System Level Simulation and Optimization of Multi-User MIMO Transmissions

Master's Thesis

for obtaining the academic degree

Diplom-Ingenieur

in the master's degree program

Telecommunications

carried out by

Alexander Bokor

matriculation number: 01608081

Faculty of Electrical Engineering and Information Technology
Institute of Telecommunications
at TU Wien

Supervision:
Associate Prof. Dipl.-Ing. Dr.techn. Stefan Schwarz
Univ.Prof. Dipl.-Ing. Dr.techn. Markus Rupp

Statement on Academic Integrity

Hiermit erkläre ich, dass die vorliegende Arbeit gemäß dem Code of Conduct – Regeln zur Sicherung guter wissenschaftlicher Praxis (in der aktuellen Fassung des jeweiligen Mitteilungsblattes der TU Wien), insbesondere ohne unzulässige Hilfe Dritter und ohne Benutzung anderer als der angegebenen Hilfsmittel, angefertigt wurde. Die aus anderen Quellen direkt oder indirekt übernommenen Daten und Konzepte sind unter Angabe der Quelle gekennzeichnet. Die Arbeit wurde bisher weder im In– noch im Ausland in gleicher oder in ähnlicher Form in anderen Prüfungsverfahren vorgelegt.

Vienna, September 2022 _____

Alexander Bokor

Abstract

Providing higher data rates in mobile communication systems without increasing bandwidth can be achieved with techniques such as multi-user multiple-input and multiple-output transmissions using linear precoding. One popular precoding strategy that is often used in practice is zero-forcing precoding, which utilises multiple antennas to cancel the interference between users. The performance of zero-forcing precoding strongly depends on the similarities of the channels of scheduled users. Therefore, a user grouping heuristic is proposed that optimises the transmission in terms of sum-throughput. In addition, an extension is introduced to trade-off throughput against fairness among users. The proposed algorithm is compared to a semi-orthogonal user selection algorithm by performing system level simulations. The performance of the algorithms is also investigated in a cellular scenario together with fractional frequency reuse.

Contents

1. Introduction **6**
- 1.1. Motivation . . . 6
- 1.2. State of the Art . . . 7
- 1.3. Structure of the Work . . . 8
- 1.4. Notation . . . 9

2. System Level Simulation **10**
- 2.1. Methodology . . . 10
- 2.2. Channel Model . . . 12
- 2.3. Channel Estimation and Feedback . . . 13
- 2.4. Base Station Handover and Scheduler . . . 14
- 2.5. Link Quality Model . . . 14
- 2.6. Link Performance Model . . . 16
- 2.7. Beamforming and Precoding . . . 16
- 2.8. Performance Metrics . . . 17

3. Proposed Multi-User Link Quality Model **19**

4. MU-MIMO User Grouping **25**
- 4.1. System Model . . . 25
- 4.2. Semi-Orthogonal User Selection . . . 29
- 4.3. Single Path Random Sampling . . . 33
- 4.4. SPRS Fairness Extension . . . 35
- 4.5. Performance of SUS and SPRS . . . 35
- 4.6. SPRS Compared to the Exhaustive Search . . . 36
- 4.7. Performance of SPRS Fairness Extension . . . 39
- 4.8. SPRS Algorithm Complexity . . . 41

5. Cellular MU-MIMO Systems **44**
- 5.1. System Model . . . 44
- 5.2. Fractional Frequency Reuse . . . 45
- 5.3. SUS and SPRS in a Cellular Scenario . . . 48
- 5.4. Performance with FFR . . . 49

6. Conclusion and Outlook **56**

A. Abbreviations 58

B. References 60

C. List of Figures 63

D. List of Tables 64

1. Introduction

In this chapter the motivation behind this work is discussed and the current state of the art is presented. Furthermore, the structure of the thesis and the mathematical notation is established.

1.1. Motivation

Due to the growth of mobile networks and the rapid evolution of technologies such as online conferences or video streaming that initiate an ever increasing demand for more data traffic, the necessity for more throughput in mobile cells arises. In [1], the authors expect that in 2022 the average traffic usage per smartphone will surpass 15 GB. Just from an economic standpoint, the classical way of increasing throughput, namely by adding more bandwidth, is connected with acquiring expensive spectrum licences. Furthermore, today there is not much unoccupied spectrum left at the frequency bands below 6 GHz, which are useful for large-area coverage. Therefore, finding other ways to push spectral efficiency without changing the bandwidth are an everlasting topic of interest.

Adding more antennas to transmitters and receivers has the potential to significantly increase throughput and reliability without additional bandwidth or power [2][3, p. 445]. It has been shown that the maximum achievable capacity for a Rayleigh-fading Gaussian channel with multiple antennas can be calculated and is depended on the number of antennas at the transmitter and the receiver side [4]. To achieve the maximum capacity, the receiver is required to have at least as many antennas as the transmitter. In terms of a mobile communication system, this presents a problem. While it is feasible to increase the number of antennas at the base station, it is challenging to add antennas at the user side. The reason for this is the limited size of a modern cell phone that imposes challenging device design constraints. Furthermore, placing antennas in close proximity to each other increases channel correlation, which decreases diversity and throughput [5][3, p. 253]. The limited number of antennas at the user limits the total throughput.

Instead of transmitting to a single user over multiple antennas, MIMO systems can be used to serve multiple users at the same time. One technique is called space-division multiple access (SDMA). It uses the available degrees of freedom of the multiple-input and multiple-output (MIMO) system to simultaneously send

symbols to multiple users at once, separated in spatial data streams [6]. In such a system, the number of users served is determined by the number of transmit antennas. SDMA allows to achieve a spatial multiplexing gain, even if each user is only equipped with a single receive antenna. By serving multiple users at once, the total throughput can be increased by serving more users instead of adding more receiver antennas at a single user.

Of course, increasing throughput with more antennas does not come free, it requires signal processing techniques to leverage the gain. In terms of that, it has been shown that dirty paper precoding (DPC) achieves the capacity of the MIMO broadcast channel [7], [8]. However, DPC is challenging to implement, since it requires non-causal knowledge about interference at the receiver and is computationally infeasible in practise. For multi-user systems there exist a sub-optimum technique for the case of single antenna users called zero-forcing dirty-paper coding that achieves asymptotically capacity with a growing number of users [9].

A reasonable trade-off for multi-user systems is linear precoding, where the signal is filtered by a linear transformation. A popular method is zero-forcing (ZF) precoding, where interference among users in the same cell is cancelled. Although it is in general suboptimal, ZF precoding has been shown to achieve asymptotically capacity with a growing number users [10] and is feasible to implement. In such a ZF system, the system throughput depends on the group of scheduled users, more specifically, on the similarity of their channels. A poorly selected user can decrease the performance of all other participants. Therefore, in addition to the regular decisions that a scheduler has to make, user grouping must be performed.

Finding the optimal user grouping is a nontrivial problem and requires significant computational effort, especially for a large number of users and transmit antennas. In fact, the optimal choice leads to a combinatorial optimisation problem of exponential complexity [11].

In this work, a novel user grouping algorithm for ZF precoding is introduced that optimises the communication system in terms of sum-throughput. The algorithm is designed to account for fairness and provide a mechanism for balancing complexity with throughput. The performance of the proposed algorithm is compared to the semi-orthogonal user selection (SUS) algorithm proposed in [10] and investigated by simulations with the Vienna 5G System Level Simulator [12]. To perform this, a system level abstraction model for multi-user transmissions is proposed.

1.2. State of the Art

Various user grouping algorithms were already proposed and classified in literature. In [13], the authors distinguish between two groups, algorithms that optimise

towards the sum-throughput and those that are based on the channel correlation and the Frobenius norm.

In [14], the authors propose a framework called G-Greedy for general greedy user grouping algorithms. They showed that, with an optimum parameter choice, such algorithms are capable of achieving capacity in the asymptotic case, where the number of users approaches infinity.

The SUS algorithm proposed in [10] can be formulated in the aforementioned G-Greedy framework and therefore achieves asymptotic capacity, while keeping computational complexity low. Here, users are selected on the basis of their channel correlation and their channel magnitude.

Other approaches, such as zero-forcing with selection (ZFS) greedily add users that maximise the group sum-rate and stop if no increase is possible [15]. In [13], the authors addresses the problem of "redundant" users. This problem occurs due to the greedy nature of ZFS, where at the end it can be beneficial to remove a user previously scheduled to further increase the sum-rate. The authors propose a modified algorithm called greedy user selection with swap (GUSS), where users can also be deleted or swapped. For DPC systems there are also capacity greedy algorithms, called CGUS, as shown in [16] and [15].

In [17], the authors propose a CGUS algorithm for zero-forcing beamforming that aims to lower the computational complexity by approximating the precoder calculation with a less complex algorithm.

The single path random sampling (SPRS) algorithm proposed in this thesis, can be classified as an algorithm that optimises towards the sum-throughput. Like many of the mentioned algorithms it is also greedy, but in contrast it additionally performs a search reduction by introducing a random sampling step to reduce the number of possible candidates. In addition it can also optimise towards fairness to achieve best-rate, proportional fair (PF), or max-min scheduling.

1.3. Structure of the Work

The work is structured as follows: In Chapter 2 the concept of system level simulation is introduced as it is the base for all simulations in this work. The basic simulation flow is described briefly, and features important for this thesis are outlined in more detail. In Chapter 3 a multi-user MIMO extension for the simulator's existing link quality model is proposed. In Chapter 4 a novel user grouping algorithm and an extension for fairness are presented, together with a performance evaluation by simulation. In Chapter 5 the scenario is extended to a cellular system and the performance of the user grouping algorithms is investigated in a fractional frequency reuse scenario. Finally, in Chapter 6 the conclusion and an outlook is presented.

1.4. Notation

x	...	scalar
$\|x\|$	...	absolute value
$\boldsymbol{x}$	...	column vector
$\mathbf{X}$	...	matrix
$[\mathbf{X}]_{mn}$	...	element of matrix $\mathbf{X}$ in the m-th row and n-th column
$\boldsymbol{x}^H$	...	conjugate transpose
$\mathbf{X}^+$	...	Moore-Penrose pseudo-inverse
$\|\boldsymbol{x}\|$	...	Euclidian norm
$\|\boldsymbol{x}\|_p$	...	p-norm
$\mathbb{E}\{X\}$	...	expectation of random variable X
$\delta_{i,j}$	...	Kronecker delta

2. System Level Simulation

A mobile communication system is defined by parameters that are either related to hardware and software specifications, or to the environment, such as the number of connected users, distances, antennas or the behaviour of the wireless channel. Our goal is to assess the performance of a system by obtaining the expected throughput for each user. With such performance metric it is possible to compare different algorithms, techniques and scenarios.

Unfortunately, due to the complexity of big systems it is infeasible to find an analytical solution and we therefore rely on simulations. Such simulation is a computational heavy problem and detailed simulations, that compute in acceptable time, are only feasible for a limited amount of users and base stations. Hence, a common technique is to rely on system level simulations, that abstract low level communication processes, to simulate larger networks in less time.

The question arises which simulator shall be used and therefore a brief overview of available simulators is presented. The Vienna 5G System Level Simulator [12] is based on object oriented MATLAB and free for academic use. The WiSE simulator [18] is implemented in C++. Also the MATLAB 5G Toolbox offers support for system level simulations. The Simu5G [19] simulator is based on the OMNeT++ [20] discret event simulator. Due to the authors previous knowledge the Vienna 5G System Level was selected and extended with new models and algorithms. Explaining the entire simulator in this thesis would go beyond the scope of the work and is already handled in [12]. Therefore, this sections briefly introduces the main concepts and the parts used to obtain simulation results.

2.1. Methodology

Simulations are based on the Monte Carlo approach by simulating over a large number of random realisations. Based on the law of large numbers the sample mean of the simulation results will approximate the expectation of the throughput if the number of samples is sufficiently large.

The simulation is performed in a time unit called slots, and for each slot throughput for each user is calculated. One slot is equivalent to a fifth generation (5G) sub-frame, assuming the default numerology. To discuss this time unit we first introduce the radio access. The 5G radio access uses orthogonal frequency-division

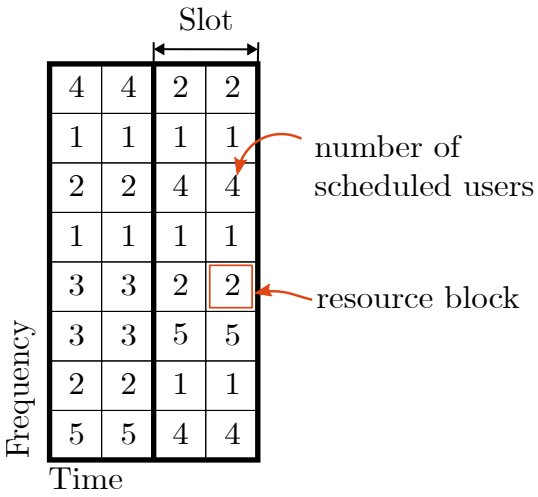


Figure 2.1.: Illustration of the resource grid. The radio resources are sliced in time and frequency and assigned by a scheduler. The smallest elements are the resource blocks. Two resource blocks form a slot, which is indicated by a thick line. With multi-user transmissions resource blocks are shared by multiple users.

multiplexing (OFDM) modulation. Here, a wideband channel is divided into many orthogonal small band channels. It is hence possible to divide radio resources in time and frequency which is called the resource grid. Since the users share the channel, a scheduler has to manage the wireless access of the users. This is done by slicing the grid into resource blocks and assigning them to users as shown in Fig. 2.1. There are multiple possible resource grid configurations, depending on the 5G numerology parameter. Here, a resource grid is assumed as shown in Fig. 2.1, with two resource blocks per slot. Hence, this grid defines some important time scales for the simulation. The smallest time period corresponds to the duration of one resource block (RB) which is 0.5 ms.

We introduce the concept of slots and segments as shown in Fig. 2.2 due to the time scale of the channel. Wireless channels are fluctuating and described by rapid changes called the small-scale fading and slow changes, called large-scale fading. We assume that the small-scale fading is much slower than the duration of a RB. This time during which the the small-scale fading is constant is called a slot and always contains two resource blocks. Multiple slots during which the macroscopic fading is constant form a segment together.

Finally, the largest time scale is the chunk. It is assumed that enough time has passed that user positions are uncorrelated between chunks.

In the system level abstraction, the transmission between a user and base station is not modelled in every detail, but abstracted by two models. The first one assess

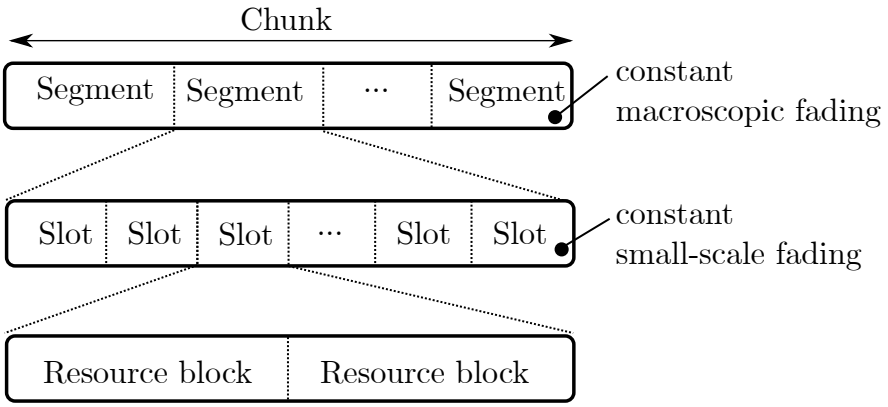


Figure 2.2.: Time line of the simulator. Chunks contain segments during which the macroscopic fading is constant. The segment consists out of various slots. The concept of slots and segments is introduced to handle times of constant small-scale and large-scale fading effects. The smallest unit is the resource block.

the link quality of the user and is called link quality model (LQM). The second one maps the link quality to a throughput and a block error rate (BLER) value and is called link performance model (LPM). The individual parts are described in the following sections.

2.2. Channel Model

The wireless channels between users and base stations are described by path loss, small-scale fading and large-scale fading. In addition, the antenna characteristics also play a role, but in scope of this work uniform antenna patterns were used.

The free space path loss (FSPL) model is used to describe the attenuation of the signal and is calculated as

$$\mathrm{FSPL}(d, \lambda) = -20 \log\left(\frac{\lambda}{4\pi d}\right), \tag{2.1}$$

where d is the distance between transmitter and receiver and λ the wave length.

The centre frequency in this work is 2 GHz and hence, λ is 155 mm.

Due to multipath propagation, the channel is frequency selective and modelled by a power delay profile. The 3GPP PedA channel model as described in [21] is used as the simulations focus mainly on the profile of slow moving pedestrians, since multi-user MIMO relies on accurate channel state information (CSI) at the transmitter to be effective and this is only achievable if the channel varies slowly,

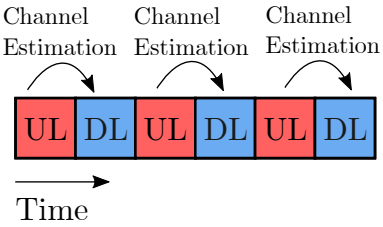


Figure 2.3.: Downlink and uplink slots are consecutive. The uplink is used to estimate the downlink channel.

as will be discussed in Section 2.3. Due to the movement of users and obstacles the channel is also time-variant. To obtain time correlation with correct statistical properties, the Rosa Zheng model is used as described in [22].

2.3. Channel Estimation and Feedback

The base station relies on accurate channel estimation and channel quality indicator (CQI) feedback from the user to select an appropriate modulation and coding scheme and to perform the user grouping. To obtain accurate CSI we assume a time division duplex (TDD) system in which the uplink and downlink slots are consecutive, as shown in Fig. 2.3. With the pilot information from the uplink, the base station can estimate the downlink channel without any additional feedback channel. Since the focus of the work is on user grouping algorithms, we assume perfect CSI knowledge. To fulfil that assumption, the channel must remain constant during uplink and downlink slots.

In addition, a limited feedback channel is required and if a user is scheduled, CQI feedback is sent to the base station to determine the modulation and coding scheme used for the next transmission. The feedback mechanism is necessary, since the signal to interference and noise ratio (SINR) is not known on the side of the base station and CSI is not sufficient to recover it. Therefore, the user grouping should not change too often, otherwise only outdated CQI information is available at the base station, since the user grouping influences the SINR. Hence, the scheduler holds the grouping constant for some number of slots.

To determine this number we require knowledge of the channel coherence time, since after some time the channel will change and the grouping will not match to the current conditions. The simulations focus on pedestrian users, but to cover the worst case, we require that the system works for users that move with a maximum speed of $100\,\mathrm{km\,h^{-1}}$. The resulting maximum Doppler spread is

$$f_\mathrm{D} = \frac{f_c v}{c_0} = 185.31\,\mathrm{Hz}, \tag{2.2}$$

where f_c is the centre frequency of 2 GHz, v is the user velocity, and c_0 the speed of light. A rough estimate of the coherence time is given by the uncertainty relation from [23] as

$$T_C \leq \frac{1}{f_D} = 5.4\,\text{ms}. \tag{2.3}$$

Since one slot has a duration of 1 ms, we can keep the grouping constant for 5 slots, which is used as the default value for the simulations. Under this assumptions the channel coherence is sufficient for accurate channel estimation and CQI feedback.

2.4. Base Station Handover and Scheduler

Users connect to base stations based on the strength of the received signal. Once connected, the base station scheduler allocates resource blocks to a user, which are used for data transmission. We distinguish between single user and multi user transmissions, where in the former only one user can be assigned to a resource block, and in the latter several users can be assigned to the same resource block. Additionally, the scheduler selects the modulation and coding schemes for each user based on their CQI feedback.

As mentioned in the introduction, and also explained in more detail in Chapter 4, it is important which users are scheduled together in a multi-user transmission. This is the task of a user grouping algorithm, which relies on accurate CSI of the users, that is obtained by the channel estimation discussed in Section 2.3.

To summarise this section the simulation flow is depicted from a scheduler perspective in Fig. 2.4. Here, we see the user grouping algorithm, the scheduler, the precoding, and its connection to the link quality and link performance model based on perfect CSI from the uplink channel and the CQI feedback.

2.5. Link Quality Model

The LQM abstraction calculates the user's SINR values based on environment parameters. In Fig. 2.5 a schematic overview of the LQM is shown. The inputs are path loss, antenna gain and the small scale fading channel and scheduling decisions such as precoding and power allocation. The output of the LQM is the post equalisation SINR for each resource block of the user. The LQM is only briefly discussed here since it's extension for multi-user transmissions is derived in Chapter 3.

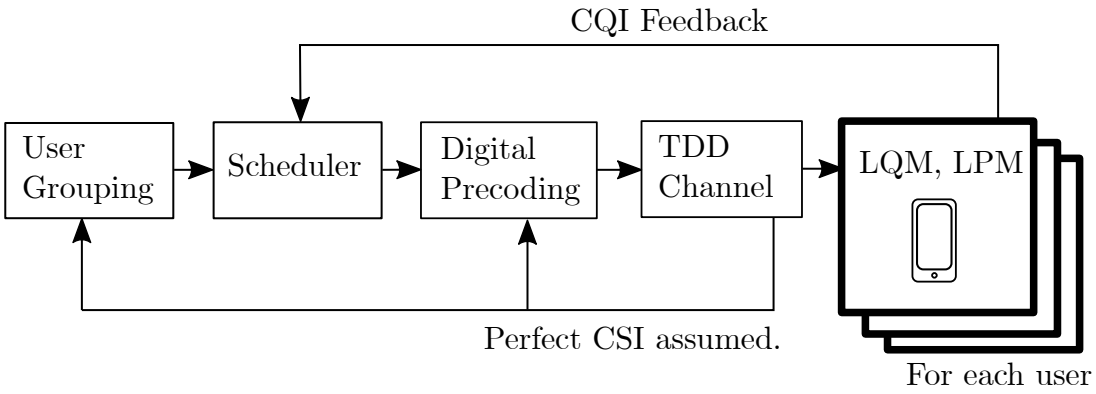


Figure 2.4.: Shows the simulation flow from a scheduler perspective. Users provide channel quality indicator (CQI) feedback to the scheduler. For user grouping and precoding perfect channel state information (CSI) is assumed.

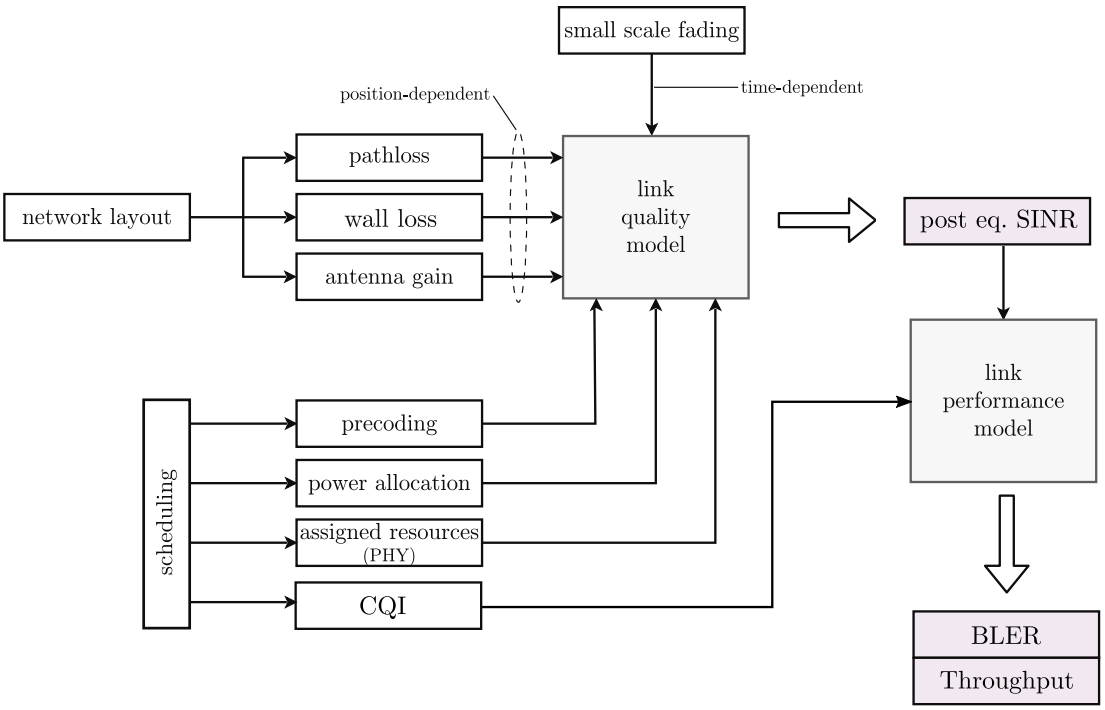


Figure 2.5.: Representation of the data flow between link quality and link performance model. The link quality model aggregates the time-dependent and position-dependent parameters together with the scheduling decisions and calculates a post equalisation SINR for each RB. The post equalisation SINR and the CQI set by the scheduler gets mapped to BLER and throughput values. Figure reproduced from [24].

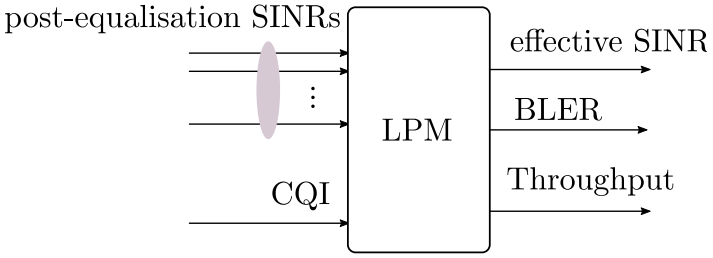


Figure 2.6.: Illustration of the link performance model The post-equalisation SINR values from the link quality model, together with a CQI value is mapped to a single effective SINR value. This single value is used to determine a BLER and throughput value for the user.

2.6. Link Performance Model

In the LPM the post-equalisation SINR values and the user CQI are assigned to a single throughput value as shown in Fig. 2.6. The CQI is set by the scheduler and depends on the user feedback. An effective SINR mapping (ESM) technique is used to compress the resource block post-equalisation SINR values and the CQI to a single effective SINR value. By performing mutual information effective SINR mapping (MIESM) [25] the resulting effective SINR value is mapped to a BLER value using an additive white Gaussian noise (AWGN) performance curve.

The BLER value is used to calculate the user throughput. The LPM operates in two modes: either the CQI value set by the scheduler is used for the calculation of the throughput or the throughput for the ideal CQI value is calculated. Therefore, the simulator always produces throughput results for the best CQI case and the case where the CQI is set by scheduler. Since this is dependent on the user feedback, it is called the feedback case.

2.7. Beamforming and Precoding

Beamforming and precoding are two techniques to steer the beams of antenna arrays. Figure 2.7 shows an antenna that steers multiple beams to groups of users. Since the beams are spatially separated, it is possible to transmit to these users at the same time. This principle is called SDMA. The more spatial separation is achieved, the less interference will impair the channel quality of other users.

In general, we distinguish between analog and digital beamforming as depicted in Fig. 2.8. With analog beamforming a single radio frequency (RF) chain supplies the antennas. Furthermore, each antenna is equipped with an analogue phase shifter that steers the overall beam in a specific direction. In contrast to that, in digital beamforming each antenna is driven by an independent RF chain and

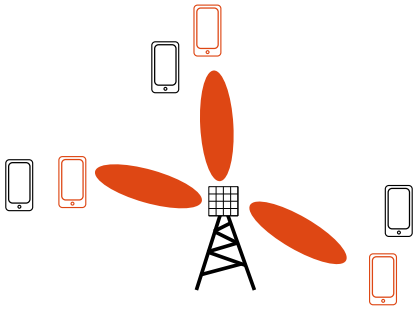


Figure 2.7.: With Space-division multiple access the base station steers radio beams to users. Therefore, multiple users can be served simultaneously. For zero-forcing precoding the situation is more complex, since beams are formed such that multi-path components add up destructively to eliminate interference. Therefore, this picture should be treated as simplistic illustration.

the signal processing is performed in the base band. Compared to the analogue case, it is possible to form complicated beam patterns and therefore apply more advanced signal processing techniques such as ZF beamforming. The scope of this work is purely on digital beamforming.

Another distinction is made with respect to the terms beamforming and precoding. Beamforming is a name that applies if each user has a single antenna. The name precoding is used if users are equipped with more than one antenna. This allows to serve them with more than one spatial data stream. The scope of this work is mainly on beamforming, since users in the simulations have one receive antenna.

2.8. Performance Metrics

This section defines metrics used later in the simulations, since the simulator produces BLER and throughput values on a slot basis and some derived metrics are used to analyse the results.

The user throughput is the number of bits that a user transmitted in a time interval. The user throughput for user k in slot n is defined as $t_{k,n}$. If we assume K users and N simulated slots, the sum over the individual user throughput in a slot n is called the slot sum-throughput and expressed as

$$T_{\text{slot},n} = \sum_{k=1}^{K} t_{k,n}, \tag{2.4}$$

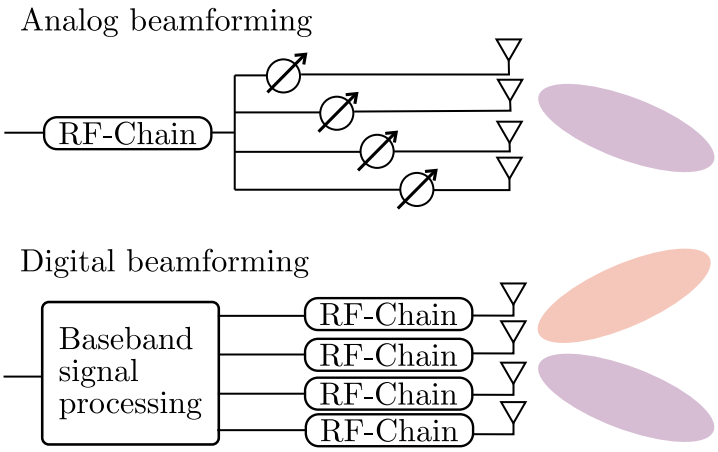


Figure 2.8.: Analog and digital beamforming. In analog beamforming one RF-chain is shared by all antennas and phase shifters are used to steer the beam. In contrast, digital beamforming spends one RF-chain per antenna and therefore allows to steer the beams by baseband signal processing. The beam patterns are a simplistic illustration.

and the total throughput of a user averaged over slots is called the user sum-throughput

$$T_{\text{user},k} = \frac{1}{N} \sum_{n=1}^{N} t_{k,n}. \tag{2.5}$$

The average sum-throughput is the total number of bits transmitted over the simulated time and expressed as

$$T_{\text{avg,sum}} = \frac{1}{N} \sum_{n=1}^{N} T_{\text{slot},n}. \tag{2.6}$$

Fairness is measured with the Jain's fairness index. The fairness in a slot is calculated by the expression

$$\mathcal{J}_n = \frac{(\sum_{k=1}^{K} t_{k,n})^2}{\sum_{k=1}^{K} t_{k,n}^2}. \tag{2.7}$$

Here, the maximum fairness of 1 is achieved if all $t_{k,n}$ have the same value. The minimum value of $1/K$ occurs if all but one user have zero throughput.

3. Proposed Multi-User Link Quality Model

The task of the LQM is to assess the signal quality in terms of a post-equalisation SINR. This is the ratio of the intended signal power to noise and interference after the receiver's equaliser. In this chapter, a LQM for multi-user MIMO transmissions is proposed.

The Vienna 5G System Level Simulator uses as LQM for single-user transmissions. The derivation of this model can be found in [24, p. 28], while this section extends the model for multi-user transmissions. Although this thesis investigates in users with single-antennas, the proposed model is formulated for an arbitrary number of user antennas.

An example of how interference can occur in a wireless network is depicted in Fig. 3.1, where a user receives the intended signal and interference from other base stations. To keep indices simple the user of interest k is connected to base station number 1 and gets interference from M base stations. Each base station has several users connected and the user indices of base station i are contained in the set $\mathcal{G}_i$. The channel between the user of interest k and base station i is denoted as $\mathbf{H}_i$. The digital precoder for any user g at base station i is $\mathbf{F}_{g,i}$. The symbols for any user g connected to a base station i are $\boldsymbol{s}_{g,i}$. The user of interest k utilises a receive filter which is described by the matrix $\mathbf{R}_k$. In addition, the user receives noise denoted as $\boldsymbol{z}_k$.

The received signal for the user of interest k results in a superposition of signals originating from the connected base station and all the other interfering base stations, at it is depicted in Fig. 3.2. This is mathematically expressed as

$$\boldsymbol{y}_k = \mathbf{R}_k \mathbf{H}_1^H \sum_{g \in \mathcal{G}_1} \mathbf{F}_{g,1} \boldsymbol{s}_{g,1} + \mathbf{R}_k \sum_{m=2}^{M+1} \mathbf{H}_m^H \sum_{g' \in \mathcal{G}_m} \mathbf{F}_{g',m} \boldsymbol{s}_{g',m} + \mathbf{R}_k \boldsymbol{z}_k. \tag{3.1}$$

The channel matrices $\mathbf{H}_i$ are defined in a conjugate transpose manner to keep consistency with Chapter 4. For sake of a simple notation the matrix dimensions are not written explicitly, since they depend on the number of user antennas and

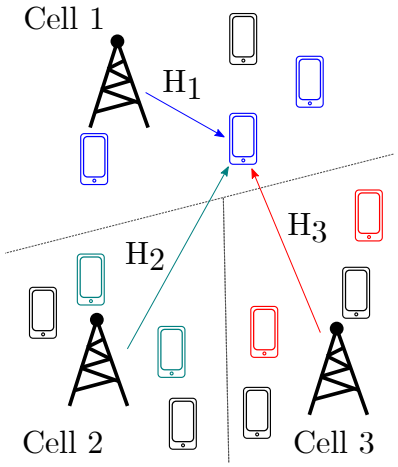


Figure 3.1.: Interference in a scenario with multiple cells. The coloured users are served simultaneously by their respective base stations. The blue users experiences interference from users in his home cell, and all users scheduled by other cells.

base station antennas, but are indicated as:

$$\mathbf{F}_k \in \mathbb{C}^{\text{nTxAntennas}\times\text{nSymbols}}$$
$$\mathbf{H}_k \in \mathbb{C}^{\text{nTxAntennas}\times\text{nRxAntennas}}$$
$$\mathbf{R}_k \in \mathbb{C}^{\text{nSymbols}\times\text{nRxAntennas}}$$

The signal of interest is at $g = k$, all other received symbols are interference. Therefore, the first term of Eq. (3.1) is split up into two parts and the received signal is written as

$$\boldsymbol{y}_k = \mathbf{R}_k\mathbf{H}_1^H\mathbf{F}_{k,1}\boldsymbol{s}_{k,1} + \mathbf{R}_k\mathbf{H}_1^H \sum_{g\in\mathcal{G}_1, g\neq k} \mathbf{F}_{g,1}\boldsymbol{s}_{g,1} + \tag{3.2}$$

$$+ \mathbf{R}_k \sum_{m=2}^{M+1} \mathbf{H}_m^H \sum_{g'\in\mathcal{G}_m} \mathbf{F}_{g',m}\boldsymbol{s}_{g',m} + \mathbf{R}_k\boldsymbol{z}_k. \tag{3.3}$$

Our goal is to calculate the SINR for which we will treat the symbols as well as the noise as random variables and calculate the expectation of the received signal's power. We assume that the user symbols are independent to each other and to the noise. In addition, the noise is assumed to be zero mean.

$$\mathbb{E}\{\boldsymbol{s}_{i,m}\boldsymbol{s}_{j,n}^H\} = \mathbf{0}, \quad i \neq j, \forall m, n \tag{3.4}$$
$$\mathbb{E}\{\boldsymbol{s}_{i,m}\boldsymbol{z}_k^H\} = \mathbf{0} \tag{3.5}$$
$$\mathbb{E}\{\boldsymbol{z}_k\boldsymbol{z}_k^H\} = \mathbf{I}\,\sigma_z^2 \tag{3.6}$$

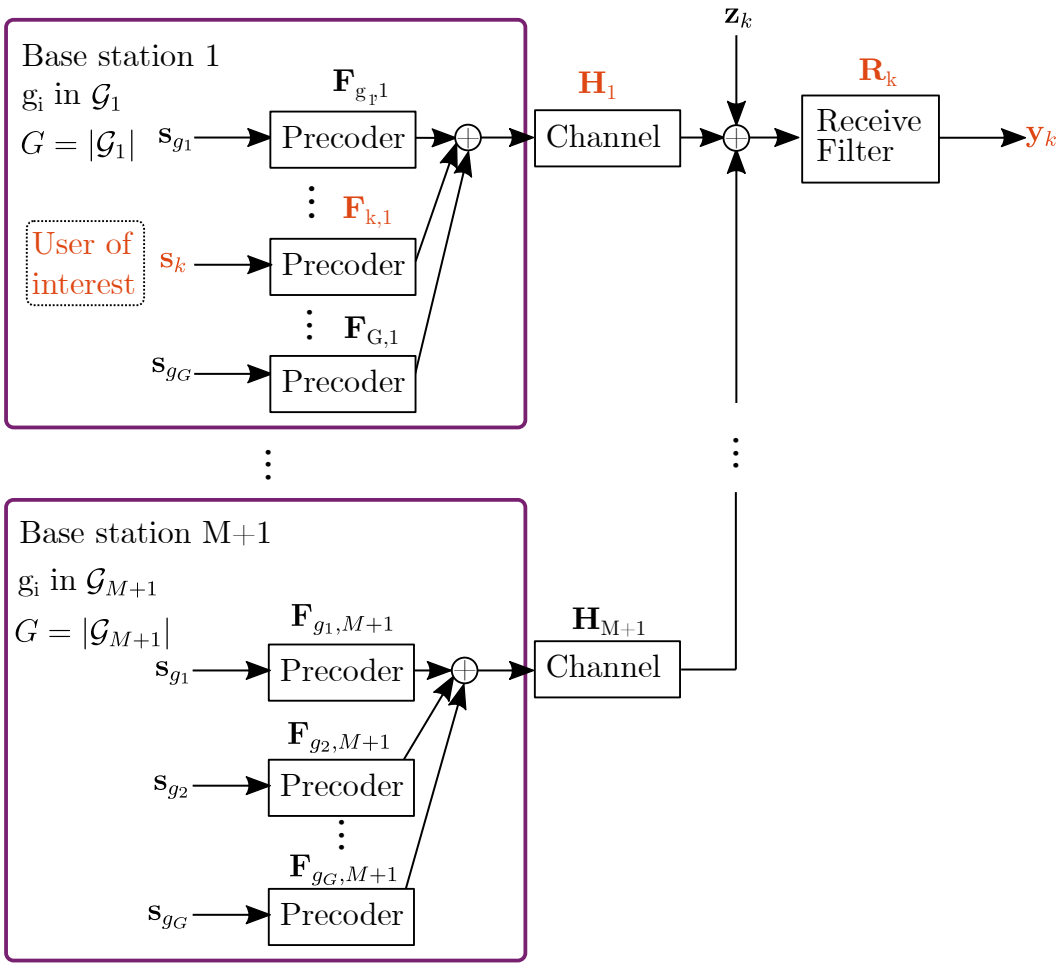


Figure 3.2.: Block diagram for the calculation of the received symbols. Each base station has to serves multiple users. The user of interest is denoted with k and is highlighted in orange to emphasise that his symbols are of interest. All other users are denoted by arbitrary index values s_{g_x}, where g_x comes from the set $\mathcal{G}_i$. It is assumed that the user k is connected to base station number 1. The received signal is a superposition of all base station signals, including each user's precoded symbols.

To obtain the expectation of the received symbols power the correlation matrix of vector $\boldsymbol{y}_k$ is calculated and expressed as

$$\mathbb{E}\{\boldsymbol{y}_k\boldsymbol{y}_k^H\} = (\mathbf{R}_k\mathbf{H}_1^H\mathbf{F}_{k,1})\,\mathbb{E}\{\boldsymbol{s}_{k,1}\boldsymbol{s}_{k,1}^H\}(\mathbf{R}_k\mathbf{H}_1^H\mathbf{F}_{k,1})^H \tag{3.7a}$$

$$+\sum_{g\in\mathcal{G}_1, g\neq k}(\mathbf{R}_k\mathbf{H}_1^H\mathbf{F}_{k,g})\,\mathbb{E}\{\boldsymbol{s}_{g,1}\boldsymbol{s}_{g,1}^H\}(\mathbf{R}_k\mathbf{H}_1^H\mathbf{F}_{k,g})^H \tag{3.7b}$$

$$+\sum_{m=2}^{M+1}\sum_{g'\in\mathcal{G}_m}(\mathbf{R}_k\mathbf{H}_m^H\mathbf{F}_{g',m})\,\mathbb{E}\{\boldsymbol{s}_{g',1}\boldsymbol{s}_{g',1}^H\}(\mathbf{R}_k\mathbf{H}_m^H\mathbf{F}_{g',m})^H \tag{3.7c}$$

$$+\mathbf{R}_k\mathbf{R}_k^H\sigma_z^2, \tag{3.7d}$$

where Eq. (3.7a) contains the intended symbol and also inter-layer interference, Eq. (3.7b) contains multi-user interference from the same cell, Eq. (3.7c) interference from other base stations and Eq. (3.7d) noise.

The layer symbols in the signal vector are uncorrelated and are the result of a bit to symbol mapping. Therefore, the expectation of the power of a symbol is denoted as

$$\mathbb{E}\{\boldsymbol{s}_{i,m}\boldsymbol{s}_{i,m}^H\} = P_{i,m}\mathbf{I}. \tag{3.8}$$

Therefore the correlation matrix is

$$\mathbb{E}\{\boldsymbol{y}_k\boldsymbol{y}_k^H\} = P_{k,1}(\mathbf{R}_k\mathbf{H}_1^H\mathbf{F}_{k,1})(\mathbf{R}_k\mathbf{H}_1^H\mathbf{F}_{k,1})^H \tag{3.9a}$$

$$+\sum_{g\in\mathcal{G}_1, g\neq k}P_{g,1}(\mathbf{R}_k\mathbf{H}_1^H\mathbf{F}_{g,1})(\mathbf{R}_k\mathbf{H}_1^H\mathbf{F}_{g,1})^H \tag{3.9b}$$

$$+\sum_{m=2}^{M+1}\sum_{g'\in\mathcal{G}_m}P_{g',m}(\mathbf{R}_k\mathbf{H}_m^H\mathbf{F}_{g',m})(\mathbf{R}_k\mathbf{H}_m^H\mathbf{F}_{g',m})^H \tag{3.9c}$$

$$+\mathbf{R}_k\mathbf{R}_k^H\sigma_z^2. \tag{3.9d}$$

To obtain the power of a received symbol we are interested in the diagonal elements of the correlation matrix. The following lemmas are used.

Lemma 1 *The diagonal elements c_{nn} of a matrix $\mathbf{C} = \mathbf{A}\mathbf{A}^H$, where $\mathbf{A} \in \mathbb{C}^{M\times N}$ are calculated by*

$$c_{mm} = \sum_{n=1}^{N}|a_{mn}|^2. \tag{3.10}$$

Lemma 2 *Using Lemma 1 the diagonal elements of a matrix $\mathbf{C} = \mathbf{A}\mathbf{A}^H + \mathbf{B}\mathbf{B}^H$ are*

$$c_{mm} = [\mathbf{A}\mathbf{A}^H]_{mm} + [\mathbf{B}\mathbf{B}^H]_{mm} = \sum_{n=1}^{N}|a_{mn}|^2 + \sum_{n=1}^{N}|b_{mn}|^2. \tag{3.11}$$

The correlation matrix from Eq. (3.9) is the sum of matrix products and hence Lemma 2 can be applied.

$$\mathbb{E}\{\boldsymbol{y}_k \boldsymbol{y}_k^H\} = P_{k,1}\mathbf{A}\mathbf{A}^H \tag{3.12a}$$
$$+ \sum_{g\in\mathcal{G}_1, g\neq k} P_{g,1}\mathbf{D}^{(g)}(\mathbf{D}^{(g)})^H \tag{3.12b}$$
$$+ \sum_{m=2}^{M+1}\sum_{g'\in\mathcal{G}_m} P_{g',m}\mathbf{C}^{(g',m)}(\mathbf{C}^{(g',m)})^H \tag{3.12c}$$
$$+ \mathbf{B}\mathbf{B}^H\sigma_z^2 \tag{3.12d}$$

With the new introduced matrices:

$$\mathbf{A} = \mathbf{R}_k\mathbf{H}_1^H\mathbf{F}_{k,1} \tag{3.13}$$
$$\mathbf{D}^{(g)} = \mathbf{R}_k\mathbf{H}_1^H\mathbf{F}_{g,1} \tag{3.14}$$
$$\mathbf{C}^{(g',m)} = \mathbf{R}_k\mathbf{H}_m^H\mathbf{F}_{g',m} \tag{3.15}$$
$$\mathbf{B} = \mathbf{R}_k\mathbf{R}_k^H \tag{3.16}$$

The diagonal element y_{ii} of the correlation matrix is then

$$y_{ii} = P_{k,1}\sum_j |a_{ij}|^2 \tag{3.17a}$$
$$+ \sum_{g\in\mathcal{G}_1, g\neq k} P_{g,1}\sum_j |d_{ij}^{(g)}|^2 \tag{3.17b}$$
$$+ \sum_{m=2}^{M+1}\sum_{g'\in\mathcal{G}_m} P_{g',m}\sum_j |c_{ij}^{(g',m)}|^2 \tag{3.17c}$$
$$+ \sigma^2\sum_j |b_{ij}|^2. \tag{3.17d}$$

Finally, the SINR $\gamma_{k,i}$ of user k in symbol i can be expressed. The power of the symbol of interest is only from element $|a_{ii}|^2$, while other contributions are inter-layer interference and all other terms are either noise, interference from signals intended for users in the same cell or interference from signals from other base stations.

$$\gamma_{k,i} = \frac{|a_{ii}|^2 P_{k,1}}{P_{k,1}\sum_{j\neq i}|a_{ij}|^2 + \sum_{g\in\mathcal{G}_1, g\neq k} P_{g,1}\sum_j |d_{ij}^{(g)}|^2 + \sum_{m=2}^{M+1}\sum_{g'\in\mathcal{G}_m} P_{g',m}\sum_j |c_{ij}^{(g',m)}|^2 + \sigma_z^2\sum_j |b_{ij}|^2} \tag{3.18}$$

$$|a_{ii}|^2 \ldots \text{signal}$$
$$|a_{ij}|^2 \ldots \text{inter-layer interference}$$
$$|b_{ij}|^2 \ldots \text{noise enhancement}$$
$$|d_{ij}^{(g)}|^2 \ldots \text{inter-cell interference}$$
$$|c_{ij}^{(g',m)}|^2 \ldots \text{intra-cell interference}$$

Furthermore, if we assume a ZF receive filter, the receiver will recover the sent symbols by suppressing the channel with

$$\mathbf{R}_k = (\mathbf{H}_1^H \mathbf{F}_{k,1})^+ \tag{3.19}$$

and hence

$$\mathbf{A} = \mathbf{I}. \tag{3.20}$$

This is equivalent to $a_{ii} = 1$ and $a_{ij} = 0$. In this case no inter-layer interference occurs. If the base station performs ZF precoding, in addition to the ZF receive filter, interference to users in the same cell will be cancelled as derived in Chapter 4. In this case $d_{ij} = 0$.

As all simulations performed in this thesis are assumed to have ZF receive filters as well as ZF precoding, the calculation of Eq. (3.18) simplifies to

$$\gamma_{k,i} = \frac{P_{k,1}}{\sum_{m=2}^{M+1} \sum_{g' \in \mathcal{G}_m} P_{g',m} \sum_j |c_{ij}^{(g',m)}|^2 + \sigma_z^2 \sum_j |b_{ij}|^2}. \tag{3.21}$$

If users are equipped with a single antenna, then only one symbol can be transmitted. Hence, only one SINR value has to be calculated and channel and precoder matrices degenerate to vectors. Anyways, the calculation of the SINR remains a computational expensive problem since it has to be performed for each user in every resource block. Hence the reduction in complexity given by Eq. (3.21) helps in decreasing the simulation duration.

4. MU-MIMO User Grouping

In this section the system model for a multi-user single cell scenario utilising ZF precoders is established. It is investigated how user grouping affects the channel conditions of users. An optimisation problem for user grouping is formulated and implementation challenges are presented. Additionally a user grouping heuristic called SPRS is proposed, and compared to the SUS algorithm from literature, to mitigate the problem of high computational complexity. An extension to SPRS is presented to account for fairness. The performance of the techniques is investigated by performing system level simulations.

4.1. System Model

We consider downlink multi-user transmissions with a single base station with N_T transmit antennas and K users with a single antenna, as shown in Fig. 4.1. To keep the notation simple the problem is considered for a single resource block. The base station transmits to G users simultaneously and applies digital precoding prior to sending the signal over the channel, where user k receives the symbol

$$y_k = \boldsymbol{h}_k^H \sum_{g \in \mathcal{G}} \boldsymbol{f}_g s_g + z_k \tag{4.1}$$

$$\boldsymbol{h}_k, \boldsymbol{f}_g \in \mathbb{C}^{N_T}, z_k \in \mathbb{C} \tag{4.2}$$

with the channel $\boldsymbol{h}_k$, the precoder $\boldsymbol{f}_g$ and the information symbol s_g. The information symbols are zero-mean random variables with a variance of P_g. The precoding vectors are of unit norm $\|\boldsymbol{f}\| = 1$, because otherwise they would contribute to the power allocation and this should only be controlled by the power of the symbols. Set $\mathcal{G}$ contains the scheduled user indices. Variable z_k denotes a zero-mean complex Gaussian random variable with a variance of σ_z^2. ZF precoding is a technique to cancel interference to other users by requiring

$$\boldsymbol{h}_i^H \boldsymbol{f}_j = c_j \delta_{i,j}, \forall i, j \in \mathcal{G}. \tag{4.3}$$

This requirement is only feasible under conditions on the set size of $\mathcal{G}$, which will be discussed later. With this, Eq. (4.1) is

$$y_k = c_k s_k + z_k. \tag{4.4}$$

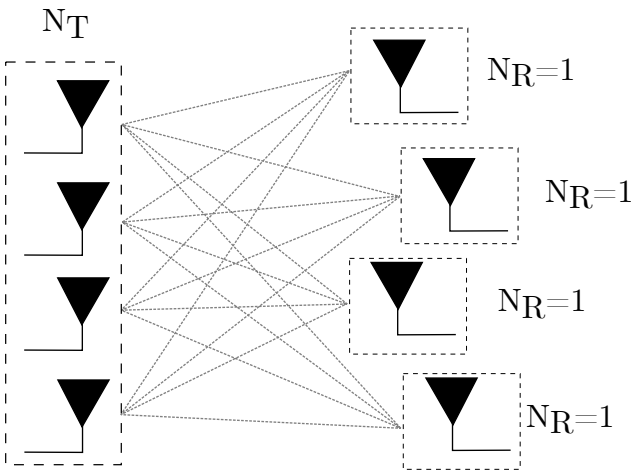


Figure 4.1.: Multi-User MIMO system, several users with $N_R = 1$ receive antennas are served by a base station with N_T transmit antennas. The users receive a mixture of signals from the transmitting antennas indicated by the dotted lines. The transmit antennas have to send signals in a way that the superposition of the transmit signals can be decoded by the receiver.

The precoders transform the effective channel such that each user is served by an independent data stream. Interference from other users is thereby cancelled.

Therefore, the signal to noise ratio (SNR) γ_k of user k depends on the inner product of the channel and the precoder.

$$\gamma_k = \frac{P_k |c_k|^2}{\sigma_z^2} \tag{4.5}$$

In general, a base station is free to choose the scheduled users. For that, a binary indicator vector $\boldsymbol{g}$ is introduced

$$\boldsymbol{g} = \begin{bmatrix} g_1 & g_2 & \dots & g_K \end{bmatrix}^T, \quad g_i \in \{0, 1\}, \tag{4.6}$$

where $g_i = 1$ means that user i is scheduled and $g_i = 0$ means not scheduled. Therefore, the set of grouped users is also described as

$$\mathcal{G} = \{i | g_i \neq 0\}, \tag{4.7}$$

and the number of grouped users is $G = |\mathcal{G}|$. For sake of simplicity we assume that after scheduling the indices 1 to G are used for the scheduled users. Defining

$$\tilde{\boldsymbol{f}}_k = \boldsymbol{f}_k / c_k, \tag{4.8}$$

we can compactly write Eq. (4.3) as

$$\begin{bmatrix} \boldsymbol{h}_1^H \tilde{\boldsymbol{f}}_1 & \boldsymbol{h}_1^H \tilde{\boldsymbol{f}}_2 & \dots & \boldsymbol{h}_1^H \tilde{\boldsymbol{f}}_G \\ \boldsymbol{h}_2^H \tilde{\boldsymbol{f}}_1 & \boldsymbol{h}_2^H \tilde{\boldsymbol{f}}_2 & \dots & \boldsymbol{h}_2^H \tilde{\boldsymbol{f}}_G \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{h}_G^H \tilde{\boldsymbol{f}}_1 & \boldsymbol{h}_G^H \tilde{\boldsymbol{f}}_2 & \dots & \boldsymbol{h}_G^H \tilde{\boldsymbol{f}}_G \end{bmatrix} = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{bmatrix}, \tag{4.9}$$

which is equivalent to

$$\begin{bmatrix} \boldsymbol{h}_1 & \boldsymbol{h}_2 & \dots & \boldsymbol{h}_G \end{bmatrix}^H \begin{bmatrix} \tilde{\boldsymbol{f}}_1 & \tilde{\boldsymbol{f}}_2 & \dots & \tilde{\boldsymbol{f}}_G \end{bmatrix} = \mathbf{I}_G, \tag{4.10}$$

where, for the sake of simplicity, $\boldsymbol{h}_i$, $\tilde{\boldsymbol{f}}_i$ denotes the channel and precoder of the i-th scheduled user. We define

$$\mathbf{H}^H \tilde{\mathbf{F}} = \mathbf{I}_G, \tag{4.11}$$

where $\mathbf{H}$ contains the stacked channels and $\tilde{\mathbf{F}}$ the stacked precoders. We emphasise here again that the channel matrix only contains the channels of the scheduled users and is therefore a function of the indicator vector written as

$$\mathbf{H} = \mathbf{H}(\boldsymbol{g}). \tag{4.12}$$

The pseudo-inverse, for $G \leq N_T$, of the combined channel matrix solves the equation and is expressed as

$$\tilde{\mathbf{F}} = \mathbf{H}(\mathbf{H}^H \mathbf{H})^{-1} = \mathbf{H}^+. \tag{4.13}$$

This defines the feasibility condition: the maximum number of users in a group G is upper bounded by the number of transmit antennas N_T. From Eq. (4.8) we conclude that

$$|c_k|^2 = \frac{1}{\|\tilde{\boldsymbol{f}}_k\|^2}. \tag{4.14}$$

We are therefore ready to express the SNR from Eq. (4.5) of a scheduled user as

$$\gamma_k = \frac{P_k}{\sigma_z^2 \left\| \tilde{\boldsymbol{f}}_k \right\|^2}. \tag{4.15}$$

It is assumed that the base station serves $G \leq \min\{K, N_T\}$ users simultaneously. Utilising Eq. (4.4) the achievable rate of a scheduled user is calculated by Shannon's AWGN channel capacity [26] and is a function of the power allocation and the user grouping vector from Eq. (4.6).

$$R_k(\boldsymbol{g}, P_k) = \log_2(1 + \gamma_k) = \log_2\left(1 + \frac{P_k}{\sigma_z^2 \left\| \tilde{\boldsymbol{f}}_k \right\|^2}\right) \tag{4.16}$$

Sum-rate R is the sum of the individual user rates and expressed as

$$R(\boldsymbol{g}, \boldsymbol{p}) = \sum_{k=1}^{K} g_k R_k(\boldsymbol{g}, P_k). \tag{4.17}$$

Vector $\boldsymbol{p}$ contains the power allocation values P_k. At the base station, a maximum transmit power P_t is imposed. The objective of maximising the sum-rate under ZF precoding is formulated in the following optimisation problem

$$\underset{\boldsymbol{g},\boldsymbol{p}}{\text{maximize}} \quad \sum_{k=1}^{K} g_k \log_2\left(1 + \frac{P_k}{\sigma_z^2 \left\|\tilde{\boldsymbol{f}}_k\right\|^2}\right) \tag{4.18a}$$

$$\text{subject to} \quad \sum_{k=1}^{K} g_k P_k \leq P_t \tag{4.18b}$$

$$\sum_{k=1}^{K} g_k \leq N_T, \tag{4.18c}$$

where Eq. (4.18a) is called the optimisation function. The constraint in Eq. (4.18b), ensures that the maximum transmit power is not exceeded and Eq. (4.18c) ensures that not more users than antennas are scheduled.

The problem has to be solved with respect to the user allocation and the power allocation. If we only look at this problem with respect to the power allocation, meaning that we assume a given user grouping, this was already solved and the solution is the water-filling algorithm [26].

Therefore, we now focus on looking at the problem only with respect to the user allocation. We impose that the transmit power is uniformly distributed between the scheduled users, since this will not artificially change the macroscopic channel conditions defined by the environment and is allocating a fair share of power to each user. In addition, it is also a simple method to keep the complexity low. Therefore, for the rest of this thesis the sum-rate is expressed as

$$R(\boldsymbol{g}) = \sum_{k=1}^{K} g_k R_k(\boldsymbol{g}, P_t/G). \tag{4.19}$$

In optimisation theory, it is convenient to have convex problems for which numerical solvers exist. To use the framework of convex optimisation the problem has to fulfil three basic properties: the optimisation variable has to come from a convex set, the objective function has to be a convex function, and also the inequality constraint functions have to be convex functions [27, p. 127].

The integer-valued user allocation variable $\boldsymbol{g}$ imposes a challenge, since it is not from a convex set. Integer relaxation provides the possibility to use the convex

optimisation framework to find a sub-optimum solution as described in [27, p. 194]. Here, the integer variable is relaxed to a real number. In the solution, the variable is rounded to the closest integer.

It was mentioned that it is also required that the objective function is convex. A function can be determined to be convex by disassembling it into basic functions and checking if each function preserves convexity. The outer function of Eq. (4.18a) is the sum of negative logarithms. The negative logarithm is a convex function, and the sum an affine mapping, both operations that preserve convexity. Inside the logarithm there is the precoder magnitude. It depends on the user selection variable and the pseudo-inverse in Eq. (4.13). This function is not convex, since its domain is not a convex set.

Therefore, problem 4.18 is neither convex nor is it obvious to find a relaxation. To find an optimum solution, all possible user allocations have to be checked. In each of these steps, the calculation of the pseudo-inverse requires one to calculate the matrix inverse of a $G \times G$ matrix, where G is the size of the user group. This imposes heavy computational effort. In [11] it is defined as a complex combinatorial problem. Therefore, we rely on heuristics to find sub-optimum solutions in feasible time.

4.2. Semi-Orthogonal User Selection

Semi-Orthogonal User Selection (SUS) [10] is a heuristic that groups users by channel correlation and magnitude. The motivation for this relies on the behaviour of the precoder with regards to correlation. Equation (4.15) shows that the SNR of a user depends on the squared inverse of the precoder magnitude. Hence, it is desired to have small precoder magnitudes to increase the SNR and therefore the rate.

We will show the dependency of the SNR to the channel correlation by a simple example, where two users are served by two transmit antennas. Let the channels be defined as

$$\boldsymbol{h}_1 = \begin{bmatrix} 1 & 0 \end{bmatrix}^T, \quad \boldsymbol{h}_2 = \begin{bmatrix} \cos\alpha & \sin\alpha \end{bmatrix}^T. \tag{4.20}$$

Using Eq. (4.13), the precoders are calculated

$$\mathbf{H} = \begin{bmatrix} \boldsymbol{h}_1 & \boldsymbol{h}_1 \end{bmatrix} \tag{4.21}$$

$$\mathbf{H}^H\mathbf{H} = \begin{bmatrix} 1 & 0 \\ \cos\alpha & \sin\alpha \end{bmatrix} \begin{bmatrix} 1 & \cos\alpha \\ 0 & \sin\alpha \end{bmatrix} = \begin{bmatrix} 1 & \cos\alpha \\ \cos\alpha & 1 \end{bmatrix} \tag{4.22}$$

$$\mathbf{F} = \mathbf{H}[\mathbf{H}^H\mathbf{H}]^{-1} = \frac{1}{\sin\alpha} \begin{bmatrix} \sin\alpha & 0 \\ -\cos\alpha & 1 \end{bmatrix}. \tag{4.23}$$

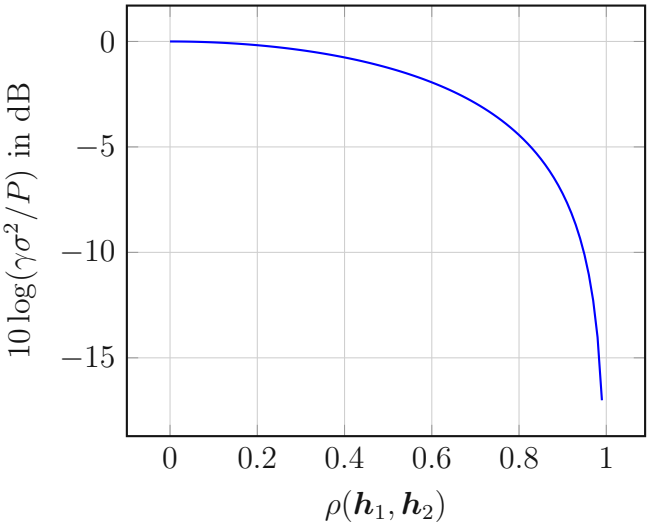


Figure 4.2.: Normalised user SNR in a two user and two antenna scenario over the channel correlation.

This allows to express the magnitude

$$\left\|\tilde{\boldsymbol{f}}_1\right\|^2 = \left\|\tilde{\boldsymbol{f}}_2\right\|^2 = \frac{1}{1-\cos^2\alpha} = \frac{1}{1-\rho^2(\boldsymbol{h}_1,\boldsymbol{h}_2)}, \tag{4.24}$$

where the channel correlation is defined as

$$\rho(\boldsymbol{h}_i,\boldsymbol{h}_j) = \frac{|\boldsymbol{h}_i^H \boldsymbol{h}_j|}{\|\boldsymbol{h}_i\|\,\|\boldsymbol{h}_j\|} = \cos\alpha. \tag{4.25}$$

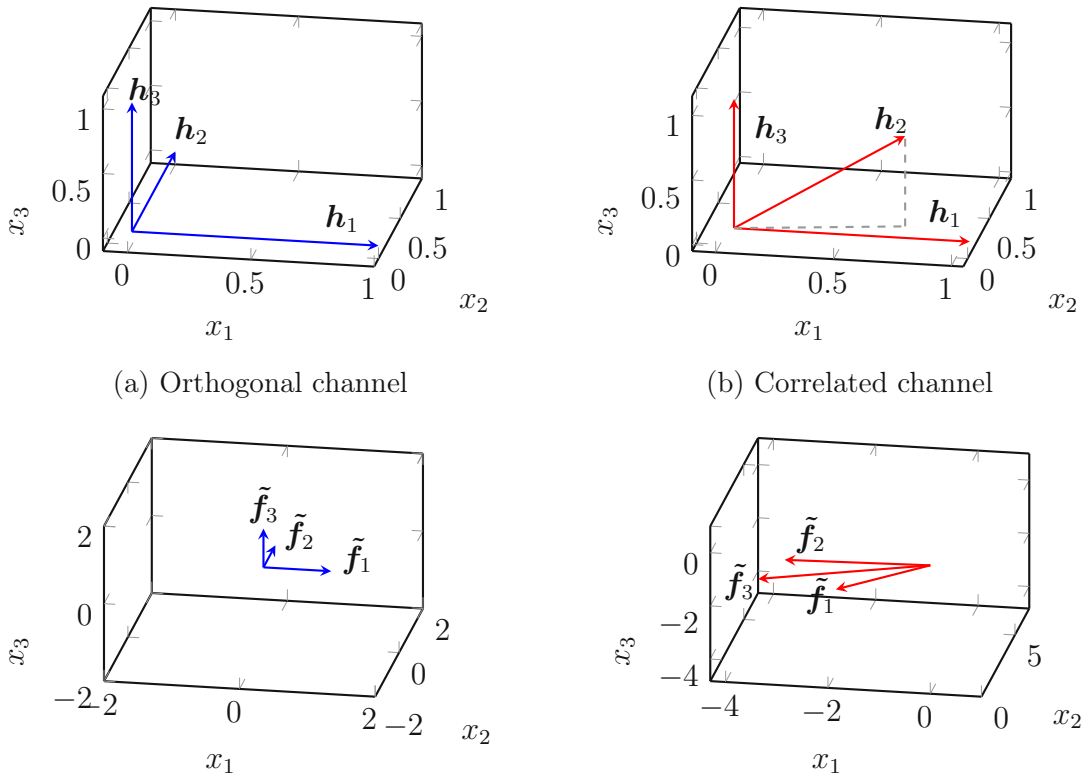
The user SNR according to Eq. (4.15) is then

$$\gamma(\rho) = \frac{P}{\sigma^2}[1-\rho^2(\boldsymbol{h}_1,\boldsymbol{h}_2)]. \tag{4.26}$$

This function is plotted in Fig. 4.2 for a logarithmic SNR, where the SNR is normalised with respect to transmit power and noise. With an increase in correlation the SNR diminishes.

For an increasing number of antennas and users, the situation becomes more complex. Here, user channels have correlation to multiple other channels. Figure 4.3 depicts the situation with three users. Although the situation is now more complex, the same principle applies and precoding magnitudes will tend to adopt larger values if channels are correlated.

Based on this insight, the SUS algorithm performs the user grouping by selecting users that are not too correlated to each other. This property of weak correlation



(a) Orthogonal channel

(b) Correlated channel

(c) Orthogonal precoders

(d) Correlated precoders

Figure 4.3.: Illustration of the behaviour of zero-forcing precoders in the 3D space. In Fig. 4.3a we see three orthogonal channels. The resulting precoders in Fig. 4.3c are also orthogonal. The situation changes in Fig. 4.3b, where channel $\boldsymbol{h}_2$ is correlated. The resulting precoders in Fig. 4.3d are much larger and result in a poor signal to noise ratio.

is called semi-orthogonality. The algorithm works in two steps: an election step, where a user is added to the group, and a filtering step, where correlated users are removed from the candidate pool.

In Algorithm 1 a pseudo-code for SUS is shown. The program starts by adding the user with the strongest channel as the first grouped user. The remaining users form the user pool. As long as there is a user in the pool, an orthonormal basis is constructed with the basis vectors of the already grouped users and the candidate. The channel vectors in the pool are projected on the basis and the user whose projected vector has the largest norm is added into the group. Afterwards, in the filtering step, the correlation between the projected vectors and the users in the pool is compared with a parameter α. Users with high correlation are removed from the pool. The algorithm ends if the pool is empty.

Algorithm 1 Semi-Orthogonal User Selection Algorithm

1: $g_1 = \arg\max_k \boldsymbol{h}_k$ ▷ Initialise with the strongest user.
2: $\mathcal{P} = \{1 \dots K\} \setminus \{g_1\}$ ▷ Rest of the users form the pool.
3: G=1
4: **while** $|\mathcal{P}| > 0$ and $G < N_T$ **do** ▷ Pool is not empty and system is feasible.
5: **for** $k \in \mathcal{P}$ **do** ▷ Election step
6: $\boldsymbol{r}_k = \boldsymbol{h}_k - \sum_{g=1}^{G} \boldsymbol{h}_k^H \boldsymbol{g}_g \boldsymbol{g}_g$ ▷ Calculate orthonormal basis.
7: **end for**
8: $G \leftarrow G + 1$
9: $\hat{k} = \arg\max_k \|\boldsymbol{r}_k\|$ ▷ Pick the strongest candidate.
10: $g_G = \hat{k}$
11: $\boldsymbol{g}_G = \frac{\boldsymbol{r}_{\hat{k}}}{\|\boldsymbol{r}_{\hat{k}}\|}$
12: $\mathcal{P} \leftarrow \mathcal{P} \setminus \{g_G\}$ ▷ Filter step. Remove candidate from pool.
13: $\mathcal{P} \leftarrow \{k \in \mathcal{P} | \frac{\boldsymbol{h}_k^H \boldsymbol{g}_G}{\|\boldsymbol{h}_k\|} < \alpha\}$ ▷ Remove correlated users from pool.
14: **end while**
15: $\mathcal{G} = \{g_1, g_2, \dots, g_G\}$ ▷ A set containing the grouped users.

Compared to the exhaustive search, the algorithm provides a significant reduction in complexity. First, because the users are added in a greedy manner. And second, its not necessary to compute any matrix inverse. A downside to this is that the algorithm does not directly maximise the sum-rate. In addition, a value for the α parameter has to be found. Its optimum value is unknown and is depending on the number of users, antennas, and the channel condition. Therefore practical simulations must be performed to find a suitable value.

4.3. Single Path Random Sampling

The section above describes that the SUS algorithm performs grouping based on the semi-orthogonality properties of the channels. In [10], the authors showed that for $K \to \infty$ this scheme is asymptotically optimal. For a finite number of K this does not hold and opens the possibility to find other heuristics.

This work proposes a novel heuristic, which is called single path random sampling (SPRS), for finding a subset of users that maximise the sum-rate. Its requirement is to get closer to the sum-rate optimisation goal.

This comes with a price, since we showed in Eq. (4.19), that the calculation of the sum-rate requires the calculation of a pseudo-inverse. Hence, to reduce time complexity, SPRS is based on four techniques: (1) The user with the strongest channel is elected as the first user in the group. (2) The algorithm is greedy. In each iteration step, a user is added to the group. (3) To reduce complexity, not all users in the pool are considered as candidates. In each iteration step, the algorithm samples candidates from the user pool as shown in Fig. 4.4. This candidate pool is used for further calculations. (4) Users with strongly correlated channels will be sorted out. Compared to SUS, this last step is not to reduce complexity, but to ensure numerical stability. Since the sum-rate calculation requires to calculate an inverse matrix, it is important to have a suitable matrix condition.

In Algorithm 2 a pseudo-code of the algorithm is presented. As described above, the algorithm works with a user pool and a candidate pool. Parameter P_S, the skip probability, is introduced to control the size of the candidate pool. P_S is a real number between 0 and 1. It represents the probability that a user will not be part of the candidate pool. The expected value of the size of the candidate pool for K users is therefore $(1 - P_S)K$. The lower the skip probability, the more users will be part of the candidate pool. This will increase the computational effort, but also provides better solutions. Therefore, the skip probability can be seen as a trade-off parameter, adjusting complexity at the cost of throughput.

Compared to SUS, the SPRS algorithm provides some advantages and disadvantages. SUS suffers from the problem that users are removed in an early iteration. Once removed, they are never considered again as possible candidates, whereas in the SPRS case, a user always has the opportunity to be reconsidered again. By design, the optimisation is performed with regards to the sum-rate. Simulations will show in the later sections that this indeed leads to higher rates. A downside is an increase in complexity, which arises from the matrix inverse calculations.

Algorithm 2 The SPRS algorithm. For sake of simplicity, the function $R(\mathcal{G})$ maps the user grouping set to the achieved sum-rate. This is equivalent to the notation in Eq. (4.19), but with set $\mathcal{G}$ as function argument.

$g_1 = \arg\max_k \|\boldsymbol{h}_k\|$ ▷ Initialise with the strongest user.
$\mathcal{G} = \{g_1\}$ ▷ Initialise group of scheduled users.
$\mathcal{P} = \{1 \dots K\} \setminus \{g_1\}$ ▷ Rest of the users form the pool.
$R_{\max} = R(\mathcal{G})$ ▷ Initialise sum-rate of grouped users.
$i = 1$ ▷ Initialise iteration index.
while $\mathcal{P} \neq \emptyset$ and $i < N_T$ **do**
  $g_i = 0$
  $\mathcal{P} \leftarrow \{k \in \mathcal{P} | \frac{\boldsymbol{h}_k^H \boldsymbol{h}_i}{\|\boldsymbol{h}_k\| \|\boldsymbol{h}_i\|} < \alpha\}$ ▷ Filter correlated users from pool.
  $i = i + 1$
  $\mathcal{C} \leftarrow$ Sample from set $\mathcal{P}$ with skip probability P_S
  for $c \in \mathcal{C}$ **do** ▷ For each user in the candidate pool
    $R' = R(\mathcal{G})$ ▷ Calculate rate
    if $R' > R_{\max}$ **then** ▷ If the rate increases
      $g_i \leftarrow c$ ▷ Save user
      $R_{\max} \leftarrow R'$
    end if
  end for
  if $g_i == 0$ **then**
    End algorithm ▷ End if no user increased the rate.
  end if
  $\mathcal{G} \leftarrow \mathcal{G} \cup \{g_i\}$ ▷ Add user that maxed out rate.
  $\mathcal{P} \leftarrow \mathcal{P} \setminus \{g_i\}$
end while

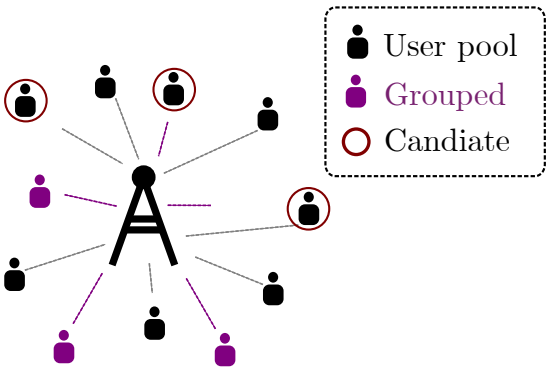


Figure 4.4.: Illustration of the SPRS algorithm. Purple users are already part of the group. The remaining black users form the user pool. Out of these, the algorithm samples the circled candidates.

4.4. SPRS Fairness Extension

The SPRS algorithm presented in the last section optimises for sum-throughput and fairness is not considered. Users with favourable channel conditions will be preferentially scheduled and users with bad channels are starved. This section aims to extend the algorithm to consider fairness. For this, instead of using the sum-rate R as shown in Algorithm 2, a weighted sum-rate metric is used. Priority P is introduced and defined as

$$P = \frac{\hat{R}^a}{R_{\text{avg}}^b}. \tag{4.27}$$

Here $\hat{R}$ is the expected throughput and R_{avg} is the past average throughput of the user. Parameters a and b tune the optimisation goal. Choosing $a = 1$, $b = 0$ leads to

$$P_{\text{best-rate}} = \hat{R} \tag{4.28}$$

and therefore to the SPRS algorithm introduced in the former section. The combination $a = 0$, $b = 1$ results in

$$P_{\text{max-min}} = \frac{1}{R_{\text{avg}}} \tag{4.29}$$

and prefers users with low average throughput. Hence, a uniform distribution of user throughput is expected. The combination $a = 1$, $b = 1$ leads to a PF scheduling with

$$P_{\text{PF}} = \frac{\hat{R}}{R_{\text{avg}}}, \tag{4.30}$$

where the expected throughput is weighted against the past average throughput. Otherwise the algorithm is identical with Algorithm 2, with the exception of interchanging the sum-rate mapping to the priority mapping. Due to the random nature of the algorithm the number of slots until a certain fairness is reached depends on the skip probability. If many users are skipped it can take some time until fairness is achieved, since in each iteration the skipped users are not considered for fairness.

4.5. Performance of SUS and SPRS

The rest of this chapter presents and discusses simulation results of various performance comparisons between grouping algorithms. Simulations are performed with the Vienna 5G System Level Simulator which is described in Chapter 2. For the following simulations users are randomly placed around the base station, as in Fig. 4.5. The high number of users will ensure a wide variety of channel strengths

due to path loss. During simulation the users will move with constant velocity in a random direction at moderate speed to create time-varying correlated channels. To ensure a fair comparison all algorithms start with the same user positions. From there users perform walks in random directions. This is important for simulations presented in Section 4.7, since fairness among individual users is examined.

The first simulation compares the performance of the SUS and SPRS algorithm, where we investigate the best-rate version of SPRS. Table 4.1 shows the key parameters of the simulation. The number of users is much higher than the number of antennas, which ensures multi-user diversity. Since the performance is dependent on the algorithm parameters, the simulation is performed for a SUS α value of 0.2 and 0.5, and for a SPRS skip probability of 0.9 and 0.7. This shall give an idea for the parameter choice of α and P_S. The performance is evaluated in terms of the sum-throughput per slot as defined in Eq. (2.4) an the results are plotted as an empirical cumulative distribution function (ECDF) in Fig. 4.6. It is shown once for the case where the users reported CQI values are used and for the case where the CQI value that maximises the throughput was chosen by the scheduler. The results show that SPRS is achieving higher sum-throughput values compared to the SUS algorithm. The influence of the grouping parameter is clearly visible, emphasising the importance of choosing a suitable value. Here, choosing α too small degrades the performance because this results in many filtered users and hence small group sizes, which reduces the rate. A similar observation holds for the skip probability of SPRS, but with a smaller influence. Choosing a skip probability of 0.9 instead of 0.7, meaning that from 100 possible users only 10 and not 30 users are searched in each iteration, degrades the throughput a bit, but still offers high sum-throughput. Hence, the skip probability is more of a throughput-to-complexity trade-off parameter.

It is observed in the difference between the feedback and best CQI case, that incorrect CQI feedback deteriorates the performance. A user has to report his CQI value for every slot. At the same time the precoder changes from slot to slot, and with it the SNR, making it hard to deliver an up-to-date value to the base station.

4.6. SPRS Compared to the Exhaustive Search

This scenario compares the performance of the exhaustive search and SPRS in the best-rate version. Although the exhaustive search is not feasible for a large number of users it can still be performed for a small number of users. In this case the performance of SPRS can be compared to the optimum solution. It is intended to show how far the result deviates from the optimal solution due to the greedy approach and the reduction of search space due to random sampling. The complexity limits the number of users and antennas to 18 with the simulation

Table 4.1.: Simulation Parameters for Section 4.5

Base stations	1
Transmit antennas	32
TX power	1 W
Users	100
Small-scale fading model	PedA 5 km h^{-1}
Bandwidth	9 MHz
Number of slots per chunk	200
Number of chunks	72

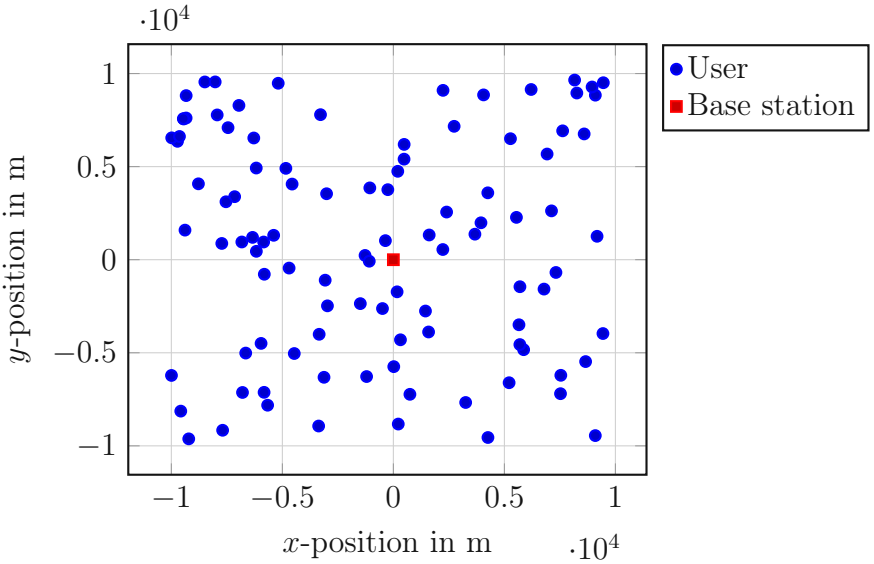
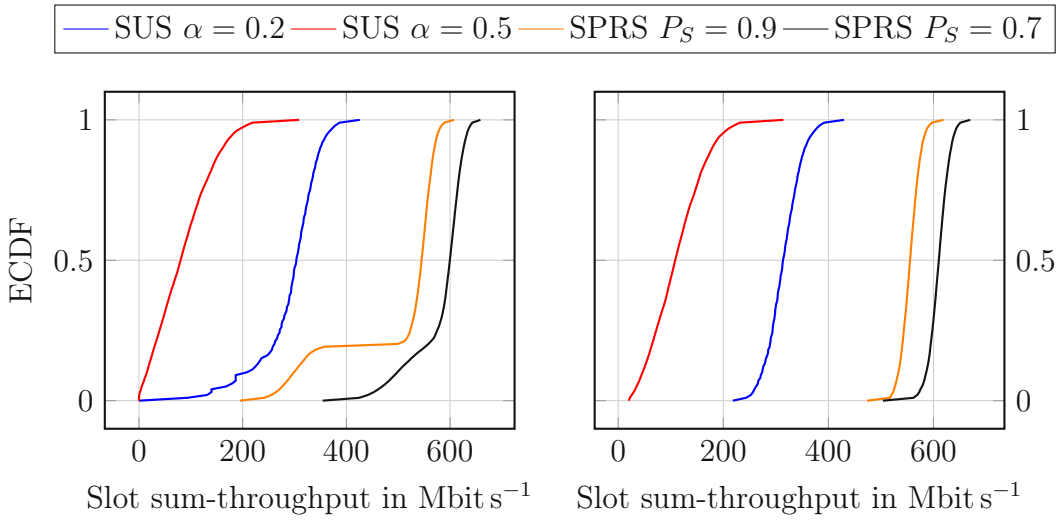


Figure 4.5.: Fixed user and base station positions of the single cell scenarios.



(a) Feedback CQI (b) Best CQI

Figure 4.6.: Sum-throughput on a slot basis comparing SUS and SPRS best-rate. In Fig. 4.6a CQI values are reported from the users. In Fig. 4.6b optimum CQI values are used.

parameters shown in Table 4.2.

In Fig. 4.7 the slot sum-throughput results are shown. What can be seen on first sight is the similar performance of all three algorithms. This concludes that the greedy approach is finding solutions close to the optimum. Also reducing the search set by 50 % still yields good results. But also the multi-user diversity is still high and the algorithm has nine candidates available in each iteration step.

Table 4.2.: Simulation Parameters for Section 4.6

Base stations	1
Transmit antennas	18
TX power	1 W
Users	18
Small-scale fading model	PedA 5 km h^{-1}
Bandwidth	4 MHz
Number of slots per chunk	150
Number of chunks	36

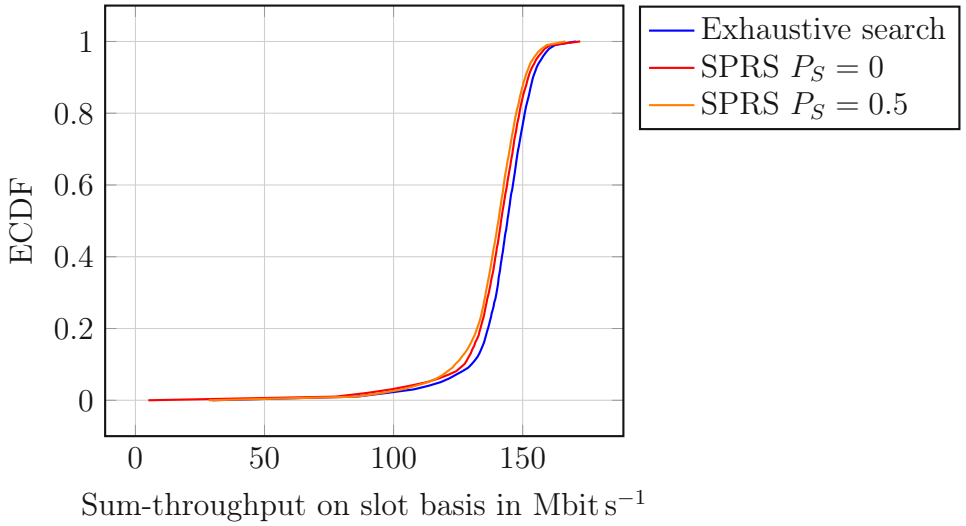


Figure 4.7.: Slot sum-throughput for SPRS best-rate and the exhaustive Search.

4.7. Performance of SPRS Fairness Extension

This scenario examines the performance of SPRS with its fairness extension as proposed in Section 4.4. Three operation modes are examined: the best-rate version that optimises for throughput, the PF version, and the max-min version. The simulation parameters are shown in Table 4.3. The number of antennas was set back to 32, but the number of users was reduced to 12 to ease the illustration. Users were sorted by their path loss in ascending order. Hence, a user with low index is close to the base station.

In Fig. 4.8 the total throughput per user averaged over slots is plotted. In case of the original SPRS algorithm the users with strong channels are preferred and users with bad channels are almost completely starved. The max-min variant balances the throughput among the users, at the cost of reducing the average sum-throughput. The PF variant operates in-between those regimes, allocating proportionally more resources to strong users, without starving weak users. In Fig. 4.9 the price, in terms of the slot sum-throughput, that must be paid for the fairness is shown.

The two parameters a and b allow an easy control of the system's fairness. Furthermore, it opens the possibility for future work to investigate in more complex priority control for scenarios where users have traffic policies. This policies could be directly implemented in the user grouping algorithm.

Table 4.3.: Simulation Parameters for Section 4.7

Base stations	1
Transmit antennas	32
TX power	1 W
Users	100
Small-scale fading model	PedA 5 km h^{-1}
Bandwidth	9 MHz
Number of slots per chunk	200
Number of chunks	72

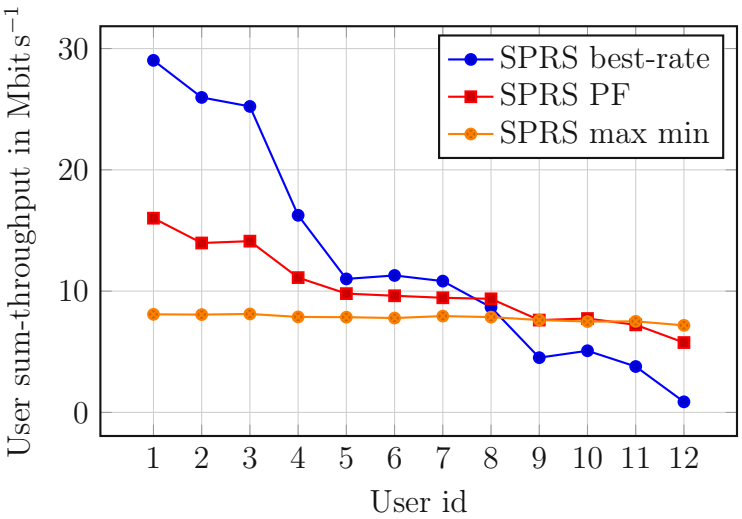


Figure 4.8.: User sum-throughput to compare SPRS fairness. Users with a lower index have higher macroscopic SNR conditions.

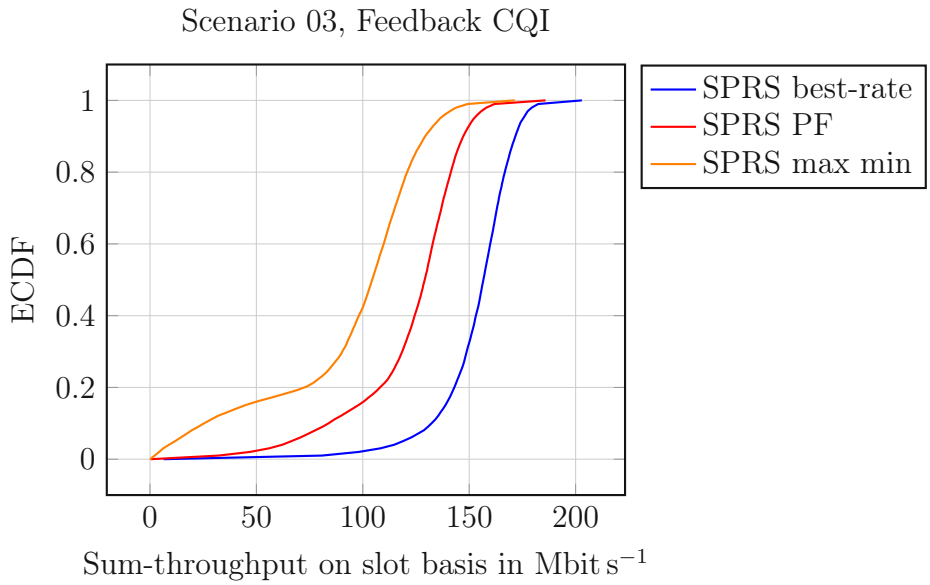


Figure 4.9.: Slot sum-throughput comparison of the different fairness options.

4.8. SPRS Algorithm Complexity

It is still an open question what value to choose for the skip probability and how it influences the complexity. Complexity is an important topic, since the grouping algorithms have to be executed periodically by the base station. The computational effort should not be too high to reduce power consumption. If the user grouping should be performed every five slots, the algorithm has to deliver a solution every 5 ms. To get an idea of its influence, simulations are performed for a skip probability ranging from 0 to 1. The average sum-throughput as well as the time spent for simulation are saved and compared to the SUS algorithm. Here, we choose the SUS parameter $\alpha = 0.2$, since it delivers the highest throughput with the lowest simulation duration.

The simulation is performed over 56 chunks with 200 slots each, as shown in Table 4.4. The user positions are randomly generated for each chunk according to a Poisson 2D process and the average sum-throughput over slots, as defined in Eq. (2.6), is calculated over all chunks.

The run-time and average sum-throughput over the slots is shown in Fig. 4.10. As can be seen, an increasing skip probability has a strong impact on the simulation time. For low skip probabilities the simulation time is high because of the large candidate sets. Large candidate sets have the advantage that the multi-user diversity is higher, but the simulation time is larger. One may think that a smaller candidate set size for higher skip probabilities is the only reason for the

Table 4.4.: Simulation Parameters for Section 4.8

Base stations	1
Transmit antennas	100
TX power	1 W
Users	100
Small-scale fading model	PedA 5 km h^{-1}
Bandwidth	9 MHz
Number of slots per chunk	200
Number of chunks	54

faster simulation time. But also a second effect comes into play: if the set is too small, the algorithm will abort with higher probability since no user can increase the sum-rate.

Although the candidate sets get smaller and the algorithm aborts earlier with increasing P_S, we observe that the throughput is still unaffected until $P_S = 0.8$. There are variations in the throughput and the curve is not monotonic, but this is explained by the random nature of the algorithm. It is not guaranteed that a global optimum is found and some solutions will be less favourable.

With an increasing P_S the algorithm runs less iterations, since the pools are smaller, but it does not affect the throughput a lot since users added at the last iteration stage usually contribute only with small throughput gains. Only if the candidate sets get too small, the throughput abruptly collapses, since too many users are skipped.

To compare it with the SUS algorithm, the graph shows two dashed lines for $\alpha = 0.2$. As it can be seen SUS has overall a lower run time complexity compared to SPRS, but also its throughput is lower. The points at $P_S = 0.8$ and $P_S = 0.9$ are especially of interest, since they achieve a significant throughput gain with moderate complexity. At $P_S = 0.8$, SPRS achieves a throughput that is about 1.74 times higher than the SUS throughput. On the other hand, it is slower by a factor of 2.76. For $P_S = 0.9$, it achieves a throughput that is about 1.5 times higher than the SUS throughput, and is slower by a factor of 2.5.

One could have the idea to find the point where SUS and SPRS produce the same throughput values. Unfortunately, this is hard to achieve since after $P_S = 0.95$, the throughput decreases significantly. Due to the many random variables, lengthy simulations are necessary to produce a stable point, which is infeasible.

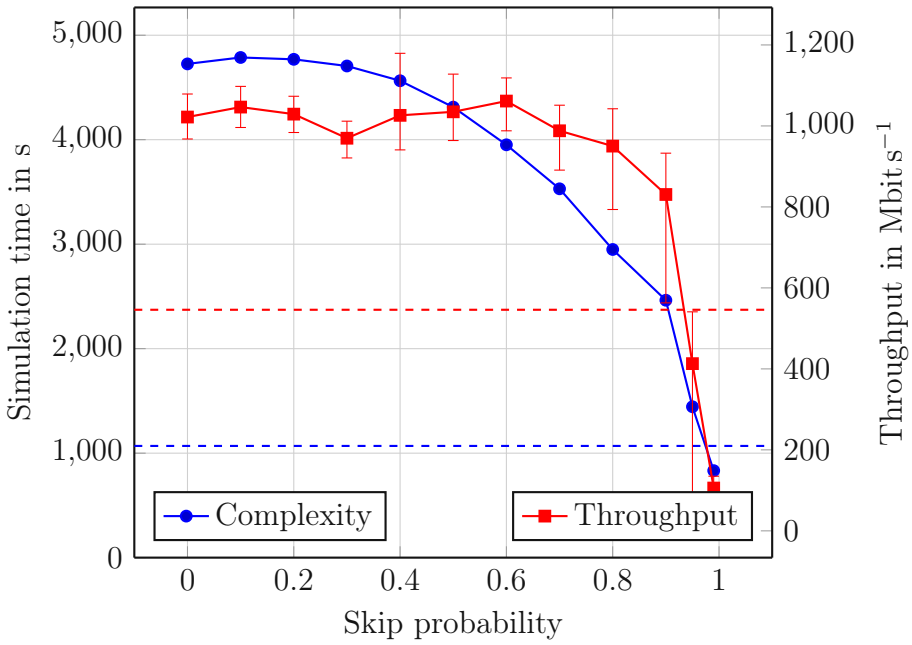


Figure 4.10.: Complexity in terms of simulation duration and average sum-throughput of the SPRS best-rate algorithm. The dashed lines show the performance of the SUS algorithm with $\alpha = 0.2$. The bars at the throughput points show the 80-th and 20-th percentiles.

5. Cellular MU-MIMO Systems

This chapter extends the previous chapter from a single cell to a cellular scenario. The optimisation problem is formulated, and difficulties due to the even higher complexity are discussed. An approach to suboptimal user grouping is investigated in combination with fractional frequency reuse (FFR). The performance of SUS and SPRS in scenarios with and without FFR is investigated with system level simulations.

5.1. System Model

In Section 4.1 a system model was presented for a scenario with a single base station and multiple users. We extend the model to the cellular case, where multiple base stations serve multiple users. It is assumed that all base stations use ZF precoding to serve their users, and therefore no intra-cell interference is created. Users not only receive noise, but also interference coming from neighbouring base stations, but in general, a base station is capable to use the available degrees of freedom to form beams that avoid interference to users of a neighbouring cell. As we are now considering multiple base stations, we assume that each user is connected to one of the B base stations. Therefore, the user SNR from Eq. (4.15) is extended with the interference from the other base stations and is now called the user SINR. Hence, the SINR of user k scheduled at base station b is expressed as

$$\gamma_k = \frac{P_{kb} / \left\| \tilde{\boldsymbol{f}}_{kb} \right\|^2}{\sigma_z^2 + \sum_{i=1, i \neq b}^{B} (1 - g_{ki}) \sum_{k' \in \mathcal{K}_i} P_{k'i} |\boldsymbol{h}_{ki}^H \tilde{\boldsymbol{f}}_{k'i}|^2}. \tag{5.1}$$

Here, P_{kb} and $\tilde{\boldsymbol{f}}_{kb}$ denote the power and precoder allocated for user k from base station b. The noise is assumed to be zero-mean Gaussian with a noise variance of σ_z^2. Variable $g_{kb} = 1$ denotes that the base station b is scheduling the user k and variable $c_{ki} = 1$ denotes that the base station i is zero-forcing the channel of user k, therefore creating no interference. This models the possibility of base stations cancelling interference to users of other cells. Set $\mathcal{K}_i$ contains indices of users served by base station i. Variable $\boldsymbol{h}_{ki}$ represents the channel between base station i and user k.

Similar to the single base station case in Chapter 4, it is again desired to formulate an optimisation problem. Therefore an extension to Eq. (4.18) is derived with objective of maximising the sum-rate in a cellular system using the Shannon channel capacity.

$$\underset{\boldsymbol{g},\,\boldsymbol{P}}{\text{maximize}} \quad \sum_{b=1}^{B} \sum_{k \in \mathcal{K}_b} g_{kb} \log_2 \left(1 + \frac{P_{kb} / \left\| \tilde{\boldsymbol{f}}_{kb} \right\|^2}{\sigma_z^2 + \sum\limits_{i=1, i \neq b}^{B} (1 - c_{ki}) \sum\limits_{k' \in \mathcal{K}_i} P_{k'i} |\boldsymbol{h}_{ki}^H \boldsymbol{f}_{k'i}|^2} \right) \tag{5.2a}$$

$$\text{subject to} \quad \sum_{k=1}^{K} g_{kb} P_{kb} \leq P_{b,\text{total}} \tag{5.2b}$$

$$\sum_{k=1}^{K} (g_{kb} + c_{kb}) \leq N_{b,T} \tag{5.2c}$$

$$\sum_{b=1}^{B} g_{kb} = 1 \tag{5.2d}$$

The objective function in Eq. (5.2a) is constrained by Eq. (5.2b) such that the total transmit power $P_{b,\text{total}}$ of a base station is not exceeded. Equation (5.2c) ensures that the number of scheduled and zero-forced users is not larger than the number of transmit antennas $N_{b,T}$ at a base station. Equation (5.2d) ensures that a user is at most scheduled by one base station.

Compared to the optimisation formulated in Eq. (4.18) the problem increased in complexity. Although a relaxation of the problem is possible by assuming a uniform power distribution and only optimise with regards to the user grouping, and removing the possibility of cancelling interference to other users from other cells, the problem is still not fitting in an optimisation framework and an exhaustive search is required. For a large number of users, this is infeasible to solve in an acceptable time. In the next section we will therefore focus on feasible user grouping heuristics to find suboptimal user groups in acceptable time.

5.2. Fractional Frequency Reuse

Modern mobile networks are operated in a single-frequency deployment, since this increases the spectral efficiency. Here, all base stations operate in the same frequency band. However, while the spectral efficiency is increased, users at the cell edge suffer from bad channel conditions and become strongly limited by interference. This creates a challenge for the grouping heuristics, since now not only noise, but also interference has to be considered. One strategy for designing a heuristic is

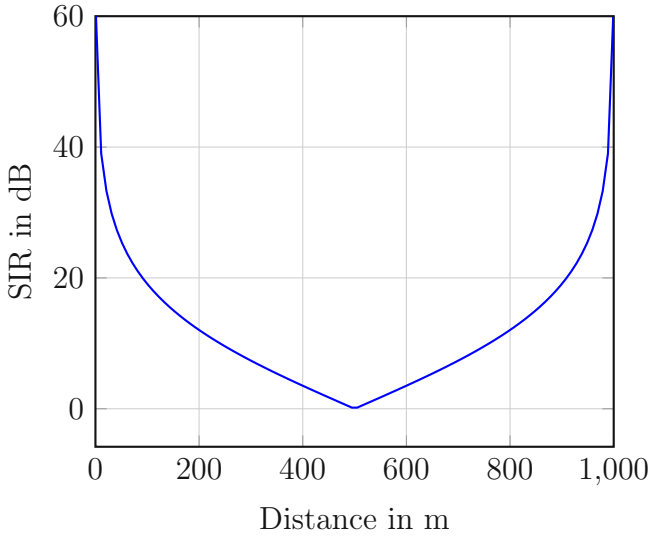


Figure 5.1.: The macroscopic signal-to-interference ratio of two base stations with 1 km distance assuming a free space path loss model. The cell-edge users are interference limited and suffer from bad reception quality due to a low macroscopic SIR.

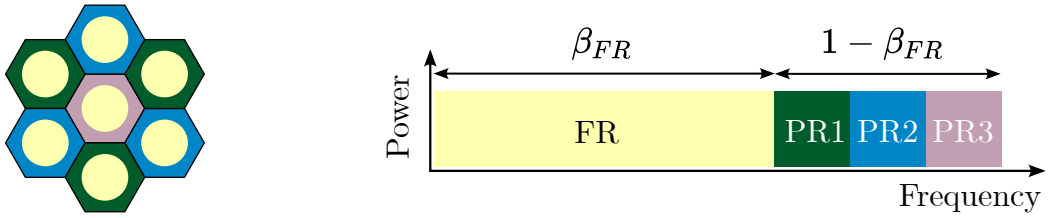
to first find a method to mitigate this additional interference. For example, having two base stations transmitting with equal power and assuming the FSPL model from Eq. (2.1), the macroscopic signal to interference ratio (SIR) is expressed by

$$\gamma(d) = 20 \log\left(\frac{D-d}{d}\right), \tag{5.3}$$

where D is the distance between the base stations and d the distance between the base station and the user along the connecting line. Figure 5.1 represents this equation with a distance D of 1 km, assuming that the user is always connected to the closer base station.

A way to mitigate this high interference, is to use different frequency bands at the cell edge. The frequency band for the inner users is common for all cells, and therefore called the full reuse band. Contrary, the cell edge users are served in a band called the partial reuse band. This strategy for interference mitigation is called fractional frequency reuse (FFR), as shown in [28].

A common example is the reuse-3 pattern for hexagonal grids, shown in Fig. 5.2. It provides three partial reuse bands, making it possible that neighbouring base stations in a hexagonal arrangement never share a partial reuse band. The bandwidth parameter β_{FR} is the fraction of the bandwidth reserved for inner users. For the reuse-3 pattern the rest of the bandwidth is shared evenly by three partial reuse bands.



(a) Hexagonal cells (b) Frequency bands

Figure 5.2.: Fractional frequency reuse-3 pattern. The hexagons from Fig. 5.2a represent cells, where each cell is served by a base station in the hexagon's centre. All base stations share a full reuse band that is represented by the circles. Users at the cell edge are assigned to partial reuse bands. With three partial bands, as shown here, a pattern can be found where base stations mitigate interference to neighbouring cell edge users. In Fig. 5.2b the full reuse and partial reuse bands are drawn in a power spectrum density graph with the threshold and the bandwidth parameter.

As users are now served in different bands the scheduler has to decide which users are at the cell edge. To achieve this, it is assumed that users report their SIR to the base station. This value is compared to a threshold γ_{thr}, and a user k is considered to be at the edge of the cell if its SIR γ_k is smaller; hence the inequality

$$\gamma_k < \gamma_{\text{thr}} \tag{5.4}$$

is fulfilled. Therefore, parameter β_{FR}, together with parameter γ_{thr} determine characteristics of the FFR network. If the threshold is high, many users are at the cell edge, diminishing the throughput as most of the users are scheduled in the smaller partial reuse band. Low thresholds will decrease the number of users in the partial reuse band, and if no fairness scheduling strategy is applied the fairness in the system decreases. The system has to be designed, such that only users that suffer from bad channel conditions are served in the less interfered partial reuse bands. Therefore, the combination of bandwidth share and threshold is expected to control the trade-off between throughput and fairness.

A network with such a reuse-3 pattern has been studied in [24] for the case of single user MIMO transmissions. Hence, only a single user per RB is served by a base station. Here, the author compares the performance of the network in terms of fairness for various values of β_{FR} and γ_{thr}. They investigate how different scheduling strategies, namely round robin and best CQI scheduling, are affecting the throughput and fairness. It was shown that for round robin scheduling it is possible to achieve a throughput and fairness gain compared to a reuse-1 network, while with PF scheduling no throughput gains are possible, without

diminishing the fairness. Nevertheless, FFR provides a way to control overall fairness by adjusting the cell edge threshold and the fractional bandwidth share [24].

Although the single-user case was extensively studied in [24], it is the goal to investigate the performance if multi-user transmissions are performed. For this we first investigate a scenario without FFR and compare how SUS and SPRS perform in a cellular scenario. Then a scenario with FFR is used with SPRS in the best-rate mode and with PF scheduling.

5.3. SUS and SPRS in a Cellular Scenario

This section investigates the performance of the SUS and SPRS algorithm in a cellular scenario without FFR. This is equivalent to defining the parameters $\beta_{\mathrm{FR}} = 1$ and $\gamma_{\mathrm{thr}} = -\infty$. Hence, all bandwidth is allocated for the full reuse band, and all users are inner users. The base stations are placed on a hexagonal grid as shown in Fig. 5.3. The users are distributed by a random Poisson point process and are surrounded by an additional ring of interfering base stations. The simulation parameters are listed in Table 5.1. Each base station is equipped with 32 antennas and the SPRS algorithm is executed with $P_S = 0.4$ and $P_S = 0.7$ and SUS with $\alpha = 0.2$ and $\alpha = 0.5$.

Before SPRS and SUS are ready to operate in a cellular scenario one additional challenge has to be solved. The cellular optimisation problem in Eq. (5.2) is more complex compared to the single cell case due to an additional interference term. The interference is dependent on the scheduling decisions of all base stations, requiring that the optimisation is performed jointly. To keep the complexity low the SUS or SPRS heuristic will be performed independently for each base station. Hence, the interference created by the neighbouring base stations is unknown. To solve this, the grouping heuristics are provided with an estimate of the macroscopic SINR, which is assumed to be delivered by the user together with the CQI. Using this modification the two algorithms are ready for cellular scenarios.

Figure 5.4 shows the ECDF of the sum-throughput per slot for SUS and SPRS where the throughput is summed over all base stations. It is visible that SPRS is delivering the highest rates and the influence of the parameters α and P_S is observable. This is similar to the behaviour of the single cell scenario in Section 4.5 as seen in Fig. 4.6. Directly comparing this scenario and the scenario in Section 4.5 reveals that the overall throughput is higher in the multi-user scenario. This is explained by the larger number of base stations and therefore antennas in the system, which enables scheduling more users at the same time. So a direct comparison between these two scenarios is unfair. For SPRS again a larger P_S decreases the throughput, while for SUS the parameter value $\alpha = 0.5$ performs

Table 5.1.: Simulation Parameters for Section 5.3

Base stations	17
Transmit antennas	32
Users	100
Small-scale fading model	PedA $5\,\mathrm{km\,h^{-1}}$
Bandwidth	9 MHz
Number of slots per chunk	200
Number of chunks	18

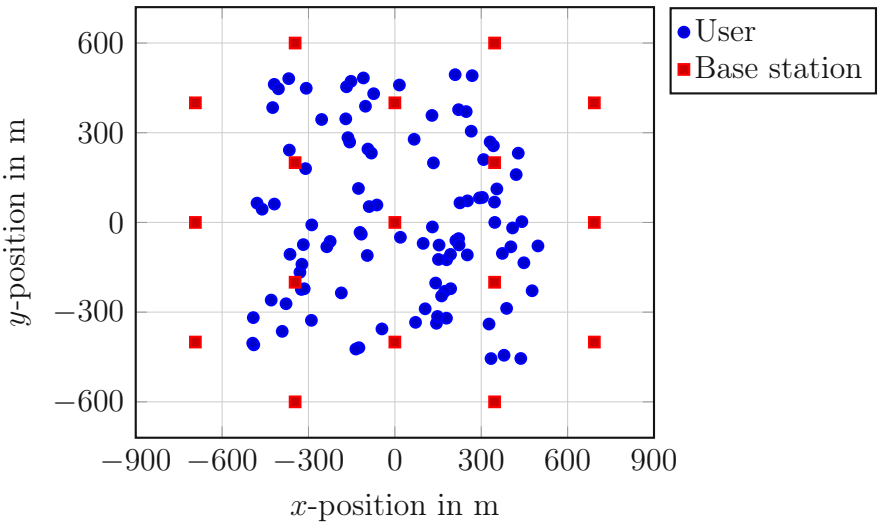


Figure 5.3.: User and base station positions. User are distributed according to a Poisson 2D process and base stations in a hexagonal grid arrangement.

better than for $\alpha = 0.2$. This is surprising as SUS performed better with $\alpha = 0.2$ in single cell case. This emphasises the sometimes surprising α dependence of the throughput when using the SUS algorithm.

5.4. Performance with FFR

We are now interested in the performance if FFR is enabled. The cells utilise a reuse-3 pattern, as described in Fig. 5.2. User grouping is performed independently for the inner users and the cell edge users. Simulations are performed with the SPRS grouping algorithm, first for best-rate and then for PF scheduling and the difference between them is analysed. The FFR bandwidth β_{FR} and the threshold

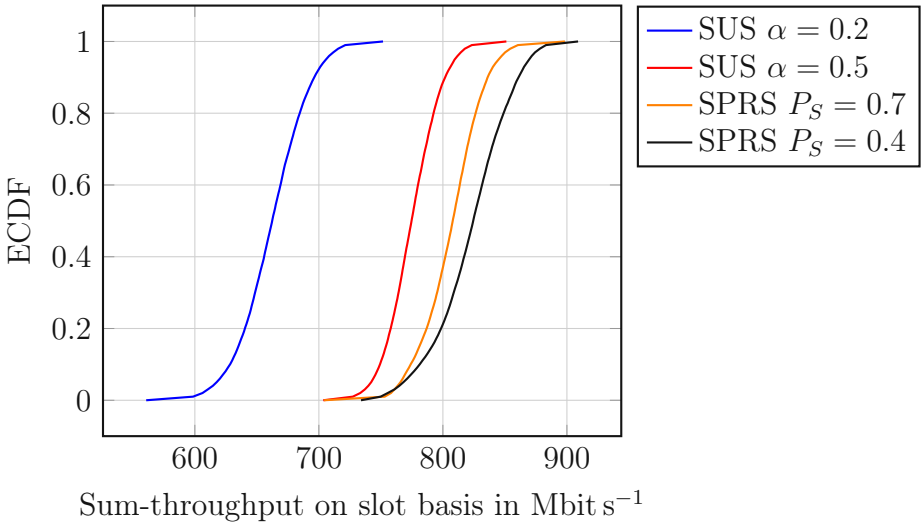


Figure 5.4.: Slot sum-throughput for the comparison of SUS and SPRS in the cellular scenario without FFR. The throughput is summed over all base stations.

γ_{thr} are parameters and swept in the intervals

$$\beta_{\text{FR}} \in [0, 1], \quad \gamma_{\text{thr}} \in [-3\,\text{dB}, 15\,\text{dB}], \tag{5.5}$$

where the step size is chosen to obtain ten values for both of them. Therefore, in total 100 simulations are performed.

Performance is evaluated using the mean throughput, Jain's fairness index, the edge throughput and the peak throughput. Edge throughput is defined as the fifth percentile (5 percent of the values are below this value), and peak throughput is the 95th percentile. This gives an idea of the throughput for the lowest and highest five percent values.

Best-Rate Scheduling

The scenario is first simulated with the SPRS best-rate algorithm for all combinations of the sweep parameters. Figure 5.5 depicts the throughput and fairness results of the 100 performed simulations over the parameters β_{FR} and γ_{thr}.

By analysing Fig. 5.5a, it can be observed that the highest mean throughput is achieved if all bandwidth is assigned to the full reuse band. This is intuitive, since we expect that the inner users, that are near to the base station and have therefore good SINR values, will significantly contribute to the mean and profit from the additional bandwidth. It is clearly visible that the throughput is also dependent on the threshold; as it grows bigger, most of the users will be considered to be at the cell edge and scheduled in the partial reuse band, resulting in a smaller

bandwidth and hence smaller throughput.

In Fig. 5.5b, Jain's fairness index is shown to be mainly controlled by the bandwidth share. The behaviour of Jain's fairness is contrary to the mean throughput, as it grows larger with an increasing partial reuse bandwidth. Therefore, this shows that FFR offers a trade-off between fairness and throughput, controlled by the parameters β_{FR} and γ_{thr}.

Furthermore, Fig. 5.5c shows that in most cases the edge throughput is low and close to $0\,\mathrm{Mbit\,s^{-1}}$, so many users are experiencing no throughput. Significant edge throughput is achieved when using a threshold of 6 dB. Beside this, the edge throughput rapidly decreases for smaller or larger threshold values. This is explained by looking at Fig. 5.6, which shows the number of users considered as cell edge users over the threshold, and reveals that for small threshold values, no users are at the cell edge. Therefore, this is a waste of bandwidth, since no users are in the partial reuse band. On the other hand, if the threshold is high, all users are scheduled in the partial reuse band. The best-rate scheduler then decides to only select the nearest users to maximise the rate, since they have the highest SINR. Also here bandwidth is wasted, since the full reuse band is mostly unused and only a third of the available bandwidth is used.

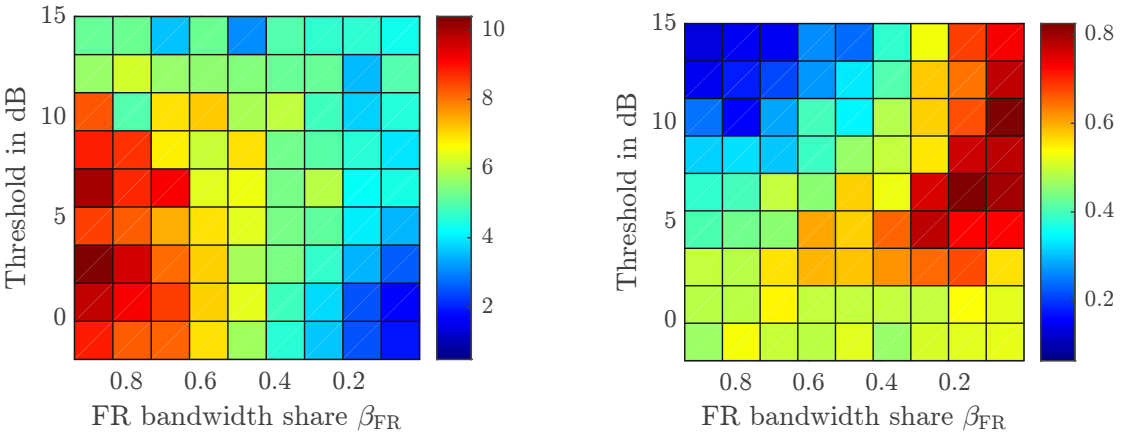
Surprisingly, Fig. 5.5d, shows that a large full reuse band and a high threshold lead to the highest peak values. This observation reveals that the highest peak throughput is not achieved with the same parameters as the highest mean throughput. At the second thought it is explained by the number of users in the full reuse band, since now only few users very close to the base station remain in the full reuse band, as shown in Fig. 5.6. In addition to the large bandwidth, which they can exclusively use, they also form smaller user groups due to the smaller number of users in the inner cell, so individual users experience high peak values. Although a few users experience high peak throughput the mean is not growing, because the majority of users are in the partial reuse band and experience low throughput.

Proportional Fair Scheduling

The simulations are repeated for the PF scheduling version of SPRS as defined in Section 4.4. Therefore, we set the parameter of the SPRS algorithm to $a = 1$ and $b = 1$ and all other parameters are left unchanged and again the simulation is performed for 100 parameter combinations and results are shown in Fig. 5.7.

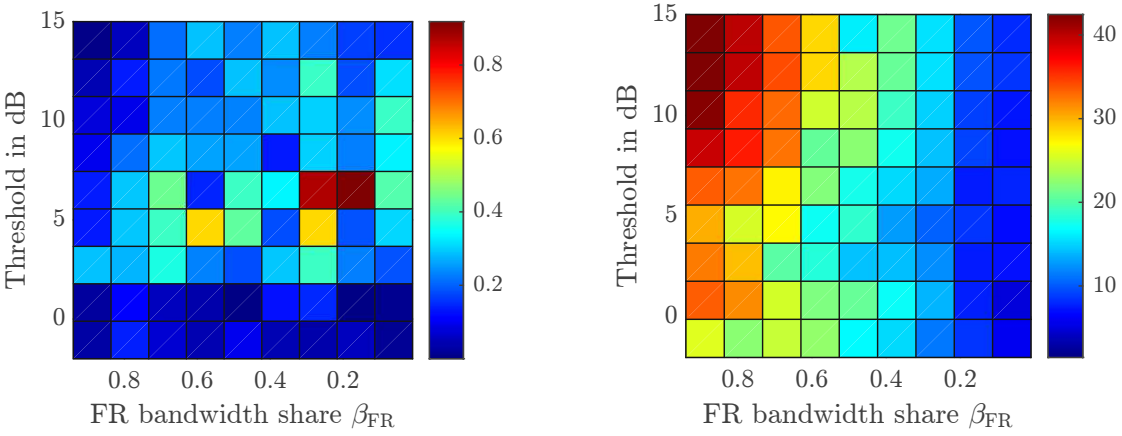
For the mean throughput in Fig. 5.7a the throughput is smaller than for the best-rate case, which is expected, as scheduling is now also performed with respect to fairness.

Also for Jain's fairness in Fig. 5.7b the results show the same trend as in the best-rate scheduling case. Again, it is seen that the parameters β_{FR} and γ_{thr} control the trade-off between throughput and fairness. The major difference is that more parameter combinations lead to a higher fairness, leading to the conclusion that



(a) Mean throughput in Mbit s^{-1}

(b) Jain's fairness



(c) Edge throughput in Mbit s^{-1}

(d) Peak throughput in Mbit s^{-1}

Figure 5.5.: Results for cellular best-rate scheduling. The graphs a,c,d show the user throughput encoded in a colour map. The graph b shows Jain's fairness index encoded in a colour map. The axes show the bandwidth share and the threshold parameter.

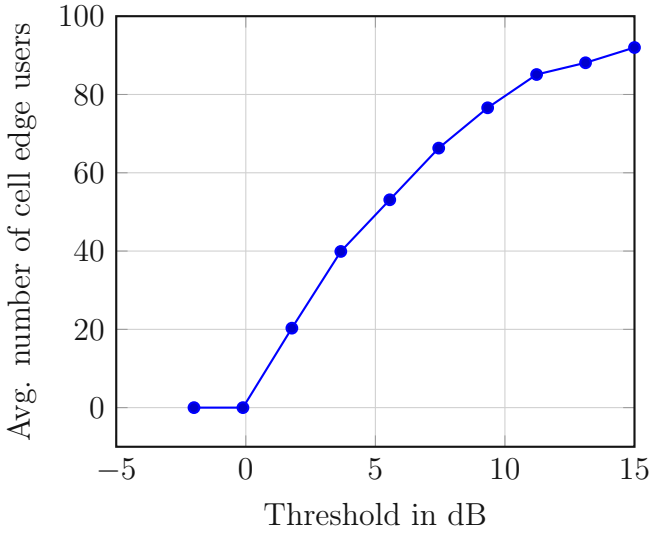


Figure 5.6.: Average number of cell edge users over the threshold γ_{thr} extracted from the simulation results.

the optimisation towards fairness is also working in a FFR deployment.

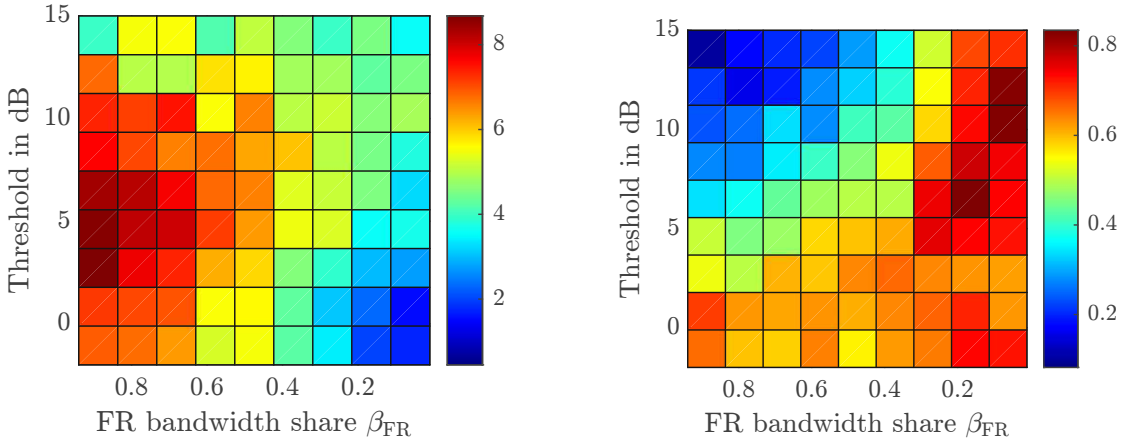
Major differences can be seen in the edge throughput results in Fig. 5.7c, where compared to the best-rate case, edge users experience significant throughput values for much more parameter constellations. Interestingly, significant contributions show up in a diagonal shape in the figure. Before, the algorithm decided to only schedule the most favourable cell edge users, which means that the closest ones were picked. Now, the balancing behaviour of the fairness algorithm also picks the weaker users at the far-cell edge, enabling more users to transmit, also with worse channel conditions.

The peak throughput in Fig. 5.7d behaves almost identically to the best-rate case, and the same reasoning as in the best-rate case can be applied to the question why maximum peak and mean throughput values appear in different parameter constellations.

Operating Points

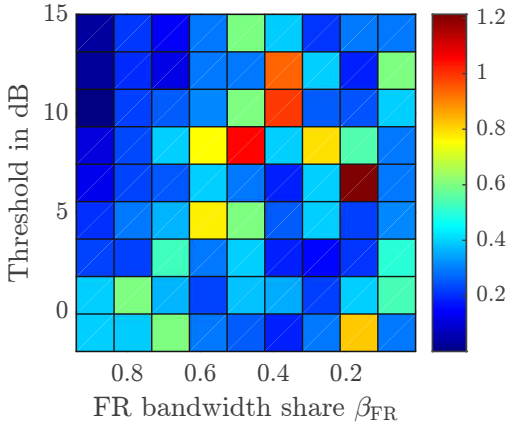
In Table 5.2 three operating points are shown with their corresponding results for best-rate and proportional fair SPRS scheduling and named (a), (b) and (c).

Operating point (a) uses approximately a third of the bandwidth for the full reuse band with a threshold of 3.6 dB. As can be seen, using the PF SPRS scheduler does not bring any benefit at this point, since it only increases the edge throughput by a small amount while decreasing the mean throughput. However, a high mean throughput can be achieved with the best-rate scheduler while maintaining a high edge and peak throughput. The Jain's fairness index for both algorithms is similar

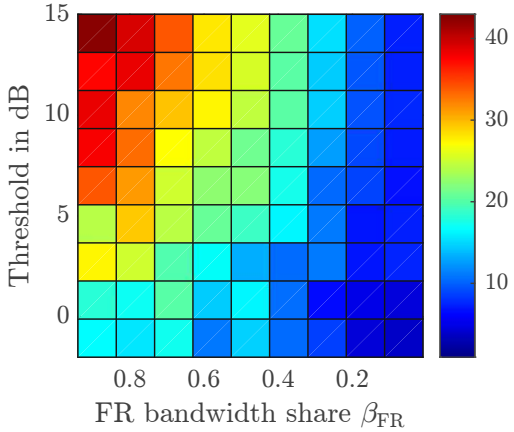


(a) Mean throughput in Mbit s^{-1}

(b) Jain's fairness



(c) Edge throughput in Mbit s^{-1}



(d) Peak throughput in Mbit s^{-1}

Figure 5.7.: Results for cellular proportional fair scheduling. The graphs a,c,d show the user throughput encoded in a colour map. The graph b shows Jain's fairness index encoded in a colour map. The axes show the bandwidth share and the threshold parameter.

Table 5.2.: Three selected operating points for best-rate and proportional fair SPRS scheduling with $P_S = 0.4$ in a fractional frequency reuse scenario.

Name	β_{FR}	γ_{thr}	Metric	Best-Rate	Proportional Fair
(a)	0.63	3.6 dB	Mean	7.15 Mbit s^{-1}	6.93 Mbit s^{-1}
			Edge	0.60 Mbit s^{-1}	0.78 Mbit s^{-1}
			Peak	17.17 Mbit s^{-1}	20.82 Mbit s^{-1}
			Fairness	0.6	0.58
(b)	0.32	9.3 dB	Mean	6.03 Mbit s^{-1}	5.15 Mbit s^{-1}
			Edge	0.29 Mbit s^{-1}	1.00 Mbit s^{-1}
			Peak	19.38 Mbit s^{-1}	19.22 Mbit s^{-1}
			Fairness	0.43	0.48
(c)	0.94	−0.1 dB	Mean	9.22 Mbit s^{-1}	7.06 Mbit s^{-1}
			Edge	0.11 Mbit s^{-1}	0.60 Mbit s^{-1}
			Peak	31.59 Mbit s^{-1}	17.18 Mbit s^{-1}
			Fairness	0.49	0.62

and has a value of 0.6.

At operating point (b), two thirds of the bandwidth are reserved for the full reuse band the threshold is at 9.3 dB. In this case, the PF scheduler is beneficial as it increases the edge throughput by a factor of 3.48, while keeping the peak throughput at the same level and decreasing the throughput by a factor of 0.8.

The majority of the bandwidth was allocated for the full reuse band at operating point (c), and the threshold is very low. Here, the best-rate algorithm achieves a high mean user throughput, while PF delivers a high Jain's fairness index.

This leads to the conclusion that FFR in combination with SPRS allows finer control and trade-off for fairness and mean, edge, and peak throughput. If the primary goal is to increase the fairness, FFR in combination with best-rate scheduling can be employed. Also, using a PF scheduler without FFR at all increases fairness. This is not the case for the edge throughput, which is influenced more by the scheduler type than the FFR parameters. If edge throughput should be increased significantly it is necessary to employ PF scheduling. The most effective increase in edge throughput is achieved at operating point (c).

6. Conclusion and Outlook

Within this thesis an analysis of multi-user MIMO transmission is given by comparing the performance of user grouping algorithms in system level simulations. Additionally, an overview of the Vienna 5G System Level Simulator is presented to ease the understanding of the utilised methodology. The main contributions of the work are a multi-user system level abstraction model to perform system level simulations and a novel user grouping algorithm called SPRS.

It is shown how the mathematical expressions of the proposed multi-user transmission abstraction model for the Vienna 5G system level simulator are simplified if ZF receiver filters and ZF precoding are used, enabling the simulation of large-scale networks.

By deriving a system model for a downlink ZF multi-user single cell scenario, an optimisation problem was formulated to optimise the sum-throughput with respect to the user grouping. After deriving the optimisation problem, it was shown that it does not classify as a convex problem and an infeasible exhaustive search is required.

This lead to the necessity of finding heuristics to obtain user groups in acceptable time. A novel heuristic called SPRS is proposed that performs a greedy search with a random sampling step to find groups that deliver high throughput. It is shown, by conducting simulations, that SPRS delivers higher sum-throughput than the well-known SUS algorithm by a slight increase in computational complexity.

Furthermore, an SPRS extension for fairness is proposed that allows to trade-off the user grouping optimisation between rate or fairness. This allows to apply scheduling strategies such as best-rate, proportional fair, or max-min scheduling.

It is also investigated how the user grouping heuristics perform in a cellular scenario with and without the deployment of FFR. Here, a scheduling strategy is proposed in which users are divided into a full reuse and partial reuse band and grouped independently with the SPRS algorithm. It was investigated how the throughput and fairness changes with respect to the FFR and SPRS parameters. It is shown that FFR in combination with a proportional fair SPRS user grouping allows to control the fairness among users.

For future work, an investigation of the proposed SPRS algorithm with respect to non-power delay profile-based channel models is of interest. Especially, models that do not only rely on statistics could be of interest, for example geometric models that take the environment and spatial correlation into account.

Furthermore, an extension of SPRS to support users with multiple antennas and multiple data streams per user is of interest. Also, supporting other precoding strategies, for example minimum mean square error (MMSE) based precoding, would be valuable.

A. Abbreviations

5G	fifth generation
LQM	link quality model
LPM	link performance model
RB	resource block
SNR	signal to noise ratio
SINR	signal to interference and noise ratio
SIR	signal to interference ratio
CQI	channel quality indicator
BLER	block error rate
OFDM	orthogonal frequency-division multiplexing
TDD	time division duplex
MIMO	multiple-input and multiple-output
DPC	dirty paper precoding
ESM	effective SINR mapping
MIESM	mutual information effective SINR mapping
AWGN	additive white Gaussian noise
SDMA	space-division multiple access
SUS	semi-orthogonal user selection
SPRS	single path random sampling
PF	proportional fair

ZF	zero-forcing
RF	radio frequency
FFR	fractional frequency reuse
ZFS	zero-forcing with selection
GUSS	greedy user selection with swap
FSPL	free space path loss
ECDF	empirical cumulative distribution function
CSI	channel state information
PF	proportional fair
MMSE	minimum mean square error

B. References

[1] "Ericsson Mobility Report June 2022," ERICCSON, Tech. Rep., Jun. 2022.

[2] J. Mietzner, R. Schober, L. Lampe, W. H. Gerstacker, and P. A. Hoeher, "Multiple-antenna techniques for wireless communications - a comprehensive literature survey," *IEEE Communications Surveys & Tutorials*, vol. 11, no. 2, pp. 87–105, 2009. DOI: 10.1109/SURV.2009.090207.

[3] A. F. Molisch, *Wireless communications*, eng, 2. ed.. Weinheim: Wiley [u.a.], 2011, ISBN: 0470741864.

[4] E. Telatar, "Capacity of multi-antenna Gaussian channels," *European Transactions on Telecommunications*, vol. 10, no. 6, pp. 585–595, 1999. DOI: 10.1002/ett.4460100604.

[5] E. Jorswieck and A. Sezgin, "Impact of spatial correlation on the performance of orthogonal space-time block codes," *IEEE Communications Letters*, vol. 8, no. 1, pp. 21–23, 2004. DOI: 10.1109/LCOMM.2003.822516.

[6] S. Schwarz and M. Rupp, "Evaluation of Distributed Multi-User MIMO-OFDM With Limited Feedback," *IEEE Transactions on Wireless Communications*, vol. 13, no. 11, pp. 6081–6094, 2014. DOI: 10.1109/TWC.2014.2346191.

[7] H. Weingarten, Y. Steinberg, and S. Shamai, "The capacity region of the Gaussian MIMO broadcast channel," in *International Symposium onInformation Theory, 2004. ISIT 2004. Proceedings.*, 2004, pp. 174–. DOI: 10.1109/ISIT.2004.1365211.

[8] M. Costa, "Writing on dirty paper (corresp.)," *IEEE Transactions on Information Theory*, vol. 29, no. 3, pp. 439–441, 1983. DOI: 10.1109/TIT.1983.1056659.

[9] G. Caire and S. Shamai, "On the achievable throughput of a multiantenna Gaussian broadcast channel," *IEEE Transactions on Information Theory*, vol. 49, no. 7, pp. 1691–1706, 2003. DOI: 10.1109/TIT.2003.813523.

[10] T. Yoo and A. Goldsmith, "On the optimality of multiantenna broadcast scheduling using zero-forcing beamforming," *IEEE Journal on Selected Areas in Communications*, vol. 24, no. 3, pp. 528–541, 2006. DOI: 10.1109/JSAC.2005.862421.

[11] Z. Wang, Y. Cao, D. Zhang, X. Hua, P. Gao, and T. Jiang, "User selection for MIMO downlink with digital and hybrid maximum ratio transmission," *IEEE Transactions on Vehicular Technology*, vol. 70, no. 10, pp. 11 101–11 105, 2021. DOI: 10.1109/TVT.2021.3110726.

[12] M. K. Müller, F. Ademaj, T. Dittrich, *et al.*, "Flexible multi-node simulation of cellular mobile communications: The Vienna 5G System Level Simulator," *EURASIP Journal on Wireless Communications and Networking*, vol. 2018, no. 1, p. 17, Sep. 2018. DOI: 10.1186/s13638-018-1238-7.

[13] S. Huang, H. Yin, J. Wu, and V. C. M. Leung, "User selection for multiuser MIMO downlink with zero-forcing beamforming," *IEEE Transactions on Vehicular Technology*, vol. 62, no. 7, pp. 3084–3097, 2013. DOI: 10.1109/TVT.2013.2244105.

[14] J. Wang, D. J. Love, and M. D. Zoltowski, "User selection with zero-forcing beamforming achieves the asymptotically optimal sum rate," *IEEE Transactions on Signal Processing*, vol. 56, no. 8, pp. 3713–3726, 2008. DOI: 10.1109/TSP.2008.919096.

[15] G. Dimic and N. Sidiropoulos, "On downlink beamforming with greedy user selection: Performance analysis and a simple new algorithm," *IEEE Transactions on Signal Processing*, vol. 53, no. 10, pp. 3857–3868, 2005. DOI: 10.1109/TSP.2005.855401.

[16] J. Jiang, R. Buehrer, and W. Tranter, "Greedy scheduling performance for a zero-forcing dirty-paper coded system," *IEEE Transactions on Communications*, vol. 54, no. 5, pp. 789–793, 2006. DOI: 10.1109/TCOMM.2006.874008.

[17] J. Liu, X. She, and L. Chen, "A low complexity capacity-greedy user selection scheme for zero-forcing beamforming," in *VTC Spring 2009 - IEEE 69th Vehicular Technology Conference*, 2009, pp. 1–5. DOI: 10.1109/VETECS.2009.5073303.

[18] C.-K. Jao, C.-Y. Wang, T.-Y. Yeh, *et al.*, "Wise: A system-level simulator for 5G mobile networks," *IEEE Wireless Communications*, vol. 25, no. 2, pp. 4–7, 2018. DOI: 10.1109/MWC.2018.8352614.

[19] G. Nardini, G. Stea, A. Virdis, and D. Sabella, "Simu5G: A system-level simulator for 5G networks," in *Proceedings of the 10th International Conference on Simulation and Modeling Methodologies, Technologies and Applications*, SCITEPRESS - Science and Technology Publications, 2020. DOI: 10.5220/0009826400680080. [Online]. Available: https://doi.org/10.5220/0009826400680080.

[20] A. Varga, "Omnet++," in *Modeling and Tools for Network Simulation*, K. Wehrle, M. Güneş, and J. Gross, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 35–59, ISBN: 978-3-642-12331-3. DOI: 10.1007/978-3-642-12331-3_3. [Online]. Available: https://doi.org/10.1007/978-3-642-12331-3_3.

[21] 3. G. P. P. (3GPP), "High Speed Downlink Packet Access: UE Radio Transmission and Reception," 3rd Generation Partnership Project (3GPP), Tech. Rep., 2002.

[22] Y. R. Zheng and C. Xiao, "Simulation models with correct statistical properties for Rayleigh fading channels," *IEEE Transactions on Communications*, vol. 51, no. 6, pp. 920–928, 2003. DOI: 10.1109/TCOMM.2003.813259.

[23] B. Fleury, "An uncertainty relation for WSS processes and its application to wssus systems," *IEEE Transactions on Communications*, vol. 44, no. 12, pp. 1632–1634, 1996. DOI: 10.1109/26.545890.

[24] J. Colom Ikuno, "System level modeling and optimization of the LTE downlink," Ph.D. dissertation, 2013.

[25] E. Tuomaala and H. Wang, "Effective SINR approach of link to system mapping in OFDM/multi-carrier mobile network," in *2005 2nd Asia Pacific Conference on Mobile Technology, Applications and Systems*, 2005, 5 pp.–5. DOI: 10.1109/MTAS.2005.243791.

[26] C. E. Shannon and W. Weaver, *The mathematical theory of communication*, eng. Urbana, Ill: The Univ. of Illinois Press, 1949.

[27] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004. DOI: 10.1017/CBO9780511804441.

[28] N. Himayat, S. Talwar, A. Rao, and R. Soni, "Interference management for 4G cellular standards [WIMAX/LTE UPDATE]," *IEEE Communications Magazine*, vol. 48, no. 8, pp. 86–92, 2010. DOI: 10.1109/MCOM.2010.5534591.

C. List of Figures

2.1. Resource Grid 11
2.2. Time Line of the Simulator 12
2.3. DL and UL Slots 13
2.4. Simulation Flow 15
2.5. LQM and LPM 15
2.6. Link Performance Model 16
2.7. Space-Division Multiple Access 17
2.8. Beamforming 18

3.1. Interference with Multiple Cells 20
3.2. LQM Received Signal Block Diagram 21

4.1. Multi-User MIMO System 26
4.2. SNR Correlation Penalty 30
4.3. Behaviour of Precoders in 3D 31
4.4. SPRS Illustration 34
4.5. Single Cell Element Positions 37
4.6. Comparison of SUS and SPRS 38
4.7. Sum-Throughput SPRS and Exhaustive Search 39
4.8. SPRS Fairness Comparison 40
4.9. Sum-throughput fairness 41
4.10. SPRS Time Complexity 43

5.1. Cell Edge Interference 46
5.2. Fractional Frequency Reuse 47
5.3. Cellular Scenario Positions 49
5.4. Slot Sum-Throughput Cellular without FFR 50
5.5. Cellular Best-Rate Scheduling Throughput 52
5.6. Number of Cell Edge Users 53
5.7. Cellular Proportional Fair Scheduling Throughput 54

D. List of Tables

4.1. Simulation Parameters for Section 4.5 37
4.2. Simulation Parameters for Section 4.6 38
4.3. Simulation Parameters for Section 4.7 40
4.4. Simulation Parameters for Section 4.8 42

5.1. Simulation Parameters for Section 5.3 49
5.2. Three Selected FFR Operation Points 55