

Radio Resource Management of Hybrid Beamforming Systems

by

Yuan Quan

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Electrical and Computer Engineering

Waterloo, Ontario, Canada, 2025

© Yuan Quan 2025

Examining Committee Membership

The following served on the Examining Committee for this thesis. The decision of the Examining Committee is by majority vote.

External Examiner: Raviraj Adve
Professor
Department of Electrical and Computer Engineering
University of Toronto

Supervisor: Catherine Rosenberg
Professor
Department of Electrical and Computer Engineering
University of Waterloo

Internal Member: Patrick Mitran
Professor
Department of Electrical and Computer Engineering
University of Waterloo

Internal Member: George Shaker
Adjunct Professor
Department of Electrical and Computer Engineering
University of Waterloo

Internal-External Member: Henry Wolkowicz
Professor
Department of Combinatorics and Optimization
University of Waterloo

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 multi-author papers, in which I am the first author. These papers include:

- Y. Quan, S. Shahsavari, and C. Rosenberg, “Planning and Operation of Millimeter-Wave Downlink Systems With Hybrid Beamforming,” in *IEEE Transactions on Wireless Communications*, vol. 23, no. 12, pp. 18214-18227, Dec.
- Y. Quan, H. Rahman, and C. Rosenberg, “Online Resource Management for the Uplink of Wideband Hybrid Beamforming System,” in *ICC 2025 - IEEE International Conference on Communications*, Montreal, QC, Canada, 2025, pp. 6492-6497.
- Y. Quan, H. Rahman, and C. Rosenberg, “Radio Resource Management for Hybrid Beamforming Uplink Systems”, submitted to *IEEE Transactions on Wireless Communications*.

Abstract

Beamforming and **MIMO (Multiple-input Multiple-output)** are two of the key technologies that enable high capacity and spectrum-efficient communications in 5G and beyond systems. Codebook-based **HBF (Hybrid Beamforming)** wherein **ABF (Analog Beamforming)** vectors are chosen from pre-designed codebooks and, optionally, **DBF (Digital Beamforming)** can be performed on the selected ABF vectors, results in lower hardware cost, training overheads, and complexity for real-time operations over **FDBF (Full Digital Beamforming)**.

We study **RRM (Radio Resource Management)** for the **DL (Downlink)** and the **UL (Uplink)** of codebook-based HBF systems assuming proportional fairness. Our study focuses on the practical multi-channel case without assuming that the number of **RF (Radio Frequency)** chains at the **BS (Base Station)**, K , is larger than the number of **UE (User Equipment)** in the cell, U . Indeed, in the highly practical yet underexplored case where $K < U$, there is a need for analog beam selection (since in most cases, not all beams with UEs can be selected together), which creates a per-time-slot constraint since the set of selected beams cannot be changed within a time slot. This constraint creates a coupling between channels in a time slot and necessitates a per-time-slot RRM solution for both DL and UL, as opposed to the per-channel approaches commonly adopted in prior work.

On the DL, we address this constraint by formulating and solving a per-time-slot joint RRM optimization problem that includes beam selection, user selection, **PD (Power Distribution)**, **MCS (Modulation and Coding Scheme)** selection and, optionally, DBF. This planning study helps us to understand the impact of the different parameters and options (e.g., non-equal or equal power distribution, DBF or not) and provides us with target performance to validate the online RRM algorithms that we design for equal PD and non-equal PD, respectively, with optional **ZF (Zero-Forcing)** DBF. The numerical results obtained for a mmWave small cell, show that our online RRM algorithms are efficient and have acceptable runtimes.

On the UL, each UE has its own power budget, in contrast to the DL where the transmit power of the base station is shared among UEs. The power budget of a UE needs to be allocated per time slot to its assigned subchannels which are not known a priori. This results in the interdependence between **PA (Power Allocation)** and user selection in addition to the inter-channel coupling caused by beam selection, making UL RRM more challenging and inherently different compared to DL RRM. We only consider the case where ZF-DBF is enabled. We formulate a per-time-slot joint RRM optimization problem, which includes beam selection, user selection, PA, ZF-DBF, and MCS selection. This problem can only be solved for small systems. In order to obtain results for larger systems (more

users and RF chains), we propose an offline heuristic that reduces the runtime necessary to obtain meaningful results while achieving performance close to the joint optimization. This offline heuristic allows us to obtain engineering insights on the impact of different system parameters as well as a target performance that we use to validate the low-complexity online heuristic that we propose (the numerical results were obtained for a mmWave small cell).

Finally, we adapt and extend our solutions to a [FWA \(Fixed Wireless Access\)](#) setting in the mid-band and in the mmWave band where the fairness criterion is max-min so that [MBR \(Minimum Bit Rate\)](#) targets can be offered to all homes. We analyze the impact of codebook size and type in this setting, where the UEs are stationary home and hence beam alignment can be simplified making large codebooks feasible. Engineering insights into the system parameters and bandwidth requirements necessary to deliver MBR targets for DL and UL are provided for both bands.

Acknowledgements

I would first like to thank my supervisor, Catherine Rosenberg, for giving me the opportunity to start my PhD in 2021. Along this journey, she taught me how to find and solve an engineering problem systematically and rigorously, as well as how to effectively communicate and present ideas to others. Her commitment to research quality has set the standard for my career. These insights and principles will be essential assets in my future.

I want to thank my committee members at the University of Waterloo, Patrick Mitran, George Shaker and Henry Wolkowicz, for their time, support and feedback since my comprehensive examination. I also sincerely thank my external committee member, Raviraj Adve, for his recognition and for taking the time to review this thesis.

Thank you to my colleague Haseen Rahman Cheriyaandilakath for the help in the UL study and also for always bringing a chill vibe to our lab. I would like to extend my thanks to João Paulo Pinto Galdino Marques for his support in the FWA study and for being such a supportive friend. Thank you to my lab mates: Maryam, Abdallah, Ahmed, Kiril, Mohammed, and Samir. Special thanks to our lab graduates, Yigit Ozcan and Andrew Lappalainen, for their professional advice and invaluable support in my career development.

I am deeply grateful to my parents for their unlimited support. None of these would have been possible without their love and sacrifices. I also want to thank my in-laws, Zhiyong and Ying, for their contributions to our family during my PhD journey.

Finally, I want to express my deepest gratitude to my love, Ziyin, and my five-year-old boy, GuaGua. Ziyin has always been there with me, from the hardest time to the happiest time. Her smile is the motivation of my life. Thank you, GuaGua, for inviting me to play with you when I was stressed. I love you. You are the gems of my life.

Dedication

To those who have encouraged me along the way.

Table of Contents

Examining Committee	ii
Author's Declaration	iii
Statement of Contributions	iv
Abstract	v
Acknowledgements	vii
Dedication	viii
List of Figures	xiii
List of Tables	xvii
Abbreviations	xviii
1 Introduction	1
1.1 HBF System	2
1.2 Multi-user Data Transmission	4
1.3 Radio Resource Management	5
1.3.1 Radio Resource Management for the Downlink	5

1.3.2	Radio Resource Management for the Uplink	7
1.4	Fixed Wireless Access	8
1.5	Contributions	9
1.5.1	Downlink Contributions	10
1.5.2	Uplink Contributions	11
1.5.3	FWA Contributions	12
1.6	Thesis Outline	13
2	Background and Literature Review	15
2.1	Analog, Fully Digital, and Hybrid BF Systems	15
2.2	Downlink Literature Review	17
2.3	Uplink Literature Review	19
2.4	Fixed Wireless Access Literature Review	21
3	System Model	23
3.1	Operational Processes	24
3.1.1	Beam Alignment	25
3.1.2	Effective Channel Estimation	25
3.2	Digital Beamforming	26
3.3	Rate Function	28
3.3.1	Logarithmic Approximation	29
3.3.2	Linear Approximation	29
3.4	Fairness	29
4	Downlink Study	32
4.1	Planning (Offline RRM Study)	32
4.1.1	Problem Formulation	33
4.1.2	Solution Method for EPD	36

4.1.3	Solution Methods for OPD	36
4.1.4	Extension to Multiple Matched BS Beams Per UE	39
4.1.5	Numerical Evaluations for Offline RRM Study	39
4.2	Operation (Online RRM Study)	47
4.2.1	Low-complexity Benchmark	48
4.2.2	Proposed Heuristic Solutions	49
4.2.3	Numerical Evaluations for Online RRM Study	53
4.3	Conclusion	53
5	Uplink Study	57
5.1	Joint RRM Optimization	57
5.1.1	Problem Formulation	58
5.1.2	Upper Bound of Joint RRM Optimization Problem	61
5.1.3	Feasible Solution to Joint RRM Optimization Problem	63
5.2	Heuristic Solutions to RRM	64
5.2.1	Offline Heuristic	64
5.2.2	Online Heuristic	66
5.3	Numerical Evaluations	69
5.3.1	System Parameters	69
5.3.2	Channel Model and Parameters	70
5.3.3	Evaluation Criterion	71
5.3.4	Numerical Evaluations for Joint RRM Optimization	71
5.3.5	Numerical Evaluation of the Offline Heuristic	72
5.3.6	Numerical Evaluations of the Online Heuristic	74
5.4	Conclusion	77

6	Fixed Wireless Access Study	79
6.1	System Model	80
6.1.1	Positioned-based Beam Matching	80
6.1.2	Max-min Fairness	81
6.2	RRM Solutions	82
6.3	Numerical Evaluations	82
6.3.1	System Parameters	83
6.3.2	Codebooks	84
6.3.3	Channel Model and Parameters	87
6.3.4	Numerical Evaluations at 3.5 GHz	88
6.3.5	Numerical Evaluations at 28 GHz	95
6.3.6	Interference Mitigation Process	101
6.4	Conclusion	106
7	Conclusion	107
7.1	Future Research Directions	109
	References	111

List of Figures

1.1	HBF architecture at the BS side and UE side (for illustration purpose $K = 4$ and $U = 6$)	3
1.2	System operational processes	5
2.1	OFDMA structure	16
2.2	ABF, FDBF and HBF systems	16
3.1	Data structure and operational procedures	24
3.2	An example of piece-wise constant rate function $f(\gamma)$ and two convex approximations $\tilde{f}(\gamma)$, a tight one and an upper bound	30
4.1	Quality of feasible solution with $N_b = 128$, $N_u = 16$, $B_b = 32$, $B_u = 4$	43
4.2	System performance vs. K for different cases EPD/OPD and ZF-DBF/N-DBF with $N_b = 128$, $N_u = 16$, $B_b = 32$, $B_u = 4$	43
4.3	Impact of ConB on system performance considering OPD, ZF-DBF with $N_b = 128$, $N_u = 16$, $B_b = 32$ and $B_u = 4$	44
4.4	System performance vs. K for different ABF codebook sizes considering OPD, ZF-DBF with $U = 10$, $N_b = 128$, and $N_u = 16$	44
4.5	(a) Impact of the number of preferred beams on system performance considering OPD, ZF-DBF with $N_b = 128$, $N_u = 16$, $B_b = 32$ and $B_u = 4$	46
4.6	WF performance for ZF-DBF with $N_b = 128$, $N_u = 16$, $B_b = 32$, $B_u = 4$	46
4.7	Building blocks of our online heuristics for a time slot (BeamS: beam selection, US: user selection, Alg.: Algorithm, N.A.: not available)	50

4.8	Performance of the proposed heuristics and benchmarks as a function of U with $K = 8$, $N_b = 128$, $N_u = 16$, $B_b = 32$ and $B_u = 4$	54
4.9	Average runtime of the proposed heuristics and benchmarks as a function of U with $K = 8$, $N_b = 128$, $N_u = 16$, $B_b = 32$ and $B_u = 4$	54
4.10	Impact of M on system performance with the proposed heuristics as a function of K with $U = 20$, $N_b = 128$, $N_u = 16$, $B_b = 32$ and $B_u = 4$	55
5.1	Performance comparison between PBO, PBS and the solution in [24] as a function of U	68
5.2	Quality of the feasible solution for $U = 6, 8, 10$ and $M = 1$ (UB: upper bound, FS: feasible solution).	72
5.3	Impact of the number, M , of matched BS beams per UE on the performance of the feasible solution for $U = 6, 8, 10$	73
5.4	Performance of the offline heuristic as a function of K for $U = 8, 10$ and $M = 1$	73
5.5	Impact of the number of matched BS beams per UE on offline heuristic performance for $U = 10$ and 30	75
5.6	Performance of the online heuristic as a function of K for $U = 10$ and 30 , and $M = 1$	75
5.7	Performance of the online heuristic as a function of U for $K = 2, 8, 16$ and $M = 1, 2$	76
6.1	Beam pattern of Song codebook for $N_b = 64$ and $B_b = 16$	85
6.2	Beam pattern of DFT codebook for $N_b = 64$ and $B_b = 50$	85
6.3	Beam pattern of CDFT codebook for $N_b = 64$, $B_b = 50$ and sidelobe controlled at -25 db.	86
6.4	Midband, DL: Impact of B_b on system performance with Song codebooks for $N_b = 64$, $N_u = 4$ and $BW = 60$ MHz.	89
6.5	Midband, UL: Impact of B_b on system performance with Song codebooks for $N_b = 64$, $N_u = 4$ and $BW = 60$ MHz.	89
6.6	Midband, DL: Impact of B_b on system performance with DFT codebook for $N_b = 64$, $N_u = 4$ and $BW = 60$ MHz.	90

6.7	Midband, UL: Impact of B_b on system performance with DFT codebook for $N_b = 64$, $N_u = 4$ and $BW = 60$ MHz.	90
6.8	Midband, DL: Impact of B_b on system performance with Chebychev-window codebook for $N_b = 64$, $N_u = 4$ and $BW = 60$ MHz.	91
6.9	Midband, UL: Impact of B_b on system performance with Chebychev-window codebook for $N_b = 64$, $N_u = 4$ and $BW = 60$ MHz.	92
6.10	Midband, DL: Impact of types of codebooks on system performance for $N_b = 64$, $N_u = 4$ and $BW = 60$ MHz.	92
6.11	Midband, UL: Impact of types of codebooks on system performance for $N_b = 64$, $N_u = 4$ and $BW = 60$ MHz.	93
6.12	Midband: Impact of antenna array placement for $N_b = 64$, $N_u = 4$, $B_b = 52$, DFT codebook and $BW = 60$ MHz on the DL and UL.	93
6.13	Midband: System performance comparison between codebook-based HBF and fully DBF architectures for $N_b = 64$, $N_u = 4$ and $BW = 40$ MHz on the DL and UL.	94
6.14	Midband, DL: Impact of bandwidth and number of antennas on system performance as a function of K	95
6.15	Midband, UL: Impact of bandwidth and number of antennas on system performance as a function of K	96
6.16	MmWave, DL: Impact of B_b on system performance with Song codebook for $N_b = 128$, $N_u = 8$ and $BW = 200$ MHz.	97
6.17	MmWave, UL: Impact of B_b on system performance with Song codebook for $N_b = 128$, $N_u = 8$ and $BW = 200$ MHz.	97
6.18	MmWave, DL: Impact of B_b on system performance with DFT codebook for $N_b = 128$, $N_u = 8$ and $BW = 200$ MHz.	98
6.19	MmWave, UL: Impact of B_b on system performance with DFT codebook for $N_b = 128$, $N_u = 8$ and $BW = 200$ MHz.	98
6.20	MmWave, DL: Impact of B_b on system performance with Chebychev-window codebook for $N_b = 128$, $N_u = 8$ and $BW = 200$ MHz.	99
6.21	MmWave, UL: Impact of B_b on system performance with Chebychev-window codebook for $N_b = 128$, $N_u = 8$ and $BW = 200$ MHz.	99

6.22	MmWave, DL: Impact of types of codebooks on system performance for $N_b = 128$, $N_u = 8$ and $BW = 200$ MHz.	100
6.23	MmWave, UL: Impact of types of codebooks on system performance for $N_b = 128$, $N_u = 8$ and $BW = 200$ MHz.	101
6.24	MmWave, DL: Impact of bandwidth and number of antennas on system performance as a function of K	102
6.25	MmWave, UL: Impact of bandwidth and number of antennas on system performance as a function of K	102
6.26	MmWave, DL: Impact of bandwidth and number of antennas on system performance as a function of K	103
6.27	MmWave, UL: Impact of bandwidth and number of antennas on system performance as a function of K	103
6.28	Midband, DL: Impact of interference mitigation on system performance as a function of K with $N_b = 64$ and $BW = 60$ MHz.	104
6.29	Midband, UL: Impact of interference mitigation on system performance as a function of K with $N_b = 64$ and $BW = 60$ MHz.	104
6.30	MmWave, DL: Impact of interference mitigation on system performance as a function of K with $N_b = 128$ and $BW = 200$ MHz.	105
6.31	MmWave, UL: Impact of interference mitigation on system performance as a function of K with $N_b = 128$ and $BW = 200$ MHz.	105

List of Tables

2.1	Main system characteristics in similar studies from DL literature (OPT: optimized, ZF: zero-forcing, WF: water-filling, ConB: beam selection constraint, CH: frequency sunchannels)	18
2.2	Main system characteristics in the HBF UL references (Eff.: effective, CH: channel, OPT/O: optimized, WF: water-filling, E: equal, Pre-CBs: predefined codebooks, FN: fairness, NI: non-iterative solution).	19
2.3	Main system characteristics in the FWA references (Stat.: statistical, Eff.: effective, CH: channel.	21
3.1	a_i and b_i values for the linear approximation	31
4.1	Simulation Parameters	42
5.1	Complexity of the online heuristic solution (LA BSel: load-aware beam selection)	77

Abbreviations

ABF Analog Beamforming [v](#), [1](#)

AD Analog-to-digital [17](#)

AP Analog Processor [2](#)

BA Beam Alignment [4](#)

BF Beamforming [1](#)

BS Base Station [v](#), [2](#)

CE Channel Estimation [3](#)

DA Digital-to-analog [17](#)

DBF Digital Beamforming [v](#), [1](#)

DL Downlink [v](#)

DP Digital Processor [2](#)

EPA Equal Power Allocation [5](#)

FDBF Full Digital Beamforming [v](#), [1](#)

FDD Frequency Division Duplex [17](#)

FWA Fixed Wireless Access [vi](#), [8](#), [9](#)

HBF Hybrid Beamforming [v](#), [1](#)

MBR Minimum Bit Rate [vi](#), [8](#)

MCS Modulation and Coding Scheme [v](#), [6](#)

MIMO Multiple-input Multiple-output [v](#), [1](#)

mmWave Millimeter Wave [1](#)

N-DBF Non-digital Beamforming [6](#)

OFDMA Orthogonal Frequency Division Multiple Access [2](#), [15](#)

PA Power Allocation [v](#), [2](#), [5](#)

PD Power Distribution [v](#), [2](#)

PF Proportional Fairness [19](#)

PRB Physical Resource Block [1](#)

PRBs Physical Resource Blocks [2](#)

PS Phase Shifter [15](#)

PSs Phase Shifters [2](#)

RF Radio Frequency [v](#)

RRM Radio Resource Management [v](#), [2](#), [3](#)

SINR Signal-to-interference-and-noise Ratio [6](#)

SNR Signal-to-noise Ratio [7](#)

TDD Time Division Duplex [2](#), [23](#)

TS Time Slot [5](#)

UE User Equipment [v](#)

UEs User Equipments [2](#)

UL Uplink [v](#)

WF Water-filling [18](#)

ZF Zero-Forcing [v](#), [6](#)

Chapter 1

Introduction

5G and emerging 6G systems aim to support ultra-high data rates and massive connectivity through access to broader bandwidths. Mid-band spectrum, spanning 1–6 GHz, offers a favorable balance between coverage and throughput, making it well-suited for wide-area deployment. However, it is over-subscribed. As a supplement, mmWave (Millimeter Wave) bands (24–100 GHz) provide abundant bandwidth but mmWave propagation is significantly impaired by high path loss and susceptibility to blockage and shadowing [1].

Large-scale antenna systems, also known as MIMO (Multiple-input Multiple-output) systems, play an important role in both bands. MIMO systems enable spatial multiplexing, allowing the transmission of multiple data streams in different spatial directions within the same PRB (Physical Resource Block)¹ to improve spectrum efficiency. Combined with BF (Beamforming) technology, they can counteract path loss and shadowing.

Different system architectures can be used for MIMO transmission and BF. FDBF (Full Digital Beamforming) provides the maximum flexibility and performance, but its deployment is costly, especially in mmWave, since it requires one radio frequency (RF) chain per antenna [2,3]. To mitigate the deployment cost and power consumption of FDBF, an HBF (Hybrid Beamforming) architecture can be used, where the BF processing is divided into per-time-slot ABF (Analog Beamforming) and per-channel DBF (Digital Beamforming) [4]. HBF utilizes significantly fewer RF chains to support multiple concurrent data streams for serving multiple UEs in the same channel. However, the ABF component imposes practical constraints that are not present in FDBF, making radio resource management (RRM) in

¹Within a time slot t , assigning a channel c is equivalent to assigning the PRB (c, t) and we will use channel and PRB interchangeably.

HBF systems challenging. More details on ABF, FDBF and HBF hardware architectures are provided in Section 2.1.

1.1 HBF System

In this thesis, we consider the RRM (Radio Resource Management) of a TDD (Time Division Duplex) and OFDMA (Orthogonal Frequency Division Multiple Access) based single cell using HBF. As will be seen in the thesis, the case of a single cell is very challenging to study due, in part, to the intra-cell interference (and in another part to the beam selection constraint). The extension to a multi-cell environment entails also taking care of the inter-cell interference. It is left for further study. The bandwidth and time are organized into PRBs (Physical Resource Blocks), each corresponding to a channel in a time slot. Some time slots are allocated to the DL and others to the UL. The cell comprises one BS (Base Station) and U UEs (User Equipments). The BS has a fully-connected HBF structure with K RF chains, each connected to each of the N_b antennas through N_b phase shifters. Each UE is equipped with N_u antennas and a single RF chain connected to the antennas via N_u phase shifters, as depicted in Fig. 1.1. Unlike most existing work in the literature that assumes $K > U$, we do not make any assumption on the relative value of K with respect to U . We also consider the multi-channel case, which is the more challenging but overlooked practical scenario.

Specifically, focusing on the BS side, the DP (Digital Processor) is connected to the AP (Analog Processor) by K RF chains as shown in Fig. 1.1. The baseband DP is used for PA (Power Allocation), PD (Power Distribution), optional DBF, and other necessary signal processing such as modulation and demodulation. The AP is used to implement ABF in the RF domain using K groups of N_b PSs (Phase Shifters). The phases of each group of PSs form a $N_b \times 1$ vector, which is the ABF vector corresponding to an analog beam of the BS. Focusing on the UE side, DBF is not available because there is only one RF chain. The RF chain will be connected to all N_u antennas by AP consisting of N_u PSs. The phases of the N_u PSs form a $N_u \times 1$ vector, which is the ABF vector corresponding to the analog beam of the UE. We consider ideal PSs in this thesis, i.e., the phase is assumed to vary continuously without quantization effects caused by limited-resolution PSs. In HBF systems, BF is divided into baseband DBF followed by RF domain ABF. DBF can vary in different channels in a time slot because it is realized in the digital domain. ABF must be fixed for the entire time slot as it is realized in RF equipment which is commonly used by all channels.

One approach to designing ABF vectors (i.e. the vector of the phase shifts applied at

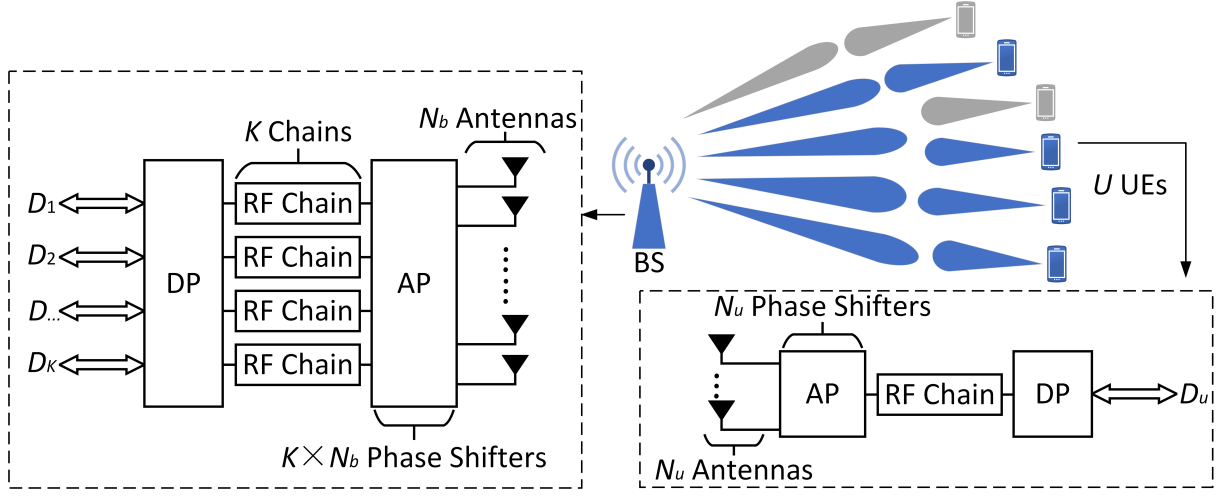


Figure 1.1: HBF architecture at the BS side and UE side (for illustration purpose $K = 4$ and $U = 6$)

the BS and UEs) is through optimization methods [4]. However, besides high computational burden, such optimizations require per-UE full CSI (i.e., MIMO channel matrix) which is typically difficult to obtain in practice [2]. In real-time operations, CE (Channel Estimation), BF and RRM (Radio Resource Management) must be executed within a fraction of a second. We use a more practical solution is to choose ABF vectors from pre-designed ABF codebooks. Inside an ABF codebook, there are B_b (resp. B_u) $N_b \times 1$ (resp. $N_u \times 1$) codewords corresponding to the phase shift configuration of a beam [5, 6]. Typically $B_b > K$ and $B_u > 1$. This approach reduces BF complexity and simplifies CE by reducing the effective channel dimensions from matrix to scalar, as discussed in [4, 7]. Finally, the DBF matrix can be generated separately using existing methods, such as zero-forcing [8], once the ABF vectors and user set have been determined. Note that RRM for a HBF single cell is very complex especially without ZF-DBF since in that case, the intra-cell interference makes the problem nonconvex. We have been able to study the case with and without ZF-DBF on the DL but not on the UL where we only studied the case with ZF-DBF because power allocation cannot be decoupled from the RRM problem and solved in advance in the case with interference on the UL as is possible on the DL.

1.2 Multi-user Data Transmission

HBF supports multi-user transmission by enabling at most K simultaneous data streams to different UEs (no more than one stream per UE since each UE has a single RF chain) in the same channel. Each stream will use one receiving beam at the BS (BS beam) and one transmitting beam at the UE (UE beam). Each BS (resp. UE) beam uses one RF chain and N_b (resp. N_u) phase shifters, whose phase shifts decide the pattern of the beam. Since the phase shifters are used in all channels, the beams cannot be changed among channels in a time slot, e.g., the data streams in different channels need to share the same K BS beams and the same UE beam. This means that beam selection is needed per time slot both at the BS and the UE. Then, a UE can transmit on multiple channels in a time slot (possibly on different selected BS beams in different PRBs if $M > 1$). After ABF, per-channel DBF is performed to mitigate the interference between streams.

In order to use a pair of BS and UE beams for transmission, CE has to be performed by the BS. A priori, the BS needs to estimate each channel for each beam pair, which would create a high signaling overhead. To reduce CE overhead, the number of pairs on which CE will be performed is limited, using a less frequent process than CE which is done every time slot, namely [BA \(Beam Alignment\)](#). Via BA, each UE selects a few “matched” BS and UE beams. Typically, each UE selects one BS beam and one UE beam. One of the objectives of this thesis is to understand how much allowing $M > 1$ matched BS beams per UE during BA (we assume one matched UE beam per UE) could improve the performance (the improvement should be significant enough to make the increase in signaling overhead worth it). We introduce more details on BA and CE in [Chapter 3](#).

Although beam selection is fixed across channels within a time slot, different UEs can be assigned to a BS beam in different channels as long as they have been matched to that beam. Hence, user selection is performed on a channel basis during RRM, given the selected beams (i.e., beam selection and user selection are coupled).

A key parameter for multi-user transmission performance is K , the number of RF chains at BS. In practice, K is often much smaller than the number of UEs (U) due to hardware limitations. However, this practical case has not been explored, particularly under multi-channel scenarios, in existing DL studies (e.g., [\[8–11\]](#)) and UL studies (e.g., [\[12–17\]](#)), which have focused on the case where $U \leq K$ or $K \leq U$ with a single channel ([\[8\]](#) and [\[11\]](#)). We will see that the case where $K < U$ is more challenging to study. Our formulations and solutions apply to both cases.

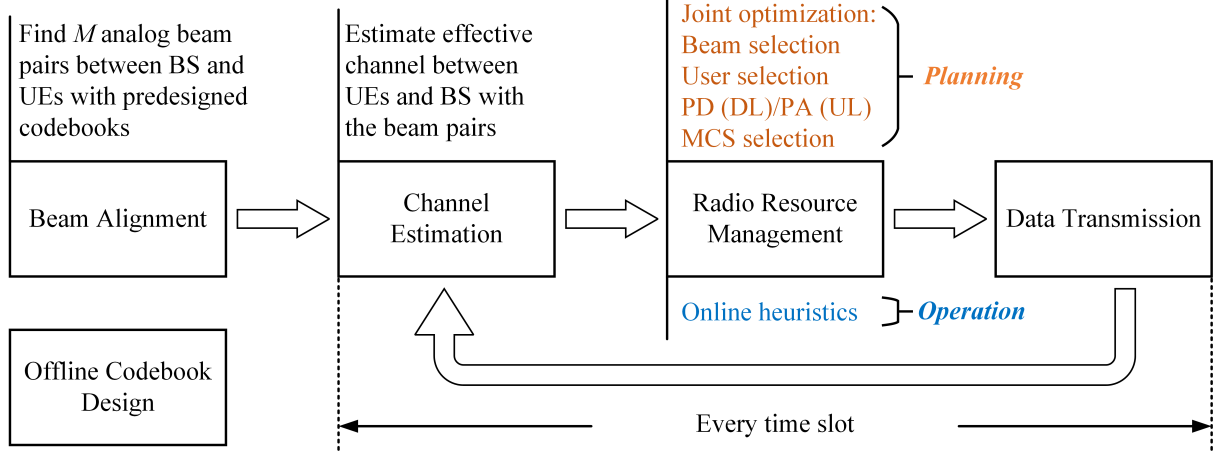


Figure 1.2: System operational processes

1.3 Radio Resource Management

This thesis focuses on the RRM for multi-user transmission of the HBF system described above, given that BA and CE have been performed. DL and UL are studied separately due to their fundamentally different transmission mechanisms. While the RRM in both cases is time slot-based, it operates under distinct sets of constraints. The following sections introduce the RRM procedures for DL and UL, respectively.

1.3.1 Radio Resource Management for the Downlink

On the DL, RRM comprises the following processes:

- **PA (Power Allocation):** The process that allocates the BS transmit power to frequency channels in a given TS (Time Slot). We consider EPA (Equal Power Allocation) among the frequency channels.
- **Per time slot beam selection:** When $U \leq K$, each UE can have its preferred matched BS beam (the one offering the highest average channel gain over PRBs in the time slot) selected all the time and hence there is no need for beam selection. When $K < U$, *per time slot beam selection* is necessary whenever the cardinality of the set containing the preferred matched BS beams for each UE is larger than K (which will happen often if the users are not clumped together). In this case, a subset of K beams must be selected

from the set of all matched BS beams and must remain fixed for all the PRBs of the time slot. Note that existing research either assumes $U \leq K$ to avoid beam selection (e.g., [9, 10]) or considers a single-channel setting and performs RRM including beam selection on a per-channel basis (e.g., [8, 11]). Neither of these approaches is applicable to our system, which must handle the practical multi-channel case where $U \geq K$ and perform beam selection on a per-time-slot basis.

- *Per PRB user selection*: Given a selected beam set, per-PRB user selection has to be performed, for all selected BS beams that have been matched with multiple UEs. At most one UE will be selected from the UEs matched with this beam in each PRB. This is illustrated in Fig. 1.1 where the blue beams and UEs are active and the gray beam and UEs are inactive. This creates an intrinsic coupling between beam selection and user selection.
- *Per PRB DBF*: To enhance the performance, one can apply DBF to the data streams of the selected UEs in a given PRB. We consider **ZF (Zero-Forcing)**-DBF as well as scenarios with **N-DBF (Non-digital Beamforming)** where only ABF is present.
- *Per PRB PD*: The way that the power allocated to a PRB is distributed among the selected UEs for that PRB. There are various approaches for PD, from simply sharing the power equally among all selected UEs in that PRB to distributing power to maximize a system utility function. We consider both scenarios in this thesis.
- ***MCS (Modulation and Coding Scheme)***: The outcome of the previous steps is an estimate of **SINR (Signal-to-interference-and-noise Ratio)** per selected UE (per PRB) which is used to select an appropriate MCS.

A rate function is required to perform some of these processes, in particular for power distribution. Typically, papers use a Shannon-based rate function which has some nice properties but is not practical. A practical rate function is MCS-based and using it makes the RRM problem more complex but more realistic. Even if PD is performed with Shannon-based rate function for simplicity in online heuristics, it is important to do user selection based on rate estimation considering MCSs and to take care of cases where the SINR is too low to deliver any positive rate. Evaluating the impact of using the MCS-based rate function on the DL is one of the research goals of this thesis.

1.3.2 Radio Resource Management for the Uplink

RRM on the UL inherently differs from RRM on the DL. On the DL, by assuming EPA to all PRBs in a time slot, the BS can distribute the power per PRB to UEs. On the UL, the power budget, that belongs to a UE cannot be allocated to PRBs within a time slot before knowing on which PRBs the UE will be transmitting, which means that power management and user selection are inter-dependent in a time slot. Consequently, designing RRM solutions tailored to the unique constraints of the UL are essential.

- *Per time slot beam selection*: The definition follows DL per time slot beam selection but unique solution is needed to address the RRM difference on DL and UL. Note that in UL literature (e.g., [12–17]), $K < U$ has not been considered and beam selection is circumvented.
- *Per PRB user selection*: The definition follows DL per time slot beam selection but unique solution is needed.
- *Per user power allocation (PA)*: RRM on the UL has to be performed on a time slot basis, not only because of beam selection but also for *power allocation* where each UE allocates its power budget to its assigned PRBs within the time slot. PA is inherently coupled with user selection, as a UE needs to know the assigned channels before allocating power to them. The interdependence of beam selection, user selection, and PA presents a significant and unique challenge when studying the RRM for the uplink of the system described above.
- *Per PRB digital BF*: To mitigate inter-beam interference, we apply zero-forcing (ZF) combining at the BS to the data streams of the selected UEs.
- *Modulation and coding scheme (MCS) selection*: Given beam selection, user selection, PA and DBF, the **SNR (Signal-to-noise Ratio)** (there is no interference because of ZF-DBF) can be estimated for each selected stream in each PRB and the appropriate MCS can finally be selected for data transmission.

As on the DL, MCS-based rate function allows each UE to allocate its transmit power across channels more precisely compared to the Shannon-based rate function in offline RRM, and it avoids selecting channels where the SINR values are too low to deliver any positive rate in online RRM. Evaluating the impact of the MCS-based rate function on the UL is one of the research goals of this thesis.

The operational processes, including BA, CE and DBF, are common for DL and UL, as will be discussed in Chapter 3. In conclusion, DL and UL differ primarily on how power is handled during RRM. On DL, the BS sends signals to UEs, so the power budget is shared among different UEs in a channel. While on UL, the UEs send signals to the BS, so each UE has its own power budget to be shared among different channels where the UE is scheduled. Correspondingly, the RRM problems and solutions vary, necessitating separate discussions on DL and UL.

1.4 Fixed Wireless Access

Wired technologies, such as fiber communication, are widely deployed to deliver high-speed internet access to households. However, the high deployment cost and the sparse distribution of homes in rural areas significantly limit their feasibility in such regions. As a result, a connectivity gap, commonly referred to as the digital divide [18], persists between urban and rural areas.

With the availability of bandwidth and a large number of antennas for increased BF gain, 5G and beyond cellular networks can offer rural households broadband Internet access, using FWA (Fixed Wireless Access) [19, 20]. In the following, we call the offering of broadband access to households via 5G FWA, FWA in short. FWA connects a household by establishing a wireless link between a BS and a fixed UE comprising an antenna array often installed on the roof. FWA offers a cost advantage over fiber networks, especially in rural areas, as it can leverage existing cellular infrastructure and available spectrum resources with affordable equipment and ease of installation. However, it also brings new challenges to cellular networks.

First, FWA requires per-household MBR (Minimum Bit Rate) for both the DL and the UL to provide a typical broadband access service. For example, MBR requirements of 50 Mbps on the DL and 10 Mbps on the UL might need to be enforced for all households [21, 22]. Guaranteeing minimum rates poses significant challenges in a wireless environment. Typical wireless services are best-effort.

Second, FWA networks must serve homes located possibly quite far from the BS, posing additional challenges, especially in mmWave bands. While prior work (e.g., [23]) showed a single-user FWA connection over 7 km at 28 GHz, the multi-user performance over long distances remains an open question.

Third, achieving high beamforming gain is critical for serving remote homes. To this end, codebooks with finer angular resolution are required to better align beams with the

homes and maximize signal strength. However, such fine-grained codebooks inherently increase BA overhead, which is a problem in general but might not be a problem in the FWA case, since UEs are fixed. Indeed, we can adopt a location-based beam matching approach, which eliminates the need for BA procedures. Hence, in this context, analyzing the impact of the codebook size is important since there is no trade-off to worry about BA overheads.

Understanding the impact of key parameters on the performance of FWA networks is critical. In this thesis, we focus on the performance analysis of codebook-based hybrid beamforming (HBF) systems for FWA in a rural area, considering key factors such as the number of antennas at the BS, the codebook type and size, the carrier frequency, and the available bandwidth.

1.5 Contributions

Our ultimate research goals for the downlink and the uplink are twofold: 1) to analyze the impact on system performance of different system parameters such as the number of RF chains (K), the number M of BS beams matched per UE, the type of codebooks and the number of beams in the codebook, as well as of different power management and DBF strategies; 2) to develop low-complexity online RRM solutions that performs well. To achieve these two objectives, we need to first develop engineering insights via offline studies that will also give us the target performance. These insights then guide the design of our online RRM heuristics, ensuring that they are simple while aiming to approach the offline targets.

We can adapt our DL and UL solutions to investigate if an HBF system can deliver MBRs on the DL and UL in a 5G [FWA \(Fixed Wireless Access\)](#) setting in the mid-band and in the mmWave band. For both bands, we study the impact of the parameters (e.g., K , the number of RF chains, and N_b , the number of antennas) and of the bandwidth, as well as of the codebook type and size on system performance. In the FWA scenario, UEs are at home with antennas at fixed locations, rendering BA overheads negligible. This allows for greater flexibility in selecting the type and size of the codebook.

We are now ready to state our contributions. As the DL and UL RRM problems and solutions are inherently distinct, as discussed in the previous section, we present the contributions separately for the DL and UL, followed by the FWA contributions.

1.5.1 Downlink Contributions

The principal contributions of our DL study include:

- We assess the performance of the uplink of the HBF system described above. In contrast to existing literature, our study considers a setting in which the number of RF chains at the BS may be smaller than the number of UEs, and multiple frequency channels are in use. Consequently, not all preferred BS beams can be selected at the same time, and the selected beam set cannot be changed from one channel to another within a time slot. To address this constraint, we formulate a joint optimization problem, which includes beam selection, user selection, optional ZF-DBF, PD and MCS selection, on a per-time slot basis rather than a per-channel basis. Our analysis shows that neglecting the beam-set constraint and solving the problem on a per-channel basis can overestimate the performance by 20%.
- We conduct a comprehensive planning study in which we propose offline solution methods to the formulated optimization problems for equal PD (EPD) and optimized PD (OPD) cases. Furthermore, we consider two scenarios for each case: without DBF (N-DBF) and with ZF-DBF. For the EPD case, we obtain the optimal solution in both scenarios. For the OPD case, we solve the problem quasi-optimally in the ZF-DBF scenario and obtain locally optimal solutions in the N-DBF scenario.
- We extend our formulations and solutions to the case where multiple BS beams are matched for each UE ($M > 1$) during BA.
- We provide engineering insights based on the numerical evaluations obtained from our planning study in a mmWave setting. For instance, we show that ZF-DBF can improve the performance by 32% compared to N-DBF. Moreover, we find that OPD can provide 22% performance increase compared to EPD. Additionally, our study reveals that matching each UE with two BS beams does not provide a large performance gain if the number of RF chains is small. Besides, we study the impact of the number of RF chains and the size of the ABF codebooks on the performance.
- We design low-complexity online heuristic RRM schemes for the OPD case with and without ZF-DBF for the case of one beam pair per UE (the case of two pairs is left for further study). Our heuristics can provide good performances compared with the performance benchmarks obtained in the planning study. The numerical results show that the performance of our best online heuristic is at least 92.3% of the best offline benchmark performance.

- We compare the power distribution using MCS-based rate function with that using Shannon-based rate function (i.e., water-filling). The results show that employing the MCS-based rate function can improve the performance by up to 10%. Although we use WF for the online heuristics due to its simplicity, considering MCSs is still important as it prevents selecting UEs whose SINR values are too low to achieve a positive rate, which improves the performance by at least 16%.

1.5.2 Uplink Contributions

As explained in Section 1.3.2, the UL RRM presents unique challenges due to the interdependency between power allocation and user selection. As a result, the solutions and insights developed for the DL do not apply to the UL. Through independent investigation, the main contributions of our UL study are summarized below.

- We assess the performance of the uplink of the HBF system described above that operates on multiple channels and aims at providing proportional fairness (PF) to its UEs. The number of RF chains, K , can take any value ≥ 1 , and each UE matches with $M \geq 1$ BS analog beams during beam alignment. To do so, we formulate a joint RRM optimization problem, over a time slot, that includes beam selection, user selection, PA, DBF, and MCS selection.
- This joint optimization problem is a large mixed-integer nonlinear programming problem. We apply relaxation and approximation techniques to transform it into an upper-bounding linear programming (LP) problem. We then show how to derive a quasi-optimal feasible solution from that LP problem.
- Obtaining this quasi-optimal feasible solution is computationally intensive due to the large size of the problem and hence it can only be done for small systems (i.e., $K \leq 10$ and $U \leq 10$). To enable system planning with a larger number of UEs and RF chains, we propose an offline heuristic based on four steps, the two main steps being Step 1 and Step 3. Step 1 optimizes user selection/PA performed per beam *in isolation* (while ignoring the other beams), considering all users that have matched with this beam. Step 3 is a “load-aware” beam selection strategy where the K beams with the largest loads are selected. We define the load of a beam as the weighted sum rate (WSR) obtained by performing user selection/PA in isolation. The weights in the WSR are there to provide proportional fairness. There are two other steps for that heuristic. Step 2 is necessary if $M > 1$, since a user might be scheduled in the same PRB on multiple beams, which is not possible. We propose a simple way to remove these conflicts and

modify results from Step 1 correspondingly. Given beam selection and user selection, Step 4 performs DBF per PRB over all selected beams for the UEs selected in that PRB and redo an optimized PA. This heuristic can achieve at least 93% of the quasi-optimal performance while reducing the computation time to obtain meaningful results (for different numbers of RF chains and UEs) from months to hours. Our formulations and solutions are independent of the frequency band on which the system operates.

- We obtain insights into the impact of system parameters from the offline optimization and the heuristic results in a mmWave setting. We analyze the impact of the number M of BS beams matched per UE during BA, and of the number of RF chains at BS (K). Results indicate that, using $M = 1$ and a number of RF chains at the BS much smaller than the number of antennas, strikes a balance between performance, complexity and cost, serving as a guideline for deployment strategies and parameters.
- Finally, our online heuristic is inspired by the offline one where we simplify Steps 1 and 4. We propose a per-beam scheduler (PBS) to replace Step 1, i.e., the per-beam user selection and PA optimization used in the offline heuristic. The proposed PBS improves upon the state-of-the-art solution for a single cell scheduler proposed in [24] by at least 22%. Finally, in Step 4, water-filling is used for PA. This online heuristic achieves at least 92% of the performance of the offline heuristic in a mmWave setting. The performance gap arises because the online heuristic allocates power based on Shannon-based rate function (i.e., water-filling), whereas the offline heuristic optimizes power allocation considering the MCS-based rate function. Although Shannon-based rate function is good for power allocation, we find that in the online heuristic, it is necessary to use an MCS-based rate function, as it prevents a UE from selecting channels whose SNR values are too low to achieve a positive rate. We see an improvement of 20% at most compared with using Shannon-based rate function.

1.5.3 FWA Contributions

We adapt our DL and UL RRM solutions to a FWA setting in both the mid-band and mmWave band. Specifically, our solutions are designed to maximize the weighted sum rate of all UEs (see Section 3.4). For the PF objective, we assign each UE u a weight of $\frac{1}{R_u}$, where R_u denotes the average rate seen by UE u in a window of W time slots. In the case of FWA, the objective is max-min fairness. Hence, we adopt a more aggressive weighting of $\frac{1}{R_u^\alpha}$, following the approach in [25, 26] with α large. In both cases, the same solution framework remains applicable. With our RRM solutions, we can analyze the deployment requirements necessary for HBF systems to meet MBR targets on both the DL and UL.

Most importantly, we can further analyze the impact of codebook size and type in HBF systems because BA overhead is no longer a problem with fixed-location UEs. Specifically, We evaluate the performance of FWA using realistic home placement data and practical system configurations based on codebook-based HBF systems. Our results show that the MBR requirements on both the DL (50 Mbps) and UL (10 Mbps) can be satisfied in both mid-band (3.5 GHz) and mmWave (28 GHz) deployments. Furthermore, we extract and present engineering insights to guide system deployment. The main contributions of our FWA study are summarized below.

- We implement and evaluate three types of codebooks: the Song codebook proposed in [27], the DFT codebook [9], and the Chebyshev-windowed DFT codebook [28]. For each codebook type, we analyze the impact of the number of beams in the BS codebook, denoted by B_b , on system performance. We evaluate a range of values for B_b under different scenarios, including codebook type, UL versus DL, and mid-band versus mmWave. Since each UE is matched with a BS beam based on its location instead of BA, we can have a large B_b without increasing training overheads. Note that on the UE side, we consider a fixed beam toward the BS.
- We compare the system performance at 3.5 GHz and 28 GHz and analyze the bandwidth and hardware requirements (e.g., the number of antennas and the number of RF chains at BS) in each band, to achieve the target MBRs. At 3.5 GHz, bandwidth is limited and expensive, but hardware components are mature due to widespread deployment and established manufacturing processes. In contrast, at 28 GHz, ample bandwidth is available; however, hardware is significantly more costly due to the higher frequency and stricter manufacturing requirements. Therefore, it is essential to identify a cost-effective trade-off between bandwidth and hardware within each band.

1.6 Thesis Outline

In this thesis, we focus on the RRM of codebook-based HBF systems. We investigate DL and UL separately because of the inherently different transmission mechanisms. For example, different UEs share the same power budget on DL while each UE has its own power budget on UL. In each part, the research starts with offline planning, which analyzes deployment strategies and the impact of parameters. With offline insights and targets, we propose online RRM solutions for real-time operations. Finally, we adapt our RRM solutions to a FWA setting to evaluate the feasibility of HBF in both the mid-band and mmWave for achieving MBR targets for home UEs on DL and UL. Particular emphasis is

placed on assessing the impact of codebook size and type on system performance in this context.

The research thesis is organized as follows. In Chapter 2, we begin by introducing different BF systems, including FDBF, ABF and HBF. Then, we conduct a thorough review of the literature and existing research on RRM with HBF systems specifically focusing on the DL and UL, respectively. Chapter 3 introduces our system model and operational processes for the DL and UL studies. Chapter 4 focuses on the RRM for DL. Chapter 5 studies the RRM for UL. In Chapter 6, we study the feasibility of HBF in both the mid-band and mmWave for the important use case of FWA in a rural area and analyze the impact of codebook type and size in this setting. Finally, Chapter 7 concludes the thesis.

Chapter 2

Background and Literature Review

In this chapter, we first introduce the background and in particular the types of beamforming architectures. Then, we present the existing research on RRM for mmWave codebook-based HBF system. In light of the differences in DL and UL RRM, we will conduct separate literature reviews for both cases. We assume that our system is based on OFDMA (Orthogonal Frequency Division Multiple Access) which is the de facto multiple access mechanism in 4G and 5G. In OFDMA, the bandwidth and time resources are organized into PRBs as shown in Fig. 2.1. Then, PRBs are allocated to users dynamically according to their channel conditions and the fairness criterion. With MIMO and BF techniques, a PRB can be shared by multiple users.

2.1 Analog, Fully Digital, and Hybrid BF Systems

Generally, there are three types of BF systems, analog BF, fully-digital BF, and hybrid BF [29], as shown in Fig. 2.2 where only the BS is represented. we assume one RF chain at the UEs.

In ABF systems, there is only one RF chain at the BS and N_b PS (Phase Shifter)s. Each PS connects the RF chain with one of the N_b antennas. As a result, ABF systems only support one stream per PRB and can be used for single-user MIMO systems. BF in this structure only includes the analog part which is realized by a group of PSs. The BS ABF vector remains constant across subchannels within a TS due to the shared utilization of common PSs.

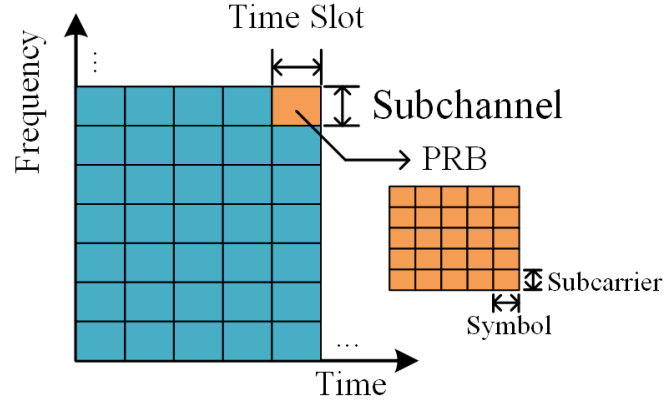


Figure 2.1: OFDMA structure

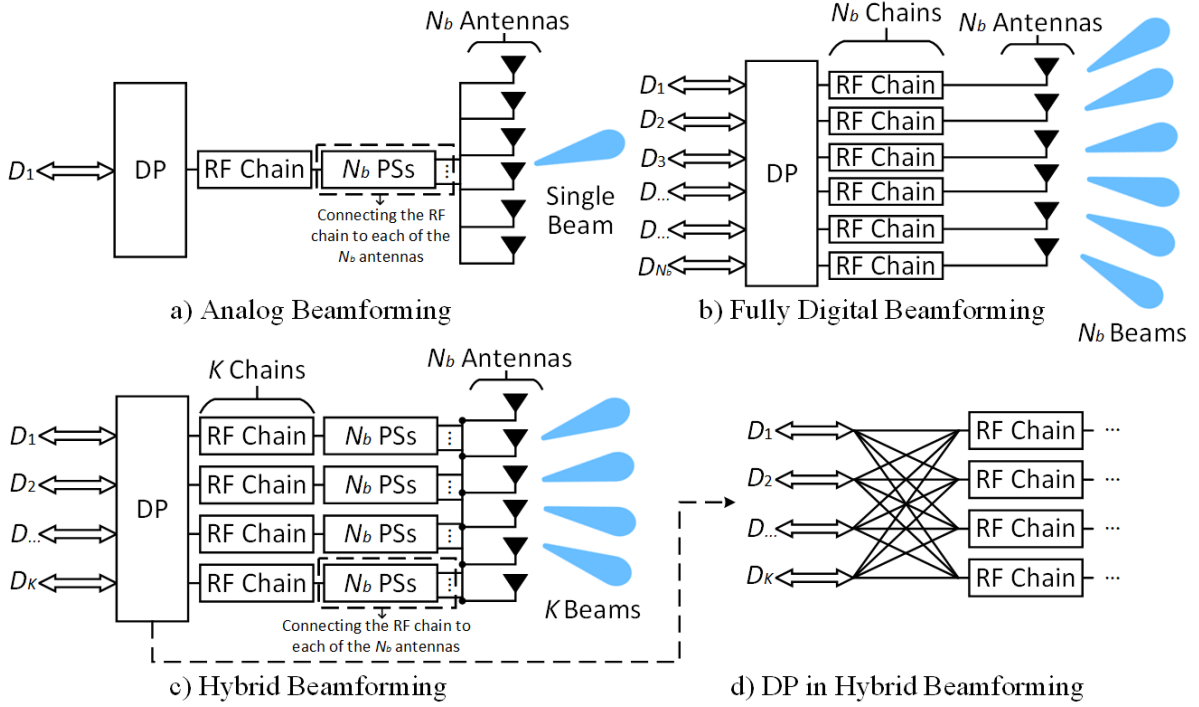


Figure 2.2: ABF, FDBF and HBF systems

In FDBF systems, each of the N_b antennas is connected with the DP by one RF chain. Therefore, FDBF systems can support as many as N_b streams. The BF vectors can vary among channels in a time slot since BF is entirely implemented in the DP.

In massive MIMO systems, N_b can be very large. Considering the cost and power consumption of RF chains including [AD \(Analog-to-digital\)](#)/[DA \(Digital-to-analog\)](#) converters, HBF systems emerge as a more practical solution in mmWave bands [30–32]. In HBF systems, there are K ($K \leq N_b$) RF chains, and each RF chain is connected with the N_b antennas by N_b PSs. This setup assumes a fully connected HBF structure with a total of $K \times N_b$ PSs. Alternatively, in a partially connected structure, RF chains can be connected to a subset of the N_b antennas to reduce the number of required PSs. In our research, we specifically consider a fully connected structure. Note that, similar to ABF systems, each of the K BS ABF vectors in HBF systems cannot be different on different channels within a TS.

We also show the interconnection of each data stream with the RF chains within the DP in Fig. 2.2. This illustrates that the DBF can be represented as a $K \times K$ matrix, which differs from the $N_b \times N_b$ precoding matrix in FDBF systems. Note that the DBF matrix can vary among channels in a time slot since DBF is implemented in the DP. In the N-DBF case, only the horizontal connections in the DP between data streams and RF chains will be utilized. In the ZF/MMSE-DBF case, all connections can be utilized for DBF.

2.2 Downlink Literature Review

In this section, we review the prior studies in the literature related to DL RRM with codebook-based HBF. Unlike our sequential approach where preferred ABF vectors are chosen in BA prior to RRM, several prior studies (e.g. [33–37]) jointly optimize RRM with ABF. However, besides high computational burden, such optimizations require per-UE full CSI (i.e., MIMO channel matrix) which is typically difficult to obtain in practice, especially for systems with frequency division [FDD \(Frequency Division Duplex\)](#) [2]. In our approach, we select the ABF vectors for each UE during BA before RRM. This in turn reduces the complexity of BF and simplifies the channel estimation by reducing the dimension of the effective channels (from matrix to scalar).

The problems studied in [8–11] are similar in nature to the one we investigate in this thesis. Additionally, BA is used to choose the ABF vectors for each UE in all of these studies. To clarify the distinctions, Table 2.1 provides a list of the main system characteristics and assumptions considered in [8–11] and our study. In this table, ‘*ConB*’ refers

Table 2.1: Main system characteristics in similar studies from DL literature
(OPT: optimized, ZF: zero-forcing, WF: water-filling, ConB: beam selection constraint, CH: frequency subchannels)

	CH	TS	$K < U$ included?	ConB	OPD	DBF	Fairness	Rate Function
[9]	1	1	×	×	×	ZF	×	Shannon
[10]	1	1	×	×	×	OPT	×	Shannon
[11]	1	1	✓	×	×	×	×	Shannon
[8]	1	≥ 1	✓	×	WF	ZF	PF	Shannon
Ours	≥ 1	≥ 1	✓	✓	OPT	ZF	PF	MCS

to the practical beam selection constraint discussed in Chapter 1, i.e., the ABF vectors cannot vary in frequency domain within a TS. Next, we elaborate on the novelties and distinctions.

In [9] and [10], it is assumed that $K \geq U$ where K is the number of RF chains at the BS and U is the number of UEs. Consequently, beam selection is not necessary as all the preferred beams (ABF vectors) can be selected at each time slot. Nevertheless, it is not typical for a small cell BS to be equipped with a large number of RF chains in practice due to the high cost and complexity, hence, $K < |\mathcal{B}_p| \leq U$ can occur especially in crowded networks, where $\mathcal{B}_p$ is the set of the indices of all the preferred BS beams. Unlike [9] and [10], we study RRM in a more general case where K can be smaller than $|\mathcal{B}_p|$ and U . In this case, not all the preferred beams can be selected and beam selection must be considered.

References [11] and [8] allow for $K < |\mathcal{B}_p|$. However, they restrict their studies to one frequency subchannel which is equivalent to performing RRM on a PRB basis, which means that the beam selection constraint ConB is ignored. We will show that the performance is significantly over-inflated when ignoring ConB and allocating the resources on a PRB basis. That is the reason why, unlike [11] and [8], we consider multiple frequency subchannels alongside beam selection constraint ConB, making the evaluations more practical.

One of our objectives in Chapter 4 is to investigate the impact of PD and DBF on performance. Among the similar papers listed in Table 2.1, only [8] adopts OPD and the rest consider EPD. To perform PD, [8] suggests using WF (Water-filling) algorithm based on data rates computed by Shannon-based rate function. Note that while Shannon-based rate function is commonly used, it may lead to misleading insights and wrong conclusions due to the fundamental shortcomings compared to practical MCSs [7]. Our evaluations also reveal that designing heuristic RRM schemes using Shannon-based rate function (as in [8]) can lead to a huge performance loss when practical MCSs are in use. In our research, we adopt a piecewise constant rate function modelling the existing practical MCSs rather

Table 2.2: Main system characteristics in the HBF UL references (Eff.: effective, CH: channel, OPT/O: optimized, WF: water-filling, E: equal, Pre-CBs: predefined codebooks, FN: fairness, NI: non-iterative solution).

	CSI	CH	$K < U$	PA	ABF	DBF	FN	NI
[12]	Full	≥ 1	$\times$	Equal	Online	MMSE	$\times$	$\times$
[13]	Full	1	$\times$	$\times$	Online	MMSE	$\times$	$\checkmark$
[14]	Full	1	$\times$	$\times$	Online	MMSE	$\times$	$\times$
[15]	Full	1	$\times$	$\times$	Pre-CBs	MMSE	$\times$	$\checkmark$
[16]	Eff.	1	$\times$	$\times$	Pre-CBs	ZF	$\times$	$\times$
[17]	Eff.	1	$\times$	$\times$	Pre-CBs	OPT	PF	$\times$
Ours	Eff.	≥ 1	$\checkmark$	O/WF/E	Pre-CBs	ZF	PF	$\checkmark$

than using Shannon-based rate function. In that case, WF algorithm is not the optimal solution [38], [39]. That is the reason we explore the optimal PD by considering the power distributed to UEs as variables in our joint optimization problem.

References [40] and [41] generate the DBF vectors with an iterative weighted sum-minimum mean square error algorithm, which introduces some complexity to both offline and online RRM schemes. Note that while there are some prior works on low complexity DBF algorithm [42], our numerical results reveal that after applying preferred ABF vectors, most of the UEs operate at high Signal-to-Noise Ratio (SNR) regime (assuming reasonable small cell radii) in which the performance of ZF is known to be close to optimum. Consequently, we adopt ZF algorithm as the DBF technique similar to [8], [9] and [10].

Fairness is an important aspect of RRM. Among the similar papers listed in Table 2.1, except [8], the rest employ sum rate as the objective in their optimizations and evaluations. Even in [8], only the proposed online algorithm is based on PF (Proportional Fairness). The derived performance upper-bound and numerical evaluations are based on the sum rate. In this thesis, We consider PF in the RRM optimization problem as well as in designing online heuristic algorithms.

2.3 Uplink Literature Review

In this section, we review previous studies in the literature related to UL RRM for HBF systems. Few studies have examined RRM in the uplink for the HBF structure based on predefined codebooks that we investigate. Table 2.2 summarizes related studies along with their system characteristics. The first thing to note is that, except for [12], none of the papers work on multiple channels and hence they cannot address the power allocation

and the beam selection correctly. [12] assumes $K > U$, thus there is no need for user selection. Secondly, [12], [13] and [14] are not based on predefined codebooks. Instead, they generate analog beams during RRM, which adds complexity and requires full channel state information (CSI). While [15] selects beams from predefined codebooks, it uses full CSI, for user selection. In contrast, our study uses HBF based on predefined codebooks where all RRM processes use effective channels instead of full CSI, reducing both complexity and overhead as in [16] and [17]. However, these two papers assume a single-channel scenario with enough RF chains to select all matched BS beams so that beam selection is not necessary, thereby bypassing one of the couplings that our work addresses. All the papers assume a rate function (i.e., one that maps an SINR into a rate) based on Shannon while we use a more practical (but also a more challenging) one based on a finite set of MCSs [43, 44]. Finally, only [17] from all the cited papers take fairness into account. We integrate fairness into our objective function, following the approach in [17, 45, 46].

One of our goals is to analyze the impact of PA strategies. Per-time-slot PA is not considered in [13–17] as they focus on single-channel scenarios. Meanwhile, [12] assumes equal power allocation across all subchannels within a time slot, which can be inefficient when the power is allocated too thinly over a large number of channels. In this thesis, we consider three PA strategies: Optimized PA (OPA) optimizes power allocation based on practical MCS levels. Water-filling PA (WFPA) allocates a UE’s power among its assigned channel in a time slot using water-filling after user selection. Equal PA (EPA) allocates a UE’s power equally across its assigned PRBs in a time slot after user selection. We use OPA in the joint optimization and offline heuristic. We use WFPA and EPA in offline and online heuristics.

Minimum mean square error (MMSE) or optimization-based digital beamforming (DBF) are employed in [12–15, 17]. However, these studies focus on scenarios where beam selection, user selection, and power allocation are not all required. For our multi-channel system, MMSE DBF makes the joint optimization highly nonconvex because of interference between data streams within a PRB. Instead, we adopt zero-forcing DBF at the BS to cancel the interference between the selected UEs within a PRB as in [16], thereby enabling the convexification of the optimization problem to obtain an optimal solution.

Iterative solutions are employed in [12, 14, 16, 17], which can lead to high runtime, making them impractical for real-time operation. In this thesis, we propose a non-iterative online solution with extremely low runtime comparable to a benchmark from the literature [24].

For completeness, we note that [43] investigates the RRM for the DL of a multi-channel HBF system based on predefined codebooks with $K < U$, but the solutions cannot be

Table 2.3: Main system characteristics in the FWA references (Stat.: statistical, Eff.: effective, CH: channel.

	BF	CSI	CH	$K < U$	mid-band	mmWave	DL	UL
[47]	FDBF	Stat.	≥ 1	$\times$	$\checkmark$	$\times$	$\checkmark$	$\times$
[48]	FDBF	Full	≥ 1	$\times$	$\times$	$\checkmark$	$\checkmark$	$\times$
[49]	FDBF	Full	≥ 1	$\times$	$\times$	$\checkmark$	$\checkmark$	$\times$
[50]	FDBF	Full	≥ 1	$\times$	$\checkmark$	$\times$	$\checkmark$	$\checkmark$
[51]	HBF	Eff.	1	$\times$	$\times$	$\checkmark$	$\checkmark$	$\checkmark$
Ours	HBF	Eff.	≥ 1	$\checkmark$	$\checkmark$	$\checkmark$	$\checkmark$	$\checkmark$

reused because RRM on the DL and the UL are inherently different due to power management.

2.4 Fixed Wireless Access Literature Review

In this section, we review previous studies in the literature related to the RRM in 5G FWA applied to broadband access. Table 2.3 summarizes related studies along with their system characteristics. Most existing works on FWA assume FDBF architecture [47–50]. As a result, their proposed RRM solutions are not applicable to our system. Moreover, unlike FDBF-based methods that require full channel state information (CSI) [48–50], our UL and DL solutions rely only on effective CSI, similar to the approach in [51]. While [51] also considers HBF architecture, it operates under the assumption that $K \geq U$ and only considers a single-channel setting,¹ ignoring the practical multi-channel case with $K < U$ and corresponding constraints on beam selection, user selection and power allocation as we discussed in Sections 2.2 and 2.3.

An important aspect of FWA system design is the choice between mid-band and mmWave spectrum. However, existing research typically considers only one of the two bands—either the mid-band (e.g., [47, 50]) or the mmWave band (e.g., [48, 49, 51]). In contrast, our study enables a direct resource requirement comparison between the two by employing the same cluster delay line channel model from 3GPP [52], while using band-specific channel parameters for the mid-band and mmWave band, respectively.

The balance between DL and UL is also critical in FWA systems, as MBR requirements apply to both directions, and DL and UL must share limited system resources, either time

¹Although a large bandwidth is considered in [51], the same single-channel RRM solution is used for all channels in the bandwidth.

(in TDD) or bandwidth (in FDD). In our study, we consider both DL and UL within the same TDD system, using a fixed DL-to-UL time slot ratio, as in [50] ([51] also evaluates both DL and UL, but independently without sharing the resource). Due to the inherent differences in the DL and UL transmission mechanisms, the resource requirements (e.g., K , N_b , B_b and bandwidth) may differ in order to satisfy the respective MBRs. Therefore, it is important to jointly consider both DL and UL when selecting system parameters for practical FWA deployment.

Chapter 3

System Model

This chapter is common to the DL and UL studies. The system model for FWA will be explained in Chapter 6. We consider an OFDMA system over a bandwidth made of C subchannels. The system is based on the fully connected HBF architecture (shown in Fig. 1.1) where the BS is equipped with N_b antennas, K RF chains and $N_b \times K$ phase shifters. Each RF chain is connected to all N_b antennas by N_b phase shifters. We use $\mathcal{U}$ and U to denote the set and the number of UEs, respectively. Typically, we have $K \ll N_b$. In addition, we do not require $K > U$. We assume that each UE has one RF chain connected with each of the N_u antennas by N_u phase shifters. BF at the BS includes ABF by the phase shifters and DBF by the digital processor. BF at the UE is purely analog because there is only one RF chain. The ABF at the BS is performed by selecting up to one beam for each RF chain from a predefined codebook. Each selected beam specifies the phase shifts between the corresponding RF chain and all BS antennas. The BS and UE codebooks are predefined prior to system operation to reduce real-time computation during RRM [27, 53]. Specifically, the BS codebook $\mathcal{C}_b$ and UE codebook $\mathcal{C}_u$ are defined as $\mathcal{C}_b = \{\mathbf{w}_j \in \mathbb{C}^{N_b \times 1} : \|\mathbf{w}_j\|^2 = 1, j = 1, \dots, B_b\}$ and $\mathcal{C}_u = \{\mathbf{v}_j(u) \in \mathbb{C}^{N_u \times 1} : \|\mathbf{v}_j(u)\|^2 = 1, j = 1, \dots, B_u\}$ respectively. B_b and B_u are the ABF codebook sizes of BS and UE, respectively. In the following, we assume that all UEs use the same codebook and remove the index u . For brevity, we also call $\mathbf{w}_j$, BS beam j , and $\mathbf{v}_j$, UE beam j .

We assume the system operates in a TDD (Time Division Duplex) mode. The bandwidth and time are organized into physical resource blocks (PRBs), each corresponding to a subchannel¹ in a time slot. Some time slots are allocated to the DL and others to the UL. The time span of a PRB is denoted by δ_T , which is the duration of a system time slot.

¹We use the terms channels and subchannels interchangeably.

3.1.1 Beam Alignment

BA matches each UE with M BS beams from $\mathcal{C}_b$ and one UE beam from $\mathcal{C}_u$. It can be realized by various techniques (e.g., [9, 54–59]) and many criteria can be used to select the matched beams for a UE. In this thesis, we adopt the average effective channel gain across all channels as the criterion, i.e., a UE ranks the beams from the highest average effective channel gain to the lowest. The RRM scheme that we will describe later in the thesis is agnostic to the BA method. Specifically, during BA, UE u selects the pair of beams (v_{l_u}, w_{b_u}) (where $v_{l_u} \in \mathcal{C}_u$ and $w_{b_u} \in \mathcal{C}_b$) that offers the highest criterion. After that, v_{l_u} is fixed and u selects the M BS beams from $\mathcal{C}_b$, which provide the top M highest criteria along with l_u . We say that these M BS beams are matched to UE u . We introduce the mapping $h(u, m) = b$ from UE u to indicate that BS beam w_b is the one that offers the m -th highest criterion to UE u with v_{l_u} . Accordingly, the set of M matched BS beams for UE u is given by $\{\mathbf{w}_{h(u,1)}, \dots, \mathbf{w}_{h(u,m)}, \dots, \mathbf{w}_{h(u,M)}\}$. Furthermore, we define $\mathcal{I}_s$ as the set of all matched BS beam indices across all UEs, that is, the set of $h(u, m) \forall u \in \mathcal{U}$ and $1 \leq m \leq M$, sorted in ascending order with duplicates removed. We also define $\mathcal{U}(b)$ as the set of UEs that are matched with BS beam b , as well as the function $g(u, b) = m$ or 0 indicating the rank of BS beam b among the BS beams matched with UE u (0 if beam b is not matched with UE u or $u = 0$).

BA is an expensive process which is done at a time scale larger than the time slot duration used by RRM, aligning with the variations of large-scale fading. In the following, we assume that BA is done once in the time horizon that we consider and hence, this first set remains fixed in our study. Note that BA can be performed in parallel, using different RF chains for different beams and separate frequency channels for different UEs, thereby accelerating the process.

3.1.2 Effective Channel Estimation

Following BA, CE is performed to obtain the effective channel state information (ECSI), represented as a $U \times M$ matrix per PRB for each UE. These matrices are derived from CE based on the UE's matched beams. Note that CE is performed frequently to take channel variations into account and that the UEs and the BS do not measure full CSI in the input generation processes. Effective channel estimation is typically done once per n_T TSs in time for each user, and if the bandwidth is large (i.e., C is large), it might be done per block of channels in frequency. In the following, we assume that effective CE is performed once per TS in time and once per frequency block, where we divide the frequency band in $Q \geq 1$ blocks.

Specifically, UE u transmits pilot signals using $\mathbf{v}_{l_u}$, $1 \leq Q \leq C$ times per time slot (i.e., for each block of $N_F = C/Q$ PRBs in a time slot) [60]. The BS measures the pilot signal of UE u in block q with each matched BS beam $\mathbf{w}_{h(u,m)}$ and computes the effective channel $g_{q,n,u,m}^{\text{eff}}$. Note that $g_{q,n,u,m}^{\text{eff}}$ can be simplified to $g_{q,n,u}^{\text{eff}}$ if $M = 1$.

For the UL, $g_{q,n,u,m}^{\text{eff}}$ represents the effective channel gain between UE u transmitting with $\mathbf{v}_{l_u}$ and the BS receiving with $\mathbf{w}_{h(u,m)}$ if $n = u$ or the effective channel interference between UE u transmitting with $\mathbf{v}_{l_u}$ and the BS receiving for UE n with $\mathbf{w}_{h(n,m)}$ if $n \neq u$, $\forall m \in \{1, \dots, M\}$. We also refer to this as the inter-user interference from UE u to UE n with $\mathbf{w}_{h(n,m)}$ on the UL.

For the DL, according to the channel reciprocity principle in TDD systems [60], $g_{q,n,u,m}^{\text{eff}}$ represents the effective channel gain between the BS transmitting with $\mathbf{w}_{h(u,m)}$ and UE u receiving with $\mathbf{v}_{l_u}$ if $n = u$ or the effective channel interference between the BS transmitting using $\mathbf{w}_{h(n,m)}$ (for UE n) to UE u receiving with $\mathbf{v}_{l_u}$ if $n \neq u$, $\forall m \in \{1, \dots, M\}$. We also refer to this as the inter-user interference from UE n with $\mathbf{w}_{h(n,m)}$ to UE u on the DL.

We define the ECSI (the effective CSI measured during CE) of UE u in block q as $\mathbf{g}_{q,u}^{\text{eff}} = [\mathbf{g}_{q,u,1}^{\text{eff}}, \dots, \mathbf{g}_{q,u,m}^{\text{eff}}, \dots, \mathbf{g}_{q,u,M}^{\text{eff}}]$ where $\mathbf{g}_{q,u,m}^{\text{eff}} = [g_{q,1,u,m}^{\text{eff}}, \dots, g_{q,U,u,m}^{\text{eff}}]^T \in \mathbb{C}^{U \times 1}$. Note that the defined effective channel gain and interference capture the impacts of the propagation channel and ABF collectively. The ECSI will be used for SINR calculation as explained in Section 5.1.1. The size of the ECSI matrix per PRB per UE increases linearly with M . Therefore, it is crucial to evaluate the impact of M on system performance to determine whether the overhead cost is justified.

We refer the reader to [5] and [7, Sec. II-C] for more details. The RRM techniques that we will introduce later are independent of the CE method. For the purposes of this thesis, we posit the perfection of CE.

Fig. 3.1 summarizes the data structure and operational procedures of our system. BA is performed once at the beginning of every N_T time slots, capturing the large-scale fading. CE, on the other hand, is executed independently for each of the Q blocks of PRBs within every time slot to track the small-scale fading variation along frequency and time. A block consists of $N_F = C/Q$ PRBs in frequency. Once BA and CE are completed, RRM and data transmission are carried out accordingly.

3.2 Digital Beamforming

DBF can enhance the performance of HBF systems by providing more control over inter-user interference. In this thesis, we adopt the well-known zero-forcing scheme whenever

DBF is present. This type of DBF removes the interference among the UE data streams after ABF, which simplifies the RRM problem. On the DL, we consider two design configurations: one without DBF, where only per-stream ABF is employed, and another with DBF. On the UL, we only consider the case with DBF because the coupling between user selection and power allocation forces us to solve the RRM problem jointly, unlike the DL where power distribution can be separated from RRM. When intra-cell interference exists, the joint problem becomes intractable on the UL. Note that the performance of ZF-DBF degrades if too many UEs are selected and the isolation between them is poor, for example, when many UEs with strong interference to each other are scheduled on the same channel. To mitigate this, we choose a BS codebook that ensures low inter-beam interference and incorporate careful UE selection in our RRM solutions.

Specifically, assume user set vector z_ℓ and the corresponding beam set vector ℓ in block q . Note that the beam set and user set are vectors because the position and order decide the mapping from a beam to a UE. Specifically, beam b_k is selected by UE u_k and a zero entry in z_ℓ indicates no UE is selected for the corresponding beam. To calculate the ECSI after ZF-DBF, we first remove the zero element(s) in z_ℓ and the corresponding beam(s) in ℓ to get $\ell' = [b_1, \dots, b_k, \dots, b_{L'}]$ and $z_{\ell'} = [u_1, \dots, u_k, \dots, u_{L'}]$ where b_k and u_k denote the indices of beams and UEs. The new ECSI vector of UE u_k is then:

$$\mathbf{g}_{\ell'}^{q,u_k}(z_{\ell'}) = [g_{q,u_1,u_k,m_{u_k}}^{\text{eff}}, \dots, g_{q,u_{L'},u_k,m_{u_k}}^{\text{eff}}], \quad (3.1)$$

where $m_{u_k} = g(u_k, b_k)$. Then the ECSI matrix for user set vector $z_{\ell'}$ is defined as

$$\mathbf{G}_{\ell'}^q(z_{\ell'}) = [\mathbf{g}_{\ell'}^{q,u_1}(z_{\ell'}), \dots, \mathbf{g}_{\ell'}^{q,u_{L'}}(z_{\ell'})]^T. \quad (3.2)$$

We also have the ABF matrix

$$\mathbf{F}_{\ell'}^{\text{ABF}}(z_{\ell'}) = [\mathbf{w}_{h(u_1,m_{u_1})}, \dots, \mathbf{w}_{h(u_k,m_{u_k})}, \dots, \mathbf{w}_{h(u_{L'},m_{u_{L'}})}]. \quad (3.3)$$

We define the ZF-DBF vector of UE $u_k \in z_{\ell'}$ as

$$\mathbf{d}_{\ell'}^{q,u_k}(z_{\ell'}) = \frac{\mathbf{a}_{\ell'}^{q,u_k}(z_{\ell'})}{\|\mathbf{F}_{\ell'}^{\text{ABF}}(z_{\ell'})\mathbf{a}_{\ell'}^{q,u_k}(z_{\ell'})\|_F} \quad (3.4)$$

where $\mathbf{a}_{\ell'}^{q,u_k}(z_{\ell'})$ is calculated from

$$\begin{aligned} & [\mathbf{a}_{\ell'}^{q,u_1}(z_{\ell'}), \dots, \mathbf{a}_{\ell'}^{q,u_k}(z_{\ell'}), \dots, \mathbf{a}_{\ell'}^{q,u_{L'}}(z_{\ell'})] \\ &= \mathbf{G}_{\ell'}^q(z_{\ell'})^H [\mathbf{G}_{\ell'}^q(z_{\ell'}) \mathbf{G}_{\ell'}^q(z_{\ell'})^H]^{-1}. \end{aligned} \quad (3.5)$$

Then the ECSI between UE u_k and u_j in user vector $z_{\ell'}$ in block q after applying both ABF and DBF is

$$h_{\ell'}^{q,u_k,u_j}(z_{\ell'}) = \mathbf{g}_{\ell'}^{q,u_j}(z_{\ell'})^T \mathbf{d}_{\ell'}^{q,u_k}(z_{\ell'}). \quad (3.6)$$

In the case of ABF only, we have

$$h_{\ell'}^{q,u_k,u_j}(z_{\ell'}) = g_{q,u_k,u_j,m_{u_j}}^{\text{eff}}. \quad (3.7)$$

Combined with zero element(s) in z_{ℓ} , we have

$$h_{\ell}^{q,u_k,u_j}(z_{\ell}) = \begin{cases} h_{\ell'}^{q,u_k,u_j}(z_{\ell'}), & \text{if } u_j, u_k \neq 0 \\ 0, & \text{otherwise} \end{cases}. \quad (3.8)$$

Note that $h_{\ell}^{q,u_k,u_j}(z_{\ell}) = 0$ for $u_k \neq u_j$ in the case of zero-forcing digital beamforming (ZF-DBF).

For the UL, $h_{\ell}^{q,u_k,u_j}(z_{\ell})$ denotes the inter-user interference from UE u_j to UE u_k , caused by UE u_j transmitting to the BS when the BS is receiving for UE u_k , or the signal of UE u_k if $u_j = u_k$.

For the DL, $h_{\ell}^{q,u_k,u_j}(z_{\ell})$ denotes the inter-user interference from UE u_k to UE u_j , caused by BS transmitting to UE u_k when UE u_j is receiving, or the signal of UE u_j if $u_j = u_k$.

3.3 Rate Function

The rate function $f(\cdot)$ translates the SINR (SNR in the case of ZF-DBF) of a UE in a given block into the achievable data rate in bps. In this thesis, we consider the rate function derived from practical MCSs instead of the widely used Shannon-based rate function. More specifically, we consider the data rate to be a piecewise constant function of SINR made of L_{MCS} levels corresponding to the L_{MCS} MCSs:

$$f(\gamma) = B_c (s_1 \mathbb{1}_{[\Gamma_1, \Gamma_2)}(\gamma) + \cdots + s_{L_{\text{MCS}}} \mathbb{1}_{[\Gamma_{L_{\text{MCS}}}, \infty)}(\gamma)), \quad (3.9)$$

where B_c is the bandwidth of a PRB in Hz, $\mathbb{1}_A(x)$ denotes the indicator function with a value of 1 if $x \in A$ and zero otherwise, Γ_l is the SINR decoding threshold for MCS l and s_l is the spectral efficiency of MCS l measured in bps/Hz.

3.3.1 Logarithmic Approximation

As it is difficult to deal with such a rate function in optimization problems due to discontinuities, we introduce a convex approximation of $f(\cdot)$ which will be used later in Chapter 4, namely:

$$\tilde{f}(\gamma) = B_c \left(\min \{ a \log (1 + \gamma/b), s_{\max} \} \right), \quad (3.10)$$

where a and b are parameters controlling the approximation and $s_{\max} = s_{L_{\text{MCS}}}$ is the spectrum efficiency of the highest MCS level. This approximation is employed to derive an upper bound and obtain an initial feasible solution for the DL RRM problem. It is necessary because the problem cannot be convexified when using an MCS-based rate function. However, once the initial solution is obtained, we further improve it by incorporating the MCS-based rate function, as discussed later in Section 4.1.3. Analyzing the improvement by considering practical MCS levels is one of research goals in this thesis.

Fig. 3.2 shows the piece-wise constant rate function, a convex tight approximation $\tilde{f}_1(\gamma)$ corresponding to $(a, b) = (1.237, 1.952)$ and an upper-bound approximation $\tilde{f}_2(\gamma)$ corresponding to $(a, b) = (1.181, 1.253)$, where the set of MCSs are adopted from 3GPP [61, Table 7.2.3-1]. We will use these functions in our evaluations.

3.3.2 Linear Approximation

We also introduce a linear approximation consisting of I linear functions, following the method in [38], to upper-bound the nonconvex rate function $f(\gamma)$ as

$$f(\gamma) \leq a_i \gamma + b_i \quad \forall i \in \{1, 2, \dots, I\}. \quad (3.11)$$

Note that $a_1 > \dots > a_I \geq 0$ and $b_I > \dots > b_1 \geq 0$. Using this linear upper bound, rather than the logarithmic upper bound in (3.10), is essential for converting the RRM problem into a LP problem, as discussed later in Chapter 5. Table 3.1 gives an example of the values of a_i and b_i for the linear upper bound.

3.4 Fairness

Fairness is crucial in RRM studies. In this thesis, we consider proportional fairness [45, 46], where the objective is to maximize the product of the UE throughputs over a certain horizon

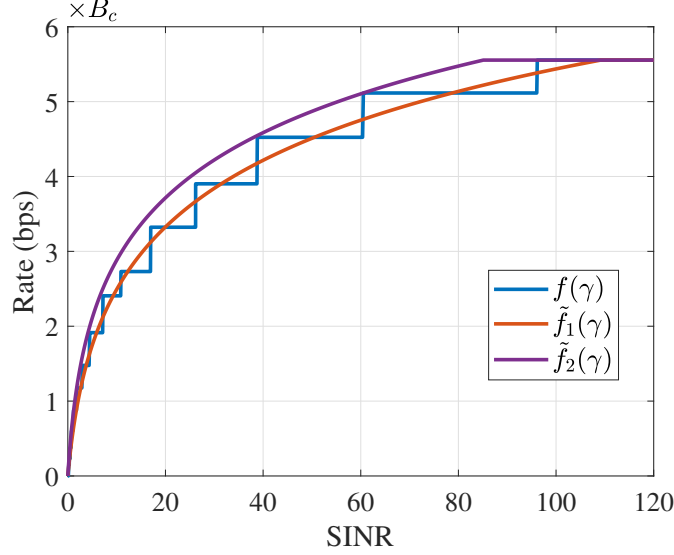


Figure 3.2: An example of piece-wise constant rate function $f(\gamma)$ and two convex approximations $\tilde{f}(\gamma)$, a tight one and an upper bound

except in Chapter 6, where we will introduce another criterion more suitable for FWA. The RRM is carried out per time slot because of beam selection (and power allocation on UL). At the beginning of time slot t , let $R_u(t)$ be the average throughput received by u over the past W slots. i.e.,

$$\begin{aligned} \max_{u \in \mathcal{U}} \prod W R_u(t) + \lambda_u(t) &\equiv \max_{u \in \mathcal{U}} \sum \log(W R_u(t) + \lambda_u(t)) \\ &\equiv \max_{u \in \mathcal{U}} \sum \log\left(1 + \frac{\lambda_u(t)}{W R_u(t)}\right), \end{aligned} \quad (3.12)$$

where $\lambda_u(t)$ is the throughput of UE u within time slot t and $R_u(t)$ is the average throughput received by u over the past W time slots at the end of time slot t , we use a moving average of size W to update $R_u(t)$ as $R_u(t+1) = (1 - 1/W)R_u(t) + \lambda_u(t)/W, \forall u \in \mathcal{U}$. Because $R_u(t)$ and the channel are time-varying, results need to be obtained for many time slots. For brevity of notation, we will remove the time slot index t when focusing on the RRM of one slot. Following the PF literature, we approximate the objective function in (3.12) by $\sum_{u \in \mathcal{U}} \frac{\lambda_u}{R_u}$ assuming $W R_u \gg \lambda_u$ and using the approximation $\log(1+x) \approx x$ for $x \ll 1$.

Table 3.1: a_i and b_i values for the linear approximation

i	1	2	3	4	5	6	7
a_i	0.6531	0.5652	0.4281	0.2758	0.2387	0.1716	0.1108
b_i	0	0.0658	0.2292	0.5129	0.6223	0.9238	1.3621
i	8	9	10	11	12	13	
a_i	0.0752	0.0637	0.0475	0.0279	0.0145	0	
b_i	1.7499	1.9448	2.3677	3.1290	3.9432	5.5547	

Chapter 4

Downlink Study

In this chapter, we present our study on DL radio resource management. The first step is to assess, offline, how various RRM procedures and system parameters impact performance. With the results from the offline study, we proceed to design low-complexity and high-performance online RRM schemes for system deployment. We assume each UE is equipped with one RF chain and matched with one pair of preferred beams ($M = 1$, one beam at the BS and the other at the UE). We only consider multiple matched BS beams per UE ($M > 1$) as an extension in Section 4.1.4.

4.1 Planning (Offline RRM Study)

In this section, we provide an offline planning RRM study where beam selection, user selection, PD, and optionally DBF, are jointly optimized. While the solutions to the RRM optimization problems can be of high complexity, the evaluation results provide engineering insights on the impact of various system parameters as well as the importance of different procedures such as DBF and PD. Further, the solutions will provide performance targets for designing online low-complexity RRM schemes which will be discussed in Section 4.2.

The RRM is performed per time slot. Following Section 3.4, our objective function in a time slot is

$$\max \sum_{u \in \mathcal{U}} \frac{\lambda_u}{R_u}, \quad (4.1)$$

where λ_u is the throughput of UE u within this time slot and R_u is the average throughput received by u over the past W time slots.

4.1.1 Problem Formulation

We formulate a generic RRM optimization problem. We will specialize it later depending if DBF is included or not and if PD is optimized or not. The problem is formulated for a given time slot, keeping track of the R_u 's. At each time slot, at most $L = \min(K, |\mathcal{I}_s|)$ beams can be selected (recall that $\mathcal{I}_s$ is the set of all preferred matched BS beams when $M = 1$) and the selected beam set should be kept the same for all PRBs in that time slot. We construct *beforehand* super set $\mathcal{L}$ which consists of all the possible beam set vectors ℓ of L preferred beams, i.e., we define

$$\mathcal{L} = \{\ell : \ell = [b_1, \dots, b_j, \dots, b_L], b_j \in \mathcal{I}_s, j \in \{1, \dots, L\}\} \quad (4.2)$$

Given that a beam set vector ℓ is selected for a time slot, the BS will select user set vectors. Note that these user set vectors could contain zero entries, as a beam may not be activated. A UE can only be in a selected user set vector, if its preferred beam is in the selected beam set vector ℓ , and no UEs in a user set vector should share the same preferred beam. Note that it is possible for multiple UEs to prefer the same beam. Therefore, a beam set vector ℓ can be mapped to multiple user set vectors. We construct *beforehand* the set of all distinct possible user set vectors corresponding to the beam set vector ℓ defined as:

$$\mathcal{M}_\ell = \{z_\ell : z_\ell = [u_1, \dots, u_j, \dots, u_L], u_j \in \{\mathcal{U}(b_j), 0\}\} \quad (4.3)$$

Note that we can compute $\mathcal{L}$ and $\mathcal{M}_\ell$ for each ℓ after BA and before RRM. Specifically, given a beam set vector ℓ , we choose at most one UE ($u = 0$ if no UE is chosen) from the set of UEs preferring beam $b \in \ell$ when constructing a user set vector z_ℓ .

Next, we introduce the variables of the joint RRM optimization problem. We first define α_ℓ as the beam selection binary variables indicating whether beam vector ℓ is selected in a given time slot. This variable does not change with q as the selected beams need to remain the same within a time slot. We have

$$\sum_{\ell \in \mathcal{L}} \alpha_\ell = 1, \quad (4.4)$$

$$\alpha_\ell \in \{0, 1\}, \forall \ell \in \mathcal{L}. \quad (4.5)$$

Given the BS selected beam set vector ℓ , various user set vectors can be scheduled in each of the QN_F vertical PRBs of each time slot. Let $\beta_\ell^q(z_\ell)$ be the fraction of PRBs within block q that user set vector z_ℓ is selected given that ℓ is active. As the channel coefficients

are the same in each PRB of a block (the bandwidth of a block is assumed to be smaller than the coherence bandwidth), it does not make any difference which PRBs are allocated to which user set vector as long as the fractions are preserved. We have

$$\sum_{z_\ell \in \mathcal{M}_\ell} \beta_\ell^q(z_\ell) \leq 1, \beta_\ell^q(z_\ell) \geq 0, \forall \ell, q, z_\ell \quad (4.6)$$

$$N_F \beta_\ell^q(z_\ell) \in \mathbb{Z}, \forall \ell, q, z_\ell, \quad (4.7)$$

where $\mathcal{Q} = \{1, 2, \dots, Q\}$ is the set of vertical block indices within a time slot. Equation (4.7) is due to the fact that we have N_F PRBs at each time slot and the number of PRBs assigned to a user set vector should be integer.

Note that the computation of α_ℓ and $\beta_\ell^q(z_\ell)$ will determine the allocation of the PRBs in a time slot to different user set vectors.

Next, we introduce the variables related to PD. Recall that we assume uniform power allocation over different frequency subchannels, i.e., we assume that the BS power budget for each PRB is $\frac{P_{\text{BS}}}{Q N_F}$, where P_{BS} is the total BS power budget. Let $p_\ell^{q,u}(z_\ell)$ be the power allocated to UE u in user set vector z_ℓ within the PRBs occupied by z_ℓ inside block q . Note that we have assigned the same per PRB power for all PRBs allocated to a set z_ℓ in block q . In other words, we consider optimizing PD with block-level granularity. While one can consider finer granularities (such as PRB-level) for the optimization of PD variables, it would significantly increase the complexity of the problem making it difficult to solve even in an offline manner.

In the design of ABF and DBF (if present), we ensure the norm of the overall HBF vector for each UE within a user set vector z_ℓ is 1, i.e., $\|\mathbf{F}_{\ell'}^{\text{ABF}}(z_{\ell'}) \mathbf{d}_\ell^{q,u}(z_{\ell'})\|_F = 1$ for all $u \in z_{\ell'}$, where $\mathbf{F}_{\ell'}^{\text{ABF}}(z_{\ell'})$ and $\mathbf{d}_\ell^{q,u}(z_{\ell'})$ (defined in Section 3.2) correspond to ABF and DBF, respectively (recall that $z_{\ell'}$ and ℓ' are z_ℓ and ℓ after removing zero elements and corresponding beams). This assumption enables us to decouple the PD from BF.

For a PRB in block q allocated to user set vector z_ℓ we have

$$\sum_{u \in z_\ell} p_\ell^{q,u}(z_\ell) \leq \frac{P_{\text{BS}}}{Q N_F} \quad \forall \ell \in \mathcal{L}, q \in \mathcal{Q}, z_\ell \in \mathcal{M}_\ell \quad (4.8)$$

$$0 \leq p_\ell^{q,u}(z_\ell), \forall \ell \in \mathcal{L}, z_\ell \in \mathcal{M}_\ell, u \in z_\ell, q \in \mathcal{Q} \quad (4.9)$$

The SINR of UE u in a PRB within block q allocated to user set vector z_ℓ is

$$\gamma_\ell^{q,u}(z_\ell) = \begin{cases} \frac{|h_\ell^{q,u}(z_\ell)|^2 p_\ell^{q,u}(z_\ell)}{\sum_{\substack{n \in z_\ell \\ n \neq u}} |h_\ell^{q,n}(z_\ell)|^2 p_\ell^{q,n}(z_\ell) + \sigma_{\text{PRB}}^2}, & \text{if } u \in z_\ell \text{ and } u \neq 0 \\ 0, & \text{otherwise} \end{cases} \quad (4.10)$$

where σ_{PRB}^2 is the noise power on a PRB. Recall that $u = 0$ in z_ℓ denotes that no UE is selected for the corresponding beam in ℓ . Note that SINR $\gamma_\ell^{q,u}(z_\ell)$ includes $h_\ell^{q,i}(z_\ell)$, $\forall i \in z_\ell$ which can be computed using (3.8). The DL data rate of UE u in a PRB within block q allocated to user set vector z_ℓ is defined as

$$r_\ell^{q,u}(z_\ell) = f(\gamma_\ell^{q,u}(z_\ell)), \forall u, q, z_\ell, \quad (4.11)$$

where $f(\cdot)$ is the rate function described in Section 3.3. Given α_ℓ , $\beta_\ell^q(z_\ell)$ and $r_\ell^{q,u}(z_\ell)$, we can compute the throughput λ_u of UE u in the time slot as

$$\lambda_u = \sum_{\ell \in \mathcal{L}} \alpha_\ell \sum_{z_\ell \in \mathcal{M}_\ell} \sum_{q \in \mathcal{Q}} N_F \beta_\ell^q(z_\ell) r_\ell^{q,u}(z_\ell), \quad \forall u \in \mathcal{U}. \quad (4.12)$$

Next, we formulate the general RRM optimization problem for a time slot. Given $h_\ell^{q,n,u}(z_\ell)$, $\mathcal{U}$, $\mathcal{Q}$, $\mathcal{L}$, $\mathcal{M}_\ell$, R_u , N_T and N_F , we formulate

$$\begin{aligned} \mathbf{\Pi} : & \max_{\beta_\ell^q(z_\ell), \alpha_\ell, p_\ell^{q,u}(z_\ell), \gamma_\ell^{q,u}(z_\ell), r_\ell^{q,u}(z_\ell), \lambda_u} \sum_{u \in \mathcal{U}} \frac{\lambda_u}{R_u}, \\ \text{s.t.} & (4.4), (4.5), (4.6), (4.7), (4.8), (4.9), (4.10), (4.11), (4.12). \end{aligned}$$

Problem $\mathbf{\Pi}$ is non-convex, non-linear and includes integer constraints. To obtain solutions, we first relax it by removing integer constraints in (4.5) and (4.7). Further, we introduce the following change of variables $\xi_\ell^q(z_\ell) \triangleq \alpha_\ell \beta_\ell^q(z_\ell)$ to eliminate the product of variables in (4.12). Subsequently, we formulate the relaxed transformed problem as

$$\begin{aligned} \mathbf{\Pi}^R : & \max_{\xi_\ell^q(z_\ell), \alpha_\ell, p_\ell^{q,u}(z_\ell), \gamma_\ell^{q,u}(z_\ell), r_\ell^{q,u}(z_\ell), \lambda_u} \sum_{u \in \mathcal{U}} \frac{\lambda_u}{R_u} \\ \text{s.t.} & (4.4), (4.8), (4.9), (4.10), (4.11), \\ & \sum_{z_\ell \in \mathcal{M}_\ell} \xi_\ell^q(z_\ell) \leq \alpha_\ell, \quad \forall \ell \in \mathcal{L}, \forall q \in \mathcal{Q}, \end{aligned} \quad (4.13)$$

$$0 \leq \xi_\ell^q(z_\ell) \leq 1, \quad \forall \ell \in \mathcal{L}, \forall q \in \mathcal{Q}, \forall z_\ell \in \mathcal{M}_\ell, \quad (4.14)$$

$$\lambda_u = \sum_{\ell \in \mathcal{L}} \sum_{z_\ell \in \mathcal{M}_\ell} \sum_{q \in \mathcal{Q}} N_F \xi_\ell^q(z_\ell) r_\ell^{q,u}(z_\ell), \quad \forall u \in \mathcal{U}. \quad (4.15)$$

The variables are all continuous in Problem Π^R . Note that the optimal objective of Problem Π^R is an upper-bound for that of Problem Π . Next, we provide methods to solve Problem Π^R for three different cases: (i) EPD (equal power distribution) with or without DBF, (ii) OPD without DBF, and (iii) OPD with ZF-DBF.

4.1.2 Solution Method for EPD

With equal power distribution, we have $p_\ell^{q,u}(z_\ell) = P_{BS}/(|z_{\ell'}|QN_F)$ for $u \neq 0$ (recall that $z_{\ell'}$ is z_ℓ after removing zero elements). Consequently, the SINRs defined in (4.10) and the data rates introduced in (4.11) can be computed prior to RRM optimization. Note that we do not set $p_\ell^{q,u}(z_\ell) = P_{BS}/(LQN_F)$ to allow for the use of user set vectors with smaller cardinality of non-zero elements if it is beneficial. Using EPD, the RRM problem can be formulated as follows. Given $h_\ell^{q,n,u}(z_\ell)$, $\mathcal{U}$, $\mathcal{Q}$, $\mathcal{L}$, $\mathcal{M}_\ell$, R_u , N_T , N_F and $r_\ell^{q,u}(z_\ell)$:

$$\Pi_{\text{EPD}} : \max_{\xi_\ell^q(z_\ell), \alpha_\ell, \lambda_u} \sum_{u \in \mathcal{U}} \frac{\lambda_u}{R_u}, \quad \text{s.t. (4.4), (4.6), (4.13), (4.14), (4.15)}.$$

Problem Π_{EPD} is a linear program (LP) which can be optimally solved using a commercial solver. Note that the existence or not of DBF does not impact the solution method. The reason is that having DBF only impacts the SINRs (see (4.10)) which are computed prior to RRM given EPD. In summary, in the EPD case, we can obtain the optimal solution for the RRM problem with or without DBF.

4.1.3 Solution Methods for OPD

While EPD simplifies the optimization, it restricts the performance. To further improve the performance, we consider the OPD case where the $p_\ell^{q,u}(z_\ell)$'s are kept as optimization variables. In this case, we need to solve Problem Π^R which can be reformulated as:

$$\begin{aligned} \Pi_{\text{OPD}} : \quad & \max_{\xi_\ell^q(z_\ell), \alpha_\ell, p_\ell^{q,u}(z_\ell), \gamma_\ell^{q,u}(z_\ell), r_\ell^{q,u}(z_\ell), \lambda} \lambda, \\ & \text{s.t. (4.4), (4.6), (4.8), (4.9), (4.10), (4.11), (4.13), (4.14),} \\ & \lambda = \sum_{\ell \in \mathcal{L}} \sum_{z_\ell \in \mathcal{M}_\ell} \sum_{q \in \mathcal{Q}} \xi_\ell^q(z_\ell) \sum_{u \in \mathcal{U}} \frac{r_\ell^{q,u}(z_\ell)}{R_u}. \end{aligned}$$

Due to the independence of $\xi_\ell^q(z_\ell)$ and $r_\ell^{q,u}(z_\ell)$, Problem Π_{OPD} can be decoupled into the two sub-problems discussed next. The first sub-problem is a weighted sum-rate optimization

which optimizes the PD variables given ℓ , z_ℓ and q :

$$\begin{aligned} \Pi_{\text{OPD}}^{\text{P1}}(\ell, z_\ell, q) : & \max_{p_\ell^{q,u}(z_\ell), \gamma_\ell^{q,u}(z_\ell), r_\ell^{q,u}(z_\ell)} \sum_{u \in \mathcal{U}} \frac{r_\ell^{q,u}(z_\ell)}{R_u}, \\ \text{s.t. } & (4.8), (4.9), (4.10), (4.11) \end{aligned}$$

We will provide solution methods to Problem $\Pi_{\text{OPD}}^{\text{P1}}(\ell, z_\ell, q)$ for scenarios with and without ZF-DBF below. Let $\Gamma_\ell^q(z_\ell)$ denote the optimal objective value of Problem $\Pi_{\text{OPD}}^{\text{P1}}(\ell, z_\ell, q)$. Given $\Gamma_\ell^q(z_\ell)$ for all $q \in \mathcal{Q}$, $\ell \in \mathcal{L}$ and $z_\ell \in \mathcal{M}_\ell$, we can then solve the second sub-problem as follows:

$$\begin{aligned} \Pi_{\text{OPD}}^{\text{P2}} : & \max_{\xi_\ell^q(z_\ell), \alpha_\ell, \lambda} \sum_{\ell \in \mathcal{L}} \sum_{z_\ell \in \mathcal{M}_\ell} \sum_{q \in \mathcal{Q}} \xi_\ell^q(z_\ell) \Gamma_\ell^q(z_\ell), \\ \text{s.t. } & (4.4), (4.6), (4.13), (4.14) \end{aligned}$$

which is an LP that can be optimally solved by efficient solvers. Consequently, to solve problem Π_{OPD} it is sufficient to solve sub-problem $\Pi_{\text{OPD}}^{\text{P1}}$ for all ℓ , q and z_ℓ to obtain the values $\Gamma_\ell^q(z_\ell)$ and then use them to solve sub-problem $\Pi_{\text{OPD}}^{\text{P2}}$. Note that no iteration is required between the sub-problems but an exhaustive computation of $\Gamma_\ell^q(z_\ell)$ for all ℓ , q and z_ℓ is.

Next, we provide solution methods for Problem $\Pi_{\text{OPD}}^{\text{P1}}(\ell, z_\ell, q)$ for the scenarios with and without ZF-DBF.

Solving $\Pi_{\text{OPD}}^{\text{P1}}(\ell, z_\ell, q)$ for the scenario with ZF-DBF

ZF-DBF eliminates the inter-beam interference, i.e., the residual interference after BA. Therefore, with ZF-DBF we have $h_\ell^{q,n,u}(z_\ell) = 0$ if $n \neq u$ and the SNR (it is now an SNR since there is no interference) expression (4.10) can be simplified as

$$\gamma_\ell^{q,u}(z_\ell) = \begin{cases} \frac{|h_\ell^{q,u,u}(z_\ell)|^2 p_\ell^{q,u}(z_\ell)}{\sigma_{\text{PRB}}^2}, & \text{if } u \in z_\ell, \\ 0, & \text{otherwise} \end{cases}, \quad (4.16)$$

In that case, the only non-convex part in $\Pi_{\text{OPD}}^{\text{P1}}(\ell, z_\ell, q)$ is Constraint (4.11), since it involves the piece-wise constant rate function f (modeling MCSs). To convexify the problem, we approximate $f(\gamma)$ by the convex rate function $\tilde{f}_1(\gamma)$ (resp. $\tilde{f}_2(\gamma)$) introduced in Section 3.3 to acquire the feasible (resp. upper-bound) solution and objective value for $\Pi_{\text{OPD}}^{\text{P1}}(\ell, z_\ell, q)$. With the feasible (resp. upper-bound) solution and objective value

for $\Pi_{\text{OPD}}^{\text{P1}}(\ell, z_\ell, q)$ for all ℓ , q and z_ℓ , we can solve $\Pi_{\text{OPD}}^{\text{P2}}$ to obtain a feasible (resp. an upper-bound) solution and objective value for Π_{OPD} .

Specifically, to obtain the feasible (resp. upper-bound) solution and objective value of Π_{OPD} , we solve $\Pi_{\text{OPD}}^{\text{P1}}(\ell, z_\ell, q)$ for all $q \in \mathcal{Q}$, $\ell \in \mathcal{L}$ and $z_\ell \in \mathcal{M}_\ell$ after replacing $f(\gamma)$ by $\tilde{f}_1(\gamma)$ (resp. $\tilde{f}_2(\gamma)$). Then, for the feasible solution, we improve the power distribution solution, i.e., the power variables, considering the MCS-based rate function. Specifically, we first calculate the corresponding MCS level of each UE and the minimum additional power required to increase this level by one. Then, we identify the UE that yields the largest increase in the weighted sum rate per unit of power and increase its power to raise its MCS level by one. This two-step procedure is repeated until the remaining power is insufficient to further increase the MCS level of any UE. We use the improved solutions to calculate the data rate according to the piece-wise constant rate function f , and use the resulting $\Gamma_\ell^q(z_\ell)$ into $\Pi_{\text{OPD}}^{\text{P2}}$ to solve it. We will compare the system performance with and without this two-step improvement process, as it shows the difference between power distribution based on MCS-based rate function and that based on Shannon-based rate function.

Our numerical evaluations provided in Section 4.1.5 reveal that the feasible solution and the upper-bound are close and hence the optimal solution to our original problem is well approximated by the feasible solution.

Solving $\Pi_{\text{OPD}}^{\text{P1}}(\ell, z_\ell, q)$ for the N-DBF scenario

In that case, there is inter-beam interference and therefore, the SINR given in Eq. (4.10) is a non-convex function of the PD variables causing the problem to be highly non-convex. One can only obtain a locally optimal solution by applying a local search around an initial solution using commercial solvers such as MINOS. This implies that the quality of the solution depends on the initial solution for the power values. We have tried various methods for finding an initial solution including (i) setting all the power variables to zero, (ii) using the solution of EPD, and (iii) using the solution of Problem $\hat{\Pi}_{\text{OPD}}^{\text{P1}}(\ell, z_\ell, q)$ as the initial solution. Our numerical evaluations revealed that using method (iii) consistently provides a better performance. Given the initial point, using an approach similar to the one used in Section 4.1.3, we transform the problem using a tight convex approximation of the rate function f and find a local optimum using MINOS in AMPL.

4.1.4 Extension to Multiple Matched BS Beams Per UE

Until now, we have assumed that one BS beam is matched for each UE during the BA process. Having more than one matched BS beams per UE increases the flexibility of beam set and user set selections which can potentially improve the performance (at a cost of higher BA and CE overhead of course). In this subsection, we investigate the RRM problem when multiple (say M) BS beams are matched with each UE during the BA process. Note that the UE is still assumed to have a single RF chain. Therefore, at most one of the matched BS beams can be selected at a time for each UE. To elaborate on the setting, we follow Section 3.1.1 considering $M > 1$.

During CE, the effective CSI of each UE for all its M matched beam pairs is estimated. Following Section 3.1.2, we have $\mathbf{g}_{q,u}^{\text{eff}}$ as the ECSI of UE u in block q as $\mathbf{g}_{q,u}^{\text{eff}} = [\mathbf{g}_{q,u,1}^{\text{eff}}, \dots, \mathbf{g}_{q,u,m}^{\text{eff}}, \dots, \mathbf{g}_{q,u,M}^{\text{eff}}]$ where $\mathbf{g}_{q,u,m}^{\text{eff}} = [g_{q,1,u,m}^{\text{eff}}, \dots, g_{q,U,u,m}^{\text{eff}}]^T \in \mathbb{C}^{U \times 1}$. Recall $g_{q,n,u,m}^{\text{eff}}$ represents the effective channel gain between the BS transmitting with $\mathbf{w}_{h(u,m)}$ and UE u receiving with $\mathbf{v}_{l_u}$ if $n = u$ or the effective channel interference between the BS transmitting using $\mathbf{w}_{h(n,m)}$ (for UE n) to UE u receiving with $\mathbf{v}_{l_u}$ if $n \neq u$, $\forall m \in \{1, \dots, M\}$.

Note that there is a special case when more than one of the matched BS beams of a UE belong to ℓ . In such cases, a UE should not be selected multiple times in z_ℓ . Hence, we remove instances of z_ℓ that include a UE more than once from $\mathcal{M}_\ell$ since we assume there is at most one stream per UE. Given beam set vector ℓ , user set vector z_ℓ and block q , following Section 3.2, we calculate $h_\ell^{q,u_k,u_j}(z_\ell)$ which denotes the inter-user interference from UE u_k to UE u_j , caused by BS transmitting to UE u_k when UE u_j is receiving, or the signal of UE u_j if $u_j = u_k$.

The problem formulations and solutions provided in the previous subsections of this section are applicable to the case of multiple matched BS beams per UE. We will provide simulation results in Sec. 4.1.5 to evaluate the impact of increasing the number of matched BS beams per UE to two.

4.1.5 Numerical Evaluations for Offline RRM Study

In this section, we present the numerical results of our offline study. The results enable us to evaluate the impact of PD, ZF-DBF as well as system parameters such as K , B_b , and B_u . The results also demonstrate the impact of beam selection constraint ConB justifying the need for performing RRM optimization jointly for a time slot as opposed to a per-subchannel optimization approach proposed in the literature.

We consider the DL of a small cell of radius $r = 75$ m within which the UEs are distributed uniformly. The BS height is set to 10 m and the BS power budget is considered to be $P_{BS} = 27$ dBm. We assume that BS and UEs are equipped with uniform linear antenna arrays (ULAs) with half-wavelength inter-element spacing. The system operates at 28 GHz and the bandwidth is assumed to be 100 MHz. Furthermore, regarding the size of the blocks, we assume $N_F = 6$, leading to $B_r = 4.32$ MHz and $Q = 22$. We have chosen the value of N_F based on the channel model introduced later such that the channel coefficients can be considered constant within each block. To elaborate, we have generated 5×10^4 *channel realizations* by generating 50 realizations with 10 users and 100 time slots and calculated the root mean squared (RMS) delay for each realization. We have chosen N_F to ensure that $\frac{1}{\bar{\tau}_{\text{rms}}} \gg B_r$ where $\bar{\tau}_{\text{rms}}$ is the mean of the RMS delay over different realizations. Additionally, we assume low-mobility scenarios in mmWave frequency bands [62] where UEs can be considered as fixed. We adopt the set of discrete MCSs from [61, Table 7.2.3-1]. We use the method in [63] to generate the ABF codebooks (necessary for BA) for BS and UE. The rest of the simulation parameters are listed in Table 4.1. During BA, we assume that the beam pairs are chosen based on the SNR they provide, i.e. the beam pairs leading to the highest SNRs are selected. Additionally, we neglect the overhead of BA and CE and the processing time required for RRM since this is an offline study. Furthermore, we assume there are frequency guard-bands to prevent inter-band interference and neglect the corresponding overhead.

For our numerical evaluations, we adopt a multi-ray wide-band channel model which incorporates multiple scattering clusters, and each cluster comprises a number of spatial paths [64]. The MIMO channel between BS and UE u at block q is defined as the following $N_u \times N_b$ matrix:

$$\mathbf{G}_{q,u} = \frac{1}{\sqrt{N_{\text{path}}}} \sum_{d=1}^{N_{\text{cluster}}} \sum_{l=1}^{N_{\text{path}}} g_{d,l}^u e^{-j2\pi\tau_{d,l}^u f_q} \mathbf{a}_{\text{RX}}(\phi_{d,l}^u) \mathbf{a}_{\text{TX}}^H(\theta_{d,l}^u),$$

where N_{cluster} is the number of clusters, N_{path} is the number of paths in each cluster, $g_{d,l}^u$ is the coefficient of the path l of cluster d , $f_q = f_c - \frac{BW}{2} + (q-1)\frac{BW}{Q}$ is the center frequency of block q , and $\tau_{d,l}^u = \tau_{0d}^u + \tau_{1d,l}^u$ in which τ_{0d}^u is the group delay of the d -th cluster and $\tau_{1d,l}^u$ is the relative path delay of the l -th path in the d -th cluster. It is assumed that $g_{d,l}^u = \sigma_d^u \sqrt{\kappa_{d,l}^u} e^{j\varphi_{d,l}^u}$ where σ_d^u models path-loss and shadowing, $\kappa_{d,l}^u$ models the fraction of the signal power carried by the l -th path within the d -th cluster, and $\varphi_{d,l}^u \sim \mathcal{U}(0, 2\pi)$ models the phase of the l -th path within the d -th cluster. We assume $\sigma_d^u = \sqrt{(v_d^u 10^{-0.1PL_u})}$, where v_d^u is the fraction of the signal power carried by the d -th cluster and PL_u (in dB) is the impact of path loss and shadowing for UE u . Moreover, we

assume τ_{0d}^u and $\tau_{1d,l}^u$ are random variables exponentially distributed with mean of 200 nsec and 20 nsec, respectively. Channel measurements indicate that the average power decays exponentially with delay [65, 66]. To model this, we use the approach described in [65, Section III-C] to obtain v_d^u from τ_{0d}^u and $\kappa_{d,l}^u$ from $\tau_{1d,l}^u$. Moreover, $\theta_{d,l}^u$ and $\phi_{d,l}^u$ represent the AoD and the AoA corresponding to the l -th path of cluster d of the channel which are Gaussian distributed around the cluster central angle as described in [65]. As mentioned earlier, we consider the system within a BA period that we assume equal to T time slots. We also assume the large-scale fading, AoAs and AoDs remain fixed within this BA period. Additionally, $\mathbf{a}_{\text{TX}}(\cdot)$ and $\mathbf{a}_{\text{RX}}(\cdot)$ are the array response vectors at the BS and UE u which depend on the antenna array type. For instance, for a ULA including N antennas with half-wavelength inter-element distance, $\mathbf{a}(\cdot)$ is given by $\mathbf{a}(\phi) = [1, e^{j\pi \sin \phi}, \dots, e^{j(N-1)\pi \sin \phi}]^T$, where ϕ is the AoD if the array is at the transmitter and it is AoA if the array is at the receiver.

Let $\Omega(U)$ be a network realization obtained by generating U uniformly distributed UE positions in the cell excluding a ring with a radius of 6 m around the BS (where path loss models are not valid). It also includes the channel parameters corresponding to these positions. For a given set of system parameters and a value of U , we generate $Z = 50$ random realizations. For each realization, we carry out BA, once in the beginning, obtain the matched beams per user, and perform RRM for all the $T = 100$ time slots of that BA period. RRM results need to be obtained for T time slots because $R_u(t)$ and the channel are time-varying. Since our focus is on PF, the performance for one realization is defined to be the Geometric Mean (GM) of the UEs' throughputs over 100 time slots, i.e.,:

$$\text{GM}(\Omega(U)) = \left(\prod_{u=1}^U \frac{1}{T} \sum_{t=1}^T \lambda_u^t(\Omega(U)) \right)^{\frac{1}{U}} \quad (4.17)$$

To obtain the $\lambda_u^t(\Omega(U))$, we solve the joint RRM problem sequentially for each time slot of the T time slots. Finally, we take the arithmetic mean of the GMs over all realizations and compute $\overline{\text{GM}} = \sum_{\Omega} \text{GM}(\Omega(U))/Z$ as the performance measure. Next we provide the simulation results.

Recall that in the EPD case, Problem $\mathbf{\Pi}^{\text{R}}$ is equivalent to LP Problem $\mathbf{\Pi}_{\text{EPD}}$ which can be solved to optimality. Our evaluations show that, for both ZF-DBF and N-DBF cases in the EPD scenario, the solutions satisfy the integer constraints (5.6) and (5.8). Therefore, the relaxation from $\mathbf{\Pi}$ to $\mathbf{\Pi}^{\text{R}}$ does not cause optimality loss in the EPD scenario.

For the OPD cases, as explained in Section 4.1.3, we evaluate the quality of our proposed feasible solution for the OPD scenario with ZF-DBF by comparing it to an upper-bound.

Table 4.1: Simulation Parameters

Parameter	Value
Number of antennas at BS N_b	128
Number of antennas at UE N_u	16
Noise power spectral density N_0	-174 dBm/Hz
Moving average parameter W	10
Initial value of R_u	2

Fig. 4.1 illustrates feasible and upper-bound values of $\overline{\text{GM}}$ as a function of the number of BS RF chains K for that scenario. Note that the gap between the feasible solution performance and the upper bound is small confirming that the proposed feasible solution is quasi-optimal. We also note that we cannot evaluate the quality of the feasible solution for the OPD scenario with the N-DBF case due to the lack of performance upper-bound. However, this is not critical since we will show that the performance of the ZF-DBF case (for which a quasi-optimal solution is found) is significantly better than the case with N-DBF justifying the use of ZF-DBF in practice.

The following remark highlights an important observation from the evaluations which will be used in the design of online RRM schemes discussed in Sec. 4.2.

Remark 1 *The optimal/feasible solutions for all scenarios, select one beam set vector per time slot and one user set vector per block.*

Next, we evaluate the impact of various RRM procedures such as PD and DBF as well as different system parameters on the system performance. Fig. 4.2 depicts the system performance measure $\overline{\text{GM}}$ as a function of K for different scenarios. The fact that the performance plateaus when K increases above the number of users U is natural since at most one stream can be used per user. Fig. 4.2 shows the impact of PD and DBF. Considering ‘EPD, N-DBF’ as the baseline scenario, the gain in $\overline{\text{GM}}$ obtained by performing OPD instead of EPD increases as K increases and reaches 17.3% (w.r.t. the baseline) when $U = 10$. We also observe that the gain by replacing N-DBF with ZF-DBF is not large with small K , because the number of selected UEs per PRB is small restricting the interference and the gain of ZF-DBF. Increasing K enhances the performance gain of ZF-DBF by up to 27.4% (w.r.t. the baseline). When both OPD and ZF-DBF are performed, the performance gain (w.r.t. the baseline) reaches 55.2% for $U = 10$. The reason for such significant performance gains is the fact that ZF-DBF mitigates the inter-beam interference and OPD improves the fairness among UEs by properly distributing transmit power within the PRBs.

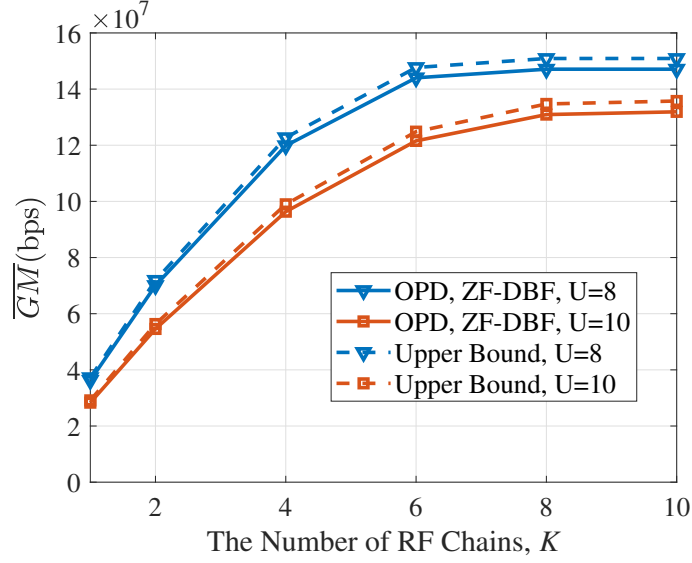


Figure 4.1: Quality of feasible solution with $N_b = 128$, $N_u = 16$, $B_b = 32$, $B_u = 4$.

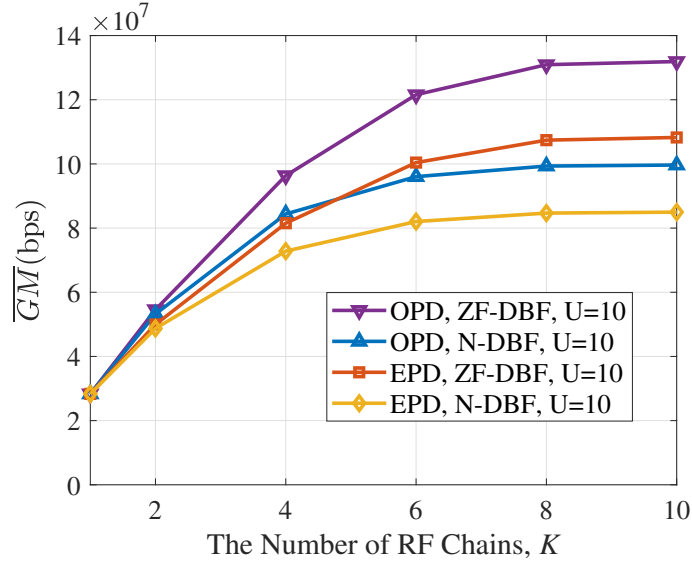


Figure 4.2: System performance vs. K for different cases EPD/OPD and ZF-DBF/N-DBF with $N_b = 128$, $N_u = 16$, $B_b = 32$, $B_u = 4$.

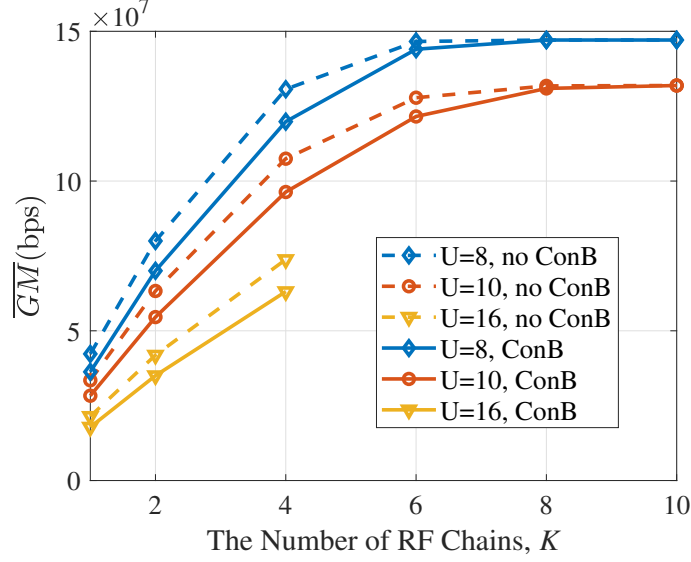


Figure 4.3: Impact of ConB on system performance considering OPD, ZF-DBF with $N_b = 128$, $N_u = 16$, $B_b = 32$ and $B_u = 4$.

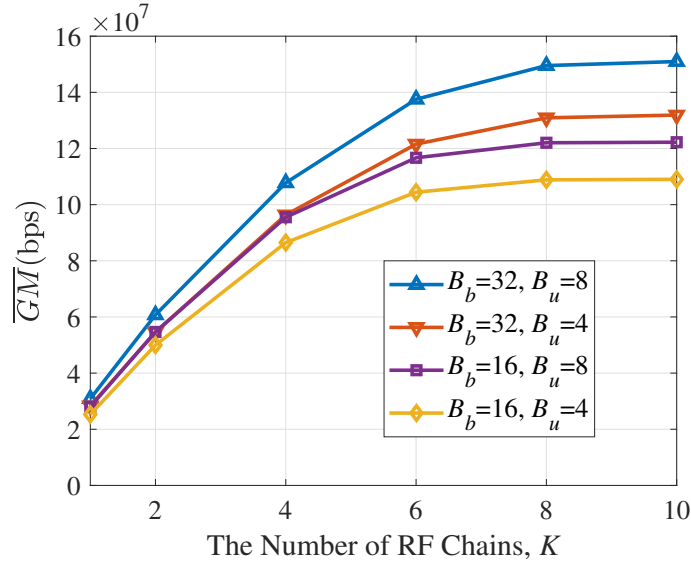


Figure 4.4: System performance vs. K for different ABF codebook sizes considering OPD, ZF-DBF with $U = 10$, $N_b = 128$, and $N_u = 16$.

Fig. 4.3 illustrates the importance of beam selection constraint ConB (i.e., that only one beam set can be selected per time slot). In this figure, we consider the OPD scenario with ZF-DBF and compare the performance when ConB is considered (i.e., performing RRM for a time slot with ConB) and when ConB is overlooked (i.e., performing RRM on a per-PRB basis neglecting ConB). First, we observe that the performances of ‘ConB’ and ‘no ConB’ are the same if $K \geq U$. This is expected since if $K \geq U$, all the preferred beams can be selected at every time slot which makes beam selection trivial. Hence constraint ConB is automatically satisfied and user selection and PD can be done on a PRB basis. However, when $K < U$, we observe that overlooking ConB overestimates the performance by as much as 16.5%, 18.2% and 20.1% for $U = 8$, $U = 10$ and $U = 16$, respectively. The over-estimation gets larger as U increases, indicating that the significance of ConB increases with U for the regime where $K < U$. Note that implementation cost and complexity limit the value of K in practice, making $K < U$ a reasonable assumption for practical scenarios. Therefore, it is misleading to overlook ConB in practical scenarios.

Next, we examine the impact of the number of beams B_b (resp. B_u) in the BS (resp. UE) ABF codebook on the system performance of the OPD scenario with ZF-DBF where $U = 10$ UEs are considered. In Fig. 4.4, considering the case with $B_b = 16$ and $B_u = 4$ as the baseline, we observe that doubling B_b can improve the performance by up to 21% while doubling B_u can at most provide 12.1% improvement. Furthermore, Doubling both B_b and B_u can improve the performance by up to 38.5%. We conclude that increasing the number of beams in the ABF codebooks can considerably boost the system performance. The reason is that increasing B_b and B_u leads to narrower beams which increases the BF gain. Moreover, it is more beneficial to increase the number of beams at BS than at the UEs since increasing B_b not only increases BF gain, but also reduces inter-beam interference due to using narrower beams. We also highlight that BA complexity and overhead increase with increasing B_b and B_u which limits the number of beams we can have during system operation.

Fig. 4.5 compares the performance of the case with one pair of preferred beams ($M = 1$) per UE and the case with two pairs ($M = 2$) for each UE after BA. We observe not much gain of performance for K small and up to 22.1% gain for larger K with $M = 2$ due to the more flexibility in beam and user selection which can reduce the interference. However, please note that the cost of such an improvement is a larger amount of ECSI (M^2 times) to be measured.

Water-filling (WF) is a well-known algorithm used for online power optimization in RRM studies. We refer the reader to [67] for more information on WF algorithm. WF is known to be optimal for such problems when a Shannon-based rate function is used. However, it is not optimal for the practical piecewise constant MCS rate function which

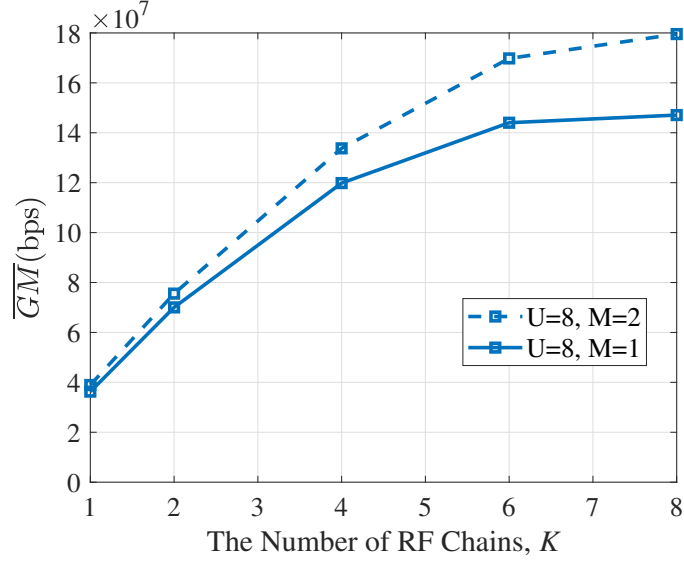


Figure 4.5: (a) Impact of the number of preferred beams on system performance considering OPD, ZF-DBF with $N_b = 128$, $N_u = 16$, $B_b = 32$ and $B_u = 4$.

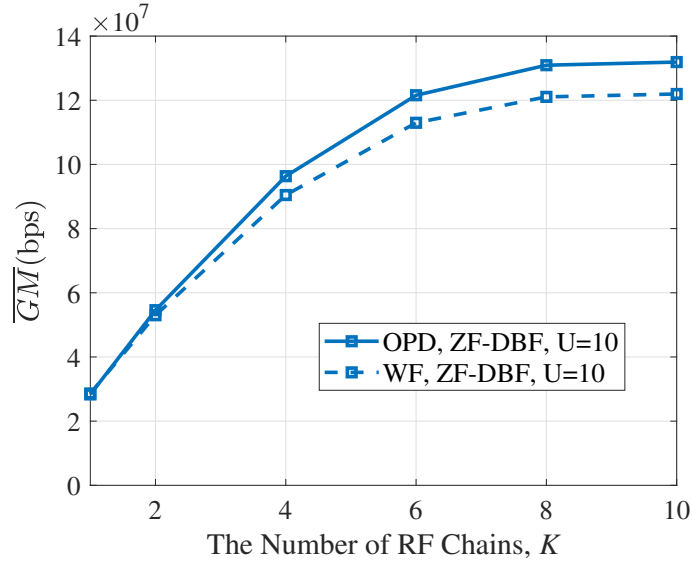


Figure 4.6: WF performance for ZF-DBF with $N_b = 128$, $N_u = 16$, $B_b = 32$, $B_u = 4$.

we adopted. We have done some simulations to verify if WF produces reasonable solutions. In the first set of simulations, we have solved the OPD problem with ZF-DBF by solving $\Pi_{\text{OPD}}^{\text{P1}}$ and modifying the power variables using the two-step MCS-based improving process followed by solving $\Pi_{\text{OPD}}^{\text{P2}}$ using our approach. In the second set of simulations, we solve the OPD problem with ZF-DBF as follows: we use WF (equivalent to solving $\Pi_{\text{OPD}}^{\text{P1}}$ without the two-step MCS-based improving process) and then use the results to solve $\Pi_{\text{OPD}}^{\text{P2}}$ using our approach. Fig. 4.6 shows the results for the case with $U = 10$ UEs. We observe that while WF is sub-optimal, the gap between WF performance and our OPD approach is not large (around 10% performance loss when $K = 10$). This reveals that a combination of WF and $\Pi_{\text{OPD}}^{\text{P2}}$ is sufficiently good in the case of ZF-DBF. We have conducted similar simulations for the case of N-DBF and obtain similar results. These observations justify the use of WF in our online heuristic algorithms in Section 4.2.

In summary, the offline study provides several important insights that help us design online heuristics for operation:

- Ignoring constraint ConB can yield significant performance over-estimation in practice.
- Good performance can be achieved even when K is not chosen to be close to U . For example, with $U = 10$ UEs, $K = 6$ yields 92% of the best performance.
- Both OPD and ZF-DBF can considerably improve the system performance, particularly when the BS can serve a large number of UEs within a PRB.
- Having more beams in the ABF codebooks can significantly boost the system performance, and it is more helpful to enlarge the codebook on the BS side than the UE side.
- EPD provides worse performance than OPD and less flexibility since the only “knobs” are beam selection and user selection which makes fairness enforcement in an online heuristic difficult.

4.2 Operation (Online RRM Study)

This section is dedicated to the design of online RRM schemes, which aim to achieve high performance with low complexity. We only consider the OPD case since it performs better than the EPD cases and provides more flexibility. Unlike offline RRM studies that focus solely on improving system performance, our objective is to develop practical schemes

that are easy to implement. We begin by introducing a low-complexity RRM benchmark scheme that can be used with ZF-DBF or without it (i.e., N-DBF). We then propose two novel and low-complexity heuristic RRM methods, one for ZF-DBF and one for N-DBF. In designing these methods, we leverage Remark 1 in Section 4.1.5, which suggests selecting a single beam set for each time slot and a single user set for each block within a time slot. Finally, we compare the performance and runtime of our heuristics and the low-complexity benchmarks. We also show the performance of the offline schemes (Section 4.1) to validate that our proposed heuristics strike an acceptable balance between performance and complexity.

In the following, we consider a particular BA period made of $T = 100$ time slots (roughly half second) and describe the heuristics for a time slot within this period. We consider one matched BS beam per UE (we will discuss the extension for $M > 1$ in Section 4.2.2). BA yields the preferred BS beam set $\mathcal{I}_s$. At the beginning of each time slot, the inputs of the online schemes are $\mathcal{I}_s$, the R_u 's (the weights of each user) and the ECSI per block per user. Beam selection is about selecting a subset of $\mathcal{I}_s$ of size L for each time slot (recall that because of ConB, the beam set cannot change within a time slot). Then, for each block, given the beam set, we perform *in parallel*, user selection, DBF optionally, and PD. At the end, we compute the rate seen by each user in the time slot and update their weights accordingly.

4.2.1 Low-complexity Benchmark

In this section, we propose a low-complexity benchmark to perform beam and user selection based on a round-robin (RR) approach. Specifically, we list all the beams in $\mathcal{I}_s$ and take L beams in a RR fashion as a beam set for each of the T time slots. Subsequently, given a time slot and its beam set, the benchmark generates all possible user sets of size L (one UE per beam) to take into account the fact that multiple UEs might have the same preferred BS beam. We then select one of these user sets per block in a RR fashion inside the time slot. In the case of the benchmark, the beam and user selections only require the preferred BS beam of each UE so it can be done for all the T time slots just after BA.

Next, DBF (if available) and PD are performed for each block of a time slot, given the selected UEs. Depending on the availability of DBF, there are two possible benchmarks: (i) $\mathcal{B}(\text{N-DBF})$, (ii) $\mathcal{B}(\text{ZF-DBF})$. To optimize PD in both (i) and (ii), similar to [8], we adopt the water-filling (WF) algorithm discussed above. Option (i) has a very low complexity since only limited computations are required per block for the WF algorithm and for computing the SINRs from the effective channels to select proper MCSs. Option (ii),

on the other hand, requires additional computations per block to compute the ZF-DBF precoding.

4.2.2 Proposed Heuristic Solutions

We propose RRM heuristics for both ZF-DBF and N-DBF cases (called $\mathcal{H}(\text{N-DBF})$ and $\mathcal{H}(\text{ZF-DBF})$ resp.). Our proposed heuristics employ smarter beam and user selections than the ones used in the benchmarks. The building blocks of our heuristics are illustrated in Fig. 4.7. The beam selection (Algorithm 1) is the same for both heuristics and is called once per time slot because of ConB. The user selection is called every block and is the same for both heuristics. PD differs for the two heuristics. Specifically, given the beam set ℓ_H for a time slot, the following steps are performed in parallel for each block in a time slot. First, we select UEs by Algorithm 2 for each block q . Given the user set $\mathcal{Z}_H^q$ for each block, we perform ZF-DBF per block for $\mathcal{H}(\text{ZF-DBF})$ (for $\mathcal{H}(\text{N-DBF})$ this step is skipped). Then, we perform PD for each block. For $\mathcal{H}(\text{ZF-DBF})$, we found that performing PD with WF works well but this is not the case for $\mathcal{H}(\text{N-DBF})$. Hence, we propose a new PD scheme that we call interference-aware WF (IAWF) (please see Algorithm 3) for that case. Note that there is no point applying IAWF to $\mathcal{H}(\text{ZF-DBF})$ since there is no interference after ZF precoding. We explain Algorithms 1, 2 and 3 next.

Our beam set selection and user selection are inspired by the heuristic proposed for the case with single frequency subchannel in [8]. In that case, there is no need to do beam selection and the whole focus is on user selection. The heuristic in [8] computes a coefficient for each UE and chooses the L UEs (recall that $L = \min(K, |\mathcal{I}_s|)$) having the largest coefficients while ensuring that the preferred beams of the selected UEs are distinct. We consider the general case where there are multiple frequency subchannels and hence, the need for beam selection and the enforcement of constraint ConB. Therefore, in our heuristics, in addition to defining UE coefficients, we propose the notion of beam coefficients. Defining both beam and UE coefficients enables us to perform beam and user selections in the multi-channel scenario.

To perform beam selection in a time slot, we will first compute beam coefficients and then select the L beams with the highest coefficients. Specifically, we first compute a coefficient per UE per block. To do so for block q , we assume that all UEs are selected in a PRB and distribute the power among UEs using WF with $|g_{q,u,u}^{\text{eff}}|^2$ as the gain and $1/R_u$ as the weight of UE u for all $u \in \mathcal{U}$. Note that $g_{q,u,u}^{\text{eff}}$ is the effective signal channel gain of UE u in block q defined in Section 3.1.2. It is simplified from $g_{q,u,m}^{\text{eff}}$ because m is fixed as one when there is only one matched BS beam per UE. Then, for each UE, we compute

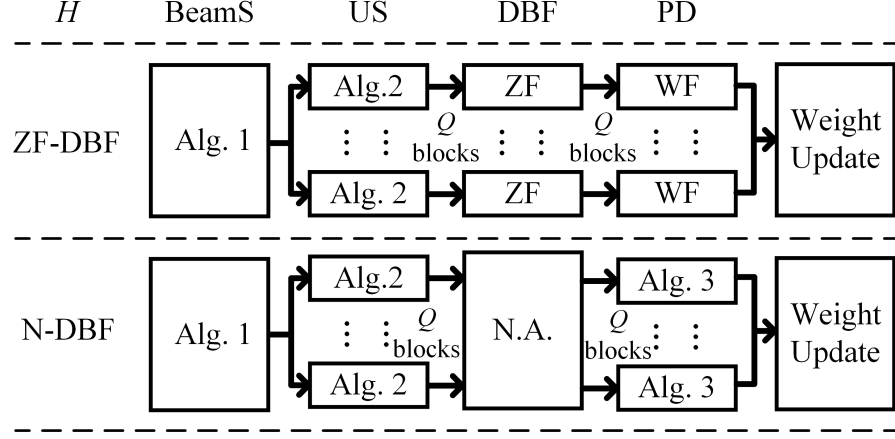


Figure 4.7: Building blocks of our online heuristics for a time slot (BeamS: beam selection, US: user selection, Alg.: Algorithm, N.A.: not available)

its SNR (i.e., we neglect interference) and map it to a data rate based on the MCSs. The obtained data rate is used as the coefficient for the UE in block q . The UE coefficients in [8] are computed assuming that the full power is allocated to a user and Shannon-based rate function. The way we compute the coefficients creates a competition in the power domain among the UEs. Also, since we employ practical MCS-based function instead of Shannon-based rate function, a UE coefficient will be zero if the SNR of the UE falls below the minimum MCS threshold.

Next, we use the UE coefficients to compute the beam coefficient c_b for each beam $b \in \mathcal{I}_s$ as $c_b = \sum_{q \in \mathcal{Q}} c_{u_{b,q}^*}^q$, where $u_{b,q}^* = \arg \max_{u \in \mathcal{U}(b)} c_u^q$. In words, to compute beam coefficient c_b , we first find, for each block q , the candidate UE $u_{b,q}^*$, that has the largest c_u^q among all the UEs whose preferred beam is b . Then, we sum up such candidate UE coefficients over $q \in \mathcal{Q}$ to obtain c_b . We then select the L beams with the highest values of beam coefficients for that time slot. Let ℓ_H denote this beam set.

User selection is performed by Algorithm 2 in each block q of a time slot given ℓ_H so that for each selected beam b^* in ℓ_H , we select UE $u_{b^*,q}^*$ for block q where $u_{b,q}^*$ was defined when defining beam coefficients above. Note that we do not select UEs with a coefficient of zero. Note also that using an MCS-based rate function in Algorithm 1 allows the coefficient to be zero if the SINR is too low to achieve a positive rate. This will avoid wasting power on UEs with low SINR. In comparison, using Shannon-based rate function will always lead to a positive rate if power is distributed to a UE. Our results show that considering practical MCS levels while calculating the coefficient improves the performance by about 16%.

Algorithm 1 Beam Selection at the beginning of a time slot

- 1: **for** $q = 1 : Q$ **do** ▷ Compute UE coefficients
 - 2: Assume all the UEs are selected in the q -th block. Distribute power to UEs using WF with $|g_{q,u,u}^{\text{eff}}|^2$ as the gain and $1/R_u$ as the weight of UE u for all $u \in \mathcal{U}$. For all $u \in \mathcal{U}$, calculate the SNR of UE u and map it to rate r_u^q using the piece-wise constant rate function.
 - 3: For all $u \in \mathcal{U}$, set c_u^q to r_u^q/R_u .
 - 4: **end for**
 - 5: **for** $b \in \mathcal{I}_s$ **do** ▷ Compute beam coefficients
 - 6: Compute $c_b = \sum_{q \in \mathcal{Q}} c_{u_{b,q}^*}^q$, where $u_{b,q}^* = \arg \max_{u \in \mathcal{U}(b)} c_u^q$.
 - 7: **end for**
 - 8: Rank beams in $\mathcal{I}_s$ in a descending order of c_b . Select the first L beams as ℓ_H , the beam set for this time slot.
-

Algorithm 2 User Selection in block q given ℓ_H

- 1: **for** $b \in \ell_H$ **do**
 - 2: Select UE u with the highest c_u^q (calculated in Algorithm 1) in $\mathcal{U}(b)$ as a UE in the user set z_H^q for this block. UE u will not be selected if $c_u^q = 0$.
 - 3: **end for**
-

Finally, we present in Algorithm 3, the PD solution for block q for $\mathcal{H}(\text{N-DBF})$, which we call Interference-Aware WF (IAWF). We first estimate the interference of each UE in z_H^q under the assumption in Algorithm 1 line 2. We assume all UEs in $\mathcal{U}$ are selected and allocated power by the WF algorithm. Note that this part has been computed by Algorithm 1 so we can use the calculated power directly. Then, we calculate the interference of UE u in block q as I_u^q with the calculated power. We will ignore the interference if it is from a UE preferring the same BS beam as UE u . With $I_u^q \forall u \in z_H^q$, we weigh their effective channel gain as $\frac{\sigma_{\text{PRB}}^2}{I_u^q + \sigma_{\text{PRB}}^2} |h_{\ell_H}^{q,u,u}(z_H^q)|^2$ and do WF algorithm based on the weighted effective channel gains. This will avoid distributing too much power to highly interfered UEs, which improves the overall rate performance.

Algorithm 3 IAWF in block q given z_H^q for $\mathcal{H}(\text{N-DBF})$

- 1: Based on the assumption in Algorithm 1 line 2, calculate the total interference of UE u from all other UEs except the ones sharing the same preferred BS beam as I_u^q for all $u \in \mathcal{U}$.
 - 2: Perform PD for UEs in z_H^q using WF with $\frac{\sigma_{\text{PRB}}^2}{I_u^q + \sigma_{\text{PRB}}^2} |h_{\ell_H}^{q,u,u}(z_H^q)|^2$ as the gain and $1/R_u$ as the weight of UE u .
-

Extension to Multiple Matched BS Beams Per UE

Similar to our offline study, we extend the online heuristics to support multiple matched BS beams per UE ($M > 1$). The following modifications are introduced to enable this extension.

First, in the beam selection stage, we consider all UEs matched with a given BS beam b , rather than only the UEs that prefer beam b . This is achieved by treating each UE as M separate virtual UEs, each associated with one of its matched BS beams, and computing M corresponding UE coefficients in Algorithm 1. Second, for each block, we check whether any UE is selected more than once due to different matched BS beams. If a UE appears more than once, we retain only the one with the best-matched BS beam. To implement these two steps, we update Algorithm 1 to Algorithm 4. Finally, we replace the UE coefficient c_u^q with $c_{u,m}^q$ in Algorithm 2, reflecting the extension from $M = 1$ to $M > 1$. In this case, selecting UE u with the highest $c_{u,m}^q$ corresponds to selecting UE u using its m -th matched BS beam.

Algorithm 4 Beam Selection at the beginning of a time slot for $M > 1$

- 1: **for** $q = 1 : Q$ **do** ▷ Compute UE coefficients
 - 2: Assume all the UEs are selected in the q -th block. Each UE is selected M times with different matched BS beams. Distribute power to UEs using WF with $|g_{q,u,m}^{\text{eff}}|^2$ as the gain and $1/R_u$ as the weight of UE u with its m -th matched BS beam for all $u \in \mathcal{U}$ and $m \in \{1, \dots, M\}$. For all $u \in \mathcal{U}$ and $m \in \{1, \dots, M\}$, calculate the SNR of UE u with its m -th matched BS beam and map it to rate $r_{u,m}^q$ using the piece-wise constant rate function.
 - 3: For all $u \in \mathcal{U}$ and $m \in \{1, \dots, M\}$, set $c_{u,m}^q$ to $r_{u,m}^q/R_u$.
 - 4: **end for**
 - 5: **for** $b \in \mathcal{I}_s$ **do** ▷ Compute beam coefficients
 - 6: Compute $c_b = \sum_{q \in \mathcal{Q}} c_{u_{b,q}^*, m_{b,q}^*}^q$, where $(u_{b,q}^*, m_{b,q}^*) = \arg \max_{u \in \mathcal{U}(b), m \in \{1, \dots, M\}} c_{u,m}^q$.
 - 7: **end for**
 - 8: Rank beams in $\mathcal{I}_s$ in a descending order of c_b . Select the first L beams as ℓ_H , the beam set for this time slot.
 - 9: **for** $q = 1 : Q$ **do**
 - 10: Check $u_{b,q}^*$ for all $b \in \ell_H$. If multiple beams have the same value of $u_{b,q}^*$ (same UE is selected multiple times with different beams), check the corresponding values of $m_{b,q}^*$. Keep the $c_{u_{b,q}^*, m_{b,q}^*}^q$ for beam b with the lowest value of $m_{b,q}^*$. Set the rest $c_{u_{b,q}^*, m_{b,q}^*}^q$ as zero.
 - 11: **end for**
-

4.2.3 Numerical Evaluations for Online RRM Study

This section presents the simulation results for our heuristics and benchmarks using the same channel model, ABF codebooks, and BA approach as in our offline study. Since our online algorithms are much faster than our offline solutions, we increase the number of realizations to $Z = 200$. All our results are for $B_b = 32$, $B_u = 8$, and $K = 8$ in this section (all other parameters consistent with Table 4.1).

We compare the performance of the heuristics and benchmarks in Fig. 4.8 where we also show the results for the offline cases. We compare the runtime of the heuristics and benchmarks in Fig. 4.9 for one time slot. We would like to highlight that the runtime of the offline solutions is in the order of multiple days which is much larger than the ones for the online algorithms.

Figs. 4.8 and 4.9 show that $\mathcal{H}(\text{ZF-DBF})$ can achieve at least 92.3% of the corresponding offline target performance and its performance is up to 338.6% of that of $\mathcal{B}(\text{ZF-DBF})$ at the cost of 53.9% higher maximum runtime. Meanwhile, $\mathcal{H}(\text{N-DBF})$ can achieve 90.2% of its target performance, and its performance is 193.4% of that of $\mathcal{B}(\text{N-DBF})$ with 124.9% higher maximum runtime (recall that $\mathcal{B}(\text{N-DBF})$ has a very low complexity). Fig. 4.8 also shows the impact of DBF. We observe that ZF-DBF significantly improves the performance compared to N-DBF in our heuristics (by 47.3%). The results confirm that our heuristics can strike a good balance between performance and runtime in both ZF-DBF and N-DBF scenarios.

With our heuristic, we further investigate the impact of M . Fig. 4.10 presents the system performance as a function of K for configurations with $U = 20$ and M up to 3. The maximum improvement, obtained by increasing $M = 1$ to $M = 2$, is 15% at $K = 20$. However, when $K < 6$, the performance gain becomes negligible due to the limited number of RF chains, which prevents the system from selecting even all of the UEs' preferred matched BS beams. Another observation from Fig. 4.10 is that increasing $M = 2$ to $M = 3$ does not yield a significant performance improvement. Having $M = 2$ is already sufficient, as the third matched BS beam for a UE typically does not offer a meaningful BF gain.

4.3 Conclusion

In this chapter, we have investigated downlink RRM in mmWave systems with codebook-based hybrid beamforming considering a practical multiple-channel scenario with a limited

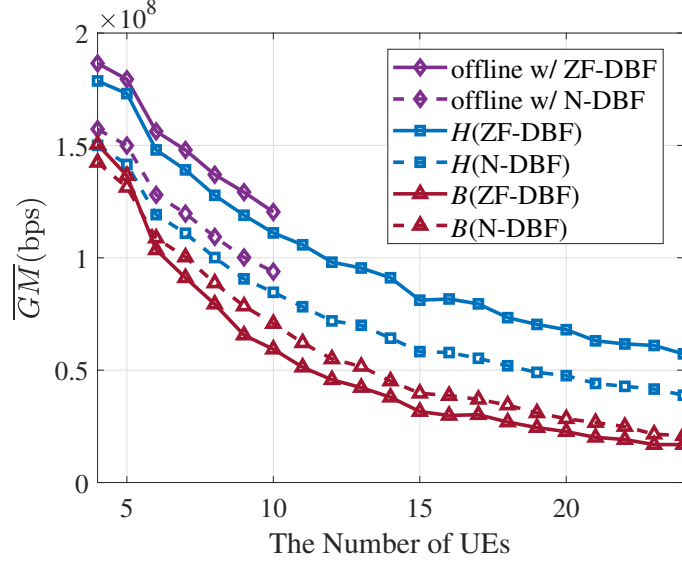


Figure 4.8: Performance of the proposed heuristics and benchmarks as a function of U with $K = 8$, $N_b = 128$, $N_u = 16$, $B_b = 32$ and $B_u = 4$

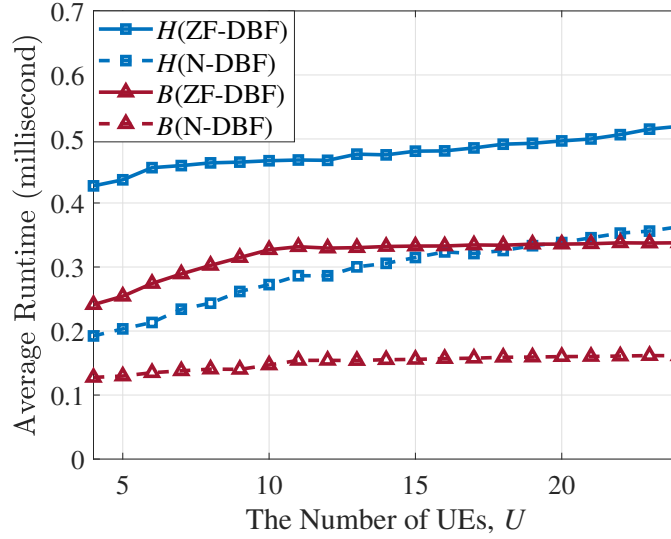


Figure 4.9: Average runtime of the proposed heuristics and benchmarks as a function of U with $K = 8$, $N_b = 128$, $N_u = 16$, $B_b = 32$ and $B_u = 4$.

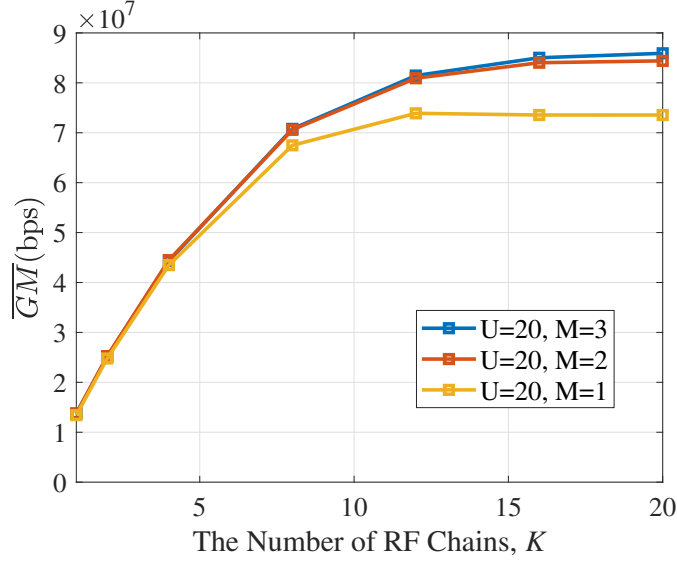


Figure 4.10: Impact of M on system performance with the proposed heuristics as a function of K with $U = 20$, $N_b = 128$, $N_u = 16$, $B_b = 32$ and $B_u = 4$

number of RF chains ($K < U$). This scenario which has been mostly overlooked in the literature, brings several new RRM challenges due to the constraint on beam selection, that require new formulations and results. By formulating and solving an offline joint RRM optimization problem for different cases, we found that ignoring the beam selection constraint overestimates the performance by 20%. We also show that good performance can be achieved even when K is smaller than U . We also found that non-equal power distribution and ZF-DBF can improve the system performance by 22% and 32% compared to EPD and N-DBF, respectively. Another finding is that having one more pair of preferred beams can improve the performance by 22% at the cost of CE overheads. The study further shows that power distribution using MCS-based rate function can improve the performance by about 10% compared with that using Shannon-based rate function (i.e., water-filling). Based on the insights gained from the offline study, we have proposed heuristic online algorithms that offer acceptable computational complexities and can reach 92% of the performance targets obtained from the offline study.

Using Shannon-based rate function in our offline optimization problem yields reasonable solutions because we have many degrees of freedom (i.e., variables) to play with. However, in an online setting, if we select beam and user sets in a non-careful way without considering practical MCS levels, water-filling (WF) performs poorly. This is why our online heuristic

uses WF on carefully chosen beam and user sets (i.e., so that the power per PRB is not spread too thin), and in that case, we can obtain good solutions.

Chapter 5

Uplink Study

In this chapter, we conduct our study on UL radio resource management under the assumption that each UE is equipped with one RF chain and matched with one pair of preferred beams (one beam at the BS and the other at the UE). We initially focus on the case with ZF-DBF at the BS, leaving the exploration of MMSE and N-DBF for future investigation. This is because the coupling between user selection and power allocation forces us to solve the RRM problem jointly, unlike on the DL where power optimization can be separated from RRM. In this case, the intra-cell interference (with MMSE and N-DBF) between UEs makes the RRM problem intractable.

We provide an offline RRM study for the uplink, when ZF-DBF is available, where beam selection, UE selection and PA are jointly considered in an optimization problem that we will solve to get an upper bound and a feasible solution. In contrast to the downlink RRM, where the BS allocates its power equally to channels in a time slot, each UE has a limited and independent power budget on the uplink. Therefore, our uplink RRM incorporates PA, which is to decide how to allocate the power of each UE to different channels in a time slot, as a crucial component. Note that there is no PD because at most one stream per user per PRB is assumed.

We also consider proportional fairness in our UL RRM study as discussed in [Section 3.4](#).

5.1 Joint RRM Optimization

In this section, we formulate and solve the joint uplink RRM problem that includes beam selection, user selection, PA, DBF and MCS selection, with an objective function based on

proportional fairness. This problem is a large-scale mixed-integer nonlinear program. Using relaxations and approximations, we reformulate it as a linear programming (LP) problem, from which we obtain an upper bound and a feasible solution to the original problem. We will show in Section 5.3, that this feasible solution is quasi-optimal by comparing it with the upper bound. It will serve as the target performance for the offline heuristic RRM scheme that we propose in Section 5.2.

The problem formulation and solution approach presented in this work are fundamentally different from those developed in [43] for the downlink. First, the power problem on the downlink is a per PRB power distribution problem for the selected streams, while the PA problem that distributes the BS power budget over all PRBs in a time slot is straightforward (each PRB receives the same share of the budget). On the uplink, there is no power distribution (since there is at most one stream per UE) and the PA is challenging. Second, to take into account the fact that more than one BS beams are matched with each UE, we need a different representation of beam and user selections. Specifically, we use a compact vector notation, where the position of a user in the user vector indicates for which beam this user is selected in a given PRB. Finally, because of the above differences, a new solution framework had to be developed, in which the RRM problem is formulated and solved jointly as a single integrated optimization problem, rather than decomposed into independent subproblems.

Following Section 3.4, our objective function considering PF in a time slot is

$$\max \sum_{u \in \mathcal{U}} \frac{\lambda_u}{R_u}, \quad (5.1)$$

where λ_u is the throughput of UE u within this time slot and R_u is the average throughput received by u over the past W time slots. As discussed in [43, 68], the natural performance metric for PF is the geometric mean of the rates, noted GM in the following, i.e., given a time horizon of T time slots,

$$GM = \left(\prod_{u=1}^U \frac{1}{T} \sum_{t=1}^T \lambda_u(t) \right)^{\frac{1}{U}}. \quad (5.2)$$

5.1.1 Problem Formulation

We now formulate the UL joint RRM optimization problem that includes beam selection, user selection, PA, DBF and MCS selection. The problem is formulated for a given time slot, and the inputs are: $K, C, N_u, N_b, M, (R_u), \mathcal{C}_b, \mathcal{C}_u, (g(u, b)), (h(u, m)), \mathcal{I}_s$, and $(\mathbf{g}_{q,u}^{\text{eff}})$.

Recall that beam selection consists in selecting $L = \min(K, |\mathcal{I}_s|)$ BS beams for each time slot (the selected beams remain unchanged in all PRBs of the time slot). We define

$$\mathcal{L} = \{\ell : \ell = [b_1, \dots, b_j, \dots, b_L], b_j \in \mathcal{I}_s, j \in \{1, \dots, L\}\} \quad (5.3)$$

as the set of all the possible beam vectors ℓ of L selected beams (where b_j is a beam index). Beams in ℓ are ranked in an ascending order of the beam indices for convenience. A beam index can appear at most once in a beam vector. For example, if $L = 4$ and $|\mathcal{I}_s| = 12$, $\ell = [3, 6, 8, 10]$ is a possible beam vector.

Beam selection consists in selecting one beam vector $\ell \in \mathcal{L}$ in the time slot. Then, user selection consists in selecting a user vector per PRB, given ℓ . Let the set of all distinct possible user vectors, given ℓ , be:

$$\begin{aligned} \mathcal{M}_\ell &= \{z_\ell : z_\ell = [u_1, \dots, u_j, \dots, u_L], u_j \in \{\mathcal{U}(b_j), 0\}, \\ &\text{s.t. } u_i = u_k \Rightarrow u_i = 0, \forall i \neq k \in \{1, \dots, L\}\}, \end{aligned} \quad (5.4)$$

where z_ℓ is a user vector in which u_j is the UE selected in the beam in position j in the beam vector ℓ (to be selected, u_j needs to be matched to that beam). If u_j is zero, no UE is selected for beam b_j . Recall that a UE can be selected at most once in z_ℓ .

Next, we introduce the beam selection binary variables α_ℓ indicating whether beam vector ℓ is selected in a given time slot. We have

$$\sum_{\ell \in \mathcal{L}} \alpha_\ell = 1, \quad (5.5)$$

$$\alpha_\ell \in \{0, 1\}, \forall \ell \in \mathcal{L}. \quad (5.6)$$

For each beam vector ℓ , we define next the user selection variables. Specifically, let $\beta_\ell^q(z_\ell)$ represents the fraction of PRBs within block q allocated to user vector z_ℓ , given that beam vector ℓ has been selected. Note that knowing the exact PRBs occupied by z_ℓ in a block is inconsequential since the channel coefficients remain the same in a block. We have

$$\sum_{z_\ell \in \mathcal{M}_\ell} \beta_\ell^q(z_\ell) \leq 1, \quad \beta_\ell^q(z_\ell) \geq 0, \quad \forall \ell \in \mathcal{L}, q \in \mathcal{Q}, \quad (5.7)$$

$$N_F \beta_\ell^q(z_\ell) \in \mathbb{Z}, \quad \forall \ell \in \mathcal{L}, q \in \mathcal{Q}, z_\ell \in \mathcal{M}_\ell, \quad (5.8)$$

where $\mathcal{Q} = \{1, 2, \dots, Q\}$. Eq. (5.8) indicates that each user vector must be allocated an integer number of PRBs.

With (α_ℓ) and $(\beta_\ell^q(z_\ell))$, beam selection and user selection are done. Next is PA. Let us introduce the PA variables, which determine how each UE's power budget, P_{UE} , is allocated across the PRBs on which the UE has been selected for transmission. Let $p_\ell^{q,u}(z_\ell)$ be the power allocated to each PRB in block q on which z_ℓ is selected. We have

$$0 \leq p_\ell^{q,u}(z_\ell) \leq P_{\text{UE}}, \forall u \in z_\ell, \ell, z_\ell, q, \quad (5.9)$$

$$p_\ell^{q,u}(z_\ell) = 0, \forall u \notin z_\ell, u = 0, \ell, z_\ell, q. \quad (5.10)$$

Unlike the DL where there is a per PRB power constraint linking all selected UEs in the PRB [43], on the uplink, the power constraint is per UE and per time slot, i.e.:

$$\sum_{q \in \mathcal{Q}} \sum_{z_\ell \in \mathcal{M}_\ell} N_F \beta_\ell^q(z_\ell) p_\ell^{q,u}(z_\ell) \leq P_{\text{UE}}, \forall u, \ell, \quad (5.11)$$

where $N_F \beta_\ell^q(z_\ell)$ is the number of PRBs allocated to vector z_ℓ in block q .

We next apply the digital BF step, i.e., ZF-DBF, and calculate $h_\ell^{q,u_k,u_j}(z_\ell)$ following Section 3.2, for $\forall u_k, u_j \in z_\ell$ given z_ℓ and ℓ . Note that $h_\ell^{q,u_k,u_j}(z_\ell) = 0$ for $u_k \neq u_j$ as a result of zero-forcing digital beamforming (ZF-DBF). Consequently, we use SNR instead of signal-to-interference-plus-noise ratio (SINR) when mapping to a UE's rate.

Assuming ℓ and z_ℓ are selected in a PRB within block q , and the allocated power to UE u is $p_\ell^{q,u}(z_\ell)$, we calculate the SNR of UE u in this PRB as

$$\gamma_\ell^{q,u}(z_\ell) = \frac{|h_\ell^{q,u,u}(z_\ell)|^2 p_\ell^{q,u}(z_\ell)}{\sigma_{\text{PRB}}^2}, \forall u \quad (5.12)$$

where σ_{PRB}^2 is the noise power on a PRB. The rate function $f(\cdot)$ maps $\gamma_\ell^{q,u}(z_\ell)$ to a rate. We use a rate function in (3.9) based on practical modulation and coding schemes (MCSs) [43, 44, 68] rather than Shannon-based rate function.

With $\gamma_\ell^{q,u}(z_\ell)$ and $f(\cdot)$, the UL data rate of UE u in a PRB within block q allocated to UE set z_ℓ is defined as

$$r_\ell^{q,u}(z_\ell) = f(\gamma_\ell^{q,u}(z_\ell)), \forall q, u, \ell, z_\ell. \quad (5.13)$$

Given α_ℓ , $\beta_\ell^q(z_\ell)$ and $r_\ell^{q,u}(z_\ell)$, we can compute the throughput λ_u of UE u in the time slot as

$$\lambda_u = \sum_{\ell \in \mathcal{L}} \alpha_\ell \sum_{q \in \mathcal{Q}} N_F \sum_{z_\ell \in \mathcal{M}_\ell} \beta_\ell^q(z_\ell) r_\ell^{q,u}(z_\ell), \forall u \in \mathcal{U}. \quad (5.14)$$

Next, we formulate the general RRM optimization problem, $\mathbf{\Omega}$, for a time slot. Given $f(\cdot)$, $\mathcal{U}$, $\mathcal{Q}$, $\mathcal{L}$, $\mathcal{M}_\ell$, R_u , W , N_F and $h_\ell^{q,u}(z_\ell), \forall z_\ell, \ell, q, u \in z_\ell$, we formulate

$$\begin{aligned} \mathbf{\Omega} : & \max_{\beta_\ell^q(z_\ell), \alpha_\ell, p_\ell^{q,u}(z_\ell), \gamma_\ell^{q,u}(z_\ell), r_\ell^{q,u}(z_\ell), \lambda_u} \sum_{u \in \mathcal{U}} \frac{\lambda_u}{R_u}, \\ \text{s.t. } & (5.5), (5.6), (5.7), (5.8), (5.9), (5.10), (5.11), (5.12), (5.13), (5.14) \end{aligned}$$

Problem $\mathbf{\Omega}$ is non-convex, non-linear and includes integer constraints.

5.1.2 Upper Bound of Joint RRM Optimization Problem

In order to provide an upper bound on the original problem $\mathbf{\Omega}$, we use the linear approximation in (3.11), following the method in [38], to upper-bound the nonconvex rate function $f(\gamma)$. By substituting approximation (3.11) in Eq. (5.13), we have

$$r_\ell^{q,u}(z_\ell) \leq a_i \gamma_\ell^{q,u}(z_\ell) + b_i, \forall i, q, u, \ell, z_\ell. \quad (5.15)$$

Using this linear upper bound, rather than the logarithmic upper bound in [43], is essential for enabling the subsequent variable transformations and to convexify the problem into a LP problem, as discussed later in this section.

We also remove the integer constraints (5.6) and (5.8) in $\mathbf{\Omega}$ to obtain the following relaxed problem $\mathbf{\Omega}^R$:

$$\begin{aligned} \mathbf{\Omega}^R : & \max_{\beta_\ell^q(z_\ell), \alpha_\ell, p_\ell^{q,u}(z_\ell), r_\ell^{q,u}(z_\ell), \lambda_u} \sum_{u \in \mathcal{U}} \frac{\lambda_u}{R_u}, \\ & \sum_{\ell \in \mathcal{L}} \alpha_\ell = 1, 0 \leq \alpha_\ell \leq 1, \forall \ell \in \mathcal{L}, \end{aligned} \quad (5.16a)$$

$$\sum_{z_\ell \in \mathcal{M}_\ell} \beta_\ell^q(z_\ell) \leq 1, \quad \beta_\ell^q(z_\ell) \geq 0, \quad \forall \ell \in \mathcal{L}, q \in \mathcal{Q}, \quad (5.16b)$$

$$0 \leq p_\ell^{q,u}(z_\ell) \leq P_{\text{UE}}, \forall u \in z_\ell, \ell, z_\ell, q, \quad (5.16c)$$

$$p_\ell^{q,u}(z_\ell) = 0, \forall u \notin z_\ell, u = 0, \ell, z_\ell, q, \quad (5.16d)$$

$$\sum_{q \in \mathcal{Q}} \sum_{z_\ell \in \mathcal{M}_\ell} N_F \beta_\ell^q(z_\ell) p_\ell^{q,u}(z_\ell) \leq P_{\text{UE}}, \quad \forall u, \ell, \quad (5.16e)$$

$$\lambda_u = \sum_{\ell \in \mathcal{L}} \alpha_\ell \sum_{q \in \mathcal{Q}} N_F \sum_{z_\ell \in \mathcal{M}_\ell} \beta_\ell^q(z_\ell) r_\ell^{q,u}(z_\ell), \forall u \in \mathcal{U}, \quad (5.16f)$$

$$r_\ell^{q,u}(z_\ell) \leq a_i \frac{|h_\ell^{q,u}(z_\ell)|^2 p_\ell^{q,u}(z_\ell)}{\sigma_{\text{PRB}}^2} + b_i, \forall i, u, \ell, q, z_\ell. \quad (5.16g)$$

Please note that Ω^R remains highly nonconvex due to two products of variables, one between the user selection and PA variables in Constraint (5.16e), and the other between beam selection, user selection, and PA variables in Constraint (5.16f). We introduce the following three changes of variables to convexify the problem:

$$\eta_\ell^{q,u}(z_\ell) \triangleq \alpha_\ell \beta_\ell^q(z_\ell) p_\ell^{q,u}(z_\ell), \quad \forall u, \ell, q, z_\ell, \quad (5.17)$$

$$\epsilon_\ell^{q,u}(z_\ell) \triangleq \alpha_\ell \beta_\ell^q(z_\ell) r_\ell^{q,u}(z_\ell), \quad \forall u, \ell, q, z_\ell, \quad (5.18)$$

$$\xi_\ell^q(z_\ell) \triangleq \alpha_\ell \beta_\ell^q(z_\ell), \quad \forall u, \ell, q, z_\ell. \quad (5.19)$$

The problem then becomes

$$\begin{aligned} \Omega^R : \quad & \max_{\eta_\ell^{q,u}(z_\ell), \epsilon_\ell^{q,u}(z_\ell), \xi_\ell^q(z_\ell), \alpha_\ell, \lambda_u} \sum_{u \in \mathcal{U}} \frac{\lambda_u}{R_u}, \\ & \sum_{\ell \in \mathcal{L}} \alpha_\ell = 1, 0 \leq \alpha_\ell \leq 1, \quad \forall \ell \in \mathcal{L}, \end{aligned} \quad (5.20a)$$

$$\sum_{z_\ell \in \mathcal{M}_\ell} \xi_\ell^q(z_\ell) \leq \alpha_\ell, \quad 0 \leq \xi_\ell^q(z_\ell) \leq 1, \quad \forall \ell \in \mathcal{L}, q \in \mathcal{Q}, \quad (5.20b)$$

$$\sum_{q \in \mathcal{Q}} \sum_{\substack{z_\ell \in \mathcal{M}_\ell \\ \text{s.t. } u \in z_\ell}} N_F \eta_\ell^{q,u}(z_\ell) \leq \alpha_\ell P_{\text{UE}}, \quad \forall u \in \mathcal{U}, \ell \in \mathcal{L}, \quad (5.20c)$$

$$0 \leq \eta_\ell^{q,u}(z_\ell) \leq P_{\text{UE}} \xi_\ell^q(z_\ell), \quad \forall u \in z_\ell, \ell, z_\ell, q, \quad (5.20d)$$

$$\eta_\ell^{q,u}(z_\ell) = 0, \quad \forall u \notin z_\ell, u = 0, \ell, z_\ell, q \quad (5.20e)$$

$$\lambda_u = N_F \sum_{\ell \in \mathcal{L}} \sum_{q \in \mathcal{Q}} \sum_{z_\ell \in \mathcal{M}_\ell} \epsilon_\ell^{q,u}(z_\ell), \quad \forall u \in \mathcal{U}, \quad (5.20f)$$

$$\epsilon_\ell^{q,u}(z_\ell) \leq a_i \frac{|h_\ell^{q,u,u}(z_\ell)|^2 \eta_\ell^{q,u}(z_\ell)}{\sigma_{\text{PRB}}^2} + b_i \xi_\ell^q(z_\ell), \quad \forall u, \ell, q, z_\ell, i.$$

Problem Ω^R then becomes a linear problem (LP) that can be efficiently solved with commercial solvers such as CPLEX, MINOS and Gurobi [69, 70]. Solving it will give us an upper bound on the original Problem Ω . However, due to the relaxations applied to the rate function and the integer constraints, this upper bound might not be directly feasible. In the next subsection, we present a method to construct a feasible solution using the values of α_ℓ and $\xi_\ell^q(z_\ell)$ obtained from solving Ω^R .

5.1.3 Feasible Solution to Joint RRM Optimization Problem

Reconstructing α_ℓ and $\beta_\ell^q(z_\ell)$

We first restore the change of variables by $\beta_\ell^q(z_\ell) = \xi_\ell^q(z_\ell)/\alpha_\ell$ if $\alpha_\ell \neq 0$ and $\beta_\ell^q(z_\ell) = 0$ if $\alpha_\ell = 0$. However, the (α_ℓ) 's and $(\beta_\ell^q(z_\ell))$'s do not necessarily satisfy Constraints (5.6) and (5.8) which were relaxed to get to Problem Ω^R . Therefore, we modify the solution to meet the constraints. Specifically, we first set $\hat{\alpha}_\ell = 1$ for the beam vector ℓ_f corresponding to the highest value of α_ℓ , and set $\hat{\alpha}_\ell = 0$ for all other ℓ . We then generate $\hat{\beta}_{\ell_f}^q(z_{\ell_f})$ by rounding down $\beta_{\ell_f}^q(z_{\ell_f})$ to satisfy the integer constraint. Then, we choose the $z_{\ell_f} \in \mathcal{M}_{\ell_f}$ such that the gap between $\beta_{\ell_f}^q(z_{\ell_f})$ and $\hat{\beta}_{\ell_f}^q(z_{\ell_f})$ is the largest, and increase $\hat{\beta}_{\ell_f}^q(z_{\ell_f})$ by $1/N_F$. We repeat the last step until $\sum_{z_{\ell_f} \in \mathcal{M}_{\ell_f}} \hat{\beta}_{\ell_f}^q(z_{\ell_f}) = 1$. $\hat{\alpha}_{\ell_f}$ and $\hat{\beta}_{\ell_f}^q(z_{\ell_f})$ are the reconstructed beam selection and user selection variables which satisfy the integral constraints.

Computing a Feasible Power Allocation

Note that the previous PA solution $p_{\ell_f}^{q,u}(z_{\ell_f})$ was calculated based on α_{ℓ_f} and $\beta_{\ell_f}^q(z_{\ell_f})$ and thus is no longer correct. Therefore, we solve Problem Ω_{LP}^R again given $\hat{\alpha}_{\ell_f}$ and $\hat{\beta}_{\ell_f}^q(z_{\ell_f})$ to get a feasible PA solution $\bar{p}_{\ell_f}^{q,u}(z_{\ell_f})$. After that, we use the MCS-based greedy algorithm proposed in [44] to improve the solution by taking the real MCS function into account.

We have now shown how to compute $\hat{\alpha}_{\ell_f}$, $\hat{\beta}_{\ell_f}^q(z_{\ell_f})$ and $\bar{p}_{\ell_f}^{q,u}(z_{\ell_f})$, which, together, form a feasible solution to the original Problem Ω . We will validate the quasi-optimality of this solution by comparing numerically its performance to the upper bound in Section 5.3. However, solving Problem Ω^R becomes impossible when the number of users, U , is larger than 10 (especially when the number of RF chains is large) due to the large number of variables. For instance, with 20 UEs, there can be 2^{20} user selection variables and 2^{20} PA variables per block. However, to provide meaningful engineering insights, we need to study systems with a larger number of users than 10 and hence, we propose an offline heuristic, in the next section, to address the computation of the performance for these systems. We will validate it by comparing its solutions to the quasi-optimal solutions for $U \leq 10$ and then provide engineering insights using it. We end the next section with a proposal for an online heuristic RRM solution obtained by simplifying the offline version. In Section 5.3, we will present our numerical results in a mmWave setting and show that the online heuristic has a low complexity suitable for real-time system operation, while achieving performance comparable to the offline heuristic.

5.2 Heuristic Solutions to RRM

In this section, we first present our offline heuristic solution, developed to deliver target performance and engineering insights in scenarios with a large number of UEs, where solving the optimization problem Ω^R becomes computationally infeasible due to the explosion in the number of variables.

5.2.1 Offline Heuristic

We have observed that simple approaches, such as round-robin beam and user selections combined with equal PA where each UE allocates equal power to its allocated PRBs, do not approach the target performance obtained from optimization. Thus, we propose an offline heuristic that achieves a performance close to the target and can be executed efficiently even with a large number of UEs.

Our offline heuristic is based on four steps, the two main steps being Step 1 and Step 3. Step 1 optimizes user selection/PA performed per beam *in isolation* (while ignoring the other beams), considering all users that have matched with this beam. Step 3 is a “load-aware” beam selection strategy where the K beams with the largest loads are selected.

Step 1: Per-beam User Selection and Power Allocation Optimization

For every $b \in \mathcal{I}_s$ and the set $\mathcal{U}(b)$ of users that are matched to b , we schedule users in a PF fashion, ignoring the other beams and hence the need for DBF. The per time slot scheduling problem can be written as problem $\mathbf{S}_b$, where we have removed the index for b

on all variables except in $\mathcal{U}(b)$ for ease of notation.

$$\mathbf{S}_b : \max_{m_u^q, p_u^q, r_u^q, \lambda_u} \sum_{u \in \mathcal{U}(b)} \sum_{q \in \mathcal{Q}} \frac{m_u^q r_u^q}{R_u},$$

$$\sum_{u \in \mathcal{U}(b)} m_u^q \leq N_F, \quad \forall q \in \mathcal{Q} \quad (5.21a)$$

$$m_u^q \in \mathbb{Z}, \quad \forall u \in \mathcal{U}(b), q \in \mathcal{Q} \quad (5.21b)$$

$$0 \leq p_u^q \leq P_{\text{UE}}, \quad \forall u \in \mathcal{U}(b), q \in \mathcal{Q} \quad (5.21c)$$

$$\sum_{q \in \mathcal{Q}} m_u^q p_u^q \leq P_{\text{UE}}, \quad \forall u \in \mathcal{U}(b) \quad (5.21d)$$

$$r_u^q \leq a_i \frac{G_u^q p_u^q}{\sigma_{\text{PRB}}^2} + b_i, \quad \forall i, u, q. \quad (5.21e)$$

where m_u^q is the number of PRBs in block q allocated to UE u , p_u^q is the power that UE u allocates to a PRB in block q , and $G_u^q = g_{q,u,u,g(u,b)}^{\text{eff}}$ is the channel gain between UE u and BS beam b . This problem is a simplified version of Problem Ω^R and hence can be relaxed and transformed as discussed earlier for Ω^R to yield an LP problem that is small in size since the number of users matched to a beam is usually small (except for M large).

The outputs of this per-beam optimization are a selected UE (or no UE) per PRB, drawn from the UEs matched with the beam, the power allocated to a UE in its allocated PRBs, and a weighted sum rate per beam. We use this weighted sum rate as the “load” of beam b .

Step 2: Only necessary if $M > 1$

Note that, if $M > 1$, a UE might be selected multiple times in a PRB (on different beams). In that case, we say that the UE is in a conflict. These conflicts have to be removed to provide a feasible solution over all beams. If a UE is in a conflict in a given PRB, we keep it on the beam providing the highest channel gain. Then, we try to assign different PRBs in the same block to the UE in the beams in conflict. For example, if UE a has been removed from beam i in the PRB. We first check if there is any empty PRB in beam i in this block, which is not in conflict with UE a . If yes, we move the UE to this PRB in beam i . If no, we check each of the other PRBs in the block in beam i one by one. As soon as we find one that is occupied by say UE b such that switching UE a and UE b gives feasible user sets in both PRBs (i.e., no new conflicts are created), we switch a and b . If there is no such b , we remove UE a . This ensures that each UE is selected at most once in

a PRB and prevents duplicate counting of a UE's contribution to the weighted sum rate calculations across beams.

Once we have removed all conflicts, we need to recompute the PA for each beam for which there were removed users to obtain the revised loads for these beams.

Step 3: Load-aware Beam Selection

We select the L beams with the highest weighted sum rates (load) as the beam set vector for the time slot. Together with the selected UE per PRB from per-beam optimization for each of the L beams, we build the user set vectors for all PRBs in the time slot. Recall that the beam set and user set are vectors because beams are one-to-one mapped to UEs in order.

Step 4: DBF and Final Power Allocation

With the selected beam set vector and the corresponding per-PRB user set vectors, we apply ZF-DBF per PRB on the selected UEs to mitigate the inter-beam interference. Since the PA calculated in Step 1 is based on effective channel gains without ZF-DBF, we need to update the PA for each UE. We follow Section 5.1.3 for the PA update. Finally, the SNR of each UE is mapped to a rate according to the MCS-based rate function.

In conclusion, we first run per-beam user selection and PA optimization and load-aware beam selection to construct a feasible beam set vector for a time slot and a user set vector for each PRB in the time slot. Next, ZF-DBF is applied per PRB to control the interference between selected UEs. Based on the resulting channel gains after ZF-DBF, per-user PA is updated using the method described in Section 5.1.3. By dividing the joint optimization in Section 5.1 into separate per-beam optimizations, we significantly reduce the computation time from months to hours to obtain a figure such as Fig. 4.

5.2.2 Online Heuristic

The offline heuristic serves as an effective tool for system planning with a large number of UEs. However, it is not suitable for online deployment since it relies on per-beam user selection and power optimization, i.e., problem $\mathbf{S_b}$. In order to design an online heuristic, we need to simplify the offline heuristic and, in particular, replace solving $\mathbf{S_b}$ for each beam b with matched users by an online scheduler. Several papers have addressed the

problem of uplink scheduling in a single cell but almost none incorporates the practical MCS-based rate function, instead relying on a Shannon-based one. Without considering practical MCS levels, a UE may be selected on too many channels and spread its power thinly among them, resulting in some channels having SINR values too low to achieve a positive rate. Therefore, we first adapt the iterative algorithm developed in [24] to the MCS-based rate function and see an improvement up to 20%, but it did not perform well when the number of UEs is small, which is commonly the case when focusing on a single beam. Hence, we propose next, a per-beam scheduler (PBS) that extends the algorithm by incorporating additional search steps considering practical MCS levels.

Per-beam Scheduler, PBS

The aim of PBS is to select, in a given BS beam in $\mathcal{I}_s$, at most one UE in each PRB of a time slot from the UEs who are matched with this BS beam. Specifically, PBS is iterative and selects UEs in a greedy way, taking their weights into account. In each step, PBS considers each UE one at a time. It gives the UE under consideration its highest gain PRB among those not yet allocated, and calculates, given all the PRBs allocated to that UE so far, the weighted rate for that UE in the time slot assuming EPA. Then, PBS finds the UE whose weighted sum rate (sum over its allocated PRBs) increases the most or decreases the least if no UE sees an increase in weighted rate and gives it its preferred PRB. The fact that we allow the increase to be negative is what makes the algorithm perform well when the number of users is small when compared to the original method in [24], which terminates immediately upon encountering a negative rate increment. The algorithm stops considering a UE when the UE has seen X steps with a negative change in the weighted rate. The algorithm also stops once there are no more PRBs to allocated or no UEs to consider. The improvement arises from the fact that while adding a single PRB may not increase a UE's rate, adding multiple PRBs can collectively yield a gain, considering the step-wise rate function. Details are presented in Algorithm 5.

To further accelerate Algorithm 5, we can allocate a block of PRBs (recall that there are Q such blocks) to a UE in each step. Once allocating a full block no longer yields performance improvement, we switch to adding one PRB per step. In this case, if we see a sum rate decrease while adding a block of PRBs, we increase the stop flag by N_F .

Fig. 5.1 compares PBS with the solution in [24] in a single-user scheduling scenario as well as with the optimal solution to $\mathbf{S}_b$ labeled PBO in the figure. We assume a flat fading channel based on the 3GPP urban macro channel model [52], with a cell radius of 800 meters. Each UE transmits at a power level of 10 dBm with a single antenna. The carrier frequency is set to 3.5 GHz, and the system operates over 100 frequency channels, each with

Algorithm 5 PBS on beam b , Inputs: $\mathcal{U}(b)$, X , R_u , P_{UE} , $g_{q,u,u,m}^{\text{eff}}$ s.t. $h(u, m) = b$

- 1: Set the stop flags of all UEs in $\mathcal{U}(b)$ to zero and all PRBs are empty.
 - 2: **while** there are any remaining PRBs to allocate in this time slot **and** UEs whose stop flag is no larger than X **do**
 - 3: **for** each UE in $\mathcal{U}(b)$ whose stop flag $\leq X$ **do**
 - 4: Select the best remaining PRB ($\max_c g_{q,u,u,m}^{\text{eff}}$).
 - 5: Calculate the sum rate of this UE in this time slot assuming EPA and that this PRB is allocated to it.
 - 6: Increase its stop flag by one if the sum rate of this UE decreases.
 - 7: **end for**
 - 8: Select the UE whose weighted sum rate increases the most and allocate its best remaining PRB to it.
 - 9: Record the user selection of beam b and corresponding weighted rate in each PRB if the weighted sum rate in this time slot is the highest over the past.
 - 10: **end while**
-

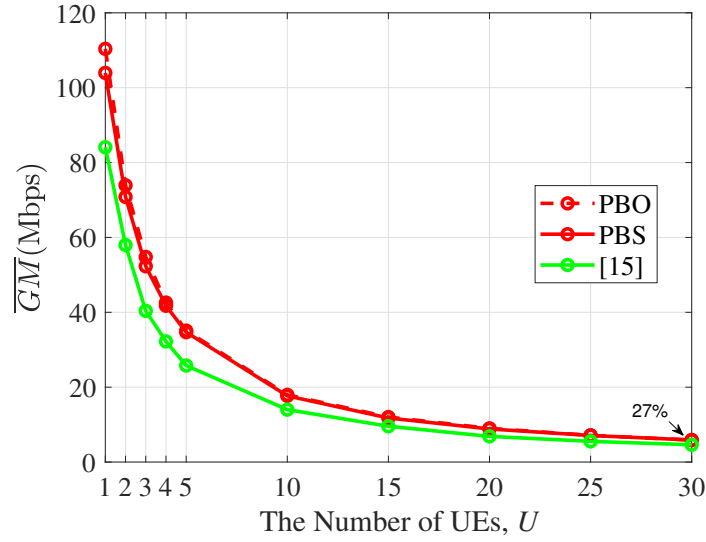


Figure 5.1: Performance comparison between PBO, PBS and the solution in [24] as a function of U .

a bandwidth of 360 kHz. The BS is equipped with 64 antennas. Performance results are averaged over 100 independent realizations, where UE locations and channel parameters are randomly generated for each realization. As shown in the figure, PBS improves the performance by at least 22%.

Fig. 5.1 also illustrates that the performance gap between PBO and PBS is between 1% and 6% relative to the performance of PBO. Hence, due to its low computational complexity, PBS serves as an efficient alternative to per-beam optimization to compute the load in the online heuristic.

The Online Heuristic

We are now ready to describe our online heuristic that mimics the offline one. PBS is used in Step 1, We do not modify Step 2 and Step 3 described above. In Step 4, we replace the last PA by one based on WF [67]. Finally, the resulting SNRs are mapped to achievable rates using MCS-based rate function. We will compare our online heuristic with our offline heuristic in the next section. Since Fig. 5.1 shows a negligible gap between PBO and PBS, this comparison reveals the impact of considering MCS-based rate function instead of Shannon-based rate function during power allocation.

5.3 Numerical Evaluations

In this section, we first present numerical results for both the upper bound and the feasible solution of our joint RRM optimization problem to validate the quasi-optimality of our feasible solution for $U \leq 10$. We also analyze the impact of the number of RF chains at the BS (K) and the number of matched BS beams per UE during CE (M) on system performance based on the feasible solution. Next, we show the effectiveness of our offline RRM heuristic by comparing it with the quasi-optimal feasible solution in scenarios with a small number of UEs. Thanks to its lower complexity, the offline heuristic can be used for scenarios with more UEs, offering engineering insights for practical deployments. Finally, we evaluate the performance and complexity of our online RRM heuristic by comparing it with the offline heuristic. All our numerical results are obtained in a mmWave setting.

5.3.1 System Parameters

We consider the uplink of a small cell network with a radius of $r = 75$ meters. UEs are uniformly distributed within this area, excluding a 6-meter radius around the BS where

standard path loss models are no longer valid. The BS is elevated at a height of 10 meters. Each UE is equipped with a transmit power budget of $P_{UE} = 7$ dBm, and the noise power spectral density is set to $N_0 = -174$ dBm/Hz. Both the BS and UEs employ uniform linear antenna arrays, comprising $N_b = 128$ and $N_u = 16$ antennas, respectively, with half-wavelength spacing. The system operates at a carrier frequency of $f_c = 28$ GHz with a total bandwidth of $BW = 100$ MHz, divided into $C = 132$ channels, each with a bandwidth of $B_c = 720$ kHz. The time slot duration is 0.25 msec. We adopt the MCS-based rate function given in [43, 44] and use the parameters in Table 3.1 for its approximation (3.11).

Analog beamforming (ABF) codebooks for both the BS and UEs are generated following the method in [27, 43]. We generate 32 beams (codewords) in the BS codebook and 4 beams (codewords) in the UE codebook.

5.3.2 Channel Model and Parameters

We adopt a wideband channel model which incorporates multipath effects, i.e., scattering clusters and spatial paths in each cluster [43, 64]. The MIMO channel between BS and UE u at block q is defined as the following $N_u \times N_b$ matrix:

$$\mathbf{G}_{q,u} = \frac{1}{\sqrt{N_{\text{path}}}} \sum_{d=1}^{N_{\text{cluster}}} \sum_{l=1}^{N_{\text{path}}} g_{d,l}^u e^{-j2\pi\tau_{d,l}^u f_q} \mathbf{a}_{\text{RX}}(\phi_{d,l}^u) \mathbf{a}_{\text{TX}}^H(\theta_{d,l}^u), \quad (5.22)$$

where N_{cluster} is the number of clusters, N_{path} is the number of paths in each cluster, $g_{d,l}^u = \sigma_d^u \sqrt{\kappa_{d,l}^u} e^{j\varphi_{d,l}^u}$ is the coefficient of the path l in cluster d in which σ_d^u models path-loss and shadowing, $\kappa_{d,l}^u$ models the fraction of the signal power carried by the l -th path within the d -th cluster and $\varphi_{d,l}^u \sim \mathcal{U}(0, 2\pi)$ models the phase of the l -th path within the d -th cluster, $\tau_{d,l}^u = \tau_{0d}^u + \tau_{1d,l}^u$ represents the delay of the path l in cluster d in which τ_{0d}^u is the group delay of the d -th cluster and $\tau_{1d,l}^u$ is the relative path delay of the l -th path in the d -th cluster, f_q is the center frequency of block q , $\mathbf{a}_{\text{TX}}(\cdot)$ and $\mathbf{a}_{\text{RX}}(\cdot)$ are the array response vectors at the BS and UE u , and finally $\theta_{d,l}^u$ and $\phi_{d,l}^u$ represent the angle of arrival (AoA) and the angle of departure (AoD) corresponding to the l -th path in cluster d of the channel. The generation of these channel parameters follows Section IV-F in [43]. Based on $\mathbf{G}_{q,u}$, BA and CE are performed as follows. Note that $g_{q,n,u,m}^{\text{eff}} = (\mathbf{v}_{l_u})^H \mathbf{G}_{q,u} (\mathbf{w}_{h(u,m)})$. Recall that $g_{q,n,u,m}^{\text{eff}}$ represents the effective channel gain between UE u and the BS using $\mathbf{w}_{h(u,m)}$ if $n = u$ or the effective channel interference between UE n and the BS due to the transmission of UE u to the BS using $\mathbf{w}_{h(n,m)}$ if $n \neq u$, for $\forall m \in \{1, \dots, M\}$.

We assume that the large-scale fading σ_d^u remains constant over T time slots and that BA is performed once at the beginning of these T time slots to capture the large-scale

fading characteristics. Specifically, for each UE, all combinations of a BS beam and a UE beam are evaluated, and the M pairs that best compensate for the large-scale fading (i.e., yield the M highest effective channel gains) are selected for that UE [9, 55].

We assume one CE block consists of $N_F = 6$ channels in frequency, corresponding to $Q = 22$ blocks in the whole bandwidth. CE is performed Q times per time slot for each block. The value of N_F was selected based on the channel model, to ensure that the channel coefficients can be considered constant within each block.

5.3.3 Evaluation Criterion

We define a network realization $\gamma(U)$ as the uniform placement of U users in the cell where, for each UE, the MIMO channel parameters for each M pairs of beams where the UE beam is fixed for each UE and there are M matched BS beams per UE are given for each Q blocks for $T = 100$ time slots. For each realization $\gamma(U)$, RRM is performed for each of the T time slots to obtain $GM(\gamma(U))$ (see Eq. (5.2)). RRM results need to be obtained for T time slots because $R_u(t)$ and the channel are time-varying. To evaluate the performance for a given value of U , we generate $Z = 50$ realizations and let $\overline{GM}(U) = \frac{1}{Z} \sum_{z=1}^Z GM(\gamma_z(U))$ denote the arithmetic mean of the GM of each realization.

Finally, the moving average parameter W is set to 10. The initial value of R_u is set to 2 for all users.

5.3.4 Numerical Evaluations for Joint RRM Optimization

We start by analyzing the quality of our feasible solution to the joint RRM optimization Problem Ω . Fig. 5.2 compares the performance of the feasible solution proposed in Section 5.1.3 with the upper bound presented in Section 5.1.2 as a function of the number of BS RF chains K . The gap between the feasible and upper-bound performances is small, at most 3% of the upper bound, confirming the quasi-optimality of the proposed feasible solution. Moreover, the performance begins to saturate way before reaching $K = U$. For instance, when $U = 10$, setting $K = 6$ already achieves 88% of the performance observed at $K = U$. We will further investigate the impact of K using the offline heuristic in scenarios with larger values of U .

Next, we analyze the impact of the number, M , of matched BS beams per UE on the performance of the feasible solution. Fig. 5.3 depicts the system performance as a function of K for $M = 1$ and $M = 2$. We observe a maximum gain of 17% at $K = 2$ and $U = 10$

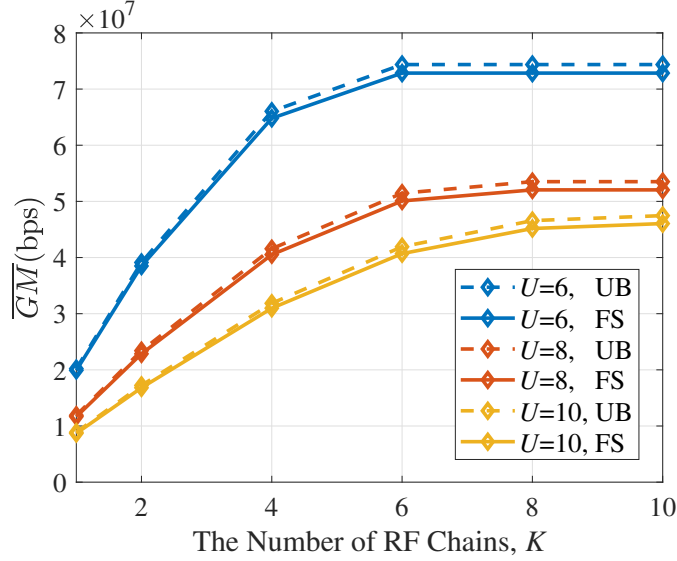


Figure 5.2: Quality of the feasible solution for $U = 6, 8, 10$ and $M = 1$ (UB: upper bound, FS: feasible solution).

by using $M = 2$ instead of $M = 1$. With larger K (e.g., $K = 6$ and $U = 6$), doubling M can increase the performance by 12%. Recall that this performance improvement is at the cost of increased channel estimation overhead. Since increasing U , K and M substantially increases the number of variables in the joint optimization, solving the problem becomes intractable for $U > 10$, $K > 4$ and $M > 2$. Therefore, we will explore the impact of M with larger U and K using the offline heuristic in the next subsection.

5.3.5 Numerical Evaluation of the Offline Heuristic

We first validate our offline heuristic by comparing it to the quasi-optimal solution. Fig. 5.4 presents the performance of our offline heuristic as a function of K , alongside the target performance obtained from the quasi-optimal solution. Our offline heuristic achieves at least 96% of the target performance for $U = 8$ and 93% for $U = 10$. Notably, the runtime for obtaining the heuristic curve with $U = 10$ is one hour and 46 minutes, compared to the 38 days required for computing the corresponding target performance via joint RRM optimization. The performance and computational efficiency of the offline heuristic allow us to obtain engineering insights for larger systems. This makes the proposed offline heuristic well-suited for system planning in large-scale deployments.

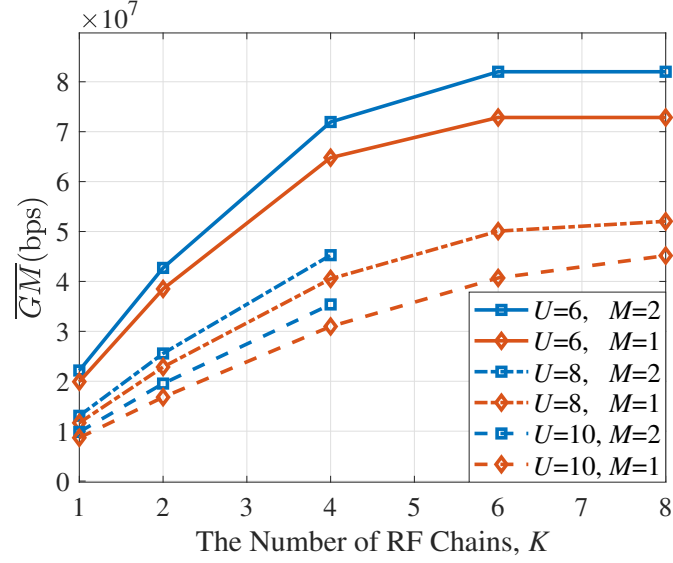


Figure 5.3: Impact of the number, M , of matched BS beams per UE on the performance of the feasible solution for $U = 6, 8, 10$.

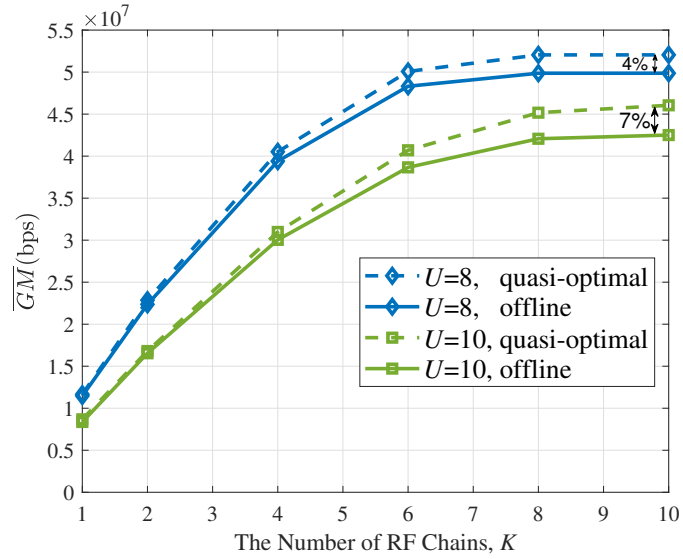


Figure 5.4: Performance of the offline heuristic as a function of K for $U = 8, 10$ and $M = 1$.

With our offline heuristic, we can investigate the impact of M under large-scale scenarios with large U and K . Fig. 5.5 presents the system performance as a function of K for configurations with up to $U = 30$, $K = 30$, and $M = 3$. The maximum improvement obtained by increasing $M = 1$ to $M = 2$ is 19% at $U = 30$ and $K = 1$. The performance gain diminishes to 12% as K increases to 30 for $U = 30$ and 8% for $U = 10$ and $K = 10$. When K is small, only a limited number of UEs can be scheduled within a time slot. Increasing M allows more UEs to be selected, thereby enabling more effective utilization of uplink power. As a result, the performance gain from increasing M is more significant when K is small. However, when comparing $M = 2$ with $M = 3$, the maximum observed gain is 7% at $U = 30$ and $K = 1$, while the improvement becomes negligible as K increases for both $U = 10$ and $U = 30$. This is because the third (and subsequent) best-matched BS beams typically offer much weaker channel gains for a UE than the first two and thus contribute minimally to performance, even with additional flexibility in user selection.

Another observation from Fig. 5.5 is that for $U = 30$, setting $K = 10$ achieves 90% of the performance at $K = U$ when $M = 1$ and 88% when $M = 2$ or $M = 3$. It further validates that good performance can be achieved even when K is significantly smaller than U . This is because, as K increases, allowing more UEs within the same PRB, the channel gains after ZF-DBF tend to decrease due to reduced spatial separation. This degradation leads to lower per-UE rates, thereby diminishing the marginal performance gains from further increasing K .

5.3.6 Numerical Evaluations of the Online Heuristic

We begin by comparing the performance of the online heuristic against that of the offline heuristic. We set the parameter X (number of extra steps in the search) of PBS to $X = N_F + 1 = 7$. Fig. 5.6 shows the performance of the online heuristic as a function of K for $U = 10$ and $U = 30$, along with the offline heuristic as the target. The performance of our online heuristic is at least 92% of the target performance for $U = 10$ and at least 94% for $U = 30$. The performance gap primarily arises from WFPA using Shannon-based rate function in the online heuristic, as opposed to the optimization using MCS-based rate function in the offline heuristic, because Fig. 5.1 shows that PBS is as good as PBO. Although WF does not achieve a similar performance as PA optimization, its millisecond-level runtime makes it suitable for operation.

We further evaluate our online heuristic with fixed values of K and M , examining its performance as a function of the number of users, U . Based on insights from the offline heuristic results, we consider $K = 2, 8, 16$ and $M = 1, 2$. The number of realizations is

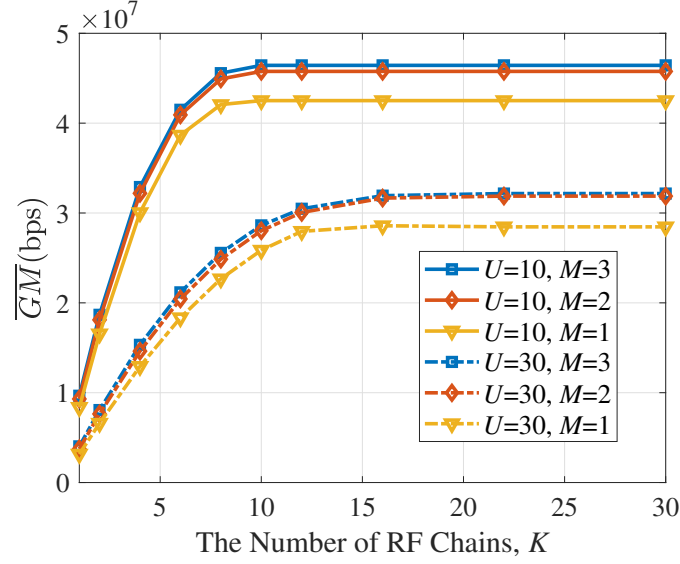


Figure 5.5: Impact of the number of matched BS beams per UE on offline heuristic performance for $U = 10$ and 30.

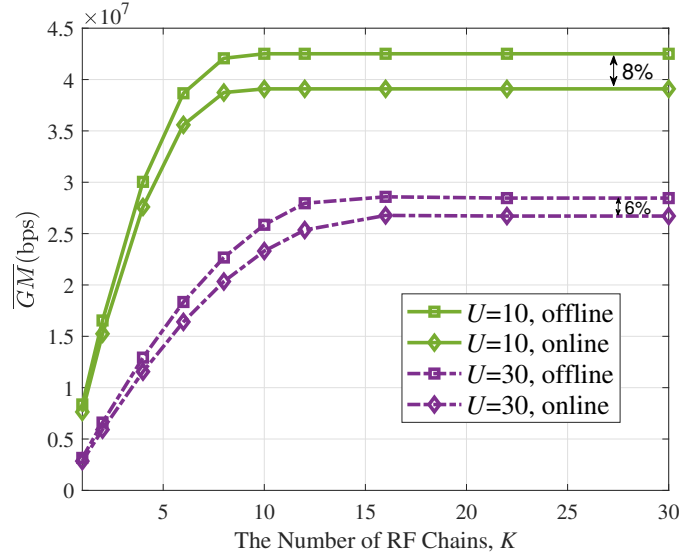


Figure 5.6: Performance of the online heuristic as a function of K for $U = 10$ and 30, and $M = 1$.

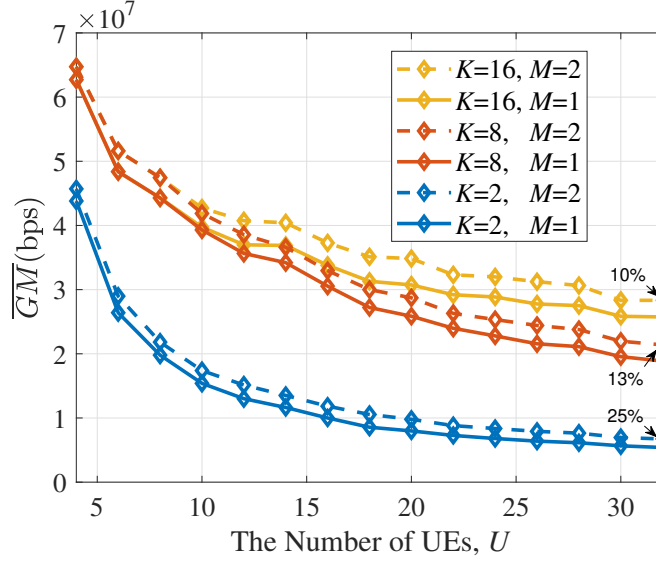


Figure 5.7: Performance of the online heuristic as a function of U for $K = 2, 8, 16$ and $M = 1, 2$.

increased to $Z = 200$ since the online heuristic is much faster than the offline solution. As illustrated by Fig. 5.7, K continues to play a critical role in determining system performance when U is large. For instance, at $U = 32$ and $M = 1$, increasing K from 2 to 8 nearly triples the system performance. While the performance gap between $K = 8$ and $K = 16$ remains under 10% for $U < 16$, it grows significantly with higher U , reaching 33% for $M = 1$ and 37% for $M = 2$ when $U = 32$. Fig. 5.7 also shows that using $M > 1$ is more beneficial when U is large and K is small. Specifically, at $U = 32$, increasing M from 1 to 2 yields gains of 25%, 13%, and 10% for $K = 2$, $K = 8$, and $K = 16$, respectively.

Finally, we present the time complexity analysis of each step of the proposed online heuristic in Table 5.1. The analysis assumes parallel execution across beams for PBS, across blocks for Step 2, across PRBs within a time slot for ZF-DBF, and across UEs for WF. For each beam, PBS searches up to Q blocks and at most another N_F PRBs if the final block is not selected. In each search step, PBS will find the best UE among at most U UEs (if all UEs are matched with this beam). Consequently, the worst-case complexity of PBS per beam is $\mathcal{O}((Q + N_F)U)$. Step 2 is performed on a per-block basis. Within each block, it checks up to N_F PRBs sequentially to remove or switch repeated UEs. This results in a worst-case complexity of $\mathcal{O}(N_F)$ per block. Load-aware beam selection has a complexity of $\mathcal{O}(|\mathcal{I}_s|)$ because it finds the L beams with the highest weighted sum rates from $|\mathcal{I}_s|$ beams.

Table 5.1: Complexity of the online heuristic solution (LA BSel: load-aware beam selection)

Step 1: PBS	Step 2 ($M > 1$)	LA BSel	ZF-DBF	WF
$\mathcal{O}((Q + N_F)U)$	$\mathcal{O}(N_F)$	$\mathcal{O}(\mathcal{I}_s)$	$\mathcal{O}(L^3)$	$\mathcal{O}(Q)$

Recall that L is the minimum between the number, K , of RF chains at BS and the number, $|\mathcal{I}_s|$, of matched BS beams (e.g., $L \leq 16$ for Fig. 5.7). Per-PRB ZF-DBF is performed on at most L UEs, hence its complexity is $\mathcal{O}(L^3)$ [44]. Finally, the complexity of WF is linear with respect to the number of channel coefficients [67] in PA, which, in our system, results in a worst-case complexity of $\mathcal{O}(Q)$ per UE. Among these steps, PBS and ZF-DBF dominate the overall complexity. Step 2, load-aware beam selection and WF have negligible complexity because $N_F \ll (Q + N_F)U$, $|\mathcal{I}_s| \propto U \ll (Q + N_F)U$ and $Q \ll (Q + N_F)U$. The overall complexity is $\mathcal{O}((Q + N_F)U + L^3)$. For a given system where K, Q , and N_F are fixed, the complexity scales linearly in U since for U large, $L = K$. Combined with its near-offline performance (92% at least), the low complexity of our online heuristic shows its suitability for system operation.

5.4 Conclusion

In this chapter, we have investigated RRM for the uplink of codebook-based HBF systems with multiple channels and no assumptions on the number of RF chains at the BS. Existing per-PRB solutions proposed in the uplink literature are not applicable in our scenario, as both beam selection and power allocation (PA) must be performed at the time-slot level. Per-time-slot solutions on the DL are also unsuitable because the per-UE PA on the UL is inherently different from that on the DL where the BS distributes power to different UEs. To address this gap, we first formulate a per-time-slot joint RRM optimization problem including beam selection, user selection and PA. We can solve this problem to quasi-optimality for a small number of UEs ($U \leq 10$). To support large system planning (e.g., 30 UEs, 30 RF chains at BS and 3 matched BS beams per UE), we propose an offline heuristic using per-beam user selection and PA optimization with a load-aware beam selection. This offline heuristic can achieve at least 93% of the quasi-optimal performance from joint optimization while reducing the runtime from months to hours. For online deployment, we propose an online heuristic by replacing the per-beam optimization with a proposed per-beam scheduler (PBS), which outperforms the state-of-art solution by 22% at least, and by substituting MCS-based power optimization with Shannon-based water-filling in the last step. Our online heuristic maintains at least 92% of the offline heuristic with very low complexity. Although Shannon-based rate function is good for power allocation,

it is necessary to consider practical MCS levels in the online heuristic to prevent selecting too many channels for a UE so that its per channel SNR is too low to achieve a positive rate.

We have also provided engineering insights into the impact of key system parameters, i.e., the number, K , of RF chains at BS and the number, M , of matched BS beams per UE. Our results indicate that setting $K = 0.6U$ (resp. $K = \frac{1}{3}U$) is sufficient to achieve strong performance, at least 88% of that at $K = U$ for $U = 10$ (resp. $U = 30$). Furthermore, increasing M from 1 to 2 can yield a substantial performance gain, up to 25%, when U is large and K is small (e.g., $U = 32$ and $K = 2$). However, increasing M from 2 to 3 results in negligible additional improvement.

Chapter 6

Fixed Wireless Access Study

In this chapter, we adopt our DL online solution from Section 4.2.2 and UL online solution from Section 5.2.2 to the use case of 5G FWA for broadband access in rural areas in the mid-band and in the mmWave band. Our goal is to use these solutions to evaluate key system parameters, such as codebook type and size, bandwidth, number of antennas, and number of RF chains, in both frequency bands to determine their ability to meet the MBR requirements on both directions.

We begin by introducing the necessary system model adaptations needed for 5G FWA in this context. In the previous chapters, we considered mobile UEs in small cell networks, where BA is necessary to match UEs with beams. In our 5G FWA use case, we consider fixed home UEs. Therefore, we use a position-based beam matching approach without relying on BA. This change allows for greater flexibility in the codebook size, since BA training constraints no longer limit the number of beams. Next, we perform a codebook study to determine appropriate configurations in terms of size and type. Note that the DL and the UL must be jointly considered when selecting codebook parameters, as the codebook type and size are determined prior to system operation. Once chosen, CE is performed only once by transmitting from the UE to the BS, using the fixed codebook. If different codebooks are applied for the DL and the UL, the measurements would be invalid for one direction. Finally, based on the selected codebook type and size, we analyze the impact of bandwidth, number of antennas, and number of RF chains on system performance in meeting the MBR targets.

6.1 System Model

The system model follows Chapter 3 except beam alignment (Section 3.1.1) and fairness (Section 3.4). In this section, we introduce a position-based beam matching instead of BA and max-min fairness instead of proportional fairness for MBR requirements in FWA.

We note that in a FWA setting, the coherence time can be very large, which may allow sufficient time for more sophisticated systems such as FDBF and optimization-based HBF (optimizing the ABF vectors instead of choosing them from codebooks) with full channel matrix estimation and iterative user selection.

6.1.1 Positioned-based Beam Matching

In FWA, we assume that all home UEs have a line-of-sight connection to the BS since the antennas are roof top-based [50, 51]. Given that the positions of both the UEs and the BS are fixed, the optimal beam pair consists of a BS beam directed toward the UE and a UE beam directed toward the BS, which together best capture the LOS path gain.

Therefore, on the UE side, we assume that the antenna array is physically oriented toward the BS, and the center beam, formed by applying a zero phase shift across all phase shifters, is used for communication. On the BS side, since the codebook is constructed without knowledge of specific UE locations and a beam cannot always be perfectly aligned with a UE's direction, we select the beam whose angular coverage includes the UE's direction. Specifically, the angular coverage of a beam refers to the interval of directions in which it outperforms all other beams in terms of beamforming gain, and is thus considered the best match for UEs within that range. For example, if a BS uses a DFT codebook with $B_b = N_b$, then the angular coverage of each beam corresponds approximately to its 3-dB beamwidth coverage. Since the position of each UE can only be mapped to one angular coverage interval and we do not have BA to evaluate multiple BS beams, we only consider one matched BS per UE ($M = 1$).

By using the position-based beam matching, we match each UE with a BS beam from BS codebook $\mathcal{C}_b$ and the central UE beam in UE codebook $\mathcal{C}_u$. This process serves the same function as BA but requires no training overhead, thereby allowing the use of larger codebooks with more beams.

6.1.2 Max-min Fairness

For FWA, we are going to assume TDD and a certain ratio between DL and UL. Our objective function both on the DL and UL is to maximize the minimum UE throughput across all UEs over a certain horizon. i.e.,

$$\max \min_{u \in \mathcal{U}} WR_u + \lambda_u, \quad (6.1)$$

where λ_u is the throughput of UE u within the current time slot and R_u is the average per-time-slot throughput received by u over the past W time slots as defined in Section 3.4. Max-min fairness can be achieved by using α -fairness [26]. Then, the objective function can be written as

$$\max \sum_{u \in \mathcal{U}} \frac{(WR_u + \lambda_u)^{1-\alpha}}{1-\alpha}, \quad (6.2)$$

where $\alpha \gg 1$ for max-min objective. It can be further approximated as

$$\max \sum_{u \in \mathcal{U}} \frac{(WR_u + \lambda_u)^{1-\alpha}}{1-\alpha} \quad (6.3)$$

$$\equiv \max \sum_{u \in \mathcal{U}} \frac{(WR_u)^{1-\alpha}}{1-\alpha} \left(1 + \frac{\lambda_u}{WR_u}\right)^{1-\alpha} \quad (6.4)$$

$$\approx \max \sum_{u \in \mathcal{U}} \frac{(WR_u)^{1-\alpha}}{1-\alpha} \left(1 + (1-\alpha) \frac{\lambda_u}{WR_u}\right) \quad (6.5)$$

$$\equiv \max \sum_{u \in \mathcal{U}} \frac{\lambda_u}{R_u^\alpha} \quad (6.6)$$

From (6.4) to (6.5), the approximation follows $(1+x)^n \approx 1+nx$ when $|x| \ll 1$ and $|nx| \ll 1$ since $|\frac{\lambda_u}{WR_u}| \ll 1$ and $|(1-\alpha)\frac{\lambda_u}{WR_u}| \ll 1$. From (6.5) to (6.6), the constant terms are omitted, as they do not affect the optimization objective. The objective function is thus reformulated as maximizing the weighted sum rate of all UEs, where the weight for UE u is given by $\frac{1}{R_u^\alpha}$. By reformulating the max-min objective as a weighted sum rate maximization objective, we can reuse the online RRM solutions proposed in Chapters 4 and 5, which were originally designed for weighted sum rate optimization. We apply the same solution framework for max-min fairness by modifying the weight from $\frac{1}{R_u}$ to $\frac{1}{R_u^\alpha}$.

6.2 RRM Solutions

Our methodology can be summarized as follows: first, we will adapt the online algorithms designed for PF to max-min as explained above. Since these online solutions assume low inter-beam interference, an assumption that may not hold when B_b is large, we introduce an interference mitigation process to mitigate the resulting interference. Specifically, our DL solution follows Section 4.2.2 where, we select the L beams having the lowest min rate, where the min rate for beam b is $R_b = \min_{u \in \mathcal{U}(b)} \{R_u\}$, followed by per-beam user selection as in Algorithm 2 for the L selected beams (recall that the L user selection processes are independent of each other), per block ZF-DBF, and water-filling PD. For UL (please refer to Section 5.2.2), we use the same beam selection as for the DL, followed by per-beam scheduler as in Algorithm 5 for the L selected beams, per-PRB ZF-DBF, and per-UE water-filling PA.

Recall that our algorithms are based on a per-beam user selection that is blind to other beams, which might not work well with all codebooks (especially large ones since in that case, the interference between adjacent beams can be high). Hence, we add a new *interference mitigation (IM)* process to the algorithms where we verify, for each PRB, if removing users might improve the utility function by decreasing the interference seen by other users. In Section 6.3.6, we will show that this new IM process improves the performance, especially for UL with large B_b , i.e., when the codebook size is larger. More precisely, at each step of the process, we deselect the UE (denoted as UE n) in a PRB that causes the strongest interference to the UE with the lowest past rate (denoted as UE u). Specifically, this is determined by evaluating $p^{q,n} g_{q,n,u}^{\text{eff}}$ on the DL or $p^{q,u} g_{q,u,n}^{\text{eff}}$ on the UL, where q is the block index of the PRB, $p^{q,n}$ is the transmit power of UE n in block q , and $g_{q,n,u}^{\text{eff}}$ is defined in Section 3.1.2. This deselection is followed by a recalculation of ZF-DBF and PD/PA. This user deselection process is repeated until no further improvement is observed in the global weighted sum rate across all UEs and all PRBs. While it improves performance in some cases, its gap to the optimal solution remains unclear.

Finally, in both the DL and the UL, we update R_u at the end of the time slot.

6.3 Numerical Evaluations

In this section, we present numerical results in FWA settings covering both DL and UL at 3.5 GHz and 28 GHz, to analyze the impact of codebook type and size, bandwidth, number of antennas, number of RF chains, and antenna array orientation on system performance

when the RRM algorithms are the ones described above. Note that it is possible that more sophisticated algorithms that would do joint user selection would yield different trends.

We begin by introducing the system parameters and the three types of codebooks evaluated in this study: the Song codebook proposed in [27], the DFT codebook [9], and the Chebyshev-windowed DFT codebook [28]. Next, we describe the 3GPP spatial channel model, which is applied to both the mid-band and mmWave evaluations using band-specific parameters. Finally, we present the numerical results along with a detailed performance analysis.

Each result is obtained by simulating 200 time slots and calculating the minimum bit rate among all UEs based on the last $T = 100$ time slots. A moving average with a window size of $W = 100$ time slots is used when updating R_u . Such a large window is necessary because of the large value of α . After evaluating various values of α , we set $\alpha = 25$ because the performance gain becomes negligible beyond this point. We have tested different values of T and W and found that after $T = 200$ and $W = 100$, the performance becomes steady. For each system setting, i.e., specific values for N_b , K , and B_b , as well as a codebook type, we generate 20 independent channel realizations and compute the average minimum bit rate across these realizations to represent the system performance.

6.3.1 System Parameters

The use case we consider has been suggested to us by one of the major Canadian operators, that has sponsored part of this FWA study. Specifically, we consider a FWA cell in rural area with a radius of $r = 3000$ meters in which 77 UEs are distributed. The BS antenna is elevated at a height of 79.2 meters, while the UE antennas are positioned at a height of 5 meters. The total transmit power of the BS is $P_{BS} = 46$ dBm. Each UE is equipped with a transmit power budget of $P_{UE} = 26$ dBm, and the noise power spectral density is set to $N_0 = -174$ dBm/Hz. Both the BS and UEs employ uniform linear antenna arrays with half-wavelength spacing. We assume that the antenna array is oriented parallel to the x-axis of the map. For the mid-band, the system operates at a carrier frequency of $f_c = 3.5$ GHz with a total bandwidth of $BW = 60$ (resp. 90) MHz consisting of 24 (resp. 36) blocks in frequency. Each block has 7 PRBs whose bandwidth is 360 kHz. The time slot duration is 0.5 msec. The number of antennas at the BS is $N_b = 32$ or 64 and the number of antennas at each UE is $N_u = 4$. For the mmWave band, the system operates at a carrier frequency of $f_c = 28$ GHz with a total bandwidth of $BW = 100$ (resp. 200) MHz consisting of 10 (resp. 20) blocks in frequency. Each block has 7 PRBs whose bandwidth is 1.44 MHz. The time slot duration is 0.125 msec. The number of antennas at the BS is

$N_b = 128$ or 256 and the number of antennas at each UE is $N_u = 8$. A fixed DL-to-UL time slot ratio of 4:1 is applied in both bands.

6.3.2 Codebooks

Our first type of codebook is the Song codebook from [27] which was used for the numerical evaluations in previous chapters with the small cell setting. With the Song codebook, we can generate $B_b = 2^B$ beams, where $1 \leq 2^B \leq N_b$ (i.e., for $N_b = 64$, the only "large" values of B_b are 64 and 32). It is constructed using a search-based algorithm that optimizes the flatness of the main lobe while simultaneously suppressing sidelobe levels. This design makes it well-suited for scenarios where B_b is constrained, such as in small cell deployments where BA is necessary for mobile UEs and a large B_b would incur significant overhead. A flat main lobe ensures consistent BF gain for UEs within the beam's coverage, while low sidelobes help reduce inter-beam interference. An example of a Song codebook is shown in Fig. 6.1. The BF gain is normalized by $10 \lg(N_b)$, which represents the maximum achievable gain of N_b antennas. We can observe a gain degradation of approximately 6 dB, which results from distributing the antenna gain across a coverage region roughly four times wider than the 3-dB beamwidth of a beam with peak gain $10 \lg(N_b)$, as achieved by applying per-antenna phase shift compensating for the propagation delay at an antenna element toward the desired direction.

Since BF gain can play a crucial role in guaranteeing the MBR, especially for remote UEs in a FWA setting, we also consider DFT codebooks [9]. Each beam in DFT codebook achieves a gain of $10 \log_{10}(N_b)$ and narrow beam width, making it suitable for high-gain transmission to remote UEs. DFT codebooks contain B_b beams that are evenly distributed in the angular domain, such that the crossing angles between adjacent beams correspond to the same BF gain. B_b can be any value (normally close to N_b) in DFT codebooks, which provides more flexibility compared to Song codebooks. We do not consider $B_b \leq \frac{N_b}{2}$, because the beamwidth of DFT beams is narrow and the BF gain becomes very low when the target angle is far from the central angle of a selected beam, i.e., a single beam does not have a good coverage. Note that when $B_b = N_b$, the Song codebook is the same as the DFT codebook. An example of DFT codebook is shown in Fig. 6.2.

One main concern with the DFT codebook is the sidelobe level of its beams, which is approximately -13 dB below the mainlobe. These sidelobes may cause interference to UEs located in adjacent beams. To address this issue, we consider a third type of codebook: Chebyshev-windowed DFT (CDFT) codebook, which is constructed by applying a Chebyshev window to each beam in a DFT codebook [28]. Chebyshev window can

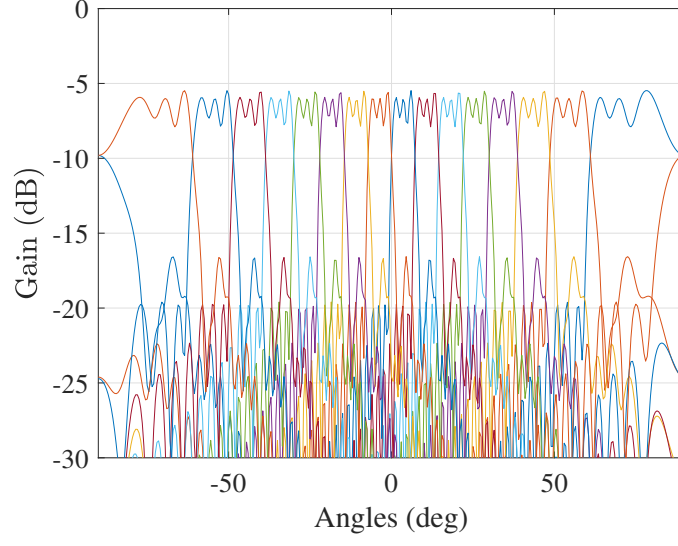


Figure 6.1: Beam pattern of Song codebook for $N_b = 64$ and $B_b = 16$.

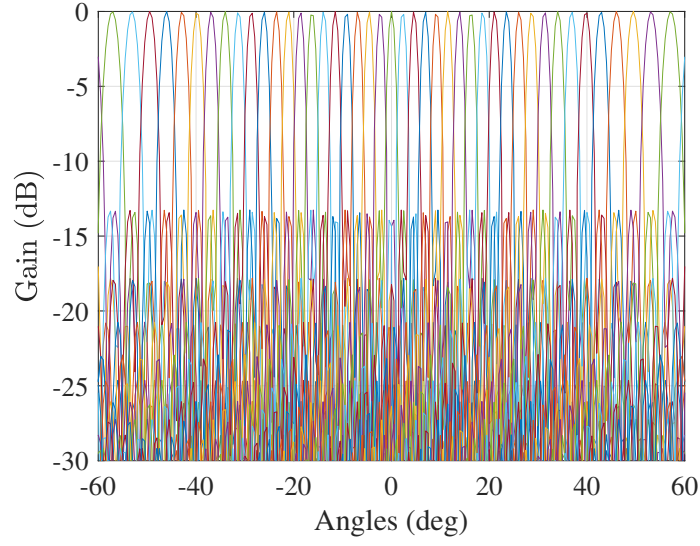


Figure 6.2: Beam pattern of DFT codebook for $N_b = 64$ and $B_b = 50$.

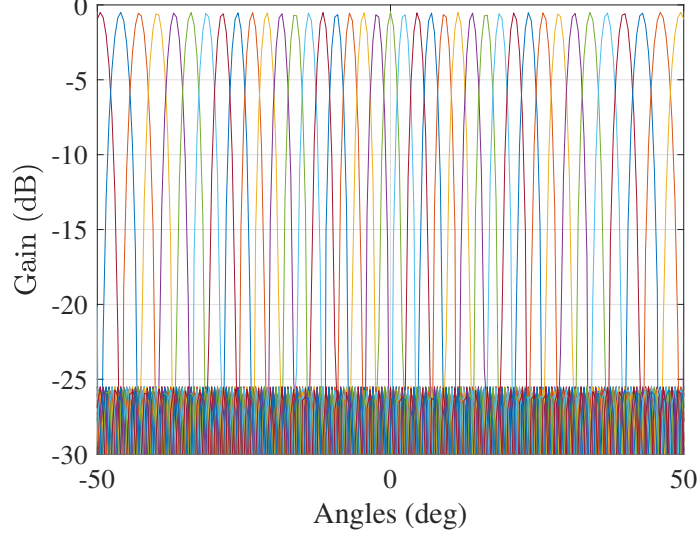


Figure 6.3: Beam pattern of CDFT codebook for $N_b = 64$, $B_b = 50$ and sidelobe controlled at -25 db.

suppress sidelobes to very low levels (e.g., -20 , -25 , or -30 dB, values above -20 dB are not considered since the sidelobes of DFT beams are already close to -15 dB), but this comes at the cost of a wider mainlobe and a slight reduction in beamforming gain compared to an unwindowed DFT beam. Note that adjusting the sidelobe level has limited effect on the BF gain and mainlobe beamwidth. We tested various sidelobe levels and observed negligible performance differences, as ZF-DBF can effectively suppress interference at the -20 dB level. Therefore, we use a sidelobe level of -20 dB for $N_b = 64$ at 3.5 GHz and -25 dB for $N_b = 128$ at 28 GHz, as these settings minimize BF gain loss. It became apparent while running numerical results, that the main concern in CDFT codebooks is not the sidelobe interference, but rather the overlap between mainlobes. As illustrated in Fig. 6.3, CDFT codebook exhibits significantly lower sidelobes but more mainlobe overlap compared to the DFT codebook in Fig. 6.2, under the same B_b and N_b . The comparison between CDFT and DFT codebook enables us to understand the trade-off between mainlobe interference and sidelobe interference control in codebook design for multi-user transmission.

6.3.3 Channel Model and Parameters

We adopt the spatial channel model from 3GPP presented in [52] for our FWA study. The channel between BS and UE u at block q is defined as the following $N_u \times N_b$ matrix [71]:

$$\begin{aligned} \mathbf{G}_{q,u} = & \frac{1}{\sqrt{10^{0.1\text{PL}_u}}} \left(\sqrt{\frac{1}{1+\kappa_u}} \sum_{d=1}^{N_{\text{cluster}}} \sqrt{\gamma_{d,u}} \sum_{l=1}^{N_{\text{path}}} \sqrt{g_{d,l,u}} e^{-j2\pi(\tau_{d,u}+\tau_{d,l,u})f_q+j\psi_{d,l,u}} \right. \\ & \times \mathbf{a}_{\text{RX}}(\phi_{d,l,u}^{\text{AOA}}, \theta_{d,l,u}^{\text{AOA}}) \mathbf{a}_{\text{TX}}^H(\phi_{d,l,u}^{\text{AOD}}, \theta_{d,l,u}^{\text{AOD}}) + \sqrt{\frac{\kappa_u}{1+\kappa_u}} e^{-j2\pi\frac{d_u}{c}f_q+j\psi_u} \\ & \left. \times \mathbf{a}_{\text{RX}}(\phi_u^{\text{AOA}}, \theta_u^{\text{AOA}}) \mathbf{a}_{\text{TX}}^H(\phi_u^{\text{AOD}}, \theta_u^{\text{AOD}}) \right), \end{aligned} \quad (6.7)$$

where, for UE u , PL_u is the path loss, κ_u is the Rician K factor, $\gamma_{d,u}$ is the power of cluster d , $g_{d,l,u}$ is the power of path l in cluster d , $\tau_{d,u}$ is the delay of cluster d , $\tau_{d,l,u}$ is the in-cluster delay of path l in cluster d , $\psi_{d,l,u} \sim \mathcal{U}[0, 2\pi)$ is the initial phase of path l in cluster d . Also, $\phi_{d,l,u}^{\text{AOA}}$, $\theta_{d,l,u}^{\text{AOA}}$, $\phi_{d,l,u}^{\text{AOD}}$, and $\theta_{d,l,u}^{\text{AOD}}$ denote the azimuth AOA, elevation AOA, azimuth AOD, and elevation AOD of path l in cluster d , respectively. The first term denotes the sum of NLOS components. d_u is the point-to-point distance from BS antenna array to UE antenna array, $\psi_u \sim \mathcal{U}[0, 2\pi)$ is the initial phase of the LOS path. Also, ϕ_u^{AOA} , θ_u^{AOA} , ϕ_u^{AOD} , and θ_u^{AOD} denote the azimuth AOA, elevation AOA, azimuth AOD, and elevation AOD of the LOS path, respectively. The second term denotes the LOS component. User-independent parameters N_{cluster} and N_{path} are the numbers of clusters and paths in a cluster, and f_q is the center frequency of block q . $\mathbf{a}_{\text{TX}}(\cdot)$ and $\mathbf{a}_{\text{RX}}(\cdot)$ are the array response vectors at the BS and UE u .

Parameters at 3.5 GHz

For 3.5 GHz, the above parameters are generated sequentially according to the procedure outlined in Section 7 of [52] based on the RMa (Rural Macro) scenario.

Parameters at 28 GHz

For 28 GHz, the RMa scenario path loss and Rician K factor provided in [52] have not been thoroughly validated [72, 73]. Therefore, we adopt the CIH model from [72] for path loss and the Rician K factor from [74]. These parameters are supported by measured data, making them more suitable for realistic evaluations at 28 GHz.

6.3.4 Numerical Evaluations at 3.5 GHz

This section presents the evaluation and analysis at 3.5 GHz. We first analyze the impact of codebook type and size on system performance. Specifically, we evaluate each codebook type individually with a range of numbers of beams B_b and find the best B_b for each codebook. Then, we compare the best performance of each codebook to decide the type for the DL and the UL respectively. Finally, we jointly consider DL and UL to choose one set of B_b and type. Recall that only one codebook type and size can be selected, as CSI is only measured once with a fixed codebook for both the DL and the UL. Note that we do not evaluate the impact of N_u since only the center beam will be used at each UE.

Subsequently, given the best codebook type and size, we analyze the impact of other key system parameters, including bandwidth BW , number of antennas N_b , and number of RF chains K .

Fig. 6.4 depicts the minimum rate on the DL as a function of the number of RF chains for different values of B_b with Song codebooks. Recall that we can use a large B_b because position-based beam matching avoids the BA overheads. Clearly, for K larger than 8, $B_b = N_b$ is key to obtaining a much higher minimum rate. Doubling B_b from 32 to 64 brings a gain of 75% when $K \geq 16$. Fig. 6.5 presents a similar trend on the UL, where increasing B_b from 32 to 64 yields a 62% performance gain when $K \geq 16$. Since BA overhead is no longer a concern, $B_b = 64$ provides the best performance for Song codebooks both on the DL and the UL. Note that the rate slightly decreases as K becomes large. This occurs because our online solutions are not optimal: with a larger K , more UEs are selected, which limits the effectiveness of ZF-DBF.

DFT codebooks offer better flexibility in selecting B_b in two ways, B_b can take any value below N_b and can be greater than N_b . Fig. 6.6 illustrates the impact of varying B_b from 46 to 70, in increments of 6, on system performance. Increasing B_b from 46 to 52 results in a 21% improvement, but beyond $B_b = 52$, further increases do not yield additional performance gains. One possible reason is that while a larger B_b enhances the BF gain per UE, it reduces the number of UEs that can be served within each beam, thereby limiting user selection flexibility. Additionally, higher B_b increases the complexity of interference mitigation. Since the interference mitigation method introduced in Section 6.2 is not optimal, its effectiveness may diminish as B_b grows. A similar trend is observed on the uplink, as shown in Fig. 6.7. Increasing B_b from 40 to 52 leads to a 43% performance gain, but no further improvement is seen for $B_b > 52$. Therefore, for DFT codebook, $B_b = 52$ achieves the best overall performance both for the DL and the UL. The comparison of the different types of codebooks is discussed later.

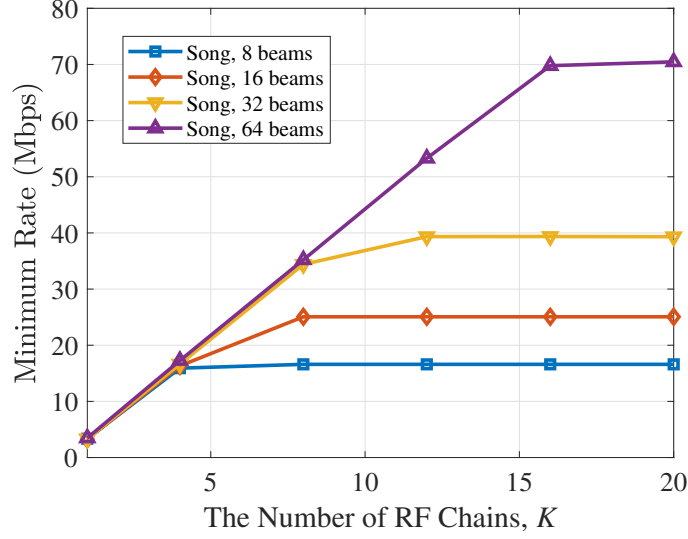


Figure 6.4: Midband, DL: Impact of B_b on system performance with Song codebooks for $N_b = 64$, $N_u = 4$ and $BW = 60$ MHz.

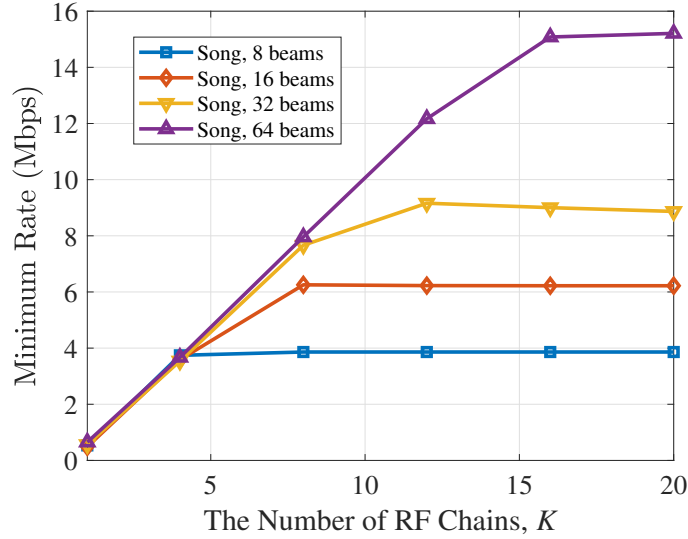


Figure 6.5: Midband, UL: Impact of B_b on system performance with Song codebooks for $N_b = 64$, $N_u = 4$ and $BW = 60$ MHz.

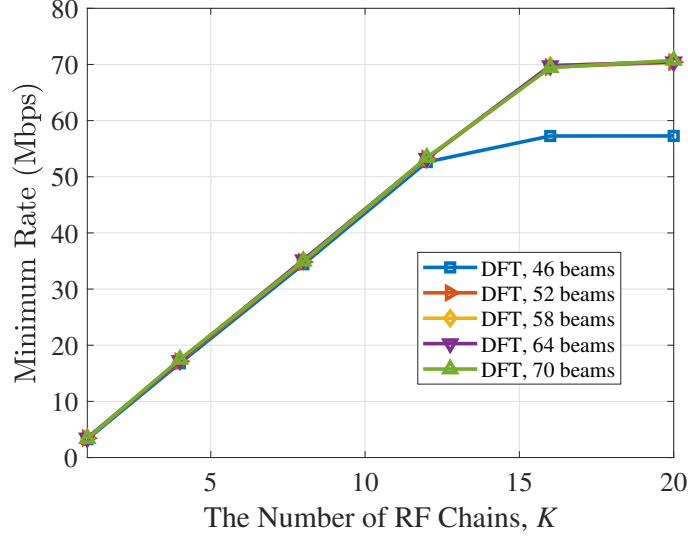


Figure 6.6: Midband, DL: Impact of B_b on system performance with DFT codebook for $N_b = 64$, $N_u = 4$ and $BW = 60$ MHz.

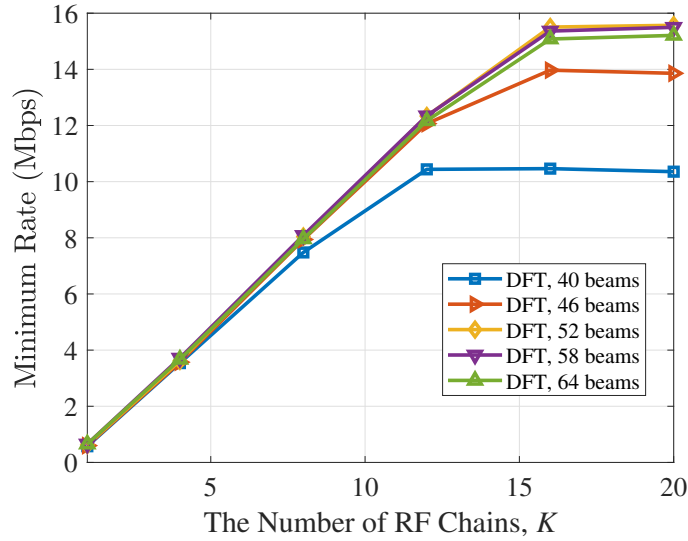


Figure 6.7: Midband, UL: Impact of B_b on system performance with DFT codebook for $N_b = 64$, $N_u = 4$ and $BW = 60$ MHz.

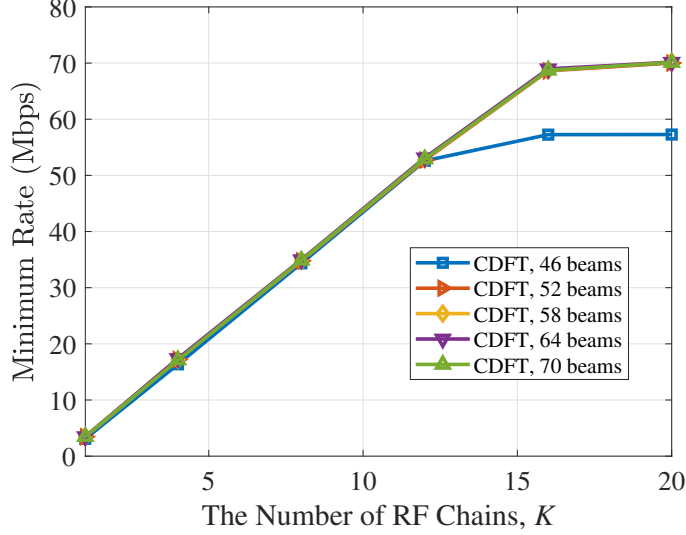


Figure 6.8: Midband, DL: Impact of B_b on system performance with Chebychev-window codebook for $N_b = 64$, $N_u = 4$ and $BW = 60$ MHz.

To evaluate the importance of sidelobe control in codebook design, we assess the performance of Chebyshev-windowed DFT codebook, with the sidelobe level set to -20 dB. Fig. 6.8 presents the performance of CDFT codebook on the DL. Similar to DFT codebook, the performance saturates beyond $B_b = 52$. On the UL, Fig. 6.9 shows that the best performance is achieved at $B_b = 58$. Considering both DL and UL results, we select $B_b = 58$ for CDFT codebook.

Given the selected values of B_b for the three codebook types, we compare performances on the DL and the UL. As shown in Fig. 6.10, only minor differences are observed among the codebooks on the DL. On the UL, Fig. 6.11 shows that DFT codebook with $B_b = 52$ achieves slightly higher performance for K large. The fact that DFT codebook outperforms CDFT codebook on the UL suggests that, at 3.5 GHz, mainlobe interference control is more important than sidelobe interference control in determining overall performance.

We also compare two antenna array placement configurations: vertical (i.e., parallel to y-axis of the map) and horizontal (i.e., parallel to x-axis of the map). As shown in Fig. 6.12, the horizontal placement achieves up to three times higher performance compared to the vertical one on both DL and UL.

Next, we consider a comparison between the best codebook and Full Digital Beamforming (FDBF). We compare the performance of the codebook-based HBF system (DFT

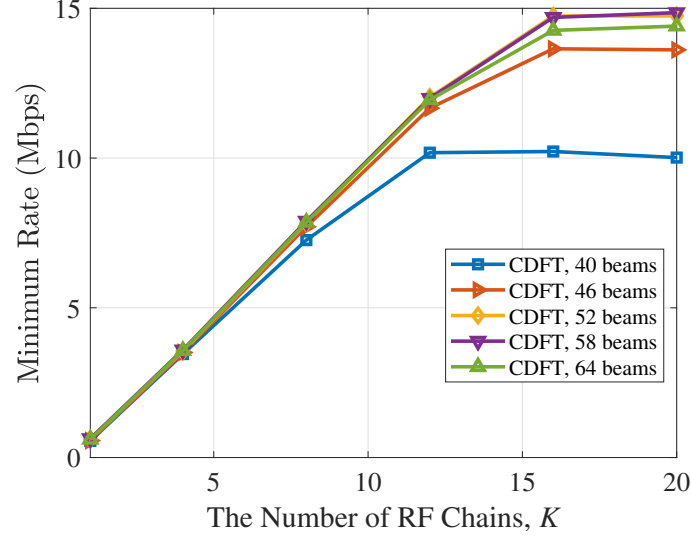


Figure 6.9: Midband, UL: Impact of B_b on system performance with Chebychev-window codebook for $N_b = 64$, $N_u = 4$ and $BW = 60$ MHz.

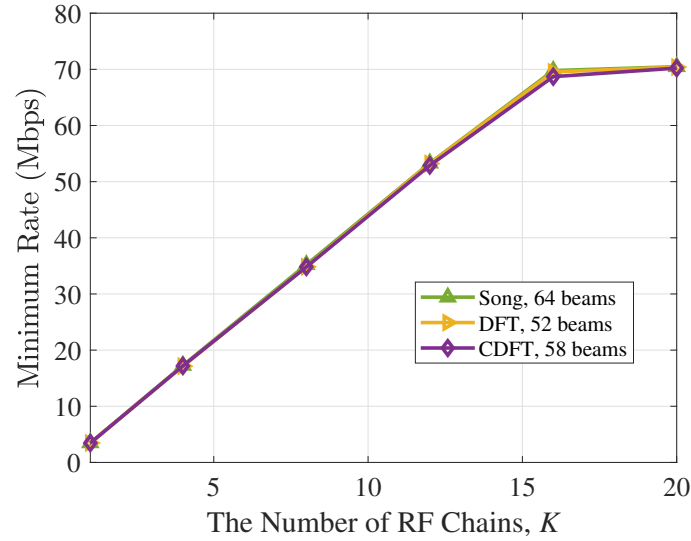


Figure 6.10: Midband, DL: Impact of types of codebooks on system performance for $N_b = 64$, $N_u = 4$ and $BW = 60$ MHz.

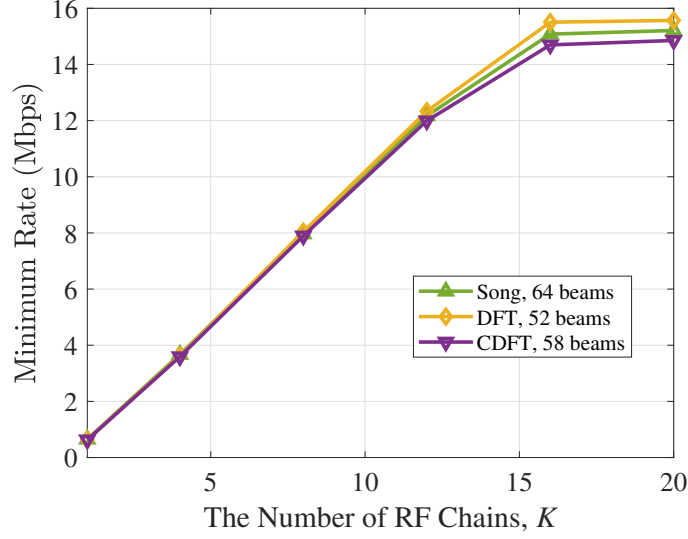


Figure 6.11: Midband, UL: Impact of types of codebooks on system performance for $N_b = 64$, $N_u = 4$ and $BW = 60$ MHz.

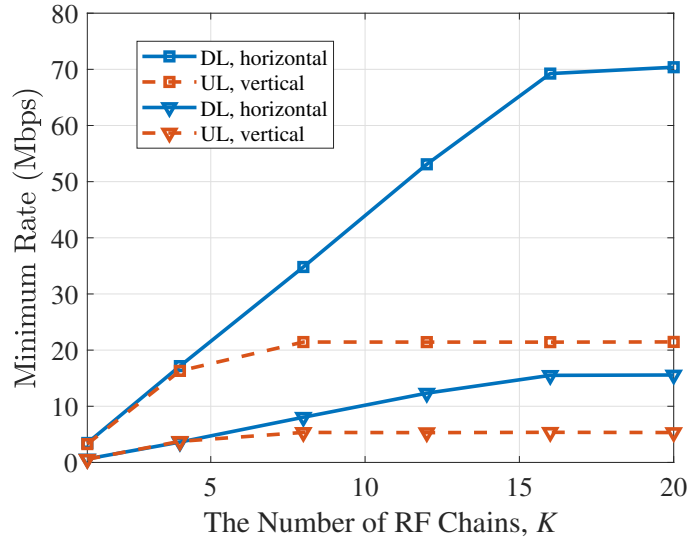


Figure 6.12: Midband: Impact of antenna array placement for $N_b = 64$, $N_u = 4$, $B_b = 52$, DFT codebook and $BW = 60$ MHz on the DL and UL.

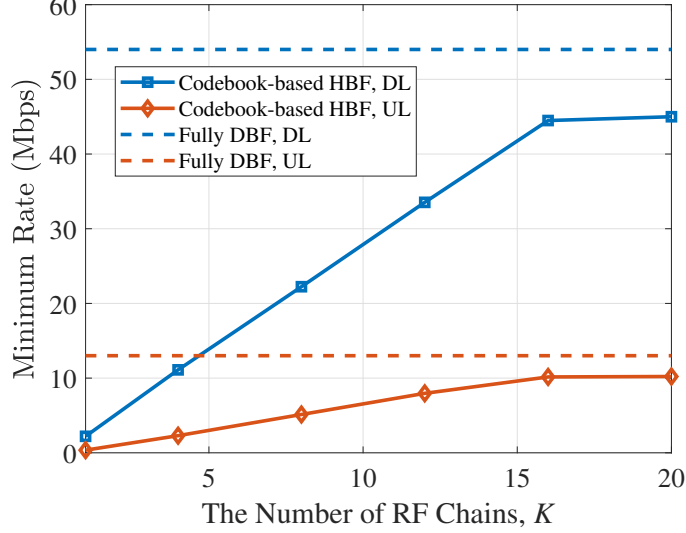


Figure 6.13: Midband: System performance comparison between codebook-based HBF and fully DBF architectures for $N_b = 64$, $N_u = 4$ and $BW = 40$ MHz on the DL and UL.

codebook with $B_b = 52$) obtained with our RRM solutions against that of a fully DBF system employing the greedy UP solution proposed in [75]. Specifically, for both DL and UL in the FDBF system, the solution evaluates all possible UE assignments in a PRB at each step, and selects the UE-PRB pair that yields the highest increase in weighted sum rate. ZF-DBF and optimized PD/PA (as in Section 4.1.3/5.1.3) are assumed throughout this process. Note that there are three system-level differences between codebook-based HBF and FDBF. First, HBF requires only the effective channel matrix as input to RRM, i.e., a $|\mathcal{I}_s| \times U$ matrix per block, while FDBF requires the full channel matrix, i.e., a $N_b \times N_u \times U$ matrix per block. Second, HBF requires significantly fewer RF chains at the BS (20 RF chains) compared to FDBF (64 RF chains) to obtain high performance. Third, each UE requires only one RF chain in the HBF system, whereas 4 RF chains are needed in the FDBF system. Fig. 6.13 presents the comparison between codebook-based HBF and FDBF. Note that we set $BW = 40$ MHz for this figure to perform the comparison, as FDBF performance results were only available for this bandwidth. With our solutions, HBF achieves 83% of the FDBF performance on the DL and 80% on the UL. This can be considered as a partial validation of our RRM solutions since it is known that FDBF is outperforming HBF.

Finally, we analyze the impact of system bandwidth and number of antennas on per-

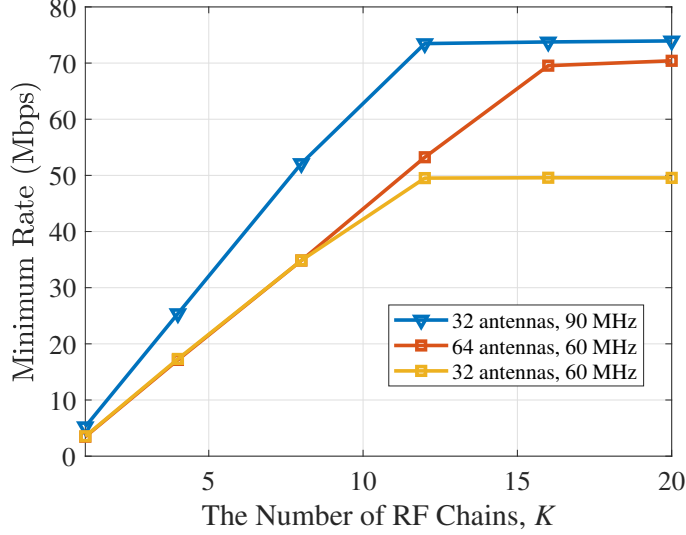


Figure 6.14: Midband, DL: Impact of bandwidth and number of antennas on system performance as a function of K .

formance with DFT codebook. We use $B_b = 52$ for $N_b = 64$ and $B_b = 26$ for $N_b = 32$. As shown in Figs. 6.14 and 6.15, increasing N_b can improve system performance by 40% on both the DL and the UL, but only when $K \geq 12$. This is because a larger N_b increases beam resolution, enabling more opportunities for spatial multiplexing. However, the gain is only realized when there are enough RF chains to exploit this additional spatial resolution. When K is small, increasing bandwidth becomes more effective. Specifically, increasing the bandwidth from 60 to 90 MHz yields a performance gain of at least 40% on both the DL and the UL for $K \geq 2$.

In general, the MBR targets of 50 Mbps on the DL and 10 Mbps on the UL can be achieved using DFT codebook with either $N_b = 64$, $BW = 60$ MHz, and $K = 12$, or $N_b = 32$, $BW = 90$ MHz, and $K = 6$.

6.3.5 Numerical Evaluations at 28 GHz

This section presents the evaluation and analysis at 28 GHz, following the same procedure used for the 3.5 GHz evaluation. We begin by analyzing the impact of codebook type and size on system performance. Subsequently, with the chosen codebook type and size (recall that we need to use the same codebook for DL and UL, as CSI is only measured once with

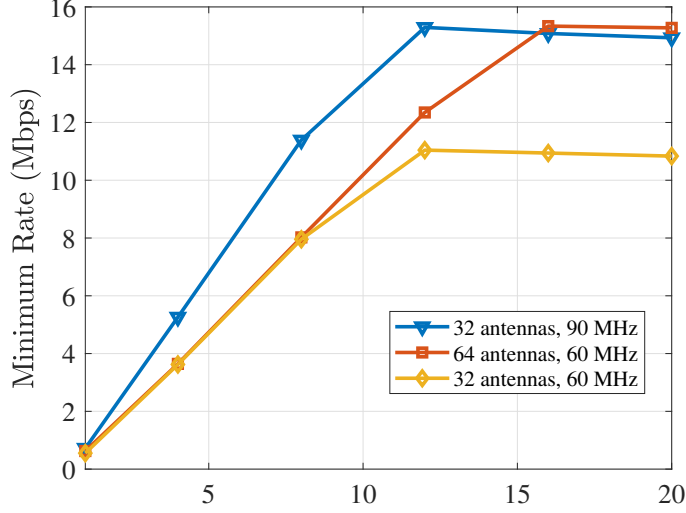


Figure 6.15: Midband, UL: Impact of bandwidth and number of antennas on system performance as a function of K .

a fixed codebook), we examine the impact of key system parameters, including B_b , N_b , and K .

Figs. 6.16 and 6.17 show the DL and UL performance of Song codebook. By evaluating B_b from 16 to 128, we observe that $B_b = 128$ yields the best performance in both directions. A larger B_b proves beneficial, improving DL performance by 50% and nearly doubling UL performance compared to $B_b = 64$.

Figs. 6.18 and 6.19 present the DL and UL performance of DFT codebook. We evaluate B_b from 104 to 140 in steps of 12. In both directions, the performance plateaus at $B_b = 128$. Notably, the difference between DL and UL performance becomes more pronounced at 28 GHz compared to 3.5 GHz. The ratio of the best DL to UL performance is approximately 8:1 at 28 GHz, compared to 5:1 at 3.5 GHz, indicating a more severe UL bottleneck in the mmWave band.

The impact of B_b with CDFT codebook is shown in Figs. 6.20 and 6.21 with the sidelobe controlled at -25 dB. While $B_b = 128$ yields the best performance on the DL, the UL performance continues to improve beyond this point. In particular, $B_b = 140$ outperforms $B_b = 128$ on the UL when $K \geq 12$. Since UL is the bottleneck for meeting the MBR target, we prioritize UL performance and select $B_b = 140$ for CDFT codebook.

Figs. 6.22 and 6.23 compare the DL and UL performance of the three types of codebooks

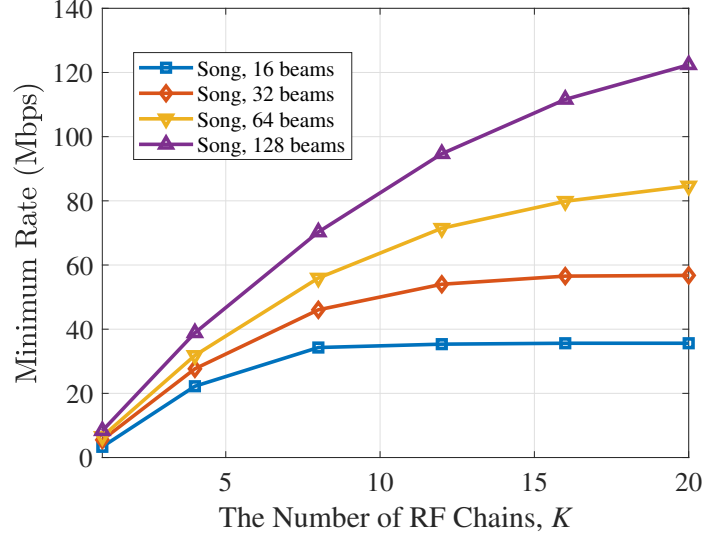


Figure 6.16: MmWave, DL: Impact of B_b on system performance with Song codebook for $N_b = 128$, $N_u = 8$ and $BW = 200$ MHz.

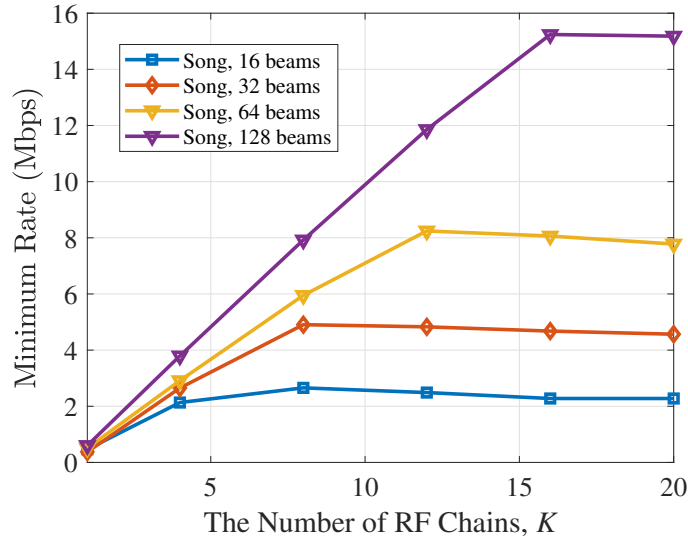


Figure 6.17: MmWave, UL: Impact of B_b on system performance with Song codebook for $N_b = 128$, $N_u = 8$ and $BW = 200$ MHz.

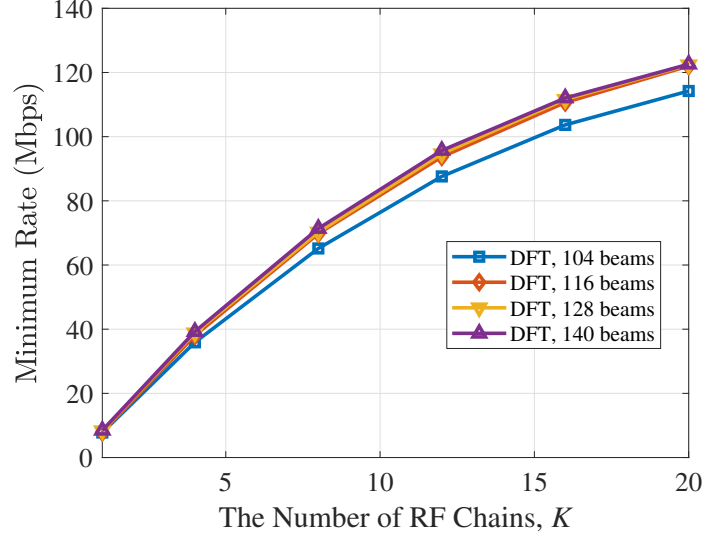


Figure 6.18: MmWave, DL: Impact of B_b on system performance with DFT codebook for $N_b = 128$, $N_u = 8$ and $BW = 200$ MHzL.

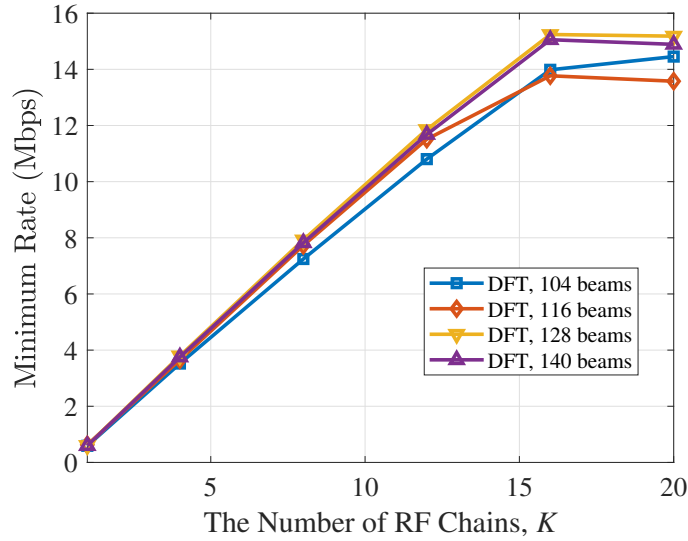


Figure 6.19: MmWave, UL: Impact of B_b on system performance with DFT codebook for $N_b = 128$, $N_u = 8$ and $BW = 200$ MHz.

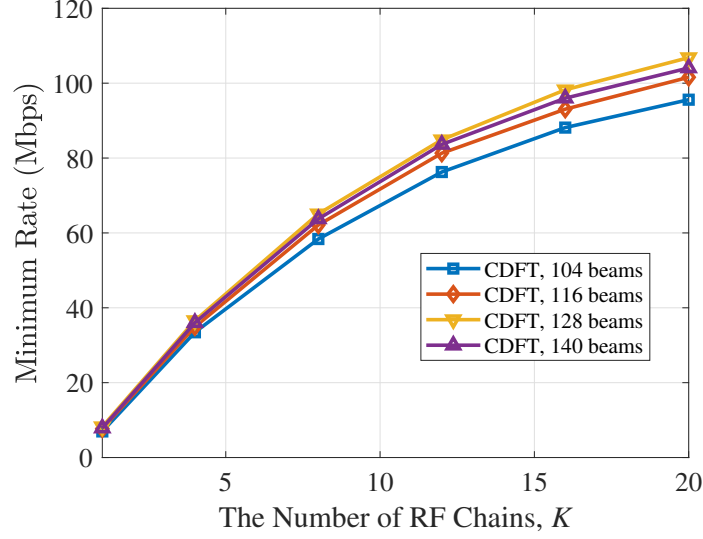


Figure 6.20: MmWave, DL: Impact of B_b on system performance with Chebychev-window codebook for $N_b = 128$, $N_u = 8$ and $BW = 200$ MHz.

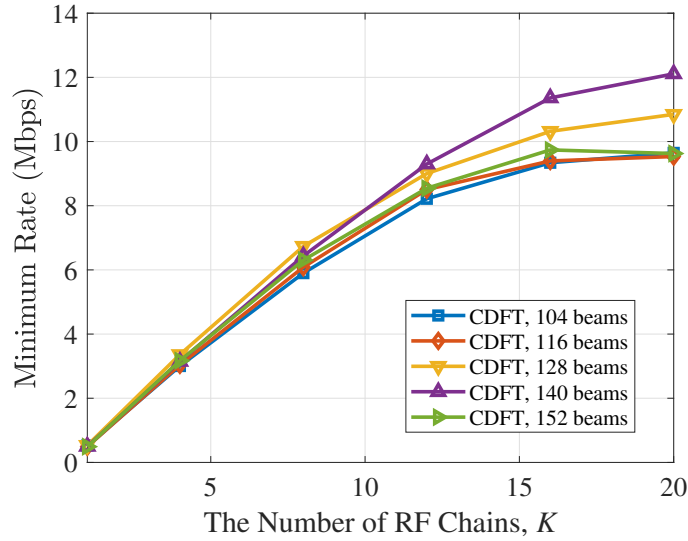


Figure 6.21: MmWave, UL: Impact of B_b on system performance with Chebychev-window codebook for $N_b = 128$, $N_u = 8$ and $BW = 200$ MHz.

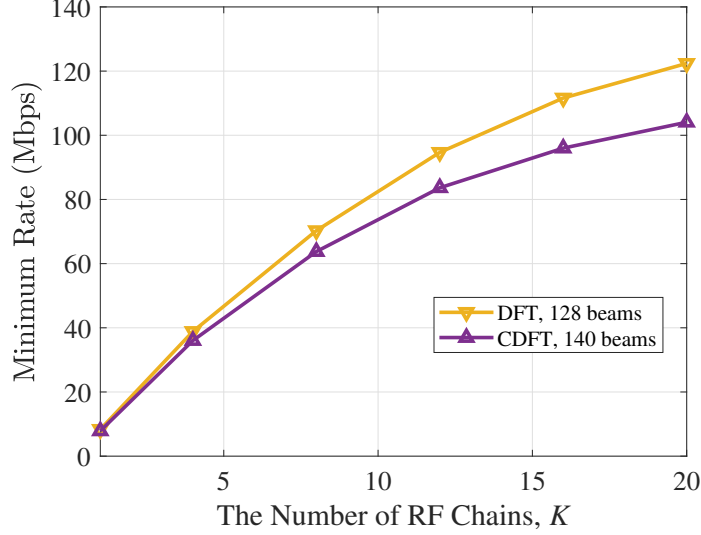


Figure 6.22: MmWave, DL: Impact of types of codebooks on system performance for $N_b = 128$, $N_u = 8$ and $BW = 200$ MHz.

with their selected B_b . Recall that DFT codebook and Song codebook are identical when $B_b = N_b$. Therefore, we only present DFT codebook results in the figures. on the DL, we notice DFT codebook outperforms CDFT codebook by 20% for large values of K . on the UL, the performance gap increases to 30% at $K = 16$. In conclusion, DFT codebook with $B_b = N_b$ outperforms the other two codebooks at 28 GHz.

Focusing now on DFT codebook, we further evaluate the impact of N_b and BW on system performance. As shown in Figs. 6.24 and 6.25, increasing N_b from 128 ($B_b = 128$) to 256 ($B_b = 232$) does not lead to a significant performance gain in either direction. This might be because $N_b = 128$ already provides sufficient angular resolution to support effective spatial multiplexing. Therefore, further increasing N_b is unnecessary from a performance perspective. We also notice that doubling the bandwidth can boost the performance by at least 30% in both DL and UL, but not as much as at 3.5 GHz. This is because we assume that the total transmit power does not change with the bandwidth, and systems operating in mmWave bands are inherently more power-limited compared to those operating in mid-bands. As shown in Figs. 6.26 and 6.27, increasing the transmit power by 3 dB together with doubling the bandwidth leads to a doubling of the minimum rate on both the DL and the UL.

Finally, we note that, in mmWave, the MBR targets of 50 Mbps on the DL and 10 Mbps

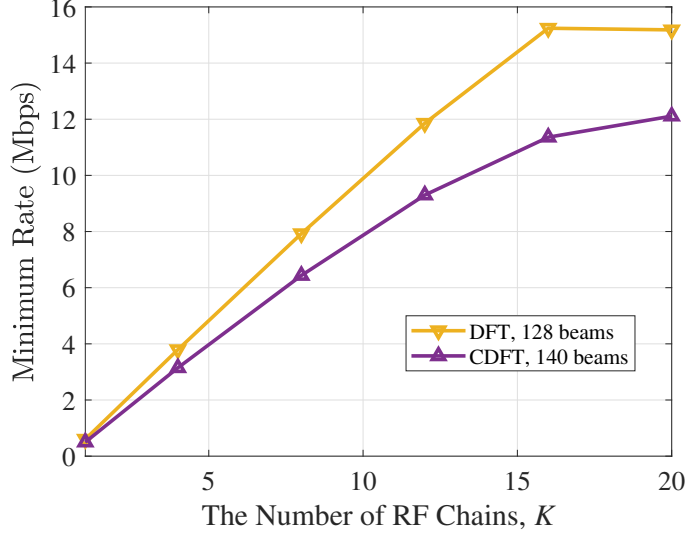


Figure 6.23: MmWave, UL: Impact of types of codebooks on system performance for $N_b = 128$, $N_u = 8$ and $BW = 200$ MHz.

on the UL can be achieved using DFT codebook with either $N_b = 128$, $BW = 200$ MHz, and $K = 12$, or $N_b = 128$, $BW = 100$ MHz, and $K = 16$.

6.3.6 Interference Mitigation Process

In this section, we analyze the impact of the IM process under different B_b values using DFT codebooks in both bands. Fig. 6.28 and Fig. 6.29 show the performance improvement from IM for $B_b = 52$ and $B_b = 64$ on the DL and UL at 3.5 GHz. On the DL, IM has a negligible impact for both B_b values. On the UL, IM yields a 7% gain for $B_b = 64$, but less than 2% for $B_b = 52$ for K large. These results indicate that with the selected B_b from Section 6.3.4, our online RRM solution can closely match the performance of the offline solution (online plus IM) at 3.5 GHz.

Figs. 6.30 and 6.31 present the improvement from IM for $B_b = 128$ and $B_b = 176$ on the DL and UL at 28 GHz. On the DL, IM provides a 7% gain for $B_b = 176$, when K is large, but offers no improvement with $B_b = 128$. The most notable gain appears on the UL with $B_b = 170$, reaching 20% for K large. However, with the selected $B_b = 128$ from Section 6.3.5, the UL gain drops below 2%. Similar to the DL case, a well-chosen B_b can eliminate the need for IM, effectively turning an offline solution into an online one.

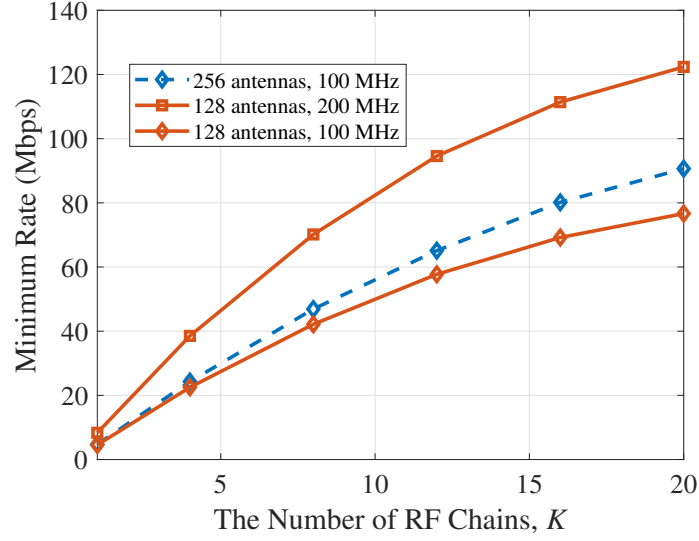


Figure 6.24: MmWave, DL: Impact of bandwidth and number of antennas on system performance as a function of K .

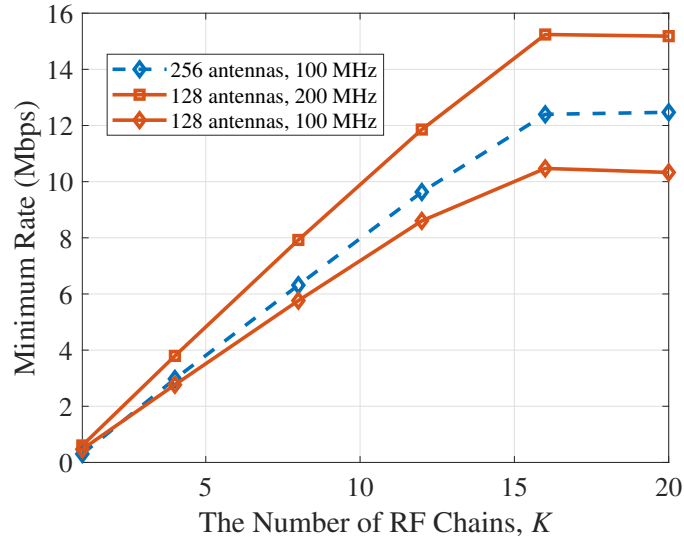


Figure 6.25: MmWave, UL: Impact of bandwidth and number of antennas on system performance as a function of K .

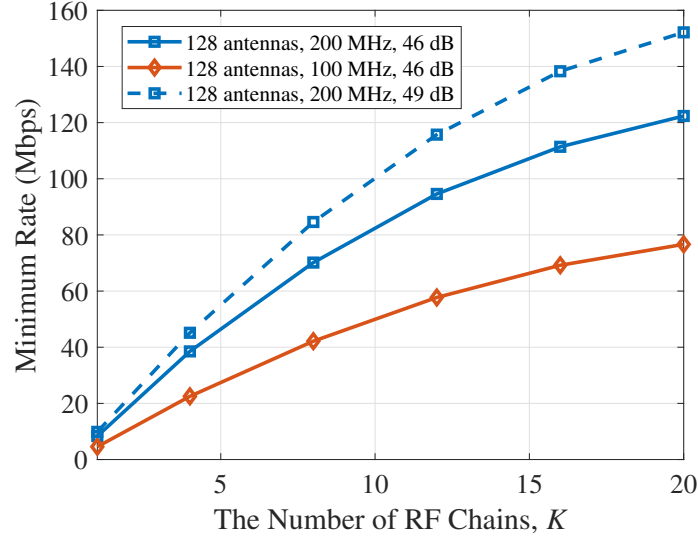


Figure 6.26: MmWave, DL: Impact of bandwidth and number of antennas on system performance as a function of K .

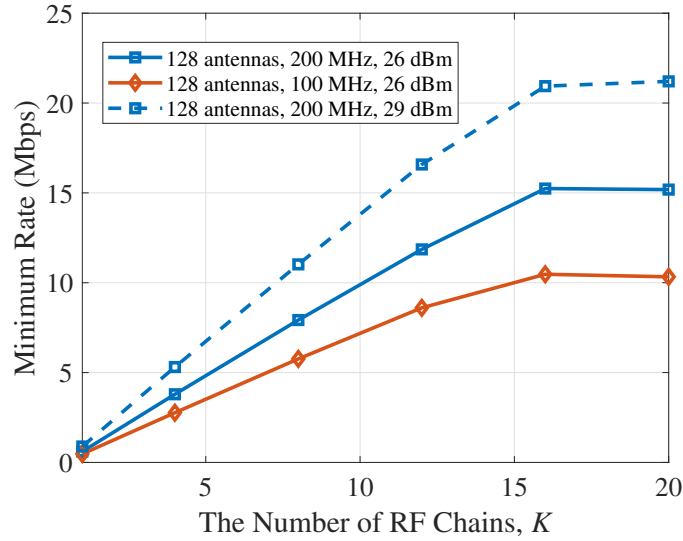


Figure 6.27: MmWave, UL: Impact of bandwidth and number of antennas on system performance as a function of K .

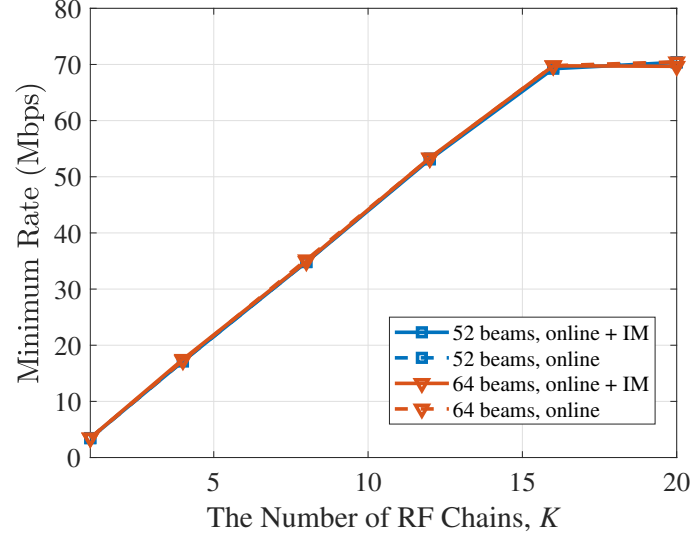


Figure 6.28: Midband, DL: Impact of interference mitigation on system performance as a function of K with $N_b = 64$ and $BW = 60$ MHz.

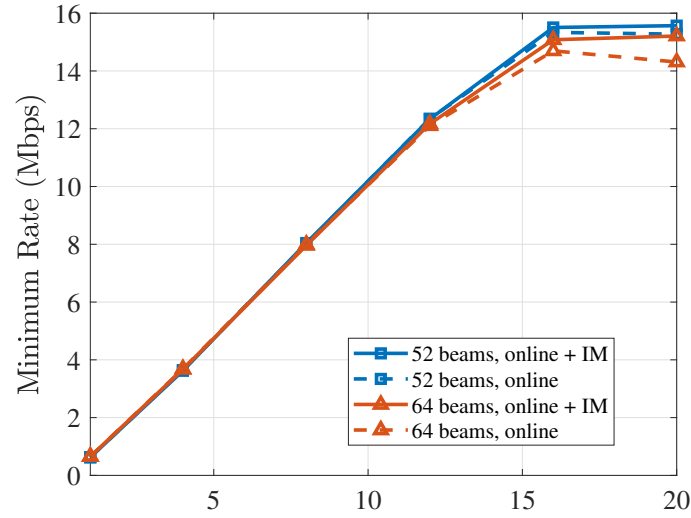


Figure 6.29: Midband, UL: Impact of interference mitigation on system performance as a function of K with $N_b = 64$ and $BW = 60$ MHz.

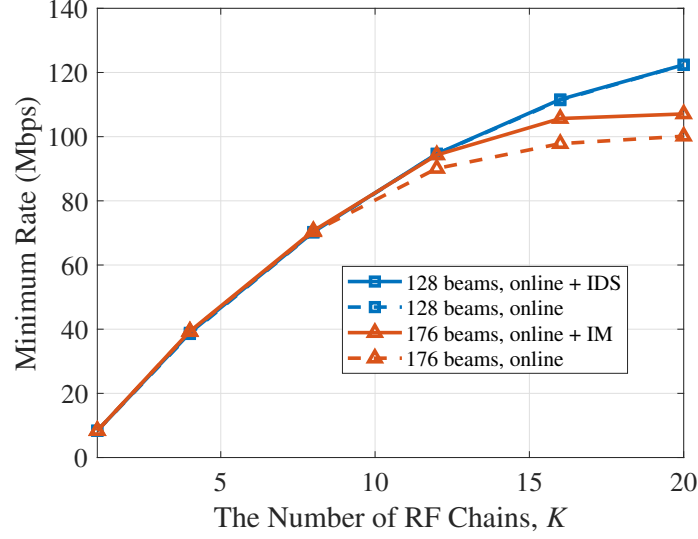


Figure 6.30: MmWave, DL: Impact of interference mitigation on system performance as a function of K with $N_b = 128$ and $BW = 200$ MHz.

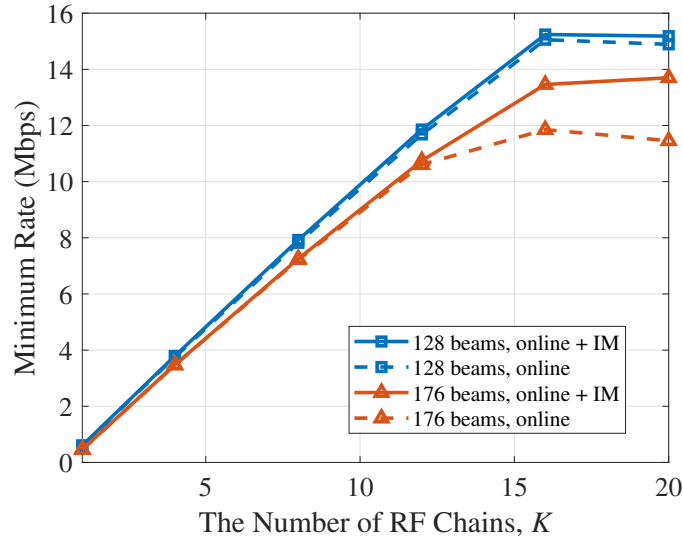


Figure 6.31: MmWave, UL: Impact of interference mitigation on system performance as a function of K with $N_b = 128$ and $BW = 200$ MHz.

6.4 Conclusion

In this chapter, we evaluate the performance of the codebook-based HBF system in rural FWA networks using our proposed RRM solutions. To eliminate the overhead associated with beam alignment, we adopt a position-based beam matching approach for fixed home UEs, which allows us to explore a wide range of codebook sizes without incurring additional training cost. The evaluation focuses on the impact of key system parameters, including codebook type and size (B_b), number of RF chains (K), number of antennas (N_b), and bandwidth (BW).

In the mid-band (3.5 GHz) with $N_b = 64$, we find that DFT codebook with $B_b = 52$ achieves the best performance in both DL and UL. Under this configuration, the HBF system attains 83% of the DL and 80% of the UL performance of a fully DBF system using a greedy up search solution from the literature, while incurring significantly less overhead and hardware cost. We also observe that increasing either N_b or BW can significantly enhance performance. A larger bandwidth improves performance across all values of $K \geq 2$, whereas a larger antenna count (N_b) only yields noticeable gains when $K \geq 16$. The results show that the MBR targets of 50 Mbps (DL) and 10 Mbps (UL) can be satisfied, in midband, under two representative configurations using DFT codebook:

- $N_b = 64$, $B_b = 52$, $BW = 60$ MHz, and $K = 12$,
- $N_b = 32$, $B_b = 26$, $BW = 90$ MHz, and $K = 6$.

In the mmWave band (28 GHz) with $N_b = 128$, DFT codebook remains the best with $B_b = 128$ in both DL and UL. Unlike the mid-band case, increasing N_b further provides limited benefits, as $N_b = 128$ already yields sufficiently narrow beams for effective spatial multiplexing. However, doubling the bandwidth leads to a performance improvement of at least 30%. The results indicate that the MBR targets can be satisfied, in mmWave, under two representative configurations using DFT codebook:

- $N_b = 128$, $B_b = 128$, $BW = 100$ MHz, and $K = 16$,
- $N_b = 128$, $B_b = 128$, $BW = 200$ MHz, and $K = 12$.

Chapter 7

Conclusion

In this thesis, we have investigated RRM for the downlink and uplink of codebook-based hybrid beamforming systems. We consider a practical but overlooked multiple-channel scenario when the number of RF chains at the BS can be lower than the number of UEs in the cell ($K < U$). In the case of $K < U$, new RRM challenges arise due to the need for analog beam selection, which necessitates per-time-slot RRM solutions for both DL and UL, an aspect not addressed in existing research.

Chapter 4 focuses on the DL study. We first formulate and solve a per-time-slot joint RRM optimization problem for different cases, including ZF-DBF versus N-DBF and NEPD versus EPD. Our optimization results show that good performance can be achieved even when K is smaller than U . Additionally, NEPD and ZF-DBF are very important for system performance, which can bring a gain of 22% and 32% compared to EPD and N-DBF, respectively. We also find that having one more pair of preferred beams ($M = 2$) can improve the performance by 22% at the cost of CE overheads. Building on the insights from the offline optimization, we have proposed heuristic online algorithms that offer acceptable computational complexities and can reach 92% of the performance targets obtained from the offline study. We further evaluate the impact of M using our online heuristic with $U = 20$. Setting $M = 2$ yields a notable performance gain of 15% at $K = 20$. However, increasing M beyond 2 to 3 results in only marginal improvement, indicating diminishing returns. We have also investigated the use of Shannon-based rate function in power distribution. Using it in our offline optimization problem yields at least 90% of that using MCS-based rate function, supporting the use of water filling in our online heuristic. However, it is important to consider practical MCS levels in the online heuristics to prevent selecting UEs whose SINR values are too low to achieve a positive rate.

In Chapter 5, we have investigated RRM for the UL. Besides per-time-slot beam selection as in DL, power allocation (PA) for each UE must also be performed at the time-slot level, resulting in a distinct RRM solution from DL. We only studied the case with ZF-DBF because power optimization cannot be decoupled from the RRM problem and solved in advance with intra-cell interference (i.e., with N-DBF) as on the DL. To address the difference, we first formulate a joint RRM optimization problem including beam selection, user selection and PA. This problem can be solved to quasi-optimality for small systems (e.g., $U \leq 10$). To enable scalable planning in larger systems (e.g., 30 UEs, 30 RF chains at the BS, and 3 matched BS beams per UE), we propose an offline heuristic that combines per-beam user selection and PA optimization with load-aware beam selection. This approach achieves at least 93% of the quasi-optimal performance while significantly decreasing the runtime. For online deployment, we propose an online heuristic that replaces the per-beam optimization with a proposed per-beam scheduler, which outperforms the state-of-the-art solution by at least 22%, and uses Shannon-based water filling instead of MCS-based power optimization in the last step. This online solution achieves at least 92% of the offline heuristic with low complexity. Although using Shannon-based rate function is acceptable for PA, it is necessary to consider practical MCS levels in the online heuristic to prevent selecting channels whose SNR values are too low to achieve a positive rate. We have also provided engineering insights into the impact of key system parameters, i.e., the number, K , of RF chains at BS and the number, M , of matched BS beams per UE. Our results indicate that setting $K = 0.6U$ (resp. $K = \frac{1}{3}U$) is sufficient to achieve strong performance, at least 88% of that at $K = U$ for $U = 10$ (resp. $U = 30$). Furthermore, increasing M from 1 to 2 can yield a substantial performance gain, up to 25%, when U is large and K is small (e.g., $U = 32$ and $K = 2$). However, increasing M from 2 to 3 results in negligible additional improvement.

Chapter 6 extends the DL and UL online heuristics from Chapter 4 and Chapter 5 to rural FWA networks in both the mid-band and the mmWave band. In previous chapters, we assumed mobile UEs, where beam alignment (BA) is necessary to match UEs with beams. This introduces training overhead that scales linearly with the codebook size, thereby limiting the size of the codebook that can be used. In FWA networks, where home UEs are fixed, we adopt a position-based beam matching approach to eliminate the training overhead from BA, allowing for the use of larger codebook sizes. A main focus of this chapter is to evaluate the impact of codebook type and size (B_b). In both bands, DFT codebook realizes strong performance with our RRM algorithms. However, the selected value of B_b varies by band. At 3.5 GHz, performance plateaus at $B_b = 52$ with $N_b = 64$, while at 28 GHz, a plateau is observed at $B_b = 128$ with $N_b = 128$. With the selected codebook type and size, the HBF system can achieve 83% of the DL and 80% of the UL

performance of a fully DBF system with significantly lower overhead and hardware costs at 3.5 GHz. We also analyze the impact of BW and N_b on performance. Increasing BW consistently improves performance in both bands for $K \geq 2$, whereas increasing N_b offers significant gains only at 3.5 GHz when $K \geq 16$. Our results show that the MBR targets of 50 Mbps (DL) and 10 Mbps (UL) can be satisfied under two representative configurations using DFT codebook:

- $N_b = 64$, $B_b = 52$, $BW = 60$ MHz, and $K = 12$ at 3.5 GHz,
- $N_b = 32$, $B_b = 26$, $BW = 90$ MHz, and $K = 6$ at 3.5 GHz,
- $N_b = 128$, $B_b = 128$, $BW = 100$ MHz, and $K = 16$ at 28 GHz,
- $N_b = 128$, $B_b = 128$, $BW = 200$ MHz, and $K = 12$ at 28 GHz.

7.1 Future Research Directions

This thesis has focused on fully-connected HBF systems. An interesting direction for future work is to investigate partially-connected HBF architectures, where each RF chain is connected to a subset of the antenna elements. This configuration can further reduce hardware complexity. Analyzing the performance gap between fully- and partially-connected HBF systems under varying system parameters would provide valuable insights. Additionally, exploring different sub-array partitioning strategies for the antenna array needs further investigation.

Another important aspect is to consider the impact of limited-resolution phase shifters in HBF systems. Although our RRM solutions are agonistic to the quality of phase shifters, reduced phase resolution can degrade codebook quality, potentially affecting overall system performance. Future work should quantify this degradation and investigate trade-offs between performance and phase shifter resolution.

Another direction for future work is to consider scenarios where each UE is equipped with multiple RF chains, thus supporting multiple data streams per UE.

Finally, the beam squint effect significantly degrades the performance when both the fractional bandwidth (bandwidth over carrier frequency) and the antenna array size are large, for instance, with 400 MHz bandwidth and 128 antennas at 28 GHz. A thorough analysis of this effect is essential for providing valuable insights into practical system deployment. Interestingly, the beam squint effect also inspires the development of a new BF

architecture, known as true-time-delay hybrid beamforming. This structure can generate distinct analog beams across different frequency blocks within the same time slot. The resulting RRM challenges are fundamentally different from conventional systems and merit dedicated investigation.

References

- [1] S. Rangan, T. S. Rappaport, and E. Erkip, “Millimeter-wave cellular wireless networks: Potentials and challenges,” *Proceedings of the IEEE*, vol. 102, no. 3, pp. 366–385, 2014.
- [2] T. L. Marzetta, E. G. Larsson, H. Yang, and H. Q. Ngo, *Fundamentals of Massive MIMO*. Cambridge, UK: Cambridge University Press, 2016.
- [3] A. Khalili, S. Shahsavari, F. Shirani, E. Erkip, and Y. C. Eldar, “On throughput of millimeter wave MIMO systems with low resolution adcs,” in *ICASSP 2020 - IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 5255–5259.
- [4] A. F. Molisch, V. V. Ratnam, S. Han, Z. Li, S. L. H. Nguyen, L. Li, and K. Haneda, “Hybrid beamforming for massive mimo: A survey,” *IEEE Communications Magazine*, vol. 55, no. 9, pp. 134–141, 2017.
- [5] M. Giordani, M. Polese, A. Roy, D. Castor, and M. Zorzi, “A tutorial on beam management for 3GPP NR at mmwave frequencies,” *IEEE Communications Surveys & Tutorials*, vol. 21, no. 1, pp. 173–196, 2019.
- [6] 3GPP, “TS 38.211 V17.4.0: NR; Physical Channels and Modulation,” 3rd Generation Partnership Project (3GPP), Technical Specification 38.211, Sep. 2024, release 17.
- [7] S. M. Hosseini, S. Shahsavari, and C. Rosenberg, “Multi-user scheduling in hybrid millimeter wave massive mimo systems,” 2021. [Online]. Available: <https://arxiv.org/abs/2109.05681>
- [8] X. Gao, X. Wu, Z. Zhang, and D. Liu, “Low complexity joint user scheduling and hybrid beamforming for mmwave massive mimo systems,” in *2020 IEEE 31st Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, 2020, pp. 1–6.

- [9] A. Alkhateeb, G. Leus, and R. W. Heath, "Limited feedback hybrid precoding for multi-user millimeter wave systems," *IEEE Transactions on Wireless Communications*, vol. 14, no. 11, pp. 6481–6494, 2015.
- [10] G. Yue, B. Zhang, and X.-F. Qi, "Joint beam management and user scheduling for massive MIMO with hybrid beamforming and limited feedback," in *2020 IEEE 92nd Vehicular Technology Conference (VTC2020-Fall)*, 2020, pp. 1–5.
- [11] S. He, Y. Wu, D. W. K. Ng, and Y. Huang, "Joint optimization of analog beam and user scheduling for millimeter wave communications," *IEEE Communications Letters*, vol. 21, no. 12, pp. 2610–2613, 2017.
- [12] Y. Zhu and T. Yang, "Low complexity hybrid beamforming for uplink multiuser mmwave mimo systems," in *2017 IEEE Wireless Communications and Networking Conference (WCNC)*, 2017, pp. 1–6.
- [13] J. Li, L. Xiao, X. Xu, and S. Zhou, "Robust and low complexity hybrid beamforming for uplink multiuser mmwave mimo systems," *IEEE Communications Letters*, vol. 20, no. 6, pp. 1140–1143, 2016.
- [14] Z. Wang, M. Li, Q. Liu, and A. L. Swindlehurst, "Hybrid precoder and combiner design with low-resolution phase shifters in mmwave mimo systems," *IEEE Journal of Selected Topics in Signal Processing*, vol. 12, no. 2, pp. 256–269, 2018.
- [15] C. Han, J. Wang, J. Wang, and L. Bai, "Hybrid beamforming design for uplink mmwave systems with a predefined low-resolution codebook," *Journal of Communications and Information Networks*, vol. 4, no. 3, pp. 1–8, 2019.
- [16] H. Li, T. Q. Wang, X. Huang, J. A. Zhang, and Y. J. Guo, "Low-complexity multiuser receiver for massive hybrid array mmwave communications," *IEEE Transactions on Communications*, vol. 67, no. 5, pp. 3512–3524, 2019.
- [17] Z. Wang, M. Li, R. Liu, and Q. Liu, "Joint user association and hybrid beamforming designs for cell-free mmwave mimo communications," *IEEE Transactions on Communications*, vol. 70, no. 11, pp. 7307–7321, 2022.
- [18] A. Lappalainen and C. Rosenberg, "Bridging the digital divide: Success depends on content provider and application developer involvement [point of view]," *Proceedings of the IEEE*, vol. 109, no. 1, pp. 2–10, 2021.

- [19] H. M. Douglas, *Key 5G Physical Layer Technologies: Enabling Mobile and Fixed Wireless Access*. Springer Nature, 2022.
- [20] A. Lappalainen and C. Rosenberg, “Can 5g fixed broadband bridge the rural digital divide?” *IEEE Communications Standards Magazine*, vol. 6, no. 2, pp. 79–84, 2022.
- [21] 3rd Generation Partnership Project (3GPP), “Study on scenarios and requirements for next generation access technologies,” 3GPP, Tech. Rep. TR 38.913, 2020, version 16.0.0. [Online]. Available: https://www.3gpp.org/ftp/Specs/archive/38_series/38.913/38913-g00.zip
- [22] Federal Communications Commission (FCC), “Fixed broadband deployment report,” FCC, Tech. Rep., 2021, fCC 21-18. [Online]. Available: <https://www.fcc.gov/reports-research/reports/broadband-progress-reports/>
- [23] A. Ericsson, L. Falconetti, H. Olofsson, J. Edstam, and T. Dahlberg, “Closing the digital divide with mmwave extended range for fwa,” *Ericsson Technology Review*, vol. 2022, no. 11, pp. 2–11, 2022.
- [24] J. Huang, V. Subramanian, R. Agrawal, and R. Berry, “Joint scheduling and resource allocation in uplink ofdm systems for broadband wireless access networks,” *IEEE Journal on Selected Areas in Communications*, vol. 27, no. 2, pp. 226–234, 2009.
- [25] C. Wengerter, J. Ohlhorst, and A. G. E. von Elbwart, “Fairness and throughput analysis for generalized proportional fair frequency scheduling in ofdma,” in *2005 IEEE 61st Vehicular Technology Conference*, vol. 3, 2005, pp. 1903–1907.
- [26] D. Nace and M. Pioro, “Max-min fairness and its applications to routing and load-balancing in communication networks: A tutorial,” *IEEE Communications Surveys & Tutorials*, vol. 10, no. 4, pp. 5–17, 2008.
- [27] J. Song, J. Choi, and D. J. Love, “Common codebook millimeter wave beam design: Designing beams for both sounding and communication with uniform planar arrays,” *IEEE Transactions on Communications*, vol. 65, no. 4, pp. 1859–1872, 2017.
- [28] L. R. Rabiner, B. Gold, and C. Aviles, “Some comparisons between chebyshev and other windows,” *Bell System Technical Journal*, vol. 49, no. 6, pp. 1513–1541, 1970.
- [29] E. Björnson, L. V. der Perre, S. Buzzi, and E. G. Larsson, “Massive MIMO in sub-6 ghz and mmwave: Physical, practical, and use-case differences,” *IEEE Wireless Communications*, vol. 26, no. 2, pp. 100–108, 2019.

- [30] A. Khalili, F. Shirani, E. Erkip, and Y. C. Eldar, "On multiterminal communication over MIMO channels with one-bit adcs at the receivers," in *2019 IEEE International Symposium on Information Theory (ISIT)*, 2019, pp. 2708–2712.
- [31] F. Sotirani and W. Yu, "Hybrid digital and analog beamforming design for large-scale antenna arrays," *IEEE Journal of Selected Topics in Signal Processing*, vol. 10, no. 3, pp. 501–513, 2016.
- [32] H. Shokri-Ghadikolaei, C. Fischione, G. Fodor, P. Popovski, and M. Zorzi, "Millimeter wave cellular networks: A MAC layer perspective," *IEEE Transactions on Communications*, vol. 63, no. 10, pp. 3437–3458, 2015.
- [33] T. E. Bogale, L. B. Le, A. Haghighat, and L. Vandendorpe, "On the number of RF chains and phase shifters, and scheduling design with hybrid analog–digital beamforming," *IEEE Transactions on Wireless Communications*, vol. 15, no. 5, pp. 3311–3326, 2016.
- [34] G. Hegde and M. Pesavento, "Joint user selection and hybrid analog-digital beamforming in massive MIMO systems," in *2018 IEEE 10th Sensor Array and Multichannel Signal Processing Workshop (SAM)*, 2018, pp. 16–20.
- [35] H. Wu, D. Liu, W. Wu, C. Na, and M. Liu, "A low complexity two-stage user scheduling scheme for mmwave massive MIMO hybrid beamforming systems," in *2017 IEEE/CIC International Conference on Communications in China (ICCC)*, 2017, pp. 1–6.
- [36] G. Kwon and H. Park, "A joint scheduling and millimeter wave hybrid beamforming system with partial side information," in *2016 IEEE International Conference on Communications (ICC)*, 2016, pp. 1–6.
- [37] L. Qianrui, "Hybrid precoding for wideband multi-user MIMO millimeter wave system," in *2019 IEEE Wireless Communications and Networking Conference (WCNC)*, 2019, pp. 1–6.
- [38] 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. 2951–2966, 2021.
- [39] A. Hussein, C. Rosenberg, and P. Mitran, "Hybrid NOMA in multi-cell networks: From a centralized analysis to practical schemes," *IEEE/ACM Transactions on Networking*, vol. 30, no. 3, pp. 1180–1193, 2022.

- [40] F. Sohrabi and W. Yu, “Hybrid analog and digital beamforming for mmwave ofdm large-scale antenna arrays,” *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 7, pp. 1432–1443, 2017.
- [41] Q. Li and H. Nouredine, “Hybrid sum rate maximization beamforming for multi-user massive mimo millimeter wave system,” in *2017 IEEE 28th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, 2017, pp. 1–6.
- [42] S. S. Christensen, R. Agarwal, E. D. Carvalho, and J. M. Cioffi, “Weighted sum-rate maximization using weighted mmse for mimo-bc beamforming design,” *IEEE Transactions on Wireless Communications*, vol. 7, no. 12, pp. 4792–4799, 2008.
- [43] 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.
- [44] 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.
- [45] Y. Özcan, J. Oueis, C. Rosenberg, R. Stanica, and F. Valois, “Robust planning and operation of multi-cell homogeneous and heterogeneous networks,” *IEEE Transactions on Network and Service Management*, vol. 17, no. 3, pp. 1769–1784, 2020.
- [46] 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.
- [47] X. Guan, Z. Chen, Y. Kang, Q. Yan, and Q. Shi, “On user scheduling for fixed wireless access via channel statistics,” *IEEE Transactions on Mobile Computing*, vol. 24, no. 5, pp. 4040–4052, 2025.
- [48] N. Chatur, T. Bose, and A. Adhya, “Planning cost-efficient fiwi access network with joint deployment of fwa and ftth,” *IEEE Transactions on Communications*, vol. 72, no. 9, pp. 5688–5703, 2024.
- [49] G. Morozov, A. Davydov, and D. Dikarev, “Line-of-sight mimo for 5g millimeter wave fixed wireless access applications,” in *GLOBECOM 2024 - IEEE Global Communications Conference*, 2024, pp. 5441–5446.

- [50] 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.
- [51] A. Brighente, J. Gambini, and S. Tomasin, “Modular hybrid beamforming for mmwave fixed wireless access,” *IEEE Transactions on Communications*, vol. 68, no. 8, pp. 5145–5158, 2020.
- [52] 3rd Generation Partnership Project (3GPP), “Study on channel model for frequencies from 0.5 to 100 ghz,” 3GPP, Tech. Rep. TR 38.901, 2022, version 17.0.0. [Online]. Available: https://www.3gpp.org/ftp/Specs/archive/38_series/38.901/38901-g00.zip
- [53] Z. Xiao, H. Dong, L. Bai, P. Xia, and X.-G. Xia, “Enhanced channel estimation and codebook design for millimeter-wave communication,” *IEEE Transactions on Vehicular Technology*, vol. 67, no. 10, pp. 9393–9405, 2018.
- [54] V. Desai, L. Krzymien, P. Sartori, W. Xiao, A. Soong, and A. Alkhateeb, “Initial beamforming for mmwave communications,” in *2014 48th Asilomar Conference on Signals, Systems and Computers*, 2014, pp. 1926–1930.
- [55] M. Giordani, M. Mezzavilla, C. N. Barati, S. Rangan, and M. Zorzi, “Comparative analysis of initial access techniques in 5g mmwave cellular networks,” in *2016 Annual Conference on Information Science and Systems (CISS)*, 2016, pp. 268–273.
- [56] C. N. Barati, S. A. Hosseini, M. Mezzavilla, T. Korakis, S. S. Panwar, S. Rangan, and M. Zorzi, “Initial access in millimeter wave cellular systems,” *IEEE Transactions on Wireless Communications*, vol. 15, no. 12, pp. 7926–7940, 2016.
- [57] S. Shahsavari, M. A. A. Khojastepour, and E. Erkip, “Beam training optimization in millimeter-wave systems under beamwidth, modulation and coding constraints,” in *2019 IEEE 30th Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, 2019, pp. 1–6.
- [58] S. Shahsavari, M. A. Amir Khojastepour, and E. Erkip, “Robust beam tracking and data communication in millimeter wave mobile networks,” in *2019 International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOPT)*, 2019, pp. 1–8.
- [59] A. Khalili, S. Shahsavari, M. A. A. Khojastepour, and E. Erkip, “On optimal multi-user beam alignment in millimeter wave wireless systems,” in *2020 IEEE International Symposium on Information Theory (ISIT)*, 2020, pp. 1590–1595.

- [60] F. Gómez-Cuba, T. Zugno, J. Kim, M. Polese, S. Bahk, and M. Zorzi, “Hybrid beamforming in 5g mmwave networks: A full-stack perspective,” *IEEE Transactions on Wireless Communications*, vol. 21, no. 2, pp. 1288–1303, 2022.
- [61] 3GPP, “LTE; Evolved Universal Terrestrial Radio Access (E-UTRA); Physical Layer Procedures (3GPP TS 36.213 version 12.4.0 Release 12),” 3rd Generation Partnership Project (3GPP), Technical Specification 36.213, Mar. 2015, release 12.
- [62] J. Du and R. A. Valenzuela, “How much spectrum is too much in millimeter wave wireless access,” *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 7, pp. 1444–1458, 2017.
- [63] J. Song, J. Choi, and D. J. Love, “Codebook design for hybrid beamforming in millimeter wave systems,” in *2015 IEEE International Conference on Communications (ICC)*, 2015, pp. 1298–1303.
- [64] Y. Chen, D. Chen, T. Jiang, and L. Hanzo, “Channel-covariance and angle-of-departure aided hybrid precoding for wideband multiuser millimeter wave mimo systems,” *IEEE Transactions on Communications*, vol. 67, no. 12, pp. 8315–8329, 2019.
- [65] M. R. Akdeniz, Y. Liu, M. K. Samimi, S. Sun, S. Rangan, T. S. Rappaport, and E. Erkip, “Millimeter wave channel modeling and cellular capacity evaluation,” *IEEE Journal on Selected Areas in Communications*, vol. 32, no. 6, pp. 1164–1179, 2014.
- [66] M. K. Samimi and T. S. Rappaport, “3-d millimeter-wave statistical channel model for 5g wireless system design,” *IEEE Transactions on Microwave Theory and Techniques*, vol. 64, no. 7, pp. 2207–2225, 2016.
- [67] 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.
- [68] 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.
- [69] R. Fourer, D. M. Gay, and B. W. Kernighan, *AMPL: A Mathematical Programming Language*. Danvers, MA, USA: Boyd & Fraser Publishing Company, 1993.
- [70] J. L. Gearhart, K. L. Adair, J. D. Durfee, K. A. Jones, N. Martin, and R. J. Detry, “Comparison of open-source linear programming solvers,” Sandia National

Laboratories, Albuquerque, NM, USA, Tech. Rep., 10 2013. [Online]. Available: <https://www.osti.gov/biblio/1104761>

- [71] C. T. Neil, M. Shafi, P. J. Smith, P. A. Dmochowski, and J. Zhang, “Impact of microwave and mmwave channel models on 5g systems performance,” *IEEE Transactions on Antennas and Propagation*, vol. 65, no. 12, pp. 6505–6520, 2017.
- [72] G. R. MacCartney and T. S. Rappaport, “Rural macrocell path loss models for millimeter wave wireless communications,” *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 7, pp. 1663–1677, 2017.
- [73] N. Saba, L. Mela, M. U. Sheikh, J. Salo, K. Ruttik, and R. Jäntti, “Rural macrocell path loss measurements for 5g fixed wireless access at 26 ghz,” in *2021 IEEE 4th 5G World Forum (5GWF)*, 2021, pp. 328–333.
- [74] M. K. Samimi, G. R. MacCartney, S. Sun, and T. S. Rappaport, “28 ghz millimeter-wave ultrawideband small-scale fading models in wireless channels,” in *2016 IEEE 83rd Vehicular Technology Conference (VTC Spring)*, 2016, pp. 1–6.
- [75] 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.