides in contrast now a significantly worse performance compared to the previous two settings.

Summary

For the rate adaptive optimization problem with dynamic power distribution the relaxation approach provides the best suboptimal performance of all three evaluated schemes. This is mainly due to the fact, that the assignment of sub-carriers and transmit power is not split (as it is for the approaches of Kim et al. and Shen et al.) but is considered jointly.

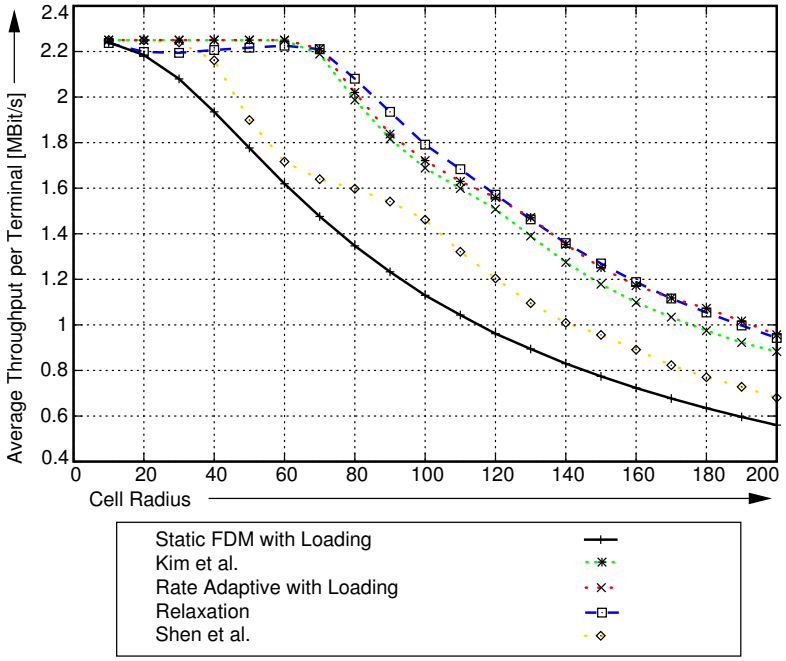
Interestingly, these results are in contrast to the application of the relaxation approach to other multiuser OFDM optimization problems, as studied in [49, 75]. There, the margin adaptive optimization problem (Section 3.2.1). has been considered. The relaxation approach performs for the margin adaptive problem well, but is significantly outperformed by other heuristic approaches. As shown above, regarding the rate adaptive problem the relaxation approach provides almost the optimum solution in all cases and can therefore be recommended for the solution of the rate adaptive problem.

Comparing the two other suboptimal schemes reveals an answer to the question: Where to spend most of the computational effort, if the joint assignment of power and sub-carriers is split into two separate steps. The approach of Kim et al. has a quite simple transmit power assignment scheme, while the sub-carrier assignment scheme is more demanding (in terms of computational resources). In contrast, Shen et al. stress the distribution of transmit power more while they spend less computational effort on the sub-carrier assignment scheme. From the above results it can be concluded clearly that spending more computational effort on the sub-carrier assignment outperforms schemes where more computational effort is spent on the power assignment.

4.3 Suboptimal Schemes for the Rate Adaptive Problem with Adaptive Modulation

As demonstrated in Section 4.1.2, the solution to the rate adaptive optimization problem with power loading achieves the best performance. Thus, dynamic multiuser OFDM systems should always try to apply a dynamic power distribution in order to achieve a high performance. However, there might be cases where a dynamic power distribution is not an option due to the absence of control circuits for a dynamic power distribution or due to regulatory issues. In these cases,

[6]The performance differences are significant, as the 0.99 confidence interval is about 0.5% (the interval is not shown in the graphs). Also, note that the maximum average transmission delay of the relaxation approach is *smaller* than the one of the optimum approach. Per down-link phase, the optimum approach achieves the best transmission delay. However, the optimal approach assigns the largest transmission delay always to the same terminal (in the simulations) while the relaxation approach distributes the maximum delay over several terminals for several down-link phases. Therefore, the maximum *average* transmission delay over multiple terminals might be smaller in the case of the relaxation approach. Per down-link, though, the optimal approach provides always the smallest delay.



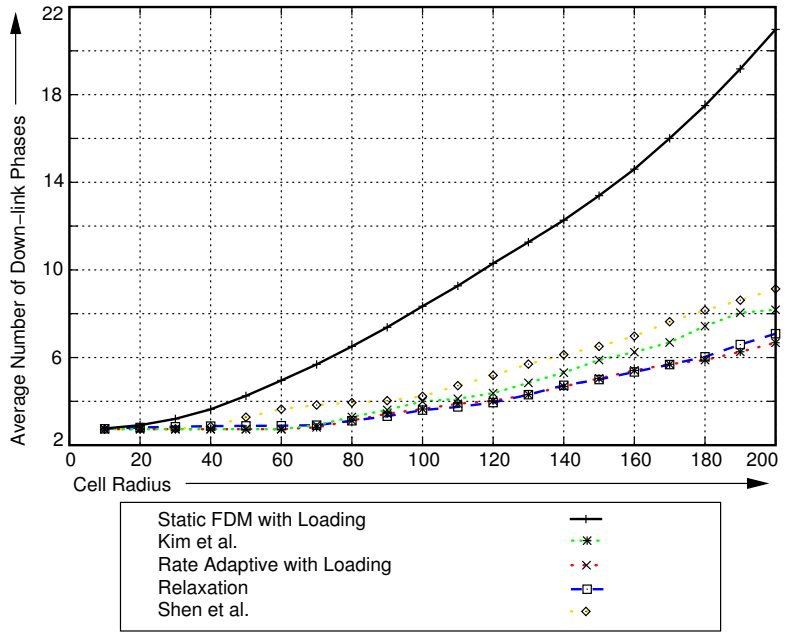


Figure 4.11: Average throughput and maximal delay results for five different system approaches with dynamic power distribution for an increasing cell radius and $J = 16$ terminals in the cell.

a static power distribution has to be applied, while still dynamic sub-carrier assignments can be performed. In addition, adaptive modulation might still be used to exploit the varying SNR values per sub-carrier and terminal pair. It has been shown in Section 4.1.1 that the corresponding rate adaptive optimization problem with a static power distribution (Problem (4.5)) is also *NP*-complete. Hence, also in this case suboptimal schemes have to be applied in practical systems.

In this section, three suboptimal algorithms for the rate adaptive optimization problem with static power distribution are evaluated (see Figure 4.12). These schemes are (simplified) counterparts of the suboptimal schemes of the previous section, where the rate adaptive optimization problem with dynamic power distribution has been discussed. In particular, the following two schemes from the literature are compared to the relaxation scheme for the rate adaptive optimization problem:

- **Two-Step Approach**: The first approach is to solve problem (4.5) by splitting the sub-carrier assignment process itself into two different ones: First, each terminal is allocated a certain number of sub-carriers. In the second step, each terminal is assigned this number of sub-carriers such that the total throughput is maximized. This was suggested first by Yin et al. [76]. Denote the number of allocated sub-carriers to terminal j at time t by $m_j^{(t)}$. For the assignment problem, the following optimization problem is obtained:

$$\begin{aligned} \max_{\mathbf{X}^{(t)}} \quad & S \cdot \sum_{j,n} b_{j,n}^{(t)} \cdot x_{j,n}^{(t)} \\ \text{s. t.} \quad & \sum_{j} x_{j,n}^{(t)} \leq 1 \qquad \forall\, n \\ & \sum_{n} x_{j,n}^{(t)} \leq m_j^{(t)} \qquad \forall\, j \end{aligned} \tag{4.9}$$

  In fact, this optimization problem can be solved in polynomial time. In graph theory, this optimization problem equals the weighted bipartite matching problem. Today, this problem is known to have a complexity of $O\left(N^3\right)$ [46].

  The sub-carrier allocation has to be done before executing the assignment algorithm (for details, see Algorithm 4). The authors of [76] propose to perform the allocation based on the rate requirements and the average sub-carrier attenuation of each terminal. However, the allocation scheme proposed in that paper suits the margin adaptive optimization problem, as a rate requirement per terminal is needed to perform the allocation. Thus, in this section an equal allocation for all terminals is assumed.

- **Heuristic by Rhee et al.**: The second approach considered is a heuristic (see Algorithm 5) developed by Rhee et al. [77]. Basically, this algorithm always "supports" the weakest terminal. After initialization, the throughput values for all terminals are compared. Then, the terminal with the lowest throughput for this down-link phase receives the next sub-carrier (chosen from a list of still available sub-carriers), which is the sub-carrier with the highest rate for the chosen terminal. The authors propose to initially assign each terminal one

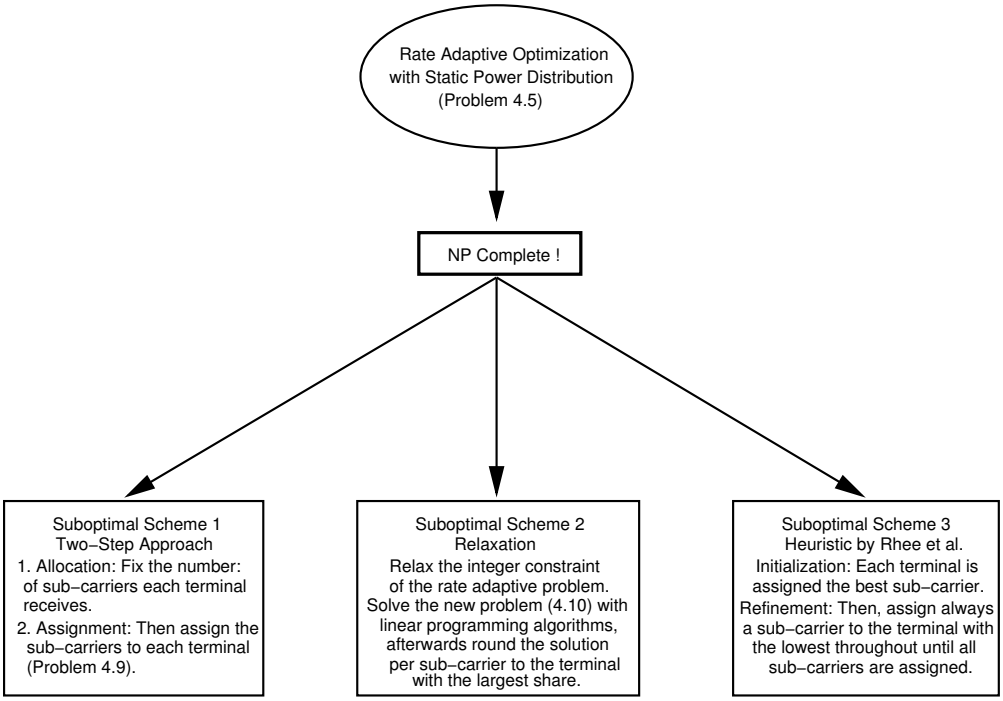


Figure 4.12: Three different suboptimal algorithms considered for the rate adaptive problem with static power distribution.

sub-carrier, the currently best one. This algorithm continues until all sub-carriers are distributed for this down-link phase. Obviously, this algorithm has a polynomial complexity of $O\left(N^2\right)$, which is the lowest one of all three considered schemes.

Relaxation: As in the previous section, it is proposed to solve the rate adaptive optimization problem by initially relaxing the integer constraint on the sub-carrier assignments (see Equation 4.10). Then, the resulting optimization problem becomes a linear programming problem, which can be solved efficiently, for example by the simplex algorithm. After the optimal real-valued solution has been generated, the solution for each sub-carrier is rounded to the terminal with the largest assignment fraction. In this case, such a rounding scheme always yields a feasible solution.

$$\begin{aligned} \max_{\mathbf{X}^{(t)}} \quad & \epsilon \\ \text{s. t.} \quad & \sum_{j} x_{j,n}^{(t)} \leq 1 \quad \forall\, n \\ & S \cdot \sum_{n} b_{j,n}^{(t)} \cdot x_{j,n}^{(t)} \geq \epsilon \quad \forall\, j \\ & 0 \leq x_{j,n}^{(t)} \leq 1 \quad \forall\, j,n \end{aligned} \tag{4.10}$$

Performance Evaluation

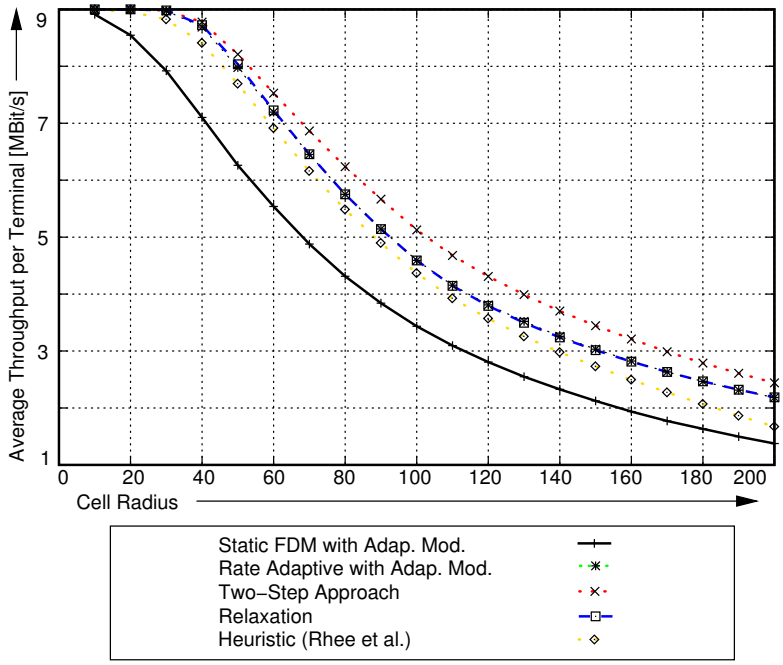
The same system model and scenario parameterization is used as in Section 4.1.2. As performance metric the average throughput and the maximum delay is considered. Apart from the three above introduced schemes, also the static scheme with adaptive modulation (**Static FDM with Adaptive Modulation**) and the optimal solution to Problem (4.5) (**Rate Adaptive with Adaptive Modulation**) are considered as comparison schemes. For the two-step approach, a static allocation is chosen: Each terminal receives exactly the same number of sub-carriers during all down-link phases. The approaches are studied for a varying number of terminals and a varying cell radius.

In Figure 4.13 the results are shown for the average throughput and the maximum transmission delay in the case of a small number of terminals ($J = 4$). While the radius increases, all schemes have a decreasing average throughput. All three suboptimal schemes have a significantly higher throughput than the static scheme. The two-step approach achieves the highest average throughput, providing about $0.3\ \mathrm{MBit/s}$ more than the optimal solution to the rate adaptive problem with adaptive modulation. Contrarily, the heuristic suggested by Rhee et al. provides especially for larger cell sizes an average throughput below the optimal solution. The third suboptimal method, the rounded solution of the relaxed problem, provides an average throughput which is almost exactly the same as the optimal one.

Considering the maximum delay (lower graph Figure 4.13) shows that the maximum delay for the relaxation approach is also quite similar to the optimal solution. Thus, the relaxation scheme works quite well in this scenario. In contrast, the maximum delay for the other two suboptimal methods are slightly higher (but still below the maximum delay of the static variant). For large cells, the two-step approach and the heuristic provide an almost identical maximum delay (even though the two-step approach provides a much higher average throughput for these cell sizes). For smaller cells, the maximum delay of the heuristic is quite similar to the optimal one, which is slightly below the maximum delay of the two-step approach.

As the number of terminals in the cell increases, the qualitative behavior of the different schemes stays the same. In Figure 4.14, the throughput and delay results are given for $J = 8$ terminals in the cell. Note that the two-step approach achieves a slightly higher average throughput than in the case of $J = 4$ terminals (compared to the optimal solution of Problem (4.5)). Again, the relaxation scheme achieves a similar average throughput to the optimal solution. The heuristic also provides a throughput comparable to the one of the optimal solution. Considering the maximum delay, the heuristic and the two-step approach provide a slightly larger maximum delay than the relaxation scheme (which provides an identical delay as the optimal integer solution). If the terminal number is even higher, this qualitative behavior continues. In Figure 4.15, the average throughput and delay results are shown for $J = 16$ terminals in the cell. First, note that the twos-step approach has an even higher throughput gain compared to all other schemes. Second, note that the relaxation scheme does not provide the same throughput any more of its integer counterpart. Instead, for smaller cell sizes the throughput is slightly higher (see also the corresponding footnote of Section 4.2).

For the maximum transmission delay, the three different schemes provide only slightly different results. The relaxation scheme has the lowest maximum delay, which is almost as good as the optimal one. The two-step approach provides a slightly higher maximum delay, while the



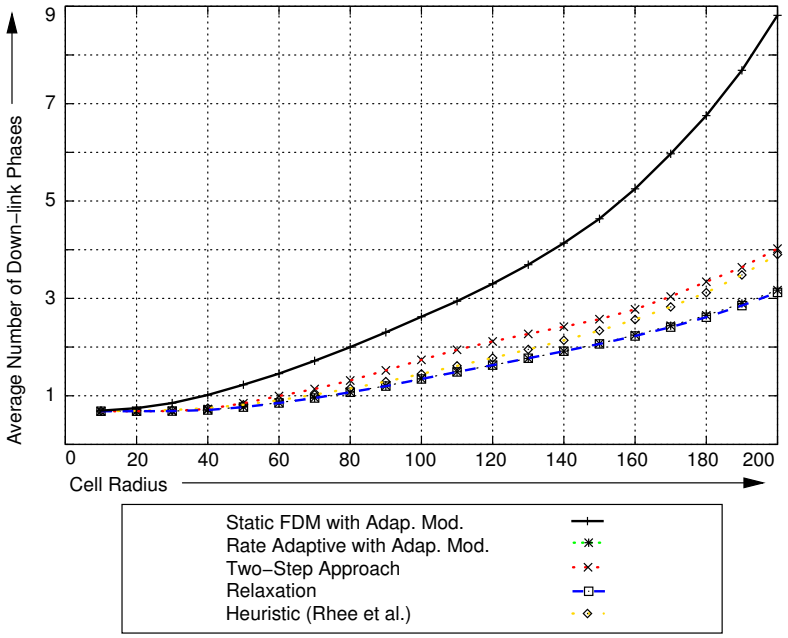
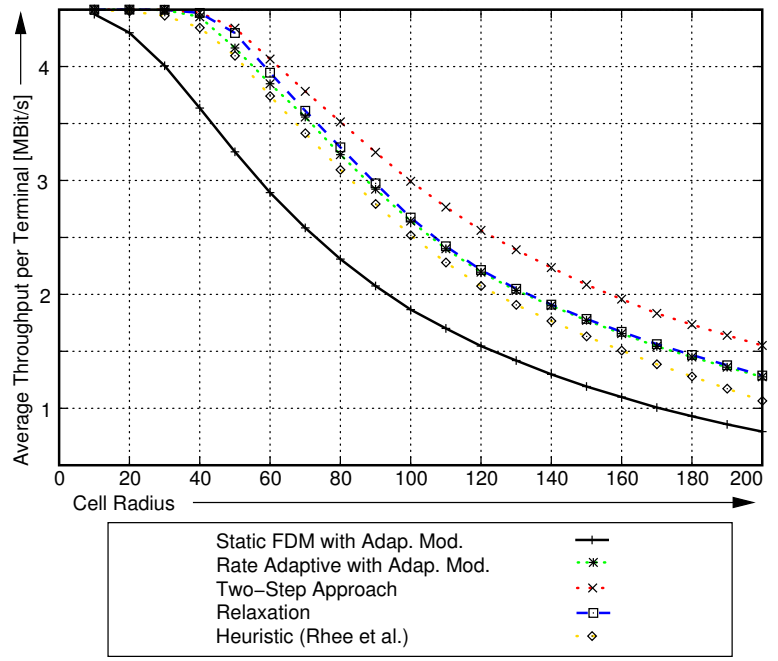


Figure 4.13: Average throughput and maximal delay results for five different system approaches for an increasing cell radius and $J = 4$ terminals in the cell.



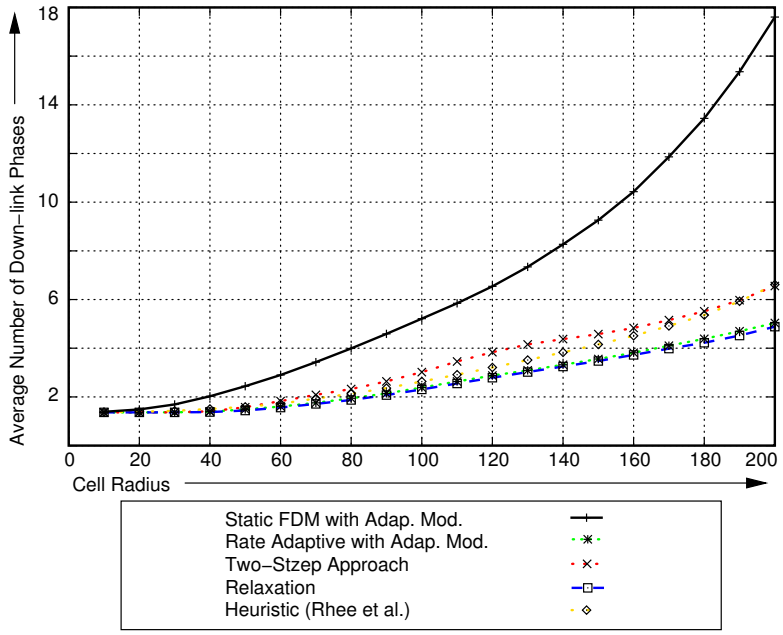


Figure 4.14: Average throughput and maximal delay results for five different system approaches for an increasing cell radius and $J = 8$ terminals in the cell.

heuristic has the highest delay. Note that all three suboptimal approaches still have a much lower delay than the static approach.

Summary

If no power loading can be performed, two suboptimal schemes qualify in particular in order to distribute sub-carriers dynamically while taking the delay of packets transmitted in the cell into account. The relaxation approach achieves a performance almost similar to the one of the optimal solution. This is true for both, the average throughput and the maximum delay of the cell. If the number of terminals is high, the performance equivalence does not hold up that tight any more, but still a satisfactory performance is provided.

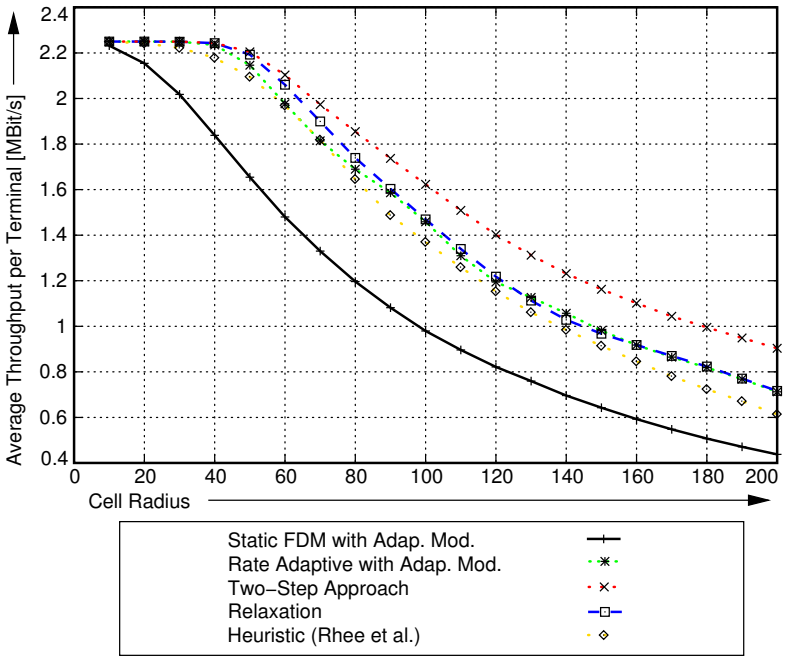
The two-step approach provides a significantly higher throughput. However, the maximum delays of packet transmissions are higher, too. This might be adjusted by a different allocation (in this study an equal sub-carrier allocation has been assumed), providing a higher throughput for terminals farer away from the access point. However, this would decrease the average throughput. Originally, the authors suggest to allocate sub-carriers according to the average attenuation, which would favor the terminals farer away of the access point. It is reasonable to assume that a different allocation could lead to a performance closer to the optimal one (which means: A lower throughput while having better, i.e lower, delays).

Compared to the other two approaches, the heuristic by Rhee et al. provides the worst performance. In all cases, a lower throughput than the one for the solution of problem (4.5) is provided, while the delay is the highest for all three evaluated schemes. Especially for a large number of terminals in the cell, the average throughput varies strongly.

4.4 Assignment Schemes for Heterogeneous Packet Streams

The rate adaptive approach aims at maximizing the cell throughput, while providing each terminal in the cell with the optimal rate in each down-link phase. However, in reality data is often transmitted in form of packets to the access point. Thus, the amount of data that has to be forwarded to each terminal varies per down-link phase as different streams might have totally different packet interarrival time statistics as well as different packet size distributions. Some terminals might even have no data queued at the access point at all at the moment. This leads to a potential under- or overprovisioning by the rate adaptive approach. Instead of achieving an equal rate for all terminals in the cell, in the following a scheme is proposed where the access point distributes the transmit resources according to the current queue size of each terminal. For two sub-optimal algorithms of Section 4.3, the two-step approach of Problem (4.9) and the relaxation approach of Problem (4.10) (and their corresponding counterparts in case of a dynamic power distribution), a queue size dependent modification is presented. This can also be applied to cases were the transmit power is distributed dynamically. Note that the discussion assumes that fragments of packets can be transmitted per down-link phase. In contrast, a packet-centric[7] transmission system would require a different optimization model.

[7] Packet centric refers to a system where only complete packets are transmitted.



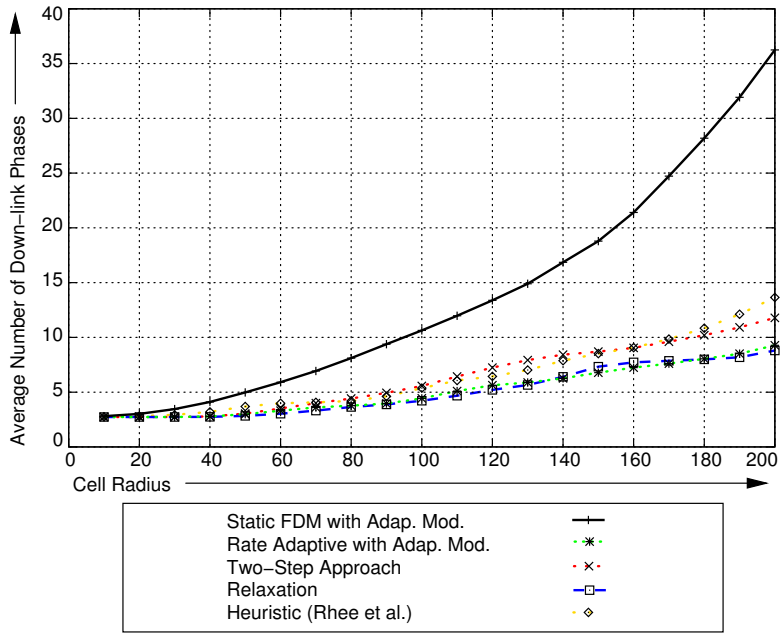


Figure 4.15: Average throughput and maximal delay results for five different system approaches for an increasing cell radius and $J = 16$ terminals in the cell.

Adapting the assignments to the queue size of each terminal has several advantages. The current queue size aggregates several parameters into a single one. It depends on the packet arrival process at the access point. However, it also depends on the sub-carrier attenuations as well as on the previous dynamic assignments. Consider a large amount of packets queued at the access point for some terminal. The queue size is large because either a lot of packets arrived recently, or because the terminal is currently at a "bad" position, meaning that the attenuation of the sub-carriers is high. As this queue size increases in comparison to the queue sizes of other terminals, this terminal should receive a higher rate.

The adaption to different queue sizes is based on the assumption that the access point knows how much data it currently queues for each terminal in the cell. Denote the amount of data queued at time t for terminal j as $q_j^{(t)}$. In order to adapt the two-step approach, the sub-carrier allocation (i.e. the number of sub-carriers) for each terminal is related to the queue size. If a terminal has a lot of data in the queue compared to other terminals, it receives more sub-carriers for the next down-link phase. The following scheme is proposed. Initially, each terminal with data in its queue larger than a certain threshold (or with a certain delay limit closer than some bound) is allocated one sub-carrier. Assume J^* such terminals to receive one sub-carrier according to this initial iteration. The remaining $N - J^*$ sub-carriers are then distributed according to the relative queue size, i.e. the fraction of each terminals queue size divided by the overall amount of data queued for the J^* terminals. This procedure yields an overall sub-carrier allocation as given below:

$$\forall\, j \in J^* \quad : \quad l_j = 1 + \left\lfloor (N - J^*) \cdot \frac{q_j^{(t)}}{\sum_{j \in J^*} q_j^{(t)}} + 0.5 \right\rfloor \quad . \tag{4.11}$$

This allocation enables a flexible way to deal with varying data streams. As such streams have quite different packet arrivals, the transmit capacities can be dynamically distributed such that the diversity of the packet arrival process of different streams can be fully exploited (as discussed below in the performance evaluation part).

For the relaxed rate adaptive approach, a different method is proposed. Here, the relaxed optimization problem itself is modified. Out of the J^* terminals which have data queued at the access point, the terminal ($\tilde{j}$) with the smallest data amount queued at the access point is chosen. Then, the queued data of each other terminal is related to the smallest queue size, yielding the factor:

$$\alpha_j^t = \frac{q_j^{(t)}}{q_{\tilde{j}}^{(t)}} \quad . \tag{4.12}$$

This factor is used in the relaxed optimal problem to scale the minimal rate assigned to each terminal (Problem 4.13).

$$
\begin{aligned}
\max_{\mathbf{X}^{(t)}} \quad & \epsilon \\
\text{s.t.} \quad & \sum_{j} x_{j,n}^{(t)} \leq 1 \quad \forall\, n \\
& S \cdot \sum_{n} b_{j,n}^{(t)} \cdot x_{j,n}^{(t)} \geq \alpha_j^t \cdot \epsilon \quad \forall\, j \\
& 0 \leq x_{j,n}^{(t)} \leq 1 \quad \forall\, j, n
\end{aligned}
\tag{4.13}
$$

As in the case of the two-step approach, a threshold should be defined when a queue is large enough to be considered. If there are many queues with a very small data amount (for example, only one voice over IP (VoIP) packet) over-provisioning might still occur because even the assignment of only one sub-carrier might yield a rate which is twice as large as required. In this case, several small packets for *different* terminals could be multiplexed on one sub-carrier.

If the frame length is rather short compared to the interarrival times of the data at the access point, the usage of these two adaption schemes for packet data transmissions will enable the flexible usage of the system resources according to the need of each terminal. No complicated "notification process" is required to register a data stream at the access point (which could be used to obtain the average bit rate per stream). Also, no physical layer information is required (for example, the average attenuation per sub-carrier per terminal) to distribute the resources among the terminals (as proposed in [76]). Simply considering the queue length aggregates all this information in a single parameter, which is easy to obtain at the access point.

Performance Evaluation

In the following the same centralized system model and parameterization as in Section 4.3 is assumed (where the transmit power is fixed; furthermore the radius is set to $100\ \mathrm{m}$). However, each terminal receives now a stream of packets. The real-time streaming of video from a remote server over the wireless access to the mobile terminals is considered. Each terminal receives a variable bit rate stream of data packets. These packets are transmitted from the source via the wired backbone and the access point to each terminal. At the access point the packets are queued and are transmitted afterwards in first-in first-out order. Each terminal has a separate queue at the access point and only receives a single stream.

As video source a **M**oving **P**icture **E**xperts **G**roup (*MPEG*) -4 video coded with an average bit rate of $723\ \mathrm{kbit/s}$ is chosen. The video file (Figure 4.16) has a frame rate of 25 Hz and a length of 3 minutes; during this time terminals roam through the cell and the sub-carrier attenuations vary constantly. Each terminal receives the same video stream, however they are randomly time-shifted at the beginning so that different parts of the video stream arrive at the access point for different terminals. A time span of $400\ \mathrm{ms}$ is assumed as maximum end-to-end delay for the transmitted video stream. The forwarding of the packets through the backbone of each stream is considered to consume some time, though. The remaining time span serves as transmission deadlines for the packets in the queue of each terminal. Three specific deadlines are chosen: $100\ \mathrm{ms}$, $175\ \mathrm{ms}$ and $250\ \mathrm{ms}$.

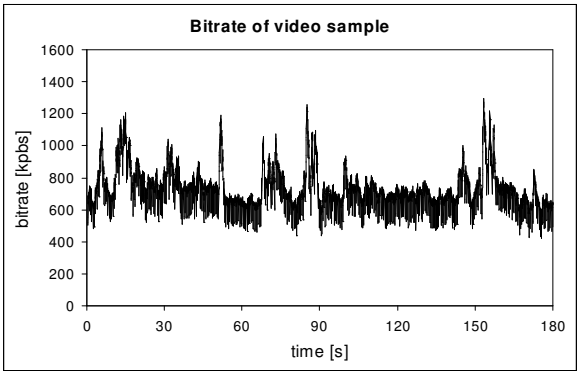


Figure 4.16: Time-varying bit rate of the encoded video source used for the simulations.

In order to battle bit errors, block codes are employed to correct a certain number of bit errors per packet. If the packet is still erroneous after applying the error correction, a simple retransmission mechanism sends the packet again from the access point to the terminal during one of the following down-link phases. Therefore, video quality degradation can only be caused by receiving outdated packets or not receiving these packets at all (if a packet is not transmitted successfully at the end of its deadline, it is dropped without being transmitted). In this section,

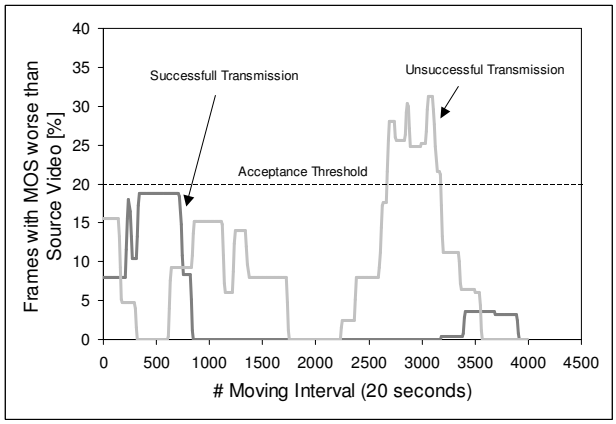


Figure 4.17: Exemplification of video quality metric; Successful transmission means the percentage of frames worse than the original must be less than 20% within all intervals

the number of "satisfied" users is considered as performance metric. A user is satisfied if the received and decoded video stream is above a certain quality threshold. This depends clearly on the subjective impression of the user. Describing the human quality impression by a subjective quality metric is usually done with a "mean opinion score" (MOS), on a scale from 5 (best) to 1 (worst) as defined by ITU [98]. This MOS is generated for every frame in the received (possibly

distorted) video and compared with the MOS values of the frames of the undistorted encoded video[8]. It must be noted that not only the overall percentage of frames with a worse MOS value influences the user impression, but also the distribution of bad phases. To address this fact, the percentage of worse frames within a window sliding over the whole duration of the video is considered. Only if the percentage of frames with a worse MOS is small in *each* interval, the video quality is rated as good (Figure 4.17). This method is explained in detail in [99]. As threshold a percentage value of 20% is chosen.

Three different assignment approaches are considered:

1. **Static Approach**: In the static case each terminal in the cell receives a fixed set of sub-carriers. The size of each sets is also fixed. If 6 terminals are located in the cell, each one receives 8 sub-carriers (as there are 48 sub-carries in total) per down-link phase. However, per down-link phase the modulation types per sub-carrier are varied, depending on the channel attenuation (as the access point has the knowledge of the sub-carrier attenuations for each down-link phase). Regarding the queue dynamics no adaptive mechanism is considered.

2. **Dynamic Assignment / Static Allocation**: In this approach the sub-carriers are assigned dynamically based on sub-carrier attenuations. However, the allocations are not dynamic. The access point allocates each terminal a fixed number of sub-carriers. The assignments are performed according to Problem 4.9. In fact, a heuristic algorithm is applied[9]. Note that this variant benefits from the sub-carrier attenuation diversity. However, it does not consider the queue dynamics, which leads sometimes to over- or underprovisioning, as packets are sometimes dropped if they have not been transmitted within the required delay or a terminal is assigned more throughput than data is queued at the access point.

3. **Dynamic Assignment / Variable Allocation**: Finally, in the fully dynamic approach the allocation and the assignment is dynamic. The allocation is performed depending on the queue sizes, as stated in Equation 4.11. Then, the assignments are performed, using the same algorithm as for variant 2.

In Figure 4.18 the capacity results — the number of maximal "satisfied" terminals — for the different simulation settings is shown. In the static case (no dynamic allocation, no dynamic assignment of subcarriers) the cell might support up to 6 terminals. Purely switching to a dynamic assignment scheme with still fixed sub-carrier allocation already results in up to 12 terminals which can be supported. The suggested sub-carrier allocation scheme in combination with the dynamic sub-carrier assignment algorithm is able to support up to 18 terminals in the cell. All three combinations obviously benefit from a longer remaining end-to-end delay in absolute terms, the capacity results are higher the longer the deadline is. Figure 4.19 shows the increase of the dynamic combinations compared to the static scheme. Here it is clearly seen

[8]The MPEG-4 coding process already introduces some degradation. Therefore, the quality is not evaluated based on the original video, but on a encoded and decoded version of it.

[9]The heuristic achieves almost the same performance as the optimal solution to Problem 4.9. The gap is found in simulations to be at most 5% [100]. However, in this section the focus is on the dynamic allocation for variable packet streams, not on developing a strict upper bound.

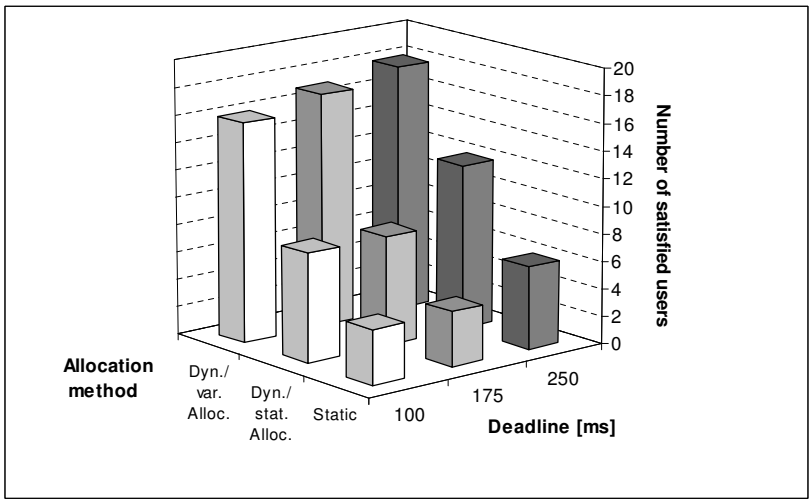


Figure 4.18: Comparison of static and variable sub-carrier allocation scheme with static and dynamic sub-carrier assignments

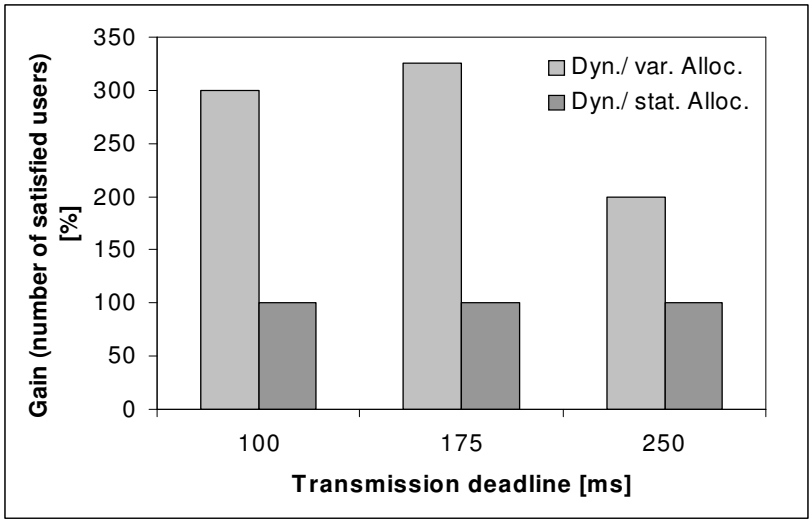


Figure 4.19: Capacity gain of the variable and static subcarrier allocation scheme combined with dynamic subcarrier assignment compared to the static allocation and assignment scheme for different deadlines

that for short deadlines the dynamic sub-carrier allocation improves the situation. Keeping the allocation method fixed and purely switching the assignment method from static to dynamic always yields a capacity increase of 100%, not depending on the deadlines. However, changing the allocation method to the one proposed here yields an additional capacity increase which is higher for short deadlines and is up to an additional 200%.

Notice that the performance results are larger than in the case of the performance results of the previous sections. This is due to the delay-sensitive data, but also due to the better throughput distribution achieved by the dynamic allocation. This gain can be improved by adaptive queuing and forwarding schemes, hence by semantic scheduling. For example, more packets carrying information of more "important" video streams could be prioritized. After these frames have been successfully forwarded, less important frames would be conveyed. This has been investigated in detail in [101].

4.5 Conclusions

As the rate adaptive and margin adaptive approach are both *NP*-complete, suboptimal schemes have to be applied to assign sub-carriers to terminals. The computational complexity of the rate adaptive problem does not change, even if the transmit power per sub-carrier is fixed: The rate adaptive problem remains *NP*-complete.

On the other hand, if only the optimal performance is considered, multiple issues have to be taken into account. In general, a dynamic power assignment pays off only if the sub-carriers are also assigned dynamically. The performance increase achieved by switching from a purely static system to a system that only assigns the transmit power dynamically is quite low. In contrast, the performance gain achieved by switching from fixed to dynamic power distribution is significant, if the sub-carriers are assigned dynamically as well. Especially for large cell sizes with a particular large attenuation spread between the terminals, a fully dynamic approach achieves a performance twice as large as the one of a fully static approach.

However, due to the complexity of the rate adaptive optimization problem, suboptimal schemes have to be applied. Three different schemes are investigated, two of them have been proposed in the literature, while the third scheme has not been previously proposed for the rate adaptive optimization problem. In general, this third scheme, referred to as relaxation scheme, is found to provide the best performance of all suboptimal schemes. For all studied scenarios, the performance of the relaxation scheme is very close to the optimal one. Therefore, the relaxation scheme can be recommended as suboptimal solution for the rate adaptive optimization problem.

In addition, a special case of the rate adaptive optimization problem is studied where the transmit power is fixed. Several constraints, such as the available hardware design or regulatory issues, may lead to this special application scenario. In this case, several suboptimal schemes perform quite well. One of them is the relaxation scheme, the other one is the two step approach.

Several issues remain as future work. First of all, the actual run times of the proposed suboptimal algorithms have to be investigated. During this study, the run times of all three schemes have been found to be around $10\ \mathrm{ms}$. The relaxation approach seems to consume the most computation time, while the other two approaches (Kim et al. and Shen et al.) require less time. Clearly, a time span of $10\ \mathrm{ms}$ is too large for the application in any dynamic OFDM

system. Even a time span of $1\ \mathrm{ms}$ seems quite large. However, the focus of this study has not been on actual run times of the algorithms. Therefore, it remains to investigate how the run times of the algorithms could be reduced. For example, in case of the relaxation approach one could imagine to exclude certain parts of the simplex algorithm, if possible. In addition, specialized hardware might speed up the process significantly.

A second issue is to develop an optimization model for dynamic OFDM systems with packet assignments, taking into account different quality of service classes as well as individual delays per packet. The optimization models developed in this thesis do not differentiate between packets of different quality of service classes. Per terminal, the overall rate is maximizes (with respect to the rates of the other terminals). Instead, if packets are to be matched to sub-carriers, a second integer variable enters the model. The advantage of such a model would be the optimal support for packet based scheduling, where the over- or underprovisioning[10] is minimal. However, it requires a complete new model, a new complexity analysis and possibly the development of other suboptimal schemes.

[10]Overprovisioning results from maximizing the rate without considering the actual data queue. Therefore, it might happen that some terminal receives more capacity per down-link phase than data is currently queued. This can be reduced by introducing weights, as proposed in Section 4.4. However, only a an optimization model which considers packet assignments can reduce overprovisioning to a minimum.

Chapter 5

The Signaling Overhead in Dynamic MU-OFDM Systems

The access point generates periodically new assignments These assignments have to be conveyed to the terminals, since the terminals do not know a priori which sub-carriers they have been assigned. Signaling is an essential necessity for any dynamic MU-OFDM system. Obviously, it decreases the performance of such a system. The performance loss depends on the update period (among a couple of other factors). In dynamic MU-OFDM systems this update period is driven primarily by the coherence time of the sub-carrier attenuations. As the coherence time is typically in the range of a few milliseconds, about 10^2 to 10^3 updates have to be conveyed per second. Another event requiring signaling is a change in the number of active terminals in the cell. For example, a terminal which is new to the cell has to be assigned system resources for payload transmission. This event causes a need for signaling in dynamic systems as well as in static systems which assigns fixed blocks (or, generally speaking, a fixed subset) of sub-carriers to terminals. However, the update frequency due to a change in the number of terminals is much lower compared to the update frequency due to a change of the wireless channels. The number of terminals in a cell varies rather in the range of seconds or even minutes and not in the range of milliseconds. Thus, the signaling overhead due to a change of the number of terminals is negligible compared to the signaling overhead due to a change of the wireless channel, which has to be considered only in dynamic systems.

The most important question regarding signaling is by how much the performance of a dynamic MU-OFDM system is decreased. Obviously, if the signaling overhead is that high that a static approach performs better, then the concept of dynamic assignment algorithms in OFDM systems is not advantageous. However, the relationship between performance gain and caused signaling overhead is more subtle. Various system parameters influence the overhead. The overhead may be represented by convenient coding schemes. Compression algorithms can be applied to reduce the signaling overhead further. Also, the signaling overhead may be included in the optimization process itself, leading to a smaller overhead.

On the other hand, these schemes may lead to significant problems. Applying convenient coding or compression schemes may increase the vulnerability of the signaling overhead to bit errors occurring during the transmission of the signaling overhead. Bit errors in the signaling

part automatically lead to the loss of the complete payload data transmission of the down-link phase. Thus, bit errors in the signaling information in general have a significant impact on the system's performance. Apart from this, bit errors in the signaling part may also lead to problems for the subsequent down-link phases if differential coding schemes are applied for the representation of the signaling data. Retransmissions of the signaling information may then be necessary, again decreasing the performance of the dynamic MU-OFDM system.

A further problem arises from potential compression and optimization schemes regarding their complexity. Compressing the signaling information requires computational capacities both at the access point *and* the terminal. As frame durations can be easily in the range of milliseconds, complex compression schemes are not likely to be applied due to the high cost of computational resources especially at the terminals. Contrarily, including the signaling cost in the optimization approach requires additional computational resources only at the access point, if any. However, the complexity of this approach may be increased significantly, even rendering an optimal solution approach impossible.

Therefore, many questions arise in the context of the signaling overhead for dynamic MU-OFDM systems. So far, these issues have not been considered in research previously. The contribution of this chapter is to first quantify the overhead due to signaling (Section 5.2 and Section 5.3). Prior to that, the modeling framework is introduced in Section 5.1. Then, multiple methods (different representations, compression, optimization) are investigated that can decrease the overhead in specific situations (Section 5.4). Finally, some conclusions are drawn (Section 5.5).

5.1 Modeling Framework for Inband Signaling Schemes

In this section, certain modeling choices are introduced. Afterwards, the basic system model is discussed. All further investigation are based on the choice of a certain dynamic MU-OFDM approach, which is presented also in this section. Note that although all following signaling schemes are discussed in the context of this specific dynamic MU-OFDM approach, they can be applied to any other dynamic MU-OFDM approach as well (for example, approaches for which the objective is to reduce the transmit power). However, the quantitative performance loss as well as the complexity consequences vary for some of the following signaling schemes.

In order to investigate the influence of the signaling overhead, a model for transmitting the signaling information is needed. A principle design question is therefore whether the signaling system is going to be an *inband system* or an *out-of-band system*.

In the case of an out-of-band system, a separate control channel is employed in order to transmit the signaling information. This separate control channel of bandwidth B_{sig} is located outside of the down-link bandwidth B. As a consequence, the performance of the dynamic MU-OFDM system is not directly degraded, as long as the transmission of the signaling information is reliable. Some additional power is required to transmit the data on the separate control channel. However, the required bandwidth depends on the data rate. For any given amount of signaling data per assignment period, a small bandwidth B_{sig} requires an earlier start of the transmission of the signaling information than in the case of a large bandwidth B_{sig}. Of course, prior to the transmission the assignments have to be generated. The generation is based on cur-

rent channel information (samples or estimates of the attenuation). Hence, if the transmission of the signaling information requires more time (due to a small control channel bandwidth), such a system can only be applied to a transmit environment with a rather low coherence time of the wireless channels. Otherwise, the channel attenuations have changed significantly once the terminals received their signaling information completely (compared to the attenuation values the assignments were generated from at the access point).

However, the transmit time is not only influenced by the bandwidth of the control channel but also by the amount of data that has to be transmitted. If the signaling overhead is not represented (encoded) efficiently, a high load has to be conveyed by the control channel, requiring more time. As this data has to be transmitted as fast as possible after the assignments were generated at the access point, the amount of signaling information to be transmitted plays also crucial role for an out-of-band signaling system. Comparing such a system to any static scheme requires to compare the above specified dynamic system with a static one with an effective system bandwidth of $B + B_{\text{sig}}$. An out-of-band signaling scheme is primarily applied to **F**requency **D**ivision **D**uplex (*FDD*) systems, as they divide the bandwidth into a down-link bandwidth and an up-link bandwidth anyway. A plausible extension to such a system would be to introduce a further bandwidth for the signaling information.

On the other hand, an in-band signaling system is more likely to be applied to **T**ime **D**ivision **D**uplex (*TDD*) systems. Here, time frames are divided into up-link and down-link phases. An additional signaling phase can be introduced. This leads to the usage of the same bandwidth B for transmission of signaling information as used for the payload data transmission in down-link and up-link direction. Thus, the performance of a dynamic system is directly degraded by the usage of an inband signaling system (even if all signaling data is received by the terminals correctly), as less time can be used to transmit down-link or up-link data. If the amount of signaling information can be reduced, the time span required for the transmission of the signaling information is reduced and ultimately, more time is available for payload data transmission.

Apart from the issue of considering an out-of-band signaling system or an inband signaling system, a second design issue is the choice of either transmitting the signaling information individually to each terminal or broadcasting it. If the signaling information is transmitted individually, each terminal basically needs its own control channel (either out-of-band or inband). If the signaling information is broadcasted, only one control channel is required.

For example, in a TDD system individual signaling information could be "piggy-backed" at the end of a down-link payload transmission. The signaling information would then be transmitted on the specific sub-carrier set previously assigned to this terminal. As a consequence, the signaling information could be transmitted probably with a quite high modulation type, as the specifically assigned sub-carriers are likely to be in a good state. However, only a few sub-carriers are available for the transmission. In contrast, broadcasting the signaling information would require to use one commonly known modulation type. As all terminals have to decode this data, the choice of this modulation type depends on the terminals with the weakest received SNR (most likely the terminals far away at the edge of the cell). However, if the data is broadcasted, all sub-carriers can be used for this transmission, not only a few sub-carriers. This trade-off exists also for FDD systems in a modified form.

Another difference between broadcasting and individual transmission exists for the conse-

quences of not assigning resources to a certain terminal for one or multiple down-link phases. Consider the case that the signaling information is transmitted individually, for example by "piggy-backing". If no resources are assigned to a certain terminal for the next down-link phase, this terminal basically looses its synchronization to the dynamic MU-OFDM assignments of the system. It has to rejoin the system first during some special phase (like an admission phase) before it regains synchronization to the dynamic assignments. In a broadcasting signaling system, this is not the case. A terminal which did not receive any resources during some down-link phase simply waits for the next broadcast of the signaling information (using the commonly known modulation type) and can rejoin the system automatically. This is a clear advantage of the broadcasting scheme.

However, broadcasting the signaling information also has a clear security draw-back. As the signaling information is announced publicly, every terminal synchronized to the physical OFDM system can potentially decode the payload data sent to each terminal in the cell. If the data is not protected by higher layer means, the down-link data transmission can be eavesdropped by any receiver. If the signaling information is transmitted individually, this is much more difficult.

Throughout this chapter, an inband, broadcasting signaling system is investigated. It is assumed that the dynamic MU-OFDM system is based on a TDD scheme. Time is divided into units (frames) of duration T_{f}. The frames are further divided into a down-link, an up-link and a signaling phase. During each signaling phase a commonly known modulation type is used for the transmission of the data. Although these investigation are based on these assumptions, qualitatively analog results can be expected for other systems (out-of-band signaling, individual transmission).

5.1.1 System Model

Consider a single cell of a wireless system with radius r_{cell}. Within this cell, one access point coordinates all data transmissions. There are J terminals located within the cell and download data continuously from the access point through the down-link. The terminals are uniformly distributed over the area of the cell. Both, up-link and down-link use OFDM as transmission scheme. However, only the down-link data transmission direction is considered further.

Physical Layer

A total bandwidth of B [Hz] at the center frequency f_{c} can be utilized for data transmission. For this bandwidth, a maximum transmit power of P_{max} is allowed to be emitted by any transmitter. The given bandwidth is split into N sub-carriers with a bandwidth of $\frac{B}{N}$ each. In order to guarantee orthogonality between the sub-carriers, the symbol length for all sub-carriers is identical and is related to the bandwidth of each sub-carrier by $\frac{N}{B} = T_{\mathrm{s}}$. Although each sub-carrier employs the same symbol rate, per sub-carrier a different amount of bits might be represented per symbol. This is realized using different modulation types out of a set of M available ones.

One OFDM symbol is generated per symbol time. Prior to the transmission of this symbol, a cyclic extension of length T_{g} is added to the OFDM symbol, the *guard interval*. The length of the complete OFDM symbol is then $T_{\mathrm{s}} + T_{\mathrm{g}}$.

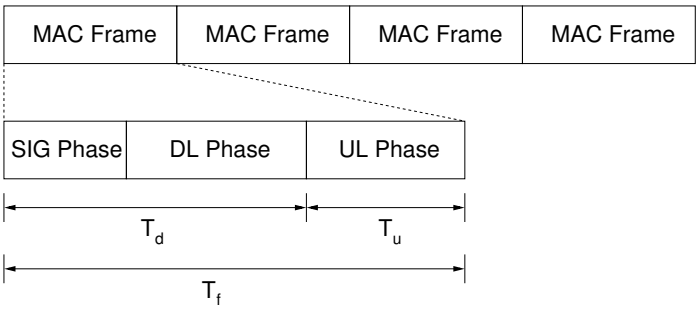


Figure 5.1: Basic medium access control layer layout

Wireless Channel Model

The perceived signal quality per sub-carrier, i.e., their SNR, varies permanently. The instant SNR of sub-carrier n for terminal j at time t is given by $v_{j,n}^{(t)} = \frac{p_n^{(t)} \cdot \left(h_{j,n}^{(t)}\right)^2}{\sigma^2}$, where $p_n^{(t)}$ denotes the transmission power, $h_{j,n}^{(t)}$ denotes the attenuation of sub-carrier n and σ^2 denotes the noise power. The attenuation is primarily responsible for the variation of the perceived SNR; it varies due to path loss, shadowing and fading[1] (see Section 2.1.1). The attenuation of each sub-carrier is assumed to be constant over the time unit T_f. Note that this time unit is considered to be smaller than the coherence time of the wireless channel (see Section 2.1.1). In addition, the attenuation of each sub-carrier is assumed to be reciprocal (the same attenuation is observed at time t on sub-carrier n in the up-link and down-link direction).

Medium Access Control Layer

Assume a system where time is divided into frames of length T_f (Figure 5.1). For each frame a fixed time span is reserved for down-link and up-link transmissions. Thus, a TDD scheme is considered. Dynamic MU-OFDM is applied during the down-link phase. Apart from reserving a certain time span in each frame, the up-link is not considered any further. The time duration of one down-link phase (including the signaling information) is denoted by T_d.

The access point generates new assignments of sub-carrier subsets for each terminal in the cell prior to each down-link phase. Delay effects due to an inadequate processing power at the access point are not considered. Since the terminals are not aware of the new assignments a priori, the access point has to transmit these assignments to the terminals first. Therefore, from the total time reserved for the down-link phase, a certain time span is subtracted for signaling, the *signaling phase*. The signaling information is transmitted through the same bandwidth which is used for payload data transmission, causing the signaling to be *inband*. In addition, the signaling information is assumed to be broadcast during the signaling phase. A broadcast of the signaling

[1] In this study the terminals are assumed to be stationary, thus the time-selective fading is caused by moving objects within the propagation environment. However, for the time-selective fading (and thus for the correlation in time) one could also assume the terminals themselves to be moving

information increases the flexibility of the system such that terminals can easily receive arbitrarily many sub-carriers without loosing synchronization to the sub-carrier assignments (which is not the case if the signaling information is "piggy-backed" to the payload data). The information is transmitted by a fixed modulation type where the number of bits represented per symbol is denoted by b_{sig}, preferably a robust (but "slow") one. Note that correctly decoding the signaling information is crucial to the system's performance.

The generation of the new assignments per frame is based on the knowledge of the sub-carrier states towards each terminal denoted by the matrix $\mathbf{H}^{(t)} = \left(h_{j,n}^{(t)} \mid \forall\, j, n\right)$. It is assumed that this knowledge is available. As a TDD system is assumed where OFDM is also applied in the up-link, the channel knowledge can be acquired by considering the sub-carrier attenuations during the previous up-link phase. If some terminals did not participate in this up-link phase they may be polled to transmit short training sequences, allowing the access point to estimate their sub-carrier attenuations. If the length of a frame is chosen sufficiently small compared to the coherence time of the sub-carriers, this channel knowledge will be quite close to the real values during the following frame, i.e. the estimate error will be rather small.

5.1.2 Dynamic MU-OFDM Approach

As discussed in Chapter 4, there are multiple realistic approaches for a dynamic MU-OFDM system. In order to study the impact of signaling, one particular suboptimal algorithm is chosen. Basically, a system is assumed where the transmit power is fixed per sub-carrier, while the sub-carriers can be assigned dynamically. Thus, the schemes discussed in Section 4.3 qualify. Out of the three discussed schemes in that section, the sub-carrier allocation and assignment approach is chosen due to its easy adaption to variable queue states at the access point (see Section 4.4).

The adaptive modulation scheme works as the following: The transmit power is fixed per sub-carrier, each sub-carrier receives a power of $p_n^{(t)} = \frac{P_{\max}}{N}$. Together with the noise power and the actual attenuation $h_{j,n}^{(t)}$ per terminal j, this fixed transmission power yields a "potential" SNR value, reflecting the SNR as it would be if sub-carrier n were assigned to terminal j during the next down-link phase. Depending on this SNR value, the modulation type with the highest number of bits transmitted per symbol is chosen out of the M possible modulation types, such that a target symbol error probability $p_{\text{sym,max}}$ is not violated. Thus, the matrix consisting of the attenuation values $h_{j,n}^{(t)}$ can be converted into a *bit matrix* with values $b_{j,n}^{(t)}$, denoting the number of bits transmittable per symbol if this pair of sub-carrier and terminal is actually chosen as assignment for the next down-link phase.

The assignment decisions for each down-link phase take the form of binary variables, denoted by $x_{j,n}^{(t)}$. They have the following semantics:

$$x_{j,n}^{(t)} = \begin{cases} 1 & \text{if terminal } j \text{ is assigned sub-carrier } n \text{ for downlink phase } t\,, \\ 0 & \text{otherwise.} \end{cases}$$

All assignment information for downlink phase t is grouped in the *assignment matrix* $\mathbf{X}^{(t)} = \left(x_{j,n}^{(t)} \mid \forall\, j, n\right)$. Obviously, each sub-carrier can be assigned to at most one terminal at a time,

resulting in the constraint

$$\sum_{j} x_{j,n}^{(t)} \leq 1 \qquad \forall n \ . \tag{5.1}$$

The total number of symbols per down-link phase is given by

$$S = \frac{T_{\mathrm{d}}}{T_{\mathrm{s}} + T_{\mathrm{g}}} \ . \tag{5.2}$$

Given a certain assignment realization $\mathbf{X}^{(t)}$, the total number of *bits* that can be transmitted in the entire down-link phase t can be calculated as

$$S \sum_{j,n} b_{j,n}^{(t)} \cdot x_{j,n}^{(t)} \ . \tag{5.3}$$

However, this has to be seen as a *gross* value as it does not contain the performance loss due to signaling.

The objective of the chosen suboptimal assignment approach is to maximize the transmitted (payload) bits per down-link phase. Prior to the assignment of the sub-carriers, each terminal is allocated a certain number of sub-carriers l_j which it may receive at most by the assignment algorithm [102]. This allocation can easily be adjusted to the current amount of queued data of each terminal at the access point. As a linear constraint, this has the form

$$\sum_{n} x_{j,n}^{(t)} \leq l_j \qquad \forall \ j \ . \tag{5.4}$$

Using the expressions introduced so far, the considered dynamic MU-OFDM approach can be formulated as MIPoptimization problem operating on the input of the bit matrix. The problem statement is thus given by:

$$\begin{aligned} \max \quad & S \cdot \sum_{j,n} b_{j,n}^{(t)} \cdot x_{j,n}^{(t)} \\ \text{s. t.} \quad & \sum_{j} x_{j,n}^{(t)} \leq 1 \qquad \forall \ n \\ & \sum_{n} x_{j,n}^{(t)} \leq l_j \qquad \forall \ j \\ & x_{j,n}^{(t)} \in \{0,1\} \quad . \end{aligned} \tag{PLAIN}$$

Problem (PLAIN) maps to a graph-theoretical problem [76] known as the *bipartite weighted matching problem* [79], where the algorithm generating the solution has a complexity of $O\left(N^3\right)$. In practice, the solution of this problem can be generated within milliseconds on standard computers, assuming reasonable system parameters (for example $J = 16$ terminals, $N = 48$ sub-carriers).

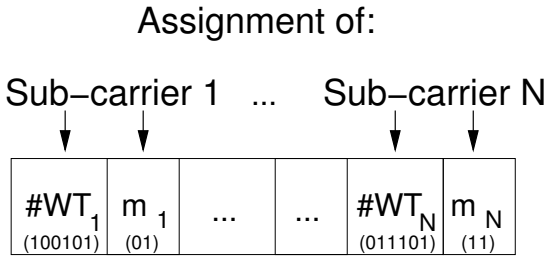


Figure 5.2: Fixed size signaling field model: Representation of assignment information for one down-link phase.

5.2 The Fixed Size Signaling Field Approach

As the dynamic algorithm generates the optimal assignments, the assignment information still has to be conveyed to the terminals prior to the down-link data transmission. In order to do so, an inband signaling system is employed where the signaling information is broadcasted, as described previously. The assignment information itself is contained in the matrix $\mathbf{X}^{(t)}$. In addition to this, also the applied modulation type has to be signaled for each assignment. Therefore the basic information unit of the signaling information consists of the triple:

$$\langle \text{sub-carrier identification}, \text{terminal address}, \text{modulation identification} \rangle \quad .$$

These triplets can be transmitted to the terminals in multiple different ways. The first, straightforward approach investigated here is called the *fixed size signaling field*. All assignments are broadcasted, regardless of whether an assignment changed from the previous to the current down-link phase or not. This scheme is initially used to quantify the overhead due to signaling.

5.2.1 Mathematical Model

The following procedure is proposed for this approach, described in Figure 5.2.1: *All* assignments are transmitted in the bit stream one after the other. Since all N assignments are transmitted in each frame, the position of the tuple

$$\langle \text{terminal address}, \text{modulation identification} \rangle$$

in the sequence already indicates the sub-carrier identification (if this tuple is the fifth one transmitted, then it relates to sub-carrier five). Therefore, per assignment a signaling cost of

$$\lceil \log_2(J) \rceil + \lceil \log_2(M) \rceil$$

bits is required. Transmitting all these assignments results in a total cost (in bits) of

$$N \cdot (\lceil \log_2(J) \rceil + \lceil \log_2(M) \rceil) \quad ,$$

which requires a total of

$$\varsigma = \left\lceil \frac{(\lceil \log_2(J) \rceil + \lceil \log_2(M) \rceil)}{b_{\text{sig}}} \right\rceil \tag{5.5}$$

symbols per down-link phase. Therefore, for payload transmission $S - \varsigma$ symbols remain. Inserting this expression (instead of only S) in Equation (5.3) yields the effective throughput of the system, referred to as *net throughput*.

Since the number of lost symbols due to signaling is constant for a certain system instance (fixed parameter set), maximizing the net throughput is equivalent to maximizing the gross throughput in the current model, given by the solution of problem (PLAIN). However, the number of symbols available for payload transmission is adjusted:

$$\begin{aligned} \max \quad & (S - \varsigma) \cdot \sum_{j,n} b_{j,n}^{(t)} \cdot x_{j,n}^{(t)} \\ \text{s. t.} \quad & \sum_{j} x_{j,n}^{(t)} \leq 1 && \forall\, n \\ & \sum_{n} x_{j,n}^{(t)} \leq l_j && \forall\, j \ . \end{aligned} \tag{5.6}$$

5.2.2 Protocol Aspects

Prior to the transmission of the signaling information, a usual preamble for synchronization and channel estimation has to be inserted. The signaling information itself has to be error protected, as it is quite sensitive information. This can be done by convolutional encoding or turbo coding. Forward error correction influences mainly b_{sig}, the number of data bits per symbol per sub-carrier (equaling 0.5 in the case of BPSK modulation in combination with a rate 12 convolutional encoder). Thus, coding is incorporated in the mathematical model already. In addition to forward error correction, a error detection scheme is also added to the signaling information. Powerfull cyclic redundancy codes (CRC) are known for this task, for example a CRC-16 code [103].

Assume that a bit error occurs in the signaling part and cannot be corrected by the forward error correction scheme. Two possible situations might occur. If the bit error can be at least detected by the error detection scheme, the terminal simply discards the payload data of this down-link phase. It then indicates the loss of the link layer data unit to the access point during the next up-link phase and receives during the next down-link phase the link layer unit (if the signaling information is received correctly). Alternatively, the terminal might try to decode the payload information, although some bit error is present in the signaling information. In fact, as a CRC relates to all signaling information, the indication of such one or more bit errors does not necessarily mean that the specific signaling information of this terminal is corrupted. However, if the data is decoded although the CRC indicates partially wrong information, the next higher instance to detect that wrong information has been decoded is the payload error protection (or detection) scheme of the link layer packet. This may span over multiple down-link phases if a large packet is fragmented. Thus, a complete retransmission of this packet might be necessary, decreasing the link performance much more than simply retransmitting one down-link frame.

However, it may also occur that the signaling information is corrupted and the error detection scheme does not indicate this. Although this is not very likely if appropriate error detection and correction schemes are applied, it still might occur. In this case, an error in the signaling information still does not automatically lead to a loss of the complete link layer packet. Only if the signaling information regarding this specific terminal is corrupted, the complete link layer packet is probably lost as either a wrong sub-carrier is used to detect data from or a wrong modulation type is assumed for the length of one down-link phase.

Obviously, the error sensitivity of the dynamic scheme depends highly on the coding and modulation combination used for the conveying the signaling information. If this combination is chosen such that a low bit error probability results even for terminals farer away from the access point, the impact of errors in the signaling information is negligible.

5.2.3 Performance Evaluation

The central question regarding signaling is by how much the performance of a dynamic MU-OFDM system is decreased. Given the above discussed model of the fixed size signaling field approach this performance loss can be studied.

The impact of the fixed size signaling field model on the throughput of a dynamic MU-OFDM system depends obviously on a couple of parameters. The overhead is related to the number of terminals in the cell J, the number of sub-carriers in the system N and the number of modulation types M. These three parameters directly influence the number of bits in the signaling bit-stream. In addition, the choice of broadcasting modulation type b_{sig} has a strong direct impact on the amount of signaling data to be transmitted per down-link phase. Note that once these parameters are given, the signaling overhead per down-link phase can be calculated precisely.

However, the loss due to signaling is also driven by other parameters of the system. For example, the length of a frame T_{f} determines how often new signaling information has to be conveyed. Thus, regarding the throughput on a larger scale (bit/s, for example), the frame duration has also a significant influence on the performance loss due to signaling. The frame length cannot be considered without taking the coherence time of the channel into account, though. Hence, the maximum speed of the propagation environment $v_{\max}$ as well as the center frequency of the system f_{c} are parameters which influence the choice of the frame length.

The loss due to signaling also depends on the absolute number of symbols available during the down-link phase S. The more symbols are available, the less the impact of signaling will be. The number of symbols per down-link phase is ruled by the bandwidth per sub-carrier (depending on the total number of sub-carriers N and the overall system bandwidth B $[\mathrm{Hz}]$) as well as the duration of the guard interval T_{g}. The choice of the guard interval depends on the expected delay spread of the propagation environment $\Delta\sigma$.

In addition, certain trade-offs can be identified which complicate the prediction of the impact of signaling even further. For example, increasing the number of sub-carriers the given bandwidth is split into increases the throughput of any OFDM system utilizing a guard duration, as more time is used for payload transmission than for transmitting the guard duration (the more sub-carriers there are, the longer the OFDM symbol gets). However, the signaling load increases as well, indicating a trade-off between these two effects. Another such effect could be expected

for the increase of the modulation types or for an increase of the terminals in the system. Both these effects increase the gross throughput or spectral efficiency, but they also lead to a higher signaling load.

In order to distinguish between parameters with a high impact and a low impact, first a sensitivity analysis has been performed following the method of the 2^k factorial design [104]. Then, the most relevant parameters are further discussed. As metric for the sensitivity analysis the average gross throughput per terminal and the average percentage of symbols per down-link phase required to transmit the signaling information (ς/S) were considered. Note that the net throughput itself could not be used directly for the sensitivity analysis, since it has maximum points within the range of the considered parameter instances, which do not allow a correct statement regarding the impact of each parameter on this metric.

For all following performance results of the fixed size signaling field approach the dynamic approach without any signaling cost (problem (PLAIN)) and a static scheme are considered as comparison schemes. The static scheme considered assigns fixed blocks of sub-carriers to each terminal. These assignments are never changed. However, per sub-carrier the same adaptive modulation scheme is applied as described in Section 5.1.2. Thus, a simple signaling scheme would be required for this approach, too. It is not taken into account for all performance results throughout this chapter, though. Hence, the static results can be viewed as rather optimistic.

Methodology

From a system point of view, metrics like the average data rate per terminal are of interest. However, the solution for an instance of one of the optimization problems discussed so far is a certain number of bits that can be transmitted per down-link phase to each terminal. In order to obtain system level results, the following procedure was conducted. Initially, channel trace files of the attenuations of each sub-carrier were generated from a computer program. Each attenuation sample consisted of path loss, shadowing and fading components. One sample was generated for every down-link phase and sub-carrier per terminal. Once the trace file of the attenuations was generated, the attenuation matrix was transformed into the bit matrix $\mathbf{B}^{(t)}$ for each down-link phase.

Specialized algorithms from graph theory are known for problem (5.6), for a survey see [46]. It is therefore not necessary to employ integer programming algorithms. However, state-of-the-art matching algorithms are complex to implement and might therefore not be the method of choice in a real system. On the other hand, due to the special properties of the so-called b-matching polytope [53], the problems can be solved with linear programming algorithms and the resulting solutions will always be integral. Thus, the generated instances of problem (5.6) were solved using linear programming software.

Based on the bit matrices, the problem instances for the corresponding optimization problems were converted into a linear integer program in standard form using the modeling language ZIMPL [97]. All resulting integer programs were solved using CPLEX 9.0, a state-of-the-art code for linear and integer programming. The software uses a Branch&Bound&Cut algorithm with an implementation of the simplex algorithm for solving linear programming subproblems; outlines of how these algorithms work can be found in [46]. For the instances of problems (5.6) only linear programming is actually used as the Branch&Bound Tree collapses to the root node

(see [53]). The resulting assignments were collected by a script. The script kept record of each frame throughput while running through the complete trace file. Typically, each trace file represented a few seconds (equaling a few thousands of down-link phases). Finally, from the total amount of bits transmitted to each terminal the throughput per terminal was obtained as well as the average signaling percentage.

Simulation Scenario

The following basic simulation scenario was chosen: The cell radius is $r_{\mathrm{cell}} = 100\ \mathrm{m}$, the center frequency is $f_{\mathrm{c}} = 5.2\ \mathrm{GHz}$. The maximum transmit power equals $P_{\max} = 10\ \mathrm{mW}$. The guard interval length is set to $T_{\mathrm{g}} = 0.8\ \mu\mathrm{s}$ (in correspondence to IEEE 802.11a [32]).

The sub-carrier attenuations $h_{j,n}^{(t)}$ are generated based on the three components path loss, shadowing and fading. For the path loss, a standard model $h_{\mathrm{pl}}^{(t)} = K \cdot \frac{1}{d^{\alpha}}$ is assumed [4]. As parameters, we use $K = -46.7\ \mathrm{dB}$ and $\alpha = 2.4$, corresponding to a large open space propagation environment. For the shadowing $h_{\mathrm{sh}}^{(t)}$, independent stochastic samples are assumed from a log-normal distribution, characterized by a zero mean and a variance of $\sigma_{\mathrm{sh}}^2 = 5.8\ \mathrm{dB}$. The samples are regenerated every second. While the path loss and shadowing affects all sub-carriers of a terminal alike, each sub-carrier experiences its own fading component. Each sample $h_{\mathrm{fad}}^{(t)}$ of the fading process is assumed to be Rayleigh-distributed. The frequency and time correlation of $h_{\mathrm{fad}}^{(t)}$ are characterized by a Jakes-like power spectrum and an exponential power delay profile. The Jakes-like power spectrum is parameterized by the maximum speed within the propagation environment, set to $v_{\max} = 1\ \mathrm{m/s}$ and the center frequency f_{c}. The exponential power delay profile is characterized by the delay spread which is set to $\Delta\sigma = 0.15\ \mu\mathrm{s}$. The noise power σ^2 is computed at an average temperature of $20^{\circ}\ \mathrm{C}$ over the bandwidth of a sub-carrier.

For the adaptive modulation scheme, a target symbol error probability of $p_{\mathrm{sym,max}} = 10^{-2}$ is chosen for the payload transmission. For the signaling phase, BPSK us used in combination with a rate $1/2$ convolutional coder (assuming soft decision at the receivers), resulting in $b_{\mathrm{sig}} = 0.5\ \mathrm{bit/symbol}$. Using the above mentioned path loss parameterization and considering a 10 dB fading margin, a SNR of at least 4 dB is thus reached at the cell border. This yields a bit error probability of 10^{-4} [2] in the signaling part; given this small value, bit errors in the signaling part are not considered any further.

Sensitivity Analysis

Initially, five parameters were identified for the sensitivity analysis. For each of these parameters two instances were chosen, as given in Table 5.1. Thus, $2^5 = 32$ different settings were evaluated. For each setting the average gross throughput per terminal and the average percentage of symbols required for signaling per frame were generated. These metrics were computed for the optimization approaches (5.6), (PLAIN) as well as for the static comparison scheme.

First, the average results for the three different approaches regarding each metric are presented, also showing the results achieved for the net throughput (for the net throughput an average value can be obtained by the method of performance analysis, however, judging on the influence of different factors on the variation of this average is not possible due to the local optima

Parameter	Instance 1	Instance 2
Terminal number J	4	16
Sub-carrier number N	64	512
Number of modulations M	2 (BPSK and QPSK)	5 (BPSK, QPSK, 16-,64, and 256-QAM)
Frame length T_{f}	1 ms	10 ms
System bandwidth B	3 MHz	50 MHz

Table 5.1: Parameter instances for the sensitivity analysis regarding the fixed signaling field approach.

within the range of the parameters). This is given in Table 5.2. Considering the net throughput per terminal, about 0.2 MBit/s (about 10%) of throughput is lost due to the fixed signaling field approach. Compared to the static scheme, the dynamic approach based on the fixed signaling field approach still provides an average net throughput which is higher by 0.4 MBit/s, which equals about 25% more throughput per terminal than the static approach achieves. Note that the net throughput reflects the average over all configurations, including configurations where the dynamic approach cannot outperform the static one very much. For example, all configurations where the number of modulations is low, $M = 2$, lead to very small performance differences between the static and dynamic approach. Also, if the number of terminals in the cell is low, $J = 4$, the performance difference is low due to a low multi-user diversity of the cell (see Section 4.3). In terms of the signaling percentage, on average 25% of each down-link phase is lost due to the signaling overhead.

	Gross throughput	Signaling percentage	Net throughput
Static assignment	1.638 MBit/s	0	1.638 MBit/s
Problem (PLAIN)	2.283 MBit/s	0	2.283 MBit/s
Problem (5.6)	2.283 MBit/s	0.267	2.059 MBit/s

Table 5.2: Average metric results stemming from the sensitivity analysis for the three different considered approaches (fixed, dynamic with no sig. cost, dynamic with the fixed signaling field model) to the dynamic MU-OFDM system.

Apart from an evaluation of the average values, the sensitivity analysis following the method of the 2^k factorial design allows also to determine factors, i.e. parameters, which have a strong (weak) impact on the variation of the considered metric. In addition to single factors (for example the number of terminals) also factor combinations might be significant (for example the number of sub-carriers and the number of terminals).

Regarding the signaling cost (percentage of symbols required per frame for signaling) Table 5.3 shows the impact of all studied factors and selected factor combinations on the variation of the signaling percentage for the fixed signaling field approach.

For the fixed signaling field approach all the factors (B , T_{f} , N , M , and J) have a certain influence on the amount of signaling overhead per down-link phase. Note that out of these the system bandwidth B , the frame length T_{f} and the number of sub-carriers N have the strongest influence on the overhead. Apart from these three, the factor combinations of them also con-

Parameter	Approach (5.6)
System bandwidth B	$35.9\ \%$
Frame length T_{f}	$25.2\ \%$
Number of sub-carriers N	$16.2\ \%$
Number of modulations M	$0.7\ \%$
Number of terminals J	$0.7\ \%$
Stochastic variation	$0\ \%$
B and T_{f}	$10.2\ \%$
B and N	$6.2\ \%$
N and T_{f}	$2.5\ \%$

Table 5.3: Variation percentage of all five factors and selected factor combinations regarding the signaling overhead for the fixed signaling field model.

tribute quite strong to the variation of the overhead results. The influence due to stochastic variations is zero, which is also quite reasonable when considering the design of the fixed signaling field approach. Interestingly, there are combinations where the total signaling overhead consumes the *complete* down-link phase resulting in a zero net throughput for these cases. This happens whenever the bandwidth is low ($B = 3$ MHz), the number of sub-carriers is high ($N = 512$) and the frame length is short ($T_{\mathrm{f}} = 1$ ms) (in fact, to transmit the complete signaling information would require in these cases a lot more than one down-link phase, however the technical upper limit of the proposed system leads to a resulting signaling ratio of 1). In the opposite cases (high system bandwidth $B = 50$ MHz, low number of sub-carriers $N = 64$ and a long frame length $T_{\mathrm{f}} = 10$ ms) the average signaling percentage is quite low at about 0.04. In these cases the influence of the number of modulation types M and the number of terminals in the cell J is higher than indicated by Table 5.3.

In the following the variation of four different parameters is studied further: Varying the system bandwidth, varying the frame length, varying the number of sub-carriers and varying the number of terminals.

Parameter Evaluation – Varying the System Bandwidth

Initially, consider the variation of the available system bandwidth. It is considered for values between $B = 1$ MHz and 50 MHz. In each case, the bandwidth is split into $N = 256$ sub-carriers. The transmit power is increased accordingly, the ratio between total transmit power and system bandwidth is kept constant at 0.6 mW/MHz. The number of terminals is fixed at $J = 8$, four different modulation types are available (BPSK, QPSK, 16-QAM and 64-QAM). A frame length of $T_{\mathrm{f}} = 2$ ms is assumed. Note that all these values are within the specification of a regular OFDM system regarding the coherence time and coherence bandwidth.

As seen from the sensitivity analysis, varying the system bandwidth has a strong impact on the variation of the signaling percentage. As the bandwidth increases, the throughput of any OFDM system will increase. As the number of sub-carriers is fixed, increasing the bandwidth leads to a higher symbol rate per sub-carrier. An increasing number of OFDM symbols can

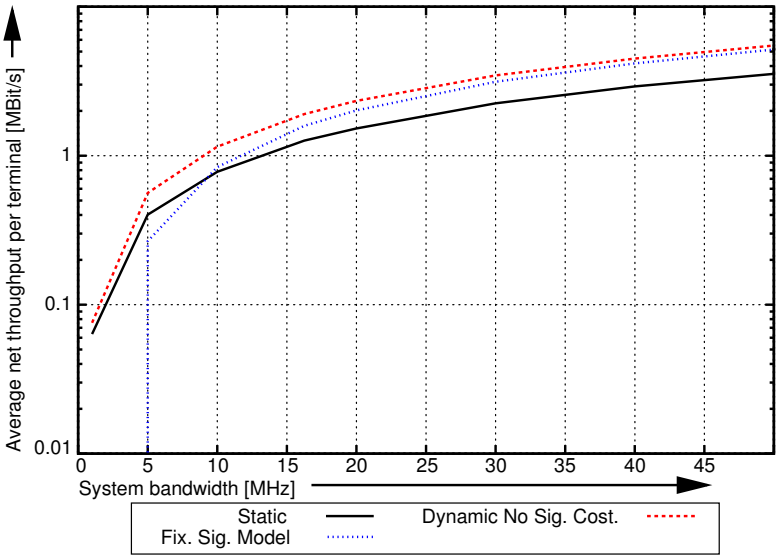
be transmitted per frame. However, as the length of a symbol becomes smaller, the length of the guard period becomes more dominant. Finally this leads to an opposite result as if the bandwidth was kept constant and the sub-carrier number was increased: As the symbol length decreases, more and more time is spent mitigating the effect of ISI by transmitting the guard. Therefore, despite the fact that the bandwidth is increased, the ratio between guard period and symbol length increases, thus per down-link phase more and more time is spent on transmitting the cyclic extension. Therefore the throughput gain by increasing the bandwidth by a certain amount is limited to some extent by the guard period. One way to deal with this performance limiting effect is to increase the number of sub-carriers, which is studied in below.

In general, the more OFDM symbols can be transmitted per frame length the lower is the impact of the signaling overhead, i.e. the signaling percentage decreases. Therefore, for a higher bandwidth, the net throughput achieved by a dynamic MU-OFDM approach with signaling should get quite close to the achieved gross throughput.

In Figure 5.3 the average net throughput per terminal and the average spectral efficiency per cell is given for an increasing bandwidth. Note that the spectral efficiency is obtained by dividing the overall net throughput of the cell (average net throughput per terminal multiplied by the number of terminals in the cell) by the overall system bandwidth. As the duplex mode assumed here is TDD (splitting each frame equally into down-link and up-link phase), any net throughput value as well as any spectral efficiency value is reduced by the factor two. As already discussed, the net throughput does increase for all considered approaches while increasing the system bandwidth. For all bandwidth values the dynamic approach potentially outperforms the static approach, where the gap between the static and dynamic approach appears to be constant from the logarithmic plot. For a very small bandwidth (1 MHz) the fixed size signaling field approach has a net throughput of 0 kBit/s due to the fact that the signaling information already consumes the complete down-link phase. At about 10 MHz the fixed size signaling field approach starts to outperform the static approach and as the overall system bandwidth increases further, the gap between the dynamic scheme with and without the signaling overhead decreases significantly.

Considering the spectral efficiency, observe at first that all graphs increase at first up to a maximum bandwidth of 30 MHz, beyond that bandwidth the spectral efficiency actually *decreases*. This behavior is due to the trade-off between two effects. As the bandwidth increases while the number of terminals is fixed, the frequency diversity increases. This leads to an increasing spectral efficiency, also observed for a moderately increasing delay spread while the bandwidth and sub-carrier number is fixed. However, as the bandwidth increase, the ratio between payload symbol duration and guard duration decreases (as discussed above). Thus, at a certain bandwidth this performance decreasing effect gets dominant compared to the performance increasing effect of the increasing frequency diversity. As a result, the spectral efficiency increases first and decreases thereafter.

Below a bandwidth of 10 MHz, there is no point in applying the dynamic MU-OFDM scheme with the fixed size signaling field model. The static scheme achieves a better spectral efficiency. Beyond that bandwidth the dynamic scheme including the signaling overhead achieves a constantly better spectral efficiency up to a bandwidth of 30 MHz. From that point on the efficiency is slightly decreasing, but not as strong as with the dynamic scheme without the signaling



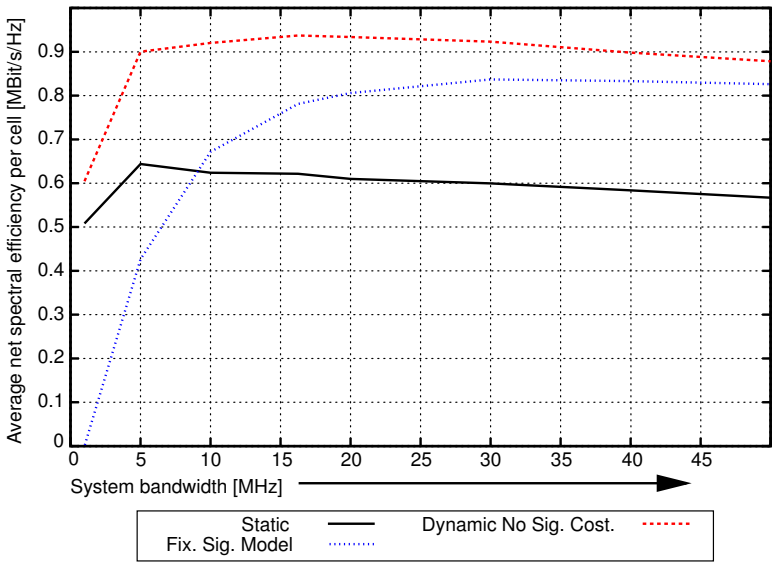


Figure 5.3: Average net throughput per terminal (upper graph - logarithmic scale) and average spectral efficiency per cell (lower graph) for an increasing system bandwidth (with a confidence level 0.99).

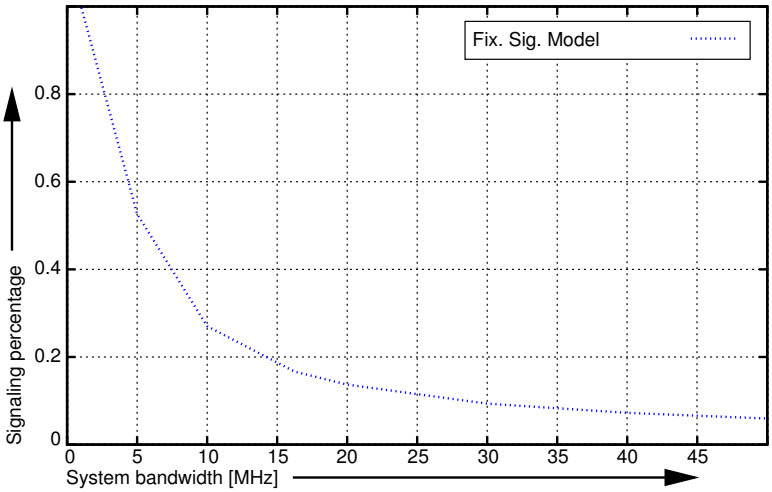


Figure 5.4: Average signaling percentage per frame for an increasing system bandwidth of the cell (with a confidence level 0.99).

overhead. This qualitative behavior of the performance curve is easily explained considering the plot of the average signaling percentage per down-link phase for an increasing bandwidth (Figure 5.4). The signaling percentage for the fixed size signaling field approach is at 100% for a bandwidth of 1 MHz. As the bandwidth increases, the signaling percentage drops down to 4%. This constantly decreasing signaling percentage leads to the almost constant spectral efficiency for bandwidths larger than 30 MHz.

Parameter Evaluation – Varying the Frame Length

Next, consider varying the length of a frame T_{f}. It is varied between 1 ms and 10 ms, always splitting the frame equally into down-link and up-link phase. These variations are performed for two different numbers of sub-carriers N: 64 and 512. The number of terminals is fixed at $J = 8$ while $M = 4$ (BPSK, QPSK, 16-QAM and 64-QAM) modulation types are available. The system bandwidth is chosen to equal $B = 16.25$ MHz while the transmit power is set to $P_{\mathrm{max}} = 10$ mW.

As the length of a frame increases, the number of symbols per down-link phase increases, too. Thus, a dynamic algorithm reassigns sub-carriers less often. This is quite fortunate since for a system with a frame length of 10 ms sub-carriers are reassigned ten times less often than in the case of a frame length of 1 ms. Apart from considering the signaling overhead, a longer frame length is quite attractive from a systems point of view. If the frame length is longer, the access point has more time to generate new assignments, thus the real-time constraint is less restrictive. The downside though is the accuracy of the channel knowledge, which will degrade in general the longer the frame length is. A lower accuracy of the channel knowledge might lead to a higher bit error probability during the payload transmission, depending on the coding and interleaving scheme used. On the other side, the usage of a longer frame length might enable

the usage of sophisticated channel estimation techniques, which would improve the accuracy of the channel knowledge again. Given the coherence time of the channel, the frame duration will be chosen as long as possible.

In Figure 5.5 the net throughput is shown for the three considered schemes in a setting with $N = 64$ sub-carriers (upper graph) and $N = 512$ sub-carriers (lower graph). In both cases the effect mentioned above can be observed. While for the static scheme as well as for the dynamic scheme without signaling cost the increasing frame length has no effect at all, for the dynamic scheme with signaling cost a longer frame length leads to a higher net throughput. This effect is much more significant in the case with a high number of sub-carriers than in the case with a low number of sub-carriers. In both cases, the potential benefit from using a dynamic approach is about 550 kBit/s on average per terminal (about 50% more throughput per terminal than in the static scheme). If the length of the frame is low, the fixed size signaling field approach yields a significantly lower net throughput than the for longer frame duration. However, the loss depends on the number of sub-carriers (apart from other parameters). If the number is high, short frame durations lead to a performance well below the static throughput, rendering the fixed size signaling field approach impossible.

As the frame length increases, the fixed size signaling field approach performs better. The highest gains though are achieved while the frame length increases up to a duration of 4 ms. Beyond that frame length, the increase is only moderate. Thus, although the net throughput increases constantly for the dynamic scheme including signaling cost, the highest gains are achieved for rather small frame lengths. In general, for a frame length longer than 5 ms, the additional net throughput gain is not that large any more. Therefore, cutting the frame length from 10 ms down to 5 ms for an increasing accuracy of the channel knowledge would not lead to a dramatic loss in terms of the net throughput. In contrast, decreasing the frame length from 4 ms down to 2 ms has much more consequences for the achieved performance.

The qualitative behavior of the dynamic scheme with the fixed signaling field model is easily explained considering the signaling percentage for both settings given in Figure 5.6. For both scenarios, the signaling percentage is exponentially decreasing, explaining the net throughput behavior observed above. Note that the quantitative percentages differ for the two graphs in Figure 5.6.

Parameter Evaluation – Varying the Number of Sub-carriers

The number of sub-carriers N is varied between 64 and 512. The available system bandwidth is fixed at $B = 16.25$ MHz while $J = 8$ terminals are present in the cell and $M = 4$ (BPSK, QPSK, 16-QAM and 64-QAM) modulation types are available. The transmit power is fixed at $P_{\text{max}} = 10$ mW. The frame length is set to $T_{\text{f}} = 2$ ms, equally divided into up- and down-link phase.

Increasing the number of sub-carriers while keeping the total transmit power and the total system bandwidth fixed, leads to an increase of the throughput for any OFDM system utilizing a fixed guard interval. The reason for this is quite simple: The more sub-carriers there are, the lower the bandwidth of each sub-carrier is, hence the symbol time per sub-carrier increases (doubling the amount of sub-carriers leads to a doubling of the symbol times). However, in order to reduce the impact of ISI, prior to each symbol the cyclic extension (guard interval) is

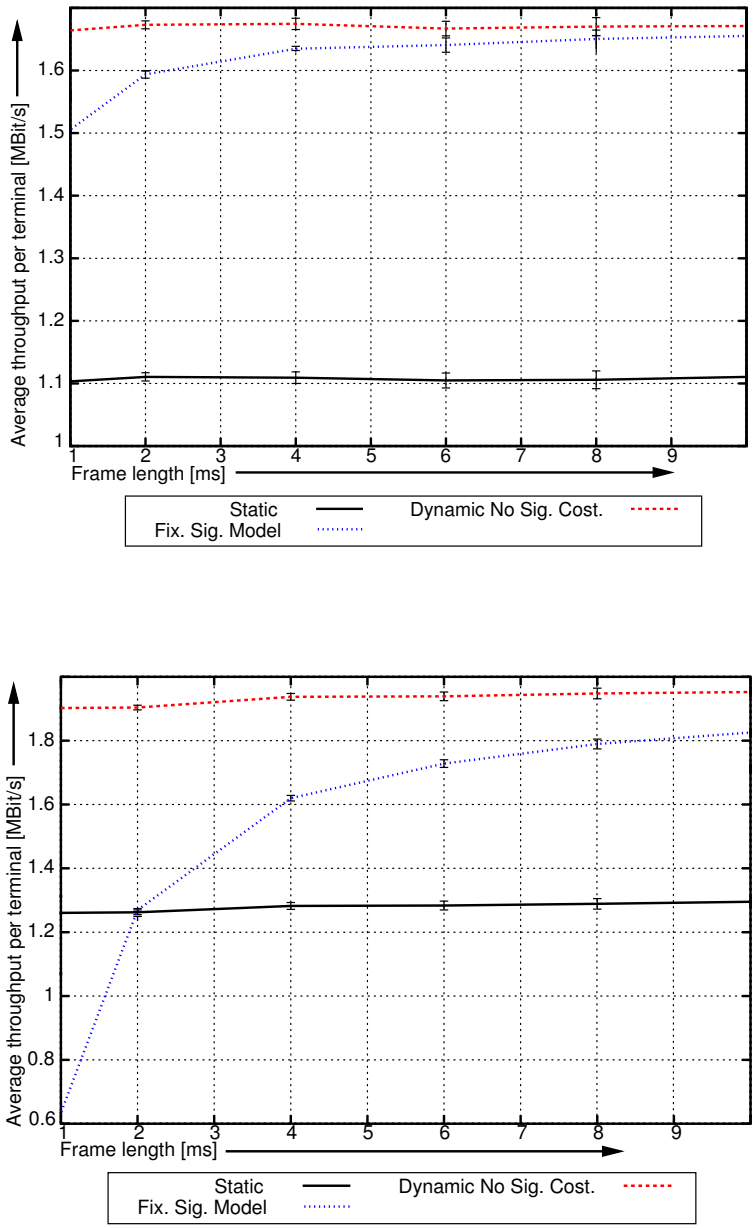


Figure 5.5: Average net throughput per terminal for increasing the frame length with $N = 64$ (upper graph) and $N = 512$ (lower graph) (both with a confidence level of 0.99).

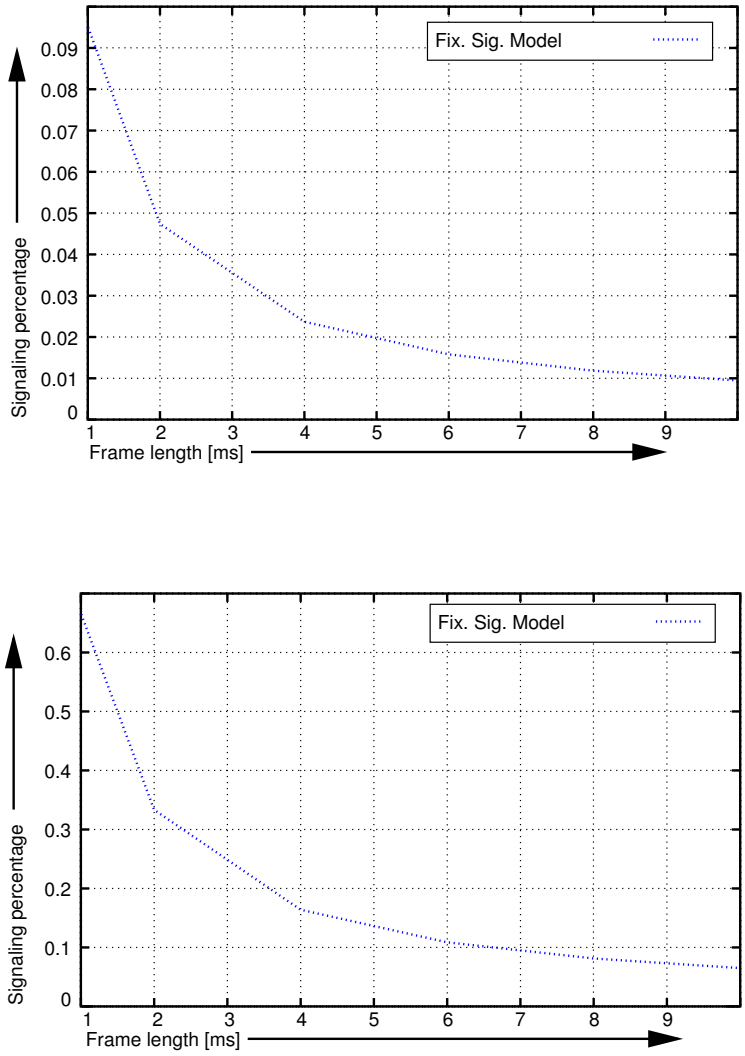


Figure 5.6: Average signaling percentage per frame for increasing the frame length with $N = 64$ (upper gaph) and $N = 512$ (lower graph) (both with a confidence level of 0.99).

sent, which is discarded at the receiver. The length of this guard period is not influenced by the number of sub-carriers, it depends directly on the delay spread of the propagation environment (which stays constant). If the symbol duration increases, the percentage of time each sub-carrier is utilized for data transmission increases compared to the time used in order to mitigate ISI, hence more time is spent on data transmission, leading to a higher system throughput.

As the number of sub-carriers increases though, the signaling cost increases also, as more sub-carriers have to be addressed. This calls for a trade-off when considering the throughput enhancing effect of increasing the number of sub-carriers.

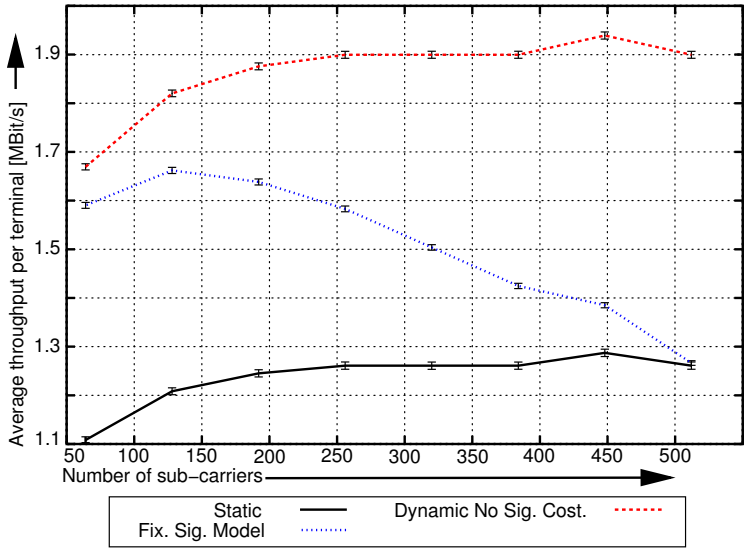
In Figure 5.7 (upper graph) the net throughput is given for the three discussed variants (dynamic MU-OFDM approach without signaling cost, fixed size signaling field model, and the static approach) while varying the number of sub-carriers. As described above, the throughput of the static approach and of the dynamic MU-OFDM approach without signaling increase as the number of sub-carriers increases. The dynamic approach gains slightly more from the increase of sub-carriers, the net throughput increases by $250\ \mathrm{kBit/s}$ on average per terminal while the static gains by $200\ \mathrm{kBit/s}$. The advantage of the dynamic approach without signaling compared to the static approach is about 50%.

However, when considering also the signaling cost the behavior of the system changes qualitatively, as also observed in the two prior sections. For the dynamic scheme with signaling cost, the net throughput first increases up to a maximum point and decrease thereafter. The reason for this is that the signaling cost increases as well as the gross throughput increases for an increasing number of sub-carriers. For low to moderate numbers of sub-carriers the throughput gain outweighs the increase in signaling cost. However, as the number of sub-carriers increases further, the signaling cost becomes the dominant effect, leading to a decreasing net throughput. While the throughput gain of the gross throughput has a logarithmic behavior (same sub-carrier increments achieve smaller and smaller throughput gains) the signaling cost increases linearly, as observed in Figure 5.7 (lower graph). At the maximum net throughput point (at about $N = 150$), the performance advantage on average per terminal is still about $450\ \mathrm{kBit/s}$.However, at the highest number of sub-carriers ($N = 512$) the fixed size signaling field approach achieves the same net throughput as the static approach.

Parameter Evaluation – Varying the Number of Terminals

Finally, the number of terminals in the cell is varied between $J = 1$ and 16. Two different scenarios are considered where the number of sub-carriers equals 64 and 512 while the given bandwidth is set to $B = 16.25\ \mathrm{MHz}$. The frame length is fixed at $T_\mathrm{f} = 2\ \mathrm{ms}$ The transmit power is fixed at $P_\mathrm{max} = 10\ \mathrm{mW}$. Four different modulation types are available: BPSK, QPSK, 16-QAM and 64-QAM.

As seen from the sensitivity analysis, varying the number of terminals is not a primary factor influencing the signaling percentage as such (see Table 5.3). However, increasing the number of terminals leads also to an increase of the spectral efficiency of a dynamic MU-OFDM system, as shown in Chapter 4. Thus, another trade-off arises due to the moderately increasing signaling cost while also the gross spectral efficiency increases as more and more terminals are present in the cell.



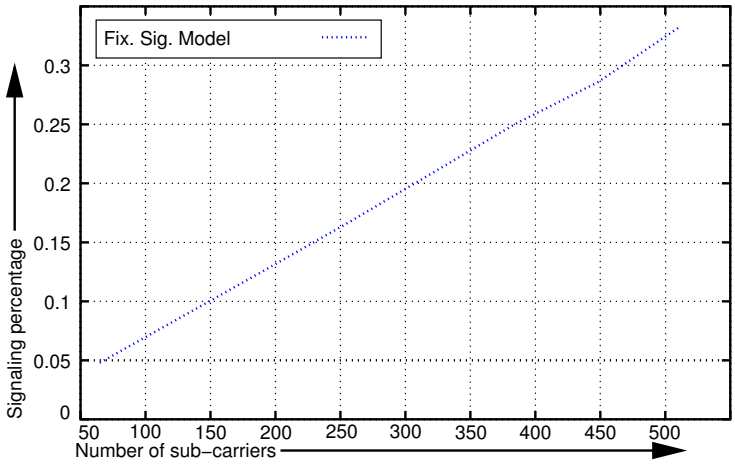


Figure 5.7: Average net throughput per terminal for an increasing number of sub-carriers (upper graph - confidence level 0.99) and average signaling percentage per frame (lower graph).

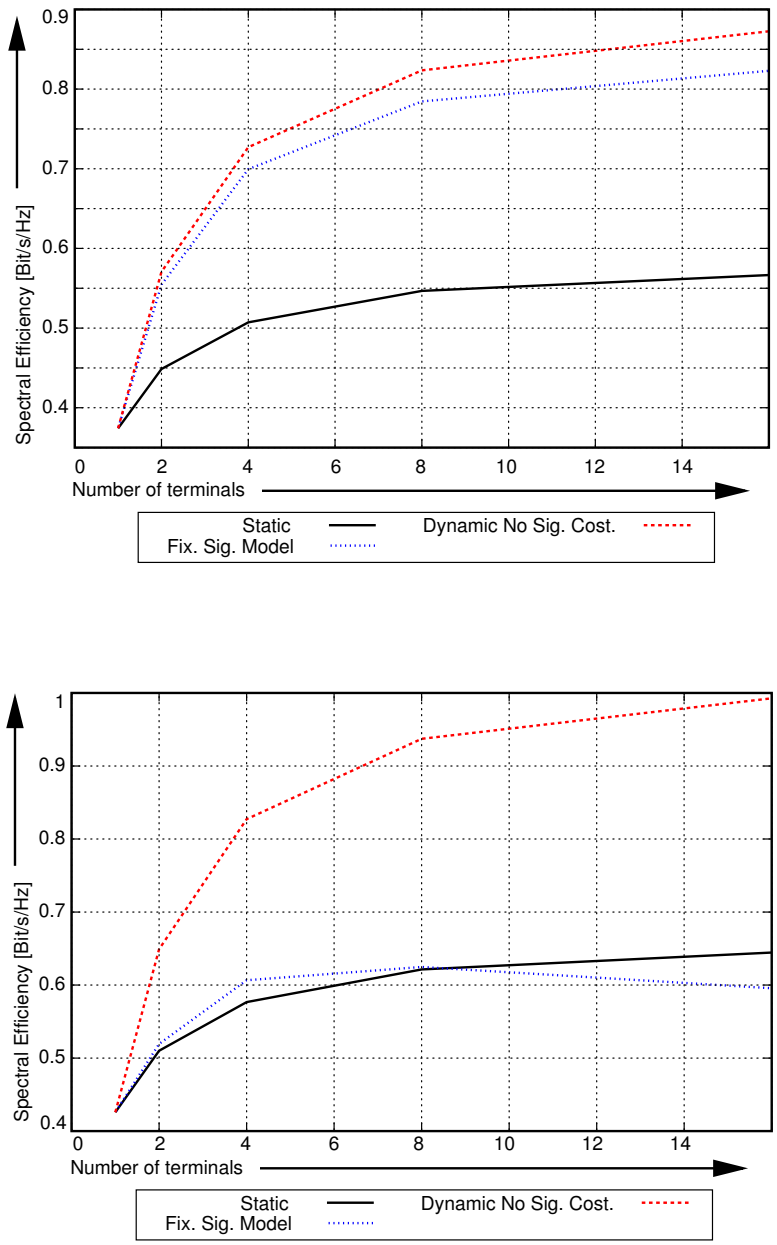


Figure 5.8: Average spectral efficiency for an increasing number of terminals in the cell with a fixed bandwidth divided into $N = 64$ sub-carriers (upper graph) and $N = 512$ sub-carriers (lower graph).

The results for the spectral efficiency are given in Figure 5.8 for a sub-carrier number of 64 (upper graph) and 512 (lower graph). As the number of terminals increases, the spectral efficiency of the dynamic scheme increases. However, the spectral efficiency increases also slightly for the static scheme. Initially consider the plot for the low number of sub-carriers. As the number of terminals in the cell increases, the overhead due to signaling increases, too. However, compared to the spectral efficiency gain due to the increasing multi-user diversity, the signaling overhead increases only slightly. Thus, a small performance loss has to be taken into account for the dynamic scheme including the fixed signaling field model. Compared to the static scheme, the net spectral efficiency of the dynamic scheme with signaling cost is still significantly higher, the performance advantage is about 50% for a high number of terminals in the cell.

However, if the number of sub-carriers is quite high (as in the lower plot of Figure 5.8), the signaling cost is unacceptably high. Still, for the dynamic scheme without the signaling cost the increase in spectral efficiency due to the multi-user diversity is quite impressive. With 16 terminals in the cell, the dynamic scheme outperforms the static scheme by about 35%. However, the net spectral efficiency is virtually equal to the one of the static scheme, for up to 8 terminals in the cell the performance is slightly better than the one of the static scheme. For higher terminal numbers the net performance becomes even worse compared to the static scheme. Thus, for a high number of terminals, using the dynamic scheme with this signaling variant does not pay off.

In Figure 5.9 the corresponding signaling percentage is given for both settings. In general, the signaling percentage increases logarithmically for an increasing number of terminals. However, while it increases between 1% and 6% for the low number of sub-carriers, this increase is between 5% and 40% for the setting with a high number of sub-carriers. As a result, a high loss due to signaling is observed in this case.

5.2.4 Summary

The fixed size signaling field model allows a first qualitative and quantitative evaluation of the performance behavior for dynamic MU-OFDM systems. This first evaluation reveals that the signaling cost influences the performance behavior of such systems noticeably, qualitatively and quantitatively.

The major parameters influencing the signaling cost are the system bandwidth and the frame length. Interestingly, these two parameters can be expected to influence the signaling cost of *any* signaling scheme quite strongly. Increasing the system bandwidth or increasing the frame length decreases the signaling cost quite drastically. However, systems are usually developed for a given system bandwidth. Thus, the frame length is the only major factor which can be influenced directly by the system design. Regarding the overhead, a longer frame duration is of interest. However, there are limits for the length of the frame duration, for example the coherence time of the propagation environment. Therefore, the frame length cannot be adjusted without taking such other constraints into account.

Apart from these two factors, the sub-carrier number and the terminal number have a certain impact on the signaling overhead. The sub-carrier number affects this overhead linearly. For a given system bandwidth, there exists a optimal choice of sub-carriers this bandwidth should

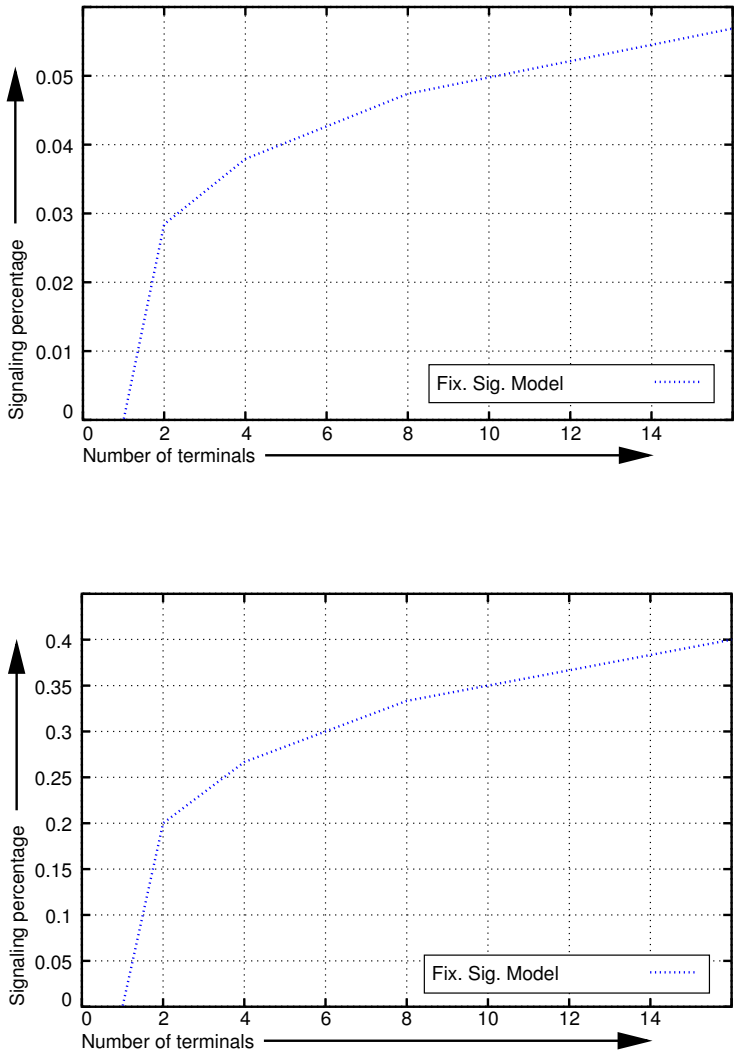


Figure 5.9: Average signaling percentage per frame for an increasing number of terminals in the cell with a fixed bandwidth divided into $N = 64$ sub-carriers (upper graph) and $N = 512$ sub-carriers (lower graph).

be split into. At this sub-carrier number the net throughput of the system is the highest. A lower number of sub-carriers leads to a loss due to the guard interval. A higher number of sub-carriers leads also to a loss, however due to the signaling overhead. In contrast, the terminal number has only a moderate impact on the signaling loss. As long as the general signaling cost is moderate, the increase or decrease of the terminal number does not change the signaling overhead dramatically. However, there exists a certain trade-off between a performance increase due to a higher multi-user diversity and a performance loss due to a higher signaling overhead.

In general, certain situations arise where the usage of a dynamic MU-OFDM system does not pay off. These situation are characterized by a high number of sub-carriers, a low system bandwidth and a short frame duration. In these situations, the signaling overhead might lead to a lower throughput than achieved by a static system. However, even in situations where the bandwidth is quite high and only a moderate number of sub-carriers exist, the signaling cost is significant (around $10 - 20\%$ of the gross throughput). Hence, more advanced methods are of interest which can decrease this signaling overhead.

5.3 Judging the Efficiency of the Fixed Size Signaling Field Model

Dynamic MU-OFDM approaches are going to be applied in rather correlated environments. This is true for the correlation in time as well as in frequency. A central requirement for the application of any OFDM system is that the system bandwidth is split into a sufficiently high number of sub-carriers. Therefore, the bandwidth of each sub-carrier is at most as large as the coherence bandwidth of the expected delay spread of the propagation environment. In general, the sub-carrier bandwidth is significantly smaller than the coherence bandwidth W_{c} in order to avoid ISI. The sub-carrier spacing, which equals the spacing of assignments in the frequency domain, is given by

$$\Delta f = \frac{B}{N} \quad . \tag{5.7}$$

For most OFDM systems the following inequality holds (otherwise, the system is strongly affected by ISI and a rather large guard duration has to be chosen, wasting a lot of transmission time for ISI mitigation):

$$R_{\mathrm{f}} = \frac{\Delta f}{W_{\mathrm{c}}} \leq 1 \quad . \tag{5.8}$$

Thus, the attenuation of adjacent sub-carriers is strongly correlated. In dynamic MU-OFDM systems this fact increases the probability that adjacent sub-carriers are *all* assigned to a certain terminal if they are in a relatively good state regarding the terminal. As a result, the coding scheme of the fixed size signaling field approach might not be efficient, leading to a rather high overhead in situations where the frequency ratio R_{f} is rather small. As seen above, these situations result in a rather low performance of the fixed size signaling field approach.

The frequency ratio is rather small either if the sub-carrier spacing Δf is small or if the coherence bandwidth W_{c} is rather large. Therefore, if the sub-carrier number is rather large or if the overall system bandwidth is small, the frequency ratio is small. The frequency ratio might also be small if the delay spread is small. Thus, the choice of certain system parameters influences the frequency ratio as well as given propagation environment parameters. As seen

in Section 5.2.3, especially scenarios with a small frequency ratio due to the choice of system parameters lead to a rather high performance loss due to signaling.

Correspondingly, the correlation in time might also lead to the possibility to decrease the signaling overhead. As discussed previously, the channel state information, which the assignments are based on, has to correspond as much as possible with the real attenuations of the sub-carriers during the following down-link phase. Thus, the coherence time of the wireless channel should be significantly larger than the length of a down-link phase. In general, it should be at least as long as the frame duration to allow for channel knowledge acquisition, channel estimation and generation of the assignments. The spacing of assignments in time is given by the frame duration, as long as one frame is not divided into sub-slots (enabling multiple assignments per frame):

$$\Delta t = T_{\mathrm{f}} \quad . \tag{5.9}$$

As for the frequency domain, a ratio between the assignment spacing in time and the coherence time can be calculated. From the above discussion, this ratio will be significantly smaller than one, otherwise the system can be expected to suffer from bit errors due to erroneous channel state information:

$$R_{\mathrm{t}} = \frac{\Delta t}{T_{\mathrm{c}}} \leq 1 \quad . \tag{5.10}$$

Whenever Inequation (5.10) holds, the fixed size signaling field approach might lead to a too high signaling overhead as its representation of assignments is not efficient. If the time ratio is significantly below one, the probability is quite high that a sub-carrier will remain in a same state regarding each terminal. Thus, a sub-carrier might be assigned to the same terminal over consecutive down-link phases. As the assignments are already known to all terminals after the first of these consecutive down-link phases, they do not have to be signaled again, unless the assignment changes. Again, scenarios where the time ratio is particularly small are characterized by a comparatively high signaling overhead. Especially, the time ratio is small if the duration of a frame is rather small which leads to many frames during one second. As seen in the previous section, as the frame length decreases the signaling load increases drastically. Therefore, in these situations a better coding might lead to a significant decrease of the signaling cost.

In order to evaluate the efficiency of the fixed size signaling field model, entropy models can be applied to the signaling information generated at the access point. The dynamic algorithm is interpreted as an information source which produces information symbols of a certain alphabet at a certain rate. Using an entropy model allows to calculate the minimal amount of information required to represent the studied source, in this case the output of the dynamic algorithm. Thus, the information rate required by the fixed size signaling field model can be compared to this minimal rate in order to estimate the efficiency. Note that the minimal rate obtained from the entropy model depends much on the stochastic characteristics of the information source included in the entropy model. Thus, the minimal rate obtained always depends on the accuracy of the applied entropy model regarding the information source.

5.3.1 Entropy Model

The basic unit of information for the fixed size signaling field approach is given by the tuple $\langle$ terminal address, modulation identification $\rangle$. As all sub-carrier assignments are transmitted,

the sub-carrier identification does not have to be signaled. For example, assume a total number of $J = 8$ terminals in the cell, and four available modulation types (BPSK, QPSK, 16-QAM and 64-QAM). In certain cases no modulation is applied to a sub-carrier as the attenuation is to high. Since this case must also be signaled to the terminals, the total number of modulation identifications used is $M = 5$. Hence, the random variables of the source $X_i^{(t)}$ $(i = 1, 2, \ldots)$ take values on the following alphabet:

$$\mathcal{A} = \{\langle 0-0\rangle, \langle 0-1\rangle, \ldots, \langle 0-4\rangle, \ldots, \langle 7-0\rangle, \langle 7-1\rangle, \ldots, \langle 7-4\rangle\}$$

The number of bits needed to signal each assignment in this example is:

$$\lceil \log_2(J) \rceil + \lceil \log_2(M) \rceil = 6 \text{ bits} \tag{5.11}$$

In this case 3 bits are used to signal both the terminal address and the modulation identification. However, since the size of the complete alphabet is $\mathcal{A} = 40$ (there are 40 different assignment tuples), a zero-order model like the one presented in Section 2.3 would need, in theory, less than 6 bits:

$$H = \log_2(m) = \log_2(J) + \log_2(M) = 3 + 2.32 = 5.32 \text{ bits} \tag{5.12}$$

In this equation, H denotes the entropy of the source, i.e. a lower bound on the number of bits required to represent the information of the source (see Section 2.3).

This would bring a compression gain of 11.33%, comparing to using 6 bits to signal the assignment information. However, a zero-order model is not accurate as the statistical properties of the signaling information source are not modeled appropriately. Besides the first order statistics, correlation in time and frequency are present. Thus, second- and third order models are chosen to investigate the compression gain achievable in time and frequency. Note that no rigorous compression limit is of interest in this section.

Three different entropy models are used here: A second order model in frequency, a second order in time, and a third order model in time and frequency. For a summary on entropy models, refer to Section 2.3.

Second Order Model in Frequency

In this model only dependencies between sub-carrier assignments in the frequency domain are modeled. The knowledge of the adjacent assignment in frequency influences the value of the entropy by reducing it (i.e. $H(X_i^{(t)}|X_{i-1}^{(t)}) \leq H(X_i^{(t)})$. The current sub-carrier assignment $X_i^{(t)}$ depends only on the adjacent sub-carrier assignment $X_{i-1}^{(t)}$, but is independent of all the other ones $X_1^{(t)}$,$X_2^{(t)}$,...,$X_{i-2}^{(t)}$. Then, the total number of bits needed to signal the information related to a certain downlink phase is:

$$H(X_1^{(t)}) + \sum_{i=2}^{N} H(X_i^{(t)}|X_{i-1}^{(t)}) = H(X_1^{(t)}) + (N-1) \cdot H(X_i^{(t)}|X_{i-1}^{(t)}) \quad . \tag{5.13}$$

The first order statistics $p_j = P(X = j)$ of each information source symbol j as well as the conditional probabilities $p_{k|j}$ are used to calculate the value of the second-order entropy in

frequency. The conditional probability $p_{k|j}$ denotes here the probability that assignment $X_i^{(t)}$ equals the k-th source symbol while the assignment of the adjacent sub-carrier $X_{i-1}^{(t)}$ equals the j-th information symbol.

$$H(X_i^{(t)}|X_{i-1}^{(t)}) = \sum_{j=1}^{m} p_j \sum_{j=1}^{m} p_{k|j} \cdot \log_2\left(1/p_{k|j}\right) \tag{5.14}$$

Second Order Model in Time

Accordingly, a second order entropy model in time can be formulated. In this case, the knowledge of the previous assignment of any sub-carrier is exploited, which reduces the amount of bits needed to represent the source: $H(X_i^{(t)}|X_i^{(t-1)}) \leq H(X_i^{(t)})$. It is assumed that the current assignment $X_i^{(t)}$ of sub-carrier n only depends on the assignment $X_i^{(t-1)}$ of this sub-carrier during the previous down-link phase. However, the assignment of the sub-carrier for previous down-link phases is assumed to be independent. This results in a number of bits spent per source symbol of:

$$\sum_{i=1}^{N} H(X_i^{(t)}|X_i^{(t-1)}) = N \cdot H(X_i^{(t)}|X_i^{(t-1)}) \quad . \tag{5.15}$$

In this case the conditional probability is calculated as given in Equation 5.14. The conditional probability $p_{k|j}$ denotes the probability that assignment $X_i^{(t)}$ equals the k-th source symbol while the assignment of the same sub-carrier during the previous down-link phase $X_i^{(t-1)}$ equaled the j-th information symbol.

Third Order Model in Time and Frequency

The last mode considered is an entropy model in time and frequency. Thus, the information symbol $X_i^{(t)}$ depends on the adjacent sub-carrier assignment $X_{i-1}^{(t)}$ as well as on the previous assignment of the same sub-carrier $X_i^{(t-1)}$. No further correlation is assumed in the model (either in time or in frequency). The number of bits required to represent a information source symbol is obtained as :

$$H(X_1^{(t)}) + \sum_{i=2}^{N} H(X_i^{(t)}|X_{i-1}^{(t)}, X_i^{(t-1)}) = H(X_1^{(t)}) + (N-1) \cdot H(X_i^{(t)}|X_{i-1}^{(t)}, X_i^{(t-1)}) \quad . \tag{5.16}$$

In order to calculate the third order entropy $H(X_i^{(t)}|X_{i-1}^{(t)}, X_i^{(t-1)})$, the first, second and third order statistics are required. The entropy is then obtained from:

$$H(X_i^{(t)}|X_{i-1}^{(t)}, X_i^{(t-1)}) = \sum_{j=1}^{m} p_j \sum_{k=1}^{m} p_{k|j} \sum_{l=1}^{m} p_{l|k,j} \cdot \log_2\left(1/p_{l|k,j}\right) \quad . \tag{5.17}$$

Here $p_{l|k,j}$ is the probability of assignment $X_i^{(t)}$ being the l-th symbol in the source alphabet $\mathcal{A}$, when $X_{i-1}^{(t)}$ is the k-th symbol and $X_i^{(t-1)}$ the j-th symbol.

5.3.2 Performance Evaluation

In order to study the efficiency of the fixed size signaling field model, different scenarios are investigated in which either R_{t} or R_{f} vary significantly. While always one of these ratios is varied, the other one is kept fixed. As performance metric the compression ratio between the entropy[2] of the signaling data source and the bit rate of the fixed size signaling field model is considered. In addition, the impact on the net throughput is also given. Apart from the three different entropy models, also the throughput for the dynamic scheme without signaling loss is given as well as the net throughput of the dynamic scheme combined with the fixed size signaling field model. The basic system model and simulation scenario is equivalent to the one discussed in Section 5.2.3.

Two scenarios are chosen in which the time ratio R_{t} was varied. Also, two different scenarios are chosen where the frequency ratio R_{f} is varied. In general, there are two ways to vary one of these ratios, either varying the numerator (the spacing of the assignment resources in the corresponding dimension) or varying the denominator (which refers to the coherence length in the corresponding dimension). For each ratio, one scenario is chosen where the numerator is changed and one where the denominator is changed.

Methodology

In order to obtain the entropy of the fixed size signaling field source, all assignments generated by the dynamic scheme with the fixed size signaling model are recorded in a file for a certain time span. This signaling record is then analyzed regarding its information alphabet symbol statistics. Thus, analyzing the complete file, a statistic of the source with respect to the entropy models is generated. For this statistic the entropy is calculated as mentioned above, yielding a required number of bits to be used per source alphabet symbol. The compression ratio is then obtained by dividing this number of bits per alphabet symbol by the required number of bits required to code a symbol by the fixed size signaling field model. This ratio is subtracted from one to obtain the compression ratio.

In order to obtain the net throughput for a certain entropy model, the ratio between the bits per symbol for the entropy model and the bits per symbol for the fixed size signaling field model is multiplied by the signaling percentage. Using this new signaling percentage, the net throughput is generated by multiplying the gross throughput with the down-link fraction available for payload data transmission.

As both metrics are calculated considering the statistics of the complete signaling record, the obtained results have to be viewed as upper limit for this dynamic algorithm. The knowledge of the statistics of the source are essential for estimating the entropy of the source. Thus, a coding scheme cannot achieve the calculated compression ratios as it has not a priori knowledge of the information source statistics in general. In order to obtain an upper limit, though, this a priori knowledge can be assumed without degradation of the results.

[2] As discussed above, the entropy of any data source is the smallest information rate at which this source can be compressed taking the so far considered statistics order into account. Thus, the entropy for a discrete source can be expressed in bits per second.

Varying the Frequency Ratio R_f

For the frequency ratio the sub-carrier number is increased for the first scenario while the delay spread and the bandwidth are constant ($\Delta\sigma = 0.15$ μs, $B = 16.25$ MHz). Thus, the frequency ratio decreases, as shown in Table 5.4, from $R_\mathrm{f} = 0.24$ down to $R_\mathrm{f} = 0.03$ due to an increasing number of sub-carriers. For this scenario the time ratio is set to $R_\mathrm{t} = 0.16$, resulting from a frame length of $T_\mathrm{f} = 2$ ms and a maximum speed in the propagation environment of $v_\mathrm{max} = 1$ m/s. In Figure 5.10 the results are given for the compression ratio (upper graph) and the

Sub-carrier Number N	**Sub-carrier Spacing** Δf	**Frequency Ratio** R_f
64	253.9 kHz	0.24
128	126.9 kHz	0.12
256	63.5 kHz	0.06
512	31.75 kHz	0.03

Table 5.4: Relationship between the sub-carrier number and the frequency ratio for a fixed bandwidth of $B = 16.25$ MHz and a delay spread of $\Delta\sigma = 0.15$ μs.

resulting net throughput (lower graph). While the compression ratio is constant for the entropy model in time, it varies strongly for the other two entropy models. Obviously, the entropy model in time cannot benefit from an increasing R_t , as it is chosen constant in this scenario. Note that the compression gain for the entropy model in time is about 25%.

The other two entropy models achieve much higher compression gains. The entropy model in frequency achieves a compression gain between 50 and 70%. Initially, the compression gain increases with the increasing number of sub-carriers. At $N = 256$ sub-carriers, the maximum compression gain is reached with 71%, thereafter the compression gain decreases despite the still decreasing frequency ratio R_f . Note that at the corresponding frequency ratio of $R_\mathrm{f} = 0.16$ the compression ratio is significantly higher for the entropy model in frequency than it is for the entropy model in time at a (always constant) time ratio of $R_\mathrm{t} = 0.16$.

The entropy model in time and frequency basically achieves the same performance as the entropy model in frequency. However, the achieved compression ratio is constantly higher by about 7%. Thus, at $N = 256$ sub-carriers this scheme reaches a maximum ratio of 78%. Thereafter, the compression ratio is also decreasing for this type of entropy model.

These results for the compression ratio translate into an improved performance of the fixed size signaling field model regarding the net throughput. As already observed in the previous section, as the sub-carrier number increases the gross throughput of any OFDM system increases, too. In contrast, the signaling load increases also the higher the sub-carrier number is chosen. As a result, the net throughput increases at first. Beyond an optimal net throughput reached for $N = 250$, the net throughput decreases again. The second scenario is based on increasing the delay spread while keeping the number of sub-carriers constant. The number of sub-carriers is set to $N = 128$, while the bandwidth is $B = 16.25$ MHz. The delay spread increases, as shown in Table 5.5, from $\Delta\sigma = 0.1$ μs up to $\Delta\sigma = 0.5$ μs[3]. Note that the guard interval

[3]In the considered indoor propagation environment a delay spread of $\Delta\sigma = 0.5$ μs is rather large. However, for illustration purposes is was chosen here to be considered. In real environments it will be observed rarely, though.

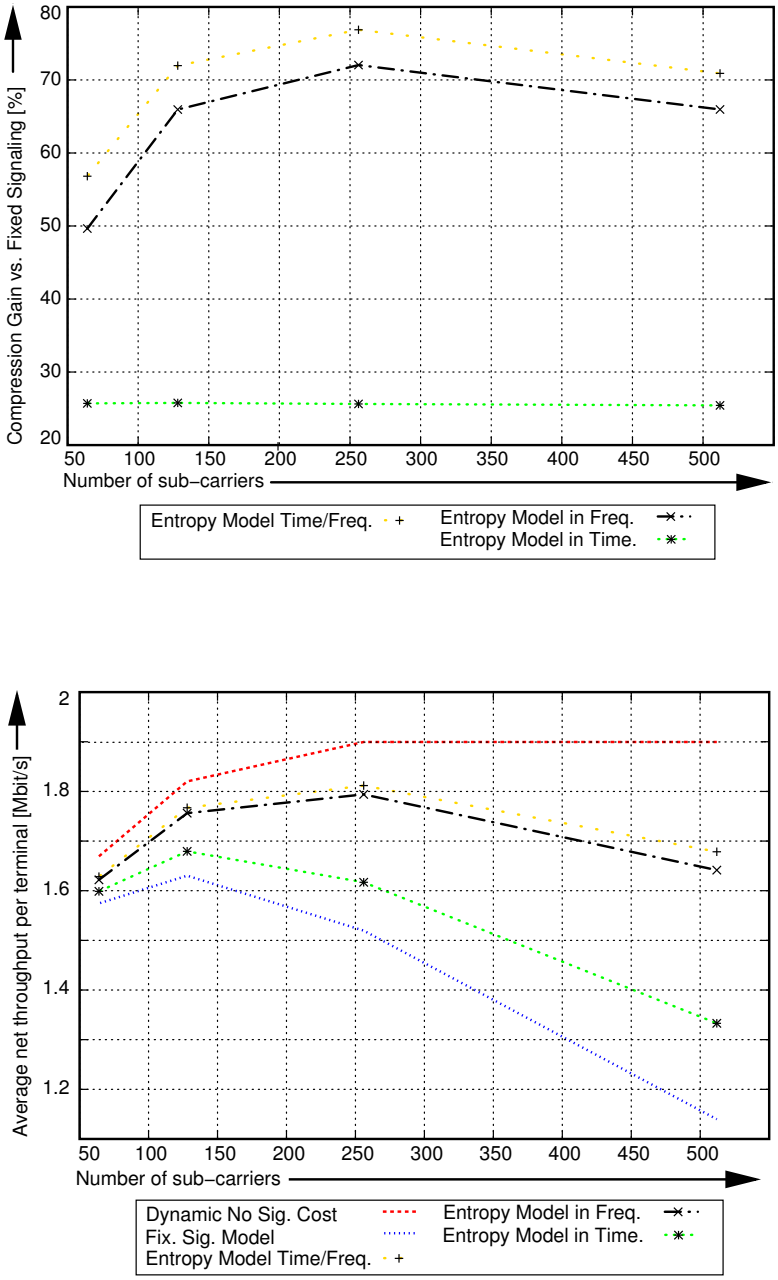


Figure 5.10: Compression rate (upper graph) and net throughput (lower graph) applying different entropy models to the output of the fixed size signaling field model for a varying sub-carrier number ($J = 8$, $\Delta\sigma = 0.15$ μs, $v_{\max} = 1$ m/s, $T_{\mathrm{f}} = 2$ ms).

was not increased, as it is chosen to be significantly larger than any chosen delay spread value ($T_\mathrm{g} = 0.8\ \mu\mathrm{s}$). Thus, the frequency ratio increases from $R_\mathrm{f} = 0.08$ up to a ratio of 0.4. The time ratio is again fixed at $R_\mathrm{t} = 0.16$, with the same parameter settings as in the previous scenario. The compression rate and net throughput results for this scenario are given in Figure 5.11. As the

Delay Spread $\Delta\sigma$	**Coherence Bandwidth** W_c	**Frequency Ratio** R_f
$0.1\ \mu\mathrm{s}$	$1591\ \mathrm{kHz}$	0.08
$0.2\ \mu\mathrm{s}$	$795\ \mathrm{kHz}$	0.16
$0.3\ \mu\mathrm{s}$	$530\ \mathrm{kHz}$	0.24
$0.5\ \mu\mathrm{s}$	$318.2\ \mathrm{kHz}$	0.4

Table 5.5: Relationship between the delay spread and the frequency ratio for a fixed bandwidth of $B = 16.25\ \mathrm{MHz}$ divided into $N = 128$ sub-carriers.

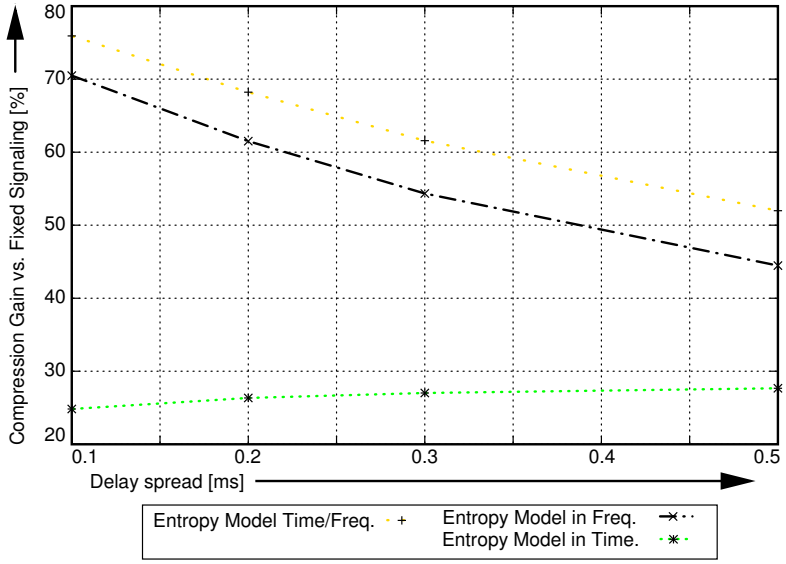
delay spread increases, the entropy models based on frequency yield a decreasing compression rate. At a low delay spread the compression rate for the entropy model in frequency is about 70%; it decreases down to a ratio of 45% for the largest considered delay spread.

The entropy model based on time and frequency also has a decreasing compression performance. However, the offset between it and the entropy model in frequency increases slightly. At a small delay spread the compression ratio is at 77%, for the largest considered delay spread the compression ratio is about 54%. In contrast, the compression ratio of the entropy model in time *increases* slightly from 25% up to a ratio of 29%.

Again, the net throughput results are effected by two different (contrary) effects. First of all, the gross throughput increases as the delay spread increases. This is due to an increasing frequency diversity as the delay spread increases. Thus, the performance of the fixed size signaling field model increases, too, as it is just offset from the gross throughput. However, for the entropy models in frequency the compression rate decreases for an increasing delay spread. As a result, the net throughput first increases up to a delay spread of $0.3\ \mu\mathrm{s}$ and stays constant thereafter. The performance advantage compared to the pure fixed signaling field model is still significant, the loss compared to the dynamic scheme without signaling cost is only about $100\ \mathrm{kBit/s}$ for the lower delay spread values. Again the entropy model in time and frequency outperforms the entropy model in frequency slightly. The entropy model in time achieves a constant improvement of $100\ \mathrm{kBit/s}$ compared to the pure fixed size signaling field model's performance. However, it is significantly outperformed by the other two entropy models.

Varying the Time Ratio R_t

The first scenario for the variation of the time ratio $R_\mathrm{t}$is chosen as described in Table 5.6. In this scenario the frame length T_f decreases, starting from a value of $T_\mathrm{f} = 3\ \mathrm{ms}$, down to a value of $0.37\ \mathrm{ms}$. The maximum speed of the propagation environment is fixed to $v_\mathrm{max} = 1\ \mathrm{m/s}$. In accordance to the first scenario (varying the frequency ratio), here the time ratio takes exactly the same values. By varying the frame length while the maximum speed is set constant, the time ratio varies due to the choice of a system parameter (in contrast to varying the speed, which is a environment parameter). The frequency ratio is set to $R_\mathrm{f} = 0.16$ by setting $N = 128$ and



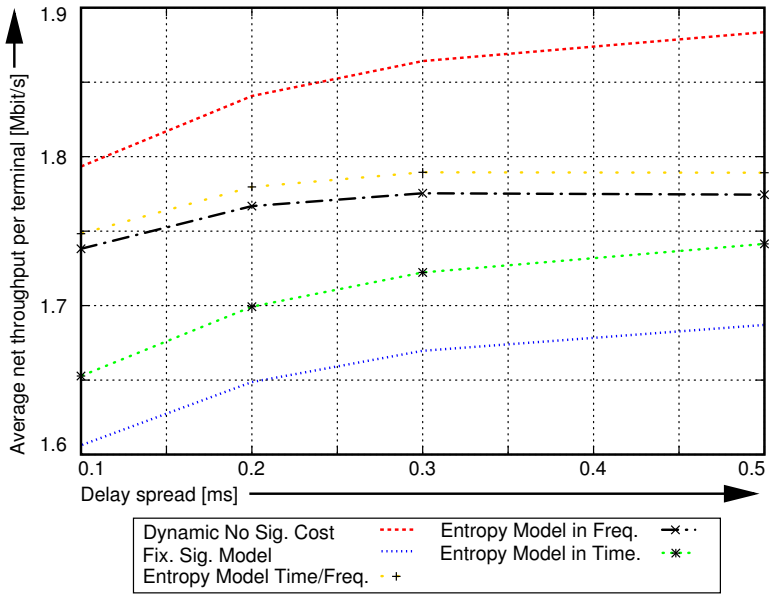


Figure 5.11: Compression rate (upper graph) and net throughput (lower graph) applying different entropy models to the output of the fixed size signaling field model for a varying delay spread ($J = 8$, $N = 128$, $v_{\max} = 1$ m/s, $T_{\mathrm{f}} = 2$ ms).

$\Delta\sigma = 0.2$ μs. In Figure 5.12 the results are shown for the compression ratio (upper graph)

Frame length (Time spacing) $T_\mathrm{f} = \Delta t$	**Time Ratio** R_t
3 ms	0.24
1.5 ms	0.12
0.75 ms	0.06
0.37 ms	0.03

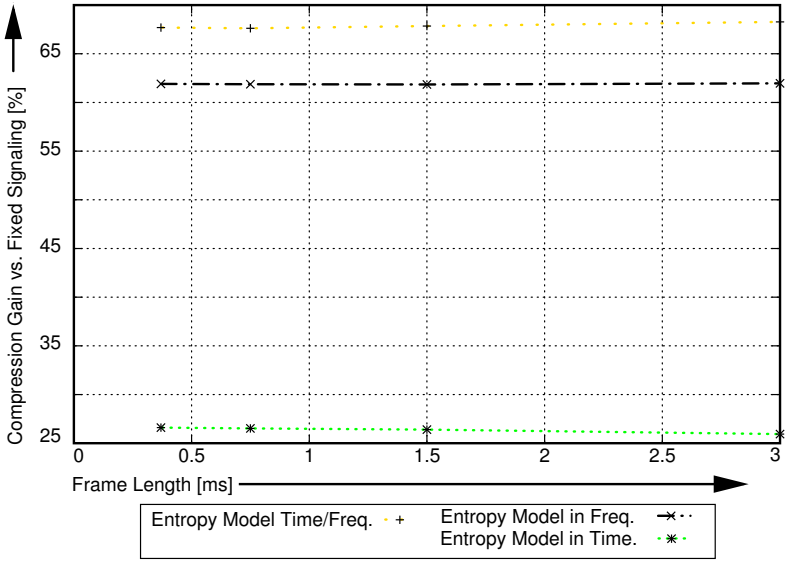
Table 5.6: Relationship between the frame length and the time ratio for a fixed center frequency of $f_\mathrm{c} = 5.2$ GHz and a maximum speed in the propagation environment of $v_\mathrm{max} = 1$ m/s.

as well as for the resulting net throughput (lower graph). As the frame length increases, the compression of all three entropy models does not change very much. The entropy model in time provides a compression ratio of about 26% which decreases slightly as the frame length increases. The entropy model in frequency provides a much higher compression ratio of 62%, which stays constant for all considered frame lengths. The entropy model in time and frequency provides the highest compression gain of about 68%. This compression gain stays constant for all considered frame lengths.

This behavior of the compression gain contrasts the observed behavior when varying the frequency ratio by increasing the number of sub-carriers. When the sub-carrier number increases, the entropy model in frequency had a significant increase in the compression ratio (in fact, the entropy model in time and frequency has such an increasing compression ratio, too). The entropy model in time had a constant behavior.

As the frame length decreases, the expectation is to observe an increasing compression ratio for the entropy model in time. However, this compression ratio stays constant, as described above.

Regarding the net throughput, the observed behavior of the compression ratios directly yields the qualitative performance difference for the three entropy models. As the frame length varies, the gross throughput stays constant. Thus, smaller frame length leads to a higher signaling loss. As all three entropy models achieve a constant compression ratio, their relative gain compared to the fixed size signaling field model is constant. Especially for a small frame length, the entropy models in frequency provide a large net throughput gain compared to the fixed size signaling field model. However, although their compression ratio is quite large, the loss compared to the dynamic scheme without signaling cost is significant (about 400 kBit/s). The entropy model in time provides only a moderate performance gain compared to the fixed size signaling field model. The achieved performance gain is rather small (in terms of net throughput) as the frame length is large. For the last scenario considered, the maximum speed of the propagation environment is varied. Here, the resulting time ratios equal the values of the frequency ratio for varying the delay spread (the second scenario for varying the frequency ratio). The exact values are given in Table 5.7. The frame length is fixed at $T_\mathrm{f} = 2$ ms, the frequency ratio equals in this setting $R_\mathrm{f} = 0.16$ with the same settings for the delay spread and the sub-carrier number as in the previous scenario. As in the case of varying the frame length, the compression ratios stay almost constant while the maximum speed increases. In Figure 5.13 (upper graph) the compression ratio is given for all three considered entropy models. While the entropy model in



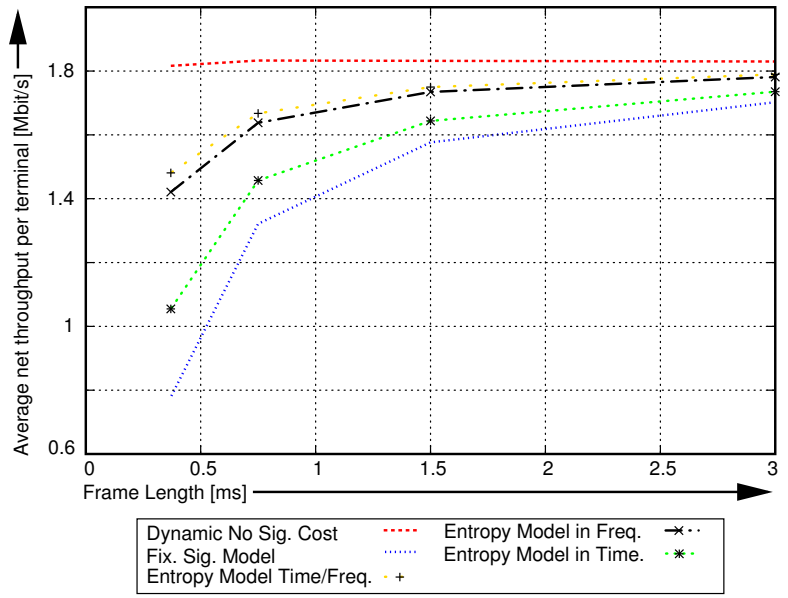


Figure 5.12: Compression rate (upper graph) and net throughput (lower graph) applying different entropy models to the output of the fixed size signaling field model for a varying frame length ($J = 8$, $N = 128$, $\Delta\sigma = 0.2$ μs, $v_{\max} = 1$ m/s).

Maximum Speed $v_{\max}$	**Coherence Time** T_{c}	**Time Ratio** R_{t}
$0.5\ \mathrm{m/s}$	$25\ \mathrm{ms}$	0.08
$1\ \mathrm{m/s}$	$12.5\ \mathrm{ms}$	0.16
$1.5\ \mathrm{m/s}$	$8.33\ \mathrm{ms}$	0.24
$2.5\ \mathrm{m/s}$	$5\ \mathrm{ms}$	0.4

Table 5.7: Relationship between the maximum speed and the time ratio for a fixed center frequency of $f_{\mathrm{c}} = 5.2\ \mathrm{GHz}$ and a frame length of $T_{\mathrm{f}} = 2\ \mathrm{ms}$.

frequency behaves as expected (constant compression gain at 62%), the compression gain of the entropy model in time decreases. However, the decrease is quite moderate starting at 26% and ending at 25%. The entropy model in time and frequency provides the best compression gain at about 68%. As the speed increases this compression ratio is also decreased by 1% for this entropy model.

Considering the corresponding net throughput results, note first that the gross throughput (dynamic approach without signaling cost) increases slightly as the speed increases. This is due to an increase in diversity (as the coherence time decrease, sub-carriers tend to have faster changing states leading to an increasing state diversity). This increasing behavior is observed for all net throughput curves, however, the increase is rather low in all cases. As already indicated by the compression ratio results, the entropy models in frequency provide a rather high throughput gain compared to the pure fixed size signaling field model of about $230\ \mathrm{kBit/s}$. The entropy model in time only provides an additional throughput of $100\ \mathrm{kBit/s}$.

5.3.3 Summary

As OFDM systems are parameterized such that the correlation in time and frequency is rather high, the fixed size signaling field approach is rather inefficient regarding its representation of the signaling data. In all above considered cases, the compression ratio for all entropy models is significant even if the corresponding correlation is rather weak (high time or frequency ratio). As substantial compression ratios are achieved, the net throughput can be increased significantly as well.

The inefficient representation by the fixed size signaling field model leads to a lower net throughput in all situations. However, this inefficiency is more evident in cases where the corresponding ratio in time or frequency is low due to a close spacing of the assignment resources (choosing a lot of sub-carriers or small frame times). As the signaling load is quite high then anyway, the loss due to the inefficient representation becomes evident in particular. In contrast, if the signaling load is rather small anyway (low number of sub-carriers, large frame times) the inefficient representation might even be tolerated.

The applied entropy models reveal a clear discrepancy between the efficiency in time and the efficiency in frequency. While the compression ratio is quite high whenever considering one of the two entropy models in frequency, the entropy model in time provides a quite constant compression ratio of about 25% regardless of the scenario. Even varying the time ratio R_{t} does not lead to a major change of the compression ratio for the entropy model in time. Thus, the fixed

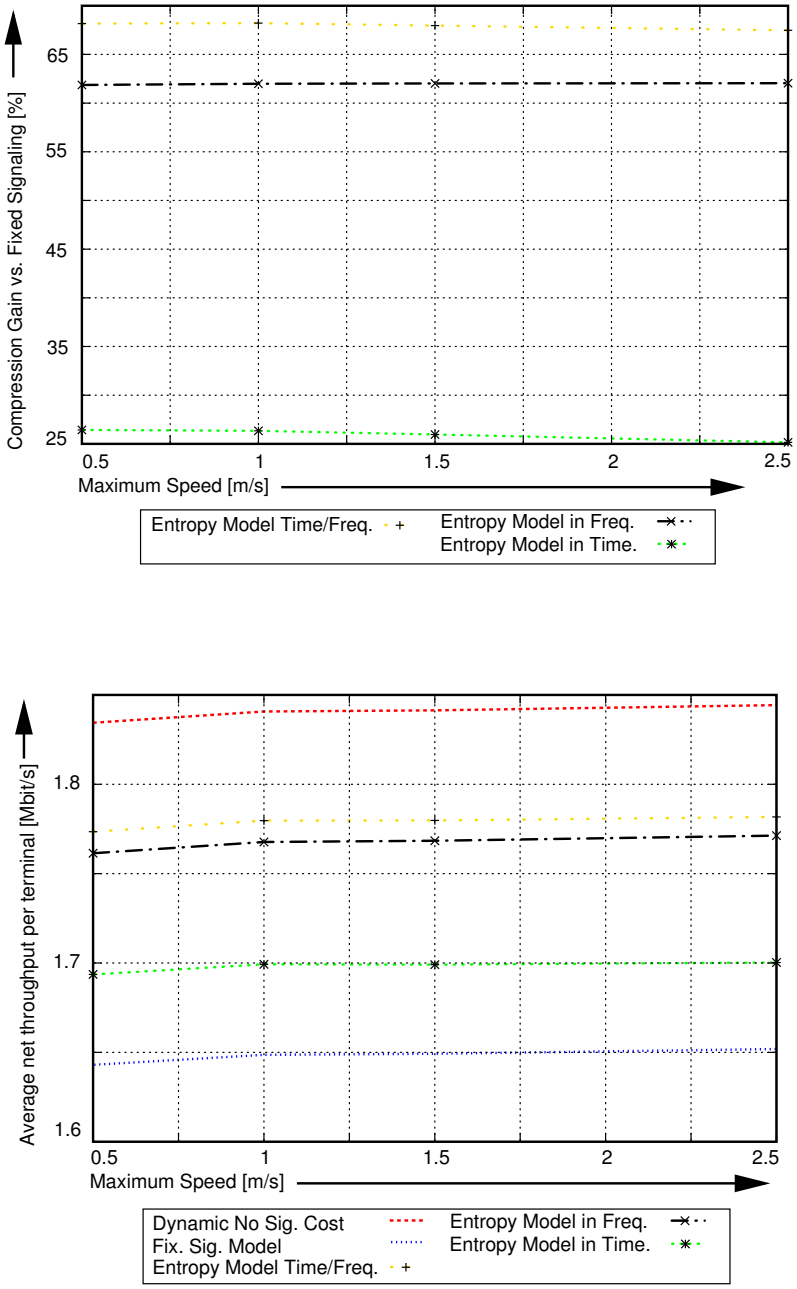


Figure 5.13: Compression rate (upper graph) and net throughput (lower graph) applying different entropy models to the output of the fixed size signaling field model for a varying maximum speed within the propagation environment ($J = 8$, $N = 128$, $\Delta\sigma = 0.2$ μs, $T_{\mathrm{f}} = 2$ ms).

size signaling field model can probably be much better compressed considering the frequency domain than the time domain. This is not related to the strength in correlation, as the above presented investigations were based on identical correlation strength in time and frequency. Note that the best performance is still achieved for the entropy model in time and frequency.

This discrepancy between the time domain and the frequency domain is probably related to the fact that the generated assignments depend on each other in frequency but not in time. The dynamic algorithm (solving problem (5.6)) which is considered in this case decides the assignment of each frequency resource (sub-carrier) for a fixed time instance. Multiple time instances are not taken into consideration. Therefore, if a certain sub-carrier is in a good state for multiple down-link phases, the probability is much lower that this sub-carrier will be assigned to the same terminal for most of these down-link phases. In contrast, if a block of sub-carriers is in a relatively good state regarding this terminal, the probability is much higher that most (or at least some) of these sub-carriers are assigned to the terminal for a single down-link phase.

Obviously, the derived entropy results (upper limits) for the fixed size signaling field model motivate the investigation of compression schemes to be applied to the fixed size signaling field model. However, a second issue arises. As the correlation in each dimension can be expected to be rather large, the signaling model itself (thus, the information source of the entropy model) might be changed. For example, the dynamic algorithm could be changed such that it also considers the assignments of the previous down-link phase in order to exploit the correlation in time. In fact, such a change of the signaling model can be driven to increase the compression rate of an entropy model as well as to minimize directly the number of signaling symbols required, as is shown in the next section.

Despite the significant compression gains achieved in the considered scenarios, still the system parameters N and T_{f} can be set to values for which the net throughput is decreased quite drastically. However, much more "aggressive" values for N and T_{f} can be overcome by the entropy models without much loss in terms of the net throughput, compared to the pure fixed size signaling field model.

5.4 Reduction of the Signaling Overhead

As the fixed size signaling field model leaves room for improvement, the signaling overhead can probably be reduced by exploiting the correlation of sub-carrier states in time and/or frequency. The entropy results of the previous section indicate that compression is one way to reduce the signaling overhead. Using a lossless compression algorithm, the fixed size signaling field information can be represented more efficient, which pays off especially in situations where the signaling load is rather high (short frame length, high sub-carrier number, low overall bandwidth).

Alternatively, the correlation in time and frequency can be exploited by including the signaling cost into the optimization model. This requires a different representation of the signaling information compared to the fixed size signaling field model, which is generally called the variable size signaling model. Depending on the correlation to be exploited, a bit cost for assigning a certain sub-carrier to a certain terminal with a certain modulation type is calculated. Using this bit cost, the required number of signaling symbols can be obtained if the number of assignments

is given. Finally, this yields an analytic expression of the net capacity, which enables in a next step its optimization.

While both schemes, compression and optimization, potentially yield a better system performance, they also have disadvantages. Primarily, both schemes demand additional computational resources. This increases the cost of the system as well as the power consumption. In case of the compression, additional computational resources are required at the access point as well as at the terminal. In contrast, the optimization of the signaling overhead requires only additional computational resources at the access point. This might lead to a higher cost of the terminal devices in case of the compression. Also, a higher power supply is in general no problem at the access point, while power is a scarce resource at the terminal. Thus, from this point of view the optimization approach seems to be more favorable.

Apart from the cost and power constraints, further issues arise regarding the protocol behavior of the signaling system. As bit errors might occur during the transmission of the signaling information, the impact of errors can be significant if compression schemes are applied or if the signaling information has been optimized. This leads to more complicated correction schemes, degrading the system performance more as in the case of the fixed size signaling field model. Thus, for each scheme a separate correction scheme has to be developed and analyzed.

In the following, all these issues are investigated regarding the compression and optimization approach. Then, the performance of the various schemes is evaluated, also taking into consideration the above discussed potential drawbacks.

5.4.1 Optimization Models for the Net Throughput

Multiple options arise in order to optimize the net throughput. Either the correlation in time is exploited to save signaling overhead (compared to the fixed size signaling field model) or the correlation in frequency is exploited. The third option, exploiting both correlations, is obtained from combining these two approaches.

In order to maximize the net throughput, two different aspects have to be considered simultaneously. The net throughput is the product of the down-link symbols remaining for payload data transmission and the throughput achieved per symbol by a certain assignment of sub-carriers. Therefore, the net throughput can be increased by decreasing the number of signaling symbols required. This increases the remaining down-link symbols for payload transmission. On the other hand, the net throughput is increased if a suitable assignment is generated, exploiting the diversity of the sub-carrier state matrix.

Unfortunately, the number of required signaling symbols depends directly on the chosen assignments. Therefore, the maximization of the net throughput requires to balance the loss due to the required number of signaling symbols and the gain due to choosing suitable assignments. As will turn out, this maximization problem belongs to the class of quadratic integer optimization problems, which are generally known to be difficult to solve.

Exploiting the Correlation in Time

In order to exploit the correlation in time it is possible to signal only the “new” assignments, which had been assigned to a different terminal in the prior down-link phase. As a consequence,

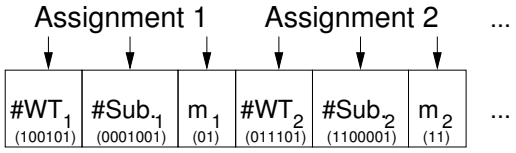


Figure 5.14: Variable size signaling field model: Representation of assignment information for one down-link phase in order to exploit the correlation in time.

the binary representation of one single assignment becomes now more "expensive", compared to the representation of a single assignment in the fixed size signaling field model. Specific sub-carriers have to be addressed, which additionally requires $\lceil \log_2(N) \rceil$ bits. Accordingly, the terminal address and modulation identification will consume $\lceil \log_2(J) \rceil$ bits and $\lceil \log_2(M) \rceil$ bits such that an assignment change from one down-link phase to the next requires

$$C_{\text{sig}} = \lceil \log_2(N) \rceil + \lceil \log_2(J) \rceil + \lceil \log_2(M) \rceil \tag{5.18}$$

bits. The layout of the resulting signaling field is shown in Figure 5.14. Using the bit size of C_{sig} enables the determination of the exact number of signaling symbols ς required for down-link phase t, once the assignments have been generated. Knowing the assignments $\mathbf{X}^{(t-1)}$ for down-link phase $t-1$, the cost for assigning sub-carrier n to terminal j in phase t is

$$c_{j,n}^{(t)} = \begin{cases} 0 & \text{if } x_{j,n}^{(t-1)} = 1, \\ C_{\text{sig}} & \text{otherwise.} \end{cases}$$

Obviously, the signaling cost of Equation 5.18 only applies if an assignment of the last down-link phase is changed. Taking into account the assignments of the previous frame when determining new assignments is the crucial difference to the fixed size signaling field approach. Per single assignment, the bit cost is higher in case of this variable size signaling cost model. However, if the correlation in time is high, significantly less assignments have to be changed, leading to the reduction of the symbols required for signaling.

Given a valid assignment $\mathbf{X}^{(t)}$ for the current down-link phase, the required number of signaling symbols is easily calculated according to Equation 5.19. It is still assumed that during the signaling phase all sub-carriers can be utilized to transmit the signaling information. In addition, the same, commonly known modulation type b_{sig} is applied during the signaling phase.

$$\varsigma = \left\lceil \frac{\sum_{j,n} c_{j,n}^{(t)} \cdot x_{j,n}^{(t)}}{N \cdot b_{\text{sig}}} \right\rceil \tag{5.19}$$

In order to optimize the net throughput, this number of signaling symbols has to be decreased while simultaneously increasing the capacity of the cell. Based on the modeling of the signaling cost for the variable size signaling field model, it is obvious that more "new" assignments can probably increase the overall capacity while it also increases the number of signaling symbols required. The corresponding formulation of the optimization problem is stated below as

problem (5.20).

$$\begin{aligned} \max \quad & \left(S - \left\lceil \frac{\sum_{j,n} c_{j,n}^{(t)} \cdot x_{j,n}^{(t)}}{N \cdot b_{\text{sig}}} \right\rceil \right) \cdot \sum_{j,n} \left(b_{j,n}^{(t)} \cdot x_{j,n}^{(t)} \right) \\ \text{s. t.} \quad & \sum_{j} x_{j,n}^{(t)} \leq 1 \qquad \forall\, n \\ & \sum_{n} x_{j,n}^{(t)} \leq l_j \qquad \forall\, j \quad . \end{aligned} \tag{5.20}$$

The objective is nonlinear and the model cannot be solved by standard algorithms for linear programming. However, the following solution approach is proposed for the problem, which is also used below to obtain the performance results.

Optimization problem (5.20) is reformulated by considering the already introduced integer variable ς, denoting the number of OFDM symbols required for conveying the signaling information:

$$\begin{aligned} \max \quad & (S - \varsigma) \cdot \sum_{j,n} \left(b_{j,n}^{(t)} \cdot x_{j,n}^{(t)} \right) \\ \text{s. t.} \quad & \text{(5.1), (5.4)} \\ & \varsigma \geq \frac{\sum_{j,n} c_{j,n}^{(t)} \cdot x_{j,n}^{(t)}}{N \cdot b_{\text{sig}}} \\ & \varsigma \text{ integral} \quad . \end{aligned} \tag{VAR}$$

This problem is quadratic since the objective involves the product of (integral) variables. In linear optimization, there are several ways of dealing with this. A quadratic programming approach does not apply here since the matrix of coefficients is not positive definite. An alternative solution approach is a selective enumeration. Consider a reformulation of problem (VAR) that essentially introduces one integral variable and two constraints for each product of ς with a binary variable. However, there are rather few possible values that ς can take: There can be at most N assignments that change with respect to the last down-link phase $t-1$ if all sub-carriers are reassigned. This implies that

$$0 \leq \varsigma \leq \left\lceil \frac{C_{\text{sig}}}{b_{\text{sig}}} \right\rceil \quad .$$

Even for large values of the sub-carrier number ($N = 1024$) and the number of terminals ($J = 64$), C_{sig} , as calculated in (5.18), is never greater than 20. Thus, enumerating the possible values for ς is a viable alternative.

To avoid enumeration of *all* possible values for ς , the following scheme can be employed. The problem is divided into subproblems by imposing bounds $0 \leq l \leq u \leq \lceil C_{\text{sig}}/b_{\text{sig}} \rceil$ on the number of symbol times used for signaling:

$$\begin{aligned} \max \quad & \text{VAR} \\ \text{s. t.} \quad & l \leq \varsigma \leq u \end{aligned} \tag{VAR'$\langle l, u \rangle$}$$

The value of the optimum regarding problem (VAR'$\langle l, u\rangle$) can be estimated by solving the easier problem (PLAIN) with some additional constraints, that is,

$$\begin{aligned} \max \quad & \text{PLAIN} \\ \text{s. t.} \quad & l - 1 < \sum_{j,n} \frac{c_{j,n}^{(t)} \cdot x_{j,n}^{(t)}}{N \cdot b_{\text{sig}}} \leq u \end{aligned} \qquad \text{(PLAIN}\langle l, u\rangle\text{)}$$

and applying the relation

$$\frac{(S-u)}{S} \cdot \text{PLAIN}\langle l, u\rangle \leq \text{VAR'}\langle l, u\rangle \leq \frac{(S-l)}{S} \cdot \text{PLAIN}\langle l, u\rangle \quad . \tag{5.21}$$

In order to solve problem (VAR) one possible solution approach works as follows:

1. Obtain a starting value by evaluating $\frac{(S-\varsigma_0)}{S} \cdot \text{PLAIN}\langle \varsigma_0, \varsigma_0\rangle$. For ς_0 the number of signaling symbols from the optimal solution of problem (5.6) can be used. Alternatively, the optimal number of signaling symbols of the previous down-link phase can be used for ς_0.

2. Check for better solutions with $\varsigma_+ = \varsigma_0 + 1,2,\ldots$ bits spent on signaling by solving (VAR'$\langle l, u\rangle$) with $l = u = \varsigma_+$, stopping as soon as $\frac{(S-\varsigma_+)}{S} \cdot \text{PLAIN}\langle \varsigma_+, C_{\text{sig}}/b_{\text{sig}}\rangle$ is less than the currently best solution value (all higher values of ς_+ can be excluded).

3. Check for better solutions with $\varsigma_- = \varsigma_0 - 1,2,\ldots$ bits spent on signaling by solving (VAR'$\langle l, u\rangle$) with $l = u = \varsigma_-$, stopping as soon as PLAIN$\langle 0, \varsigma_-\rangle$ is less than the currently best solution value (all lower values of ς_- can be excluded).

Empirically, the optimal solution is often close to ς_0 and enumeration of most possible values for ς can be skipped by the above method. However, this solution method still requires to solve multiple instances of problem PLAIN$\langle l, u\rangle$. Thus, the complexity of the approach presented above depends on the complexity of problem PLAIN$\langle l, u\rangle$.

Problem PLAIN$\langle l, u\rangle$ is a special form of the weighted bipartite matching problem where the number of matches to be changed from a given (perfect) matching is restricted by an upper and lower limit. Although several approaches have been taken to construct a polynomial (or weakly-polynomial) time algorithm for this problem, no such algorithm has been found. Hence, the complexity of problem PLAIN$\langle l, u\rangle$ is not known. It might be *NP*-hard.

Exploiting the Correlation in Frequency

Next, consider a representation of the signaling information such that multiple adjacent sub-carrier assignments regarding the *same* terminal can be summarized in one single signaling representation. One possibility is to add a field to the fixed size signaling field model which indicates the size of such a *bundle* assignment. A bundle is a set of adjacent sub-carriers which are all utilized with the same modulation type and assigned to the same terminal. Following this model, for each down-link phase still the complete signaling information is conveyed but bundles of equivalent, neighboring assignments are summarized in one signaling element (symbol). Then, per assignment the bit cost per signaling symbol is obtained as

$$C_{\text{sig}} = \lceil \log_2(J) \rceil + \lceil \log_2(M) \rceil + \lceil \log_2(A) \rceil \quad . \tag{5.22}$$

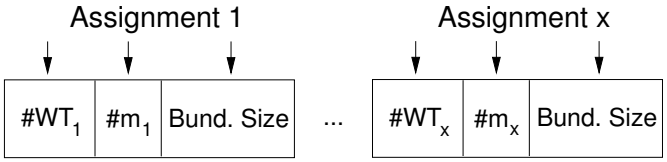


Figure 5.15: Variable size signaling field model: Representation of assignment information for one down-link phase in order to exploit the correlation in frequency.

As mentioned, all assignments are still transmitted. Given the position of a single bundle assignment information (shown in Figure 5.15) and the bundle sizes of all previous assignments in the bit stream, the sub-carrier identification of this assignment can be generated. In this sense, this way of representing the signaling information provides already a compression in the frequency domain if a lot of adjacent sub-carriers are assigned to the same terminal with the same modulation time (thus, if the correlation in frequency is high). In Equation 5.22 A denotes the largest number of sub-carriers which can be combined to one bundle. Instead of deciding on the assignment of single sub-carriers, the dynamic algorithm at the access point decides now on the assignment of bundles instead. This leads to a major change of the optimization problem formulation (even if the signaling cost is not taken into consideration) compared to the basic optimization problem (PLAIN).

Based on the bit matrix $\mathbf{B}^{(t)}$ a new bit matrix $\tilde{\mathbf{B}^{(t)}}$ is built, holding the symbol capacity $b_{j,g}^{(t)}$ of each bundle g regarding each terminal j. The number of all bundles G depends on the maximum size A of sub-carriers a bundle can combine. If $A = 1$, there are $G = N$ bundles. In general, the number of bundles computes to

$$G = \sum_{i=0}^{A-1} (N - i) \quad . \tag{5.23}$$

The assignment matrix $\mathbf{X}^{(t)}$ changes in an equivalent way, holding binary entries on the assignment decision $x_{j,g}^{(t)}$ of bundle g to terminal j at time t.

The number of required signaling symbols is obtained by Equation (5.24). Note that in this case the number of required signaling symbols is always at least one, as all assignments are still signaled. However, it might occur that each terminal is assigned a bundle of the size of its sub-carrier allocation l_j, such that for each terminal only one signaling symbol has to be spent.

$$\varsigma = \left\lceil \frac{\sum_{j,g} C_{\text{sig}} \cdot x_{j,g}^{(t)}}{N \cdot b_{\text{sig}}} \right\rceil \tag{5.24}$$

The resulting optimization problem is given in (5.25). In order to state the original constraints of problem (PLAIN), a "inclusion" matrix has to be constructed with binary entries $i_{n,g}^{(t)}$, indicating

if sub-carrier n is an element of bundle g or not.

$$\begin{aligned} \max \quad & \left(S - \left\lceil \frac{\sum_{j,g} C_{\text{sig}} \cdot x_{j,g}^{(t)}}{N \cdot b_{\text{sig}}} \right\rceil \right) \cdot \sum_{j,g} \left(b_{j,g}^{(t)} \cdot x_{j,g}^{(t)} \right) \\ \text{s. t.} \quad & \sum_{j,g} i_{n,g}^{(t)} \cdot x_{j,g}^{(t)} \leq 1 \qquad \forall\, n \\ & \sum_{n,g} i_{n,g}^{(t)} \cdot x_{j,g}^{(t)} \leq l_j \qquad \forall\, j \ . \end{aligned} \tag{5.25}$$

The choice of the bundles to be considered is of some relevance. A simple design would only build bundles if all adjacent sub-carriers happen to have the same modulation type due to the adaptive modulation. Alternatively, bundles can also be created artificially. In this case adjacent sub-carriers of the same bundle are all utilized with one modulation type. If their attenuation differs strongly, this modulation type has to be adjusted such that the target symbol error probability is even met on the sub-carrier with the highest attenuation. In this case, the choice of A can be obtained from the coherence bandwidth W_{c}.

The resulting maximization problem for the net throughput is a nonlinear problem. In fact, the structure is quite similar to the optimization problem (5.20), therefore this is a quadratic optimization problem. In order to solve this problem, a similar iteration technique as in the case of exploiting the time correlation can be applied. However, the complexity of a bounded instance of this problem, given in the following equation (PLAIN BUNDLE $\langle l, u \rangle$), is not known as well. Hence, it is open if this problem can be solved to optimality in practical systems.

$$\begin{aligned} \max \quad & S \cdot \sum_{j,g} \left(b_{j,g}^{(t)} \cdot x_{j,g}^{(t)} \right) \\ \text{s. t.} \quad & \sum_{j,g} i_{n,g}^{(t)} \cdot x_{j,g}^{(t)} \leq 1 \qquad \forall\, n \\ & \sum_{n,g} i_{n,g}^{(t)} \cdot x_{j,g}^{(t)} \leq l_j \qquad \forall\, j \\ & l \leq \left\lceil \sum_{j,g} \frac{C_{\text{sig}} \cdot x_{j,g}^{(t)}}{N \cdot b_{\text{sig}}} \right\rceil \leq u \ . \end{aligned} \tag{PLAIN BUNDLE $\langle l, u \rangle$}$$

Exploiting the Correlation in Time and Frequency

Finally, it is also possible to develop an optimization model which exploits the correlation in time and frequency simultaneously. In this case the basic signaling symbol is composed of the terminal address, the sub-carrier identifier (at which the bundle starts), the used modulation type as well as the size of the bundle (see Figure 5.16). Therefore, the signaling cost per single bundle assignment is obtained by

$$C_{\text{sig}} = \lceil \log_2(J) \rceil + \lceil \log_2(N) \rceil + \lceil \log_2(M) \rceil + \lceil \log_2(A) \rceil \quad . \tag{5.26}$$

Bundle assignments are only changed if it pays off in terms of the net throughput. If the

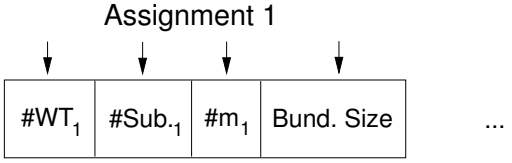


Figure 5.16: Variable size signaling field model: Representation of assignment information for one down-link phase in order to exploit the correlation in frequency as well as in time.

correlation in time is high and the previous assignment provides a sufficient throughput, no signaling information at all is transmitted. On the other hand, if multiple sub-carriers are all newly assigned to some terminal this can be indicated by a relatively few number of signaling symbols. In contrast, if many sub-carriers have to be reassigned and only small sub-carrier blocks can be built, this variable size signaling representation is rather inefficient, compared to the fixed size signaling field model. Per single assignment, this variable size signaling approach clearly requires the largest amount of bits.

$$c_{j,g}^{(t)} = \begin{cases} 0 & \text{if } x_{j,g}^{(t-1)} = 1, \\ C_{\text{sig}} & \text{otherwise.} \end{cases} \tag{5.27}$$

The resulting optimization problem is given in (5.28). It is the combination of problem (5.20) and (5.25). It remains to state that the cost matrix has to be adjusted to bundles rather than sub-carriers, which is given in Equation 5.27.

$$\begin{aligned} \max \quad & \left(S - \left\lceil \frac{\sum_{j,g} c_{j,g}^{(t)} \cdot x_{j,g}^{(t)}}{N \cdot b_{\text{sig}}} \right\rceil \right) \cdot \sum_{j,g} \left(b_{j,g}^{(t)} \cdot x_{j,g}^{(t)} \right) \\ \text{s. t.} \quad & \sum_{j,g} i_{n,g}^{(t)} \cdot x_{j,g}^{(t)} \leq 1 \qquad \forall\, n \\ & \sum_{n,g} i_{n,g}^{(t)} \cdot x_{j,g}^{(t)} \leq l_j \qquad \forall\, j \ . \end{aligned} \tag{5.28}$$

In order to solve this quadratic optimization problem, an adjusted iteration can be applied as in the case of exploiting the correlation in time. However, the complexity of this approach is open, as the solution of a bounded instance equivalent to the combination of
problem (PLAIN BUNDLE $\langle l, u \rangle$) and (PLAIN$\langle l, u \rangle$) is not known.

Protocol Aspects of the Variable Size Signaling Field Approaches

If bit errors occur in the signaling part, for some of the presented optimization approaches more complicated correction schemes are required. Note that the variable size signaling field approach exploiting the frequency correlation does not require any modification compared to the fixed size signaling field model as for each down-link phase all assignments are signaled.

However, if only the assignment differences are signaled (as with the signaling models exploiting the time correlation and the time/frequency correlation) the situation is more complicated. In this case, if a bit error occurs that cannot be corrected by the terminal, it has to discard the transmitted updates of this signaling information. During the next down-link phase, this terminal will not automatically resynchronize with all currently valid assignments, as only the differences are signaled and the loss of the assignment updates of any previous down-link phase discards all further assignment updates. Thus, the terminal with the bit error has to receive the complete signaling information before it can resynchronize. Different approaches can be chosen for this process.

After the bit error has been detected by the terminal, it has to indicate the loss of the current assignment update to the access point during the next up-link phase. Until this up-link phase, all payload data received might be corrupted due to various different consequences of the erroneous signaling update. However, there is still a probability that the bit error occurred without any consequence for the affected terminal, as the bit error changed an assignment update related to a different terminal. Once the access point received the indication of the corruption of the signaling information, it transmits during the next down-link phase all current assignments using the encoding of the fixed size signaling field model (or a compressed version of it). This leads to the resynchronization of the terminal with the erroneous signaling information to the complete assignment information. Alternatively, the access point might also just transmit all current assignments of the specific terminal in addition to the (in any event transmitted) assignment changes during the next down-link phase. In this situation, the terminal with the error in some of its signaling information would at least be able to obtain all current assignments for itself. Still, the assignment information regarding other terminals could be corrupted. However, this would have no impact on the payload data transmission to the specific terminal. Obviously, the second method degrades the overall system performance less than the first method. In general, the impact of any method depends on the error probability regarding the signaling part. A rather strong forward error correction scheme should be used to reduce the amount of bit errors in the signaling part.

Still, a rather low probability exists for a bit error to occur that is not detected. In this case, the bit error might have no consequences if it changes the assignment information regarding a different terminal. If it manipulates the information regarding this specific terminal, such a bit error can only be detected by higher layer error correction and detection schemes. Unfortunately, the higher layer cannot distinguish between regular packet errors and packet errors due to wrong signaling information. One option would be in this case to retransmit all signaling information for some terminal if a higher layer entity was not received correctly after three times. Then, an undetected bit error in the signaling update might have occurred. A different option is to attach the current assignment information to the negative acknowledgement of a higher layer data entity. In this case, the access point could check this assignment information and indicate an undetected error to the affected terminal. This would require that the assignment information is correctly received by the access point. However, note that an undetected bit error in the signaling part occurs extremely seldom if the error correction and detection schemes have been designed well.

5.4.2 Other Reduction Schemes for the Signaling Overhead

Beyond the optimization of the net throughput, the correlation in time and frequency can also be exploited by compression schemes. In Section 5.3 we discussed that signaling information in the fixed size signaling field model could potentially be compressed quite well in certain cases. However, comparing the entropy of an information source and the compression gain achieved by a certain algorithm can yield different results as the entropy is an upper bound on the compression gain in general (if the entropy model is appropriate). Hence, in this section the performance due to applied compression schemes is studied.

Three schemes are considered. The first scheme is the application of the BSTW algorithm to the fixed size signaling field model. The compression exploits the correlation in frequency only. The next scheme is one which exploits the correlation in time. The representation of assignments is similar to the representation used in Section 5.4.1 for the optimization exploiting the correlation in time. The used representation in that section can also be applied as compression scheme. However, in Section 5.3 it was seen that the pure compression in the time domain does not yield compression ratios as good as in the frequency domain. Therefore, in addition to the application of the new representation, the information source is modified to keep as more assignments fixed if the correlation in time is high. This is done by a rather simple modification of the optimization model. In contrast to the optimization models in Section 5.4.1 no additional computational complexity is added. The third scheme considered employs the BSTW compression algorithm in combination with the optimization and compression approach described previously. Thus, this last model exploits the correlation in time and frequency.

Exploiting the Correlation in the Frequency Domain

In this case, the output of the fixed size signaling field approach is compressed, using an algorithm introduced by Bentley, Sleator, Tarjan and Wei [105] – hence the name BSTW. The BSTW scheme has the property that it does not need a dictionary, unlike the Lempel-Ziv compression scheme for example. Apart from that this algorithm presents the additional benefit of taking advantage of *locality of reference*, which is the tendency for a certain source message to occur frequently for a short period of time and then fall into long periods of disuse. This is exactly the mechanisms that can exploit the correlation of assignments in frequency. Trains of the same assignment tuple for adjacent sub-carriers can be compressed. This algorithm uses a self-organizing list as an auxiliary data structure, and employs shorter encodings for words near the front of the list. BSTW uses the move-to-front heuristic as strategy for maintaining the self-organizing list [106].

BSTW works as follows: As in other adaptive schemes, sender and receiver maintain identical representations of the coding. In this case, the code is given by assignment lists which are updated at each transmission, using move-to-front. These lists are initially empty. When an assignment i is transmitted, if i is on the senders list, its current position in the table is transmitted. Then the list is updated by moving i to position 1 and shifting each of the other messages down one position. The receiver similarly alters the word list. If i is being transmitted for the first time, then $k + 1$ is the "position" transmitted, where k is the number of distinct messages transmitted so far. Some representation of the assignment itself must then be transmitted as well, but just

Integer	Elias Code
1	1
2	0100
3	0101
4	01100
5	01101
6	01110
7	01111
8	00100000
16	001010000
17	001010001
32	0011000000

Table 5.8: Elias Codes.

for this first time (for example, the fixed size signaling field approach representation). Again, i is moved to position 1 by both sender and receiver subsequent to its transmission.

The following example shows a group of 10 sub-carrier assignments in the fixed size signaling field approach, and how this information is coded using BSTW:

$$\langle 1-3\rangle - \langle 1-3\rangle - \langle 1-3\rangle - \langle 2-4\rangle - \langle 2-4\rangle - \langle 1-3\rangle - \langle 1-3\rangle - \langle 4-2\rangle - \langle 4-2\rangle - \langle 4-3\rangle$$

$$1\langle 1-3\rangle - 1 - 1 - 2\langle 2-4\rangle - 1 - 2 - 1 - 3\langle 4-2\rangle - 1 - 4\langle 4-3\rangle$$

This example needs 60 bits if the usual fixed size signaling field approach is used, while with BSTW 47 bits are needed. As the example shows, algorithm BSTW transmits each source message only once. The rest of the transmission consists of encodings of list positions. This means that after a couple of down-links only positions will be transmitted. Hence, an essential feature of this algorithm is a reasonable scheme for representation of the integers. The method here used is known as Elias codes [107]. This scheme needs a number of bits to encode an integer n given by Equation 5.29. In Table 5.8, an example of mapping from integer numbers to Elias codewords is given.

$$1 + \lfloor \log_2 n \rfloor + 2 \cdot \lfloor \log_2 (1 + \lfloor \log_2 n \rfloor) \rfloor \quad \text{bits} \tag{5.29}$$

Algorithm BSTW can be easily implemented, and it presents the advantage of being quite simple, hence requiring few computational resources. BSTW presents also an important disadvantage. If an error occurs, sender and receiver will rapidly loose synchronization, since the receiver will update the assignment list wrongly and both instances will maintain from then on different representations of the code. This might make it necessary to use stronger error correction techniques in addition to compression, since correctly decoding the signaling information is crucial for the system's performance. However, errors in the compressed information have not been considered further.

Computational power has to be present at the transmitter as well as at the receiver to employ this scheme, but it is rather low. This is a significant difference to the above presented optimization schemes.

Exploiting the Correlation in the Time Domain

As in the frequency domain, a compression scheme could be used to reduce the signaling overhead by exploiting the correlation of assignments in time. However, it has been shown (see Section 5.3) that the correlation in time of the assignments generated by the fixed size signaling approach is rather low. This is rather surprising as the correlation of the sub-carrier attenuations in time is equally strong as the correlation in frequency. However, the correlation of the assignments is much stronger in frequency than in time.

As result, a different approach is taken to exploit the correlation in time. First of all, the information source (the generation of assignments) is modified such that a stronger correlation in time of the assignments is obtained. This is done by solving a slightly modified optimization problem, compared to the optimization approach of the fixed size signaling field model.

Given the assignment matrix of the previous down-link phase $\mathbf{X}^{(t-1)}$, the bit matrix is modified by a cost term $c_{j,n}^{(t)}$. This value is zero if terminal j was assigned to sub-carrier n during the previous down-link phase, hence if $x_{j,n}^{(t-1)} = 1$. Otherwise (in case of $x_{j,n}^{(t-1)} = 0$) the cost term is set to some value $c_{j,n}^{(t)} = C_{\text{sig}}$. The resulting bit matrix entry is given by the values $b_{j,n}^{(t)} - c_{j,n}^{(t)}$, which is lower if this sub-carrier /terminal was not assigned in the previous down-link phase. This yields the optimization problem:

$$\begin{aligned} \max \quad & \sum_{j,n} \left(S \cdot b_{j,n}^{(t)} - c_{j,n}^{(t)} \right) \cdot x_{j,n}^{(t)} \\ \text{s. t.} \quad & \sum_{j} x_{j,n}^{(t)} \leq 1 && \forall\, n \\ & \sum_{n} x_{j,n}^{(t)} \leq l_j && \forall\, j \ . \end{aligned} \tag{5.30}$$

Note that (5.30) is equivalent to (5.6) with the difference that assignment coefficients are adapted to reflect the cost of signaling. The resulting system throughput of (5.30) is not the net throughput, since this again has to be computed from the assignments and the number of signaling symbols ς required. However, by choosing C_{sig} appropriately, assignments are much more correlated in time than observed for the fixed size signaling field approach.

In addition, the resulting assignment information is represented differently. An equivalent representation as with the optimization model exploiting the correlation in time is applied. Assignments are only signaled if they change from one down-link phase to the next one. However, this requires to transmit an information triple for each assignment, consisting of

$$\langle \text{sub-carrier identification}, \text{terminal address}, \text{modulation identification} \rangle \ ,$$

which requires

$$\lceil \log_2(N) \rceil + \lceil \log_2(J) \rceil + \lceil \log_2(M) \rceil$$

bits per assignment. The combination of this representation and the modified optimization approach of (5.30) is studied as "compression" scheme exploiting the correlation in time. Note that this scheme does not require additional computational resources at the access point (compared to applying the fixed size signaling field approach) as well as no additional resources at the receivers.

Exploiting the Correlation in the Time and Frequency Domain

Both methods described above can be combined. Thus, first the assignments are generated according to the modified optimization problem in Equation 5.30. Next, the assignments are represented according to the coding presented above. Finally, these assignments are taken and compressed in frequency by the BSTW algorithm. The resulting scheme exploits the correlation in time and frequency at a much lower complexity than the optimization schemes presented in Section 5.4.1. However, this scheme requires small additional resources at the transmitter (access point) as well as at the receivers (terminals) due to the application of the BSTW algorithm.

5.4.3 Performance Evaluation

The presented reduction schemes are evaluated for the scenarios that have been introduced in Section 5.3.2. Four different scenarios are considered, two for which the correlation in time varies due to a change of the speed in the propagation environment and, for a different scenario, in which the frame length varies. In two other scenarios, the correlation of the attenuation values varies due to a variation of the delay spread and due to an increasing number of sub-carriers. For details, refer to Section 5.3.2.

Methodology

In the case of the optimization approaches the same general methodology as in Section 5.2 was applied. A channel trace file was read by a script and transformed into attenuation matrices. These were then used to as input for the optimization problems. However, the optimization problems themselves (problem (5.20), problem (5.25) and problem (5.28)) could not be solved as such by the software used in Section 5.2.3, as all three are quadratic optimization problems. Instead, the problems were solved by the iterative approaches presented with each of the corresponding optimization problems in Section 5.4.1. These iterative solution algorithms are based on linear, integer programming problems which could be solved by the standard software.

Varying the Frequency Ratio R_{f}

Initially, consider the results for the optimization approaches and the comparison scheme regarding a varying frequency ratio (as considered in Section 5.3.2). The frequency ratio was varied for two different scenarios: Varying the number of sub-carriers (between $N = 64$ and $N = 512$) and varying the delay spread (between $\Delta\sigma = 0.1$ μs and $\Delta\sigma = 0.5$ μs). All other system parameters are set to the corresponding values of Section 5.3.2.

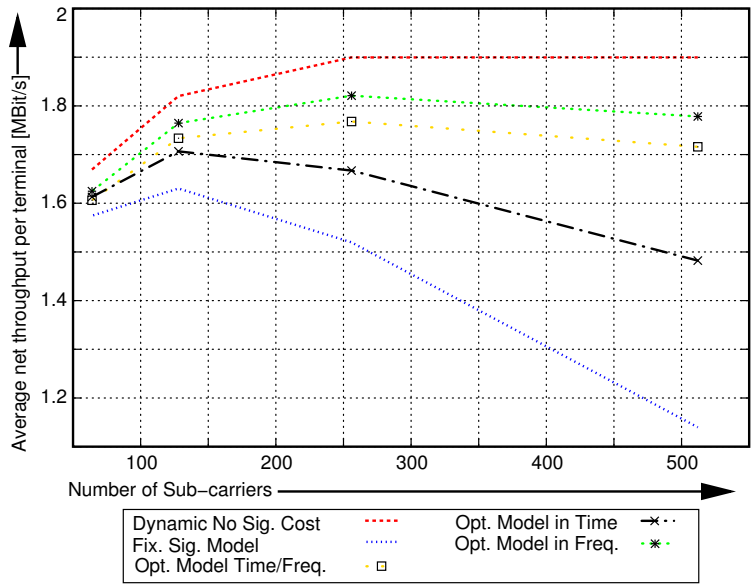
In Figure 5.17 the net throughput is shown for a varying number of sub-carriers (upper graph : optimization approaches – lower graph : comparison schemes). Increasing the number of sub-carriers leads to a stronger correlation of sub-carriers as the sub-carrier spacing is reduced. As the number of sub-carriers increases, the optimization approaches outperform the fixed size signaling field model significantly. Even for the smallest number of sub-carriers the performance advantage is already about 50 KBit/s. For a sub-carrier number of $N = 512$ the net throughput of the best optimization approach is only about 120 KBit/s lower than the dynamic approach without signaling cost.

Out of the three different optimization approaches, the one exploiting the correlation in frequency performs best. The scheme exploiting the correlation in time and frequency is a bit lower in terms of net throughput for all sub-carrier numbers considered. In contrast, the scheme exploiting only the correlation in time loses quite a bit of performance, especially as the number of sub-carriers is large. Still, this scheme outperforms the fixed size signaling field approach.

Considering the net throughput results of the compression scheme, observe that almost the same performance can be achieved. However, while for the optimization approaches the one exploiting the correlation in frequency performed best, regarding the compression schemes the one exploiting the time and frequency correlation achieves the highest performance. Compared to the fixed size signaling field model all compression schemes achieve a higher net throughput. Comparing the two approaches (optimization and compression scheme) which exploit the correlation in time, the optimization approach outperforms the other one. Regarding the frequency correlation only, also the optimization approach outperforms the compression scheme (the BSTW compression scheme). However, regarding the combination of time and frequency the comparison scheme outperforms the optimization. Note that all comparison scheme require a significantly lower computational power at the access point than in case of the optimization schemes.

In Figure 5.18, the achieved net performance of the optimization approaches is studied in detail. The upper graph shows the gross throughput for all three optimization approaches. Observe that the optimization approach exploiting only the frequency achieves a similar gross throughput as the fixed size signaling field model. On the other hand, the other two optimization approaches achieve a significantly smaller gross throughput. Especially the optimization approach exploiting only the correlation in time has a gross throughput much lower than the one of the other three schemes. As the number of sub-carriers increases, the gross throughput increases for all schemes except for the optimization approach exploiting the correlation in time. While the other schemes benefit from an increasing throughput (due to the invreasing number of sub-carriers), the optimization approach in time has to decrease the gross throughput as the signaling cost per reassignment increases (as the number of sub-carriers increases).

The lower graph of Figure 5.18 shows the corresponding percentage of symbols per downlink phase required for transmitting the signaling information. It can be seen that all three optimization approaches have a quite low signaling symbol percentage (always below 8%). For all three schemes, the signaling percentages increases as the number of sub-carriers increases. Hence, an increasing correlation of the sub-carriers cannot be exploited efficiently by a scheme like the optimization approach in frequency, as the signaling percentages increases even for this scheme beyond $N = 128$. For the other two schemes, the increasing signaling percentage is



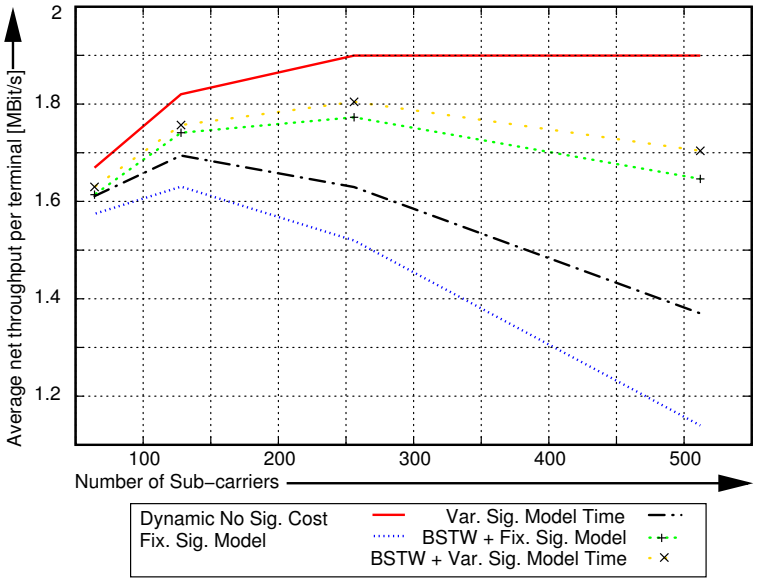


Figure 5.17: Net throughput results for all overhead reduction schemes (Upper graph: Optimization Approaches – Lower graph: Comparison Schemes) while varying the number of subcarriers for a fixed bandwidth.

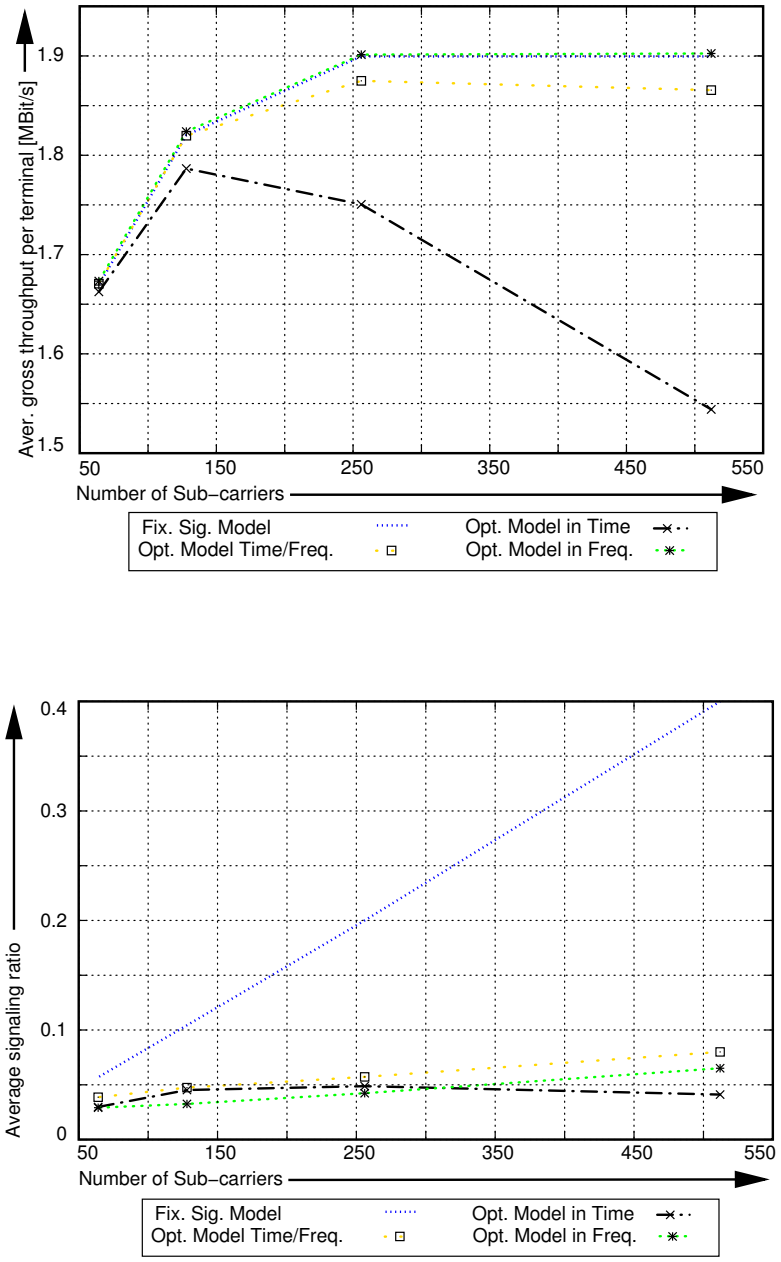


Figure 5.18: Gross throughput (upper graph) and average signaling percentage (lower graph) for the optimization approaches while the number of sub-carriers increases for a fixed bandwidth.

not very surprising as the cost per reassignment (in bits) increases significantly for an increasing number of sub-carriers.

Next, consider the net throughput for all optimization approaches and comparison schemes as the delay spread increases. An increasing delay spread reduces the correlation between adjacent sub-carriers. Therefore, the attenuation diversity in the system increases which leads to a larger throughput in general for any OFDM system.

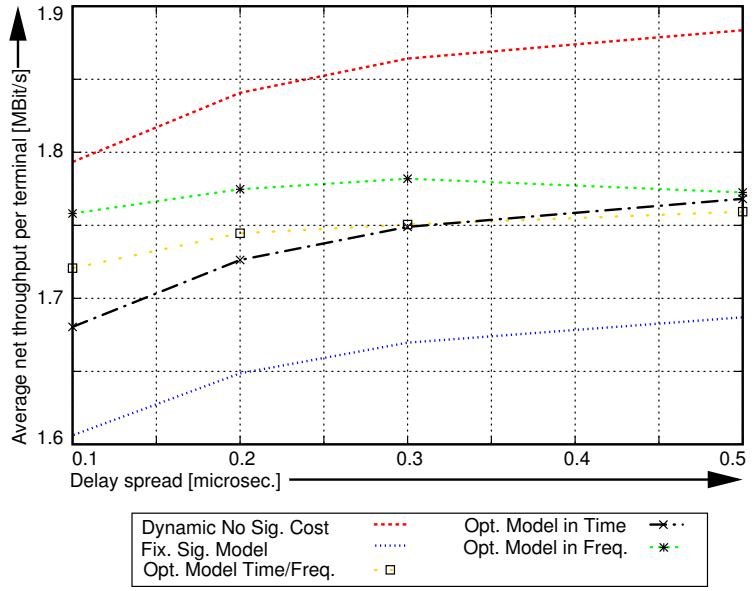
In Figure 5.19, the net throughput is shown for the optimization approaches (upper graph) as well as for the comparison schemes (lower graph). As the delay spread increases, the net throughput increases for some scheme but decreases for other schemes. It increases for the fixed size signaling field approach as well as for the optimization approach which only exploits the correlation in time. However, the net throughput decreases the optimization approach which exploits the correlation in frequency. For the third optimization approach the net throughput increases first and stays constant thereafter. For a delay spread of $\Delta\sigma = 0.5$ μs the net throughput is basically the same for all three optimization approaches. Thereafter, it can be expected that the optimization in frequency achieves a significantly lower net throughput as the correlation of the attenuation values decreases and hence, not enough bundles can be assigned for different terminals in order to reduce the signaling overhead.

As the delay spread varies, the comparison schemes achieve in fact a higher performance than the optimization approaches. In particular the comparison scheme exploiting the time and frequency correlation achieves for higher delay spread values the highest net throughput. For all three comparison schemes, the net throughput increases constantly as the delay spread increases. Again, the optimization approach in time outperforms the comparison scheme in time while for the other two combinations the optimization approaches outperform their counterparts as long as the delay spread is small. If the delay is large, the comparison schemes outperform the optimization approaches. The net throughput performance can again be explained by considering the gross throughput and the signaling percentage. While the gross throughput (not shown in a figure) does not differ strongly for all three optimization approaches, the signaling percentage increases significantly for the optimization approach in frequency as the delay spread increases. This is shown in Figure 5.20. Note in particular that the signaling percentage of the optimization approach in time stays constant and the percentage of the optimization approach in time and frequency increases only slightly. Thus, as the correlation in frequency becomes weaker, the optimization approach in time and frequency can rely more on exploiting the correlation in time while the optimization approach in frequency suffers from an steadily increasing signaling percentage.

Varying the Time Ratio R_{t}

The two scenarios for varying the time ratio R_{t} are given by varying the frame length and varying the maximum speed within the propagation environment. As in Section 5.3.2, the frame length was varied between $T_{\mathrm{f}} = 3$ ms and $T_{\mathrm{f}} = 0.37$ ms while the maximum speed was varied between $v_{\max} = 0.5$ m/s and $v_{\max} = 2.5$ m/s.

First, consider the variation of the frame length. As the frame length decreases, the correlation in time of the sub-carrier attenuations increases. However, the total number of symbols



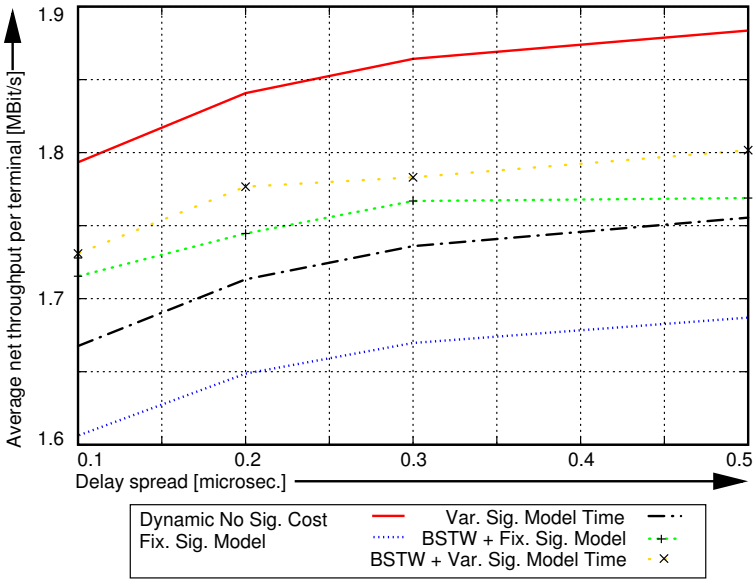


Figure 5.19: Net throughput results for all overhead reduction schemes (Upper graph: Optimization approaches – Lower graph: Comparison schemes) while varying the delay spread of the propagation environment.

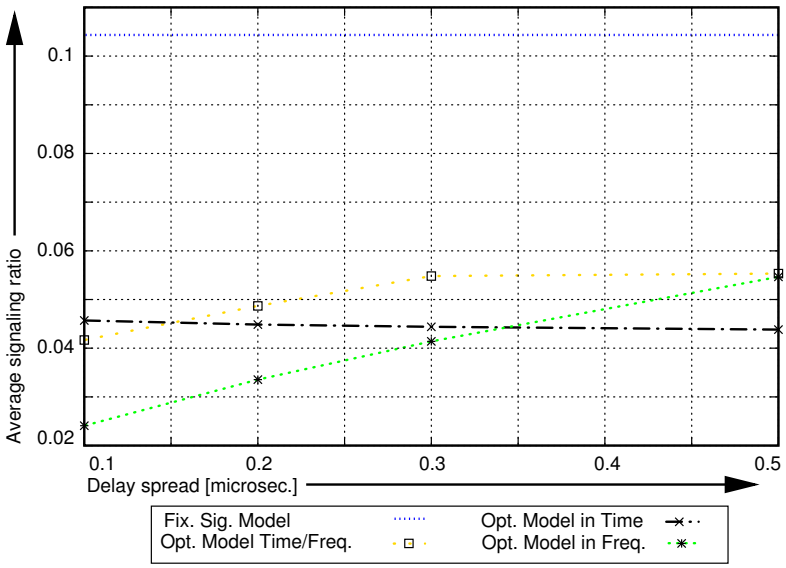


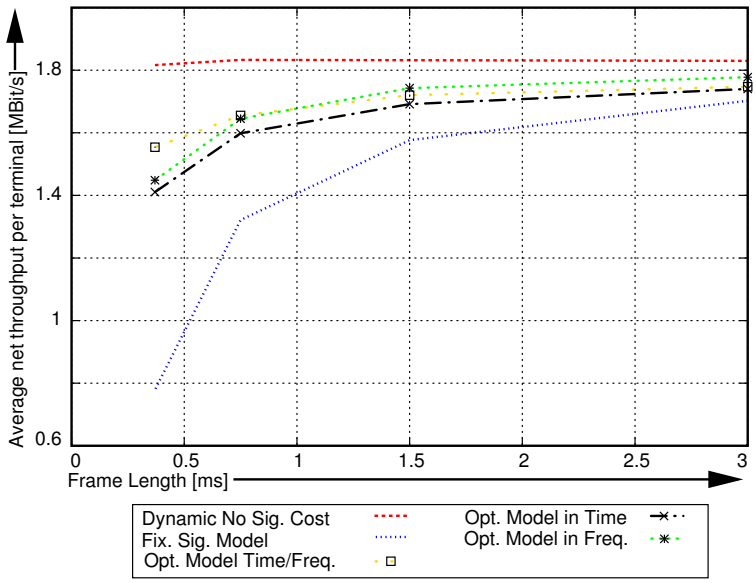
Figure 5.20: Average signaling percentage for the optimization approaches while varying the delay spread of the propagation environment.

per down-link phase decreases as well. Thus, requiring an absolute amount of symbols for the signaling information has a decreasing impact on the net throughput.

In Figure 5.21, the net throughput results are given for the optimization approaches (upper graph) and the comparison schemes (lower graph). As the frame length decreases, all optimization approaches have a decreasing net throughput. However, compared to the fixed size signaling field model the optimization approach can achieve a significantly larger performance, especially as the frame length is rather short. In this case the optimization approach in time and frequency outperforms the other two approaches. Again, the optimization approach in time provides the lowest net throughput, slightly outperformed by the optimization approach in frequency. Compared to the dynamic approach without signaling cost, the largest performance difference is about $100\ \mathrm{KBit/s}$ which corresponds a 6% performance loss.

The comparison schemes achieve a comparable performance. The comparison scheme in time and frequency yields the highest net throughput which is equal to the one of the optimization approach in time and frequency. However, the other two comparison schemes do not provide a comparable performance to their optimization counterparts, especially for a short frame length these schemes are significantly outperformed by the optimization approaches. Note that for the shortest frame length the comparison scheme in time outperforms the comparison scheme in frequency (compression of the fixed size signaling field model).

As the frame length decreases, the gross throughput of the different optimization approaches changes quite interestingly. In Figure 5.22 the gross throughput is shown for all three optimization approach as well as for the fixed size signaling field approach. While for the fixed size signaling field approach and the optimization approach in frequency the gross throughput decreases



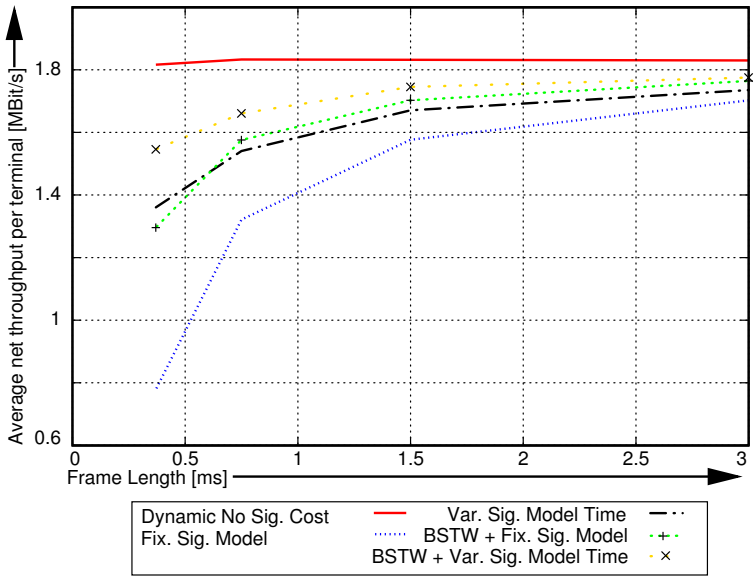


Figure 5.21: Net throughput results for all overhead reduction schemes (Upper graph: Optimization approaches – Lower graph: Comparison schemes) while varying the frame length.

only slightly, it is much lower (especially for short frame lengths) for the other two optimization approaches, which can exploit the correlation in time. However, as the gross throughput decreases for the two optimization approaches in time, the signaling percentage (Figure 5.22 lower graph) also decreases strongly, while it increases for the fixed size signaling field model and the optimization approach in frequency. Finally, in Figure 5.23 the net throughput is shown for the optimization approaches (upper graph) and the comparison schemes (lower graph) as the maximum speed of the propagation environment increases. An increasing speed leads to a weaker correlation of the sub-carrier attenuations in time.

As the speed increases, both optimization approaches in time are outperformed by the optimization approach in frequency. Even for the smallest speed value, the optimization approach in frequency provides a better net throughput. For higher speeds, the optimization approach in time provides almost the same net throughput as the fixed size signaling field approach. The comparison schemes provide a slightly higher net throughput, at least in case of the comparison scheme which exploits the correlation in time and frequency. As the speed increases, the comparison scheme in time has a decreasing net throughput while the other two schemes have an almost steady net throughput.

5.4.4 Summary

As indicated by the results of Section 5.3, the performance of the fixed size signaling scheme can be significantly improved. The improvement is always based on exploiting the correlation of sub-carriers and of assignments. Two basic approaches can be taken: Optimizing the net throughput and compressing the signaling overhead.

Optimization of the net throughput is done by considering a variable size signaling representation. Depending on the correlation to be exploited, a basic representation for a block assignment (either in time or frequency) can be found and a bit amount is obtained for such a representation. Then, an optimization problem is formulated for which the assignment number is minimized while also taking into account the achieved gross throughput. As a result, a maximum net throughput with respect to the chosen assignment representation is obtained. However, compared to the fixed size signaling field model the computational complexity of the resulting optimization problem increases. In fact, the computational complexity is not clear for all three optimization models considered.

In contrast, the compression schemes investigated are quite simple to implement. For the frequency domain, the application of a well known compression algorithm (the BSTW algorithm) is investigated. In order to exploit the correlation in time, a slightly modified optimization problem of the fixed size signaling field model is proposed in combination with a suitable representation of assignments which are change from one down-link phase to the next. The two schemes (the compression and the slightly modified optimization approach) can also be combined, exploiting the correlation in time and frequency.

Interestingly, the performance of these two highly different methods is almost the same. For all investigated scenarios the fixed size signaling field model could be outperformed significantly by an optimization approach as well as by a comparison scheme. Especially in settings where the signaling burden had decreased the performance of the fixed size signaling field approach well below the throughput of the static scheme, the reduction schemes discussed here superior

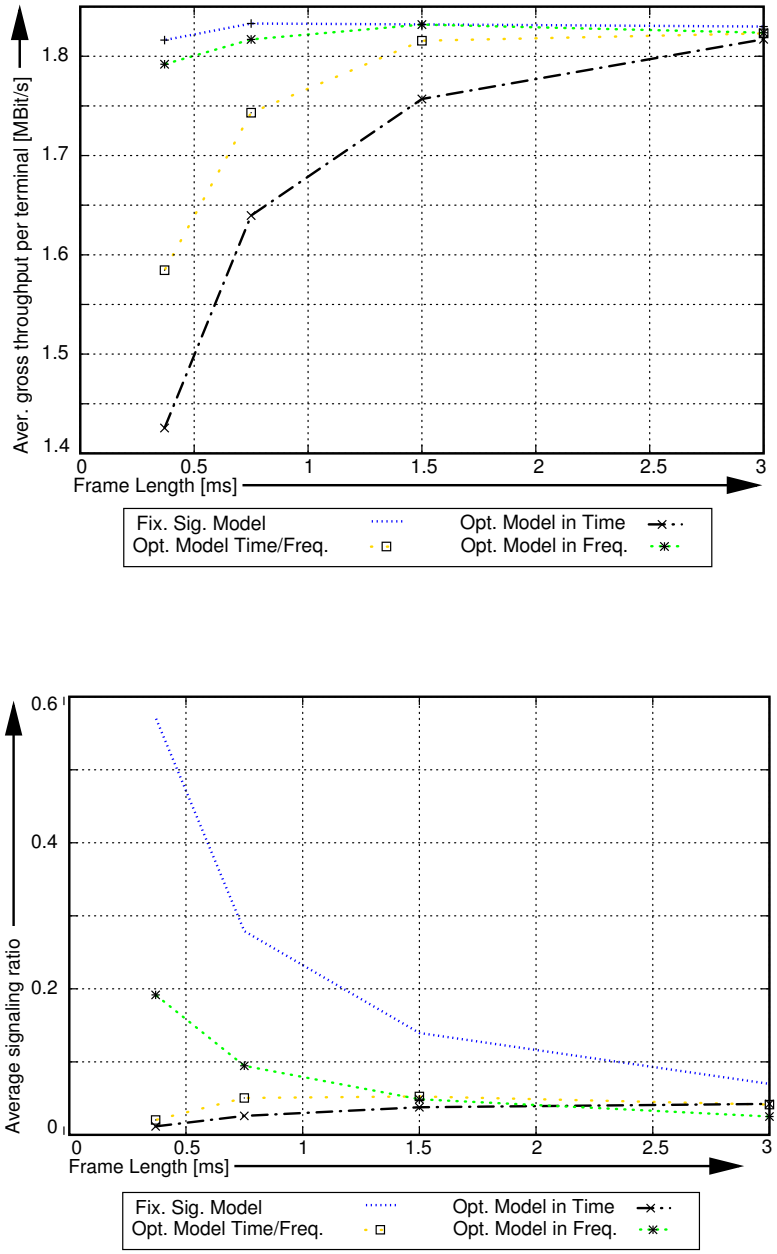
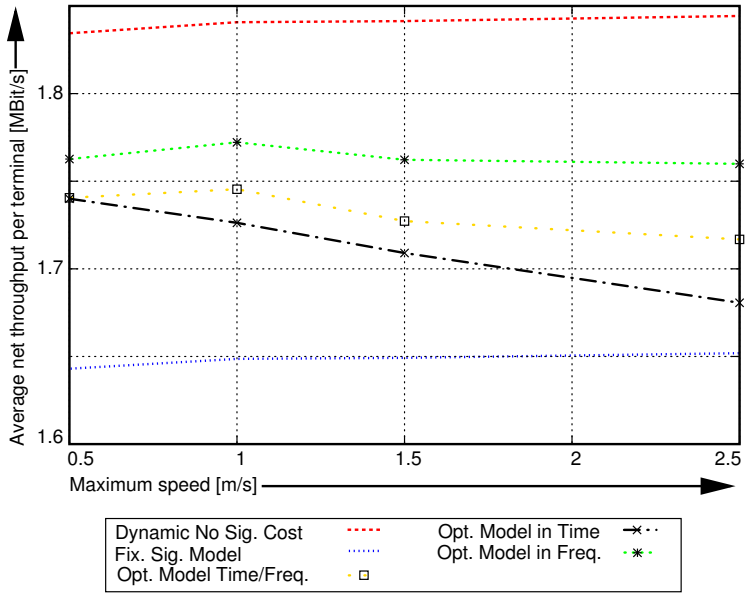


Figure 5.22: Gross throughput (upper graph) and average signaling percentage (lower graph) for a varying frame length.



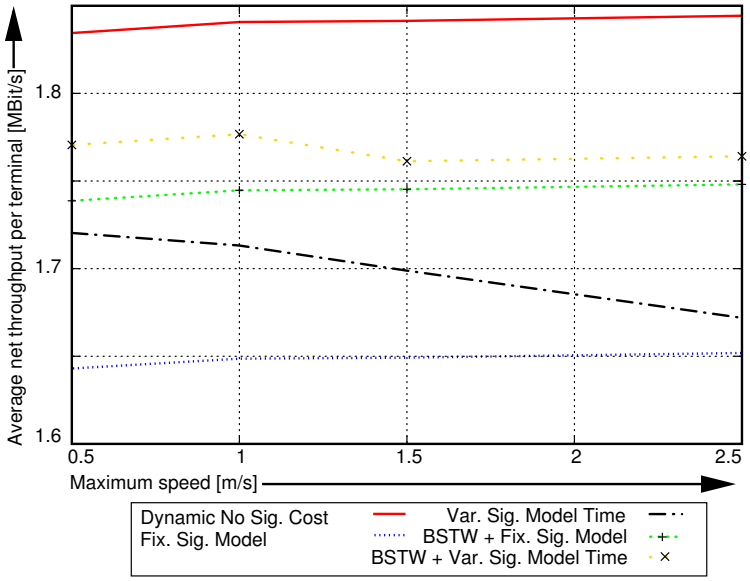


Figure 5.23: Net throughput results for all overhead reduction schemes (Upper graph: Optimization approaches – Lower Graph: Comparison schemes) while varying the maximum speed within the propagation environment.

performance results. Thus, even in such "harsh" conditions a dynamic MU-OFDM approach can be applied as efficient schemes are available reducing the caused signaling overhead.

However, as the optimization approaches do require much higher computational resources at the access point, the compression schemes are more likely to be applied. Still, compression schemes are more vulnerable to bit errors occurring in the transmission of the signaling information. But even if a stronger error correction scheme is applied leading to a slight decrease of the net throughput for the compression schemes, the advantage in terms of computational power required is that large that the compression scheme have to be chosen over the optimization approaches. Note that for an implementation generation times of the assignments below $1\ \mathrm{ms}$ are likely to be required.

5.5 Conclusions

The issue of signaling is not a particular problem of dynamic MU-OFDM systems. In general, the loss due to signaling is rather low (about 5% of the gross throughput). Compared to the performance of static comparison schemes, a significant performance advantage remains even if the signaling overhead is taken into account.

However, the behavior of dynamic MU-OFDM systems changes qualitatively (as well as quantitatively). Most importantly, for any scenario setting (bandwidth, transmit power, number of terminals, number of modulation types, chosen signaling scheme, coherence in time and frequency) an optimal number of sub-carriers exists for which the system performance is optimal (in terms of net throughput). This implies a system design in which a central controller can switch between various sub-carrier numbers depending on the current scenario.

In addition, in certain situations the signaling overhead can lead to a very high loss if no countermeasures are taken. This is observed especially if the system bandwidth is rather small, the number of sub-carriers is high and the frame length is short. In such situations though, the correlation of sub-carrier attenuations is always remarkably high. This can be exploited by different schemes reducing the signaling overhead drastically. As a result, even in these situations dynamic MU-OFDM systems provide a much higher performance than static schemes. On the other hand, additional complexity has to be spent either at the transmitter (access point) or at both, the transmitter and receiver. This might result in a higher cost of the system as well as in a higher power consumption.

As future work it remains to investigate the computational complexity of the three optimization problems, which exploit the correlation. In particular, the complexity of
Problem (PLAIN$\langle l, u\rangle$) (which then yields the complexity of Problem (5.20)) is of interest. Furthermore, also the complexity of problem (5.25) and of problem (5.28) should be obtained, as these two approaches provide a better performance than the solution to Problem (PLAIN$\langle l, u\rangle$). If these problems turn out to be difficult, suboptimal algorithms should be considered, as the compression approach always requires additional computational power at the terminal as well. Finally, an adaptive coding scheme for the signaling information can further reduce the overhead. For example, if all terminals are close to the access point, a weaker coding scheme can be applied which increases the net throughput. The choice of the "best" code depends on the terminal attenuations and could be chosen dynamically.

Chapter 6

Proposal for an IEEE 802.11 Extension Based on Dynamic OFDM

In this chapter an outline of a proposal is presented for a new standard within the family of IEEE 802.11, incorporating a dynamic OFDM approach. Stations and access points modified according to this proposal benefit from a higher throughput. The proposal is based on the IEEE 802.11a OFDM standard and provides full backward compatibility to legacy IEEE 802.11a systems. In the case of a transmission from the access point to stations which support the new proposal, the system performance is enhanced by dynamic power, sub-carrier and modulation assignments. The access point can switch to a "multi-user" mode during which multiple stations receive packets in parallel according to a dynamic MU-OFDM scheme. In case of an up-link data transmission (one station transmits to the access point) the link performance is still improved by dynamic power and modulation assignments (bit and power loading), if the transmitter supports the new proposal. Note, that in this section only infrastructure BSS systems (as described below) are considered. However, the new concept can also be extended to the "ad-hoc" mode of IEEE 802.11. Below, first an overview of the IEEE 802.11a system is given (Section 6.1). Then, the new proposal is presented in Section 6.2. Finally, the performance is evaluated in Section 6.3.

6.1 Overview of the IEEE 802.11 Standard

Most of the discussion in here follows the standard as published in [108].Today, the IEEE 802.11 standard describes a medium access control sublayer (MAC), MAC management protocols and services, and multiple different physical layers (PHY) for wireless local area networking. The original version of 1997 included three different PHYs, namely an infrared based PHY, a frequency hopping spread spectrum PHY in the 2.4 GHz band, and a direct sequence spread spectrum PHY in the 2.4 GHz band, all of them supporting two transmission rates, 1 Mbit/s and 2 Mbit/s. Meanwhile, several PHY extensions have been amended: 802.11b increases the transmission rate in the 2.4 GHz band to up to 11 Mbit/s using advanced modulation and coding schemes, i.e.: Complementary code keying modulation and direct sequence spread spectrum packet binary convolutional coding; 802.11a provides PHY rates of up to 54 Mbit/s using the

5 GHz band and orthogonal frequency division multiplexing (OFDM).

An IEEE 802.11 network consists in general of (wireless) stations, which comprise of a MAC and a PHY entity according to the standard.

6.1.1 Architecture and Services

There are several ways of organizing the stations as network in IEEE 802.11:

- *Independent BSS* : This is an organization form where each station can directly communicate with each other one. No access point is present. An IBSS is mostly a network which is built sporadic: Users wish to exchange data quickly between each other, for example, but do not intend to do this constantly at this place. Services which require the connection to a larger network (the Internet) are mostly not of interest. All stations communicate directly with each other, i.e. there is no distinguished station that performs relaying services for other stations. Thus, an IBSS is a rather short-lived network architecture, also referred to as an "ad-hoc" network.

- *Infrastructure **B**asic **S**ervice **S**et (BSS)* : An infrastructure BSS [1] always contains an access point. The access point provides distribution services to the stations in the BSS. All communication within this BSS is relayed via the access point. Even if one station of the BSS has data to be sent to another station in the same BSS, the data is first sent to the access point and then to the specific station.

- *Extended Service Set (ESS)*: In addition to these two forms of a BSS, IEEE 802.11 also provides an architecture to support mobility of stations. In this case the network obviously has to consist of multiple BSS. The access points of each BSS communicate with each other through an own communication channel (referred to as distribution system), which is still part of the link layer. The main purpose of the ESS architecture is to hide the mobility of stations within the ESS to network protocols outside of the ESS (higher layer protocols with no support of mobility, for example).

There are nine services specified by IEEE 802.11: Four services are related to stations and five to access points. The four services regarding stations are authentication, deauthentication, privacy and data delivery.

The authentication service provides means to prove the identity of a certain station to another one. In contrast, the deauthentication service removes a (previously authorized) station from the list of authorized stations. The privacy service protects data while being conveyed through the wireless medium. It is designed to provide a level of privacy granted by a wired network with restricted access to it. It does not provide full privacy for higher layer protocols, for example, between active applications on each station of a BSS. The data delivery service provides the same MAC functionality given by any 802 LAN standard. The main goal of this service is to provide reliable data delivery of frames from the MAC in one station to another stations MAC or even to multiple stations MAC.

[1] If in the following only a BSS is mentioned, this refers to an infrastructure BSS

The five services of of the access point are association, reassociation, disassociation, distribution and integration. They are also called distribution services, as they are all related to the functionality of a BSS or an ESS – allowing stations to roam freely within a connected collection of access points. Association, reassociation and disassociation are all related to the maintenance of a logical connection between stations and the access points. In a BSS or ESS each station always has to be logically connected to one access point. Thus, the association, reassociation and disassociation services provide the specific means for this administration of roaming stations. The distribution is invoked by an access point if it receives a MAC frame to be forwarded to another station. In this case, the access point either transmits the frame directly to the station (if the station is associated to this access point) or it forwards the frame on the distribution system. Finally, the integration service provides means to connect the IEEE 802.11 WLAN to another LAN. Basically, the integration service translates IEEE 802.11 frames into other frame formats and allows thus the forwarding of the IEEE 802.11 frame to other networks. In addition, other frame types are also translated into the IEEE 802.11 frame format.

A simple state machine describes the invocation of the required services to join a BSS or an ESS. After a station has been powered on (or the network adapter has been plugged to the computer), the station is unauthenticated and unassociated. Thus, the station has to authenticate itself to the access point. This is done by the authentication service described above. Only a restricted set of frames is allowed to be used by such a station. After the station is authenticated, it has to associate with one (and only one) access point. This is done by the association service, based on other specific frames. After that the station is authenticated and associated and can use the data distribution service of the BSS. In contrast, in the IBSS mode (the "ad-hoc" mode) the data distribution service is immediately available to all stations. Even no authentication is required.

6.1.2 Medium Access Control Layer

Three functionalities are provided by the IEEE 802.11 MAC. The first one is reliable data communication. As the underlying transmission medium is particularly unreliable, this functionality is quite important. A specific frame exchange protocol grants this function. The second issue is the access of the shared wireless medium. Regarding this issue, the fairness of the access is quite important, giving each associated station the same chance to acquire the transmission medium. Two access schemes are part of the standard: A centralized mode called the **P**oint **C**oordination **F**unction (*PCF*) and a distributed mode, called the **D**istributed **C**oordination **F**unction (*DCF*) . The DCF scheme is mandatory in all organization forms, while the PCF is only applied in infrastructure BSS networks. However, even in those networks, the DCF is applied far more often than the PCF. The third function is to protect the data conveyed within the 802.11 network. This function is granted by a privacy service, referred to as wired equivalent privacy (WEP) which encrypts the data transmitted over the wireless medium using a (within the network) commonly known key.

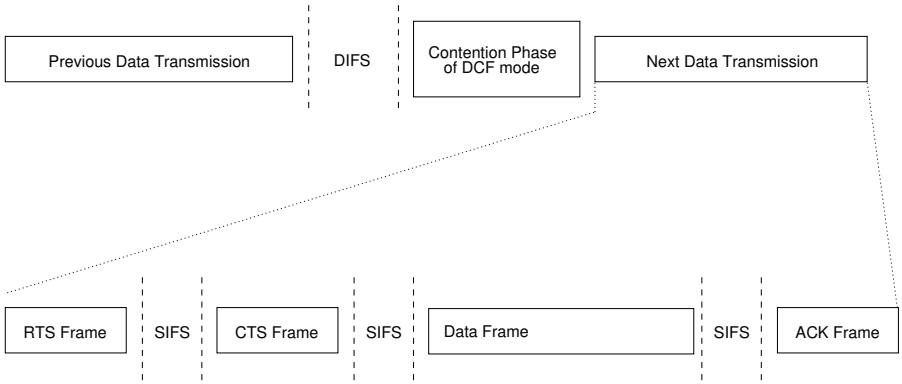


Figure 6.1: DCF mode – medium access and atomic unit of data transmission

Medium Access and Operation Modes

Once a station is authenticated and associated, it might have data to be transmitted. In order to do so, the IEEE 802.11 MAC layer has to acquire the channel. In an infrastructure BSS network the station can subscribe to a poll list (if the network is organized by the PCF) or is has to try to acquire the channel directly (if the network is organized by the DCF). In contrast, in an IBSS network the station always has to acquire the channel itself. In case of the DCF (see Figure 6.1), the channel acquisition is performed by the **C**arrier **S**ense **M**ultiple **A**ccess with **C**ollision **A**voidance (*CSMA/CA*) scheme with an exponential backoff. A station holding data to be transmitted checks the wireless channel for an ongoing data transmission. If the channel is currently not in use, the station may enter the contention phase (Figure 6.1). Then, each such station generates a random deferral period. If the channel is still not in use after this deferral period, the station may start a data transmission. If the channel is still in use after the deferral period, the station increments a retry counter (which rules the life time of a frame to be transmitted) and generates a backoff duration (randomly chosen) after which the channel is sensed again. Eventually, the medium is observed unused and the frame transmission starts or the retry counter reaches its maximum value such that the frame is discarded. After the medium has been in use, all stations are required to refrain from a medium access for the duration of a distributed interframe space (DIFS). This time span takes different values depending on the PHY. Then, all stations actually wishing to transmit data again generate their deferral periods and enter the contention phase.

In addition to this physical carrier sensing, 802.11 introduces a virtual carrier sensing mechanism: The network allocation vector (NAV). The NAV is a time period in which the wireless medium must be treated as busy even if the physical carrier sensing does not indicate this situation. The NAV is set according to the duration field found in the MAC header and usually indicates the remaining time until the ongoing transmission (sequence) is finished. In the DCF mode, the NAV is updated every time a frame is transmitted.

In case of an infrastructure BSS network, the channel access might also follow the PCF mode. In this mode, the access point transmits data currently queued to multiple terminals

sequentially, while stations can also be polled to transmit data to the access point. Shorter timing intervals are used in this mode to assure that all other stations cannot take over control of the channel as long as the access points has data to be transmitted.

Data Transmission

A regular data transmission within an IEEE 802.11 network in the DCF mode consists usually of four exchanged frames. Each frame is separated by a time span of a short interframe space (SIFS), which is also defined depending on the PHY.

Initially, the station which acquired the channel transmits a Ready-to-Send (RTS) frame to the station that it wants to transmit data to. If the RTS frame was received successfully, the receiving station transmits a Clear-to-Send (CTS) frame back. This RTS/CTS scheme is a common method to overcome problems arising from the hidden terminal problem. Note that the RTS/CTS handshake is not mandatory, however, today almost all products apply the RTS/CTS handshake.

As both stations involved in the transmission indicate that a data transmission will occur immediately, all stations in their environment most likely receive at least one of the exchanged frames and do not start transmission of data themselves (which would interfere with the other data transmission). Once the RTS/CTS handshake was successful, data is transmitted. As the wireless medium is rather unreliable, any data transmission has to be acknowledged by the receiving station. This fourth message finalizes the successful transmission. If any of these four frames is corrupted or not received at all, the data transmission is supposed to have failed and the transmitting station tries to acquire the channel again if the retry counter allows another attempt.

In the PCF mode, the RTS/CTS handshake is not required as it is assumed that all stations associated to the access point of the BSS receive frames from it. Instead, the access point immediately starts the transmission of a frame to a certain station. As stations receiving data from the access point might also have data to be transmitted to the access point, the exchange of certain frames can be combined. For example, payload transmission and the poll of a station can be transmitted in one frame rather than using two. The access point might also combine the transmission of an acknowledgement with a poll. Also, the station can transmit data back to the access point together with the acknowledgement of the previous data transmission.

Frame Types and Formats

There are three different subtypes of frames exchanged between stations of an IEEE 802.11 network: Control frames, data frames and management frames. Control frames are, among others, mainly acknowledgements and RTS or CTS indications. Data frames always carry the payload. Management frames are used to execute the authentication and association process. In addition, certain frames are used to broadcast the existence of the BSS (referred to as Beacon frames). All these frames follow a fixed general format (see Figure 6.2) with a MAC header and trailer. These two form the MAC protocol data unit (MPDU), which might also contain payload data, referred to as MAC service data unit (MSDU). The MAC header of any frame includes a Frame Control field, a Duration field, an Address field and a Sequence Control field. The Frame Control field contains the following sub-fields: Protocol version, frame type and subtype,

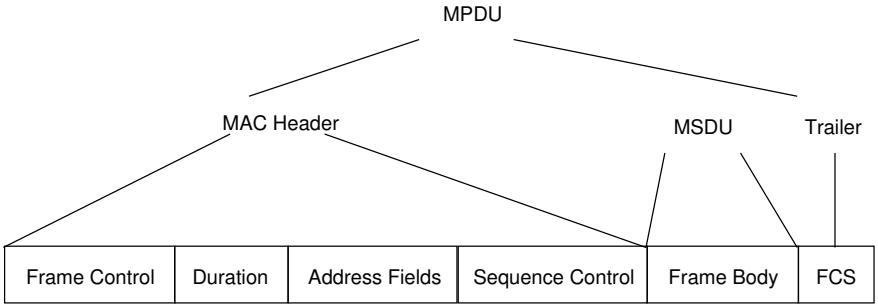


Figure 6.2: General structure of a MAC frame

fragmentation information, privacy information and queue information of the transmitter. The other three fields (Duration, Address and Sequence Control) do not feature sub-fields. The complete MAC header is at most 30 Bytes long. The trailer only contains a frame check sequence (FCS) field (4 Bytes) based on a CRC-32 polynomial for error detection. A typical RTS frame has a length of 20 bytes, a CTS frame has a length of 14 bytes. A data frame usually has a MAC header of 24 bytes while an acknowledgement has a length of 14 bytes.

The MSDU may be anything between 0 and 2312 bytes long. Payload data which is longer than 2312 is split into fragments, which are transmitted separately. Fragmentation is mainly introduced due to interference. As the IEEE 802.11 standard is based on the use of the ISM bands, interference of many different sources, which cannot be controlled by the network, might occur (microwave ovens for example). Thus, splitting large payload data chunks into multiple smaller units increases the probability of a successful transmission. After the first fragment was successfully acknowledged by the receiver, the transmitting station immediately transmits the next fragment. Special fields within the MAC header allow the transmitter to indicate the transmission of multiple fragments. Each fragment is acknowledged separately by the receiver until the last fragment was received.

6.1.3 The IEEE 802.11a Physical Layer

One primary design principle of the IEEE 802.11 standard is that the above described MAC scheme works with any specified PHY. Meanwhile, there exist multiple physical layers based on different transmission schemes.

The IEEE 802.11a PHY [32] is based on the OFDM transmission scheme. It utilizes $N = 52$ sub-carriers spread over a bandwidth of $B = 16.25$ MHz. Thus, each sub-carrier has a bandwidth of 312.5 kHz, which results in a raw symbol length of 3.2 μs. Each symbol is protected by a guard interval of 0.8 μs, such that the resulting OFDM symbol has a total length of 4 μs. For the center frequency several bands between 5 GHz and 6 GHz exist, which are also allocated a different transmit power, varying between 40 mW and 800 mW. Four different modulation types (BPSK, QPSK, 16-QAM and 64-QAM) are available, combined with 3 different convolutional encoder schemes: $R = 1/2, 2/3$ and $3/4$. This results in a total of eight possible modulation/coding combinations, providing a physical layer throughput between 6 MBit/s and

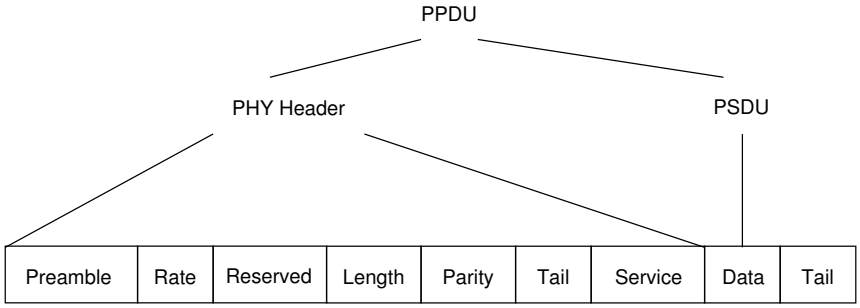


Figure 6.3: Structure of a IEEE 802.11a PHY frame

$54\ \mathrm{MBit/s}$. These different modulation/coding combinations are applied by a technique referred to as link adaption: Depending on a quality metric for the link, the same modulation/coding scheme is applied to all sub-carriers. However, the standard does not define a criteria to be used. One possibility is to use the error rate observed at the receiver, another possible criteria is the instantaneous SNR. The coding and modulation type are chosen in correspondence to the rate requested by the MAC layer. Hence, the PHY does not decide on the applied rate itself. The IEEE 802.11a standard does not provide the possibility of adaptive modulation, all sub-carriers are utilized with the same modulation type and with the same transmit power.

The interframe spaces are defined as follows. A SIFS has a time duration of $16\ \mu\mathrm{s}$. The length of a DIFS is $34\ \mu\mathrm{s}$. In order to transmit data, the PHY also adds a header and a trail to the MPDU to be transmitted (Figure 6.3, here the MPDU is denoted as PHY service data unit – PSDU). Among other functions, the header enables the synchronization between transmitter and receiver in time and frequency, provides the receiver the chance to estimate the attenuation per sub-carrier, indicates the length of the following data packet and indicates the rate (the modulation/coding combination) to be used for this packet. The total header has a length of 13 OFDM symbols (where the first 12 symbols – the PHY preamble – only require a transmit time of $16\ \mu\mathrm{s}$ such that the total PHY header requires a transmit time of $20\ \mu\mathrm{s}$). The Rate field indicates the modulation/coding combination used for the transmission of the following payload data transmission (see Table 6.1). The PHY header is always transmitted using the fixed modulation/coding combination of BPSK and a rate $1/2$ convolutional encoding. Immediately after this information has been transmitted, the payload data is transmitted using the modulation/coding scheme indicated in the PHY header. For payload transmission only 48 sub-carriers are used, the other 4 sub-carriers are used for carrier synchronization. The process of encapsulating an MPDU into the PHY protocol data unit (PPDU) is performed by the PHY layer convergence procedure (PLCP). The processing chain of the PLCP contains the following steps. After the PHY knows the rate requested by the MAC layer, it first scrambles the bits of the MPDU. This is done in order to prevent long sequences of 1s or 0s in the bit stream. Then, the complete MPDU is convolutionally encoded with the corresponding rate. Next, the resulting bit stream is interleaved and finally coded into modulation symbols per sub-carrier. At the receiver, this complete process is reverted, (hopefully) yielding the original bits of the MPDU.

Signal Bits in the Rate Field	Associated Modulation/Coding Combination
1101	BPSK & $R = 1/2$
1111	BPSK & $R = 3/4$
0101	QPSK & $R = 1/2$
0111	QPSK & $R = 3/4$
1001	16 QAM & $R = 1/2$
1011	16 QAM & $R = 3/4$
1101	64 QAM & $R = 2/3$
1101	64 QAM & $R = 3/4$

Table 6.1: Signal bits of the Rate field and associated modulation/coding combination according to IEEE 802.11a.

6.2 Integration Concept of Dynamic MU-OFDM Schemes

Some previous work on the topic of link adaption exists. In [109], the authors show that there exists an *optimal* link adaption scheme if the current channel SNR is known. This technique is extended to a dynamic programming approach by the same authors in [110], which determines the best PHY mode based on the current channel state as well as on the frame retry counter and the payload length. However, the authors constrain these approaches to link adaption, i.e. setting the modulation type for all sub-carriers equally, and not adapting the transmit power. In contrast, such a full dynamic OFDM approach is evaluated in [68] in the context of IEEE 802.11a. There, the authors find a significant performance increase for adapting the transmit power and modulation types per sub-carrier. However, the authors do not consider the integration of such a dynamic OFDM scheme into the standard. It is open how the transmitting station can acquire the required channel knowledge. Also, no signaling scheme is considered. Finally, the authors do not consider the point-to-multi-point OFDM communications case, where the degree of adaptions increases even further.

The goal of this proposal is to suggest a modification which changes as little as possible compared to legacy IEEE 802.11a. The major changes relate to the physical layer, however, also the MAC layer has to be slightly modified. The proposal is based on the IEEE 802.11a PHY specification and also supports stations compliant to legacy IEEE 802.11a. All stations compliant to the proposed new standard benefit from a higher throughput. This throughput advantage is a result of applying dynamic MU-OFDM schemes in the down-link. However, also the throughput in the up-link is improved if the transmitting station and the receiver (access point) both support the proposed standard. This is due to performing dynamic power allocation in combination with adaptive modulation. Accordingly, the throughput is still improved for any point-to-point connection between stations which support the proposed standard. Finally, for any point-to-point connection between a transmitting station, which supports the here proposed standard, and a station compliant to legacy IEEE 802.11a, the bit error rate is significantly improved by applying dynamic power distribution only.

The IEEE 802.11a standard is changed mainly regarding aspects of the framing, i.e. a new PLCP frame structure is introduced which allows the incorporation of more signaling informa-

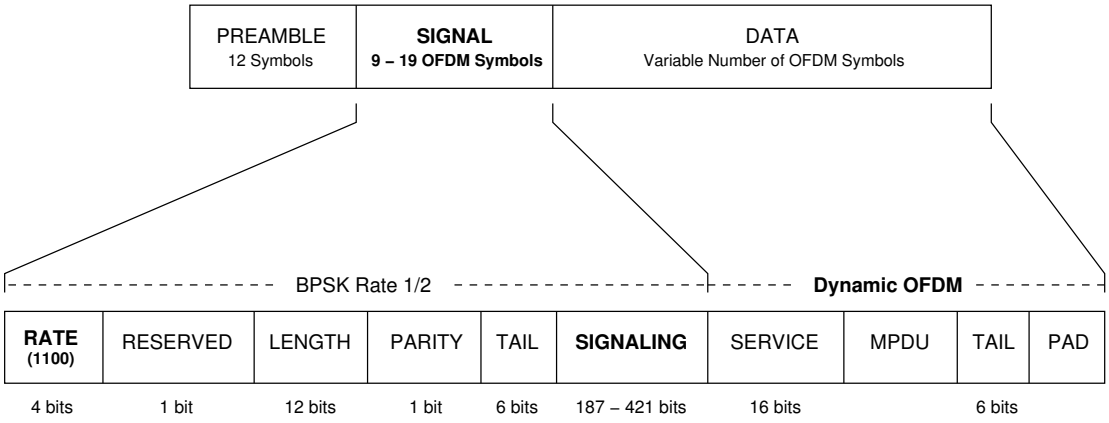


Figure 6.4: Structure of a PHY frame according to the new proposal. Note in particular that the only changes affect the Rate field and a new Signaling field. All data of the header is transmitted using BPSK with a rate $1/2$ convolutional encoder.

tion. In addition, the sequence of some frame exchanges is slightly modified. First, the framing changes are discussed, then the protocol aspects (the modified frame exchange sequences) are presented. Finally, some comments on the required hardware changes are provided.

6.2.1 Physical Layer Protocol Data Unit Layout

The most important change is a modified layout of the physical layer protocol data unit header (see Figure 6.4). Mainly, a legacy header differs from the new proposal in the additional signaling field (compare Figure 6.3 and Figure 6.4). However, a station/access point which receives a PLCP according to the new proposal recognizes the new layout from the value in the Rate field. The Rate field is 4 bits in length. In IEEE 802.11a it indicates the modulation/coding combination used for transmitting the payload data. Eight different combinations are specified in IEEE 802.11a (see Table 6.1). This leaves eight bit combinations unused. In the new proposal one of these unused bit combinations is employed in the Rate field of the new PLCP header. Stations detecting this unused rate then switch to the processing mode for the new proposal. If a station does not support this mode, the physical layer simply discards the physical layer protocol unit as it can not decode it. Thus, the usage of an unused Rate field bit combination serves as exit point for the new proposal. As bit combination the bits 1100 are suggested. Apart from the following proposal for dynamic OFDM schemes in IEEE 802.11, this exit point could also be used by other extensions (a certain further field is included in this proposal to distinguish between different extensions). After this special Rate field bit combination has been transmitted, the following four fields of the header (Reserved, Length, Parity, Tail) are unchanged. After the Tail field, the new Signaling field is introduced. This field contains the assignments of terminals and modulation types to sub-carriers.

The Signaling field has the following structure (see Figure 6.5). At the beginning, a short identification field of 2 bits is transmitted. The purpose of this field is to provide a further

characteristic in order to identify the new proposal. As other (future) proposals might also use the special bit combination 1100 of the Rate field as exit point, the ID field serves to distinguish between these extensions. For this proposal, the ID field always holds the values 11. A station thus only switches to the operation mode of the new proposal if the Rate field holds the specific value and if the ID field is set accordingly. Otherwise, the frame is discarded. Then, the Length field (9 bits in length) is transmitted. This field describes the length (in bits) of the complete Signaling field except for the Pad field. It is proposed here that the maximal length of the signaling field is 412 bits (which requires 9 bits in the Length field). However, note that this is only a proposal based on certain assumptions regarding the representation of the assignment information. The Length field always includes itself as well as the ID field in the numeric value indicating the data amount. The next field following is the Representation field. This is a 4 bit field which encodes the chosen representation for the assignment information. 4 different types are suggested (for the different signaling types compare also the signaling schemes investigated in Chapter 5), as well as a fifth type for management purposes:

- 1010 Point-to-multi-point signaling using the fixed size signaling field model (for all sub-carriers the assigned station and modulation types are indicated).
- 1011 Point-to-multi-point signaling using the variable size signaling field model and compression in the frequency domain (only assignments of sub-carriers which change from one transmission period to the next are indicated).
- 0101 Point-to-point signaling using the fixed size signaling field model (for each sub-carrier only the used modulation type has to be indicated).
- 0100 Point-to-point signaling using the variable size signaling field model exploiting frequency correlation (only indication of changes to the assignment of modulation types of the previous transmission phase).
- 0011 Indication of a station identification list only contained in the assignment field. This is used in the case of point-to-multi-point connections such that the access point receives CTS frames from each station involved in a MU-OFDM burst transmission without polling each station separately by a RTS frame (see Section 6.2.2).

Thus, for each transmit scenario an appropriate representation can be applied. For the identification of terminals an address space of 4 bits is used. However, for the point-to-multi-point case the multiplexing of only up to 8 terminals is suggested. Still an address space for 16 terminals should be available in order to supply each associated station one unique identification bit sequence. For a detailed discussion of the Assignment field, refer to Section 6.2.2.

For the modulation types an identification space of 3 bits is used. Note that in this proposal only the modulation types are adapted, not the coding. Therefore, per sub-carrier only one out of 5 different combinations has to be indicated (as a sub-carrier might also be employed with the "zero" modulation type). If the variable size signaling field model is applied, bundle sizes are identified by 3 bits, enabling the bundling of up to eight sub-carriers.

Finally, for the coding only 2 bit are required. Note that the coding is not adapted per sub-carrier, the complete MPDU is encoded equally. Therefore, only 3 coding schemes are available,

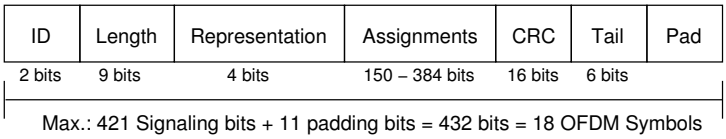


Figure 6.5: Layout of the Signaling field in the physical layer header of the new proposal.

which can be handled by 2 bits. After transmitting the assignment information, a 16 bit CRC field is transmitted which guards the assignment information. Finally, six tail bits (all set to zero) are transmitted in order to let the convolutional decoding process return to the zero state. At the very end, a varying number of 1 bits are appended to fill up the OFDM symbol.

The maximum size of the Signaling field occurs in case of a point-to-multi-point transmission and results in 421 uncoded bits (plus some padded bits). This requires 18 OFDM symbols to be transmitted using BPSK and a rate $1/2$ convolutional encoder. Thus, the total physical layer header is transmitted within 92 μs at most (16 μs for the preamble, 4 μs for the legacy IEEE 802.11a header, 72 μs for the Signaling field).

6.2.2 Protocol Aspects

After a station (or the access point) compliant to the new proposal has acquired the medium in order to transmit data, the transmission procedure follows a slightly changed message flow. The access point as well as the stations require additional functionalities on the MAC layer. In the following these changes are discussed for the up-link transmit direction (point-to-point) and for the down-link transmit direction (point-to-multi-point).

Up-link Data Transmission

The basic flow of an up-link data transmission is given in Figure 6.7. After the station acquired the medium, it transmits an RTS frame to the access point. The access point responds with a CTS frame if it received the RTS correctly. The PLCP preamble of this CTS response is used to estimate the attenuation of all sub-carriers as the wireless channel can be assumed to be symmetric [95]. Both frames, RTS and CTS, are transmitted using the regular IEEE 802.11a physical layer packet layout (modulation set to BPSK and the coding rate set to $1/2$). This ensures that all stations in the environment recognize the transmission and can set their NAV variables accordingly. The calculation of the NAV is performed at the station. Given the amount of data to be transmitted, the required time length is estimated assuming that all sub-carrier can only transmit with BPSK while the strongest coding is employed. Usually, this will result initially in a too long NAV setting). However, the correct NAV setting cannot be generated until the exact sub-carrier attenuations have been obtained (which is only the case after reception of the CTS frame). However, this pessimistic NAV setting is corrected after the payload transmission (see below).

Next, the bit loading algorithm determines the modulation type and transmit power spent per sub-carrier depending on the sub-carrier states estimated from the received CTS frame (for

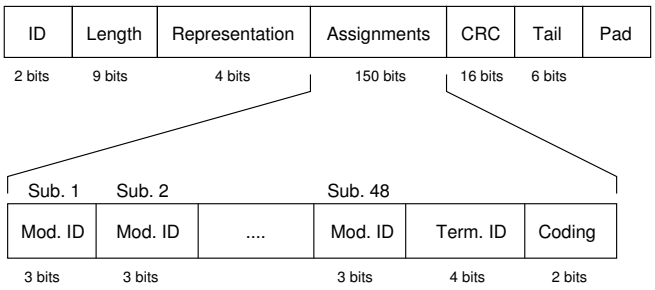


Figure 6.6: Structure of the Signaling field in the point-to-point communication mode (up-link, i.e. station to access point).

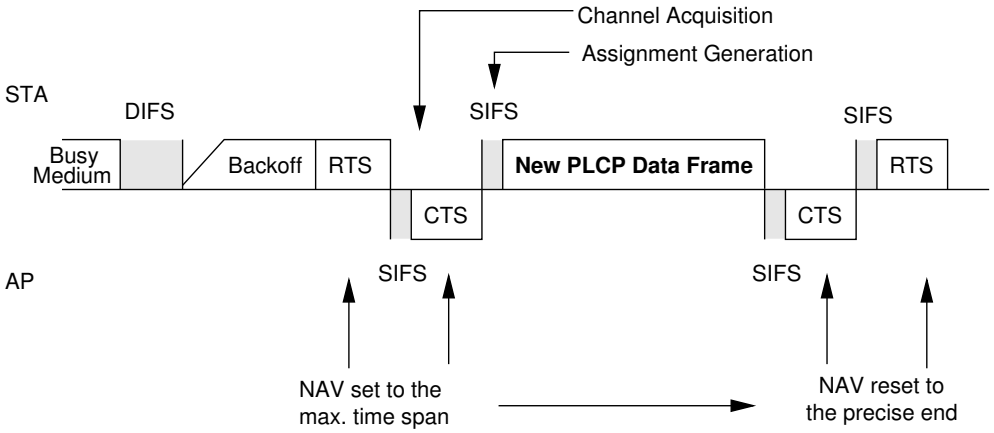


Figure 6.7: Transmission of data according to the new proposal in a point-to-point connection including the channel estimation and the generation of new resource assignments.

a discussion on the choice of the loading algorithm, see Section 6.2.3). After this assignment has been generated, the physical layer packet is generated were especially the Length field and the Signaling field have to be set to the values corresponding to the computed assignments. The Assignment field contains all modulation types chosen per sub-carrier. In addition, the transmitter also has to indicate the used coding scheme. Therefore, at the end of the Assignment field 6 more bits are transmitted indicating the binary identifier of the receiver as well as the used coding scheme (note that the coding scheme is identical for all sub-carriers). An example layout of the Signaling field in case of a point-to-point transmission is shown in Figure 6.6. The access point receives the data, decodes it according to the signaling information and (in case of a successful reception) transmits an acknowledgement in form of a legacy CTS frame. This CTS frame contains an update of the NAV setting. It can be recognized by all stations, including legacy stations. Finally, the transmitter conveys an RTS frame, destined for itself, to definitely update the NAV of any station which did not receive the update from the CTS frame previously. This frees the wireless medium. Therefore, three significant differences arise in the

case of a dynamic OFDM point-to-point communications, namely the generation of resource assignments, the signaling of these assignments and the reset of the NAV value after receiving an ACK frame. The different modulation types per sub-carrier require in total 8 OFDM symbols of signaling information ($= 32$ μs) using BPSK and a rate $1/2$ convolutional encoder for the fixed size signaling field approach. Also, the final RTS frame conveyed by the transmitter consumes another 60 μs. Therefore, the total overhead of the new concept in this point-to-point mode equals 92 μs. This assumes, that the bit and power allocations can be generated within a time span of a SIFS (16 μs – see discussion in Section 6.2.3).

In the case that the generation of the resource assignments requires more time, a busy tone has to be transmitted in order to indicate that the wireless channel is still in use. If the signaling information is corrupted, the complete packet has to be discarded at the receiver. It is rather unlikely that the signaling information is corrupted and not indicated by the error detection scheme. In this case, the MAC layer still can perform error detection of the payload by considering the checksum at the end of the MAC packet.

Down-link Data Transmission

In the case of a point-to-multi-point connection the access point might decide to initiate a dynamic MU-OFDM transmission phase. The start of such a phase depends on the data currently queued at the access point. If some data is queued for at least two stations which support the new proposal, a dynamic MU-OFDM phase can be initiated. However, the access point can also transmit the data to the stations sequentially, only boosting the point-to-point connections as described above.

A dynamic MU-OFDM phase is started like any other transmission by acquiring the wireless medium. After the access point obtained the medium, a Beacon frame is transmitted using the legacy IEEE 802.11a physical layer packet format. This is done primarily to have all associated devices set the NAV value. As in the point-to-point mode, the correct NAV setting is not known yet, as the channel states have not been acquired yet. Therefore, the access point announces a pessimistic NAV, which can certainly be achieved (for example, the NAV is computed by assuming all sub-carriers to be in the lowest modulation type). Then, the access point transmits a RTS frame using the new physical layer frame format. The RTS frame is addressed to the access point itself, however, the Signaling field of the physical layer packet contains a sorted list of terminal identifications which is interpreted by the stations as polling sequence. These stations will receive data during the following payload data transmission. The position of a stations identification in the list serves as allocation when to transmit a regular CTS by this station (as the access point has to acquire the channel knowledge). This special terminal list is indicated by a certain bit combination (0011) in the Representation field of the Signaling part.

After the stations supporting the new standard received this special RTS frame, the involved stations reply (depending on their position in the station identification list) a legacy CTS frame to the access point, setting the NAV value according to the previously received Beacon. The PLCP preambles of these CTS frames are used to estimate the sub-carrier attenuation for each station. Once the channel knowledge is obtained, the access point starts to generate the assignments (see Section 6.2.3). If the generation of the assignments is not done within the time span of DIFS, the access point can poll stations to transmit data in the up-link. Alternatively, the access point

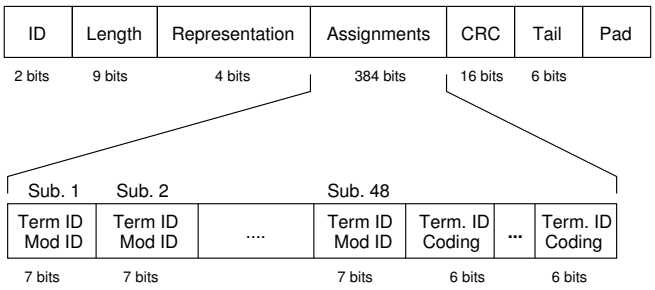


Figure 6.8: Structure of the Signaling field in the point-to-multi-point communication mode (down-link, i.e. access point to stations).

can transmit data to legacy IEEE 802.11a stations which cannot participate in the dynamic MU-OFDM phase. However, if none of the two options can be done (as no station has data for the up-link or no legacy IEEE 802.11a station is present), the access point has to guarantee that the medium is not accessed by stations which did not receive the NAV value previously. Hence, a busy tone must be transmitted.

Once the assignments have been generated, the down-link transmission starts using the new physical layer packet format, including the signaling information (for example, represented in the form of the fixed size signaling field model). The access point also has to indicate which station receives packets with which coding scheme. Therefore, several tuples of station identifications and coding identifications are added to the end of the Assignment field (Figure 6.8). Assignments are generated such that the parallel packet transmission requires an almost equal amount of time (see Section 6.2.3). If the station conveys a normal CTS to the access point, the data was received correctly. Otherwise, the station transmits nothing. Then, the access point knows that the payload data for this station has been lost. For the ACK frame transmission, the same sequence as in the initial CTS frame transmission is used by the stations. If the signaling information is lost due to bit errors, no frame is transmitted back to the access point. Then, the access point has to address this station again in a later transmission phase. The ACK frames are conveyed using the legacy PLCP format of IEEE 802.11a. Within each ACK frame, the NAV value is set to the correct end of the multi-user mode transmission phase. To assure that all stations receive a correct NAV setting, the access point finally also transmits an RTS frame in the legacy format, addressing itself. Afterwards, the wireless medium is freed. An example transmission sequence is shown in Figure 6.9.

Recognition of Stations and Access Points Compliant to the New Proposal

Another important issue to resolve is the question how stations and access points can get the information if the intended receiver supports the new standard or not. This can be handled during the association phase. The access point simply announces the support of the new concept in the (extended) capabilities field of the Beacon frame. Any Beacon contains in its frame body codes for multiple optional capabilities, defined by the standard [108]. One of these codes has to be added for the new concept. A station associating to such an access point may then announce

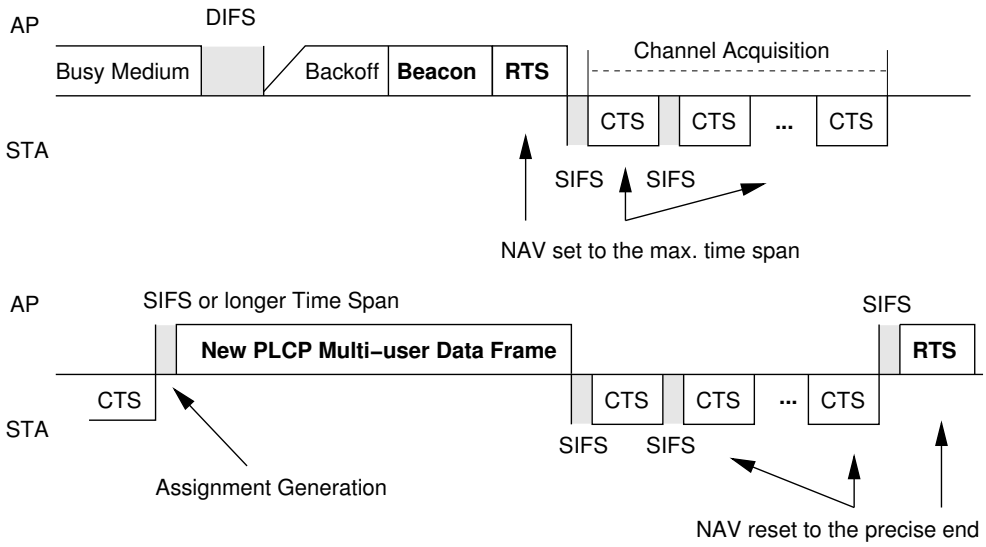


Figure 6.9: Transmission sequence of the new concept in the case of the point-to-multi-point communication mode (down-link, i.e. access point to stations).

its compliance to this new concept in the association frame transmitted (this association frame contains a field where stations can announce their compliance to several different rates, therefore, one of these rates would have to be reserved to indicate the support of the new concept).

Finally, the question arises how stations can obtain their Signaling identification bit sequence. This could also be done during the association phase. Alternatively, the access point could transmit some new management frame, announcing the current relation between stations MAC addresses and signaling bit sequences.

6.2.3 Transmitter and Receiver Hardware

Several changes are required for the transmitter and receiver hardware. First of all, the new proposal applies adaptive modulation with a varying amount of transmit power to each sub-carrier. This is not possible in the current IEEE 802.11a mode as only one modulation/coding combination can be applied to all sub-carriers. Control circuits are required which indicate which sub-carrier has to be modulated with which type. In addition, the transmit power can be adapted dynamically in the new proposal, therefore control circuits are also required for this. In the case of a dynamic point-to-multi-point transmission, different data streams for different stations have to be managed according to the generated assignments.

All these changes have to be added to the receiver structure as well. Separate demodulation is required per sub-carrier as well as separate decoding, according to the information received during the signaling phase. Obviously, once such a transmitter/receiver structure is given, legacy IEEE 802.11a can easily be implemented as well. The mode, the receiver is in, is determined by the reception of the identification bit fields (Rate field and ID field with predefined bit sequences)

for the new proposal. For the access point as transmitter the new mode is always switched on if data for various stations supporting the new mode is queued. Then, if the medium is acquired, the new mode is activated. For any other station as transmitter, the new mode is activated if the corresponding receiver is able to switch to the new mode. Lists and certain identification schemes are required such that the transmitter knows which station can switch to the new mode (see below). In case data is transmitted to a legacy IEEE 802.11a station, the transmitter can still optimize the distribution of the transmit power. It is well known that this decreases the error probability significantly (see Section 3.1).

In addition to the changes of the transmitter and receiver structure, any transmitter also has to be able to estimate the attenuation of each sub-carrier given the transmitted preamble of legacy IEEE 802.11a. The precise algorithm for this is not part of the proposal. However, the estimation is subject to a timing constraint.

Furthermore, all stations and access points compliant to the new standard perform dynamic assignments of transmission resources. Stations only adapt the transmit power and the modulation type per sub-carrier. In contrast, an access point generates disjunct sets of sub-carriers assigned to different terminals. For each of these sets the transmit power and modulation type is varied as well.

Once the sub-carrier attenuations have been acquired, a station or the access point start the computation of the assignments. This computation has to be performed quite fast due to several issues. First of all, the sub-carrier attenuations change in time. In a typical indoor scenario the coherence time of a sub-carrier at $5\ \mathrm{GHz}$ is about $6\ \mathrm{ms}$. The faster the assignments are computed, the more valid the sub-carrier estimates are for the following down-link transmission.

Another constraint is given by potential legacy IEEE 802.11a stations in the system. While a transmitter generates the assignments, the medium is not busy. Thus, legacy IEEE 802.11a systems might interpret a too long time span required for computation as a free medium for which they start the channel acquisition (assuming that they did not receive the NAV setting previously). In IEEE 802.11 a time span of DIFS is defined as the minimum time the medium has to be observed as unused before stations can start to acquire the medium. For legacy systems, DIFS is set to $34\ \mu\mathrm{s}$. Either the generation of the new assignments as well as the processing of the data (generating the physical layer protocol unit including the signaling information) can be performed in this time span or the station generating the assignments has to transmit a short busy tone.

A third issue is that the generation of the assignments simply requires time. During this time, the channel is not used in this new proposal. In contrast, in an IEEE 802.11a system this time span is already used to transmit data if the computation requires more than the time span of a SIFS ($16\ \mu\mathrm{s}$). Therefore, also from this point of view sufficient computational resources are required at the stations as well at the access point. The computational load is the highest at the access point, in particular if many stations are involved in the next transmit phase.

Therefore, the choice of objective and the choice of the used assignment algorithm is important. For the point-to-point dynamic OFDM transmission mode, it is suggested to perform a loading which maximizes the throughput for a given target BEP. There is evidence that this can be performed within the time span of a SIFS [56].

In the case of the multi-user dynamic OFDM transmission mode, the issue of the computa-

tional complexity is more difficult. The best choice for the assignment would be to solve the rate adaptive optimization problem (see Section 4). This would ensure that the parallel MPDU transmissions are finished more or less simultaneously, as each station would have been assigned the maximum equal transmission rate (or the adjusted equivalent in case that the stations have different MPDU sizes queued). However, this problem cannot be solved to optimality in the short amount of time. The approximation schemes discussed in Section 4.2 could be applied, especially if one considers that specialized hardware could be built for these approximation schemes. As the multi-user dynamic OFDM transmit mode is restricted to a maximum of 8 stations involved, some approximation scheme could probably be implemented on specialized hardware, requiring at most 100 μs or less to perform the assignments. Today, the discussed relaxation scheme requires only a few milliseconds running on Linux userspace to perform the power and sub-carrier assignments for up to 8 stations, applying a full blown LP algorithm. Furthermore, after the first few sub-carrier attenuations have been estimated, one might be able to pipeline the problem, such that the first assignments are already computed prior to the complete reception of all sub-carrier attenuations of all stations.

6.3 Performance Evaluation

The new concept has been evaluated by means of simulation. Only the DCF infrastructure mode of IEEE 802.11 has been considered. As performance metric, the time span required to transmit an MPDU successfully (several MPDU's in the case of the point-to-multi-point mode) by either legacy IEEE 802.11a or some variant according to the new concept is considered. This takes into account the increase in spectral efficiency as well as all additional overhead, as discussed previously. Note that collisions or the impact due to the backoff scheme are not considered in this metric, as the new concept behaves similar to legacy IEEE 802.11a in terms of the backoff scheme and the collision occurrence (see also the discussion below).

In the following, first the methodology and the results for the point-to-point communications case (station to access point) are discussed. Afterwards, the point-to-multi-point case (access point to multiple stations) is considered.

6.3.1 Up-Link (Point-to-Point) Communications

Simulation Scenario

In the up-link case the following scenario is considered: Some station j has a packet, i.e. MPDU, of a certain size to be transmitted to the access point. Therefore, it waits until the medium is free and starts the transmission after some deferral period. As this medium acquisition process is not affected by the fact that the station supports the new concept, it is not consider any further. Then, the station starts the transmission of the PLCP frame either by applying legacy IEEE 802.11a or by applying the new concept. In detail, four different comparison schemes are considered:

1. Legacy IEEE 802.11a without RTS/CTS handshake.
2. Legacy IEEE 802.11a with RTS/CTS handshake.

3. Dynamic OFDM according to the new concept, but with adaptive modulation only; the transmit power is distributed equally.

4. Dynamic OFDM according to the new concept with bit and power loading.

Notice the specific adaption scheme for each comparison variant. In the case of legacy IEEE 802.11a, it is well known that there exists an optimal link adaption strategy [109]. Unfortunately, the transmitter requires the current average SNR in order to perform the optimal link adaption. In case of comparison scheme 2, this knowledge can be assumed to be present at the station (due to the RTS/CTS). In contrast, in comparison scheme 1 the station could only utilize its average SNR value of the last received Beacon frame of the corresponding access point. This information is most likely outdated, as Beacon frames are transmitted often with a period of $100\ \mathrm{ms}$. However, in this study it is assumed that the station can adapt the transmission rate *optimally*, as described in [109]. Recall that this is an assumption which strongly favors the legacy mode, at least regarding variant 1.

In case of the dynamic schemes 3 and 4, the sub-carrier modulation types are adapted such that the resulting average BEP equals the average BEP of a legacy transmission. Consider a current "snap-shot" of the 48 sub-carriers. Clearly, the attenuation varies per sub-carrier. Hence, as legacy IEEE 802.11a applies the same transmit power and modulation type to each sub-carrier, the resulting BEP varies per sub-carrier as well. Given the BEP per sub-carrier, an average over all sub-carriers, $\overline{p_{\max}}$, can be obtained for the legacy system. This average value is used in the dynamic variants to perform the adaptive modulation. In comparison case 4, the station applies the Hughes-Hartogs algorithm to perform the bit and power loading. During the loading process, the power amounts for each modulation type and sub-carrier are generated such that any sub-carrier features at least $\overline{p_{\max}}$. The difference between comparison case 3 and 4 is that in 3 no power loading is applied. Instead, the transmit power is distributed statically and for each resulting SNR per sub-carrier the corresponding modulation type is chosen (with respect to $\overline{p_{\max}}$). In contrast, in comparison scheme 4 a full blown bit and power loading algorithm is applied.

As the four schemes are identical regarding their average bit error probability in the physical layer, the total frame error rate is assumed to be equal as well for a given (stable) attenuation per sub-carrier. Again, this is in favor of the legacy system, as the link adaption strategy leads to an uneven bit error distribution in the physical layer. Perfect interleaving is assumed, such that any impact due to the different bit error distributions regarding the four different comparison schemes are not taken into account.

Finally, after the station transmitted the PLCP frame it waits for an acknowledgement from the access point. It is always assumed that packets are transmitted correctly, as the focus is only on the time span difference between a successful transmission due to legacy IEEE 802.11a compared to the new concept. As both transmission schemes have the same average BEP in the physical layer, they will have the same frame error rates as well. Therefore, considering frame errors and the corresponding retransmissions would not change the qualitative difference between the four schemes investigated here.

Parameterization and Methodology

The following basic simulation scenario is chosen: The cell radius is set to $r_{\text{cell}} = 80\ \text{m}$, the center frequency is $f_{\text{c}} = 5.2\ \text{GHz}$. The maximum transmit power equals $P_{\text{max}} = 10\ \text{mW}$. The bandwidth, the number of sub-carriers, the symbol duration and the guard interval are all chosen in accordance to IEEE 802.11a (see Section 6.1.3).

The sub-carrier attenuations $h_n^{(t)}$ are generated based on path loss, shadowing and fading. For the path loss, a standard model $h_{\text{pl}}^{(t)} = K \cdot \frac{1}{d^{\alpha}}$ is assumed [4], parameterized by $K = -46.7\ \text{dB}$ and $\alpha = 2.4$ (corresponding to a large open space propagation environment). For the shadowing $h_{\text{sh}}^{(t)}$, independent stochastic samples from a log-normal distribution are assumed, characterized by a zero mean and a variance of $\sigma_{\text{sh}}^2 = 5.8\ \text{dB}$. The shadowing component changes every second. Each sample $h_{\text{fad}}^{(t)}$ of the fading process is assumed to be Rayleigh-distributed. The frequency and time correlation of $h_{\text{fad}}^{(t)}$ are characterized by a Jakes-like power spectrum and an exponential power delay profile (parameterized by the maximum speed within the propagation environment, set to $v_{\text{max}} = 1\ \text{m/s}$, the center frequency and the delay spread $\Delta\sigma = 0.15\ \mu\text{s}$). It has been shown that for wireless local area networks the channel can safely be assumed to be constant over a time horizon of multiple milliseconds [95]. Thus, in case of the dynamic variants the measured sub-carrier attenuations during the CTS are assumed to be stable during the complete PLCP frame transmission. The noise power σ^2 is computed at an average temperature of $20^{\circ}\ \text{C}$ over the bandwidth of a sub-carrier.

According to these parameters, sub-carrier attenuation traces are generated by a computer program for several stations at different positions. These channel traces are used to obtain the performance results for each comparison scheme. Every $10\ \text{ms}$ it is assumed that an MPDU of a certain size is passed to the physical layer. Then, another computer program reads the corresponding sub-carrier attenuations from the trace file and generates the throughput over all sub-carriers according to the considered scheme. Thus, per MPDU a total time span can be obtained required for a successful transmission, including the reception of the acknowledgement. For each station and comparison scheme, this is repeated over a quite long time span (a couple of seconds), yielding a sequence of time spans. These sequences are recorded in a file and afterwards statistically analyzed.

Results

In Figure 6.10 the average performance results are given for each considered scheme. Each point represents the average number of OFDM symbols over 16 different station positions that is required to transmit the corresponding MPDU. The confidence intervals (95% level) are below 1% for each point. Comparing the different schemes, the new concept outperforms the legacy system with RTS/CTS for MPDU sizes larger than $500\ \text{byte}$. Compared to the legacy system without RTS/CTS, the new concept pays off for MPDU sizes larger than $1400\ \text{byte}$. Note that the majority of IEEE 802.11 implementations apply the RTS/CTS handshake. At an MPDU size of $1500\ \text{byte}$ (which is the typical size of IP packets), the new concept transmits packets faster by 20% (on average). Also note that the difference between the loading scheme and adaptive modulation is rather small considering the overall average transmit time of the cell.

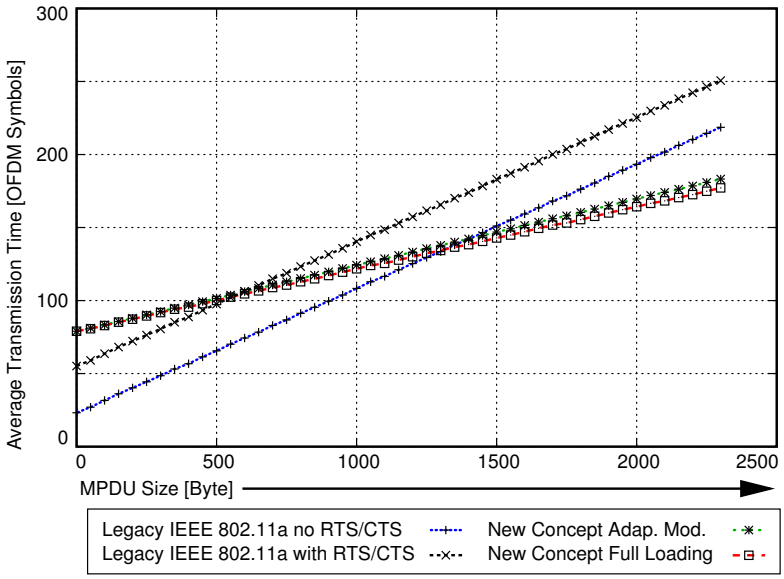


Figure 6.10: Performance comparison of the new concept and legacy IEEE 802.11a regarding an increasing MPDU size.

Figure 6.11 shows details which give a better understanding of the performance differences. The upper graph of Figure 6.11 shows the average number of bits transmitted per symbol and sub-carrier during the payload phase of the four different schemes (note that the two legacy variants apply the same link adaption, therefore their results are the same in this figure). While the optimal link adaption scheme from [109] achieves on average a throughput of 3 bit per sub-carrier and symbol, the throughput of the two dynamic schemes is much higher with about 5.8 bit per symbol and sub-carrier. Therefore, the link adaption scheme of the legacy system applies on average either the QPSK or the 16-QAM modulation type, while the adaptive schemes of the new concept almost always utilize each sub-carrier with the 64-QAM modulation type. Note that all schemes achieve the same average bit error probability. The lower graph of Figure 6.11 compares the MAC overhead of the different schemes. It is determined considering the transmission of a MPDU of size of 0 byte. Hence, the resulting time span represents mostly the MAC overhead. The new concept has the highest overhead, requiring about 20 OFDM symbols more than the legacy scheme with RTS/CTS. This difference of the overhead results in a better average performance of the legacy schemes for smaller MPDUs (see Figure 6.10). In the upper graph of Figure 6.12 more specific results are shown for a station which is rather close to the access point (distance of 20 m). In this case, the overall performance gain of the new concept is rather small compared to legacy IEEE 802.11a. In fact, the legacy variant without RTS/CTS achieves a better performance for any MPDU size. However, the new concept still achieves a performance gain compared to legacy 802.11a with RTS/CTS for MPDUs larger than 1000 byte (considering the loading scheme). Note that the legacy system benefits from a short distance (for the scheme with RTS/CTS handshake, the transmission of an MPDU of 2300 byte

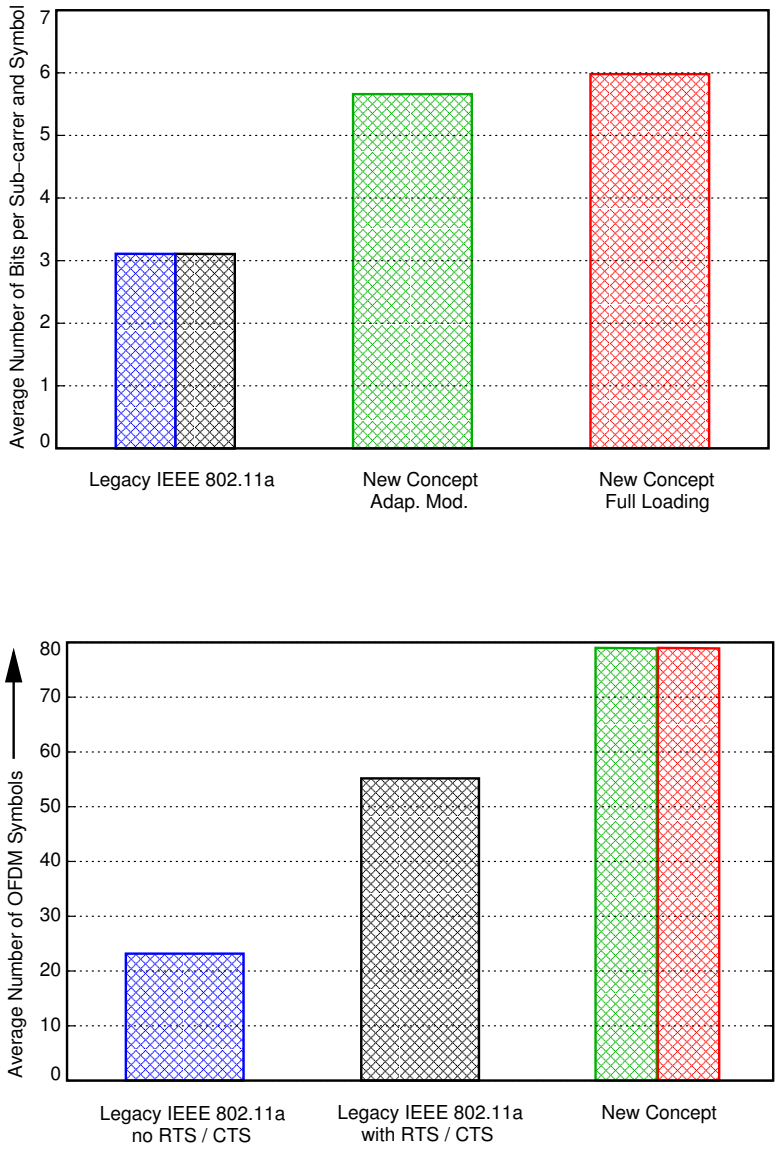


Figure 6.11: Efficiency of the various variants. Upper graph: Average throughput per symbol and sub-carrier achieved on average during the MPDU transmission – Lower graph: Average number of OFDM symbols required to transmit an MPDU of 0 bit.

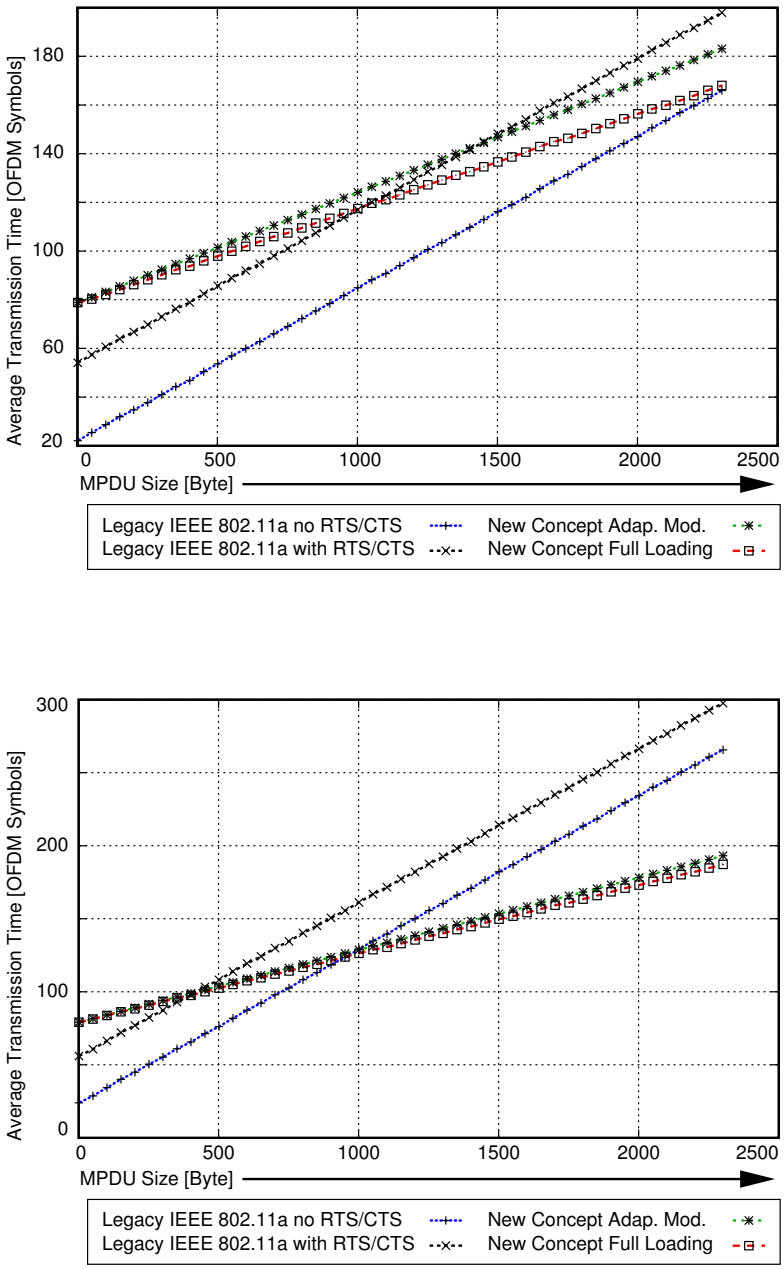


Figure 6.12: Performance comparison of the new concept and legacy IEEE 802.11a regarding an increasing MPDU size for a station close to the access point (upper graph) and far away from the access point (lower graph).

requires about 200 symbols, whereas the average for all stations – Figure 6.10 – is 250 symbols). In contrast, both schemes of the new concept benefit only marginally from the short distance (for an MPDU of 2300 byte the loading scheme requires 168 symbols if the access point is close, while 180 symbols are required on average – Figure 6.10).

Furthermore, Figure 6.12 also shows the results for a setting with a large distance between access point and station (lower graph – distance of 65 m). In this case, both legacy schemes have very large transmit durations (about 300 symbols for an MPDU of 2300 byte considering the legacy scheme with RTS/CTS). Interestingly, the transmit times of the new concept remain close to the average of all stations (Figure 6.10). Therefore, the performance gain is larger for stations far away from the access point, i.e. for stations which have a large attenuation on all sub-carriers regarding the access point.

6.3.2 Down-Link (Point-to-Multi-Point) Communications

Simulation Scenario

In this section the access point is considered to hold MPDUs for several different stations in its transmit memory. Then, it can perform a sequence of point-to-point transmissions where each single transmission is compliant either to the new concept or to legacy IEEE 802.11a. However, it is proposed that the access point initiates in this case a point-to-multi-point transmission.

In order to evaluate the performance of this transmission mode, the following scenario is considered: The access point has several packets for different stations in its transmit queue. For illustration purposes the MPDU of each packet are set to the same size. However, note that the new concept can be extended to any other packet size mixture without loss in performance. Five different comparison schemes are considered. The first four of them are equal to the four schemes of the point-to-point case, however, the MPDUs are transmitted sequentially to the corresponding stations. As in the previous section, it is assumed that the packets are received successfully and no packet collisions occur. Thus, between two successful point-to-point transmissions the access point only waits for the duration of a DIFS (as the backoff behavior is not captured in the simulation, which would affect all "sequential" comparison schemes equally).

The fifth scheme is the point-to-multi-point transmission mode of the new concept. As in the point-to-point case, the dynamic assignments are tuned such that a BEP is achieved which equals the BEP in case of an optimal legacy transmission (according to the optimal link adaption described in [109]). The dynamic assignment algorithm considered in the simulations is the optimal assignment maximizing the throughput of all terminals equivalently, thus, the optimal solution to the corresponding rate adaptive problem. Note that in Chapter 4 several suboptimal scheme for the rate adaptive problem have been investigated, which approximately achieve the same performance while featuring a lower complexity. In addition to these, also the modifications of Section 4.4 for a varying packet size per terminal can easily be adopted. In fact, that approach can be coupled with a packet scheduler.

The same parameterization is assumed and the same methodology is applied as in the point-to-point case. The optimal assignments in the case of the multi-user transmission are generated by CPLEX 9.0, a state-of-the-art solver for linear integer programming problems.

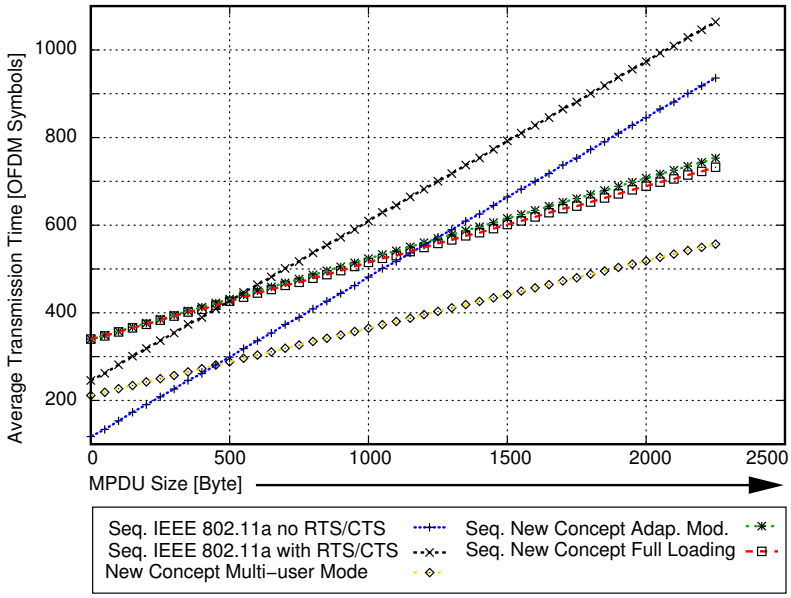


Figure 6.13: Performance comparison of the multi-user mode of the new concept vs. "sequential" MPDU transmission either by the point-to-point mode of the new concept or by legacy IEEE 802.11a for an increasing MPDU size per station (4 stations).

Results

Figure 6.13 shows the performance results in the case that the access point has packets for 4 stations queued. Clearly, the dynamic schemes outperform the legacy schemes. Compared to the legacy "sequential" transmission with RTS/CTS, the dynamic schemes pay off for most MPDU sizes. The difference between the multi-user mode and all other "sequential" schemes is large. Note that all four sequential schemes are simulated rather optimistically, as only a time span of a DIFS is assumed to be between two consecutive MPDU transmissions. Usually, the time span between two successive transmissions will be larger, as the deferral period has to be taken into consideration as well. In addition, other stations could acquire the medium in the mean time, which would further prolong the time span between two consecutive MPDU transmissions. Therefore, in reality the multi-user mode is likely to pay off more than indicated in Figure 6.13. Figure 6.14 shows results for two other station settings. In the upper graph, the number of stations included in the transmission scenario is set to 2. In this case, the performance advantage of all variants of the new concept are rather small. For MPDU sizes below 350 byte, the new concept does not outperform any "sequential" legacy transmission. However, recall that the simulation scenario for all "sequential" schemes is rather optimistic. For an MPDU size of 1500 byte, the multi-user mode outperforms legacy IEEE 802.11a with RTS/CTS already by about 30%. For that MPDU size, the sequential scheme with bit and power loading is outperformed by about 10%. The larger the MPDU size, the more all "sequential" schemes are outperformed by the multi-user mode. The lower graph of Figure 6.14 shows results for the

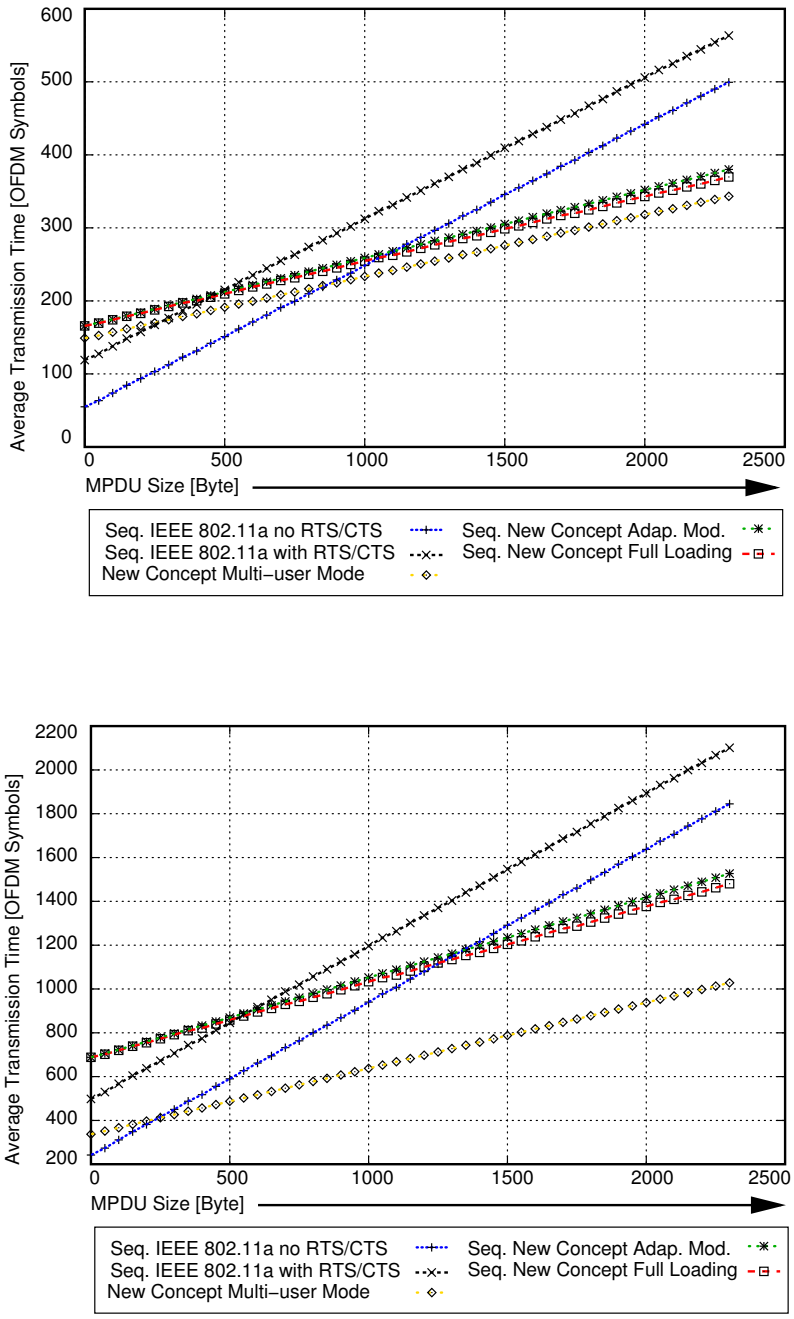


Figure 6.14: Performance comparison of the multi-user mode of the new concept vs. "sequential" MPDU transmission either by the point-to-point mode of the new concept or by legacy IEEE 802.11a for an increasing MPDU size per station (Upper graph: 2 stations are assumed – Lower graph: 8 stations are assumed) .

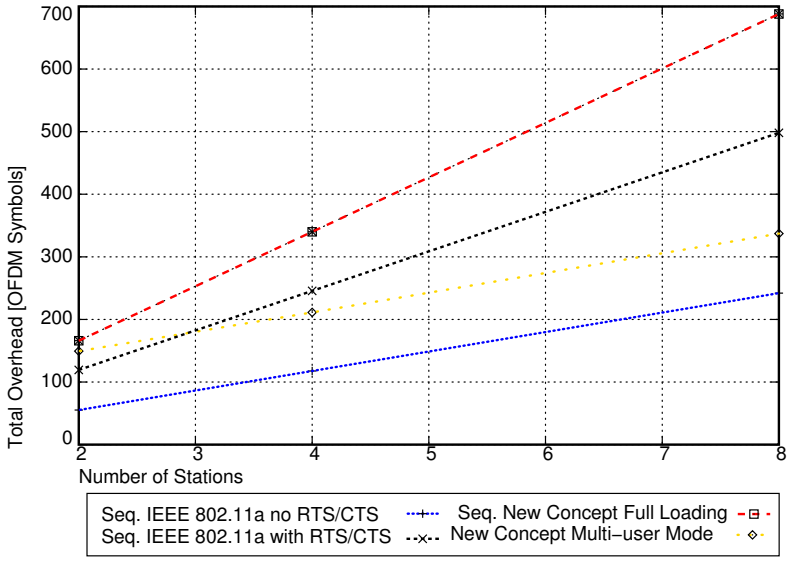


Figure 6.15: Management overhead for all five variants for an increasing number of stations included in a parallel or sequential MPDU transmission. The two sequential variants of the new concept have the same performance.

highest possible number of stations (8 stations). If the number of stations is high, the multi-user mode outperforms all "sequential" schemes (except the legacy mode without RTS/CTS) for all MPDU sizes. For a typical IP packet size, the legacy scheme with RTS/CTS is outperformed by 100% considering the multi-user mode. All other "sequential" scheme are outperformed by 50% for this MPDU size. Figure 6.15 shows the management overhead per variant as the number of stations included in a multi-user transmission increases. As more and more stations are part of the parallel multi-user transmission, the overhead increases only slightly. In contrast, the overhead increases more for the sequential legacy IEEE 802.11a transmission with RTS/CTS. The sequential variants of the new concept are in between. Hence, a large percentage of the performance gain achieved by the multi-user mode of the new concept results from a much lower management overhead. On top of this advantage, the multi-user mode also features a higher throughput during the payload transmission.

6.4 Conclusions

In this Chapter, a concept is proposed which incorporates dynamic OFDM schemes into IEEE 802.11a. This new concept fully supports legacy IEEE 802.11a systems. On the other hand, it exploits the performance advantage of dynamic OFDM schemes if transmitter and receiver both support the new concept. The IEEE 802.11a standard has to be changed only regarding a few aspects. A new PLCP frame is introduced, as more signaling information has to be transmitted. Also, the sequence of exchanged frames is modified.

The new concept supports two different modes. In the first one, dynamic OFDM for point-to-point data communications is applied. For each sub-carrier, the modulation type and transmit power is adapted individually. This mode can be applied by stations which have to transmit MPDUs to the access point. Channel knowledge is acquired by a mandatory RTS/CTS handshake prior to the payload transmission. In order to assure the correct NAV setting in all stations, transmitter and receiver both have to convey one more management frame after the payload frame exchange. Apart from that, the data transmission is the same as used by legacy systems in the IEEE 802.11 DCF mode.

The second mode which is supported by the new concept, is a multi-user OFDM mode. The key feature of this mode is the parallel transmission of MPDUs to several stations. This requires some more changes to the legacy transmission scheme, especially regarding the way the sub-carrier attenuations can be obtained at the access point.

Regarding the performance of the new variants, both modes improve the throughput of legacy IEEE 802.11a systems. The point-to-point mode benefits only from a higher throughput during the payload transmission. Therefore, the point-to-point mode is recommended only for larger MPDU sizes or for stations that have a low average SNR (due to a high path loss, for example). For small MPDU sizes, the point-to-point mode does not pay off due to the additional overhead required for signaling, channel acquisition and the correction of the NAV value.

However, the multi-user mode benefits from two effects. First of all, the throughput during the payload transmission is further enhanced, which boosts the performance. On the other hand, the overhead is significantly lower for the multi-user mode, as the wireless medium is acquired only once, apart from other issues. Therefore, especially the multi-user mode can be recommended for application in practical systems regardless of the MPDU size. The performance gain due to the multi-user mode is up to 100%.

As future work it remains to perform network simulations of the new concept in comparison to legacy IEEE 802.11a. These simulations can take several other aspects into account. In the performance evaluation above, frame errors, collisions and backoff periods have not been considered. Also, different packet sizes and flow dynamics between up-link and down-link have not been taken into consideration. All these aspects are likely to have an impact on the observed performance and can be investigated in detail by such simulations. However, as the above performance evaluation was based on very "optimistic" assumptions regarding the legacy system variants, the network simulations are expected to verify the performance gain of the new concept vs. legacy IEEE 802.11a.

As second issue, the usage of different queue-adaption schemes can be considered, as the dynamic multi-user transmission mode provides the flexibility to serve several stations in parallel. This issue has already been addressed in Section 4.4 in the context of centralized scheme. There the performance evaluation has shown that the capability of an OFDM system can be increased even further by exploiting the queue dynamics. As it has been shown in this section that the new concept saves MAC overhead, it can be modified for the case that packets with different sizes are queued. It might even be possible to extend the concept such that multiple MPDUs for different stations are transmitted on one single sub-carrier, if several large MPDUs (for example, IP packets) are queued together with several very small packets (like VoIP packets). This would require additional changes but could further increase the MAC efficiency.

Chapter 7

Conclusions

This thesis considers dynamic multi-user OFDM systems. In these systems the transmit power and sub-carriers are assigned dynamically to different terminals depending on the current state of the OFDM sub-carriers. The analysis assumes a single wireless cell where the access point performs these assignments periodically for the down-link transmission direction. It is assumed that the access point has sufficiently accurate channel knowledge, while the wireless channels are stable over the time span of one assignment period.

Initially, the trade-off between computational complexity and system performance is studied. In general, the down-link of a dynamic multi-user OFDM system can be modeled as an (mixed integer) optimization problem, where the assignments, i.e. the transmit power and the sub-carriers, are the variables. However, the thesis shows that the resulting optimization problems are *NP*-complete. Hence, solving them to optimality cannot be recommended for practical systems. Several alternative options are further studied, which lead to a lower computational load but provide also a lower performance. Instead of assigning the transmit power and the sub-carriers dynamically, the transmit power could be fixed, leaving only the sub-carriers to be assigned (optimally). For cells with a smaller radius, this approach is found to provide quite good performance results. However, even this approach is *NP*-complete. Therefore, suboptimal assignment algorithms are studied. Specifically, the approach of relaxing the integer constraint on the sub-carrier assignments is proposed and evaluated in comparison to two other schemes suggested recently. It is found that the relaxation strategy provides the best performance of all the schemes proposed so far.

The second issue considered in this thesis is the effect of signaling. As the access point periodically generates the assignments for the down-link transmissions, the terminals have to be informed of the assignments in order to decode the correct data. Thus, a signaling system is required to convey this information, which obviously consumes system resources. It is found that the qualitative behavior of the system performance changes if the signaling overhead is taken into consideration. Especially the number of sub-carriers the system bandwidth is split into is a sensitive parameter regarding the performance. It turns out that for a given system and propagation environment an optimal number of sub-carriers exists which provides a maximum performance. This motivates the idea to have access points adapt the sub-carrier number depending on the propagation environment and number of terminals currently associated to them. In addition,

situations may arise where the overhead can become very large. In these cases, the overhead can be reduced effectively by the application of either optimization schemes or compression. Both reduction schemes exploit the correlation of assignments in time or/and frequency. However, the considered compression schemes have a lower complexity while providing a similar performance compared to the optimization approaches. In almost all scenarios the performance of a dynamic MU-OFDM system is reduced at most by 10% if one of the overhead reduction schemes is applied.

Finally, an outline of a new standard proposal is presented. It is based on IEEE 802.11a systems and incorporates a dynamic multi-user OFDM approach to increase the system performance of such wireless local area networks. This proposal features a scheme to acquire the channel knowledge. It enhances the system performance for point-to-point communications as well as for point-to-multi-point communications by dynamic power and sub-carrier assignments, while also serving legacy IEEE 802.11a systems.

Several issues remain as future work. Regarding the complexity of the assignment strategies it is open if the relaxation approach can be performed faster than a milliseconds, for example. The core problem is to consider if the corresponding linear program can be solved within this time span by some algorithm. Using commercial standard software, computation times on fast machines are at about $20\ \mathrm{ms}$ and below. However, a different approach could be to investigate the characteristics of the resulting polytope more closely and maybe strip off several components of open source solvers for linear programs. This could ultimately lead to very fast computation times if also specialized hardware can be assumed.

As second issue it remains to investigate the nature of optimization problem (5.25) (and the other two problems, (5.25) and (5.28)) more closely. While compression schemes can achieve a similar system performance, applying these optimization approaches has several advantages in order to reduce the signaling overhead. In particular, the memory and processing power requirements for terminals are higher if the compression approach is applied. In addition, undetected bit errors can lead to much more serious problems in the case of compression than in the case of the optimization approach. Thus, if an efficient algorithm can be found for optimization problems (5.25), (5.25) and (5.28), this will probably be the algorithm of choice for future systems.

Finally, an implementation of a dynamic multi-user OFDM system would be the ultimate proof of the superior performance of such a system. For example, such an implementation could be based on the outline of a standard proposal for IEEE 802.11a systems, presented in Chapter 6.

Appendix A

Publications

Journals

- J. Gross, H. Geerdes, H. Karl, and A. Wolisz
  "Performance Analysis of Dynamic OFDMA Systems with Inband Signaling", *IEEE Journal on Selected Areas in Communications (JSAC)*, Special Issue on 4G Wireless Systems, to appear

- J. Gross, S. Valentin, H. Karl, and A. Wolisz
  "A study of impact of inband signaling and realistic channel knowledge for an example dynamic OFDM-FDMA system", *European Transactions on Telecommunications*, vol. 16, pp. 37-49, 2005

- J. Gross, J. Klaue, H. Karl, and A. Wolisz
  "Cross-Layer Optimization of OFDM Transmission Systems for MPEG-4 Video Streaming", *Computer Communications*, vol. 27, pp. 1044-1055, 2004

Conference Proceedings

- M. Bohge, J. Gross, and A. Wolisz
  "The Potential of Dynamic Power and Sub-carrier Assignments in Multi-User OFDM-FDMA Cells", *Proc. of IEEE Globecom 2005*, St. Louis, MO, USA, November 2005

- J. Gross, H. Geerdes, and A. Wolisz
  "On the Signaling Overhead in Dynamic OFDMA Wireless Systems", *Proc. of 35. Jahrestagung der Gesellschaft für Informatik 2005*), Bonn, Germany, September 2005, "Special session: PhD Students in Graduate Colleges"

- S. Valentin, J. Gross, H. Karl, and A. Wolisz
  "Adaptive scheduling for heterogeneous traffic flows in cellular wireless OFDM-FDMA systems", *Proc. of 10th IFIP International Conference on Personal Wireless Communications / PWC'05*, Colmar, France, August 2005

- J. Gross, I. Paoluzzi, H. Karl, and A. Wolisz
  "Throughput Study for a Dynamic OFDM-FDMA System with Inband Signaling", *Proc. of Vehicular Technology Conference (VTC Spring)*, Milan, Italy, May 2004

- J. Gross, H. Karl, and A. Wolisz
  "Throughput Optimization of Dynamic OFDM-FDMA Systems with Inband Signaling", *Proc. of 2nd WiOpt'04 (Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks)*, Cambridge,United Kingdom, March 2004

- J. Gross, H. Karl, and A. Wolisz
  "On the effect of inband signaling and realistic channel knowledge on dynamic OFDM-FDMA systems", *Proc. of European Wireless 2004*, Barcelona, Spain, February 2004

- J. Gross, J. Klaue, H. Karl, and A. Wolisz
  "Subcarrier allocation for variable bit rate video streams in wireless OFDM systems", *Proc. of Vehicular Technology Conference (VTC Fall)*, Orlando, Florida,USA, October 2003

- J. Klaue, J. Gross, H. Karl, and A. Wolisz
  "Semantic-Aware Link Layer Scheduling of MPEG-4 Video Streams in Wireless Systems", *Proc. of Applications and Services in Wireless Networks (ASWN)*, Bern, Switzerland, July 2003

- J. Gross, H. Karl, F. Fitzek, and A. Wolisz
  "Comparison of heuristic and optimal subcarrier assignment algorithms", *Proc. of Intl. Conf. on Wireless Networks (ICWN)*, June 2003

- J. Gross, F. Fitzek, A. Wolisz, B. Chen, and R. Gruenheid
  "Framework for Combined Optimization of DL and Physical Layer in Mobile OFDM Systems", *Proc. of 6th Intl. OFDM-Workshop 2001*, pp. 32-1-32-5, Hamburg, Germany, September 2001

Appendix B

Acronyms

ADSL **A**symmetric **D**igital **S**ubscriber **L**ine

ARQ **A**utomatic **R**epeat **R**equest

BEP **B**it **E**rror **P**robability

BER **B**it **E**rror **R**ate

BPSK **B**inary **P**hase **S** **K**eying

BSS **B**asic **S**ervice **S**et

CDMA **C**ode **D**ivision **M**ultiple **A**ccess

CCI **C**o**c**hannel **I**nterference

CNR **C**hannel–to–**N**oise **R**atio

CSMA **C**arrier **S**ense **M**ultipe **A**ccess

CSMA/CA **C**arrier **S**ense **M**ultiple **A**ccess with **C**ollision **A**voidance

CTS **C**lear-**t**o-**S**end

DAB **D**igital **A**udio **B**roadcasting

DCF **D**istributed **C**oordination **F**unction

DFT **D**iscrete **F**ourier **T**ransform

DIFS **D**istributed **I**nter**f**rame **S**pace

DMT **D**iscrete **M**ulti-**t**one

DSL **D**igital **S**ubscriber **L**ine

DVB **D**igital **V**ideo **B**roadcasting

FCFS **F**irst–**C**ome **F**irst–**S**erve

FCS **F**rame **C**heck **S**equence

FDD **F**requency **D**ivision **D**uplex

FDMA **F**requency **D**ivision **M**ultiple **A**ccess

FDM **F**requency **D**ivision **M**ultiplexing

HF **H**igh **F**requency

GMSK **G**aussian **M**inimum **S**hift **K**eying

GSM **G**obal **S**ystem for **M**obile Communications

ICI **I**nter**c**hannel **I**nterference

IP **I**nteger **P**rogrammig

ISI **I**nter**s**ymbol **I**nterference

ITU **I**nternational **T**elecommunication **U**nion

LAN **L**ocal **A**rea **N**etwork

LOS **L**ine-**o**f-**S**ight

MAC **M**edium **A**ccess **C**ontrol

MCM **M**ulti-**c**arrier **M**odulation

MIP **M**ixed **I**nteger **P**rogramming

MOS **M**ean **O**pinion **S**core

MPDU **M**AC **P**rotocol **D**ata **U**nit

MPEG **M**oving **P**icture **E**xperts **G**roup

MSDU **M**AC **S**ervice **D**ata **U**nit

MU-OFDM **M**ulti**u**ser-**OFDM**

NAV **N**etwork **A**llocation **V**ector

NLOS **n**o **L**ine-**o**f-**S**ight

OFDM **O**rthogonal **F**requency **D**ivision **M**ultiplexing

OFDMA **O**rthogonal **F**requency **D**ivision **M**ultiple **A**ccess

PARP **P**eak-to-**A**verage **R**atio **P**roblem

PCF **P**oint **C**oordination **F**unction

PEP **P**acket **E**rror **P**robability

PER **P**acket **E**rror **R**ate

PLCP **P**hysical **L**ayer **C**onvergence **P**rocedure

PPDU **P**hysical layer **P**rotocol **D**ata **U**nit

PSDU **P**hysical layer **S**ervice **D**ata **U**nit

QAM **Q**uadrature **A**mplitude **M**odulation

QPSK **Q**uadrature **P**hase **S** **K**eying

RTS **R**eady-**t**o-**S**end

SCM **S**ingle**c**arrier **M**odulation

SEP **S**ymbol **E**rror **P**robability

SER **S**ymbol **E**rror **R**ate

SIFS **S**hort **I**nter**f**rame **S**pace

SNR **S**ignal–to–**N**oise **R**atio

SNIR **S**ignal-to-**N**oise-and-**I**nterference **R**atio

SS **S**pread **S**pectrum

TDMA **T**ime **D**ivision **M**ultiple **A**ccess

TDM **T**ime **D**ivision **M**ultiplexing

TDD **T**ime **D**ivision **D**uplex

WEP **W**ired **E**quivalent **P**rivacy

WFFQ **w**ireless **F**luid **F**air **Q**ueuing

WLAN **W**ireless **L**ocal **A**rea **N**etwork

WMAN **W**ireless **M**etropolitan **A**rea **N**etwork

WPS **W**ireless **P**acket **S**cheduling

Appendix C

Acknowledgements

This thesis has been based on the physical and moral support of a lot of people. After all, I believe that I wouldn't have been able to complete it without this help.

First of all, I would like to mention Prof. Dr.-Ing. Adam Wolisz here for his continuous, strong support during my years at the Telecommunication Networks Group. Often his advice has not been comfortable, but after all exactly this has encouraged me to work harder on many aspects.

During most of the time I collaborated with Prof. Dr. Holger Karl from the University of Paderborn, who also contributed to many issues in the thesis while always providing valuable input during many, many discussions on dynamic OFDM systems. In this context, I would also like to thank Hans-Florian Geerdes (Zuse Institute Berlin), Mathias Bohge (TU Berlin), Frank Fitzek (University of Aalborg), Ana Aguiar (TU Berlin), Jirka Klaue (TU Berlin), Tobias Harks (Zuse Institute Berlin) and Gregor Wünsch (TU Berlin) for their valuable input and for helping me by many (sometimes tough) discussions. During most of my time as Ph.D. student I was supported by the "DFG Graduiertenkolleg 621" and I would like to explicitly thank all professors related to this group and also all other Ph.D. students for their contributions to my thesis. The "Graduiertenkolleg" has been one of the major factors which lead to the moderate duration it took me to finish my thesis. Last but not lest, from the technical point of view, I would like to thank several reviewers, namely Barbara van Schewick (TU Berlin), Daniel Willkomm (TU Berlin), Lars Wischhof (TU Hamburg-Harburg), Steve Thompson (UC San Diego), and Michael Jaeger (TU Berlin) for helping me finalize the style and the wording of my thesis.

Finally, many friends and relatives have helped me throughout the time. Primarily, my wife Theresa Claus has supported me whenever necessary. Next, my parents, grandparents and my sister have offered me valuable advice as well as moral and financial support, which has been the precondition of this thesis. I am very grateful for this support. My daughter Helena gave me the final motivation to complete my thesis. Furthermore, many friends have been supportive and I would like to thank them as well.

Bibliography

[1] A. Papoulis, *Probability, Random Variables and Stochastic Processes*, McGraw-Hill Book Company, 1965.

[2] J. Proakis, *Digital Communications*, McGraw-Hill, 1995.

[3] C. Shannon, "A mathematical theory of communication," *Bell Sys. Tech. Journal*, 1948.

[4] J. Cavers, *Mobile Channel Characteristics*, chapter 1.3, Kluwer Academic, 2000.

[5] R. Steele and L. Hanzo, Eds., *Mobile Radio Communications*, J. Wiley & Sons Ltd, 2000.

[6] T. Rappaport, *Wireless Communications*, Prentice Hall, 1999.

[7] W. Jakes, *Microwave Mobile Communications*, IEEE Press, Wiley Interscience, 1994.

[8] 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 *Proc. IEEE Vehicular Technology Conference (VTC)*, 2001, vol. 1, pp. 337–341.

[9] W. Lee, *Mobile Cellular Telecommunications*, McGraw-Hill International Editions, 1995.

[10] A. Neskovic, N. Neskovic, and G. Paunovic, "Modern approaches in modeling of mobile radio systems propagation environment," IEEE Communications Surveys and Tutorials, Third Quarter 2000.

[11] J. Zander and S. Kim, *Radio Resource Managements for Wireless Networks*, Mobile Communications Series. Artech House Publishers, 2001, page 330.

[12] S. Yang, *CDMA RF System Engineering*, Mobile Communications Series. Artech House Publishers, 1998, page 19.

[13] A. Aguiar and J. Gross, "Wireless channel models," Tech. Rep. TKN-03-007, Telecommunication Networks Group, Technische Universität Berlin, Apr. 2003.

[14] R. Steele, *Mobile Radio Communications*, Pentech Press, 1992.

[15] K. Blackard, T. Rappaport, and C. Bostian, "Measurements and models of radio frequency impulsive noise for indoor wireless communications," *IEEE Journal on Selected Areas in Communications*, vol. 11, no. 7, pp. 991–1001, Sep 1993.

[16] M. Paetzold, *Mobile Fading Channels*, chapter 4.1, J. Wiley & Sons, Inc., 2002.

[17] J. Linnartz, *Special Issue on Multi-Carrier Modulation*, Number 1-2 in Wireless Personal Communications. Kluwer, 1996.

[18] R. Chang, *Orthogonal Frequency Division Multiplexing*, U.S. Patent No. 3, 488,4555, filed Nov. 14, 1966, issued Jan. 6, 1970.

[19] R. Chang, "Synthesis of band limited orthogonal signals for multichannel data transmission," *Bell Systems Technical Journal*, vol. 45, pp. 1775–1796, December 1966.

[20] R. Chang and R. Gibby, "A theoretical study of performance of an orthogonal multiplexing data transmission scheme," *IEEE Transactions on Communication Technology*, vol. COM-16, pp. 529–540, August 1968.

[21] B. Salzberg, "Performance of an efficient parallel data transmission system," *IEEE Transactions on Communication Technology*, vol. COM-15, pp. 805–811, December 1967.

[22] R. van Nee and R. Prasad, *OFDM Wireless Multimedia Communications*, Artech House, 2000.

[23] B. Hirosaki, "An Orthogonal-Multiplexed QAM System Using the Discrete Fourier Transform," *IEEE Transactions on Communication Technology*, vol. COM-29, no. 7, pp. 982–989, 1981.

[24] S. Weinstein and P. Ebert, "Data Transmission by Frequency-Division Multiplexing Using the Discrete Fourier Transform," *IEEE Transactions on Communication Technology*, vol. 19, no. 5, pp. 628–634, 1971.

[25] E. Ayanoglu, V. Jones, G. Raleigh, J. Gardner, D. Gerlach, and K. Toussi, "VOFDM Broadband Wireless Transmission and its Advantages over Single Carrier Modulation," in *Proc. IEEE Int. Conference on Communications (ICC)*, 2001.

[26] T. Schmidl and D. Cox, "Robust frequency and timing synchronisation for OFDM," *IEEE Transactions on Communications*, vol. 45, pp. 1613–1621, 1997.

[27] H. Rohling, *Lecture Notes on Mobile Communications*, TU Hamburg–Harburg.

[28] H. Schmidt and K. Kammeyer, "Reducing the peak to average power ratio of multicarrier signals by adaptive subcarrier selection," in *Proc. IEEE International Conference on Universal Personal Communications ICUPC'98*, 1998, vol. 2, pp. 933 – 937.

[29] J. Bingham, *ADSL, VDSL and Multicarrier Modulation*, J. Wiley & Sons, Inc., New York, USA, 2000.

[30] ETSI ETS 300 401, *Digital Audio Broadcasting (DAB); DAB to mobile portable and fixed receivers*, February 1995.

[31] ETSI EN 300 744, *Digital Video Broadcasting (DVB); Framing structure, channel coding, and modulation for digital terrestrial television*, August 1997.

[32] IEEE P802.11a/D7.0, *Supplement to Standard for Telecommunications and Information Exchange Between Systems - LAN/MAN Specific Requirements - Part 11: Wireless MAC and PHY Specifications: High Speed Physical Layer in the 5-GHz Band*, July 1999.

[33] ETSI, *BRAN HIPERLAN Type 2, DLC Layer, Part 1: Basic Data Transport Functions*, ts 101-761-1 edition, December 2001.

[34] IEEE 802.16-2001, *IEEE Standard for Local and Metropolitan Area Networks Part 16: Air Interface for Fixed Broadband Wireless Access Systems*, April 2002.

[35] T. Cover and J. Thomas, *Elements of Information Theory*, J. Wiley & Sons, Inc., 1991.

[36] D. Lelewer and D. Hirschberg, "Data compression," *ACM Computing Surveys*, 1987.

[37] R. Lucky, *Silicon Dreams: Information, Man and Machine*, St. Martin's Press, 1989.

[38] D. Huffmann, "A method for the construction of minimum-redundancy codes," *Proc. IRE*, pp. 1098–1101, September 1952.

[39] M. Pioro and D. Medhi, *Routing, Flow, and Capacity Design in Communication and Computer Networks*, Morgan Kaufmann, 2004.

[40] P. Varaiya, "Lecture notes on optimization," http://paleale.eecs.berkeley.edu/ varaiya/papers_ps.dir/NOO.pdf, 1998.

[41] R. Luce and H. Raiffa, *Games and Decision*, John Wiley, 1957.

[42] M. Osborne and A. Rubinstein, *A Course in Game Theory*, MIT Press, 1994.

[43] H. Kushner, *Introduction to Stochastic Control*, Holt, Rinehart and Winston, Inc., 1971.

[44] D. Bertsekas and S. Shreve, *Stochastic Optimal Control: The Discrete-Time Case*, Athena Scientific, 1997.

[45] C. Papadimitriou and K. Steiglitz, *Combinatorial Optimization*, Dover Publications, 1998.

[46] B. Korte and J. Vygen, *Combinatorial Optimization*, Springer, 2000.

[47] M. Grötschel, L. Lovász, and A. Schrijver, *Geometric algorithms and combinatorial optimization*, Springer-Verlag, 2 edition, 1993.

[48] R. Fischer and J. Huber, "A new loading algorithm for discrete multitone transmission," in *Proc. of IEEE Global Telecommunications Conference (Globecom)*, November 1996, pp. 724–728.

[49] C. Wong, R. Cheng, K. Letaief, and R. Murch, "Multiuser OFDM with adaptive subcarrier, bit and power allocation," *IEEE Journal on Selected Areas of Communications*, vol. 17, no. 10, pp. 1747–1758, October 1999.

[50] A. Jeffrey, *Mathematics for Engineers and Scientists*, Chapman & Hall, 5 edition, 1996.

[51] I. Kalet, "The multitone channel," *IEEE Transactions on Communications*, vol. 37, no. 2, pp. 119–124, February 1989.

[52] P. Chow, J. Cioffi, and J. Bingham, "A practical discrete multitone transceiver loading algorithm for data transmission over spectrally shaped channels," *IEEE Transactions on Communications*, vol. 43, no. 2, February 1995.

[53] A. Schrijver, *Combinatorial Optimization*, Springer, 2003.

[54] D. Hughes-Hartogs, *Ensemble Modem Structure for Imperfect Transmission Media*, U.S. Patents 4,679,227 (July 1987); 4,731,816 (March 1988); 4,833,706 (May 1989).

[55] J. Campello, "Optimal discrete bit loading for multicarrier modulation systems," in *Proc. IEEE Int. Symp. on Information Theory*, 1998.

[56] L. Piazzo, "Fast optimal bit-loading algorithm for adaptive OFDM systems," Tech. Rep. 002-04-03, University of Rome, 2003.

[57] S. Lai, R. Cheng, K. Letaief, and C. Tsui, "Adaptive Tracking of Optimal Bit and Power Allocation for OFDM Systems in Time-Varying Channels," in *Proc. of IEEE Wireless Communications and Networking Conference (WCNC)*, 1999, vol. 2, pp. 776 – 780.

[58] J. Campello, "Practical bit loading for DMT," in *Proc. IEEE Int. Conference on Communications (ICC)*, 1999, vol. 2, pp. 801 – 805.

[59] B. Krongold, K. Ramchandran, and D. Jones, "Computationally efficient optimal power allocation algorithms for multicarrier communication systems," *IEEE Transactions on Communications*, vol. 48, no. 1, pp. 23–27, January 2000.

[60] C. Mutti, D. Dahlhaus, T. Hunziker, and M. Foresti, "Bit and Power Loading Procedures for OFDM Systems with Bit-Interleaved Coded Modulation," in *Proc. International Conference on Telecommunications (ICT)*, 2003, vol. 2, pp. 1422 – 1427.

[61] R. Sonalkar and R. Shively, "An efficient bit-loading algorithm for DMT applications," *IEEE Communications Letters*, vol. 4, no. 3, pp. 80–82, March 2000.

[62] T. Keller and L. Hanzo, "Adaptive multicarrier modulation: A convenient framework for time-frequency processing in wireless communications," *IEEE Proceedings of the IEEE*, vol. 88, no. 5, pp. 611–640, May 2000.

[63] T. Hunziker and D. Dahlhaus, "Optimal Power Adaptation for OFDM Systems with Ideal Bit-Interleaving and Hard-Decision Decoding," in *Proc. IEEE Int. Conference on Communications (ICC)*, 2003, vol. 5, pp. 3392 – 3397.

[64] T. Willink and P. Wittke, "Optimization and performance evaluation of multicarrier transmission," *IEEE Trans. on Information Theory*, vol. 43, no. 2, pp. 426 – 440, 1997.

[65] A. Czylwik, "OFDM And Related Methods For Broadband Mobile Radio Channels," in *Proc. of Broadband Communications*, 1998, pp. 91 – 98.

[66] H. Rohling and R. Gruenheid, "Performance of an OFDM-TDMA mobile communication system," in *Proc. IEEE Vehicular Technology Conference (VTC)*, May 1996, vol. 3, pp. 1589–1593.

[67] A. Czylwik, "Adaptive OFDM for wideband radio channels," in *Proc. of IEEE Global Telecommunications Conference (Globecom)*, 1996, vol. 2, pp. 713 – 718.

[68] A. Barreto and S. Furrer, "Adaptive Bit Loading for Wireless OFDM Systems," in *Proc. IEEE Int. Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, 2001, vol. 2, pp. G–88 – G–92.

[69] D. Love and R. Heath, "OFDM power loading using limited feedback," *IEEE Transactions on Vehicular Technology*, vol. 54, no. 5, pp. 1773 – 1780, 2005.

[70] H. Nguyen, J. Brouet, V. Kumar, and T. Lestable, "Compression of Associated Signaling for Adaptive Multi-Carrier Systems," in *Proc. IEEE Vehicular Technology Conference (VTC)*, May 2004.

[71] J. Ziv and A. Lempel, "Compression of individual sequences via variable rate coding," *IEEE Transactions on Information Theory*, vol. 24, pp. 530–536, 1978.

[72] M. Lei and P. Zhang, "Subband Bit and Power Loading for Adaptive OFDM," in *Proc. IEEE Vehicular Technology Conference (VTC)*, October 2003.

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

[74] I. Kim, H. Lee, B. Kim, and Y. Lee, "On the use of linear programming for dynamic subchannel and bit allocation in multiuser OFDM," in *Proc. of IEEE Global Telecommunications Conference (Globecom)*, November 2001.

[75] M. Ergen, S. Coleri, and P. Varaiya, "QoS aware adaptive resource allocation techniques for fair scheduling in OFDMA based broadband wireless access systems," *IEEE Transactions on Broadcasting*, vol. 49, no. 4, 2003.

[76] H. Yin and H. Liu, "An efficient multiuser loading algorithm for OFDM-based broadband wireless systems," in *Proc. of IEEE Global Telecommunications Conference (Globecom)*, 2000.

[77] W. Rhee and J. Cioffi, "Increase in capacity of multiuser OFDM system using dynamic subchannel allocation," in *Proc. IEEE Vehicular Technology Conference (VTC)*, 2000, pp. 1085 – 1089.

[78] S. Pietrzyk and G. Janssen, "Multiuser subcarrier allocation for QoS provision in the OFDMA systems," in *Proc. Vehicular Technology Conference*, 2002, vol. 2, pp. 1077–1081.

[79] H. Kuhn, "The hungarian method for the assignment problem," *Naval Research Logistics Quarterly*, , no. 2, pp. 83–97, 1955.

[80] S. Lai, R. Cheng, K. Letaief, and R. Murch, "Adaptive trellis coded MQAM and power optimization for OFDM transmission," in *Proc. IEEE Vehicular Technology Conference (VTC)*, May 1999.

[81] C. Wong, C. Tsui, R. Cheng, and K. Letaief, "A real-time sub-carrier allocation scheme for multiple access downlink OFDM transmission," in *Proc. IEEE Vehicular Technology Conference (VTC)*, Sep 1999, vol. 2, pp. 1124–1128.

[82] D. Kivanc and H. Liu, "Subcarrier allocation and power control for OFDMA," in *Proc. Conference on Signals, Systems and Computers*, 2000, vol. 1, pp. 147 –151.

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

[84] H. Rohling, K. Bruninghaus, and R. Gruenheid, "Comparison of multiple access schemes for an OFDM downlink system," *Multi-Carrier Spread Spectrum*, pp. 22–30, 1997, K. Fazel and G. Fettweis, Eds., Kluwer.

[85] H. Kim, J. Kwak, J. Choi, and J. Lee, "Efficient subcarrier and bit allocation algorithm for OFDMA system with adaptive modulation," in *Proc. IEEE Vehicular Technology Conference (VTC)*, May 2004.

[86] S. Pfletschinger, G. Muenz, and J. Speidel, "Efficient subcarrier allocation for multiple access in OFDM systems," in *Proc. of the 7th Int. OFDM Workshop, Hamburg, Germany*, Sep 2002.

[87] G. Muenz, S. Pfletschinger, and J. Speidel, "An efficient waterfilling algorithm for multiple access OFDM," in *Proc. of IEEE Global Telecommunications Conference (Globecom)*, 2002.

[88] Y. Zhang and K. Letaief, "Adaptive resource allocation and scheduling for multiuser packet-based OFDM networks," in *Proc. IEEE Int. Conference on Communications (ICC)*, Paris, France, 2004.

[89] L. Xiaowen and Z. Jinkang, "An adaptive subcarrier allocation algorithm for multiuser OFDM system," in *Proc. IEEE Vehicular Technology Conference (VTC)*, October 2003.

[90] Z. Shen, J. Andrews, and B. Evans, "Optimal power allocation in multiuser OFDM systems," in *Proc. of IEEE Global Telecommunications Conference (Globecom)*, 2003, vol. 1, pp. 337 – 341.

[91] C. Mohanram and S. Bhashyam, "A Suboptimal Joint Subcarrier and Power Allocation Algorithm for Multiuser OFDM," *IEEE Communications Letters*, vol. 9, no. 8, pp. 685–687, 2005.

[92] R. Gruenheid and H. Rohling, "Performance comparison of different multiple access schemes for the downlink of an OFDM communication system," in *Proc. IEEE Vehicular Technology Conference (VTC)*, May 1997, vol. 3, pp. 1365 –1369.

[93] W. Wang, T. Ottosson, M. Sternad, A. Ahlen, and A. Svensson, "Impact of multiuser diversity and channel variability on adaptive OFDM," in *Proc. IEEE Vehicular Technology Conference (VTC)*, October 2003.

[94] T. Alen, A. Madhukumar, and F. Chin, "Capacity enhancement of a multi-user OFDM system using dynamic frequency allocation," in *Proc. IEEE Vehicular Technology Conference (VTC)*, October 2003.

[95] A. Aguiar and J. Klaue, "Bi-directional WLAN channel measurements in WLAN environments with mobility," Tech. Rep. TKN-04-002, Telecommunication Networks Group, Technische Universität Berlin, 2004.

[96] S.A. ILOG, *ILOG CPLEX 9.0 - User's Manual*, Paris, France, 2004.

[97] T. Koch, *ZIMPL User Guide 2.02*, Zuse Institute Berlin, May 2004.

[98] ITU-R Recommendation BT.500-10, "Methodology for the subjective assessment of the quality of television pictures," March 2000.

[99] J. Klaue, B. Rathke, and A. Wolisz, "EvalVid - A Framework for Video Transmission and Quality Evaluation," in *In Proc. of 13th International Conference on Modelling Techniques and Tools for Computer Performance Evaluation*, 2003.

[100] J. Gross, H. Karl, F. Fitzek, and A. Wolisz, "Comparison of heuristic and optimal subcarrier assignment algorithms," in *Proc. Int. Conference on Wireless Networks (ICWN)*, June 2003.

[101] J. Klaue, J. Gross, H. Karl, and A. Wolisz, "Semantic-aware link layer scheduling of mpeg-4 video streams in wireless systems," in *In Proc. of Applications and Services in Wireless Networks (ASWN)*, 2003.

[102] J. Gross and H. Karl, "Comparison of different fairness approaches in wireless OFDM-FDMA cells," Tech. Rep. TKN-04-003, Telecommunication Networks Group, Technische Universität Berlin, 2004.

[103] A. Tanenbaum, *Computer Networks*, Prentice Hall PTR, 1996.

[104] R. Jain, *The art of computer systems performance analysis*, J. Wiley & Sons, Inc., 1991.

[105] J. Bentley, D. Sleator, R. Tarjan, and V. Wei, "A locally adaptive data compression scheme," *Communication of the ACM*, vol. 29, no. 4, pp. 320–330, 1986.

[106] J. Hester and D. Hirschberg, "Self-organizing linear search," *ACM Computer Survey*, 1985.

[107] P. Elias, "Universal codeword sets and representations of the integers," *IEEE Trans. Inform. Theory*, vol. 21, no. 2, pp. 194–203, 1975.

[108] "IEEE 802.11-1999 – Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications," 1999.

[109] D. Qiao and S. Choi, "Goodput enhancement of IEEE 802.11a wireless LAN via link adaptation," in *Proc. IEEE Int. Conference on Communications (ICC)*, 2001, vol. 7, pp. 1995 – 2000.

[110] D. Qiao, S. Choi, and K. Shin, "Goodput analysis and link adaptation for IEEE 802.11a wireless LANs," *IEEE Transactions on Mobile Computing*, vol. 1, no. 4, pp. 278 – 292, 2002.