Performance Analysis of Zero-Forcing Beamforming Strategies for the Uplink of an MU-MIMO System with Multi-Antenna Users

João Paulo P. G. Marques, Catherine Rosenberg *Fellow, IEEE*

***Abstract*—We conduct a comprehensive evaluation of the performance of the uplink of OFDMA-based MU-MIMO systems with multi-antenna users, for three Zero-Forcing (ZF) Beamforming (BF) strategies: Coordinated-Transmit-Receive-1 (CTR_1), where only the strongest data stream is enabled per scheduled user; Block Diagonalization (BD), where all possible streams are enabled per scheduled user; Coordinated-Transmit-Receive-Flexible (CTR_F), which allows a flexible stream allocation per user. The Radio Resource Management (RRM) of the uplink of all OFDMA-based systems must be done over an entire Time-Slot (TS) due to power management, making it challenging. To enable this study, we propose an efficient heuristic based on greedy-up searches for stream-sets that provides feasible solutions. It operates over the TS and considers fairness, practical Modulation and Coding Schemes and all RRM processes. The results show that, for Rural Macro scenarios, BD (resp. CTR_1) could replace the more complex CTR_F if the number of users is small (resp. large), while for Urban Macro scenarios, CTR_1 emerges as an alternative to CTR_F due to its similar performance. We also show that the system parameters can substantially impact the performance of the ZF strategies and that BD performance is more impaired with a simpler power management scheme than CTR_1 and CTR_F.**

***Index Terms*—MU-MIMO, Uplink, Radio Resource Management, Zero-Forcing Beamforming, Multi-Antenna Users.**

I. Introduction

MULTI-user Multiple-Input Multiple-Output (MU-MIMO) is a key technology to enhance the performance of wireless cellular systems by allowing transmissions of multiple data streams at the same time and frequency [1]. This work considers the Uplink (UL) of an MU-MIMO single-cell system with Multi-Antenna (MA) users and Orthogonal Frequency-Division Multiple Access (OFDMA). Most of the studies on MU-MIMO have focused on Single-Antenna (SA) users and the Downlink (DL) [2]. With MA users, there is greater potential for performance improvement as the multiple antennas can be used to enable one strong data stream for many users or several data streams per user for fewer users [3]. Addressing the UL has become increasingly critical given the growth in the UL traffic and the UL impact on the perceived quality of service [4], [5].

Radio Resource Management (RRM) must be carefully addressed to achieve the MU-MIMO potential both on the DL and on the UL. RRM with MA users is composed of the following inter-related processes: Stream Selection (SS) (as opposed to User Selection with SA users), Beamforming (BF), power management and Modulation and Coding Scheme (MCS) selection. Power management with MA users comprises Power Allocation (PA) and Power Distribution (PD) on both the UL and the DL as opposed to PD (resp. PA) alone for the DL (resp. UL) for SA users.

SS is the process that selects a subset of streams for transmission in each Physical Resource Block (PRB)[1]. On the DL, SS can be done independently on each PRB, while on the UL, it has to be done on all the PRBs within a Time-Slot (TS) as will be explained shortly.

BF is the process that enables the Base-Station (BS) to send to or receive from multiple users at once on the DL or the UL, respectively. Zero-Forcing (ZF) BF comprises a family of techniques that eliminates inter-user interference [6], [7]. They also enable the decoupling of BF and power management, simplifying the RRM procedure. ZF BF techniques have been studied extensively in the literature and several efficient computational methods have been developed [8]–[11]. They are the BF techniques used in this paper. There are three primary ZF BF strategies based on the number of data streams used per scheduled user: 1) Coordinated-Transmit-Receive-1 (CTR_1) that uses exactly one stream per selected user (the strongest) [9]; 2) Block Diagonalization (BD) that uses all possible streams per selected user [7] (this number depends on the number of antennas at the user); 3) Coordinated-Transmit-Receive-Flexible (CTR_F) that allows flexible stream allocation [6]. CTR_F can deliver better performance as any stream combination can be enabled per user at the price of higher BF computational complexity compared to the other strategies [6], [7]. Also, the SS process under CTR_F is more complex due to the flexibility in the stream allocation, while for CTR_1 or BD, the stream allocation is fixed for each scheduled user [12]. Investigating offline (i.e., without time constraints) when the simpler strategies CTR_1 and BD perform comparably to CTR_F is paramount to the design of real-time RRM solutions for MU-MIMO systems with MA users. While this was studied on the DL in [3], [6], [9], [13], [14], there is no equivalent study on the UL. In fact, there are no studies on UL RRM for CTR_F to the best of our knowledge and only very limited studies for BD.

The UL RRM problem is more challenging than its DL counterpart because of power management. On the UL, the power belongs to each user individually and the PA process allocates the power budget in a given TS across the PRBs for which the user is selected for transmission. Then, with MA users, PD is required for distributing the power allocated to each PRB by PA among the user's data streams enabled in that PRB. On the DL, in contrast, the power belongs to the BS. This difference between the UL and DL is key from an RRM

This work was conducted when all authors were at the Department of Electrical and Computer Engineering at the University of Waterloo, Canada.

[1]A pair made of a subchannel and a time-slot is the smallest scheduling unit and is called a PRB.

perspective as we shall see and this means that the numerous results and techniques developed for the DL cannot be used as such on the UL. RRM can be done on a per-PRB basis on the DL by allocating, a priori, the BS power equally among the PRBs in a TS [15], while on the UL, the RRM must be done over all PRBs in a TS at once [4]. This is because PA on the UL for a given user cannot be carried out before knowing in which PRBs (over the entire TS) the user is selected for transmission.

An RRM process that is often ignored in the literature is MCS selection, i.e., the process by which rate adaptation is performed by selecting an MCS out of a finite set of pre-defined MCSs based on the outcomes of SS and power management. Instead, papers use a Shannon-based rate adaptation, which is much less realistic [15].

Solving the RRM problem for MU-MIMO systems to optimality is impractical even for a moderately sized system (e.g., for 20 users) on both the DL [16] and the UL [4], due to the need for exhaustive searches to find the best SS. Previous works have compared CTR_1, BD and CTR_F on the DL by proposing and employing offline *Greedy-Up Search (GUS)* heuristics [3], [6], [9], [13], [14].

To the best of our knowledge, no work has studied the strategies CTR_1, BD and CTR_F and compared their performance within the same UL framework while taking Proportional Fairness (PF) and all RRM processes into account. We propose to adapt GUS used on the DL to the UL given its proven performance on the DL [8] and wide applicability. However, such a search must be done over all the PRBs at once in a TS (there are in general many dozens of PRBs in a TS), which increases the search complexity greatly compared to the DL. Another distinction between the DL and UL appears when considering fairness among users; since the RRM procedure is done over the entire TS at once on the UL, weights for fairness are updated every TS as opposed to every PRB on the DL [4]. Many TSs must be then executed to allow the evolution of the weights on the UL, which means that UL heuristics have to be designed for speed. Indeed, heuristics on the DL had to be run over hundreds of PRBs while on the UL, they will need to be run over thousands.

We are now ready to state our Research Questions (RQs):

RQ 1: How to formulate the UL RRM problem for the BF strategies BD, CTR_1 and CTR_F, considering MCS-based rate adaptation, PF and power management over a TS? Can this problem be solved as such?

RQ 2: If it cannot be solved as such, can we extend, in a computationally efficient way, the GUS heuristic developed for the DL to the UL?

RQ 3: What are the performance gains brought by CTR_F compared to the simpler strategies BD and CTR_1 on the UL for different 3rd Generation Partnership Project (3GPP) scenarios and system parameters? Are those gains scenario-dependent?

RQ 4: How do different power management schemes compare on the UL of a MU-MIMO system with MA users?

To answer these RQs, we make the following contributions:

- **Formulation of the UL RRM problem over a TS:** this formulation applies to the three BF strategies BD, CTR_1 and CTR_F, considering MCS-based rate adaptation, PF and power management. The problem is extremely large and cannot be solved even for small but meaningful systems and hence we need to resort to heuristics.
- **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 ZF BF computation methods developed in [8], [9] and efficient power management.
- **Judicious Computation Technique:** Despite the efforts to make the search efficient by using fast ZF BF and efficient power management, the number of rate computations at each search iteration can be very large as will be seen in Section IV. 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 system 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 Rural Macro (RMa) and Urban Macro (UMa), across different system 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 system 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 system 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 ZF BF strategies.
- **Impact of Simpler Power Management Scheme:** Finally, we consider a simple power management scheme that shares the power equally among all streams of a user over a TS and compare its performance, for each ZF BF strategy, with the more complex power management scheme used so far. We show that BD performance is the most impacted by the simpler scheme (decrease of 35% in performance) relative to the other strategies (decrease of 15% for CTR_1 and CTR_F).

TABLE I: Main notations: subscript u, s means stream s of user u; superscript c (resp. t) refers to PRB c (resp. TS t)

Notation	Description
M_B and M_U	Number of antennas at the BS and at the users
$\mathcal{U}$, $\mathcal{M}_U$ and $\mathcal{C}$	Set of users, stream indices and subchannels in a TS
$v_{u,s}^c$	Binary variable for stream selection
$\mathbf{V}^c$; $\mathcal{V}$	Allocation matrix; Allocation set over the TS
$\mathbf{H}_u^c$; $\mathcal{H}$	Channel matrix; Set of channel matrices over the TS
w_u^t and R_u^t	Weight and moving average per-TS rate for a TS
r_u^t	Rate transmitted in a TS
W	Window of TSs for fairness
P_U	Per-TS power budget per user
$P_{u,s}^c, r_{u,s}^c, E_{u,s}^c$	Power, rate and Effective channel
L; Γ_l and ζ_l	Number of MCSs; SNR and rate for MCS l
$\tau_{u,s}^c$	Power for the highest MCS level
$f(\cdot)$	Function mapping SNR to rate
A and D	Coefficients of the fitted MCS-based rate function

The paper is organized as follows. Section II introduces the system model and presents the related work. Section III describes the UL RRM in detail. Section IV presents the proposed UL GUS heuristic. Finally, Section V contains the numerical results and discussions. Section VI concludes the paper with final remarks and avenues for future research.

Notation: Calligraphic letters, such as $\mathcal{A}$, denote sets, with $|\mathcal{A}|$ indicating the cardinality of set $\mathcal{A}$. Upper case bold letters, such as $\mathbf{M} \in \mathbb{C}^{M\times N}$, indicate a matrix of dimension M by N with complex numbers. Table I presents the main notation used throughout this work.

II. System Model and Related Works

In this section, we present the system model and discuss related works and how our research fills some of the gaps in the literature.

A. System Model

We consider the UL of a single MU-MIMO OFDMA-based cell. OFDMA divides the system bandwidth into subchannels and the time into TSs. The bandwidth is divided into C subchannels, each with a bandwidth of B_C, where $\mathcal{C} = \{1, \ldots, C\}$ is the set of subchannels. A PRB consists of a subchannel $c \in \mathcal{C}$ in a TS t. The BS and users have M_B and M_U antennas, respectively[2]. Under ZF BF with MA users, each user can transmit up to M_U independent data streams in a given PRB, while the BS can receive up to M_B streams in a PRB [17]. The users transmit at a carrier frequency f_c with a per-TS power budget of P_U. We assume no signaling overhead, error-free Channel State Information (CSI), and full buffers at the users so that data is always available for transmission. The channel model is described in Section V. Table I summarizes the notations used throughout this work.

B. Related Works

As discussed previously, the UL RRM procedure should be done over all PRBs of a TS at once due to the power management, which increases the difficulty when compared to the DL (which can be done on a PRB basis). In the following, we discuss works on the UL of MU-MIMO systems.

The most relevant works to ours were conducted in [4], [17] and [18]. In [4], an offline Greedy-Down Search (GDS) heuristic over the TS was proposed for SA users with ZF BF. GDS 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, GDS has a constraint on the number of users in the cell which must not be greater than the number of BS antennas. A GUS heuristic based on ZF with multiple streams per user, where the BF matrices were optimized during the search, was proposed in [17]. 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. [18] 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.

The UL was also addressed in other significant works, which are less relevant to our study but are reported for completeness. For instance, in [19], the sum-power minimization problem with rate constraints was investigated with Maximum Ratio Transmission (MRT) BF for MA users (i.e., each user utilized only its strongest stream) and Successive Interference Cancellation (SIC) at the BS. [20] proposed an unsupervised learning user selection method with Minimum Mean Square Error (MMSE) receivers at the BS, but it did not consider fairness and was designed for SA users. In [21], the scheduling UL problem over multiple PRBs in a TS was addressed with MRT BF for MA users and MMSE receivers at the BS, but the solution design was based on the Shannon rate function and single-stream users. In [22], a grouping-based user scheduling method for PF and SA users with MMSE-based decoding at the BS was proposed. [23] examined the minimization of overall transmit power over multiple PRBs while guaranteeing the fulfillment of each user's Quality-Of-Service (QoS) requirements with MRT BF for MA users and different receivers, such as MMSE and ZF.

Table II summarizes the main characteristics of the papers discussed above. From this discussion, it is clear that no work has evaluated the performance of the ZF BF strategies on the UL of a multi-channel MU-MIMO system with MA users and considering all key RRM processes carefully. Our work intends to answer the research questions above.

III. Uplink Radio Resource Management

In this section, we introduce the UL RRM over an arbitrary TS t, omitting index t and referring to a subchannel as a PRB for ease of notation. Specifically, we discuss each RRM process independently from SS, ZF BF, power management to MCS selection. Also, we formulate and discuss the general *joint* UL RRM problem over the TS. We explain why this joint problem cannot be solved for realistic systems and then

[2]We assume an equal number of antennas per user for simplicity, but this study can be extended to a non-equal setting.

TABLE II: Synthesis of related works. Notation: Shannon-based (SB), User Antenna Setting (UAS), Sum-Power Minimization (SPM).

Work	Rate Adaptation	Subchannels	Criterion	BF Strategy	UAS
[4]	MCS	**Multiple**	PF	CTR_1=BD	SA
[17]	SB	Single	WSR	Similar to CTR_F	**MA**
[18]	MCS	**Multiple**	Minimum Rate	BD	**MA**
[19]	SB	Single	SPM	MRT/SIC	**MA**
[20]	SB	**Multiple**	SR	MMSE	SA
[21]	SB	**Multiple**	PF	MMSE/MRT	**MA**
[22]	SB	**Multiple**	PF	MMSE	SA
[23]	MCS	**Multiple**	SPM/QoS	MRT/MMSE/ZF	**MA**
Ours	MCS	**Multiple**	PF	CTR_1, BD, CTR_F	**MA**

propose heuristics to obtain good feasible solutions using the problem structure.

We start by presenting some important notation. Let $\mathcal{U}$ be the set of users. Let $\mathbf{H}_u^c \in \mathbb{C}^{M_U \times M_B}$ be the channel matrix for user $u \in \mathcal{U}$ in PRB $c \in \mathcal{C}$ and $\mathcal{H} = \{\mathbf{H}_u^c\}_{u\in\mathcal{U},c\in\mathcal{C}}$ be the channel matrices set for all users and PRBs in the TS.

A. Stream Selection

The SS process selects which users and streams will be allowed to transmit in each PRB. On the UL, SS has to be done over all PRBs at once as opposed to sequentially one PRB at a time as in the DL, because power management, which will be discussed below, depends on which PRBs are allocated to a user [4], [15]. The stream selection is reduced to a user selection problem for CTR_1 and BD, since for each scheduled user $u \in \mathcal{U}$ in a PRB $c \in \mathcal{C}$, the BS enables only the strongest stream (resp. all M_U streams) for CTR_1 (resp. BD). Under CTR_F, any combination of streams can be enabled for each user in any PRB (up to M_U streams per user per PRB).

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 [6], [9] for details on stream strength) and let $\mathcal{M}_U = \{1,\ldots,M_U\}$ be the stream indices set. 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. For any ZF BF strategy, another restriction is on the number of scheduled streams per PRB, which must be no greater than M_B. Then, in a given PRB $c \in \mathcal{C}$, we have the following constraints:

- For CTR_1, $v_{u,s}^c = 0, \forall u \in \mathcal{U}$ and $1 < s \leq M_U$;
- For BD, $\sum_{s\in\mathcal{M}_U} v_{u,s}^c = v_{u,1}^c M_U, \forall u \in \mathcal{U}$;
- For CTR_F, $\sum_{s\in\mathcal{M}_U} v_{u,s}^c \leq M_U, \forall u \in \mathcal{U}$;
- For CTR_1, BD and CTR_F, $\sum_{u\in\mathcal{U}}\sum_{s\in\mathcal{M}_U} v_{u,s}^c \leq M_B$.

We define the matrix $\mathbf{V}^c = (v_{u,s}^c), \forall c \in \mathcal{C}, \forall u \in \mathcal{U}, \forall s \in \mathcal{M}_U$, and $\mathcal{V}$ to be the allocation set $\mathcal{V} = \{\mathbf{V}^c\}_{c\in\mathcal{C}}$ over the whole TS. In the following, if $v_{u,s}^c = 1$, we will say that stream $s \in \mathcal{M}_U$ of user $u \in \mathcal{U}$ has been selected (equivalently scheduled) in PRB $c \in \mathcal{C}$. Likewise, we will say that a user u is selected (scheduled) in PRB c if at least one of its streams is selected for transmission (i.e., $\exists s \in \mathcal{M}_U$ such that $v_{u,s}^c = 1$).

B. Zero-Forcing Beamforming

ZF BF enables the transmission of interference-free data streams[3] from multiple users in the same PRB (at most M_B streams) on the UL, where, with MA users, each user may have multiple streams (at most M_U) [17]. Inter-user interference is nullified to the detriment of the co-scheduled users' effective channels. The more correlated the selected users in the same PRB are, the more their effective channels are reduced to achieve interference-free transmission, hence the need for appropriate SS to schedule users with low mutual correlation [16], [24].

Given an allocation set $\mathcal{V}$, ZF BF is performed for each PRB $c \in \mathcal{C}$ independently (and hence can be done in parallel). For a given PRB, it is performed jointly over the selected streams to obtain the effective channels $E_{u,s}^c(\mathbf{V}^c, \{\mathbf{H}_k^c\}_{k\in\mathcal{U}}), \forall u \in \mathcal{U}, \forall s \in \mathcal{M}_U$. We use the convention that $E_{u,s}^c(\mathbf{V}^c, \{\mathbf{H}_k^c\}_{k\in\mathcal{U}}) = 0$ if a stream is not selected, i.e., if $v_{u,s}^c = 0$. The effective channels depend on which streams are selected and on the channel matrices of the users scheduled for transmission in the PRB. For brevity, we will write the effective channels as $E_{u,s}^c(\mathbf{V}^c)$ in the following.

Unfortunately, computing the effective channels for a given PRB, given a set of scheduled streams, is cumbersome. It requires singular value decompositions of matrices, which makes it difficult to incorporate as such in an optimization problem. Several studies have proposed numerical algorithms to efficiently compute the effective channels under ZF BF [6]–[9], [11]. In this work, we employ the approach developed in [8], [9] in our GUS heuristic in Section IV. This method fits the iterative nature of our heuristic since it reuses BF computations from previous iterations to speed up the ZF BF computations in the current iteration, as will be discussed in Section IV. Also, the method can be used for CTR_1, CTR_F and BD. Other ways of computing the effective channels can be found in [6], [7], [11].

C. Power Management

Given an allocation set $\mathcal{V}$ and a BF strategy, we can compute the effective channels $E_{u,s}^c(\mathbf{V}^c)$ for all streams $s \in \mathcal{M}_U$ of all users $u \in \mathcal{U}$ in each PRB $c \in \mathcal{C}$ and then execute power management over the TS to obtain the allocated power $P_{u,s}^c$ for stream s of user u in PRB c, with the convention that $P_{u,s}^c = 0$ if $v_{u,s}^c = 0$. Then, since the product between $P_{u,s}^c$ and $E_{u,s}^c(\mathbf{V}^c)$ is the Signal-to-Noise Ratio (SNR) for stream s of user u in PRB c, a rate function $f(.)$ can be applied to the SNR to obtain the rate $r_{u,s}^c$. Note that there is no intra-user interference due to ZF BF. We now present the rate function.

[3]Under error-free CSI, ZF BF nullifies interference.

D. Rate Function and MCS Selection

In most papers, the rate function is chosen to be based on the Shannon formula [17], [21]. However, in a practical system, there is a set of available MCSs for selection. Each MCS requires a minimum SNR to achieve a fixed rate for a target Block-Error Rate (BLER) [25]. MCS selection is about choosing the highest MCS for each selected stream over a PRB given its SNR. Let L be the number of available MCSs, let ζ_l and Γ_l be the achievable rate and required SNR for MCS $l \in \{1, \ldots, L\}$, where ζ_l and Γ_l are increasing sequences on l, then the MCS-based rate function $f(\cdot)$ can be written as

$$f(SNR) = \begin{cases} 0 & SNR < \Gamma_1; \\ \zeta_l & \Gamma_l \leq SNR < \Gamma_{l+1},\ l \in \{1, L-1\}; \\ \zeta_L & SNR \geq \Gamma_L. \end{cases}$$

Note that $f(\cdot)$ is a piecewise constant function, and is neither concave nor differentiable.

In the following, we formulate the joint UL RRM problem over a TS for our system. This problem is more general than the one in [4] because MA users are considered instead of SA users. Considering MA users makes the problem more complex since each user can have multiple streams (with BD and CTR_F) and in this case, power management includes not only PA but also PD (while only PA is needed for SA users).

E. Problem Formulation

We are now ready to formulate the joint RRM problem on the UL, for a given TS, where fairness is taken into account by using weights w_u's that reflect how much rate has been received by each user u in a window of size W TSs (please see Section IV in [4] for more details). These weights are computed from previous TSs results as will be discussed in Section V. Let $\mathcal{W} = \{w_u\}_{u \in \mathcal{U}}$ be the set of all weights. Given $\mathcal{H}$ and $\mathcal{W}$ as well as a sufficiently large positive constant A, under ZF BF, the UL RRM over a TS can be written as the following Weighted Sum-Rate (WSR) maximization problem:

$$\max_{(P_{u,s}^c),(v_{u,s}^c),(r_{u,s}^c),(E_{u,s}^c(\mathbf{V}^c))} \sum_{u \in \mathcal{U}} \sum_{c \in \mathcal{C}} \sum_{s \in \mathcal{M}_U} w_u r_{u,s}^c \quad (1)$$

$$\sum_{c \in \mathcal{C}} \sum_{s \in \mathcal{M}_U} P_{u,s}^c \leq P_U, \ \forall u \quad (2)$$

$$P_{u,s}^c \geq 0, \ \forall u, c, s \quad (3)$$

$$r_{u,s}^c = f\left(P_{u,s}^c E_{u,s}^c(\mathbf{V}^c)\right) \ \forall u, c, s \quad (4)$$

$$E_{u,s}^c(\mathbf{V}^c) \leq v_{u,s}^c A, \ \forall u, c, s \quad (5)$$

$$v_{u,s}^c \in \{0, 1\}, \ \forall u, c, s \quad (6)$$

$$\sum_{u \in \mathcal{U}} \sum_{s \in \mathcal{M}_U} v_{u,s}^c \leq M_B, \ \forall c \quad (7)$$

$$\text{CTR}_1 : \ v_{u,s}^c = 0, \ \forall u, c, s \neq 1 \quad (8)$$

$$\text{BD} : \sum_{s \in \mathcal{M}_U} v_{u,s}^c = 0 \text{ or } \sum_{s \in \mathcal{M}_U} v_{u,s}^c = M_U, \ \forall u, c \quad (9)$$

$$\text{CTR}_\text{F} : \sum_{s \in \mathcal{M}_U} v_{u,s}^c \leq M_U, \ \forall u, c \quad (10)$$

Recall that $\mathbf{V}^c$ is the matrix $(v_{u,s}^c)$ of dimension $|\mathcal{U}| \times M_U$ representing the stream selection for PRB c. The variables $v_{u,s}^c$'s are binary, which makes the problem an integer one. Constraint (2) is the sum-power constraint per user over the TS, coupling the PRBs in the TS for each user. This constraint is the reason why the UL RRM has to be done over the entire TS as opposed to per PRB on the DL. Equation (4) maps the SNR of a stream to a rate with the rate function $f(\cdot)$. Recall that the MCS-based function $f(\cdot)$ is not concave, which makes the problem non-convex. Implicit in this formulation are the computations of the effective channels, which, as discussed earlier, cannot be expressed in a simple algebraic form. This also makes the problem non-convex. Constraint (5) guarantees that $E_{u,j}^c(\mathbf{V}^c) = 0$ if stream s of user u is not selected in PRB c. Constraint (7) originates from the ZF BF restriction on the number of scheduled streams per PRB, which must be no greater than M_B. Constraint (8) is only used for CTR_1 and ensures that only the strongest stream of each selected user is enabled, while Constraint (9) is introduced for BD, guaranteeing that all the streams are enabled for each selected user. For CTR_F, Constraint (10) limits the sum of $v_{u,s}^c$ for each user and PRB to be less than or equal to M_U.

This problem is a large integer and highly non-convex problem that needs computations of effective channels that can only be done using a numerical approach. In the next section, we discuss why obtaining feasible solutions for such a problem is impractical even for small systems.

F. Why a straightforward solution is not applicable

We note that, given a stream selection, i.e., a matrix $\mathbf{V}^c$ for all $c \in \mathcal{C}$, then the effective channels can be computed on each PRB, and then the problem becomes a set of $|\mathcal{U}|$ much simpler **per-user** power management (joint power allocation and power distribution) problem (JPM in short), i.e.,

$$\textbf{JPM}(u, \{E_{u,s}^c(\mathbf{V}^c)\}_{c,s}) : \max_{P_{u,s}^c, r_{u,s}^c} \sum_{c \in \mathcal{C}} \sum_{s \in \mathcal{M}_U} r_{u,s}^c \quad (11)$$

$$\text{s.t.} \sum_{c \in \mathcal{C}} \sum_{s \in \mathcal{M}_U} P_{u,s}^c \leq P_U \quad (12)$$

$$P_{u,s}^c \geq 0, \ \forall c, s \quad (13)$$

$$r_{u,s}^c = f\left(P_{u,s}^c \times E_{u,s}^c(\mathbf{V}^c)\right), \ \forall c, s \quad (14)$$

JPM is no longer an integer problem. It is non-convex because of the MCS-based rate function $f(.)$. Nevertheless, it can easily be convexified as done in [18], [26] by fitting the rate function with a concave function (either an upper-bound or a tight approximation).

Hence, solving the original problem could in theory be done by doing an exhaustive search on the set of matrices $\{\mathbf{V}^c\}_{c \in \mathcal{C}}$ and by computing, for each such set, the effective channels for each stream of each user in each PRB and then solving the convexified version of the per-user JPM problem to finally compute the WSR over the TS. Comparing the upper-bound with the feasible solution obtained with a tight approximation of the rate function would allow us to validate the goodness of the feasible solution. Clearly, this is too time-consuming, especially since the computations of the effective channels are cumbersome and exhaustive searches on $\{\mathbf{V}^c\}_{c \in \mathcal{C}}$ are impractical even for small systems since there are too many different stream selections over the multiple PRBs of a TS. Instead of an exhaustive search, we will do a directed search, which will still be cumbersome but will be tractable and will allow us to obtain a feasible solution to the problem.

IV. Uplink Greedy-Up Search Heuristic

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 CTR_1, 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 CTR_1 (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 CTR_1, only the strongest stream of each user can be selected.

Efficiency is at the core of the heuristic design and implementation, as it must be tractable over many TSs to allow the weights to evolve for performance evaluations under fairness. 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 ZF BF 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 ZF BF and power computation techniques as well as a novel iterative rate-reusing method.

A. Heuristic Design and Implementation

We use CTR_F in the following to present GUS, since it is the most intricate case. We will discuss GUS for BD and CTR_1 in Section IV-C. In the case of CTR_F, the heuristic executes an unconstrained greedy-up search on the stream sets per PRB over the TS. The heuristic starts by initializing $v^b_{k,j} \leftarrow 0, \forall k \in \mathcal{U}, \forall b \in \mathcal{C}, \forall j \in \mathcal{M}_U$, 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 $\hat{v}^b_{k,j}$ be the best stream allocation over the TS yielding WSR_{max}. They are initialized as $WSR_{max} \leftarrow 0$ and $\hat{v}^b_{k,j} \leftarrow 0, \forall k \in \mathcal{U}, \forall b \in \mathcal{C}, \forall j \in \mathcal{M}_U$. 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 $v^c_{u,s} = 0$ (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 $WSR^c_{u,s}$ be the WSR over the TS when candidate (u, c, s) is assessed. For the sake of computing $WSR^c_{u,s}$, assume stream s of user u has been selected temporarily in PRB c (i.e., $v^c_{u,s} \leftarrow 1$).

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

- **ZF BF Computation:** To improve the search computational efficiency, we reuse effective channel values (when possible) and employ fast ZF BF computation methods [8], [9].
- **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. $\exists j \in \mathcal{M}_\mathcal{U}$ s.t. $v^c_{k,j} = 1$) and obtain the power allocated for each of its selected stream $P^b_{k,s}, \forall j \in \mathcal{M}_U, \forall b \in \mathcal{C}$ 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.
  To execute power management, we could solve JPM defined in Section III-F for each of those users. However, due to the nature of the MCS-based function $f(\cdot)$, the problem is non-convex, preventing the use of efficient algorithms such as Water-Filling (WF) [27]. Instead, we follow a similar approach to [18], [26] 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, JPM over the TS can be rewritten as, for a user $k \in \mathcal{U}$,

$$\overline{\textbf{JPM}}(k, \{E^b_{k,j}(\mathbf{V}^b)\}_{b,j}) :$$

$$\max_{P^b_{k,j}} \sum_{b\in\mathcal{C}} \sum_{j\in\mathcal{M}_U} A\log(1 + DE^b_{k,j}(\mathbf{V}^b)P^b_{k,j}) \tag{15}$$

$$\text{s.t.} \sum_{b\in\mathcal{C}} \sum_{j\in\mathcal{M}_U} P^b_{k,j} \le P_U; \tag{16}$$

$$0 \le P^b_{k,j} \le \tau^b_{k,j}, \forall b \in \mathcal{C}, \forall j \in \mathcal{M}_U, \tag{17}$$

  which can be solved efficiently by WF using Algorithm 6 from [27]. The power is upper bounded in Constraint (17) to avoid power waste since the MCS-based rate function is upper bounded by ζ_L for any SNR greater or equal to Γ_L (the required SNR for L, the highest order MCS). Specifically, $\tau^b_{k,j}$ is the power necessary for stream j to achieve an SNR equal to Γ_L, since allowing a larger SNR would be wasteful.
  For each user $k \in \mathcal{U}$ that has a stream selected in PRB c, we compute the rates $r^b_{k,j}$ 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** $WSR^c_{u,s}$: Finally, the WSR when adding candidate (u,c,s) is

$$WSR^c_{u,s} = \sum_{k\in\mathcal{U}}\sum_{b\in\mathcal{C}}\sum_{j\in\mathcal{M}_U} w_k r^b_{k,j},$$

where we reuse the rates $r^b_{k,j}$ 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 $\overline{\text{JPM}}$ 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., $(u^*, c^*, s^*) = \text{argmax}_{(u\in\mathcal{U}, c\in\mathcal{C}, s\in\mathcal{M}_U)} WSR^c_{u,s}$. Note that the weights w_u are used to compute the weighted sum rates. If $WSR^{c^*}_{u^*,s^*} > WSR_{max}$, then the candidate (u^*, c^*, s^*) is definitely selected (i.e., $v^{c^*}_{u^*,s^*} \leftarrow 1$) and we update $WSR_{max} \leftarrow WSR^{c^*}_{u^*,s^*}$ and $\hat{v}^{c^*}_{u^*,s^*} \leftarrow 1$, and go to the next search iteration. However, if $WSR^{c^*}_{u^*,s^*} < WSR_{max}$ but $WSR_{max}/WSR^{c^*}_{u^*,s^*} \leq \beta$ (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 $\hat{v}^{c^*}_{u^*,s^*}$. Finally, if $WSR_{max}/WSR^{c^*}_{u^*,s^*} > \beta$, 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., $\sum_{k\in\mathcal{U}}\sum_{j\in\mathcal{M}_U} v^b_{k,j} = M_B, \ \forall b \in \mathcal{C}$, or when all streams of all users have been selected in all PRBs, i.e., $\sum_{j\in\mathcal{M}_U} v^b_{k,j} = M_U, \forall k \in \mathcal{U}, \forall b \in \mathcal{C}$. Upon stopping, the search returns the selected streams over the TS corresponding to the largest WSR_{max} observed during the search (given by $\hat{v}^b_{k,j}$), 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 $\overline{\text{JPM}}$) and using the solution as an input to the greedy MCS-aware heuristic from Algorithm 1 in [15]. 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 [15]. We do remove all zero-rate streams at once, redo the ZF BF 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, $r_k = \sum_{b\in\mathcal{C}}\sum_{j\in\mathcal{M}_U} r^b_{k,j}$. This ends the GUS heuristic for CTR_F.

B. Speeding Up the Computations: Rate-Reusing Method

To compute $WSR^c_{u,s}$ for a candidate (u,c,s), we execute ZF BF computations using a fast computation method, then the rate computations are done with WF-based PAs and PDs (i.e., by solving $\overline{\text{JPM}}$) 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 ZF BF 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 in Appendix A 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 V, while not affecting the performance as we are only being judicious in what to recompute.

C. GUS for CTR₁ and BD

GUS for both CTR_1 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 CTR_1, 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 CTR_1 are not defined by specific streams; a candidate for BD and CTR_1 is (u,c) defined by an unselected user $u \in \mathcal{U}$ in PRB $c \in \mathcal{C}$.

V. Numerical Results

The numerical results are presented in this section. We start by defining the experimental setup, presenting how the weights for PF are updated between TSs and the procedure used to execute many TSs, which is necessary to allow the weights to evolve.

A. Experimental Setup

We consider the 3GPP-based RMa and UMa scenarios. Following the 3GPP documentation [28], we model the system as a sector of a three-sector hexagonal cell, where the users are fixed and uniformly distributed within the sector. The cell radii for RMa and UMa are obtained from Tables 7.2-1 and 7.2-3 in [28] and are around 1000 m and 300 m, respectively. Similar to [13], [18], we adopt a 3GPP-based tapped-delay-line channel model containing 8 taps based on Table 7.7.2-5 from [28] where the first tap can be either Line-Of-Sight (LOS) or Non-Line-Of-Sight (NLOS), and the remaining taps are NLOS. The channel model has all key practical components, such as path loss and shadowing, small-scale fading, and transmit and receive correlation based on an exponential model (see [18] and references therein for the exact channel formulation). The channel parameters are taken from Tables 7.2-1, 7.2-3, 7.4.1-1, 7.4.2-1, 7.5-6 in [28].

The 5G MCS mapping employed in this work is shown in Table III and is taken from [29]. As discussed in Section IV, we fit the MCS-based rate function with a function of the form $A\log(1+D\gamma)$, with coefficients $D = 0.5191$ and $A = 1.389$.

To obtain robust results and conclusions, we have to consider the randomness in the channel and user positions within the cell. For that, we introduce the concept of realization. We define a *realization* $\Omega(\Delta)$ over a horizon of Δ TSs as a set of fixed users $\mathcal{U}$ with their channel matrices for all ΔC PRBs. Within a realization, the varying channel element for a user is the small-scale fading, which changes independently across distinct reporting blocks consisting of C_B subchannels and T_B TSs. A reporting block represents a set of PRBs for which the CSI of each user is given and constant. We will evaluate many realizations and for each realization, many TSs will be executed to allow for the proper evolution of the channels and the weights for PF.

Regarding the fairness weights used to compute WSR for a TS, in line with the literature [4], let R_u^t be the average per-TS rate seen by user $u \in \mathcal{U}$ in the past W TSs at the beginning of TS t where W is a parameter to decide on which horizon fairness needs to be enforced. The weight of user u at the beginning of TS t is simply $w_u^t = 1/R_u^t$. At the end of TS t, we have

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

where r_u^t is the rate transmitted by user u in TS t and the weights are updated accordingly [4], [16].

For each realization $\Omega(\Delta)$, given that we use PF, the metric of interest is the Geometric Mean (GM) of the rates seen by the users in the Δ TSs [4], i.e.,

$$GM(\Delta) = (\prod_{u\in\mathcal{U}} \sum_{t=1}^{\Delta} r_u^t)^{1/|\mathcal{U}|}$$

Hence, the computations for each realization run for Δ TSs.

The values of the other parameters for the evaluations are shown in Table IV. The system bandwidth is set to 30 MHz, which yields 78 subchannels per TS [30], each with a bandwidth of 360 kHz following the New Radio (NR) numerology 1 [31]. We set the GUS stopping parameter β to 1.05 (higher values for β were tested without performance improvement) and set the window $W = 6$.

The evaluations are conducted using GUS equipped with the rate-reusing method, except when indicated otherwise. 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 ZF BF strategies CTR_F, BD and CTR_1 in both 3GPP scenarios for different system parameters. We then study the impact of power management and end this section by presenting the run time reduction in our GUS heuristic when using the proposed rate-reusing method.

B. ZF BF Strategies Evaluation

Fig. 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. 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. 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. 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. 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. 3 and 4 when $P_U = 1$ mW. Compared to the results for RMa (see Fig. 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

TABLE III: 5G MCS mapping [29].

MCS l	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
ζ_l (rate in 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
Γ_l (SNR in 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

TABLE IV: Parameters (RB: reporting block, #: number).

Parameter	**Value**
Maximum # of TSs per realization Δ	66
Window length in TSs (W)	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
Antennas at the BS (M_B); at UE (M_U)	100, 64; 4, 2, 1
Power budget per user in mW (P_U)	50 (RMa); 1 and 5 (UMa)
GUS stopping parameter (β)	1.05
Channel correlation (BS and users)	0.4
Adjacent antenna distance (BS and users)	$0.5 \times 3 \times 10^8/f_c$
Noise power density	-174 dBm/Hz
Noise figure	9 dB
Coefficients for MCS fitting A; D	1.389; 0.5191

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 ZF BF 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. 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. 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.

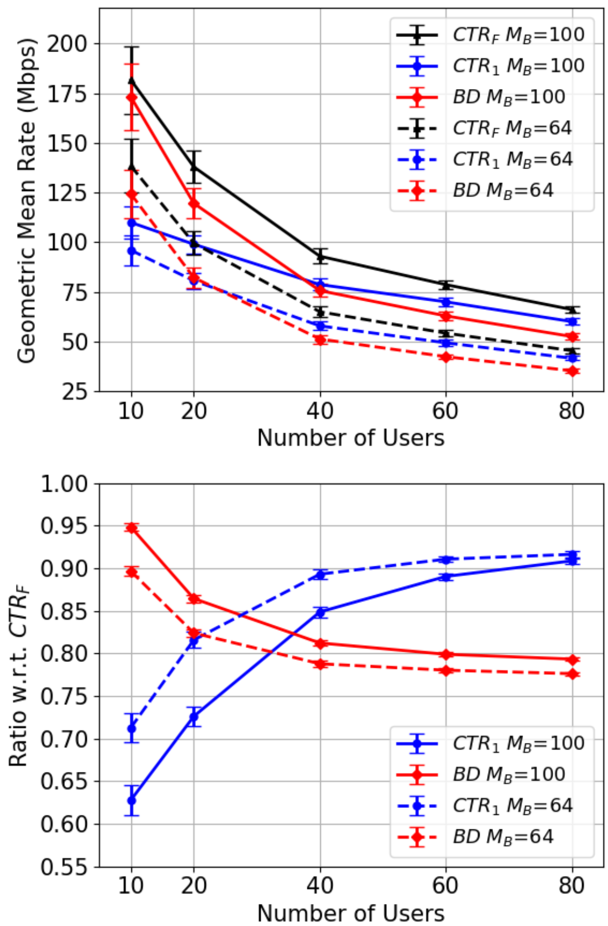

Fig. 1: RMa scenario: $M_B = 100$ and $M_B = 64$ antennas at the BS, $M_U = 4$ antennas at the users, $P_U = 50$ mW. Top figure: GM rates in Mbps per TS vs. number of users for CTR_F, BD and CTR_1. Bottom figure: ratios in the GM rates between BD and CTR_F, and between CTR_1 and CTR_F.

We also evaluate the performance of the ZF BF strategies for different per-user power budgets, P_U. Fig. 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. 6 (bottom figure) shows the corresponding ratios for BD and CTR_1 relative to CTR_F. The GM rates for all ZF BF 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.

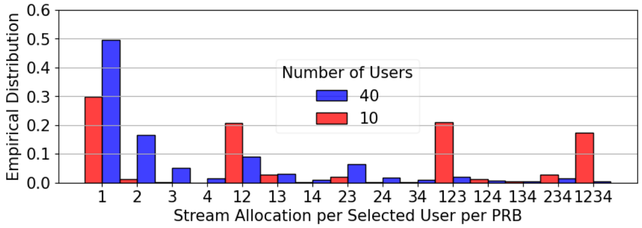

Fig. 2: RMa scenario: empirical distribution of the stream allocation per scheduled user per PRB given by CTR_F for $|\mathcal{U}| \in \{10, 40\}$ users with $M_B = 64$ BS antennas, $M_U = 4$ antennas at the users, and $P_U = 50$ mW.

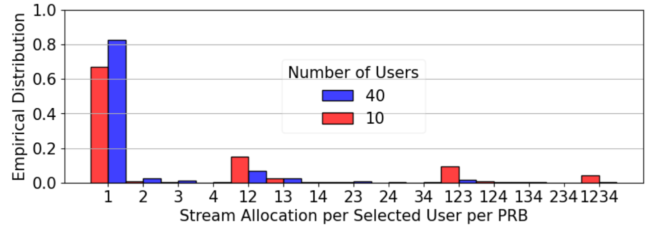

Fig. 4: UMa scenario: empirical distribution of the stream allocation per selected user per PRB given by CTR_F for $|\mathcal{U}| \in \{10, 40\}$ users with $M_B = 64$ BS antennas, $M_U = 4$ antennas at the users, and $P_U = 1$ mW.

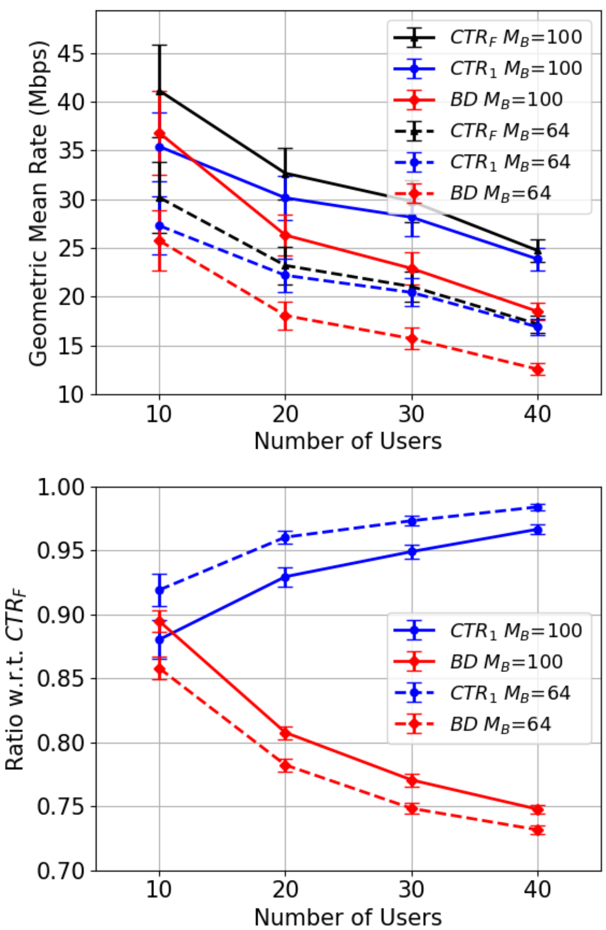

Fig. 3: UMa scenario: $M_B = 100$ and $M_B = 64$ antennas at the BS, $M_U = 4$ antennas at the users, and $P_U = 1$ mW. Top figure: GM rates in Mbps per TS vs. number of users for CTR_F, BD and CTR_1. Bottom figure: ratios in the GM rates between BD and CTR_F, and between CTR_1 and CTR_F.

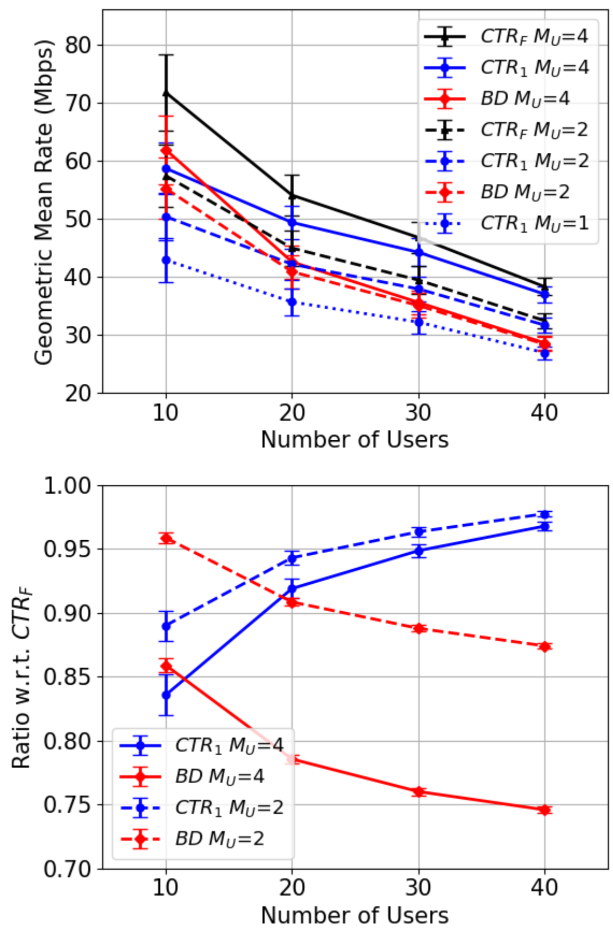

Fig. 5: UMa scenario: $M_B = 64$ antennas at the BS, $M_U = 4$, $M_U = 2$ and $M_U = 1$ antennas at the users, $P_U = 5$ mW. Top figure: GM rates in Mbps per TS vs. number of users for CTR_F, BD and CTR_1. Bottom figure: ratios in the GM rates between BD and CTR_F, and between CTR_1 and CTR_F.

C. Power Management Schemes Evaluation

All the results presented above were obtained with the complex power management defined in Section IV as TPM. Next, we study how a simpler power management scheme, called Equal Power Management (EPM), performs for each ZF BF strategy. This scheme allocates, for each user, the power budget equally among its selected streams over the TS. This scheme is used during and at the end of the greedy-up search explained in Section IV.

Fig. 7 (top figure) presents, for the UMa scenario, the GM rates per TS for the three ZF strategies vs. number of users for TPM and EPM with $M_B = 64$ BS antennas, $M_U = 4$ antennas at the users, and a per-user power budget of $P_U = 5$ mW. Fig. 7 (bottom figure) shows the corresponding ratio between EPM and TPM for each ZF BF strategy. The performance

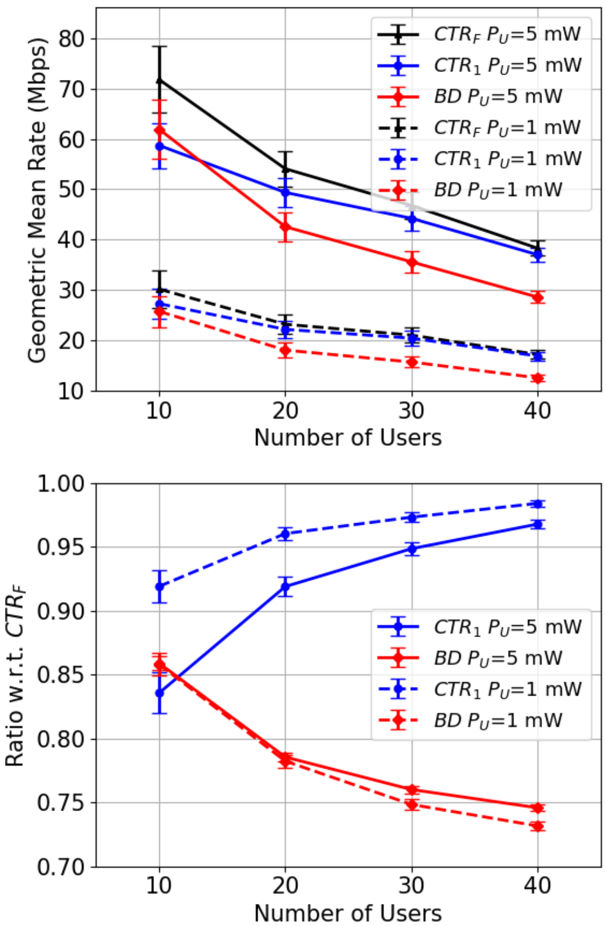

Fig. 6: UMa scenario: $M_B = 64$ antennas at the BS, $M_U = 4$ antennas at the users, $P_U = 5$ mW and $P_U = 1$ mW. Top figure: GM rates in Mbps per TS vs. number of users for CTR_F, BD and CTR_1. Bottom figure: ratios in the GM rates between BD and CTR_F, and between CTR_1 and CTR_F.

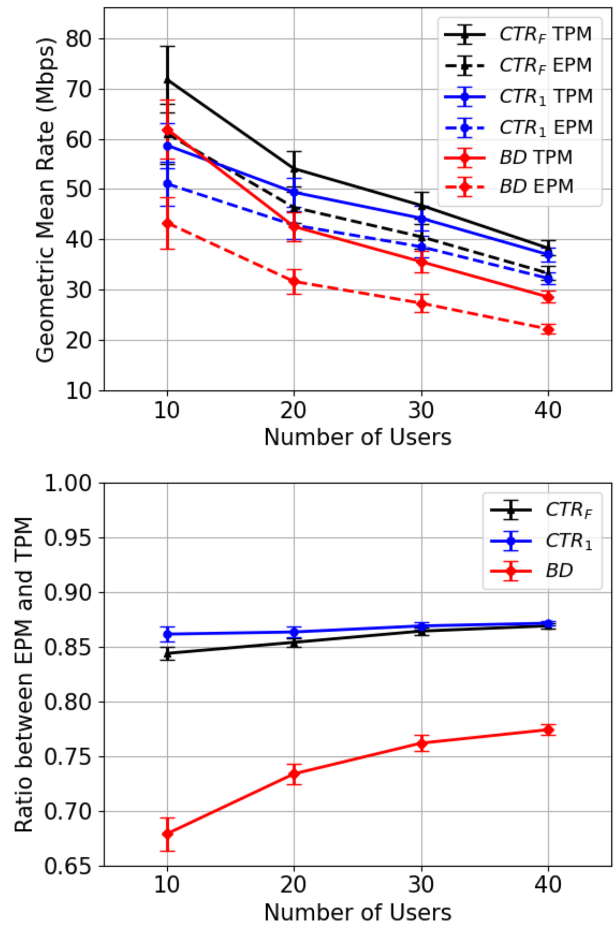

Fig. 7: UMa scenario: $M_B = 64$ antennas at the BS, $M_U = 4$ antennas at the users, $P_U = 5$ mW. Top figure: GM rates in Mbps per TS vs. number of users for CTR_F, BD and CTR_1 for the power management schemes TPM and EPM. Bottom figure: ratio in the GM rates between EPM and TPM for each ZF BF strategy.

drop for BD using the simpler EPM is larger compared to the one for CTR_1 or CTR_F. For example, the ratio between EPM and TPM for BD is just above 75% for 40 users but can reach nearly as low as 65%. In contrast, the ratios for CTR_1 and CTR_F are approximately 85% for all numbers of users. An explanation for that is that, with BD, all M_U streams are enabled per scheduled users per PRB, and these streams have distinct strengths. TPM distributes the power (i.e., PD) taking into account the streams' gains, while that does not occur with EPM. CTR_1 and CTR_F do not suffer from this issue because in the former, only the strongest stream is enabled, while the latter allows a flexible stream allocation. In conclusion, in case a simpler power management scheme, such as EPM, is required for real-time settings, the outcomes suggest that CTR_F or CTR_1 are better suited but the performance cost is not negligible.

D. 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 IV-B w.r.t. GUS without it as proposed in Section IV-A. 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 ZF BF 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 ZF BF strategies, across different 3GPP-based scenarios and system parameters, highlighting the importance of the proposed rate-reusing method for run time reduction.

VI. Conclusion

This study is the first of its kind. Very few papers have addressed the uplink of an MU-MIMO/OFDMA single cell with multi-antenna users and multiple channels. Of those, none have considered different ZF BF strategies and all RRM processes. We have evaluated the performance that can be achieved on the uplink of an MU-MIMO/OFDMA single cell with multi-antenna users for three ZF BF strategies: BD, where all possible data streams are enabled per scheduled user; CTR_1, in which only the strongest streams of the scheduled users are used; CTR_F, which allows a flexible stream allocation per user. CTR_F offers potentially better performance due to its flexibility, albeit at the price of higher complexity compared to other strategies. This study was based on a system with proportional fairness, practical MCSs for 3GPP-based scenarios (RMa and UMa). The UL RRM must be done over the entire TS at once due to the UL power management and many TSs must be executed to take fairness into account. Hence, there is a need for efficient tools to evaluate the performance of such a system. We have proposed a heuristic based on a greedy-up search for stream-sets that operates over the TS and considers an MCSs-based rate function, all key RRM processes and PF. To make the heuristic tractable for this study, we equipped it with efficient ZF BF and power computation methods and proposed a novel rate-reusing method where rates computed in previous search iterations are reused whenever possible.

The results and discussions presented in Section V provide valuable information that can guide the design of real-time UL RRM heuristics, which is an important avenue for future research. Specifically, the results showed that CTR_F outperformed BD and CTR_1, but in some cases, BD or CTR_1 could replace CTR_F in real-time settings given their comparable performance and lower complexity. For instance, for RMa, either BD or CTR_1 could be an alternative to CTR_F depending on the number of users, while for UMa, CTR_1 performed similarly to CTR_F across all system parameters. We also showed that the numbers of antennas at the BS and at the users, and the power budget per user considerably impacted the performance of the ZF strategies. Finally, we compared the performance of each ZF BF strategy with a complex power management scheme to that with a simpler scheme (based on equal power management), where we saw that the BD performance was the most impacted by the simpler scheme relative to the other ZF BF strategies. Another avenue for future work is to consider a multi-cell setting, where the interference from neighbouring cells must be taken into account.

References

[1] T. L. Marzetta, “Massive MIMO: An introduction,” *Bell Labs Technical Journal*, vol. 20, pp. 11–22, 2015.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[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] 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.

[8] 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.

[9] 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.

[10] 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.

[11] L.-N. Tran, M. Juntti, and E.-K. Hong, “On the precoder design for block diagonalized MIMO broadcast channels,” *IEEE Communications Letters*, vol. 16, no. 8, pp. 1165–1168, 2012.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] Y. J. Zhang and K. Letaief, “An efficient resource-allocation scheme for spatial multiuser access in MIMO/OFDM systems,” *IEEE Transactions on Communications*, vol. 53, no. 1, pp. 107–116, 2005.

[24] X. Yi and E. K. S. Au, “User scheduling for heterogeneous multiuser MIMO systems: A subspace viewpoint,” *IEEE Transactions on Vehicular Technology*, vol. 60, no. 8, pp. 4004–4013, 2011.

[25] 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.

[26] 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.
[27] 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.
[28] 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
[29] ——, "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
[30] ——, "5G; NR; User equipment (UE) radio transmission and reception; Part 1: Range 1 Standalone," 3rd Generation Partnership Project, TS 38.101-1, Version 16.2.0, 2020. [Online]. Available: https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3283
[31] ——, "5G; NR; Physical channels and modulation," 3rd Generation Partnership Project, TS 38.211, Version 16.2.0, 2020. [Online]. Available: https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3213

Appendix A
Rate-reusing Rules for BD

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 IV-A, 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 IV-A.
- **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 $\overline{\text{JPM}}$) 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 $\overline{\text{JPM}}$) 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.