

Performance Analysis of Zero-Forcing Beamforming Strategies for the Uplink and Downlink of MU-MIMO Systems with Multi-Antenna Users

by

Joao Paulo Pinto Galdino Marques

A thesis
presented to the University of Waterloo
in fulfilment of the
thesis requirement for the degree of
Master of Applied Science
in
Electrical and Computer Engineering

Waterloo, Ontario, Canada, 2025

© Joao Paulo Pinto Galdino Marques 2025

Author's Declaration

This thesis consists of material all of which I authored or co-authored: see Statement of Contributions included in the thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Statement of Contributions

This thesis contains some material that was taken from a multi-author paper, in which I am the first author. The paper is:

- J. P. P. G. Marques, K. Danilchenko, and C. Rosenberg, “Performance Evaluation of MU-MIMO Systems with Multi-Antenna Users for Different Precoding Strategies,” in *2025 IEEE Wireless Communications and Networking Conference (WCNC)*, 2025, pp. 1-6.

Abstract

Multi-User Multiple-Input Multiple-Output (MU-MIMO) systems provide high performance through the transmission of multiple data streams to users (resp. from users) on the Downlink (DL) (resp. Uplink (UL)) at the same frequency and time. Multiple antennas at the users can further improve performance by enabling the choice between diversity (using the multiple antennas per user to create one data stream with potentially higher rates) and multiplexing (utilizing the antennas to create multiple independent data streams per user). To achieve the potential of MU-MIMO systems with Multi-Antenna (MA) users, the Radio Resource Management (RRM) processes must be carefully executed. In this work, we focus on the DL and UL of realistic MU-MIMO systems with MA users.

The necessary RRM processes are user selection, digital beamforming (called here precoding), Power Allocation (PA), Power Distribution (PD) and Modulation and Coding Scheme (MCS) selection. We consider the well-known family of precoding strategies called Zero-Forcing (ZF), which nullifies inter-stream interference. There are three primary ZF-based precoding strategies: Coordinated-Transmit-Receive-1 (CTR₁), where we enable only the strongest stream per scheduled user, Block Diagonalization (BD), where all possible streams are enabled per selected user, and Coordinated-Transmit-Receive-Flexible (CTR_F), which allows a flexible stream allocation per scheduled user. The latter has the potential for increased performance compared to the other strategies at the cost of higher complexity.

Throughout this work, we conduct novel offline (where run time is unimportant) studies to compute the performance of MU-MIMO systems under Proportional Fairness (PF) with those precoding strategies on the DL and UL in realistic systems characterized by 3rd Generation Partnership Project (3GPP)-based scenarios where the Base-Station (BS) has a large number of antennas and employs practical MCSs. To enable the offline study on the DL, we employ heuristics from the literature and adapt them to consider practical systems with MCSs and PF, whereas we propose the necessary tools for the offline UL study. The offline studies allow us to execute performance comparisons of MU-MIMO systems with the precoding strategies (BD, CTR₁ and CTR_F), which can guide the design of real-time RRM heuristics for both the DL and UL. The DL results indicate that CTR_F outperforms the other strategies under PF. However, depending on the network parameters, either CTR₁ or BD could replace CTR_F given their comparable performance. On the UL, the conclusions are similar, but in the 3GPP Urban Macro scenario, CTR₁ presents a comparable performance to CTR_F over all considered network parameters, emerging as an alternative to CTR_F.

Acknowledgements

First, I would like to express my sincere gratitude to my supervisor, Professor Catherine Rosenberg, for her invaluable guidance. Thanks to her mentorship, I was able to grow academically and become a more well-rounded person. I am also thankful to all the professors and colleagues at the University of Waterloo, whose support and shared experiences made this journey truly memorable.

I am deeply grateful to my mother, Rosineide, my sister, Maria Eduarda, and my partner, Haein. I have come this far owing to their support, encouragement, and presence.

I also want to thank all my friends and relatives, especially Belarmino, Elly, Euan, Felipe, Fernanda, Gustavo, Igor, Jingyeong, Kim, Kimia, Maryam, Matheus, Milena, Nathalia, Netinho, Paolo, Pedro, Samir, and many others, whose friendship and support were irreplaceable in this achievement.

Table of Contents

List of Figures	viii
List of Tables	x
List of Abbreviations	xi
1 Introduction	1
1.1 Motivation and Research Questions	3
1.2 Contributions	5
1.3 Outline	6
2 Literature Survey and Background	7
2.1 Downlink	7
2.2 Uplink	9
2.3 System Model	10
2.4 Scenarios and Channel Model	11
3 Downlink	12
3.1 Introduction	12
3.2 Per-PRB RRM Procedure	14
3.3 Long-term Performance Computation	17
3.4 Numerical Results	19

3.4.1	Proportional Fairness	20
3.4.2	Sum-rate Maximization	21
3.5	Conclusion	23
4	Uplink	25
4.1	Introduction	25
4.2	Uplink Greedy-up Search Heuristic	28
4.3	Long-term Performance Computation	32
4.4	Search Rate-reusing Method	34
4.5	Numerical Results	36
4.5.1	Precoding Strategies Evaluation	37
4.5.2	Miscellaneous	39
4.6	Conclusion	39
5	Conclusion	47
5.1	Summary	47
5.2	Potential Research Directions	48
5.2.1	Extending The Offline Study	49
5.2.2	Real-time RRM Study	49
	Letter of Copyright Permission	50
	References	51

List of Figures

3.1	Flow diagram of the PRB-level procedure for a given scenario, precoding strategy, network parameters $\mathcal{N}$, realization $\omega(\Delta)$, and PRB.	17
3.2	Top figures: geometric mean of the per-TS rates vs. numbers of users for the three precoding strategies for RMa (a), UMa (b), UMi (c) for 100, 64 and 16 antennas at the BS. Bottom figures: ratios in the geometric mean rate between BD and CTR_F , and between CTR_1 and CTR_F are shown. . .	21
3.3	Empirical distribution of the stream allocation per selected user per PRB for 20, 40 and 80 users, for CTR_F , PF, 100 BS antennas, and UMa.	22
3.4	Ratios in the sum rate per TS between BD and CTR_F , and between CTR_1 and CTR_F for different numbers of users for RMa (a), UMa (b) and UMi (c) with 100, 64 and 16 BS antennas.	22
3.5	Empirical distribution of the stream allocation per selected user per PRB for 20, 40 and 80 users, for CTR_F , SR maximization, 100 BS antennas, and UMa.	23
4.1	RMa scenario: $M_B = 100$ and $M_B = 64$ antennas at the BS, $M_U = 4$ antennas at the users, $P_U = 50$ mW. Top figure: GM rates in Mbps per TS vs. number of users for CTR_F , BD and CTR_1 . Bottom figure: ratios in the GM rates between BD and CTR_F , and between CTR_1 and CTR_F	41
4.2	RMa scenario: empirical distribution of the stream allocation per scheduled user per PRB given by CTR_F for $ \mathcal{U} \in \{10, 40\}$ users with $M_B = 64$ BS antennas, $M_U = 4$ antennas at the users, and $P_U = 50$ mW.	42
4.3	UMa scenario: $M_B = 100$ and $M_B = 64$ antennas at the BS, $M_U = 4$ antennas at the users, and $P_U = 1$ mW. Top figure: GM rates in Mbps per TS vs. number of users for CTR_F , BD and CTR_1 . Bottom figure: ratios in the GM rates between BD and CTR_F , and between CTR_1 and CTR_F	43

4.4	UMa scenario: empirical distribution of the stream allocation per selected user per PRB given by CTR_F for $ \mathcal{U} \in \{10, 40\}$ users with $M_B = 64$ BS antennas, $M_U = 4$ antennas at the users, and $P_U = 1$ mW.	44
4.5	UMa scenario: $M_B = 64$ antennas at the BS, $M_U = 4$, $M_U = 2$ and $M_U = 1$ antennas at the users, $P_U = 5$ mW. Top figure: GM rates in Mbps per TS vs. number of users for CTR_F , BD and CTR_1 . Bottom figure: ratios in the GM rates between BD and CTR_F , and between CTR_1 and CTR_F	45
4.6	UMa scenario: $M_B = 64$ antennas at the BS, $M_U = 4$ antennas at the users, $P_U = 5$ mW and $P_U = 1$ mW. Top figure: GM rates in Mbps per TS vs. number of users for CTR_F , BD and CTR_1 . Bottom figure: ratios in the GM rates between BD and CTR_F , and between CTR_1 and CTR_F	46

List of Tables

2.1	Related works on the DL. Notation: Rate Function (RF), Greedy-up Search (GUS), Shannon Rate Function (SRF), Subchannels (SBs), Criterion (CR), Sum-rate Maximization (SRM).	8
2.2	Related works on the UL. Notation: Rate Function (RF), Greedy-up Search (GUS), Shannon Rate Function (SRF), Subchannels (SBs), Criterion (CR), Sum-power Minimization (SPM), Precoding Strategy (PS), User Antenna Setting (UAS), Minimum Rate (MR).	10
2.3	Employed MCS-based rate mapping. Notation: Spectral Efficiency (SE). .	11
3.1	Computation parameters for the DL (RB: reporting block, #: number). . .	20
4.1	Computation parameters for the UL (RB: reporting block, #: number). . .	36

List of Abbreviations

- 3GPP** 3rd Generation Partnership Project.
- BD** Block Diagonalization.
- BS** Base-Station.
- CSI** Channel State Information.
- CTR₁** Coordinated-Transmit-Receive-1.
- CTR_F** Coordinated-Transmit-Receive-Flexible.
- DL** Downlink.
- GM** Geometric Mean.
- GUS** Greedy-Up Search.
- GUS-BD** Greedy User Selection - Block Diagonalization.
- LOS** Line-Of-Sight.
- MA** Multi-Antenna.
- MCS** Modulation and Coding Scheme.
- MDET** Multiple Dominant Eigenmode Transmission.
- MET** Multiple Eigenmode Transmission.
- MMSE** Minimum Mean Square Error.
- MRT** Maximum Ratio Transmission.

MU-MIMO Multi-User Multiple-Input Multiple-Output.

NLOS Non-Line-Of-Sight.

OFDMA Orthogonal Frequency-Division Multiple Access.

PA Power Allocation.

PD Power Distribution.

PF Proportional Fairness.

PRB Physical Resource Block.

RMa Rural Macro.

RQ Research Question.

RRM Radio Resource Management.

SA Single-Antenna.

SIC Successive Interference Cancellation.

SNR Signal-to-Noise Ratio.

SR Sum-Rate.

TDD Time-Division Duplexing.

TM Target Metric.

TPM Two-Step Power Management.

TS Time-Slot.

UL Uplink.

UMa Urban Macro.

UMi Urban Micro.

WF Water-Filling.

WSR Weighted Sum-Rate.

ZF Zero-Forcing.

Chapter 1

Introduction

Wireless Multi-User Multiple-Input Multiple-Output (MU-MIMO) systems offer enhanced performance compared to their single-user counterparts by allowing simultaneous transmissions of data streams to or from multiple users at the same time and frequency [1]. This work focuses on the Downlink (DL) and Uplink (UL) of MU-MIMO systems when user devices are equipped with multiple antennas. Let M_U be the number of antennas per user (for simplicity, we assume that all users have the same number of antennas). Multi-Antenna (MA) users are desirable from a performance perspective because their multiple antennas can be used to receive (on the DL) or transmit (on the UL) M_U independent data streams simultaneously or possibly fewer data streams with potentially higher rates [2]. We focus on single-cell MU-MIMO (i.e., there is only one Base-Station (BS)). The BS uses Orthogonal Frequency-Division Multiple Access (OFDMA), which divides the system bandwidth into subchannels and the time into Time-Slots (TSs). A pair made of a subchannel and an OFDMA TS is the smallest scheduling unit and is called a Physical Resource Block (PRB).

In an MU-MIMO system with OFDMA and MA users, the necessary Radio Resource Management (RRM) processes are user selection, digital beamforming, Power Allocation (PA), Power Distribution (PD) and Modulation and Coding Scheme (MCS) selection. The user selection process decides which streams of each user are selected for transmission on the UL or reception on the DL in each PRB. It is a crucial step in MU-MIMO systems because the channel correlations among the users selected together affect system performance [3]. Digital beamforming is done per PRB over the selected streams and is related to steering signals to achieve a particular goal, such as interference mitigation. In this work, we focus on linear beamforming techniques for interference mitigation called Zero-Forcing (ZF) [4]. In linear beamforming for the DL, the signals intended for a particular user are multiplied

by a precoding matrix (precoder) at the BS for transmission and by a combining matrix (combiner) at the user for decoding, while for the UL, the precoding and combining order are reversed. For brevity, we will refer to the ZF-based linear beamforming process as just precoding for both the DL and UL.

The number of streams that are used per user can be either fixed or variable depending on the ZF-based precoding strategy. There are three primary precoding strategies based on the number of data streams used per selected user in a PRB: 1) Coordinated-Transmit-Receive-1 (CTR₁) that uses only the strongest stream per selected user [5]; 2) Block Diagonalization (BD) that uses all possible streams (i.e., M_U streams) per selected user [4]; 3) Coordinated-Transmit-Receive-Flexible (CTR_F) that allows flexible stream allocation per selected user [4, 6]. These precoding strategies can be used for both the DL and UL. Systems with CTR_F have the potential to offer better performance due to the flexibility in allowing different numbers of streams per user compared to the other strategies, but it is more complex to compute.

The PA and PD processes are related to power management. PA is about allocating the power budget to each PRB in a given TS, while PD is about distributing the power within each PRB individually to the selected streams using the per-PRB power given by PA. On the UL, the power belongs to each user individually and the PA process allocates the power in a given TS across the PRBs for which the user is selected for transmission. Then, PD distributes the power allocated to each PRB by PA among the user's data streams enabled in the PRB. Given the user selection per PRB in a given TS and the ZF-based precoding, joint PA and PD on the UL can be easily done for each user. For brevity, we will then refer to the power management on the UL for each user as just PA (although recall that there is a PD process for the user's streams in each PRB). In contrast, the power on the DL belongs to the BS. PA on the DL is then about allocating the BS power to each PRB in a given TS. PD distributes the allocated power to each PRB by PA among the streams of the selected users for reception in each PRB individually.

Finally, given the user selection per PRB in a TS, precoding, PA and PD, we can obtain the Signal-to-Noise Ratio (SNR) for each stream of each selected user in each PRB for both the DL and UL. ZF-based precoding nullifies interference in each PRB under perfect Channel State Information (CSI), hence interference does not play a role in the SNR computation. Given the SNR for each stream, we can compute its rate. Most works consider the smooth Shannon rate function for rate computation even though in a practical system, there is a set of MCSs where each MCS yields a fixed rate as long as the SNR is larger than a certain threshold [7, 8]. MCS selection incurs additional difficulty in the RRM process optimization as it yields a piece-wise constant (i.e., non-differentiable at the steps) rate function.

1.1 Motivation and Research Questions

The increased flexibility of CTR_F comes at the price of higher complexity. The precoding computations in CTR_F are more complex than in BD and CTR_1 [2, 4, 5, 6]. Furthermore, the user selection for MU-MIMO with CTR_F becomes essentially stream selection since we can enable any combination of streams per user. Even the optimal user selection (e.g., under BD or CTR_1) becomes prohibitively complex for just one PRB on the DL for a moderate number of users due to the necessity for exhaustive searches [9]. The additional degree of freedom brought by CTR_F with streams makes the selection problem more difficult. Hence, it is paramount to investigate when systems with the simpler strategies CTR_1 and BD perform similarly to CTR_F to guide the design of real-time RRM heuristics for the DL and UL of MU-MIMO systems with MA users.

Previous works have compared the performance of MU-MIMO systems with MA users employing these precoding strategies on the DL [2, 5, 6]. Nevertheless, the studies were limited because they only considered the Sum-Rate (SR) maximization criterion, excluding fairness among the users. In addition, the comparisons were carried out using the Shannon rate function in systems where the number of BS antennas was small, which is not in line with the large BS antenna arrays expected in MU-MIMO deployments for increased performance [10]. On the UL, there is no previous work comparing these important precoding strategies. Our goal is to compare these precoding strategies considering fairness in realistic systems for both the DL and UL. We define a realistic system as a system deployed in one of the 3rd Generation Partnership Project (3GPP)-compliant scenarios, namely Rural Macro (RMa), Urban Macro (UMa) and Urban Micro (UMi), with a large number of BS antennas, and where the BS uses a practical set of 3GPP-based MCSs. Then, we pose the first of our Research Questions (RQs):

RQ 1.1 What are the performance gains in realistic MU-MIMO systems with MA users brought by CTR_F (that uses a flexible stream allocation per user) compared to the simpler strategies BD and CTR_1 on the UL and DL?

To answer RQ 1.1 comprehensively, we have to study the RRM processes for the different 3GPP scenarios over distinct network parameters (e.g., number of BS antennas). Focusing first on the DL, finding the optimal user (resp. stream) selection for CTR_1 and BD (resp. CTR_F) is not feasible for realistically sized systems even for just one PRB on the DL. To obtain a selected set for a PRB for those precoding strategies, several works have followed a suboptimal strategy based on offline (where run time does not matter) greedy-up searches [5, 6, 9]. These searches start with no selected user (resp. stream) for

CTR₁ and BD (resp. CTR_F), and add iteratively a user (resp. stream) at a time, where the addition offers the largest sum rate. Put differently, these greedy searches explore the stream/user solution set smartly focusing on SR maximization. Albeit not optimal, this greedy-up search principle has been extensively used and has been shown to perform well in the context of MU-MIMO systems with both Single-Antenna (SA) and MA users [9, 11, 12]. These greedy-up searches were also used to compare the target precoding strategies in the aforementioned limited settings [2, 5, 6]. However, there is an additional problem to be addressed: these offline searches were defined for a PRB but an OFDMA-based system possesses multiple PRBs in a given TS over which the BS power budget has to be allocated by the PA process. Following previous works [13], we consider that the BS power is allocated equally across the PRBs. Then, the RRM processes can be considered on a per-PRB basis since the per-PRB power is given. In this work, we consider Proportional Fairness (PF), which is enforced for each PRB through weights reflecting the user priorities [13]. Then, the system utility maximized by the searches becomes a Weighted Sum-Rate (WSR) (for SR maximization, the weights can be set to one). To obtain meaningful comparisons of the precoding strategies over different network parameters and scenarios, another challenge is to run a large number of PRBs to allow the weights to evolve (for PF) and to account for the inherent wireless channel variations.

Addressing [RQ 1.1](#) for the UL is more complicated. The main challenge is that the power of a given user should not be allocated to the PRBs in advance in a given TS as done for the DL. This is because we need to know in which PRBs a user is selected for transmission before doing any PA. Thus, the RRM processes on the UL should be done jointly over all PRBs of a TS at once, which increases the RRM complexity substantially when compared to the DL (which can be done on a PRB basis). In [14], an offline greedy-down search over the TS was proposed for SA users, approach that can be used to compute the performance of MU-MIMO systems with CTR₁¹ but not with CTR_F or BD. Moreover, the proposed search has a constraint on the number of users, which has to be no greater than the number of BS antennas. The extension of the offline greedy-up search used on the DL to the UL would be advantageous given its wide application and satisfactory performance. However, such a search has to be done over all the PRBs at once in a TS, which greatly increases the search complexity compared to the DL. Another distinction between the DL and UL appears when considering fairness; since the RRM processes are done over the entire TS at once on the UL, the weights are updated on a TS basis. Many TSs have then to be run to enable the evolution of the weights and consider channel variations. The implementation of the offline greedy-up search over several TSs calls for efficient methods

¹MA users under CTR₁ can be seen as SA users where the multiple antennas per user are combined to create a virtual antenna.

of precoding and rate computations to reduce computational costs. This yields our next RQ:

- RQ 1.2 How to extend in a computationally efficient way the greedy-up search developed for the DL to the UL for MA users under the precoding strategies BD, CTR_1 and CTR_F ?

1.2 Contributions

In the following, we present the general contributions of this work, where we separate the contributions for the DL and UL. The contributions are:

- **Downlink:** We conduct extensive computations on the DL to evaluate offline the performance of MU-MIMO systems under the precoding strategies CTR_1 , CTR_F and BD in realistic systems for PF and SR maximization. In addition to PF, we also consider SR maximization to investigate whether previous conclusions drawn in the literature hold for realistic systems as well. For that, we use the greedy-up searches proposed in the literature and adapt them to consider practical MCS-based rate functions and PF (using WSR). Furthermore, to allow comprehensive comparisons and consider many PRBs, we use efficient precoding computations proposed in the literature [5, 12]. For PF, we show that the choice of precoding strategy should depend on the network parameters such as the number of BS antennas and the number of users, and that results obtained in previous papers for SR maximization do not always hold in larger systems with many BS antennas.
- **Uplink:** Similar to the DL, we execute comprehensive computations to compare offline the performance of MU-MIMO systems with the precoding strategies CTR_F , CTR_1 and BD in realistic systems for PF on the UL. To achieve that, we propose extensions of the greedy-up search developed for the DL to the UL. However, the search on the UL must be over the entire TS consisting of many PRBs and must be done for many TSs. The search over the TS induces a higher complexity on the UL, which we address by proposing a computationally efficient search, where we reuse, when possible, results computed in previous search iterations. The results show that the choice of precoding strategy should depend on the network parameters and the 3GPP scenario. More specifically, CTR_1 emerges as a suitable alternative to CTR_F in the UMa scenario on the UL due to its lower complexity and comparable performance, whereas for RMa, the choice of precoding strategy should depend on the number of users.

1.3 Outline

This manuscript is organized as follows. In Chapter 2, we present a literature survey on the DL and UL of MU-MIMO systems focusing on MA user settings. We also describe in Chapter 2 the system and channel models as well as the 3GPP-based scenarios employed throughout the work. In Chapters 3 and 4, we address the studies on the DL and UL, respectively. In Chapter 5, we conclude the work with final remarks and potential research directions. The DL study in Chapter 3 is adapted from the work [15] published in the IEEE Wireless Communications and Networking Conference (WCNC) Proceedings. The UL study in Chapter 4 has been submitted for publication in IEEE Transactions on Wireless Communications.

Notation: Calligraphic letters, such as $\mathcal{A}$, denote sets, with $|\mathcal{A}|$ indicating the cardinality of set $\mathcal{A}$. $\mathbb{C}$ represents the set of complex numbers. The bold letter $\mathbf{M} \in \mathbb{C}^{M \times N}$ indicates a matrix of dimension $M \times N$ with complex numbers.

Chapter 2

Literature Survey and Background

In this chapter, we discuss related works for the DL and UL of MU-MIMO concentrating on MA users with ZF processing. All the works address MA users unless otherwise specified. Then, we describe the system and channel models used throughout this manuscript, as well as the 3GPP scenarios.

2.1 Downlink

As mentioned in Chapter 1, we use offline greedy-up searches to look for suitable user-sets (resp. stream-sets) for BD and CTR_1 (resp. CTR_F) per PRB to compare the performance of MU-MIMO systems with those precoding strategies. Greedy-up search heuristics were proposed in [9], [6] and [5] employing the precoding strategies BD, CTR_F and CTR_1 , respectively. In [2, 5, 6], performance comparisons of MU-MIMO systems with those precoding strategies were studied for small systems (w.r.t. the number of BS antennas) for SR maximization with the Shannon rate function (i.e., the function that maps SNR to rate).

Many other works have used the same greedy-up search principle for user/stream selection for the DL of MU-MIMO systems with MA users under ZF processing. For example, in [16, 17, 18], offline heuristics that are variants of the greedy-up search for BD from [9] were proposed. As discussed in Chapter 1, the BD-based greedy search from [9] operates by adding users iteratively to the selected set where the selected user is the one that yields the largest sum rate. A selection metric based on the sum rate requires precoding computations and power optimization. The heuristics from [16, 17, 18] are based on lower-complexity selection metrics other than the sum rate, but the final sum rates given by these variants are

worse compared to the proposal from [9], which uses the sum rate directly as the selection metric. In [19], another offline heuristic based on greedy-up search was proposed to select streams while optimizing the combining matrices during the search to enhance the system performance at the expense of additional complexity. The proposal from [19] considered the smooth Shannon rate function and the extension to non-smooth (i.e., MCS-based) rate functions is not clear. In [8], the search was based on a simpler variant of the search from [6] for CTR_F , where at a given search iteration, only the next nonselected strongest stream of each user was assessed for selection, as opposed to assessing all nonselected streams in [6]. Such a proposal decreases the search complexity to the detriment of a less thorough exploration and hence, worse performance than the approach in [6]. [20] proposed a search based on the users' antennas instead of streams, where a subset of the antennas is selected per user. Selecting a subset of antennas per user can reduce the system complexity, like signalling overhead, but at the price of reduced performance.

To summarize, the offline greedy-up search that we use to compare the precoding strategies has been used in many works in the literature. These papers have focused mainly on reducing the search complexity. Just a few papers [2, 5, 6] are comparing the precoding strategies CTR_1 , CTR_F and BD for the DL of MU-MIMO systems with MA users and none in realistic large systems for PF and over multiple PRBs.

Table 2.1 summarizes the works presented in this section. The second column gives a brief description, the third indicates which rate function is considered (MCS-based or Shannon), the fourth whether multiple subchannels are considered in a TS, the fifth indicates the criterion being optimized, while the last column gives how many BS antennas (M_B) are used in the numerical results.

Table 2.1: Related works on the DL. Notation: Rate Function (RF), Greedy-up Search (GUS), Shannon Rate Function (SRF), Subchannels (SBs), Criterion (CR), Sum-rate Maximization (SRM).

Paper	Description	RF	SBs	CR	M_B
Ours	Comparison of CTR_F , CTR_1 , BD	MCS	Multiple	SRM/PF	16/64/100
[6]	GUS for CTR_F , comparison of CTR_F , BD	SRF	Single	SRM	4/12
[2]	Comparison of CTR_F , BD	SRF	Single	SRM	8
[5]	GUS for CTR_1 , comparison of CTR_F , BD, CTR_1	SRF	Single	SRM	12
[9]	GUS for BD and simpler variant	SRF	Single	SRM	4/8/12
[17]	Simpler Variant of GUS for BD	SRF	Single	SRM/PF	8
[16]	Simpler Variant of GUS for BD	SRF	Single	SRM	4/6
[18]	Simpler Variant of GUS for BD	SRF	Single	SRM/PF	4/8
[20]	GUS for antenna selection	SRF	Single	SRM	4/10/12
[19]	GUS with optimized equalizers	SRF	Single	WSR	4
[8]	GUS with simpler variant of CTR_F	MCS/SRF	Multiple	WSR	4

2.2 Uplink

In the following, we discuss works that addressed offline problems for the UL of MU-MIMO, where we do not limit ourselves to search-based heuristics and MA users under ZF processing given the scarce literature on the UL.

In the context of WSR maximization, [21] proposed an offline greedy-up search on the streams of the users where the precoders (at the user side since it is the UL) were optimized during the search and the BS used ZF processing to decode the signals. Nevertheless, the search only considered one PRB and hence, there was no PA over multiple PRBs which is critical on the UL. In [22], a precoding strategy similar to CTR_1 at the user side (in the sense that the users transmit at their strongest stream) with Successive Interference Cancellation (SIC) at the BS was investigated for the sum-power minimization problem with rate constraints. For the UL of MU-MIMO with SA users under ZF processing, [14] proposed an offline search on the users over a given TS but using a greedy-down strategy. The search proposed in [14] can be readily extended to CTR_1 but not to CTR_F or BD. Furthermore, the proposed search has two key drawbacks: i) it assumes that the number of users is no greater than the number of BS antennas; ii) the search breadth is limited, as we will see further in Chapter 4. In [23], the user selection for WSR maximization was addressed where the precoding strategy Maximum Ratio Transmission (MRT) was used at the users and Minimum Mean Square Error (MMSE) at the BS, with the constraint of contiguous allocation of PRBs over a given TS to each user. Also under the contiguous PRB allocation constraint, in [24], a grouping-based user scheduling method for PF and SA users with MMSE for decoding at the BS was proposed. [25] considered an MU-MIMO/OFDMA system with BD and MA users, focusing on maximizing the minimum rate offered to all users. However, the work considered a very simple heuristic where user selection is based on allocating the same number of subchannels to each user and power management is based on equal PA.

In conclusion, there is no existing offline heuristic that can be employed to compare the target precoding strategies (i.e., CTR_F , CTR_1 and BD) on the UL of practical (e.g., with multiple PRBs per TS) MU-MIMO systems for PF. In this work, we intend to fill that gap by extending well-established greedy searches from the DL to the UL so that we can compare the precoding strategies.

Table 2.2 summarizes the discussed works on the UL. The table layout is similar to the one from Table 2.1.

Table 2.2: Related works on the UL. Notation: Rate Function (RF), Greedy-up Search (GUS), Shannon Rate Function (SRF), Subchannels (SBs), Criterion (CR), Sum-power Minimization (SPM), Precoding Strategy (PS), User Antenna Setting (UAS), Minimum Rate (MR).

Paper	Description	RF	SBs	CR	PS	UAS	M_B
Ours	GUS proposal and comparison of CTR_F , CTR_1 , BD	MCS	Multiple	PF	ZF	MA	64/100
[21]	GUS with optimized precoders	SRF	Single	WSR	ZF	MA	8
[22]	SPM with rate constraints	SRF	Single	SPM	MRT/SIC	MA	up to 8
[14]	Greedy-down search	MCS	Multiple	PF	ZF	SA	up to 128
[23]	Iterative UL user selection	SRF	Multiple	WSR	MMSE/MRT	MA	32
[24]	Grouping-based user selection	SRF	Multiple	PF	MMSE	SA	64/128
[25]	Grouping-based user selection	MCS	Multiple	MR	ZF	MA	64/128

2.3 System Model

We consider both the DL and UL of a single-cell MU-MIMO OFDMA-based system. OFDMA divides the system bandwidth into subchannels and the time into TSs. Time-Division Duplexing (TDD) can be used to separate the DL and UL transmission in time, where at a given time, the whole system bandwidth (in terms of subchannels) is allocated to either the DL or the UL. We assume no signalling overhead, error-free and full CSI, and full buffers at the BS (resp. users) for the DL (resp. UL) so that data is always available for transmission. Both the BS and the users are equipped with square planar antenna arrays, with M_B and M_U antennas¹, respectively. A TS has a duration of d seconds and the bandwidth is divided into C subchannels, each with a bandwidth of B_C . We define $\mathcal{C} = \{1, \dots, C\}$ to be the set of subchannels. The smallest scheduling unit is the PRB consisting of a subchannel $c \in \mathcal{C}$ in a TS t and is represented by the pair (c, t) . Under ZF-based precoding with MA users, each user can receive (resp. send) up to M_U independent data streams at the same time on the DL (resp. UL) in a given PRB. In addition, the BS can transmit (resp. receive) up to M_B streams simultaneously in a given PRB on the DL (resp. UL) [4, 21]. For the DL, the BS transmits at a carrier frequency f_c with a power budget P_{DL} , which is equally distributed across the PRBs in a TS resulting in a per-PRB power budget of $P_{\text{PRB}} = P_{\text{DL}}/C$. For the UL, the users transmit at the same carrier frequency of f_c with a per-TS power of P_U , which has to be allocated among the PRBs allocated to each user.

¹Although we assume an equal number of antennas per user for simplicity, this study can be extended to non-equal settings.

Table 2.3: Employed MCS-based rate mapping. Notation: Spectral Efficiency (SE).

SE (bps/Hz)	0.15	0.38	0.88	1.48	1.91	2.41	2.73	3.32	3.90	4.52	5.12	5.55	6.23	6.91	7.40
SNR (dB)	-6.82	-3.44	-0.53	3.79	5.80	8.08	9.76	11.72	13.49	15.87	17.73	19.50	21.32	23.51	25.15

2.4 Scenarios and Channel Model

We consider three 3GPP-based scenarios, namely RMa, UMa, and UMi with fixed users. Each scenario models the system as a sector of a three-sector hexagonal cell, with users being uniformly distributed within the sector. The cell radius changes depending on the scenario and is taken from Tables 7.2-1 and 7.2-3 in [26]. Similar to [25], we adopt a 3GPP-based tapped-delay-line channel model with 8 taps based on Table 7.7.2-5 from [26], where the first tap can contain a Line-Of-Sight (LOS) component or only Non-Line-Of-Sight (NLOS) components, while the remaining taps are NLOS. The channel model incorporates all essential practical elements, such as path loss and shadowing (large-scale fading), small-scale fading, and transmit and receive correlation based on an exponential model (see [25] and references therein for the detailed channel formulation). The channel parameters, namely large-scale fading, LOS probability, and Rician coefficients, are dependent on the scenarios and can be found in Tables 7.4.1-1, 7.4.2-1, 7.5-6 from [26]. The employed 3GPP-compliant MCS-based rate mapping is based on [27] and given in Table 2.3. We define the channel model and MCS mapping here for organizational purposes since they are used for both the DL and UL in the numerical sections of Chapters 3 and 4, respectively. Nonetheless, the methodology is agnostic to them.

To evaluate and compare different precoding strategies in both the DL and UL, we have to introduce the concept of realization. We define a *realization* $\omega(\Delta, U, \text{Scenario}, \mathcal{N})$ over a horizon of Δ TSs as a set of fixed users $\mathcal{U}$ with their channel matrices for all ΔC PRBs, where $U = |\mathcal{U}|$ is the number of users, Scenario is one of the three scenarios (i.e., UMa, UMi, RMa), and $\mathcal{N}$ is a set of network parameters described in Section 2.3, such as the user (resp. BS) power budget, P_U (resp. P_{DL}), number of subchannels (i.e., PRBs) per TS, C , and number of antennas per user, M_U (resp. at the BS, M_B). Within a realization, the only channel aspect that varies for a given user $u \in \mathcal{U}$ is the small-scale fading across PRBs. Hereinafter, we will refer to a realization as $\omega(\Delta)$ for simplification and to highlight that the number of TSs executed per realization (i.e., Δ) may differ. Many TSs must be run per realization since the channel varies and the PRBs (resp. TSs) for the DL (resp. UL) are coupled through weights with PF. We will explain how to obtain the value of Δ for each realization in the following chapters.

Chapter 3

Downlink

3.1 Introduction

Recall that for MA users under ZF-based precoding, there are three primary precoding strategies based on the number of streams used per selected user in a given PRB, namely CTR_F , BD and CTR_1 . Systems with CTR_F have the potential to offer better performance compared to the other strategies at the cost of higher complexity in the precoding and user selection (stream selection under CTR_F). Hence, it is crucial to investigate when systems with those simpler strategies perform similarly to CTR_F to guide the design of real-time RRM heuristics for the DL of MU-MIMO systems with MA users.

We perform an offline study to evaluate and compare the performance of MU-MIMO systems with these precoding strategies on the DL. In line with the literature, we need to search the user space for a user-set of dimension at most M_B users for CTR_1 and of dimension at most M_B/M_U for BD for each PRB of an OFDMA TS, for a sufficiently large number of TSs to make the results stationary. For CTR_F , we need to search the stream space for a stream-set of dimension at most M_B streams with at most M_U streams per user. For each user-set or stream-set, we need to compute the precoding and then distribute power among the streams (i.e., PD) to maximize the utility function corresponding to the selected criterion. We consider SR maximization and PF, for which the RRM processes can be performed on a per-PRB basis as we discussed in Chapter 1 (the PF problem is transformed into a WSR problem with weights being updated every PRB [13]). Running an extensive comparison of the three precoding strategies is difficult because of the computational burden of precoding and PD for multiple user-sets or stream-sets for each PRB of

several TSs. Indeed, finding the optimal set even for a single PRB becomes prohibitively complex for a moderate number of users as it requires exhaustive searches [9].

Previous works have proposed offline greedy search heuristics that explore user-sets (resp. stream-sets) under CTR_1 and BD (resp., CTR_F). These searches explore the solution space smartly to obtain a suitable set while maximizing the sum rate for a PRB. Specifically, Multiple Eigenmode Transmission (MET) was proposed using CTR_F [6], Greedy User Selection - Block Diagonalization (GUS-BD) using BD [9], and Multiple Dominant Eigenmode Transmission (MDET) employing CTR_1 [5]. The searches operate similarly in a greedy up fashion, starting with an empty set and iteratively adding users or streams to the scheduled set, selecting the one at a given iteration that yields the largest sum rate when combined with the already selected set. To compute the rates at each iteration given a SNR, all papers utilized the Shannon rate function and PD based on Water-Filling (WF) [28].

Preceding research using these offline greedy search heuristics has shown that the more complex strategy, CTR_F , significantly outperforms BD, and that CTR_1 performs comparably to CTR_F while also outperforming BD [2, 5, 6]. However, these comparisons were limited in three main ways: 1) they only considered SR maximization over one PRB; they did not consider the more relevant PF criterion for which weights need to be introduced and updated every PRB and the computations have to be done over many PRBs since the PRBs are coupled through the weights; 2) the computations were conducted with a small number of BS antennas, not reflecting the large antenna arrays expected in future MU-MIMO deployments [10]; 3) they did not consider a realistic rate function and used instead the Shannon rate function for rate computation. The system performance comparison using these precoding strategies under PF and in large realistic systems remains unexplored. Recall that a realistic system belongs to one of the 3GPP-based scenarios (i.e., RMa, UMa and UMi) with a large number of BS antennas, where the BS uses a 3GPP-compliant MCS-based rate function instead of the commonly used Shannon function. To evaluate comprehensively these precoding strategies for SR maximization and PF over many PRBs in realistic large systems, a large number of precoding computations and PD are needed (since the search is deep if not exhaustive), which is not tractable using traditional precoding computation techniques. Therefore, smarter iterative precoding techniques such as those proposed in [5, 12] must be used.

In this chapter, we address the following RQs:

- RQ 3.1 What performance gains can be obtained by using the more complex precoding strategy CTR_F as opposed to the simpler CTR_1 and BD strategies considering both SR maximization and PF in realistic systems?

RQ 3.2 Do the conclusions drawn from previous works for SR maximization with a few BS antennas also hold in realistic systems with large numbers of BS antennas?

RQ 3.3 Are the trends and conclusions for PF similar to those for SR maximization?

To do so, we make the following key contributions:

- **Extensive Comparisons:** We extend previous studies and conduct comprehensive offline evaluations of the performance of MU-MIMO systems with MA users for the precoding strategies CTR_1 , BD, and CTR_F for both SR maximization and PF. We use greedy-up searches for user/streams-sets, as well as an advanced precoding computation method and an efficient PD heuristic, over multiple PRBs in realistic systems. Specifically, we employ the advanced precoding computation method from [5, 12] to make the evaluations feasible for realistic systems featuring large numbers of users and antennas.
- **Practical Considerations:** We adapt the greedy searches for user/stream-sets by considering practical MCS-based rate functions and to account for PF.
- **Fairness Analysis:** We consider both SR maximization and PF, showing that these criteria yield substantially different conclusions. We show that the BD strategy performs as well as CTR_F and emerges as a suitable choice for SR maximization with a sufficiently large number of BS antennas, whereas for PF, the choice of precoding strategy should depend on the number of users and BS antennas. This highlights that evaluating precoding strategies based solely on SR maximization can be misleading and a comprehensive evaluation should also consider fairness such as PF.

The remainder of this chapter is organized as follows. Section 3.2 describes the per-PRB RRM procedure necessary to obtain the system performance for each precoding strategy, including the practical adaptations in the search methods to consider an MCS-based rate function and PF. In Section 3.3, the procedure for obtaining the long-term performance over many TSs is explained. Section 3.4 presents the numerical results and discussions. Finally, Section 3.5 concludes the chapter with key findings.

3.2 Per-PRB RRM Procedure

We describe here the per-PRB RRM offline procedure required to obtain the MU-MIMO system performance given one of the three precoding strategies. It is known that for both

SR maximization and PF, the DL RRM problem over a horizon Δ of TSs can be decoupled in per-PRB WSR problems, where the weights for SR maximization are constant and equal to one across PRBs and change every PRB for PF [13, 29]. Introducing weights and updating them every PRB to study PF is the first adaptation we make to the comparison procedure used in [2, 5, 6]. To obtain meaningful results for the research questions, many PRBs must be considered since the channel varies across PRBs and for PF, to allow the weights to evolve. Performing computations over many PRBs for both SR maximization and PF is the second adaptation we make to the technique used in [2, 5, 6]. We will start by describing the PRB-level procedure and then, in the next section, we describe the procedure to obtain the long-term performance over Δ TSs (we also discuss how we select Δ).

Given a 3GPP scenario, a precoding strategy, a set of network parameters $\mathcal{N}$, described in Section 2.3, a realization $\omega(\Delta)$, a PRB (c, t) , and for each user $u \in \mathcal{U}$, the channel matrix $\mathbf{H}_u$ and the weight w_u at the beginning of the PRB, we present next the PRB-level procedure employed to obtain the selected set and compute the WSR for the given PRB on the DL (we remove the dependence of $\mathbf{H}_u$ on the PRB index (c, t) for ease of notation).

For CTR₁ (resp. BD), we need to select a user-set of at most M_B users (resp. M_B/M_U) since each user will receive exactly one (resp. M_U) stream(s). For CTR_F, we need to select a stream-set of at most M_B streams with at most M_U streams per user. An exhaustive search of the user-sets (resp. stream-sets) for CTR₁ and BD (resp. CTR_F) is prohibitively expensive for reasonably sized systems (i.e., $M_B > 60$, $M_U > 1$, $U > 10$), hence we adopt a greedy up search of the user/stream-sets as done in previous works [2, 5, 6, 9].

We describe the per-PRB *iterative* procedure using the precoding strategy CTR_F, which is the most complex of the three strategies. It is iterative because we try multiple stream-sets. Let $\mathcal{S}_u$ be the stream-set containing the indices of the current selected streams for each user $u \in \mathcal{U}$ (the indices can take any integer value from 1 to M_U with stream 1 being the strongest and stream M_U the weakest, see [5] for more details on stream strength). We initialize each $\mathcal{S}_u$ as the empty set and the WSR as zero. The union of all sets $\mathcal{S}_u$ is the current selected stream-set for the PRB. Then, at each search iteration, we assess each unselected stream for all users one at a time to obtain the candidate stream that yields the largest WSR, $\overline{WSR}$, when added to the current selected stream-set.

At a given iteration of the search, $\overline{WSR}$ for a candidate stream is obtained as follows. Let $\mathcal{K}$ be the set of active users for which at least one stream has been selected (i.e., if $\mathcal{S}_u \neq \emptyset$, then $u \in \mathcal{K}$, $\forall u \in \mathcal{U}$). Also, assume the candidate stream for which $\overline{WSR}$ is being computed is temporarily added to the stream-set of the user it belongs to. Then:

1. Compute the *joint precoding* for all the streams in $\{\mathcal{S}_u\}_{u \in \mathcal{K}}$ to obtain the effective

channel, $E_{u,s}$, of stream $s \in \mathcal{S}_u$ of user $u \in \mathcal{K}$, $\forall u \in \mathcal{K}, \forall s \in \mathcal{S}_u$. The precoding computations for MA users under ZF were performed in [2, 6] very inefficiently as they recomputed precoding from scratch when assessing a new candidate. We employ an iterative precoding computation method proposed in [5, 12] that reuses results from previous search iterations to make precoding computations much faster. Considering this method is the third adaptation that we make to the techniques developed in [2, 6] to enable comprehensive evaluations.

2. Compute the *joint PD* on streams in $\{\mathcal{S}_u\}_{u \in \mathcal{K}}$, using $\{E_{u,s}\}_{u \in \mathcal{K}, s \in \mathcal{S}_u}$, $\{w_u\}_{u \in \mathcal{K}}$, and P_{PRB} . Then, obtain the power allocated to stream $s \in \mathcal{S}_u$ of user $u \in \mathcal{K}$, $P_{u,s}$, $\forall u \in \mathcal{K}, \forall s \in \mathcal{S}_u$. For PD, considering the appropriate rate function is essential [13]. Due to the piece-wise constant nature of the MCS-based rate function, the selection process requires careful consideration, as it may identify multiple candidates yielding equivalent WSR. In such instances, we propose an enhancement to the searches, where we select the candidate that not only maximizes the WSR when added to the already selected set but also minimizes power usage. Specifically, the PD using the MCS-based rate function is as follows. We first fit the rate function with a logarithmic function, which is then used in conjunction with WF (Algorithm 6 of [28]) to provide an initial PD solution efficiently. Subsequently, the greedy MCS-aware Algorithm 1 from [13] refines the solution by considering the MCS-based rate function.
3. *Select the MCS* for each stream s in $\{\mathcal{S}_u\}_{u \in \mathcal{K}}$, using the SNRs computed as $P_{u,s} \times E_{u,s}$, $\forall u \in \mathcal{K}, \forall s \in \mathcal{S}_u$, and obtain its corresponding rate, $r_{u,s}$. Finally, the WSR is $\overline{WSR} = \sum_{u \in \mathcal{K}, s \in \mathcal{S}_u} w_u r_{u,s}$.

Recall that we are describing the per-PRB procedure for CTR_F . After assessing all unselected streams across the users and obtaining the stream that, when combined with the set of already selected streams, yields the largest WSR $\overline{WSR}$ (with the lowest power usage in case of a tie), we add such a stream to the selected set if $\overline{WSR} > WSR$ (i.e., there is an increase in the WSR upon adding the new stream). Otherwise, the search stops, as done in [2, 5, 6, 9]. Other obvious stopping criteria for the search are either when M_B streams have been selected, i.e., when $\sum_{u \in \mathcal{K}} |\mathcal{S}_u| = M_B$, or when all the streams of all users have been selected (i.e., $|\mathcal{S}_u| = M_U, \forall u \in \mathcal{U}$). After stopping, the search returns the final rate received by each stream in the PRB, $r_{u,s}, \forall u \in \mathcal{U}, \forall s \in \mathcal{S}_u$. Let the final rate seen by user $u \in \mathcal{U}$ in that PRB be $r_u = \sum_{s \in \mathcal{S}_u} r_{u,s}$. This ends the per-PRB procedure, which is illustrated in Figure 3.1.

The per-PRB procedures for CTR_1 and BD are similar, with the key difference being that the search scope is over users instead of streams: for BD, we use all M_U streams when

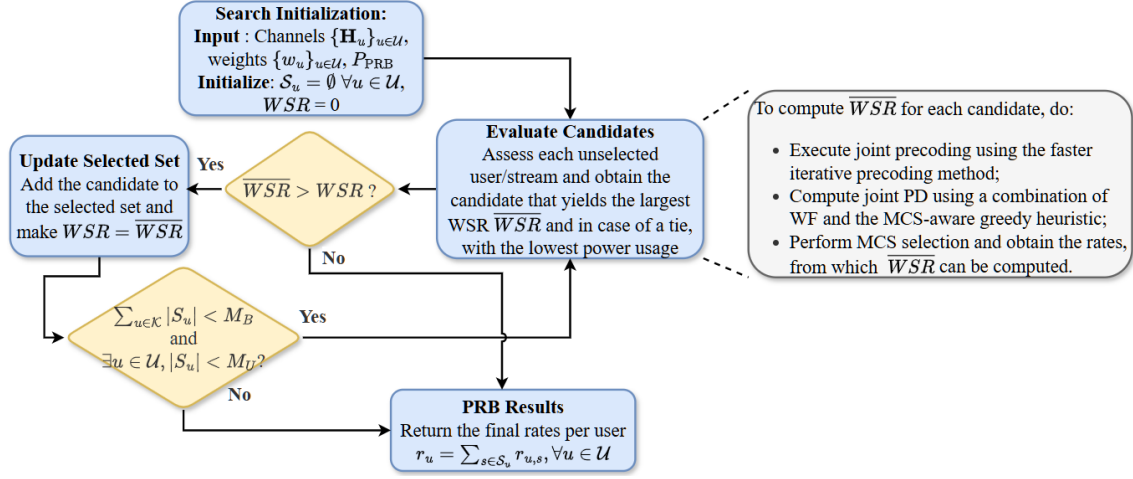


Figure 3.1: Flow diagram of the PRB-level procedure for a given scenario, precoding strategy, network parameters $\mathcal{N}$, realization $\omega(\Delta)$, and PRB.

assessing and selecting a user, whereas, for CTR_1 , we assess and select only the strongest stream of a user. In the following, we present the procedure for obtaining the long-term performance over many TSs for a given realization $\omega(\Delta)$.

3.3 Long-term Performance Computation

At the end of the computations for the PRB, we need to update the average per-PRB rates received by each user so far as well as the weights $\{w_u\}_{u \in \mathcal{U}}$ for PF. Recall that the weights remain constant and equal to one across PRBs for SR maximization. For PF, it is standard practice to do the following [13, 29]. Let W be a certain window size (a number of PRBs). Let R_u be the moving average per-PRB rate seen by user $u \in \mathcal{U}$ in the past W PRBs. Then, the R_u 's are updated for both SR maximization and PF at the end of the computations for a PRB as:

$$R_u = \frac{(W-1)R_u + r_u}{W}, \quad \forall u \in \mathcal{U},$$

and the weights w_u 's are updated as: $w_u = 1/R_u$ for PF.

As discussed previously, performance needs to be computed on a large number of TSs for both PF and SR maximization. This is clear for PF since the computations for one

PRB are dependent on the computations from the previous PRBs through the weights $\{w_u\}_{u \in \mathcal{U}}$, hence the need for many PRBs to allow these quantities to evolve and become stationary (this is possible because the small-scale fading variations are of mean zero for each user). For SR maximization, many PRBs are also needed to take into account the channel variations. In this context, two issues need to be taken care of: i) how to initialize the weights for PF? ii) on how many TSs should the performance computation be based?

Given a 3GPP scenario, a precoding strategy, network parameters $\mathcal{N}$ and a realization $\omega(\Delta)$ (where Δ will be discussed later in this section and is a large enough horizon in TSs to obtain meaningful results), we initialize $R_u, \forall u \in \mathcal{U}$, to be the maximum possible per-PRB rate, which depends on the set of MCSs. The computations across PRBs start with a warm-up period of $\hat{T}_d$ TSs (i.e., $\hat{T}_d \times C$ PRBs) to prevent the initial values of w_u from influencing the end results, i.e., we select $\hat{T}_d$ such that $W \leq \hat{T}_d \times C$. After the warm-up period, we will proceed with enough TSs, so that the Target Metric (TM) converges (note that TM is *not* the long-term performance of the system but just a metric to verify convergence of the algorithm). The TM is the geometric mean of the average rates R_u for PF and the sum over R_u for SR maximization. After the warm-up period of $\hat{T}_d$ TSs, we run the computations for T_d TSs and collect the TM at the end of this period. We then run the computations for an extra T_d TSs and check the TM convergence. If the relative difference in the TM is within a certain tolerance tol , we stop the computations; otherwise, we run the computations for another T_d TSs. Once we have reached convergence, we have the value of Δ . We also keep track of the aggregated rates λ_u for each user $u \in \mathcal{U}$ over the PRBs executed after the warm-up period of $\hat{T}_d \times C$ PRBs. Algorithm 1 summarizes the procedure for the long-term performance computations with many TSs under PF for a realization $\omega(\Delta)$ (with the channel model for each user $u \in \mathcal{U}$ across the PRBs defined in Section 2.4), a set of network parameters $\mathcal{N}$, a 3GPP scenario and a precoding strategy. The output of Algorithm 1 is the geometric mean of the aggregated rates per user $\{\lambda_u\}_{u \in \mathcal{U}}$ normalized by the total number of executed TSs (essentially Δ). In Algorithm 1, n represents the PRB index and $GM(\cdot)$ computes the geometric mean of a group of values. Algorithm 1 contains two main loops and we check the TM convergence at the end of each loop. The first loop runs the first block of computations comprising $(\hat{T}_d + 2T_d) \times C$ PRBs and the second loop runs additional blocks of computations if the first block is insufficient for the TM convergence. For SR maximization, the algorithm can be adapted by replacing $GM(\cdot)$ by $SR(\cdot)$, which computes the sum of a group of values. Also, recall that there are no weight updates for SR maximization since they are always set to 1. In the next section, we will show how many TSs are necessary for the TM to converge depending on the precoding strategy and the 3GPP scenario.

Algorithm 1 Algorithm for computing the long-term performance over many TSs on the DL for PF.

Input: Realization $\omega(\Delta)$, network parameters $\mathcal{N}$, 3GPP scenario, precoding strategy, numbers of TSs for computations T_d , $\hat{T}_d$, window W and convergence tolerance tol

Output: $GM(\{\lambda_u\}_{u \in \mathcal{U}})$

```

1: Initialize  $R_u$  with the maximum per-PRB rate,  $w_u \leftarrow 1$ ,  $\lambda_u \leftarrow 0$ ,  $\forall u \in \mathcal{U}$ 
2: for PRB  $n = 1$  to  $(\hat{T}_d + 2T_d) \times C$  do ▷ First block of computations
3:   Obtain  $r_u, \forall u \in \mathcal{U}$ , by executing the procedure in Section 3.2 and update  $R_u, w_u, \forall u \in \mathcal{U}$ 
4:   if  $n > \hat{T}_d \times C$  then
5:     Make  $\lambda_u \leftarrow \lambda_u + r_u, \forall u \in \mathcal{U}$ 
6:     if  $n = (\hat{T}_d + T_d) \times C$  then
7:       Make TM  $\leftarrow GM(\{R_u\}_{u \in \mathcal{U}})$ 
8:   if  $|GM(\{R_u\}_{u \in \mathcal{U}})/TM - 1| \leq tol$  then
9:     return  $GM(\{\lambda_u\}_{u \in \mathcal{U}})/(n/C)$ 
10: Make TM  $\leftarrow GM(\{R_u\}_{u \in \mathcal{U}}), n' \leftarrow n$ 
11: for PRB  $n = n' + 1$  to  $n' + T_d \times C$  do ▷ Additional blocks of computations
12:   Obtain  $r_u, \forall u \in \mathcal{U}$ , by executing the procedure in Section 3.2 and update  $R_u, w_u, \forall u \in \mathcal{U}$ 
13:   Make  $\lambda_u \leftarrow \lambda_u + r_u, \forall u \in \mathcal{U}$ 
14: if  $|GM(\{R_u\}_{u \in \mathcal{U}})/TM - 1| \leq tol$  then
15:   return  $GM(\{\lambda_u\}_{u \in \mathcal{U}})/(n/C)$ 
16: else
17:   go to Line 10

```

3.4 Numerical Results

We present results for each precoding strategy for each scenario, with varying numbers of users and BS antennas. The results are averaged over 50 realizations for a given number of users with confidence intervals computed using the t -distribution with a confidence level of 99%. Table 3.1 presents the key parameters used in the computations. Within a realization $\omega(\Delta)$, the small-scale fading for a user changes independently across distinct reporting blocks consisting of $C_B = 19$ subchannels and $T_B = 2$ TSs while the other channel elements remain constant (see Section 2.4 for the channel model). In the following, the results for PF are presented first, followed by those for the SR maximization.

Table 3.1: Computation parameters for the DL (RB: reporting block, #: number).

Parameter	Value
# of TSs T_d and $\hat{T}_d$	8 and 2
# of channels (C_B) and TSs (T_B) in a RB	19 and 2
Subchannels per TS (C)	133
Carrier frequency (f_c)	3.5 GHz
Subchannel bandwidth (B_C)	360 kHz
TS duration (d)	0.5 ms
Antennas at the BS (M_B); at UE (M_U)	100, 64, 16; 4
Transmit power in Watts (P_{DL})	40 (UMa, RMa), 1 (UMi)
Convergence tolerance (tol)	10^{-2}
Window length in PRBs (W)	100
Channel correlation (BS and users)	0.4
Adjacent antenna distance (BS and users)	$0.5 \times 3 \times 10^8 / f_c$

3.4.1 Proportional Fairness

Figure 3.2 (top figures) presents the geometric mean of the per-TS rates for each precoding strategy, across different scenarios, as a function of the number of users, for different numbers of BS antennas. Also, Figure 3.2 (bottom figures) shows the corresponding ratios between BD and CTR_F, and between CTR₁ and CTR_F.

The first four comments are that: i) the number of BS antennas has a huge impact on the performance and the conclusions are different for small and large numbers of antennas; ii) CTR_F performs better than the other two precoding strategies; iii) BD is, somewhat surprisingly, doing much worse than CTR₁ in all scenarios for a large number of users; iv) the trends are similar for the three 3GPP scenarios.

The relative performance gap between BD and CTR_F increases with the number of users while it is the reverse for CTR₁. For a large number of BS antennas ($M_B = 100$), BD has a ratio of around 0.8 or above relative to CTR_F for all numbers of users and scenarios, while CTR₁ only does the same for a large number of users. On the other hand, for a small number of BS antennas ($M_B = 16$), BD has a ratio relative to CTR_F below 0.8 whereas CTR_F presents a ratio above 0.9 for all scenarios and numbers of users.

The results can be further explained by Figure 3.3, showing the empirical distribution of the stream combination per selected user per PRB for CTR_F with $U \in \{20, 40, 80\}$ users, for $M_B = 100$ BS antennas in UMa. For example, for $U = 40$, we see that streams 1 and 3 (i.e., combination 13) have been selected about 15% over all scheduled users and PRBs, which is interesting in itself since it means that sometimes it might be beneficial

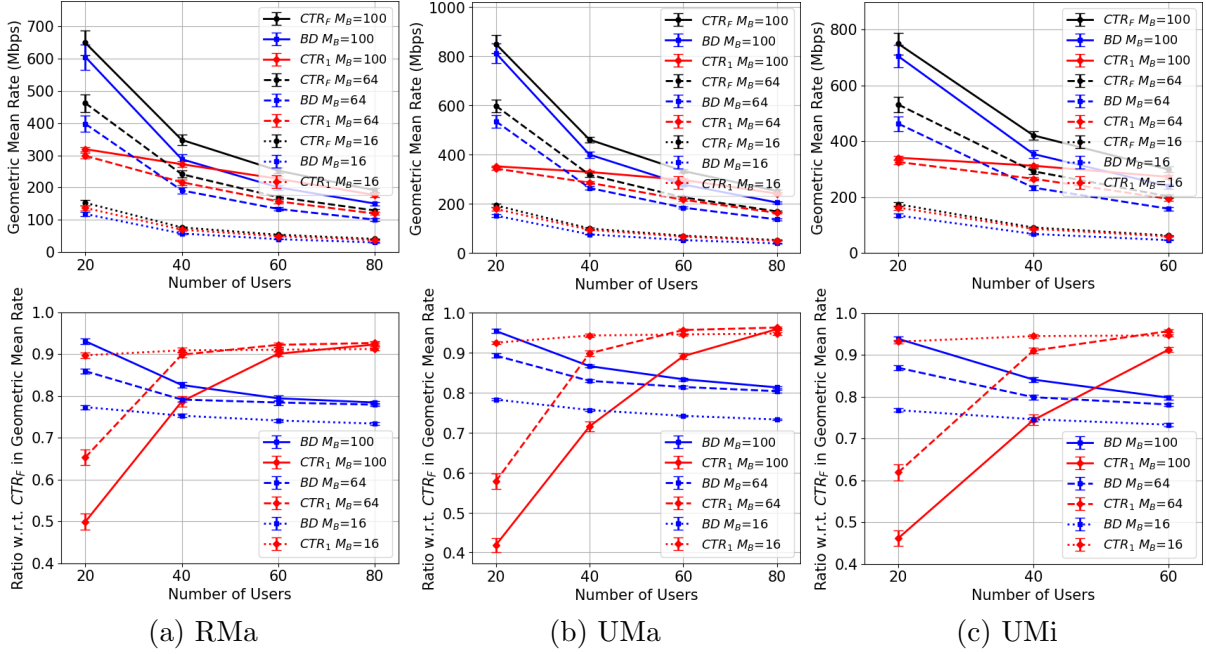


Figure 3.2: Top figures: geometric mean of the per-TS rates vs. numbers of users for the three precoding strategies for RMa (a), UMa (b), UMi (c) for 100, 64 and 16 antennas at the BS. Bottom figures: ratios in the geometric mean rate between BD and CTR_F , and between CTR_1 and CTR_F are shown.

not to select the strongest streams. For $U = 20$, most of the selected users have been allocated their $M_U = 4$ streams, which is consistent with Figure 3.2, where the disparity between CTR_F and BD is small (ratio above 0.9). For $U = 80$, CTR_1 dominates with more than 60% of the selected users being allocated only stream 1, reflecting the similarity in performance between CTR_F and CTR_1 . The performance difference between CTR_F and the other strategies increases for $U = 40$, as seen in Figure 3.2, and this can be explained by the stream allocation in Figure 3.3, which does not align significantly with BD or CTR_1 . Similar trends are observed for RMa and UMi.

3.4.2 Sum-rate Maximization

For SR maximization, we show only in Figure 3.4 the ratios in the sum rate per TS between BD and CTR_F , and between CTR_1 and CTR_F across different scenarios, numbers of users and BS antennas. For $M_B = 16$, previous research findings hold [2, 5, 6], with BD being

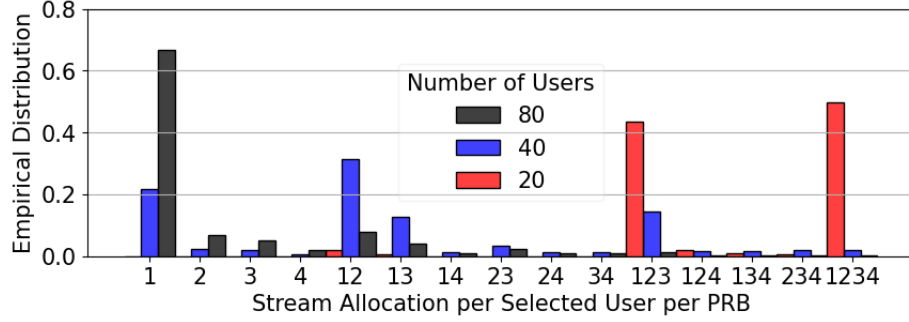


Figure 3.3: Empirical distribution of the stream allocation per selected user per PRB for 20, 40 and 80 users, for CTR_F , PF, 100 BS antennas, and UMa.

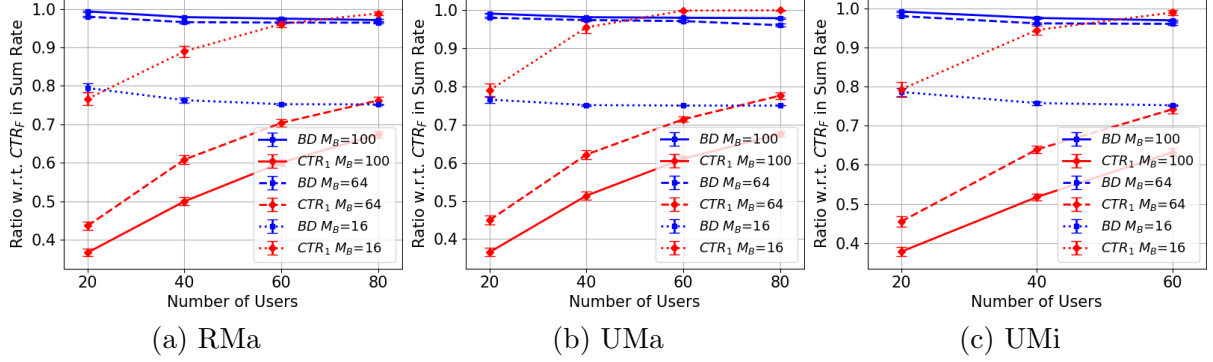


Figure 3.4: Ratios in the sum rate per TS between BD and CTR_F , and between CTR_1 and CTR_F for different numbers of users for RMa (a), UMa (b) and UMi (c) with 100, 64 and 16 BS antennas.

outperformed by CTR_F (ratio below 0.8), and CTR_1 showing similar performance to CTR_F , with the performance disparity decreasing as the number of users increases. However, as the number of BS antennas increases to $M_B = 64$ or $M_B = 100$, the difference between CTR_F and BD becomes negligible, while BD significantly outperforms CTR_1 . The performance of BD can be explained by Figure 3.5 that shows that, for $M_B = 100$ BS antennas and UMa, the most frequent allocation in CTR_F involves activating all $M_U = 4$ streams, which is essentially BD. The outcomes are similar for $M_B = 64$ and for the other scenarios as well. These results suggest that for SR maximization and a sufficiently large number of BS antennas, BD is the best choice as it is simpler than CTR_F and performs comparably. These findings indicate distinct trends for SR maximization and PF and also show the importance of studying systems with a larger number of antennas since the trends are also

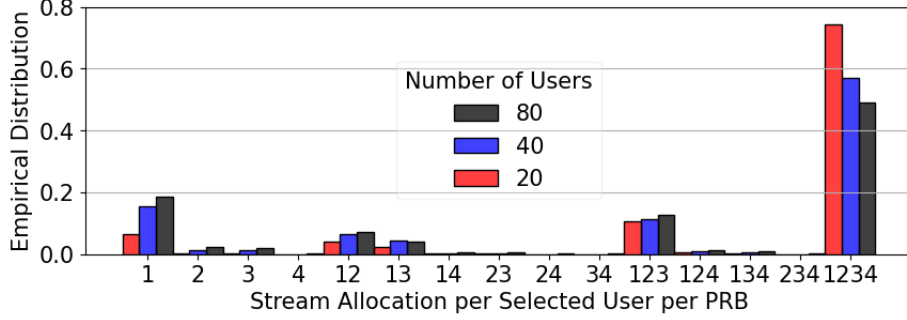


Figure 3.5: Empirical distribution of the stream allocation per selected user per PRB for 20, 40 and 80 users, for CTR_F , SR maximization, 100 BS antennas, and UMa.

significantly different for $M_B = 16$ and $M_B = 100$.

Lastly, we report that for most realizations in all scenarios and precoding strategies, 18 TSs are sufficient to achieve convergence of the metric of interest. However, for a few realizations, up to 82 TSs are necessary for CTR_F (for RMa) and for CTR_1 (for UMi and UMa), while up to 42 TSs are required for BD (for RMa and UMi). Hence, it is important not to take convergence for granted.

3.5 Conclusion

We evaluated the performance of the DL of an MU-MIMO single cell with MA users for three precoding strategies within the ZF scope. We extended previous research by comparing the precoding strategies BD (all streams active per selected user), CTR_1 (strongest stream per selected user), and CTR_F (flexible number of streams per selected user) using greedy searches across multiple PRBs considering SR maximization and PF in realistic systems, characterized by 3GPP-based scenarios with practical MCSs and large number of BS antennas. We enhanced the searches by accounting for PF and the piece-wise constant behaviour of the MCS-based rate function.

We are now ready to answer the RQs presented in Section 3.1. It was shown that the conclusions drawn from previous research for SR maximization only held for a small number of BS antennas (RQ 3.2). Concerning RQ 3.1, the results showed that for SR maximization, BD performed similarly to the more complex CTR_F for a sufficiently large number of BS antennas while outperforming CTR_1 substantially. For PF and a large number of BS antennas, CTR_1 and BD performed closely to CTR_F for many and a few

users, respectively. For a small number of BS antennas, CTR_1 performed comparably to CTR_F under PF. Addressing [RQ 3.3](#), the outcomes for SR maximization and PF presented distinct trends. In conclusion, for SR maximization, BD emerges as a suitable choice over CTR_F for a sufficiently large number of BS antennas. For PF, CTR_1 could replace CTR_F for a small number of BS antennas, while for larger BS antenna arrays, the strategy that could replace CTR_F depends on the number of users. These findings can guide the design of real-time RRM heuristics for the DL of MU-MIMO systems with MA users.

Chapter 4

Uplink

4.1 Introduction

In this chapter, we develop the necessary tools to perform the *offline* UL study where we evaluate the performance of MU-MIMO systems with MA users on the *uplink* under PF with the three ZF-based precoding strategies CTR_F , BD, and CTR_1 . Recall that most works on MU-MIMO systems focus on the DL. However, studies on the UL have become increasingly critical due to the growth in the UL traffic and the impact of the UL on the perceived quality of experience [7, 14]. To the best of our knowledge, no previous work has compared the performance of MU-MIMO systems on the UL with those precoding strategies. This is crucial for guiding the design of real-time UL RRM heuristics.

Recapitulating the discussion in Chapter 1, there are some fundamental differences between the DL and UL of an MU-MIMO OFDMA-based system with MA users. First, the power belongs to each user on the UL and has to be allocated (via power allocation) among the PRBs allocated to the user in a given TS. Then, as for the DL, the per-PRB power has to be distributed among the streams in the PRB. We cannot execute PA per user on the UL before knowing in which PRBs (over the entire TS) the user is selected for transmission. Hence, the RRM processes have to be done jointly over the entire TS on the UL (as opposed to per-PRB on the DL given (equal) PA). In addition, the user weights used to deliver fairness can only be updated at the end of a TS instead of after each PRB on the DL.

User (resp. stream) selection on the UL for BD and CTR_1 (resp. CTR_F) over the TS is especially complicated. As mentioned in the previous chapter, many works have addressed

greedy searches for user/stream-sets on the DL [6, 9]. Developing greedy searches for the UL is a more difficult task since the search must be over all PRBs in a TS and very few papers have focused on it. The most relevant works to ours were conducted in [14], [21] and [25]. In [14], an offline greedy-down search heuristic over the TS was proposed for SA users with ZF precoding. It can be readily used for CTR_1 as MA users with one stream can be interpreted as SA users, but its extension to BD or CTR_F is not clear. Also, this heuristic has a constraint on the number of users in the cell which must not be greater than the number of BS antennas. A Greedy-Up Search (GUS) heuristic based on ZF with multiple streams per user, where the precoding matrices were optimized during the search, was proposed in [21]. However, it was based on a single PRB, ignoring completely the major issue of power management on the UL, and its design was based on the Shannon rate function. [25] considered a multi-channel MU-MIMO system employing BD with MA users, focusing on maximizing the minimum rate offered to all users. However, the work considered a very simple heuristic where user selection is based on allocating the same number of subchannels to each user, and power management is based on equal PA.

To enable the evaluation of the three precoding strategies, we propose to extend the offline GUS developed for the DL to the UL. The proposed search operates over the TS with an objective to maximize the WSR while taking into account the practical MCS-based rate function (PF is achieved through the weights in the WSR). Note that this extension is not trivial since we need to make sure that the search will be tractable over a full TS. Indeed, since the UL GUS operates over the TS, its computational cost is much higher compared to the search developed for the DL, which works over a PRB. To illustrate that, consider CTR_F in a system with 40 users, each with 4 antennas, and 78 PRBs in a TS. Consider a search iteration in which a total of 30 streams have already been selected across the 78 PRBs. Then, to select just one new stream in a PRB, we have to assess 130 streams (40×4 streams in total minus 30 already selected streams) in 78 PRBs, one at a time (more than 10^3 assessments), and choose the stream that provides the highest WSR over the TS. In each of those assessments, possibly one precoding computation and several power computations (one per user affected by the assessment) have to be carried out, as will be discussed shortly. Therefore, to reduce its computational cost, we have equipped GUS with efficient precoding and power computation techniques as well as a novel iterative rate-reusing method. Our GUS on the UL has three key features that make it efficient (in decreasing order of importance):

1. Rates computed in previous search iterations are reused whenever possible, i.e., we judiciously compute rates only when strictly necessary;
2. The efficient iterative precoding computation method proposed in [5, 12] is used;

3. The rates are computed using a logarithmic fitting of the MCS-based rate function without MCS selection (MCS selection is executed at the end of the search). Thanks to this, we can perform PA and PD using WF (better PA and PD are executed at the end of the search).

This set of features makes it possible to compute offline the performance of large realistic MU-MIMO systems with any of the precoding strategies CTR_1 , BD and CTR_F over many TSs on the UL. Recall that a realistic system is one that belongs to one of the 3GPP-compliant scenarios (i.e., RMa, UMa and UMi), and features a large number of BS antennas, where the BS uses a set of 3GPP-based MCSs.

On the UL, we address the following RQs:

- RQ 4.1 Can we extend, in a computationally efficient way, the GUS heuristic developed for the DL to the UL?
- RQ 4.2 What are the performance gains brought by CTR_F compared to the simpler strategies BD and CTR_1 on the UL for different 3GPP scenarios and network parameters? Are those gains scenario-dependent?

To answer these RQs, our contributions are:

- **Efficient per-TS GUS Design and Implementation:** We propose an efficient per-TS GUS heuristic for the UL, considering all key RRM processes, that can be used for BD, CTR_1 and CTR_F . The search takes into account practical aspects, such as PF and a MCS-based rate function. Furthermore, the search utilizes the faster precoding computation methods developed in [5, 12] and efficient power management.
- **Judicious Computation Technique:** Despite the efforts to make the search efficient by using fast precoding and efficient power management, the number of rate computations at each search iteration can be very large as will be seen in Section 4.2. Then, we propose a smart computation method to reduce the computational cost of GUS by judiciously reusing rates computed in past search iterations whenever possible. This aggressive reusing method can make the heuristic faster while not changing the performance (in terms of rates) since it is only smart in what to recompute. For example, for CTR_1 and BD, we show that the run time can be reduced by 11% and 42%, respectively, when the search is equipped with the proposed rate-reusing method. This allows us to study the impact of different network parameters and scenarios in a reasonable time.

- **Extensive Numerical Study:** Using the proposed GUS heuristic, we conduct a comprehensive evaluation of CTR_F , CTR_1 and BD over many TSs for two different 3GPP scenarios, namely RMa and UMa, across different network parameters. We provide valuable information that can be used for the design of real-time UL RRM heuristics. Particularly, we show that, while CTR_F outperforms BD and CTR_1 across all scenarios and network parameters, CTR_1 or BD could replace CTR_F in some cases if CTR_F is deemed too complex for real-time settings. For instance, for RMa, BD and CTR_1 are alternatives to CTR_F for small and large numbers of users, respectively, given their comparable performance. For UMa, CTR_1 presents a similar performance to CTR_F across all network parameters, emerging as a choice over CTR_F . We also show that the numbers of antennas at the BS and at the users, and the power budget per user can significantly impact the performance of the precoding strategies.

This chapter is organized as follows. Section 4.2 presents the proposed GUS heuristic on the UL. We then explain how we evaluate the long-term performance of our systems with many TSs in Section 4.3. In Section 4.4, we discuss the reuse of rates during the greedy-up search to reduce the search computational cost. Finally, Section 4.5 and Section 4.6 present the numerical results and conclusions, respectively.

4.2 Uplink Greedy-up Search Heuristic

The PF RRM procedure for the UL, over a horizon of Δ TSs, can be decoupled in per-TS WSR maximization problems where the weights per user are updated after every TS (the weights are specific to PF). To allow the weights to evolve and to consider the channel variations, several TSs (Δ TSs) must be run for each precoding strategy, network parameter set $\mathcal{N}$, 3GPP scenario, and realization $\omega(\Delta)$. We will first define the TS-level procedure, where the proposed GUS is executed, and then, the procedure required to execute many TSs and select Δ is described in Section 4.3. Finally, we present our novel rate-reusing method to reduce the complexity of the GUS heuristic in Section 4.4.

Given a 3GPP scenario, a precoding strategy, a set of network parameters $\mathcal{N}$ (described in Section 2.3), a realization $\omega(\Delta)$, and a TS t , let $\mathbf{H}_u^{c,t} \in \mathbb{C}^{M_U \times M_B}$ be the channel matrix in PRB $c \in \mathcal{C}$ of TS t for user $u \in \mathcal{U}$, and $w_u(t)$ be its weight at the beginning of TS t . We present next the TS-level procedure employed to obtain the selected user (resp. stream) set per PRB $c \in \mathcal{C}$ in TS t under CTR_1 or BD (resp. CTR_F) and compute the WSR *over the TS* on the UL. This procedure uses the proposed GUS that we describe next. We will omit the index t representing TS t for ease of notation.

We present in this section our proposed offline per-TS greedy up heuristic, GUS, that considers all key RRM processes and can be used for CTR₁, BD and CTR_F. The heuristic incorporates realistic considerations, i.e., an MCS-based rate function and PF. It operates in an iterative greedy-up fashion to build a feasible solution over the TS including a selection of streams per PRB, along with their allocated power and rate. In each search iteration, for each PRB, we try all streams (resp. all users) one by one among those not yet selected in that PRB for CTR_F and CTR₁ (resp. for BD). At the end of the iteration, the pair among all the stream (resp. user) and PRB pairs that yields the largest WSR over the TS when combined with the already selected streams (resp. users) in all PRBs is added to the set of scheduled pairs of streams (resp. users) and PRBs. Note that for CTR₁, only the strongest stream of each user can be selected.

We use CTR_F in the following to present GUS, since it is the most intricate case. We will discuss GUS for BD and CTR₁ later. In the case of CTR_F, the heuristic executes an unconstrained greedy-up search on the stream sets per PRB over the TS. Let $v_{u,s}^c \in \{0, 1\}$ be the variable indicating stream selection, where $v_{u,s}^c = 1$ if stream $s \in \mathcal{M}_U$ of user $u \in \mathcal{U}$ is selected in PRB $c \in \mathcal{C}$ and $v_{u,s}^c = 0$ otherwise. Let the stream indices take any integer value from 1 to M_U with stream 1 being the strongest and stream M_U the weakest (see [5, 6] for details on stream strength) and let $\mathcal{M}_U = \{1, \dots, M_U\}$ be the stream indices set. The heuristic starts by initializing $v_{k,j}^b \leftarrow 0, \forall k \in \mathcal{U}, \forall b \in \mathcal{C}, \forall j \in \mathcal{M}_U$, meaning that, initially, there is no scheduled stream. Let WSR_{max} be the largest WSR over the TS registered so far during the search, and $\hat{v}_{k,j}^b$ be the best stream allocation over the TS yielding WSR_{max} . They are initialized as $WSR_{max} \leftarrow 0$ and $\hat{v}_{k,j}^b \leftarrow 0, \forall k \in \mathcal{U}, \forall b \in \mathcal{C}, \forall j \in \mathcal{M}_U$. At each search iteration, we add one stream in one of the PRBs. Specifically, at the beginning of a search iteration, we assess each candidate for selection (u, c, s) where s is a stream of user u in PRB $c \in \mathcal{C}$ such that $v_{u,s}^c = 0$ (i.e., s has not yet been scheduled in c). Clearly, the number of such candidates for selection can be very large, especially at the beginning of the search. The selected candidate for scheduling is the one that yields the largest WSR over the TS when scheduled together with the already selected streams over all PRBs.

Let $WSR_{u,s}^c$ be the WSR over the TS when candidate (u, c, s) is assessed. For the sake of computing $WSR_{u,s}^c$, assume stream s of user u has been selected temporarily in PRB c (i.e., $v_{u,s}^c \leftarrow 1$).

To compute $WSR_{u,s}^c$ for candidate (u, c, s) , the steps are:

- **Precoding Computation:** To improve the search computational efficiency, we reuse effective channel values (when possible) and employ fast precoding computation methods [5, 12]. Then, we obtain the effective channels, $E_{u,s}^b, \forall u \in \mathcal{U}, \forall s \in \mathcal{S}_u^b, \forall b \in \mathcal{C}$.

- **Rate Computation:** To obtain the rates for each selected stream over the TS, given the addition of candidate (u, c, s) , we need to execute power management (i.e., PA and PD) for each user $k \in \mathcal{U}$ that has been selected in PRB c (i.e., $\forall k \in \mathcal{U}$ s.t. $\exists j \in \mathcal{M}_U$ s.t. $v_{k,j}^c = 1$) and obtain the power allocated for each of its selected stream $P_{k,s}^b, \forall j \in \mathcal{M}_U, \forall b \in \mathcal{C}$ over all channels c of the TS. Executing power management for each user $k \in \mathcal{U}$ selected in PRB c is necessary because the effective channel values for the currently selected streams in PRB c most likely change when adding stream s of user u .

Due to the nature of the MCS-based function $f(\cdot)$, the power management problem is non-convex [14], preventing the use of efficient algorithms such as WF [28]. Instead, we follow a similar approach to [25, 30] and fit $f(\cdot)$ with a smooth concave function, specifically $A \log(1 + D\gamma)$, where A and D are non-negative coefficients, where γ is the SNR. Using the fitted function, the power management (joint power allocation and power distribution) problem (JPM in short) over the TS can be written as, for a user $k \in \mathcal{U}$,

$$\begin{aligned} & \mathbf{JPM}(k, \{E_{k,j}^b\}_{b,j}) : \\ & \max_{P_{k,j}^b} \sum_{b \in \mathcal{C}} \sum_{j \in \mathcal{M}_U} A \log(1 + DE_{k,j}^b P_{k,j}^b) \end{aligned} \quad (4.1)$$

$$\text{s.t.} \quad \sum_{b \in \mathcal{C}} \sum_{j \in \mathcal{M}_U} P_{k,j}^b \leq P_U; \quad (4.2)$$

$$0 \leq P_{k,j}^b \leq \tau_{k,j}^b, \forall b \in \mathcal{C}, \forall j \in \mathcal{M}_U, \quad (4.3)$$

which can be solved efficiently by WF using Algorithm 6 from [28]. The power is upper bounded in Constraint (4.3). Specifically, $\tau_{k,j}^b$ is the power necessary for stream j to achieve an SNR equal to the SNR corresponding to the MCS of highest order, since allowing a larger SNR would be wasteful as there is no rate increase.

For each user $k \in \mathcal{U}$ that has a stream selected in PRB c , we compute the rates $r_{k,j}^b$ using the fitted MCS-based rate function. Note that the rate computations have to be done very often and hence we are only using some of the characteristics of the MCS-based rate function at this stage. We will use it fully at a later stage.

- **Computation of $WSR_{u,s}^c$:** Finally, the WSR when adding candidate (u, c, s) is

$$WSR_{u,s}^c = \sum_{k \in \mathcal{U}} \sum_{b \in \mathcal{C}} \sum_{j \in \mathcal{M}_U} w_k r_{k,j}^b,$$

where we reuse the rates $r_{k,j}^b$ for users $k \in \mathcal{U}$ that have no streams selected in PRB c , computed in previous search iterations (the rates of users that have not been selected

in PRB c are not affected by scheduling a new stream in that PRB). Note that we do not execute an exact MCS selection during the search to decrease its computational cost. Although MCS selection generally requires a small binary search, executing it several times slows the search down.

It is important to highlight the computational burden caused by this step, as many instances of JPM must be executed for each candidate assessment.

After assessing all unselected streams across the users and PRBs in a given search iteration, we can determine the candidate (u^*, c^*, s^*) that, when selected together with the already selected streams in all PRBs, yields the largest WSR, i.e., $(u^*, c^*, s^*) = \arg\max_{(u \in \mathcal{U}, c \in \mathcal{C}, s \in \mathcal{M}_U)} WSR_{u,s}^c$. Note that the weights w_u are used to compute the weighted sum rates. If $WSR_{u^*,s^*}^{c^*} > WSR_{max}$, then the candidate (u^*, c^*, s^*) is definitely selected (i.e., $v_{u^*,s^*}^{c^*} \leftarrow 1$) and we update $WSR_{max} \leftarrow WSR_{u^*,s^*}^{c^*}$ and $\hat{v}_{u^*,s^*}^{c^*} \leftarrow 1$, and go to the next search iteration. However, if $WSR_{u^*,s^*}^{c^*} < WSR_{max}$ but $WSR_{max}/WSR_{u^*,s^*}^{c^*} \leq \beta$ (where β is a threshold that we choose by trial and error), we still select the candidate (u^*, c^*, s^*) and continue the search (this is because the WSR may start to increase again at some future search iteration), but we do not update WSR_{max} and $\hat{v}_{u^*,s^*}^{c^*}$. Finally, if $WSR_{max}/WSR_{u^*,s^*}^{c^*} > \beta$, we stop the search. Two other evident stopping criteria for the search are when M_B streams have been selected in all PRBs, i.e., $\sum_{k \in \mathcal{U}} \sum_{j \in \mathcal{M}_U} v_{k,j}^b = M_B, \forall b \in \mathcal{C}$, or when all streams of all users have been selected in all PRBs, i.e., $\sum_{j \in \mathcal{M}_U} v_{k,j}^b = M_U, \forall k \in \mathcal{U}, \forall b \in \mathcal{C}$. Upon stopping, the search returns the selected streams over the TS corresponding to the largest WSR_{max} observed during the search (given by $\hat{v}_{k,j}^b$), along with their effective channels.

Next, we perform a last round of power management with a focus on the exact MCS-based rate function. Specifically, we execute PA and PD for each user $k \in \mathcal{U}$ individually over its selected streams over the TS using first WF (by solving JPM) and using the solution as an input to the greedy MCS-aware heuristic from Algorithm 1 in [13]. This MCS-aware heuristic can improve the rates for each user by selecting better MCSs for its scheduled streams over the TS. We call this two-step joint PA and PD, Two-Step Power Management (TPM). It is computationally efficient (due to WF) while considering the MCS mapping (through the MCS-aware heuristic) for performance enhancement. With this MCS selection, some streams in some PRBs may receive a zero rate if their SNRs are below Γ_1 . Unselecting zero-rate streams in PRBs can improve performance since the effective channels of the remaining streams may increase [13]. We do remove all zero-rate streams at once, redo the precoding computations in the PRBs where streams were removed, and redo TPM for the users. Then, we obtain the final rate seen by user $k \in \mathcal{U}$ in the TS, $r_k = \sum_{b \in \mathcal{C}} \sum_{j \in \mathcal{M}_U} r_{k,j}^b$. This ends the GUS heuristic for CTR_F .

GUS for both CTR₁ and BD is similar but simpler, with the key difference being that the search scope is over users instead of streams: for BD, we use all M_U streams when assessing and selecting a user in a particular PRB, whereas, for CTR₁, we assess and select only the strongest stream of a user in a certain PRB. Put differently, the candidates in a given search iteration for both BD and CTR₁ are not defined by specific streams; a candidate for BD and CTR₁ is (u, c) defined by an unselected user $u \in \mathcal{U}$ in PRB $c \in \mathcal{C}$.

4.3 Long-term Performance Computation

At the end of the computations for the TS, in line with the literature [14, 31], we update the moving average per-TS rate transmitted by each user $u \in \mathcal{U}$ as well as the weights under PF. Let W be a certain window size (a number of TSs) and R_u be the moving average per-TS rate seen by user $u \in \mathcal{U}$ in the past W TSs. Then, the R_u 's are updated at the end of the computations for a TS as:

$$R_u = \frac{(W-1)R_u + r_u}{W}, \quad \forall u \in \mathcal{U},$$

and the weights w_u 's are updated as: $w_u = 1/R_u, \forall u \in \mathcal{U}$. As discussed, a large number of TSs must be run for each realization under PF for two reasons: i) the TSs computations are coupled through the weights, hence it is necessary to allow the weights to evolve across many TSs; ii) wireless channels are inherently variable over time and frequency and it is important to take the variations into account in the performance evaluations of the UL of MU-MIMO systems.

Given a 3GPP scenario, a precoding strategy, a set of network parameters $\mathcal{N}$ (described in Section 2.3) and a realization $\omega(\Delta)$ (where Δ is a sufficiently large number of TSs for performance convergence and will be discussed shortly), we initialize $R_u, \forall u \in \mathcal{U}$, to a per-TS rate corresponding to a certain MCS. The computations across TSs commence with a warm-up period of $\hat{T}_u$ TSs to mitigate the initial influence of w_u values on the final results, i.e., we select $\hat{T}_u$ such that $W \leq \hat{T}_u$. After the warm-up period, we will proceed with enough TSs, so that the TM converges. The TM is the Geometric Mean (GM) of the average rates R_u . After the warm-up period of $\hat{T}_u$ TSs, we collect the TM at the end of this period. Then, we run the computations for an extra T_u TSs and check the TM convergence. If the relative difference in the TM is within a certain tolerance tol , we stop the computations; otherwise, we run the computations for more T_u TSs. Once we have reached convergence, we have the value of Δ . We also keep track of the aggregated per-TS rates λ_u for each user $u \in \mathcal{U}$ over the TSs executed after the warm-up period of

$\hat{T}_u$ TSs. Algorithm 2 summarizes the procedure on the UL for the long-term performance computations with many TSs under PF for a realization $\omega(\Delta)$ (with the channel model for each user $u \in \mathcal{U}$ across the TSs defined in Section 2.4), a set of network parameters $\mathcal{N}$, a 3GPP scenario and a precoding strategy. The output of Algorithm 2 is the GM of the aggregated per-TS rates per user $\{\lambda_u\}_{u \in \mathcal{U}}$ normalized by the total number of executed TSs (essentially Δ). In Algorithm 2, t represents the TS index and $GM(\cdot)$ computes the GM of a group of values. Algorithm 2 contains two main loops, and we check the TM convergence at the end of each loop. The first loop runs the first block of computations comprising $\hat{T}_u + T_u$ TSs and the second loop runs additional blocks of computations if the first block is insufficient for TM convergence.

Algorithm 2 Algorithm for long-term performance computations with many TSs on the UL for PF.

Input: Realization $\omega(\Delta)$, network parameters $\mathcal{N}$, 3GPP scenario, precoding strategy, numbers of TSs for computations T_u , $\hat{T}_u$, window W and convergence tolerance tol

Output: $GM(\{\lambda_u\}_{u \in \mathcal{U}})/t$

- 1: Initialize $R_u, w_u \leftarrow 1, \lambda_u \leftarrow 0, \forall u \in \mathcal{U}$
 - 2: **for** TS $t = 1$ to $\hat{T}_u + T_u$ **do** ▷ First block of computations
 - 3: Obtain $r_u, \forall u \in \mathcal{U}$, by executing the procedure in Section 4.2 and update $R_u, w_u, \forall u \in \mathcal{U}$
 - 4: **if** $t > \hat{T}_u$ **then**
 - 5: Make $\lambda_u \leftarrow \lambda_u + r_u, \forall u \in \mathcal{U}$
 - 6: **if** $t = \hat{T}_u$ **then**
 - 7: Make TM $\leftarrow GM(\{R_u\}_{u \in \mathcal{U}})$
 - 8: **if** $|GM(\{R_u\}_{u \in \mathcal{U}})/TM - 1| \leq tol$ **then**
 - 9: **return** $GM(\{\lambda_u\}_{u \in \mathcal{U}})/t$
 - 10: Make TM $\leftarrow GM(\{R_u\}_{u \in \mathcal{U}}), t' \leftarrow t + 1$
 - 11: **for** TS $t = t'$ to $t' + T_u$ **do** ▷ Additional blocks of computations
 - 12: Obtain $r_u, \forall u \in \mathcal{U}$, by executing the procedure in Section 4.2 and update $R_u, w_u, \forall u \in \mathcal{U}$
 - 13: Make $\lambda_u \leftarrow \lambda_u + r_u, \forall u \in \mathcal{U}$
 - 14: **if** $|GM(\{R_u\}_{u \in \mathcal{U}})/TM - 1| \leq tol$ **then**
 - 15: **return** $GM(\{\lambda_u\}_{u \in \mathcal{U}})/t$
 - 16: **else**
 - 17: go to Line 10
-

4.4 Search Rate-reusing Method

In this section, we address the reuse of rates computed in previous iterations during the greedy-up search to improve the computational efficiency of the proposed search described in Section 4.2. More specifically, we describe which rates must be computed in a given search iteration and which rates can be reused from previous search iterations.

To recapitulate what we discussed in Section 4.2, to compute $WSR_{u,s}^c$ for a candidate (u, c, s) , we execute precoding computations using a fast computation method, then the rate computations are done with WF-based PAs and PDs (i.e., by solving JPM) using a smooth fitted rate function (no search is necessary on the MCSs). Nonetheless, the number of candidates in a particular search iteration can be very large, as discussed previously. We saw that for each candidate assessment, possibly one precoding computation is needed, whereas several rate computations (where PA and PD are the sources of complexity) are required. The high computational cost from rate computations can hinder performance evaluations on the UL for large systems over many TSs. To reduce the number of rate computations in each search iteration and make the search more efficient, we have established rules to decide precisely which rates must be recomputed when assessing the candidates in a given search iteration and which rates can be reused from previous search iterations. These rules are given here for BD. They can easily be extended to CTR_1 and CTR_F . They can reduce the heuristic run time by about 42%, as shown in Section 4.5, while not affecting the performance as we are only being judicious in what to recompute.

At the end of an iteration, we have a newly selected candidate, and we will classify the candidates for the next iteration into four types (some types are exclusive and some are not) w.r.t. the newly selected candidate. We are at the beginning of an iteration, and the selected candidate at the previous iteration was in PRB $\hat{c}$, which we call the reference PRB. Recall that for a candidate (u, c) , according to Section 4.2, we perform rate computations through PA and PD for users that have already been selected in PRB c as well as for user u , and we can reuse the rates of the other users. What we are trying to do next is limit further the number of those users for whom we do indeed perform PA and PD for rate computations and instead use rate computations done earlier. The candidate types are:

- **Type 1:** This type includes all the candidates $(u, \hat{c})$. These candidates cannot be classified into other types and no further rate reuse can be done other than the ones defined in the rate computations of Section 4.2.
- **Type 2:** This type includes all candidates with a user already selected in the reference PRB, i.e., all candidates (u, c) such that user u has been selected in PRB $\hat{c}$ (note that

$c \neq \hat{c}$ by definition). To compute the WSR over the TS for each such candidate, we have to execute one rate computation through PA and PD for the candidate user u to obtain its rates. The rates of the other users selected in the other PRBs, excluding the reference PRB, can be reused from previous iterations. The rates of the users selected in the reference PRB (other than the candidate user u) might have to be updated. This last case occurs when the candidate (u, c) is not only a candidate of type 2 but also type 3. Type 3 candidates are discussed below.

To compute the WSR WSR_u^c for a candidate (u, c) of type 2, assuming we temporarily select user u in PRB c , we execute a rate computation by doing a PA and PD (through JPM) for user u and obtain the rates $r_{u,j}^b, \forall b \in \mathcal{C}, \forall j \in \mathcal{M}_U$ for its selected streams. Whether we can reuse or have to compute the rates for users $k \in \mathcal{U}$ selected in PRB $\hat{c}$ with $k \neq u$ is tackled below for candidates of type 3.

- **Type 3:** Candidates of type 3 are characterized as follows: consider the set $\mathcal{K}$ of all users that have been selected in the reference PRB $\hat{c}$. Let $\hat{\mathcal{C}}_k \subseteq \mathcal{C}$ be the set of PRBs for which user $k \in \mathcal{K}$ has been selected. Candidates (u, c) , with $c \neq \hat{c}$, such that $\exists k \in \mathcal{K}$ with $c \in \hat{\mathcal{C}}_k$ is a candidate of type 3, for which we have to execute one rate computation through PA and PD for each $k \in \mathcal{K}$ such that $c \in \hat{\mathcal{C}}_k$. If the candidate is also of type 2, an extra rate computation for user u is required over its selected PRBs plus PRB c as seen previously. To compute the WSR WSR_u^c for a candidate (u, c) of type 3, for each user $k \in \mathcal{K}$ such that $c \in \hat{\mathcal{C}}_k$, we execute a rate computation by PA and PD (through JPM) and obtain the rates $r_{k,j}^b, \forall b \in \mathcal{C}, \forall s \in \mathcal{M}_U$ for its selected streams.
- **Type 4:** Candidates of type 4 do not require new rate computations for the WSR computation. A type-4 candidate is any candidate that is not categorized into any of the previously described types. In this candidate type, we can reuse all the rates computed in previous search iterations.

A candidate can be of types 2 and 3 simultaneously, as discussed, but candidates of type 1 or 4 cannot be of other types. The reuse of rates for all candidate types in a given search iteration relies on the assumption that we computed those rates previously in some search iteration. We have presented above the rules governing which rates have to be computed and which rates can be reused for WSR computation in a given search iteration depending on the candidate type for BD.

Table 4.1: Computation parameters for the UL (RB: reporting block, #: number).

Parameter	Value
# of TSs T_u and $\hat{T}_u$	6 and 6
# of channels (C_B) and TSs (T_B) in a RB	13 and 2
Subchannels per TS (C)	78
Carrier frequency (f_c)	3.5 GHz
Subchannel bandwidth (B_C)	360 kHz
TS duration (d)	0.5 ms
Antennas at the BS (M_B); at UE (M_U)	100, 64; 4, 2, 1
Transmit power per user in mW (P_U)	50 (RMa), 1 and 5 (UMa)
Convergence tolerance (tol)	10^{-2}
Window length in TSs (W)	6
Channel correlation (BS and users)	0.4
Adjacent antenna distance (BS and users)	$0.5 \times 3 \times 10^8 / f_c$
Greedy-up search stopping parameter (β)	1.05
Coefficients for MCS fitting A ; D	1.389; 0.5191

4.5 Numerical Results

In this section, we present the numerical results for the performance of realistic MU-MIMO systems with the precoding strategies CTR_F, CTR₁ and BD for PF in the 3GPP scenarios RMa and UMa for different network parameters. The results are obtained using the procedures described in Sections 4.2 and 4.3 with the rate-reusing method proposed in Section 4.4. We also present the run time reduction in the GUS heuristic when using the proposed rate-reusing method. Table 4.1 shows the key parameters used in the computations. Within a realization $\omega(\Delta)$, the small-scale fading for a user changes independently across distinct reporting blocks consisting of $C_B = 13$ subchannels and $T_B = 2$ TSs while the other channel elements remain constant (see Section 2.4 for the channel model). Finally, we set the greedy-up search stopping parameter β to 1.05 (higher values for β were tested, but no performance improvement was observed). The results are averaged over 50 realizations for a given number of users, and confidence intervals are computed using the t -distribution with a confidence level of 90%.

We will start our evaluation with a comparison of the performance of the precoding strategies CTR_F, BD and CTR₁ in both 3GPP scenarios for different network parameters. We then present the run time reduction in our GUS heuristic when using the proposed rate-reusing method.

4.5.1 Precoding Strategies Evaluation

Fig. 4.1 (top figure) presents, for the RMa scenario, the GM rates per TS for CTR_F , BD and CTR_1 as a function of the number of users for $M_B = 100$ and $M_B = 64$ antennas at the BS. The number of antennas, M_U , at the users is 4 and $P_U = 50$ mW. The ratios in the GM rates between BD and CTR_F , and between CTR_1 and CTR_F are also shown in Fig. 4.1 (bottom figure). The main observation is that BD is not always better than CTR_1 . Their relative performance depends on the values of M_B and $|\mathcal{U}|$. Typically, BD is significantly better for a small number of users while the reverse is true for a large number of users. Clearly, as expected, CTR_F outperforms the other strategies across all numbers of users and BS antennas and the performance gap can be significant. More BS antennas reduce the gap between BD and CTR_F , while it is the reverse for CTR_1 . We can also see in the top figure that M_B has a significant impact on performance for the three strategies for all values of the number of users. Specifically, for CTR_F and BD, the improvement in the GM rates for $M_B = 100$ relative to $M_B = 64$ starts at around 30% for 10 users and grows with the number of users to above 40%. For CTR_1 , similar trends are observed, except that the improvement for 10 users is smaller at roughly 16%.

These observations can be understood better by referring to Fig. 4.2, which shows the empirical distribution of the stream allocation pattern per selected user per PRB given by CTR_F for different numbers of users and $M_B = 64$ BS antennas. In Fig. 4.2, the x axis represents which stream pattern is enabled per scheduled user per PRB. Recall that the streams are numbered from 1 to $M_U = 4$ in order of strength. For example, referring to the $|\mathcal{U}| = 10$ case, 30% of the users have only stream 1 scheduled in a PRB allocated to them, and around 20% have streams 1, 2, and 3. For $|\mathcal{U}| = 10$ users, around 40% of the time, at least three streams are enabled per scheduled user, whereas for $|\mathcal{U}| = 40$ users, users with high stream counts barely occur and one-stream users are observed in almost 80% of the cases. These results are in line with the performance given by Fig. 4.1, where CTR_1 , which only uses one stream per scheduled user, performs comparably to CTR_F for 40 users, and BD, presenting four streams per scheduled user, does similarly for 10 users.

Next, we show the equivalent results for UMa in Figs. 4.3 and 4.4 when $P_U = 1$ mW. Compared to the results for RMa (see Fig. 4.1), the main differences are that the gap between CTR_1 and CTR_F is small across all values of the number of users for the two values of M_B and that CTR_1 outperforms BD, except for $|\mathcal{U}| = 10$ and $M_B = 100$. These observations are corroborated by Fig. 4, where the most common stream allocation given by CTR_F is stream 1 enabled per scheduled user, which is essentially CTR_1 . Similar to the RMa case, we can also observe in the top figure that M_B also has a considerable impact on performance for the three strategies for all values of the number of users. Particularly,

for all strategies, the improvement in the GM rates with $M_B = 100$ compared to $M_B = 64$ is at least 30% for a few users (e.g., 10 users) and increases with the number of users to reach over 40%.

In summary, depending on the 3GPP scenario, the strategy that could replace CTR_F , if deemed too complex to operate in real-time, varies. For RMa, considering that the number of users is generally not large (below 30) in a sector, BD could replace the more complex CTR_F given its comparable performance. For larger user counts, CTR_1 could be an alternative to CTR_F because of the small performance gap. For UMa, CTR_1 emerges as a suitable option over CTR_F due to its lower complexity and similar performance.

Next, we evaluate the performance of the three precoding strategies for different numbers of antennas, M_U , at the users in the UMa scenario. For this study, we employ a per-user power budget of $P_U = 5$ mW because for $M_U = 1$, $P_U = 1$ mW can yield unconnected users. Fig. 4.5 (top figure) presents the GM rates per TS for the ZF strategies vs. number of users for different values of M_U with $M_B = 64$ BS antennas. Fig. 4.5 (bottom figure) shows the corresponding ratios for BD and CTR_1 relative to CTR_F . For $M_U = 1$, only results for CTR_1 are provided, given that all strategies are equivalent. We can observe that more antennas at the users produce a significant improvement in the GM rates delivered by CTR_F and CTR_1 . Specifically, the improvements in the GM rates for CTR_F (from $M_U = 2$ to $M_U = 4$) and CTR_1 (from $M_U = 1$ to $M_U = 2$ and from $M_U = 2$ to $M_U = 4$) fluctuate at approximately 20% for all numbers of users. Conversely, the improvement for BD is low for a small number of users and is nonexistent when the number of users increases. For instance, for 10 users, the improvement from $M_U = 2$ to $M_U = 4$ for BD is around 12% but drops to nearly zero for 40 users. Another interesting observation is that the performance disparities for CTR_1 and BD w.r.t. CTR_F decrease for $M_U = 2$ compared to $M_U = 4$. This can be explained by the fact that with fewer antennas at the users, CTR_F has less flexibility in the stream allocation and loses part of its potential.

We also evaluate the performance of the precoding strategies for different per-user power budgets, P_U . Fig. 4.6 (top figure) presents the GM rates per TS for the ZF strategies vs. number of users for different values of P_U with $M_B = 64$ BS antennas, $M_U = 4$ antennas at the users in the UMa scenario. Fig. 4.6 (bottom figure) shows the corresponding ratios for BD and CTR_1 relative to CTR_F . The GM rates for all precoding strategies approximately double for $P_U = 5$ mW compared to $P_U = 1$ mW, which is very significant. Also, the performance gap for BD w.r.t. CTR_F remains nearly the same for the two power budgets. On the other hand, the difference between CTR_1 and CTR_F increases with more power. One reason for this is that, owing to the extra capacity provided by increased power, more streams per scheduled user can be enabled with CTR_F .

4.5.2 Miscellaneous

Finally, we investigate how much decrease in run time we can obtain by executing GUS with the more aggressive rate-reusing method proposed in Section 4.4 w.r.t. GUS without it as proposed in Section 4.2. We consider the RMa scenario with $M_B = 64$ BS antennas, $|\mathcal{U}| \in \{20, 40, 60, 80\}$ users, $M_U = 4$ antennas at the users, and $P_U = 50$ mW. We use CTR_1 and BD to illustrate that. We do not employ CTR_F as an example, since it can take dozens of hours for just one realization, even with the rate-reusing method. The complexity of CTR_F comes from the increased granularity in the stream selection process (i.e., flexible stream allocation per user) and the fact that the precoding computations are more costly. The run times are averaged over at least 650 TSs for a given number of users. For CTR_1 , the decrease in run time is up to roughly 11% and for BD, the run time reduction ranges approximately from 26% to 42%. Computational efficiency is critical due to the long period needed to conduct the comprehensive evaluations in this work. For instance, more than four hours were required for only one realization of CTR_1 with 80 users without the rate-reusing method. In this study, we had to execute a large number of realizations for different precoding strategies, across different 3GPP-based scenarios and network parameters, highlighting the importance of the proposed rate-reusing method for run time reduction.

4.6 Conclusion

We evaluated the performance of an MU-MIMO single cell on the UL with MA users for three ZF-based precoding strategies: CTR_F , BD and CTR_1 . The study was conducted across multiple TSs considering PF in realistic systems, characterized by 3GPP-based scenarios with practical MCSs and a large number of BS antennas. To enable comprehensive evaluations in large realistic systems, we proposed an efficient heuristic based on a greedy-up search for user-sets with CTR_1 or BD (resp. stream-sets with CTR_F) on the UL. To reduce the heuristic run time, we proposed a novel rate-reusing method where rates computed in previous search iterations are reused whenever possible. The proposed greedy-up search equipped with the rate-reusing method answers [RQ 4.1](#).

The results showed that CTR_F outperformed the other strategies across all scenarios and network parameters. However, depending on the 3GPP scenario, either CTR_1 or BD could replace CTR_F . For RMa, considering that the number of users is generally not large (below 30) in a sector, BD could replace the more complex CTR_F given its comparable performance. For larger user counts, CTR_1 could be an alternative to CTR_F because of

the small performance gap. For UMa, CTR_1 emerged as a suitable option over CTR_F due to its lower complexity and similar performance across the considered network parameters. We also demonstrated that the numbers of antennas at the BS and at the users, and the power budget per user substantially impacted the performance of the precoding strategies. These observations address [RQ 4.2](#) and can guide the design of real-time RRM heuristics for the UL of MU-MIMO systems with MA users under PF.

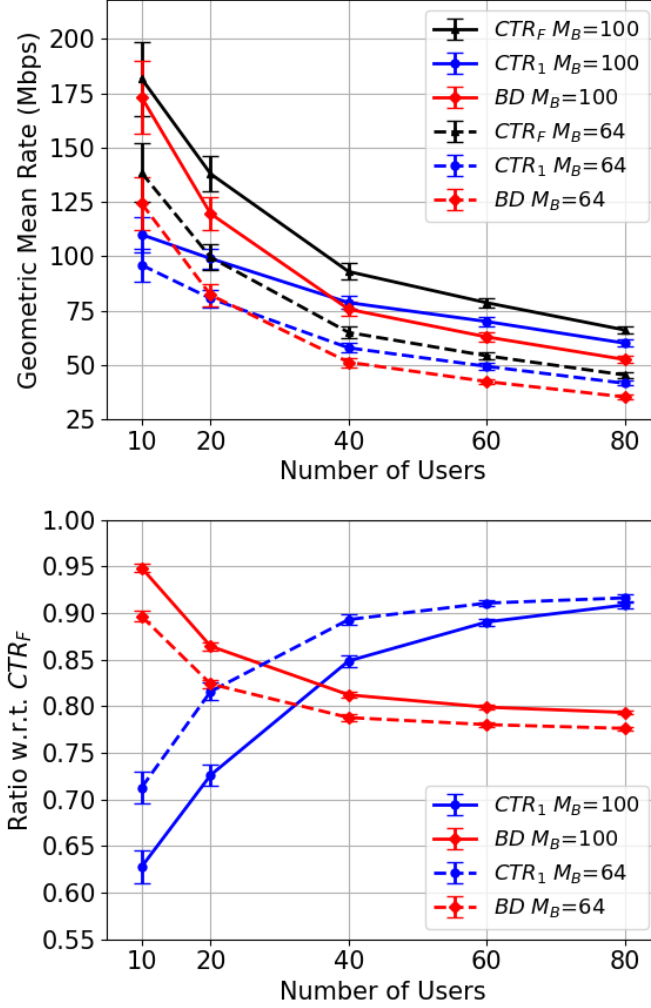


Figure 4.1: RMa scenario: $M_B = 100$ and $M_B = 64$ antennas at the BS, $M_U = 4$ antennas at the users, $P_U = 50$ mW. Top figure: GM rates in Mbps per TS vs. number of users for CTR_F , BD and CTR_1 . Bottom figure: ratios in the GM rates between BD and CTR_F , and between CTR_1 and CTR_F .

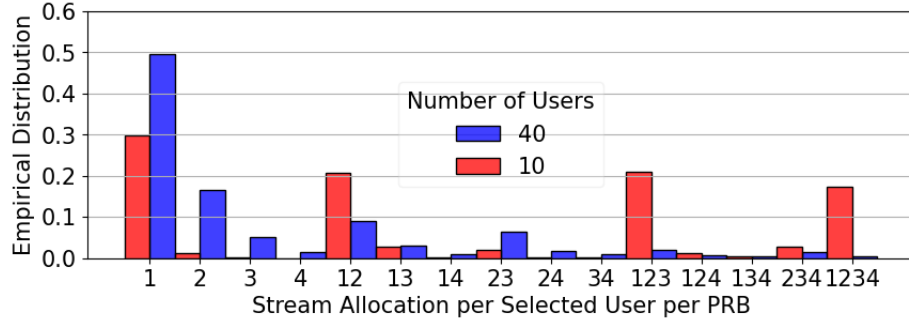


Figure 4.2: RMa scenario: empirical distribution of the stream allocation per scheduled user per PRB given by CTR_F for $|\mathcal{U}| \in \{10, 40\}$ users with $M_B = 64$ BS antennas, $M_U = 4$ antennas at the users, and $P_U = 50$ mW.

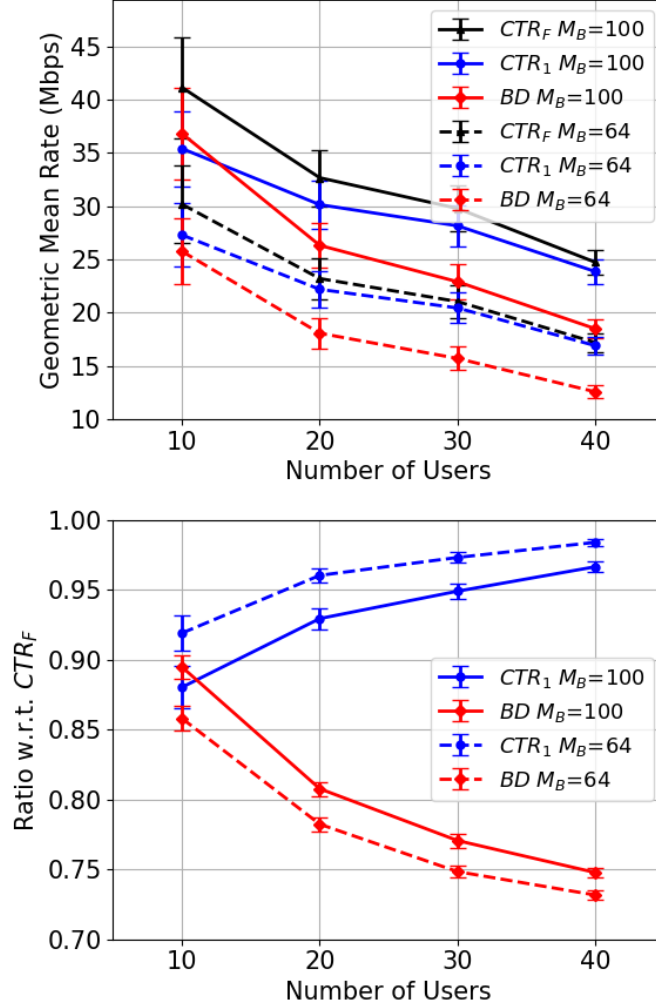


Figure 4.3: UMa scenario: $M_B = 100$ and $M_B = 64$ antennas at the BS, $M_U = 4$ antennas at the users, and $P_U = 1$ mW. Top figure: GM rates in Mbps per TS vs. number of users for CTR_F , BD and CTR_1 . Bottom figure: ratios in the GM rates between BD and CTR_F , and between CTR_1 and CTR_F .

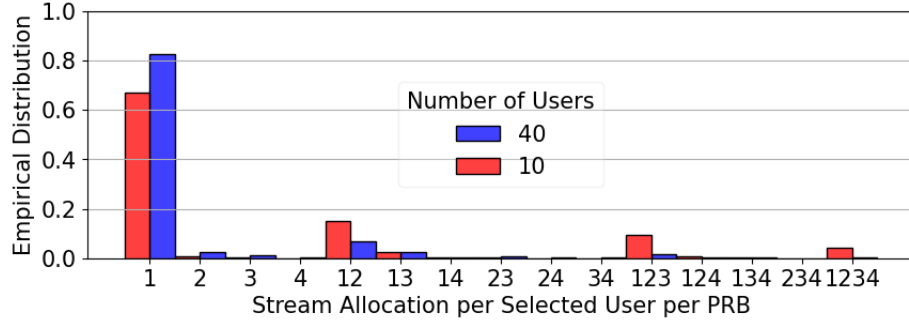


Figure 4.4: UMa scenario: empirical distribution of the stream allocation per selected user per PRB given by CTR_F for $|\mathcal{U}| \in \{10, 40\}$ users with $M_B = 64$ BS antennas, $M_U = 4$ antennas at the users, and $P_U = 1$ mW.

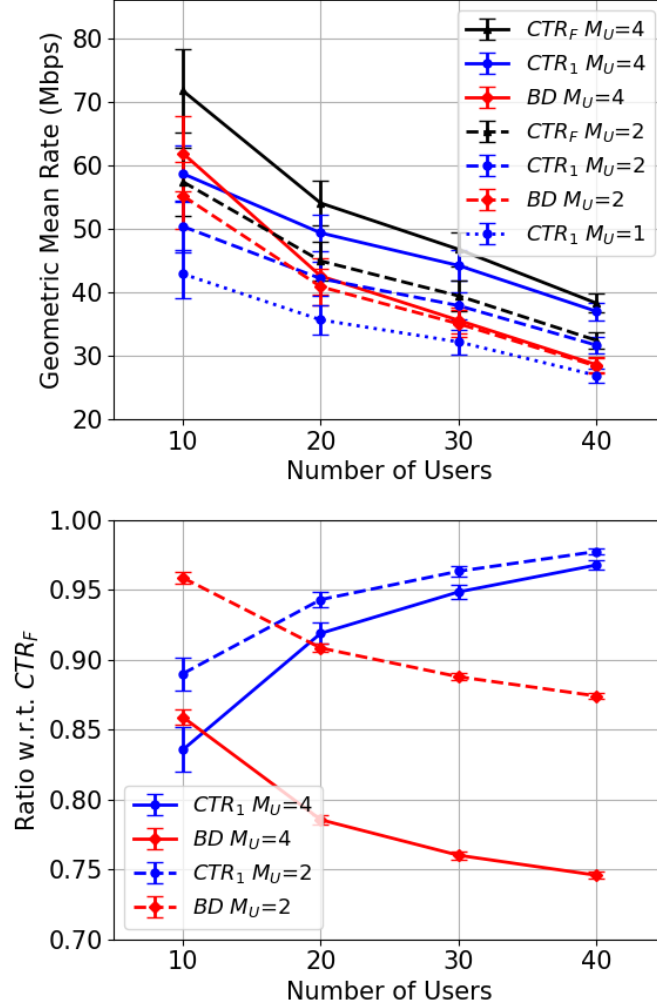


Figure 4.5: UMa scenario: $M_B = 64$ antennas at the BS, $M_U = 4$, $M_U = 2$ and $M_U = 1$ antennas at the users, $P_U = 5$ mW. Top figure: GM rates in Mbps per TS vs. number of users for CTR_F , BD and CTR_1 . Bottom figure: ratios in the GM rates between BD and CTR_F , and between CTR_1 and CTR_F .

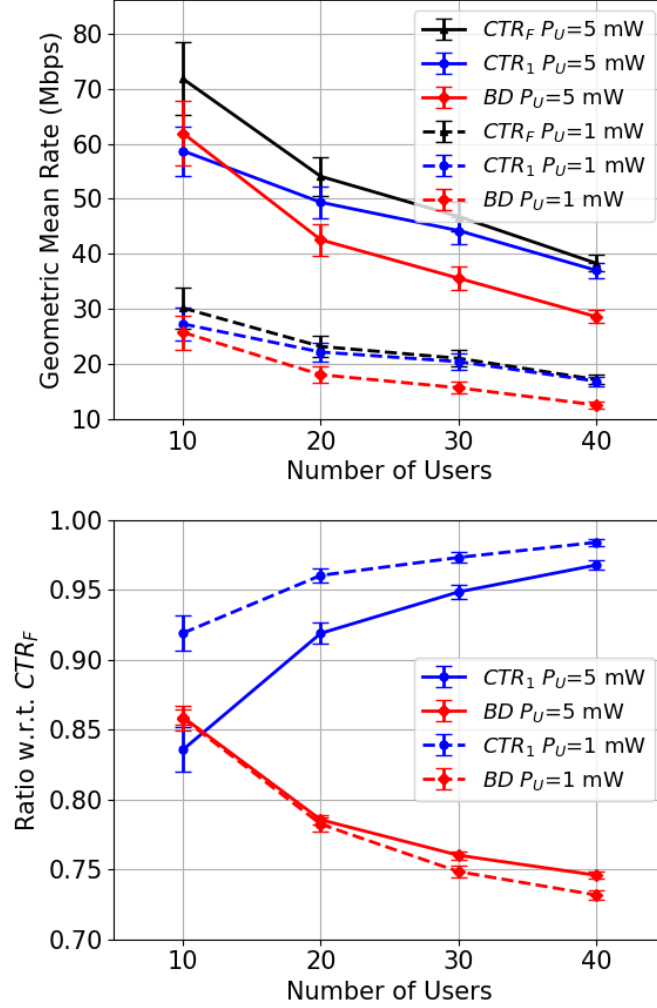


Figure 4.6: UMa scenario: $M_B = 64$ antennas at the BS, $M_U = 4$ antennas at the users, $P_U = 5$ mW and $P_U = 1$ mW. Top figure: GM rates in Mbps per TS vs. number of users for CTR_F , BD and CTR_1 . Bottom figure: ratios in the GM rates between BD and CTR_F , and between CTR_1 and CTR_F .

Chapter 5

Conclusion

5.1 Summary

In this work, we considered both the DL and UL of MU-MIMO/OFDMA systems with MA users and ZF processing, along with their RRM processes, namely user selection, precoding, PD, PA and MCS selection. The DL and UL were studied separately, and in a practical system, such a decoupling can be done in time using TDD. Under the ZF scope, we considered three primary precoding strategies: CTR_F (flexible stream allocation per selected user), CTR_1 (only the strongest stream used per selected user) and BD (all streams enabled per selected user). CTR_F is more complex than BD and CTR_1 since the precoding computations are more costly and requires stream selection as opposed to the simpler user selection for BD and CTR_1 . We focused on scenarios with fairness, specifically PF, and conducted offline (i.e., without time constraints) studies on both the DL and UL, where we aimed to answer pertinent questions and obtain valuable information that can facilitate the design of real-time RRM heuristics. More specifically, we focused on examining the performance gains (in terms of geometric mean rate for PF) brought by the more complex CTR_F over the simpler strategies BD and CTR_1 . All numerical computations were carried out for 3GPP-based scenarios (i.e., RMa, UMa and UMi), with 3GPP-compliant MCSs and a large number of BS antennas.

The DL was studied in Chapter 3. The DL study was based on multiple PRBs for fairness considerations, where in each PRB, we employed greedy-up searches for user-sets for BD and CTR_1 (resp. user-streams for CTR_F). The searches were adapted from the literature to consider MCSs and PF and were equipped with efficient precoding computation methods to enable comprehensive evaluations. We showed that CTR_F outperformed the

other strategies for PF across all 3GPP-based scenarios and network parameters. However, CTR_1 could replace CTR_F for a small number of BS antennas given its comparable performance, while for higher BS antenna counts, the strategy that could replace CTR_F depended on the number of users. We also showed that preceding research under SR maximization, which indicated that BD was outperformed by CTR_F for a few BS antennas, did not hold for larger numbers of BS antennas. In this case, BD performed similarly to CTR_F and emerged as a suitable choice over CTR_F .

The UL was addressed in Chapter 4. To enable the UL study, we extended the greedy-up search heuristic developed for the DL to the UL. Since the UL RRM must be done over the entire TS (as opposed to per PRB on the DL), the per-TS search was equipped with efficient precoding and rate computation methods and with a novel rate-reusing technique where rates computed in previous search iterations were reused whenever possible for computational efficiency. The UL study was conducted across different 3GPP scenarios (RMa and UMa) and network parameters, and was based on multiple TSs to consider PF. We observed that CTR_F outperformed the other strategies in all scenarios, but different trends were observed between the RMa and UMa scenarios. In RMa, the strategy that could replace CTR_F depended on the number of users, with BD and CTR_1 performing comparably to CTR_F for a few and many users, respectively. In UMa, CTR_1 performance was similar to CTR_F across the considered network parameters, hence CTR_1 could be an alternative to CTR_F .

Finally, throughout this work, we addressed RQ 1.1, where we saw the performance gains brought by allowing a flexible stream allocation per user (i.e., CTR_F) compared to the simpler strategies BD and CTR_1 on both the DL and UL. Also, to enable extensive comparisons of the precoding strategies on the UL, we extended the greedy-up search developed for the DL to the UL in an efficient manner, with which we addressed RQ 1.2.

5.2 Potential Research Directions

In the following, we address potential research directions. We first discuss extending the offline study to consider multi-cell settings. Next, we investigate potential directions for real-time RRM studies considering the conclusions drawn in this work.

5.2.1 Extending The Offline Study

In this study, we considered single-cell systems, where, thanks to ZF precoding, the interference among users (streams) scheduled in the same PRB is nullified, and the rates depend only on the SNR. Nonetheless, practical systems often consist of multiple cells, where interference from neighbouring cells can impact performance and the RRM for each cell should be adjusted accordingly. Hence, a compelling extension to this work is to evaluate the performance of the precoding strategies CTR_F , CTR_1 and BD in multi-cell settings on both the DL and UL.

5.2.2 Real-time RRM Study

From the conclusions drawn in this work, there are some potential paths for real-time RRM studies under PF, depending on whether the priority is performance (in terms of rates) or complexity. In the following, we discuss possible approaches from each perspective.

We saw that CTR_F outperformed BD and CTR_1 across all scenarios and network parameters on both the DL and UL. Therefore, from a performance perspective, a RRM heuristic based on CTR_F is worth investigating. Nevertheless, selecting an adequate stream-set under CTR_F for reception in a PRB on the DL depends on the intricate relations between the streams (e.g., correlation and strength) and the users' weights (for fairness). The problem is exacerbated on the UL as stream-sets must be selected for each PRB over the TS at once. Learning-based methods are emerging as promising candidates to solve these complex problems in real time. They have previously been applied in the context of user selection for SA users [32, 33].

On the other hand, we also observed that either BD or CTR_1 could replace CTR_F in some cases due to the small performance gap. For example, on the DL, CTR_1 performed similarly to CTR_F for many users, and BD did the same for a few users. On the UL, similar conclusions were drawn for the RMa scenario, while for the UMa scenario, CTR_1 presented a similar performance to CTR_F across all network parameters. Thus, from a complexity perspective, real-time RRM heuristics based on either CTR_1 , BD, or a combination thereof, could be proposed. Heuristics based on these strategies, instead of CTR_F , are simpler because they require user selection instead of the more complex stream selection. Furthermore, the precoding computations are also less computationally demanding. In [13] and [14], real-time RRM heuristics for the DL and UL, respectively, with SA users (i.e., with user selection) were proposed.

Letter of Copyright Permission

Chapter 3 was adapted from a work previously published in Proceedings of the IEEE. © 2025 IEEE. Reprinted, with permission, from [15].

In reference to IEEE copyrighted material which is used with permission in this thesis, the IEEE does not endorse any of the University of Waterloo's products or services. Internal or personal use of this material is permitted. If interested in reprinting/republishing IEEE copyrighted material for advertising or promotional purposes or for creating new collective works for resale or redistribution, please go to http://www.ieee.org/publications_standards/publications/rights/rights_link.html to learn how to obtain a License from RightsLink. If applicable, University Microfilms and/or ProQuest Library, or the Archives of Canada may supply single copies of the dissertation.

References

- [1] E. Castañeda, A. Silva, A. Gameiro, and M. Kountouris, “An overview on resource allocation techniques for multi-user MIMO systems,” *IEEE Communications Surveys & Tutorials*, vol. 19, no. 1, pp. 239–284, 2017.
- [2] E. Björnson, M. Kountouris, M. Bengtsson, and B. Ottersten, “Receive combining vs. multi-stream multiplexing in downlink systems with multi-antenna users,” *IEEE Transactions on Signal Processing*, vol. 61, no. 13, pp. 3431–3446, 2013.
- [3] F. Rusek, D. Persson, B. K. Lau, E. G. Larsson, T. L. Marzetta, O. Edfors, and F. Tufvesson, “Scaling up MIMO: Opportunities and challenges with very large arrays,” *IEEE Signal Processing Magazine*, vol. 30, no. 1, pp. 40–60, 2013.
- [4] Q. Spencer, A. Swindlehurst, and M. Haardt, “Zero-forcing methods for downlink spatial multiplexing in multiuser MIMO channels,” *IEEE Transactions on Signal Processing*, vol. 52, no. 2, pp. 461–471, 2004.
- [5] Z. Shi, C. Zhao, and Z. Ding, “Low complexity eigenmode selection for MIMO broadcast systems with block diagonalization,” in *2008 IEEE International Conference on Communications*, 2008, pp. 3976–3981.
- [6] F. Boccardi and H. Huang, “A near-optimum technique using linear precoding for the MIMO broadcast channel,” in *2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07*, vol. 3, 2007, pp. III–17–III–20.
- [7] Y. Özcan and C. Rosenberg, “Uplink scheduling in multi-cell OFDMA networks: A comprehensive study,” *IEEE Transactions on Mobile Computing*, vol. 20, no. 10, pp. 3081–3098, 2021.
- [8] G. Femenias and F. Riera-Palou, “Scheduling and resource allocation in downlink multiuser MIMO-OFDMA systems,” *IEEE Transactions on Communications*, vol. 64, no. 5, pp. 2019–2034, 2016.

- [9] Z. Shen, R. Chen, J. Andrews, R. Heath, and B. Evans, “Low complexity user selection algorithms for multiuser MIMO systems with block diagonalization,” *IEEE Transactions on Signal Processing*, vol. 54, no. 9, pp. 3658–3663, 2006.
- [10] T. L. Marzetta, “Massive MIMO: An introduction,” *Bell Labs Technical Journal*, vol. 20, pp. 11–22, 2015.
- [11] 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.
- [12] 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.
- [13] A. Hussein, P. Mitran, and C. Rosenberg, “Operating multi-user massive MIMO networks: Trade-off between performance and runtime,” *IEEE Transactions on Network and Service Management*, vol. 21, no. 2, pp. 2170–2186, 2024.
- [14] H. Rahman and C. Rosenberg, “Resource allocation for the uplink of a multi-user massive MIMO system,” *IEEE Transactions on Mobile Computing*, vol. 24, no. 5, pp. 4326–4338, 2025.
- [15] J. P. P. G. Marques, K. Danilchenko, and C. Rosenberg, “Performance evaluation of MU-MIMO systems with multi-antenna users for different precoding strategies,” in *2025 IEEE Wireless Communications and Networking Conference (WCNC)*, 2025, pp. 1–6.
- [16] K. Ko and J. Lee, “Multiuser MIMO user selection based on chordal distance,” *IEEE Transactions on Communications*, vol. 60, no. 3, pp. 649–654, 2012.
- [17] L.-N. Tran, M. Bengtsson, and B. Ottersten, “Iterative precoder design and user scheduling for block-diagonalized systems,” *IEEE Transactions on Signal Processing*, vol. 60, no. 7, pp. 3726–3739, 2012.
- [18] S. Nam, J. Kim, and Y. Han, “A user selection algorithm using angle between subspaces for downlink MU-MIMO systems,” *IEEE Transactions on Communications*, vol. 62, no. 2, pp. 616–624, 2014.
- [19] C. Guthy, W. Utschick, R. Hunger, and M. Joham, “Efficient weighted sum rate maximization with linear precoding,” *IEEE Transactions on Signal Processing*, vol. 58, no. 4, pp. 2284–2297, 2010.

- [20] R. Chen, Z. Shen, J. G. Andrews, and R. W. Heath, "Multimode transmission for multiuser MIMO systems with block diagonalization," *IEEE Transactions on Signal Processing*, vol. 56, no. 7, pp. 3294–3302, 2008.
- [21] C. Guthy, W. Utschick, R. Hunger, and M. Joham, "Weighted sum rate maximization in the MIMO MAC with linear transceivers: Algorithmic solutions," in *2009 Conference Record of the Forty-Third Asilomar Conference on Signals, Systems and Computers*, 2009, pp. 1474–1478.
- [22] L. Ping and P. Wang, "Multi-user gain and maximum eigenmode beamforming for MIMO systems with rate constraints," in *2007 IEEE Information Theory Workshop on Information Theory for Wireless Networks*, 2007, pp. 1–5.
- [23] L. Zhang, A. Liu, and X. Chen, "A WMMSE-based contiguous resource scheduling algorithm for 5G-NR uplink," *IEEE Wireless Communications Letters*, vol. 13, no. 2, pp. 466–470, 2024.
- [24] X. Meng, Y. Chang, Y. Wang, and J. Wu, "Multi-user grouping based scheduling algorithm in massive MIMO uplink networks," in *2018 IEEE 4th International Conference on Computer and Communications (ICCC)*, 2018, pp. 409–413.
- [25] A. Lappalainen, Y. Zhang, and C. Rosenberg, "Planning 5G networks for rural fixed wireless access," *IEEE Transactions on Network and Service Management*, vol. 20, no. 1, pp. 441–455, 2023.
- [26] 3GPP, "5G; Study on channel model for frequencies from 0.5 to 100 ghz," 3rd Generation Partnership Project, TR 38.901, Version 16.1.0, 2020. [Online]. Available: <https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3173>
- [27] —, "5G; NR; Physical layer procedures for data," 3rd Generation Partnership Project, TS 38.214, Version 15.2.0, 2018. [Online]. Available: <https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3216>
- [28] C. Xing, Y. Jing, S. Wang, S. Ma, and H. V. Poor, "New viewpoint and algorithms for water-filling solutions in wireless communications," *IEEE Transactions on Signal Processing*, vol. 68, pp. 1618–1634, 2020.
- [29] Y. Zhang, P. Mitran, and C. Rosenberg, "Joint resource allocation for linear precoding in downlink massive MIMO systems," *IEEE Transactions on Communications*, vol. 69, no. 5, pp. 3039–3053, 2021.

- [30] Y. Quan, S. Shahsavari, and C. Rosenberg, “Planning and operation of millimeter-wave downlink systems with hybrid beamforming,” *IEEE Transactions on Wireless Communications*, vol. 23, no. 12, pp. 18 214–18 227, 2024.
- [31] Y. Chen, Y. Wu, Y. T. Hou, and W. Lou, “mCore: Achieving sub-millisecond scheduling for 5G MU-MIMO systems,” in *IEEE INFOCOM 2021 - IEEE Conference on Computer Communications*, 2021, pp. 1–10.
- [32] C. Feres and Z. Ding, “An unsupervised learning paradigm for user scheduling in large scale multi-antenna systems,” *IEEE Transactions on Wireless Communications*, vol. 22, no. 5, pp. 2932–2945, 2023.
- [33] C.-H. Hsu and Z. Ding, “Unsupervised learning for resource allocation and user scheduling in wideband MU-MIMO systems,” *IEEE Open Journal of the Communications Society*, vol. 5, pp. 2240–2256, 2024.