Downlink Resource Allocation in Cooperative Wireless Networks

JINGYA LI

Communication Systems Group
Department of Signals and Systems
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden 2014
Thesis for the degree of Doctor of Philosophy

Downlink Resource Allocation in Cooperative Wireless Networks

Jingya Li

Communication Systems Group
Department of Signals and Systems
Chalmers University of Technology
Gothenburg 2014
Li, Jingya
Downlink Resource Allocation in Cooperative Wireless Networks.

ISBN: 978-91-7597-123-0
Doktorsavhandlingar vid Chalmers tekniska högskola
Ny Serie No. 3804
ISSN 0346-718X

Communication Systems Group
Department of Signals and Systems
Chalmers University of Technology
SE-412 96 Gothenburg, Sweden
Telephone: + 46 (0)31-772 4826

Copyright ©2014 Jingya Li except where otherwise stated.
All rights reserved.

This thesis has been prepared using \LaTeX.

Front Cover: Multi-cell joint transmission in a centralized coordinated multi-point cluster.

Printed by Chalmers Reproservice,
Gothenburg, Sweden, December 2014.
To my parents and Gongpei
Abstract

Wireless cooperative networks, which is based on exploiting coordination among multiple access nodes, has been considered as a promising approach to improve the spectral efficiency, reduce the energy consumption, and extend the network coverage of future wireless communication systems. In practice, the actual benefit of multi-node cooperation is affected by a variety of factors, including the quality of channel state information (CSI), the constraints on the feedback and backhaul links, hardware impairments, resource allocation and data processing schemes. This thesis investigates the design of resource allocation algorithms and the performance of different coordinated transmission schemes for downlink wireless cooperative systems under practical constraints.

First, we consider multi-node cooperation in homogeneous cooperative networks. In [Paper A], a power allocation scheme is proposed for a worst case scenario, where the carrier phases between base stations (BSs) are un-synchronized so that joint transmission must be performed without precoding. We show that in this scenario, joint transmission happens with higher probability when the maximum transmit power is high, or the users are in the overlapped cell-edge area. In practice, the network is divided into clusters of coordinated BSs, and the cooperation gain is limited by the inter-cluster interference. In [Paper B], different fractional frequency reuse schemes are proposed to coordinate inter-cluster interference. Simulation results show that the proposed schemes can efficiently reduce the inter-cluster interference and provide considerable performance improvement in terms of both the cell-edge and cell-average user data rate. [Paper C] compares different coordinated transmission schemes, considering the effects of the feedback and backhaul latency. Compared to zero-forcing coherent joint transmission, we show that non-coherent joint transmission and coordinated scheduling are more robust to channel uncertainty.

The second part of the thesis focuses on heterogeneous cooperative networks. In [Paper D], we analyze the performance of amplify-and-forward two-way relaying with in-phase and quadrature-phase imbalance (IQI) at the relay node. Different design guidelines and power allocation schemes are proposed to improve the system reliability under a total transmit power constraint. [Paper E] investigates adaptive power allocation for hybrid automatic repeat request based relay networks. Our results demonstrate that depending on the relay positions and the total power budget, the system should switch between the single-node transmission mode and the joint transmission mode, in order to minimize the outage probability. Finally, [Paper F] studies the joint design of precoding and load balancing for energy-efficient small cell networks with imperfect CSI. An optimal BS association condition is parameterized, which reveals how it is impacted by different system parameters. Our results also show that putting BSs into sleep mode by proper load balancing is an important solution for energy savings in heterogeneous networks.

Keywords: Coordinated multi-point transmission, fractional frequency reuse, imperfect channel state information, imperfect synchronization, I/Q imbalance, load balancing, precoding design, relaying, resource allocation.
List of Included Publications

This thesis is based on the following appended papers:


List of Additional Related Publications

Publications by the author not included in this thesis:


Contents

Abstract i
List of Included Publications iii
List of Additional Related Publications v
Acknowledgements xiii
Acronyms xv

I Overview 1
1 Introduction 3

2 Cooperative Wireless Networks 7
  2.1 Cooperative Scenarios 7
  2.1.1 Homogeneous Cooperative Networks 7
  2.1.2 Heterogeneous Cooperative Networks 8
  2.2 Cooperative techniques 9
    2.2.1 Inter-Cell Interference Coordination 10
    2.2.2 Multi-Node Cooperative Transmission 12
    2.2.3 Load Balancing 13
    2.2.4 Cell On/Off 14
    2.2.5 Cell Clustering 15
  2.3 Challenges and Difficulties 17
    2.3.1 Non-Ideal Feedback 17
    2.3.2 Non-Ideal Backhaul 18
    2.3.3 Hardware Impairments 18
    2.3.4 Resource Optimization 19
    2.3.5 Cluster-Edge Effect 20

3 Radio Resource Allocation 21
  3.1 System Model 21
  3.2 Problem Formulation 23
    3.2.1 Objectives 23
    3.2.2 Constraints 24
  3.3 Radio Resource Optimization 25
3.3.1 Power Minimization ........................................... 25
3.3.2 Worst SINR Maximization .................................... 27
3.3.3 Sum Rate Maximization ....................................... 28
3.4 Remarks on Practical Constraints ............................... 28

4 Conclusions and Future Work .................................. 31
4.1 Contributions ................................................... 31
4.1.1 Resource Allocation in Homogeneous Cooperative Networks ... 31
4.1.2 Resource Allocation in Heterogeneous Cooperative Networks ... 33
4.2 Future Work .................................................... 36

References ......................................................... 37

II Included papers ............................................... 45

A Power Allocation for Two-Cell Two-User Joint Transmission .......................... A1
A.1 Introduction .................................................... A2
A.2 System Model .................................................. A3
A.3 Optimal Transmit Power Allocation ................................... A4
A.4 Joint Transmission Analysis ...................................... A7
A.5 Numerical Results ............................................... A8
A.6 Conclusions ..................................................... A10
References ........................................................ A10

B Resource Allocation for Clustered Network MIMO OFDMA Systems ................... B1
B.1 Introduction ..................................................... B2
B.2 System Model and Problem Formulation ................................... B5
B.3 Inter-Cluster Interference Mitigation .................................... B8
B.3.1 Cooperative Frequency Reuse Scheme 1 ........................ B8
B.3.2 Cooperative Frequency Reuse Scheme 2 ........................ B10
B.4 Joint Scheduling and Power Allocation .................................. B10
B.5 Simulation Results ............................................... B12
B.5.1 Frequency Partition and User Partition .......................... B13
B.5.2 Performance Analysis .......................................... B17
B.6 Conclusions ..................................................... B22
References ........................................................ B23

C Performance Evaluation of Coordinated Multi-Point Transmission Schemes with Predicted CSI ........................................... C1
C.1 Introduction ..................................................... C2
C.2 System Model ................................................... C3
C.3 CoMP Transmission Schemes ....................................... C4
C.3.1 Coherent Joint Transmission .................................... C4
C.3.2 Non-coherent Joint Transmission .............................. C5
C.3.3 Coordinated Scheduling ........................................ C6
C.4 Simulation Results ............................................... C7
C.4.1 Sum Rate Performance with Perfect CSI ......................... C7
Joint Precoding and Load Balancing Optimization for Energy-Efficient Heterogeneous Networks

F.1 Introduction ................................................. F2
F.2 System and Signal Model .................................. F5
   F.2.1 Power Consumption Model ....................... F6
   F.2.2 Aggregated Received SINR ....................... F6
   F.2.3 Problem Formulation .............................. F7
F.3 Optimal Precoding and Load Balancing ................. F8
   F.3.1 Structure of the Optimal Load Balancing .... F11
F.4 Iterative Heuristic Algorithm Design ................ F14
F.5 Numerical Results ........................................ F16
F.6 Conclusions .............................................. F21
References .................................................... F23
I would like to express my deepest gratitude to my main supervisor, Associate Prof. Tommy Svensson, for giving me this opportunity to pursue doctoral studies in the Communication Systems group, and the freedom and encouragement to work on so many interesting research problems. Your knowledge, guidance, and constant support have been fundamental to my growth in every aspect during these four years at Chalmers. This gratitude also goes to my co-supervisors, Prof. Thomas Eriksson, for always pointing me in good directions, for your kind support, guidance and all the nice research discussions. Many thanks to the head of our group, Prof. Erik Störm, for creating such a friendly and joyful research atmosphere.

I would also like to use the opportunity to send my gratitude to some distinguished researchers who have mentored me over the years. Thank you, Prof. Xiaofeng Tao and Associated Prof. Xiaodong Xu, for your guidance and support during my master study period, and for the continuous collaboration during these years. Thank you, Associated Prof. Carmen Botella, for taking great care of me during the first year of my PhD and shared your insights on limited feedback and backhaul design for CoMP systems. Thank you, Dr. Agisilaos Papadogiannis, for the stimulating discussions and collaborations during the second and third years of my PhD study. Special thanks to Prof. Michail Matthaiou, for our fruitful collaboration over the last two years, and for all the help he has provided me outside of my research. I am also grateful to Prof. Mikael Sernad, for creating such a nice discussion and learning opportunity during our VR project meetings. Thank you, Agisilaos, Rikke, Tilak, Behrooz, Nima, Anna and Annika, for all the discussions and fruitful collaborations we have had during the VR project meetings. I would also like to mention my deepest gratitude to Prof. Mérouane Debbah, for giving me the opportunity to visit your group at Supelec. Your sharp mind in research discussions has always been a source of inspiration for me. I am also thankful to Assistant Prof. Emil Bjöösson, who was a post-doc in Supelec during my visit, for the stimulating discussions and fruitful collaborations, for reading the rough draft of my papers and taking care of me during my visit.

I would also like to thank the current and former members of the Communication Systems group for creating a pleasant research atmosphere. In particular, I would like to thank my office-mates Erik, Tilak, Rajet and Reza for all the nice discussions we have had inside or outside of research, and for providing me all kinds of information. Many thanks to Lars for the computer support and to Agneta, Natasha and Madeleine for all their help. I would also like to thank all my Chinese friends in Gothenburg for all the great moments we have experienced together.

Finally, I would like to express my sincerest gratitude to my family for their constant support, love and encouragement over the years. My warmest appreciation belongs to
my husband Gongpei. His understanding, support and love has been the source of my courage and joy.

Jingya Li
Gothenburg, October 2014

This work has been supported by Swedish Governmental Agency for Innovation Systems (VINNOVA), the Swedish Research Council (VR), and the Seventh Framework Program (EU FP7).
Acronyms

ABS: Almost blank subframe
3GPP: 3rd Generation Partnership Project
BBU: Baseband unit
bps: bit per second
BS: Base station
CoMP: Coordinated multi-point
CRS: common reference signals
CSI: Channel state information
CSIT: Channel state information at the transmitter
CU: Control unit
eICIC: Enhanced inter-cell interference coordination
FDD: Frequency division duplex
FeICIC: Further enhanced inter-cell interference coordination
HARQ: Hybrid automatic repeat request
ICI: Inter-cell interference
ICIC: Inter-cell interference coordination
IQI: In-phase and quadrature-phase imbalance
KKT: Karush-Kuhn-Tucker
LTE: Long term evolution
LoS: Line-of-sight
MIMO: Multiple input multiple output
OFDMA: Orthogonal frequency division multiple access
PDCCH: Physical downlink control channel
PDSCH: Physical downlink shared channel
PSS: primary synchronization signals
QoS: Quality of service
RF: Radio frequency
RRH: Remote radio head
SFR: Soft frequency reuse
SINR: Signal to interference plus noise ratio
SSS: Secondary synchronization signals
TDD: Time division duplex
ZF: Zero-forcing
Part I
Overview
Chapter 1

Introduction

Over the past few decades, wireless communication has played an important role as a way to let people get and share information with each other anywhere and anytime. Studies have shown that the data traffic volume in wireless communication networks continues to grow at an impressive rate, mainly driven by the uptake of smart devices and apps. Based on extrapolations from Ericsson [1] and Cisco [2], it is possible to conclude that beyond 2020, wireless communication systems will have to support more than 1,000 times today's traffic volume.

The raised user expectations of quality of service (QoS) as well as the rapid growth of the data traffics impose very different requirements on future wireless communication networks, such as higher system throughput and spectral efficiency, sufficient data rate and speed to run apps with an affordable price. At the same time, the network also needs to provide a homogenous QoS distribution over the communication area in order to guarantee fairness among the users independent of their locations [3, 4]. Moreover, energy efficiency also becomes one of the important key performance indicators for the design of future wireless communication systems, in order to achieve a low cost and green networked society.

In traditional multi-cell communication networks, system spectral efficiency is mainly limited by inter-cell interference (ICI), which is caused by the transmission from other cells on the same time-frequency resource block [5]. The presence of ICI especially degrades the performance and affects the experience of the users located in the cell-edge areas. This is because that the cell-edge users typically receive weak desired signals from their connected cell, while suffering strong interference from neighboring cells.

Under current macro-cell network deployment, one of the promising approaches for combating ICI is to introduce coordination between base stations (BSs). In the 3rd generation partnership project (3GPP) long term evolution (LTE) systems, inter-BS signaling can be accomplished over the X2 interface between BSs. Hence, inter-cell interference coordination or avoidance was proposed as a key technique to deal with the ICI issue [3, 6]. The common theme of inter-cell interference coordination or avoidance in LTE is to apply restrictions to the time or frequency or power resources available in a cell in a coordinated way. Such restrictions provide improvement in the ratio of the desired received signal power over interference and noise power on the corresponding resource blocks in the neighboring cells. Consequently, the cell-edge data rates and the cell coverage can be improved. However, it should be pointed out that the ICI is reduced at the expense of the available resources that can be scheduled in each cell, leading to a degradation in the system peak
Introduction

or sum throughput. Instead of coordinating ICI by restricting how radio resources are used in each cell, multi-cell advanced coordination and joint transmission can be used as a more proactive way to handle the ICI issue with much tighter multi-BS cooperations [7, 8]. The technology component “Coordinated multi-point (CoMP) transmission/reception”, which has the same basic principles, is considered in 3GPP LTE-Advanced [9]. Based on the channel state information (CSI) and/or the user data shared via backhaul links between multiple transmission nodes, CoMP operation performs dynamic coordination among multiple BSs. Depending on the levels of multi-BS cooperation, CoMP techniques can either coordinate or exploit the interference in order to improve the coverage of high data rates, the cell-edge throughput, as well as the system throughput.

Besides further enhancing the performance of existing homogeneous macro-cell networks, new heterogeneous deployment scenarios, such as relay-assisted networks and dense small-cell networks, have also been considered for future wireless communication networks to address the challenge of being able to provide very high data rate in specific scenarios, e.g., in shopping malls, in dense urban environments, or big events in stadiums [10–12]. Future heterogeneous dense networks will consist of low-cost and low-power access nodes, being densely deployed and coexisting with the traditional macro BSs. By creating a large number of small cells, these low-power access nodes have the potential to offload traffic from macro BSs, reduce the average distance between users and transmitters, and thereby improving the data rates and/or reducing the transmit power. In order to fully exploit the benefits offered by all these low-power nodes, a mechanism to efficiently coordinate their transmissions will be needed. Different from traditional homogeneous macro-cell networks, the densely deployed access nodes will be heterogeneous in the number of antennas, transmit power, backhaul capacity and reliability, and coverage area, etc. Moreover, the CSI at each BS are highly likely to be different and imperfect. All these heterogeneous properties make the multi-node coordination and ICI mitigation more challenging.

In an ideal and global cooperative system, where the CSI and the data of all users are perfectly shared between all transmission nodes, all the communication links can be exploited to provide joint data transmission to all users. The ICI can be completely eliminated, and hence the system throughput as well as cell-edge data rates can be significantly improved [7]. However, implementing multi-node cooperation in practical wireless communication systems faces some major challenges. Firstly, multi-node cooperation may require large amount of CSI and control signaling overhead placed on the over-the-air feedback links and the backhaul links between transmission nodes. Secondly, the quality of feedback and backhaul links in terms of capacity, latency and reliability cannot be granted. Therefore, information sharing between multiple transmission nodes can introduce errors and delays, and thereby affecting the transmission decisions made at the network side. Thirdly, in practice, most communication systems suffer from hardware impairments, e.g., phase noise, power amplifier nonlinearities, and in-phase and quadrature-phase imbalance (IQI), etc. Moreover, synchronization errors between multiple BSs can significantly impair the effectiveness of the most advanced coordinated transmission techniques. Heterogeneous wireless networks can be more prone to hardware impairments, since the hardware of the low-cost access nodes is most likely to be of low quality compared to the macro BSs. Without careful compensation, these hardware impairments can result in significant performance degradation. Finally, the involved large number of transmission nodes and users, as well as the increased spatial degrees of freedom, make the radio resource management that performs scheduling, power control and precoding
design more difficult in a cooperative system to achieve the promising cooperation gain.

Motivated by the above discussion, in this thesis, we investigate different multi-node cooperative techniques in the downlink of realistic wireless communication networks. The aim is to study the performance of different multi-node coordinated transmission schemes under the real-world impairments, and to develop efficient radio resource allocation algorithms for overcoming these limitations. In particular, the consequences of imperfect BS synchronization on multi-node joint transmission, the impact of IQI on relaying-assisted transmission, and the impact of imperfect CSI on the design of energy-efficient heterogeneous networks are studied. In addition, practical and efficient resource allocation algorithms are proposed to overcome the impact of these impairments, and to reduce the network overhead and the complexity for different cooperation schemes.

The thesis is organized as follows. Chapter 2 gives a brief introduction of cooperative wireless networks, where different cooperative scenarios, different multi-node coordination strategies, and the associated challenges for practical implementations are discussed. Chapter 3 presents the system model considered for the downlink multi-node joint transmission, and illustrates a general way to formulate and solve different radio resource optimization problems. Finally, in Chapter 4, we summarize the contributions of the thesis. The future work are also presented in this chapter.
Chapter 2

Cooperative Wireless Networks

In traditional wireless communication systems as shown in Fig. 2.1, each macro BS transmits desired signals only to users within its coverage area, namely a cell. For each user, the signals received from other cells on the same time-frequency resource will be treated as interference. The presence of ICI limits the system throughput, and it especially degrades the performance and affects the experience of the users located in the cell-edge areas, e.g., UE1 in Fig. 2.1.

In cooperative wireless systems, a users communication link is enhanced in a cooperative way or in a supportive way by other transmission nodes, e.g., neighbouring macro BSs, relay nodes, small cell BSs, and nearby users. These cooperative nodes either contribute to making decisions on scheduling/beamforming/power control in the time-frequency resource, or directly participate in data transmission. Compared to the traditional wireless communication systems, a properly designed cooperative communication system has the potential to improve the system capacity and reliability, reduce the energy consumption, and expand the network coverage.

This chapter gives a brief introduction of cooperative wireless networks. In particular, different cooperative scenarios, cooperative transmission schemes, and the associated challenges posed by practical constraints are discussed.

2.1 Cooperative Scenarios

Depending on what and how the transmission nodes are deployed, cooperative wireless networks can be divided into two categories, i.e., homogeneous cooperative networks and heterogeneous cooperative networks.

2.1.1 Homogeneous Cooperative Networks

In homogeneous cooperative networks, the multiple geographically separated transmission nodes are typically within the same type; that is all transmission nodes have the similar transmit power levels, antenna patterns, receiver noise floors and backhaul connectivity, etc. Fig. 2.2 illustrates an example of the downlink of a multi-cell cooperative system, where all macro BSs are inter-connected via wired backhaul links for information sharing. If infinite cooperation between the BSs is enabled (i.e., with ideal feedback, backhaul and synchronization), the system is effectively a network multiple-input-multiple-output (MIMO) broadcast channel, where all communication links, including the interfering ones,
can be exploited to coherently transmit user data via joint processing of information. Therefore, the ICI can be completely eliminated, and the system throughput as well as cell-edge data rates can be significantly improved [7, 8, 13].

2.1.2 Heterogeneous Cooperative Networks

A heterogeneous cooperative network may include a number of low-cost low-power access nodes which are deployed to coexist with the traditional macro BSs. A low-power node refers to a node whose transmit power is lower than macro BS classes. There are various types of low-power nodes including relay nodes, micro BSs, Pico BSs or femto BSs (also called home eNBs). These access nodes create multiple tiers of the network and differ in terms of the number of antennas, transmit power, backhaul reliability and coverage area, etc. The low-power nodes have the potential to offload traffic from macro BSs, bring the transmitter closer to the users, and thereby improving spectral efficiency per unit area and providing a uniform QoS experience to users anywhere in the network. Two most commonly considered heterogeneous cooperative networks are relay-assisted networks and small-cell networks.

A Relay-assisted network can be considered as a two-tier heterogeneous network, consisting of two types of transmission nodes, i.e., source nodes (e.g., macro BSs) and relay nodes. In such a heterogeneous network, if a communication cannot be established between a source and a destination due to severe fading, a third party device (i.e., the relay node) that receives the information from the transmitter can help forward the information via a relaying channel that is independent from the source-to-destination link. Moreover, by processing and retransmitting the received signals from the source to the destination, relay nodes have the potential to extend the coverage area, improve system reliability and the link-level performance. Relay-assisted networks do not need wired connection between the transmission nodes, and thereby reducing operators’ backhaul cost. However, relaying techniques do little to increase the performance of the users placed in severe ICI-dominated areas, such as the cell-edge areas of current cellular networks [8].

Figure 2.1: Illustrations of a traditional cellular system. The solid lines represent the useful signals, while the dashed lines denote the interference signals.
Small-cell networks can be referred to as multi-tier heterogeneous networks, with different low-cost low-power access nodes deployed in the presence of an overlaid macro cellular network. The macro cells provide wide coverage, while the low-power nodes are deployed in a more targeted manner to offload traffic from macro BSs, or to alleviate coverage dead zones, or to improve the data rate in hot spot areas [15, 16]. The macro cells and the small cells may share the same spectrum (a co-channel deployment), or different frequency bands could be assigned for the macro layer and small cell layers. The backhaul links between macro and small cells, as well as between small cells can be different. For example, the macro and small cells can be connected via very high throughput and low latency backhaul such as dedicated point-to-point optical fiber, while the small cell nodes are inter-connected via xDSL, microwave or wireless self-backhauling. The small cell nodes can be deployed indoors or outdoors, and in either case they could provide service to indoor or outdoor users. Depending on the scenarios, the deployment of small cell nodes can be either sparse or dense. For the indoor or outdoor hot spots, a few low-power nodes could be sparsely deployed to cover the areas. On the other hand, for the dense urban area or large shopping malls, a large number of small cell nodes needs to be densely deployed to boost the capacity and support large traffic over a relatively wide area. Fig. 2.3 illustrates an example of the downlink of a small cell network, where all BSs are deployed outdoors and sharing the same spectrum.

2.2 Cooperative techniques

In order to achieve the promising cooperation gain in terms of the spectral efficiency, cell-edge data rate, system reliability and potentially the energy efficiency, cooperative schemes need to be properly designed and selected for different scenarios. This section introduces some key technical components that have been widely considered in cooperative
wireless networks.

### 2.2.1 Inter-Cell Interference Coordination

Traditional techniques for combating ICI have focused on either allocating orthogonal radio resources to different transmit signals, for example, frequency reuse, cell sectoring, or canceling interference via signal processing [5, 6]. These interference mitigation approaches can be characterized as passive. In the 3rd generation partnership project (3GPP) long term evolution (LTE) systems, inter-BS signaling can be accomplished over the X2 interface between BSs. Hence, inter-cell interference coordination (ICIC) was proposed in LTE Release 8 as a key technique to deal with the ICI issue between neighboring macro BSs in a proactive way [3]. ICIC is a power and frequency domain interference coordination scheme. In particular, it restricts the available frequency resources of different cells through a predefined frequency reuse rule or through appropriate power control. Fig. 2.4 shows a classic ICIC approach, named as soft frequency reuse (SFR) [17]. In the SFR scheme, the spectrum in each cell are divided into two sets, i.e., major sub-band and minor sub-band. Normally, the maximum allowed transmit power for major sub-band is higher than that of the minor sub-band. The major sub-band (high-power sub-band) can be used to cover the whole cell area, while the minor sub-band (low-power sub-band) is used only in cell-center area. Major sub-bands in the neighboring cells are non-overlapping. Therefore, for a user located at the cell-edge area, the received ICI comes only from the low-power sub-band in the neighboring cells. This provides improvement in the ratio of the desired received signal power over interference and noise power (SINR). Consequently, the cell-edge data rates and the cell coverage can be improved. However, it should be pointed out that the SINR for the cell-edge users is improved at the expense of degrading SINR for the cell-center users, which will result in a degradation in the system peak or sum throughput.

In heterogeneous networks, different types of cells are randomly deployed and over-
lapping in many scenarios. Having a highly heterogeneous network topology makes it more difficult to control and coordinate the ICI in co-channel deployed scenarios. The power and frequency domain based ICIC strategy can effectively reduce the ICI on data channels between macro BSs, however, it cannot mitigate the ICI on control channels. This is because for downlink data transmission, each subframe consists of two parts: one is used for control channel (physical downlink control channel (PDCCH)) and the other is for data channel (physical downlink shared channel(PDSCH)). The data channels are used for transmitting user data, while the control channels are used for delivering the resource allocation information and other signalling information. ICIC can assign different frequency sub-bands to cell-edge users only for data channels. The control channels have to be distributed over the entire frequency bandwidth. This causes ICI on the control channels of neighboring cells. ICI on control channels is not a big issue in homogeneous networks, since all BSs’s transmit power are at the same level. However, in heterogeneous networks, there is a large difference in transmission power between macro BSs and low-power nodes. Therefore, control channels of small cells can be significantly interfered by the macro cells, resulting in ineffective ICIC on the data channels.

To cope with the above problem, enhanced inter-cell interference coordination (eICIC) was introduced in Release 10 to reduce ICI on both data and control channels in co-channel deployed heterogeneous networks [11, 16, 18, 19]. Compared to ICIC, in addition to coordinate the ICI in power and frequency domain, eICIC introduces a new concept of almost blank subframe (ABS) to coordinate ICI in the time domain. ABS subframes do not transmit any power on the data channels and carry only some necessary control signals, including common reference signals (CRS), primary and secondary synchronization signals (PSS and SSS), paging information, etc.) with low power. eICIC configures ABS subframes in the interfering macro BS. By sharing this ABS pattern information between the macro BSs and small cell access nodes over X2 interface, the interfered small cells can perform data transmission to their users during these ABS subframes. In this way, the ICI is effectively coordinated by assigning different subframes to different types of cells. To further reduce the ICI, enhanced receivers can perform interference cancellation of the residual control signals transmitted by macro cells in ABS subframes. This technique is named as further enhanced ICIC (FeICIC) in LTE Release 11 [20, 21].
2.2.2 Multi-Node Cooperative Transmission

Instead of coordinating ICI by restricting how radio resources are used in each cell, multi-cell advanced coordination and joint transmission can be used as a more proactive way to handle the ICI issue with much tighter multi-node cooperation. In the literature, a family of cooperative communication techniques have emerged and gained significant interest, such as “network MIMO” [6], “network coordination” [7], “multi-cell processing” [8], “multicell multiuser MIMO” [13, 22, 23], “distributed antenna systems” [24] and “group cells” [25]. In the 3GPP standard development organization, this concept is referred to as “Coordinated multi-point (CoMP) transmission/reception”, which is considered as a dedicated study item in LTE-Advanced Release 11 [26] and Release 12 [27].

In 3GPP LTE-Advanced, a CoMP cooperating set is defined as a set of nodes that directly participate in data transmission or contribute to making decisions on scheduling/beamforming in the time-frequency resource [26]. A CoMP cooperating set typically consists of multiple geographically separated transmission nodes from either a homogeneous or a heterogeneous network. In the literature, cooperating sets are usually labeled as “CoMP cluster”.

Depending on whether the user data is shared among all the transmission points within a CoMP cluster, downlink CoMP schemes can be divided into two main categories: joint processing and coordinated scheduling/beamforming.

Joint Processing

In the CoMP joint processing approach, user data is available simultaneously at all transmission points within the CoMP cluster. By sharing both the CSI and the data of all users in the cluster, coordinated multiple points can act as a single and distributed antenna array. Simultaneous joint data transmission can then be performed coherently or non-coherently to a single user or multiple users from multiple transmission points in a time-frequency resource. In this way, the ICI is mitigated as signals transmitted from other points assist the transmission rather than acting as interference. This network MIMO technique falls into one subset of joint processing, labeled as joint transmission, see Fig. 2.5 a).

Another subset of joint processing, which is shown in Fig. 2.5 b), is dynamic point selection/muting, where the data of a user is only transmitted from one of the points within the CoMP cluster in a certain time-frequency resource. However, user data is
available at multiple points and the transmission/muting point may change from one subframe to another via dynamic scheduling by exploiting changes in the channel fading conditions.

Note that dynamic point selection may be combined with joint transmission, that is, multiple points can be selected for data transmission in a time-frequency resource. In this case, data to a single user can be transmitted non-coherently from the selected multiple points without phase adjustment. Even with perfect CSI available at the transmitter side, this non-coherent joint transmission scheme can not completely mitigate ICI unless the unselected points are muted. However, it might be more robust to channel uncertainty than coherent joint transmission [28].

Coordinated Scheduling/Beamforming

In the coordinated scheduling/beamforming approach, as shown in Fig. 2.5 c), the data symbols to a user is only available at and transmitted from one point from the CoMP cluster on a time-frequency resource. However, by sharing the CSI of all users among multiple transmission points, user scheduling and beamforming can be coordinated in order to control ICI. Note that this CoMP approach can only mitigate ICI rather than exploiting it.

2.2.3 Load Balancing

In many relevant wireless communication scenarios, e.g., in shopping malls, in dense urban environments, or during the occurrence of traffic jams, users are typically non-uniformly distributed over the network, yielding load imbalance between cells. Even with a uniform user distribution, in heterogeneous networks, large disparity in transmit power between macro BS and low-power nodes, as well as different BS capabilities can still lead to a major load imbalance problem. In particular, if the traditional user association metric is applied, i.e., each user selects the cell associated with the highest downlink received power, most of the users may connect to the macro BSs even when some small cells have no user to serve [10, 29]. As a result, some users in overloaded macro cells could be unable to get services due to lack of resources, and the advantages of deploying low-power nodes cannot be fully utilized.

Load balancing is referred to as an approach to balance the load (e.g., users, data traffic) across different cells in order to optimize the system metrics such as sum throughput, user fairness, resource utilization, energy efficiency, etc. This is usually achieved by adjusting the network control parameters in such a way that overloaded cells can offload the excess load to low-loaded adjacent cells, whenever it is available. In the following, we discuss two key techniques used for load balancing: cell range expansion and centralized optimization.

Centralized Optimization

Depending on the objectives and the practical constraints, the design of wireless communication networks can be formulated into different system-level optimization problems, where load balancing, transmission scheme and power control are coupled with each other. By solving these large optimization problems, different load balancing strategies can be
obtained for optimizing different system metrics [30–38]. In general, these system-level optimization problems are combinatorial optimization problems, which are non-convex and difficult to solve. The computational complexity can grow exponentially with the number of users or the number of cells in the network. This makes it difficult to implement the designed algorithms in large-scale networks. Moreover, these proposed algorithms are typically sensitive to the changes in the networks, e.g., the deployment of access nodes, the user locations, the channel conditions, etc. Therefore, the algorithms need to rerun every time slot in order to keep track of these changes. Nevertheless, for a given system optimization problem, the obtained optimal solution can be used to serve as the upper bound for any other suboptimal load balancing solutions.

**Cell Range Expansion**

Biasing-based cell selection is a suboptimal but simple technique to balancing the load among high and low power BSs. Traditionally, in practical systems, cell selection is determined by the strongest received signal power. Based on this user association metric, the coverage of low-power nodes (e.g., micro or pico BSs) can be very small such that most of the users get connected to high-power nodes (e.g., macro BSs). In order to offload users from high-power nodes to low-power nodes and hence make better use of the network resources, different biasing-based cell selection schemes have been proposed, e.g., biased-received-power-based, biased-SINR-based, biased-rate-based schemes [10, 39–42]. These schemes allows users to connect to a low-power node by adding a multiplicative bias to each tier of BSs. For example, consider a network consisting of $M$ macro BS and $N$ pico BSs. The $N$ pico BSs are deployed as an overlay to the macro cell layer. Let $P_{k,M}^r$ be the received power from macro BS $i$ to user $k$ for $i = 1, \ldots, M$, and $P_{k,P}^r$ denotes the received power from pico BS $j$ to user $k$ for $j = 1, \ldots, N$. Then, utilizing a biased-received-power-based load balancing strategy, the selected BS for user $k$ will be the one that provides the maximum biased received signal power, i.e., $\arg \max_{M,P} \left\{ B_M P_{k,M}^r, B_P P_{k,P}^r \right\}$, where $B_M$ and $B_P$ are the bias for macro BSs and pico BSs, respectively. If $B_M = 1$ and $B_P = 10$, then a user would connect to a pico BS until its received power from a macro BS was $10\text{dB}$ higher than the pico BS. Biasing-based load balancing can effectively expand the coverage of low-power nodes, therefore, it is referred to as cell range expansion in 3GPP standardization work.

**2.2.4 Cell On/Off**

Except from the spectral efficiency and user experience, energy efficiency is also an important performance metric for the design of future green wireless networks. The global energy consumption of a network can be roughly divided into two parts: a circuit part that depends on the transceiver hardware and a dynamic part which is a function of the transmission power [43–46]. Cooperative communication networks have the potential to achieve substantial energy savings for the dynamic part. In particular, multi-node transmission can provide better energy-focusing to the desired users. Moreover, adding more low-power nodes reduces propagation losses between the users and the transmitters. Therefore, the total transmit power in cooperative networks can be effectively reduced. However, multi-node cooperation also requires extra power on feedback and backhaul information exchange. Furthermore, since more access nodes means that more hardware
is required, increasing the number of small cells will thus increase the circuit power consumption, and therefore it may increase the global energy consumption.

The circuit power consumption depends on the operation mode of each access node. In general, there are three different operation modes considered for a cell, i.e., the on, sleep and off modes [47]. When a cell is turned on, it can transmit the data signals and also the control signals necessary for data transmission, such as reference signals used for measurements and demodulation. Users may be connected to and receive data transmission from the cell. In the sleep mode, a cell cannot send or receive any signals, and users cannot access the cell and will not receive data transmission from the cell, but the cell can be woken up by a control unit (CU) (e.g., a macro BS) though backhaul links. In the sleep mode, the power amplifier and the radio frequency (RF) components are deactivated. This provides significant energy savings compared to the on mode. For example, the circuit power consumption of a macro BS in the sleep mode is only 57% of the power consumption in the on mode [43, 44]. Note that, in the sleep mode, in order for enabling fast transition to the on mode, some hardware elements, such as the power supply and the baseband processing components, are not switched off. In order to further reduce the energy consumption, a cell can be put into the off mode, in which the cell is completely deactivated. However, with current technology, the transition time from the off to the on mode is non-negligible.

Since the circuit power consumption under the sleep or off mode is much lower compared to the one under the on mode, the increase in the circuit part from the extra power consumed by activating BSs clearly outweighs the decrease in the dynamic part. Therefore, putting a BS into the sleep or off mode by proper coordination strategies is an important solution for energy savings, especially for dense cooperative network scenarios or when the network traffic is low. Note that cell on/off may also provide benefits for interference avoidance and coordination [48–50]. In particular, a network may turn off certain cells to reduce ICI, especially the interference caused by common channel transmissions such as CRS in heterogeneous networks.

### 2.2.5 Cell Clustering

In practice, the cooperation area is limited by the feedback, backhaul, synchronization, and complexity constraints in a real system deployment. Thus, the network is typically divided into clusters of coordinated access point so that cooperative transmission can be independently implemented within each cluster. The cluster formation becomes an important issue that affects the cooperation performance.

There are various ways to divided the network into different cooperation clusters. Based on the cluster reconfiguration time scale, the cluster formation can be characterized into two categories:

- **Static clustering**, which specifies a predefined set of disjoint clusters of cells that does not change in time or changes over a very long time scale (e.g., based on traffic statistics) [51–53]. The static cluster formation is easily implementable, and it requires very limited inter-cluster information exchange.

- **Dynamic clustering**, where the clusters are formed based on the varying channel conditions of the users [54, 55] or uneven traffic load of the system [56]. With
higher flexibility, theoretically, the dynamic cluster formation can provide more co-operation gains compared to the static clustering approach. However, large amount of signaling information exchange between different BSs is needed for cluster reconfiguration decisions, which is infeasible in large networks. Examples of exchanged information can be traffic-distribution within different cells, downlink interference contribution from cell A to cell B, user channel conditions, etc.

Depending on where the cluster formation decision is made, cell clustering can also be classified as network-specific, user-specific or hybrid.

- Network-specific clustering: disjoint clusters of cells are formed by the network based on e.g., the dominating interference cells and/or the traffic-distribution within different cells, regardless of the channel condition of each individual user. Users belonging to the same cell are assigned to the same cluster. The cluster construction can be performed either in a static or a dynamic fashion. Static network-specific clustering can only mitigate the interference within the cluster. The performance is mainly limited by the inter-cluster interference, especially for the users located at the cluster edge area, referred to as cluster edge effect. An example of network-specific dynamic clustering is shown in Fig. 2.6 where, in a certain time frame, cell 1 is grouped with cell 2 and cell 3 as a cluster; In the next time frame, according to the network traffic distribution cell 2 will be replaced by cell 4 to form a new cluster.

- User-specific clustering: each user selects a set of BSs that are suitable to form a cluster. The clusters of different users may overlap. The construction of the cluster for each user can be semi-static or changed dynamically based on the channel conditions between the user and the BSs. This way, users are guaranteed to be always located at the cluster center to avoid the cluster edge effect. However, user-specific clustering requires joint scheduling across BSs, which increases the inter-BS information exchange as well as the resource allocation complexity. Fig. 2.7 illustrates a user-specific clustering method, named as slide Group Cell, proposed in [57]. In the current time slot, cells 1, 3 and 4 are selected by UE 1 as a CoMP cluster, while the cluster for UE 2 is formed by cells 1, 2 and 3. In the next time slot, with the possible move of UE 1 and UE 2, UE 1 will select cells 1, 2 and 3 as its cluster and UE 2 will choose cells 1, 3 and 4 to form a new cluster.
2.3 Challenges and Difficulties

Theoretically, global network cooperation can provide significant performance gains both in terms of system spectrum efficiency and the cell-edge throughput. However, the realistic gains can be limited by many practical constraints for a cooperative network deployment. In this section, we discuss some of the most important challenges that need to be considered in the design of cooperative wireless networks.

2.3.1 Non-Ideal Feedback

In order to enable multi-node cooperative transmission, the CSI of all users in the network (named as full CSI) is required at the transmitter side. In a time division duplex (TDD) system, each transmission node can obtain the CSI from the users belonging to its coverage (named as local CSI) by exploiting channel reciprocity. Then, the local CSI can be exchanged via backhaul links with other coordinated nodes or be forwarded to a CU, where the resource allocation and/or data processing take place. In frequency division duplex (FDD) systems, each point acquires local CSI through feedback from its users instead. In contrast to the traditional non-cooperative networks, each user within the FDD-mode CoMP cluster needs to not only estimate and feedback the CSI related to the strongest transmission node, but also related to the other coordinated nodes. Therefore, the feedback load grows proportional to the number of transmission nodes in the network, which poses tight capacity requirements on the feedback channels \([61, 62]\).

In practice, the feedback CSI acquired at the serving BS can never be perfect due to hardware deficiencies, channel estimation errors, quantization errors, and the feedback latency \([16, 27, 63–65]\). It should be pointed out that different levels of channel knowledge may be required by different categories of cooperation techniques. For example, for inter-cell interference coordination, coordinated scheduling/beamforming, and dynamic point selection schemes, no inter-point phase information is needed. However, for coherent joint transmission schemes, inter-point phase information is required, inter-point amplitude information may also be needed. Therefore, the feedback overhead might be different.
when considering different cooperation techniques.

### 2.3.2 Non-Ideal Backhaul

All cooperation techniques rely on information exchange between cooperative nodes through the backhaul network, i.e., a network interconnecting the transmission nodes. For multi-node cooperative transmission, the fed back local CSI needs to be shared over backhaul links among multiple access nodes in order to gather CSI at the transmitter side. For the case of joint processing, user data also needs to be available simultaneously at the cooperative nodes. The control signaling information may need to be exchanged between different nodes in order to support (e)ICIC. In addition, resource utilization information exchange may be needed to assist load balancing decision making. All those inter-node information exchange places a large amount of overhead on the backhaul links [61, 66]. Therefore, in order to achieve the potential cooperation gain, high capacity backhaul links are required, especially for the joint processing schemes.

Depending on the backhaul network deployment of a realistic system, e.g., the transport technology and the network topology, the overall latency introduced by only one hop backhaul link can range from hundreds of microseconds to 20 ms, and the capacity requirement ranges from a few Mbps to 10 Gbps [16, 67]. It is also pointed out in [67] that, even with point-to-point fiber technology, inter-BS information exchange may require the X2 logical link to go through several aggregation routers, hence, the normal latency between two BSs (eNBs) would be 10-20 ms. In 3GPP LTE-Advanced, the latency values of 2, 10, and 50 ms are recommended for evaluation of both homogeneous and heterogeneous cooperative networks [27, 48].

Besides capacity and latency constraints, the reliability of the backhaul links also plays an important role when performing different cooperation techniques. In heterogeneous cooperative networks, the backhaul links interconnecting multiple nodes are highly likely to be wireless and unreliable. This is because the high number of access nodes cannot be accompanied by a proportionally high financial investment in order to build high quality wireline backhaul. Furthermore, the flexible deployment of heterogeneous access points, i.e., some will be mounted on high towers (macro BSs), some will be deployed on the street level below roof tops (pico BSs and relays) and others will be indoors (femto BSs), requires that backhaul links interconnecting access nodes are wireless and without guaranteed line-of-sight (LoS) [68–71]. In other words, the backhaul links of future dense heterogeneous networks will be unreliable to some degree and the system designers should optimize the system by taking this constraint into account.

### 2.3.3 Hardware Impairments

In practice, most communication systems suffer from hardware impairments, e.g., phase noise, power amplifier nonlinearities, and in-phase and quadrature-phase imbalance (IQI) [72, 73]. Here, we discuss two main hardware impairments that have been considered in this thesis.
2.3 Challenges and Difficulties

Synchronization

Multi-node cooperative transmission requires tight time and frequency synchronization between different transmission points. The synchronization constraint is most challenging for the joint transmission approach, since the carrier phases between coordinated nodes also need to be synchronized, which can be extremely difficult mainly due to the effect of carrier frequency offset, or/and phase noise from local oscillators in each BS [74–76]. An example in [76] shows that for a small relative velocity (5 km/h), the phase noise process can be assumed to vary much faster than the channel. This is because the bandwidth of the phase noise (arising from the local oscillators) is much higher than the Doppler spread. Note that the user velocity of interest in 3GPP LTE-A CoMP scenarios is 3 km/h [26], which makes the difference even larger. In a worst case scenario, where the phase shift of each link (arising from the oscillator) has a random uniform distribution and varies much faster than the channel fading, the effect of phase adjustment by joint precoding will be averaged out. In this case, if phase uncertainty is not handled, joint precoding cannot contribute to the performance improvement when performing joint processing.

I/Q Imbalance

In general, I/Q imbalance (IQI) refers to the phase and/or amplitude mismatch between the in-phase (I) and quadrature (Q) signals at the transmitter and receiver sides. With perfect I/Q matching, the I and Q branches of the transceiver will have exactly the same amplitude and a 90° phase shift. However, in practice, due to the limited accuracy of the used analog components at the transceivers, IQI is always present in the up- and down-converters. The sources of IQI include: errors in the nominal 90° phase shift between the local oscillator signals during up- and down-conversion, and the difference in amplitude transfer of the I and Q arms [73, 77]. Such imbalance results in an additional image signal, which appears as interference on top of the desired signal.

The mixture of the image and the desired signals can lead to significant performance loss, especially in high-rate wireless communications systems. In heterogeneous networks, the good channel connection experienced by the users from their nearby access nodes provides the possibility for using higher order modulation scheme (i.e., 256QAM) for the downlink data transmission, and thereby improving system spectrum efficiency [48]. However, it has been shown in [26] that the real gains of 256 QAM are dependent on the transceiver IQI, especially the IQI at the receiver. When modelling of receiver IQI with -25dB image rejection ratio, no gains from 256QAM were observed.

Note that compared to the traditional macro-cell cellular networks, heterogeneous wireless networks are more prone to hardware impairments, since the hardware of the low-cost access nodes is most likely to be of low quality compared to the macro BSs.

2.3.4 Resource Optimization

In order to achieve a certain network design objective, the available resources need to be efficiently utilized. Resource allocation may include user scheduling, subchannel allocation, node selection, power control, as well as precoding/beamforming design. In general, the optimization problems in a multi-cell multi-user system are non-convex and difficult to solve. The computational complexity for a centralized resource allocation increases with the number of subchannels, users and transmission nodes in the system. Moreover,
resource allocation algorithms are designed based on the CSI available at the transmitter side, which is never perfect due to the non-ideal feedback and backhaul. Therefore, the algorithms relying on perfect CSI may lose significant cooperation gains under practical scenarios, especially for the heterogeneous cooperative networks. Developing practical resource allocation solutions for cooperative wireless systems, taking different practical constraints into account, is a difficult task. This is the main focus of this thesis.

2.3.5 Cluster-Edge Effect

The CoMP performance gains highly rely on the accuracy of CSI at the transmitter side. How to acquire the full CSI and design CoMP transmission parameters is an important issue for the system level design. Regarding this aspect, different centralized and decentralized CoMP architectures are proposed [78, 79]. As the size of cluster increases, i.e., the number of transmission nodes and the number of users increase, the inter-point information exchange via backhaul links, as well as the amount of CSI fed back from users over the feedback links will increase. In addition, the complexity for resource allocation will become prohibitively high. Therefore, the cluster size is limited by the feedback, backhaul, synchronization, and complexity constraints in a real system. A cooperative network is typically divided into clusters of coordinated nodes so that cooperation techniques can be independently implemented within each cluster. Note that a coordinated cluster may also cause inter-cluster interference to the users in the neighboring clusters, especially to users in the cluster-edge area, named as *cluster edge effect*. As we will show later in this thesis, fractional frequency reuse can be considered as a promising technique to coordinate the inter-cluster interference for static CoMP clusters, thus, mitigating this so-called cluster edge effect.
Chapter 3

Radio Resource Allocation

Resource allocation plays an important role in wireless communication networks as a way of optimizing the assignment of available resources to achieve a network design objective and at the same time guarantee the QoS for all users. All cooperative transmission techniques require the network to jointly design the user scheduling, subchannel allocation, power control, and/or the precoding/beamforming matrices for all transmission nodes within the cooperative network. In general, the optimization problems are non-convex and difficult to solve. In a cooperative system, where large number of transmission nodes and users are involved, the resource allocation and data processing problems can be more complex and challenging, especially for the case of multi-node joint processing. In addition, the design of resource allocation algorithms highly rely on the CSI at the transmitter side, which can be corrupted by various practical constraints as mentioned in Chapter 2. Developing practical resource allocation solutions for cooperative wireless systems, taking different practical constraints into account, is a difficult task. This is the main focus of this thesis.

In this chapter, by assuming ideal global cooperation between all transmission nodes, a system model for the multi-node coherent joint transmission is presented in Section 3.1. Then, a general way of formulating resource allocation problems is illustrated in Section 3.2. In Section 3.3, we introduce basic concepts of convex optimization. Then, different resource optimization problems that can be formulated in convex forms are discussed. Finally, the effect of different practical constraints on the design of resource allocation algorithms for multi-node coordinated transmission is discussed in Section 3.4.

3.1 System Model

We consider the downlink of a cooperative network, which consists of $M$ BSs and $K$ single-antenna users. The system spectrum is universally reused by each BS. The $K$ users don’t belong to any particular cell and are randomly deployed in the network. BS $v$ is assumed to have $N_v$ antennas. The total number of transmit antennas is $N = \sum_{v=1}^{M} N_v$.

At each time slot, the channels from all BSs to user $k$ is denoted by $h_k^H = [h_{k,1}^H, \ldots, h_{k,M}^H] \in \mathbb{C}^{1 \times N}$, where $h_{k,v}^H \in \mathbb{C}^{1 \times N_v}$ is the channel from BS $v$ to user $k$ for $v = 1, \ldots, M$ and $k = 1, \ldots, K$. Similarly, let $x = [x_1^H, \ldots, x_K^H]^H \in \mathbb{C}^{N \times 1}$ denote the signal vector transmitted from all $M$ BSs, where $x_v \in \mathbb{C}^{N_v \times 1}$ is the transmitted signal from BS $v$. Then,
the received signal at user $k$ is
\[ y_k = h_k^H x + n_k, \] (3.1)
where $n_k \sim \mathcal{C}\mathcal{N}(0, \sigma_k^2)$ is the independent additive receiver noise at user $k$.

Assume that the CSI and data symbols of the $K$ users are perfectly known at each BS. In addition, the BSs are connected via backhaul links, and all BSs are able to coherently transmit to all users at the same time-frequency resource. By using linear precoding, the transmitted signal vector $x$ can be expressed as
\[ x = Wb, \] (3.2)
where $b \in \mathbb{C}^{K \times 1}$ denotes the normalized complex data symbols for the $K$ users, with $b_k \sim \mathcal{C}\mathcal{N}(0, 1)$. Here, $W \in \mathbb{C}^{N \times K}$ is the precoding matrix used to map the data symbol vector to the transmit signal vector. The $k$-th column of $W$ corresponds to the aggregated precoding vector for user $k$ from all BSs, and is denoted by $w_k$ with $w_k = [w_{k,1}, \ldots, w_{k,M}]^T \in \mathbb{C}^{N \times 1}$, where $w_{k,v} \in \mathbb{C}^{N_v \times 1}$ is referred to as the precoding vector for user $k$ from BS $v$.

Substituting (3.2) into (3.1), the received signal of user $k$ is
\[ y_k = h_k^H w_k b_k + \sum_{i \neq k} h_k w_i b_i + n_k, \] (3.3)
We assume that data symbols for different users are independent. Then, the signal to interference plus noise ratio (SINR) of user $k$ is
\[ \gamma_k = \frac{||h_k^H w_k||^2}{\sum_{i \neq k} ||h_k^H w_i||^2 + \sigma^2} = \frac{||h_k^H u_k||^2 p_k}{\sum_{i \neq k} ||h_k^H u_i||^2 p_i + \sigma^2}. \] (3.4)
Here, $w_k = p_k u_k$ with the normalized precoding vector $||u_k|| = 1$ for $k = 1, \ldots, K$. By treating interference as noise, the achievable data rate of user $k$ is
\[ R_k = \log_2(1 + \gamma_k). \] (3.5)
According to (3.2), the transmit power of BS $v$ can be derived as
\[ P_{\text{trans},v} = \sum_{k=1}^{K} \left( \|w_{k,v}\|^2 \right) \mathbb{E}\{\|b_k\|^2\} = \sum_{k=1}^{K} \|w_{k,v}\|^2. \] (3.6)

From (3.5) we see that the beamforming vectors $u_k$ and the power control solution $p_k$ can be immediately obtained from the precoding vectors $w_k$. Moreover, based on (3.2), if $w_{k,v} = 0$, then, the transmitted signals from BS $v$ (i.e., $x_v$) will not contain the data symbols of user $k$, i.e., BS $v$ will not provide data transmission to user $k$. Hence, by finding the precoding solution, we can also obtain the load balancing and cell selection solution; that is, the set of users assigned to BS $v$ is $\mathcal{U}_v \trianglerighteq \{k|w_{k,v} \neq 0, k \in \{1, \ldots, K\}\}$, and the set of BSs that provide data transmission to user $k$ is $\mathcal{V}_k \trianglerighteq \{v|w_{k,v} \neq 0, v \in \{1, \ldots, M\}\}$. Therefore, the joint design of beamforming, power control and user scheduling can be
3.2 Problem Formulation

A generic optimization problem has the standard form

\[
\begin{align*}
\text{minimize} & \quad f_0(x) \\
\text{subject to} & \quad f_i(x) \leq 0, \ i = 1, \ldots, m, \\
& \quad h_i(x) = 0, \ i = 1, \ldots, p,
\end{align*}
\]  

(3.8)

where \( x \in \mathbb{R}^n \) is the optimization variable, the function \( f_0 \) is the objective function, \( f_1, \ldots, f_m \) are the \( m \) inequality constraint functions and \( h_1, \ldots, h_p \) are the \( p \) equality constraint functions.

Depending on the targets and the requirements for the design of the networks, different resource optimization problems can be formulated. In this section, based on the system model in Section 3.2, we present a number of different objectives and constraints that have been widely considered in radio resource allocation problems.

3.2.1 Objectives

In general, the optimization objectives can be divided into the following three categories:

- **Transmit power minimization**: minimize \( f_0 (P_{\text{trans}, 1}, \ldots, P_{\text{trans}, v}, \ldots, P_{\text{trans}, M}) \).
  
  Typical examples for the utility functions of \( f_0 () \) for transmit power minimization are weighted sum and max. The corresponding objective functions are
  
  \[
  \sum_{v=1}^{M} \alpha_v P_{\text{trans}, v}
  \]  
  
  (3.9)
  
  and
  
  \[
  \max_v P_{\text{trans}, v}
  \]
  
  (3.10)

  respectively. Here, \( \alpha_v \geq 0 \) denotes the weight assigned to BS \( v \), which can be used to balance the power consumptions of different BSs.

- **Rate maximization**: maximize \( f_0 (R_1, \ldots, R_k, \ldots, R_K) \).
  
  For this category, the most common utility functions of \( f_0 () \) are weighted sum and \( \min \). The corresponding objective functions are
  
  \[
  \sum_{k=1}^{K} \beta_k R_k
  \]
  
  (3.11)
  
  and
  
  \[
  \min_k R_k
  \]
  
  (3.12)
respectively. Similarly, $\beta_k \geq 0$ is the weight assigned to user $k$, which is introduced to compensate for the heterogeneous QoS requirements at the users.

- Energy efficiency maximization: maximize $f_0 (P_{\text{trans},1}, \ldots, P_{\text{trans},M}, R_1, \ldots, R_K)$.
  If the energy efficiency of a network is measured by bit/Joule delivered to the users, the objective function can then be expressed as

$$\frac{\sum_{k=1}^{K} R_k}{P_{\text{tot}}},$$

(3.13)

where $P_{\text{tot}} = g (P_{\text{trans},1}, \ldots, P_{\text{trans},M})$ is the total power consumption of the network, which is a function of the transmit powers $P_{\text{trans},v}$ for $v = 1, \ldots, M$.

### 3.2.2 Constraints

The constraints of downlink radio resource optimization problems include:

- Power constraints, for example, the total transmit power constraint

$$\sum_{v=1}^{M} P_{\text{trans},v} \leq P_{\text{trans},\text{tot}}$$

(3.14)

and the per-BS power constraints

$$P_{\text{trans},v} \leq P_{v,\text{max}}, \forall v,$$

(3.15)

where $P_{\text{trans},\text{tot}}$ and $P_{v,\text{max}}$ denote the maximum total transmit power and the maximum transmit power for BS $v$, respectively.

- The QoS expectations of the users, which is usually modeled as a function of the SINR

$$\gamma_k \geq \Gamma_k, \forall k,$$

(3.16)

or a function of the user data rate as

$$R_k \geq r_k, \forall k,$$

(3.17)

where $\Gamma_k$ and $r_k$ denote the target SINR and the target data rate for user $k$.

Depending on the network architectures and traffic models, some resource optimization problems may also include backhaul capacity constraints [80–83], CSI feedback constraints [84, 85] and transmission latency constraints [86], etc. Finding the optimal resource allocation solution for a multi-cell multi-user cooperative system is generally NP hard. However, as will be shown in Section 3.3, some resource optimization problems have hidden convexity structures, and therefore can be formulated or transformed into convex problems. Thus, these problems can be efficiently solved by using standard optimization techniques.
3.3 Radio Resource Optimization

An optimization problem in (3.8) is a convex problem, if the equality constraint functions \((h_1,\ldots,h_p)\) are affine, and the objective function \((f_o)\) and the inequality constraint functions \((f_1,\ldots,f_m)\) are convex, i.e.,

\[
f_i(\alpha_1x_1 + \alpha_2x_2) \leq \alpha_1 f_i(x_1) + \alpha_2 f_i(x_2), \quad i = 0, 1, \ldots, m,
\]

for all \(x_1, x_2 \in \mathbb{R}^n\) and all \(\alpha_1, \alpha_2 \in \mathbb{R}\) with \(\alpha_1 + \alpha_2 = 1, \alpha_1, \alpha_2 \geq 0\). A fundamental property of convex problems is that the local optimal point is also globally optimal. Another attractive property is that the optimal solution of the primary convex problem (3.8) can be obtained by solving its Lagrange dual problem, leading to decomposable structures and distributed algorithm design \[87, 88\]. In general, a convex optimization problem can be solved efficiently and reliably by using interior-point methods or other methods. There are many standard convex optimization software, such as CVX \[89\], YALMIP \[90\], developed for solving different classes of convex problems.

The benefits of convex optimization only come when the problem is a convex problem. In many cases, the original optimization problem does not have a standard convex form. Recognizing a convex problem or transforming an original problem into a convex problem is a big challenge. In the following, we will introduce three widely considered resource allocation problems in wireless cooperative systems. Some of them can be reformulated to a convex problem, thus, efficiently solved via convex optimization approaches.

3.3.1 Power Minimization

The first category of the optimization problems is to minimize some functions of transmit power of \(M\) BSs subject to SINR constraints for the selected users. We assume that the target SINR values are feasible. The weighted sum transmit power minimization problem can be formulated as

\[
\begin{align*}
\text{minimize} & \sum_{v=1}^{M} \alpha_v P_{\text{trans},v} \\
\text{subject to} & \quad 1) P_{\text{trans},v} \leq P_{v,\text{max}}, \forall v \\
& \quad 2) \gamma_k \geq \Gamma_k, \forall k.
\end{align*}
\]

Plugging (3.4) and (3.7) into (3.19), problem (3.19) becomes

\[
\begin{align*}
\text{minimize} & \quad \sum_{v=1}^{M} \alpha_v \sum_{k=1}^{K} \|w_{k,v}\|^2 \\
\text{subject to} & \quad 1) \sum_{k=1}^{K} \|w_{k,v}\|^2 \leq P_{v,\text{max}}, \forall v \\
& \quad 2) \frac{\|h_k^H w_k\|^2}{\sum_{i \neq k} \|h_k^H w_i\|^2 + \sigma^2} \geq \Gamma_k, \forall k.
\end{align*}
\]

The SINR constraints in (3.20) are complicated functions of the precoding vectors, making the problem non-convex in its original formulation. However, we can prove that problem (3.20) can be reformulated as a convex problem by using semi-definite relaxation from \[91\].
Recall that \( w_k \triangleq [w_{k,1}^T, \ldots, w_{k,M}^T]^T \in \mathbb{C}^{N \times 1} \). We define \( W_k \triangleq w_k w_k^H \succeq 0 \), \( R_k \triangleq h_k h_k^H \), and 
\[
Q_v \triangleq \text{diag} (Q_{1,v}, Q_{2,v}, \ldots, Q_{M,v})
\]  
where 
\[
Q_{i,v} = \begin{cases} 
I_{N_v}, & \text{if } i = v \\
0_{N_v \times N_v}, & \text{otherwise}.
\end{cases}
\]  
By gathering the power weights in a diagonal matrix form as 
\[
A \triangleq \text{diag} (\alpha_1 \Delta_1 I_{N_1}, \alpha_2 \Delta_2 I_{N_2}, \ldots, \alpha_M \Delta_M I_{N_M})
\]  
and noting that \( w_k^H Q w_k = \text{Tr} (Q W_k) \) for any matrix \( Q \), problem (3.20) can be reformulated as 
\[
\begin{align*}
\text{minimize} & \quad \sum_{k=1}^K \text{Tr} (A W_k) \\
\text{subject to} & \quad \sum_{k=1}^K \text{Tr} (Q_v W_k) \leq P_{v,\text{max}}, \forall v \\
& \quad \left(1 + \frac{1}{\Gamma_k}\right) \text{Tr} (R_k W_k) - \sum_{i=1}^K \text{Tr} (R_k W_i) \geq \sigma_k^2, \forall k
\end{align*}
\]  
with the additional constraints rank \((W_k) = 1, \forall k\). The problem (3.24) is convex except for the rank constraints, but based on [92, Theorem 1], it can be shown that (3.24) always has a rank one solution, if the problem is feasible. Therefore, the rank-one constraints can be dropped without loss of optimality. Thus, the optimization problem (3.19) can be transformed into a convex form, which can be efficiently solved via standard convex optimization techniques [93].

In case it is difficult to choose the weights for all BSs, an alternative is to minimize the maximum transmit power over all the \( N \) coordinated BSs as [94]
\[
\begin{align*}
\text{minimize} & \quad \max_{v} P_{\text{trans},v} \\
\text{subject to} & \quad P_{\text{trans},v} \leq P_{v,\text{max}}, \forall v \\
& \quad \gamma_k \geq \Gamma_k, \forall k.
\end{align*}
\]  
which is equivalent to 
\[
\begin{align*}
\text{minimize} & \quad \rho \\
\text{subject to} & \quad P_{\text{trans},v} \leq P_{v,\text{max}}, \forall v \\
& \quad \gamma_k \geq \Gamma_k, \forall k \\
& \quad P_{\text{trans},v} \leq \rho, \forall v.
\end{align*}
\]  
Similar to (3.19), problem (3.25) can be transformed into a convex form, thus solved via convex optimization.

When our target is to improve the energy efficiency of networks, the objective can be changed to minimize the total power consumption of the network instead of the total transmit power. The total power consumption can be modeled with a circuit part that
depends on the transceiver hardware and a dynamic part which is a function of the transmitted signal power. The circuit power consumption also depends on the operational mode of each BS, i.e., whether the BS is active or in the sleep mode. This objective function is typically non-convex, and it can lead to a hard combinatorial problem. One way to solve this problem is to perform iterative convex approximations of the nonconvex power consumption functions, and thereby find a local optimum to the original problem. In Paper F, by considering a linear approximated, non-continuous power consumption model proposed in [44], we illustrate how to design an iterative heuristic algorithm to efficiently obtain a local optimal solution for the design of downlink precoding and load balancing.

### 3.3.2 Worst SINR Maximization

Another category of joint precoding problems is to maximize the worst data rate or SINR subject to per-BS power constraints in order to guarantee the user fairness [95–97]. The optimization problem can be written as

\[
\begin{align*}
\text{maximize} & \quad \min_k \gamma_k \\
\text{subject to} & \quad P_{\text{trans},v} \leq P_{v,\text{max}}, \forall v.
\end{align*}
\]

Using the fact that for any given target SINR value \( \Gamma \), similar to (3.24), \( \gamma_k \geq \Gamma \) can be reformulated as a second order cone constraint, which is convex. Therefore, the objective function \( \min(\gamma_k) \) is quasi-concave in \( w_k, v \). Hence, (3.27) can be efficiently solved by using the bisection method, which is illustrated in Algorithm 1 [93].

**Algorithm 1** Bisection method for maximization of the worst SINR

given \( l \leq \gamma^* \) and \( u \geq \gamma^* \), tolerance \( \epsilon > 0 \).

repeat
1: \( \Gamma = (l + u)/2 \).
2: Solve the convex feasibility problem:
\[
\begin{align*}
\text{find} \quad & w_k, \forall k, \\
\text{subject to} & \quad 1) \sum_{k=1}^{K} p_{k}^{n} \leq \rho, \forall n, \\
& \quad 2) \gamma_k \geq \Gamma, \forall k.
\end{align*}
\]
3: if (3.28) is feasible then
4: \( l := \Gamma \);
5: else
6: \( u := \Gamma \).
7: end if
until \( u - l < \epsilon \).

In a system where some users have different QoS requirements, the design objective can be modified by replacing the \( \gamma_k \) with \( \gamma_k \beta_k \) in (3.27). Here, \( \beta_k \) is the weight for user \( k \) used to prioritize different users. In this case, the solution of (3.27) ensures a weighted user fairness among users.
3.3.3 Sum Rate Maximization

In general, the weighted sum rate maximization problem subject to per-BS power constraints can be formulated as

\[
\begin{align*}
\text{maximize } & \sum_{k=1}^{K} \beta_k R_k \\
\text{subject to } & P_{\text{trans},v} \leq P_{v,\text{max}}, \forall v.
\end{align*}
\]

This problem is not convex. Finding the optimal solution of (3.29) is typically non-tractable. However, some iterative algorithms can be designed to obtain the local optimal solutions, for example, by iteratively solving a set of Karush-Kuhn-Tucker (KKT) conditions of the non-convex problem [98], or by iteratively solving the problem in each step with respect to one variable keeping the other variables fixed [96].

Note that the precoding matrix specifies both the beamforming vectors and the allocated power to each data symbol. Thus, \( w_k \) can be further divided into two parts, i.e., the normalized beamforming vector \( u_k \) and the symbol power \( p_k \) allocated for the \( k \)th user, with \( w_k = p_k u_k \) and \( \|u_k\| = 1 \). A simple linear beamforming scheme for joint transmission is known as zero-forcing (ZF) beamforming, where the beamforming matrix \( U \) is firstly calculated as the pseudo-inverse of the channel matrix \( H^H \), then the columns of \( U \) are normalized to have a unit norm. With ZF beamforming, the inter-user interference within the cooperation cluster can be eliminated, that is

\[
h_k^H u_i = 0, \ i \neq k.
\]

The problem of maximizing the weighted sum rate in (3.29) is reduced to a joint power allocation problem given by

\[
\begin{align*}
\text{maximize } & \sum_{k=1}^{K} \beta_k \log_2 \left( 1 + \frac{\|h_k^H u_k\|^2 p_k}{\sigma^2} \right) \\
\text{subject to } & \sum_{k=1}^{K} \|u_{k,v}\|^2 p_k \leq P_{v,\text{max}}, \forall v.
\end{align*}
\]

where the beamforming vector \( u_k \) are fixed. The problem (3.31) is convex since the objective function is concave in \( p_k \) and the constraints are linear. Therefore, it can be effectively solved by standard convex optimization techniques [99, 100].

3.4 Remarks on Practical Constraints

The resource allocation problems discussed above are based on the assumption that the data symbols of all \( M \) users are perfectly synchronized at each BS. However, as mentioned in Chapter 2, the phase synchronization between coordinated BSs can be extremely difficult in practice. Imperfect phase synchronization can significantly reduce the joint transmission gain. In the worst case, as shown in Paper A, the random phase shift arising from the oscillators of different BSs can average out the effect of phase adjustment provided by joint precoding, thus, resulting in a different power allocation solution. Moreover,
in practice, the channel vectors $h_{k,v}$ are imperfectly known to the user $k$ and the BSs. This can be modeled as $h_{k,v} = \hat{h}_{k,v} + e_{k,v}$ with $v = 1, \ldots, M$, where $\hat{h}_{k,v}$ is the known channel estimate (at both the corresponding transmitter and receiver). The error vectors $e_{k,v} \sim \mathcal{CN}(0, \mathbf{E}_{k,v})$ are assumed to be zero-mean with covariance matrix $\mathbf{E}_{k,v} \in \mathbb{C}^{N_v \times N_v}$. The errors can, for example, originate from channel estimation/prediction errors [101]. Based on this channel model, in Paper F, we illustrate how to design the precoding vectors under imperfect CSI.
Chapter 4

Conclusions and Future Work

4.1 Contributions

The purpose of this thesis is to investigate the design and the performance of downlink cooperative wireless communication systems. To achieve this, a number of contributions (see the list of included papers and the list of additional related publications) are introduced considering different cooperative techniques in different network scenarios. This section summarizes the contributions of this thesis, which are divided into two parts: resource allocation in homogeneous cooperative networks, and resource allocation in heterogeneous cooperative networks.

4.1.1 Resource Allocation in Homogeneous Cooperative Networks

The first part of the contributions focus on the design of resource allocation algorithms for homogeneous cooperative networks. In particular, we consider the scenarios where different cooperative transmission schemes are performed between multiple macro BSs.


This paper considers the downlink of a two-cell two-user joint transmission system. We study a worst case scenario where the carrier phases between the BSs are un-synchronized so that multi-cell joint transmission must be performed without precoding. An optimal power allocation scheme is proposed to maximize the sum rate under per-BS power constraints. The derived power allocation scheme is remarkably simple, i.e., each cell transmits with full power to only one user. Note that joint transmission is still possible, when two cells select the same user for data transmission. In addition, we show that, in this scenario, the joint transmission case happens with higher probability when the maximum transmit power is high, or the two users are in the overlapped cell-edge area.

Paper B: “Resource Allocation for Clustered Network MIMO OFDMA Systems”

In this paper, we address the resource allocation problem for the downlink of a large network MIMO orthogonal frequency division multiple access (OFDMA) system with
3-sector BSs. The system is statically divided into a number of disjoint clusters of sectors. A two-step resource allocation scheme is proposed involving the inter-cluster and the intra-cluster levels. As a first step or inter-cluster level, the inter-cluster interference is mitigated by two fractional frequency reuse approaches, which restrict the available frequency resources for cluster-edge users in a cooperative way. Then, as a second step or intra-cluster level, a utility-based joint scheduling and power allocation algorithm is proposed for each cluster, to maximize the sum utility of all users in the cluster under per-sector power constraints. Zero-forcing based coherent joint transmission is used across multiple sectors within the same cluster. Our simulation results show that the proposed scheme can efficiently reduce the inter-cluster interference and provide considerable performance improvement in terms of both the cell-edge and cell-average user data rate. The proposed two-step resource allocation scheme can be implemented independently in each cluster without inter-cluster information exchange, which is an attractive property for practical systems, since it reduces both the network signaling overhead and the computational complexity.

Paper C: “Performance Evaluation of Coordinated Multi-Point Transmission Schemes with Predicted CSI”

In this paper, we consider the downlink of a CoMP cluster and compare three different CoMP transmission schemes: zero-forcing coherent joint transmission, non-coherent joint transmission and coordinated scheduling. For each of the analyzed schemes, the performance in terms of average sum rate of the CoMP cluster is studied with predicted CSI, considering the effects of the feedback and backhaul latency, as well as the user mobility. Compared to zero-forcing coherent joint transmission, we show that non-coherent joint transmission and coordinated scheduling are more robust to channel uncertainty. In addition, depending on the latency, user mobility and user locations, different schemes would achieve the highest average sum rate performance. Hence, a system could switch between the transmission schemes to improve the sum rate.

Related contributions

In [C6], we investigate multi-cell resource allocation with zero-forcing coherent joint transmission for a downlink OFDMA system with universal frequency reuse. Joint optimization of the user selection and power allocation is studied with the objective of maximizing the weighted sum rate under per-BS power constraints. Our results in [C6] show that joint user set selection across multiple subchannels significantly improves the system performance in terms of the weighted sum rate.

Coherent joint transmission requires tight phase synchronization between coordinated BSs and the knowledge of network CSI at the transmitters with sufficiently high quality. When those constraints are not satisfied, non-coherent joint transmission together with coordinated scheduling can be used as a fallback transmit mode to harvest the macro-diversity gains. Considering non-coherent joint transmission, in [C7], we propose two joint user scheduling and power allocation algorithms for the downlink of a CoMP cluster with 3 neighboring base station sectors. We prove that, in this scenario, binary power control is the optimal strategy for maximizing the cell-edge sum rate under per-sector transmit power constraints. Moreover, the simulation results in [C7] demonstrate that the
4.1 Contributions

proposed algorithms achieve a good trade-off between joint transmission and interference coordination, which helps to improve the cell-edge performance. The results in [C7] are limited to the flat-fading channel case. In [C12, J2], we investigate non-coherent joint transmission in downlink multi-cell OFDMA systems. In particular, an efficient optimal power allocation algorithm is derived for maximizing the sum rate of a two-cell system under per-cell transmit power constraints in [C12]. In [J2], utility based coordinated scheduling schemes are proposed to maximize the users sum utility. The utility function is based on the user average throughput, which provides a good balance between the system spectrum efficiency and fairness between the users.

Apart from the backhaul latency, backhaul capacity can also become a constraint restricting multi-cell coordination. This issue is partly addressed in [C11] and [C12], where different backhaul load reduction schemes are proposed for zero-forcing joint transmission. In particular, a decentralized network architecture is considered in [C11] for backhaul load reduction. In this setting, users broadcast the quantized CSI such that the coordinating BSs could simultaneously receive the CSI. The advantage of this decentralized architecture is that it does not require a CU, and each BS can design its own precoding matrix and perform data transmission locally with minimum amount of information sharing over the backhaul links. [C12] considers the feedback and backhaul load reduction problem in a centralized CoMP architecture. In this work, a feedback load reduction technique is employed via partial joint processing to alleviate the CSI feedback overhead. Similarly, to reduce the backhaul load due to precoding weights forwarded from the CU to the coordinated BSs, different scheduling approaches are proposed, which choose the best subset of the BSs and UEs at the CU that yields the best sum rate under different constraints of efficient backhaul use. Simulation results in [C12] show that scheduling a smaller subset of BSs and users can potentially achieve a better tradeoff in terms of the sum rate per backhaul use.

All the above contributions adopt the assumption of full buffer traffic model for all users, however, in practical communication systems, different users may have different traffic patterns, resulting in diverse QoS requirements. Taking this into account, different utility based joint resource allocation algorithms are proposed in [J4, C14, C15], considering mixed real-time voice over IP and best-effort services. System level simulation results show that, by exploiting multi-cell joint transmission and utility diversities, the proposed algorithms can provide a great improvement in the average throughput of best-effort users, and meanwhile substantially reduce the average packet drop ratio and call outage ratio of voice over IP users.

4.1.2 Resource Allocation in Heterogeneous Cooperative Networks

The second part of the contributions considers the design of resource allocation algorithms and performance analysis for heterogeneous cooperative networks. The studies mainly focus on two scenarios, i.e., relay-assisted networks and multi-tier small cell networks, where different challenges, e.g., the hardware impairments in low-power low-cost relay nodes, unreliable backhaul links between different types of BSs, random node activities, and load balancing, have been addressed.
Paper D: “I/Q Imbalance in Two-Way AF Relaying”

This work analyzes the performance of dual-hop, two-way amplify-and-forward relaying in the presence of IQI at the relay node. In particular, two power allocation schemes, namely, fixed power allocation (FPA) and instantaneous power allocation (IPA), are proposed to improve the system reliability and robustness against IQI under a total transmit power constraint. For each proposed scheme, the outage probability is investigated over independent, non-identically distributed Nakagami-m fading channels, and exact closed-form expressions and bounds are derived. Our theoretical analysis indicates that without IQI compensation, IQI can create fundamental performance limits on two-way relaying. However, these limits can be avoided by performing IQI compensation at source nodes. Compared to the equal power allocation scheme, our numerical results show that the two proposed power allocation schemes can significantly improve the outage performance, thus reducing the IQI effects, especially when the total power budget is large. We also observe that IPA is particularly effective for the symmetric channel case. On the other hand, for the asymmetric channel case, it is better to select the FPA scheme to mitigate I/Q imbalance and keep the signaling overhead as low as possible.


In this paper, adaptive power allocation is investigated for hybrid automatic repeat request (HARQ) based relay networks, where the helping relay node becomes active only if it successfully decodes the data transmitted from the source node. For HARQ-based relaying techniques, relay-assisted cooperative transmission can be divided into two categories, namely, Single-Node Transmission (SNT) and multi-node Joint Transmission (JT). In the SNT mode, only one node (either the source or the relay) is active in each retransmission round. In the JT mode, once the relay decodes the data correctly, the source and the relay use, e.g., distributed space-time coding, to provide joint retransmission to the destination. Here, we study a cooperation mode switch in HARQ-based relaying by using adaptive power allocation. The outage probability is minimized under a total power constraint. Our results demonstrate that adaptive power allocation reduces the outage probability significantly. Moreover, depending on the channel conditions and the total power budget, switching between the SNT mode and the JT mode is the optimal way for minimizing the outage probability. Another important observation is that, in order for minimizing the power-limited outage probability, the optimal relay position with equal power allocation is closer to the source. On the contrary, when performing adaptive power allocation, the optimal relay position is closer to the destination.


This paper considers the downlink of a heterogeneous network, where multiple BSs can serve the users by non-coherent multiflow beamforming. We assume imperfect channel state information at both BSs and users. The objective is to jointly optimize the precoding, load balancing, and BS operation mode (active or sleep) for improving the energy efficiency of the network. The considered problem is to minimize the weighted total power consumption (both circuit power and dynamic transmit power), while satisfying per-user
quality of service constraints and per-BS transmit power constraints. This problem is non-convex, but we prove that for each combination of BS modes, the considered problem has a hidden convexity structure. Thus, the global optimal solution is obtained by an exhaustive search over all possible BS mode combinations. Furthermore, by iterative convex approximations of the non-convex power consumption functions, a heuristic algorithm is proposed to obtain a local optimal solution with low complexity. We show that although multi-cell joint transmission is allowed, in most cases, it is optimal for each user to be served by a single BS. An optimal BS association condition is parametrized, which reveals how it is impacted by different system parameters. Simulation results illustrate that our proposed algorithms significantly reduce the total power consumption, compared to the scheme where all BSs are continuously active. This implies that putting BSs into sleep mode by proper load balancing is an important solution for energy savings in heterogeneous networks. Moreover, the BS activation probability depends on the target QoS requirements, as well as the ratio between the circuit power consumed in the active mode and that consumed in the sleep mode.

Related contributions

In heterogeneous dense networks, the access points will be heterogeneous in transmit power, coverage area, and activation probability. In this complex scenario, high-capacity and reliable feedback links are unlikely to be available due to the limited bandwidth and high inter-cell interference. Furthermore, the backhaul links interconnecting multiple nodes are highly likely to be wireless and unreliable. In [C10], we propose a backhauling model by assigning a link failure probability to each backhaul link. The performance of various multi-node coordinated transmission schemes is investigated under unreliable backhaul. We show that the performance gains offered by multi-node coordination quickly diminish, as the unreliability of the backhaul links grows. This work is extended in [J4], where the impact of both the feedback and the backhaul channel reliability on multi-node coordination is studied under three different network architectures. In particular, two transmission schemes, coherent joint transmission and coordinated scheduling, are evaluated under the centralized, semi-distributed, and fully-distributed CoMP architectures. Numerical results show that cooperative transmission techniques have the potential to improve the performance of the cellular system, in terms of sum rate. However, the performance of the system highly depends on the reliability of the control channels and, more importantly, on the probability of successful channel state information exchange. Moreover, we show that the semi- and fully distributed architectures are more robust to LFP, and the performance of the considered cooperative transmission schemes under these architectures will converge to traditional single cell transmission, as the link failure probability grows.

Due to some practical constraints, e.g., energy saving, bad channel quality, and time constraint, the coordinated nodes may not always be available for helping data transmission. Also, the receiving nodes may not request data all the time. In [C5], we study the power allocation problem for the downlink of a cooperative system with different node activeness; that is, each receiving node requests for data transmission according to a certain probability. Data symbols of the active nodes are jointly transmitted from cooperative transmission points using zero-forcing precoding. The problem is cast in form of maximizing the achievable sum rate subject to per-transmit-point average total power.
Conclusions and Future Work

The optimal power allocation solution is proved to fall into two categories: 1) Greedy power allocation, when two receiving nodes are located close to one transmission node and the per-transmit-point power budget is below a threshold; 2) Power sharing solution, i.e., both receiving nodes receive the data from the transmission nodes, when different receiving nodes are close to different transmission nodes or the per-transmit-point total power budget is above a threshold. Moreover, we show that depending on the channel condition and the RNs’ activation probability, a system should switch between cooperative joint transmission and non-cooperative transmission to improve the sum rate.

4.2 Future Work

In this thesis, we have investigated different multi-node cooperative techniques in the downlink of realistic wireless communication networks. We showed that, by proper design of radio resource allocation algorithms, multi-node cooperation can provide promising performance gains in terms of system throughput, user fairness, system reliability, and/or energy efficiency.

Throughout the thesis, we have made a simplifying assumption that each user has only a single receiving antenna. As future users are high likely to be equipped with multiple antennas, interference cancellation can be performed at the user side to further improve the received SINR, and thereby improving the overall system performance. Future work should look into the joint design of the cooperative transmission schemes at the transmitters and the interference cancellation schemes at the receivers.

As shown in Paper F, putting BSs into sleep mode by proper load balancing is an important solution for energy savings in heterogeneous networks. The joint precoding and load balancing algorithms designed in Paper F are based on the assumption of a full buffer traffic model for all users. However, in order to provide a realistic analysis of the network energy efficiency, it is essential to know what kind of traffic demands are supported by the network. The traffic load situation of a network may vary radically over the time. Therefore, in order to always operate in an energy efficient mode, network resource optimization solutions should be derived by taking realistic traffic models into account.
Bibliography


[18] 3GPP R1-104968, “Summary of the description of candidate eICIC solutions,” CMCC.


[20] 3GPP R1-112037, “Considerations on further enhancements of Rel-10 eICIC,” Huawei, HiSilicon.


[49] 3GPP R1-133456, “Views on small cell on/off mechanisms,” NTT DOCOMO.


Part II

Included papers
Power Allocation for Two-Cell Two-User Joint Transmission

Jingya Li, Thomas Eriksson, Tommy Svensson and Carmen Botella

Published in
IEEE Communication Letters,
vol. 16, Issue 9, pp. 1474-1477.
Sept. 2012
©2012 IEEE
The layout has been revised.
Abstract

In this paper, we develop a power allocation scheme for the downlink of a two-cell two-user joint transmission system. The objective is to maximize the sum rate under per-cell power constraints. We study a worst case scenario where the carrier phases between the two base stations are un-synchronized, so that joint transmission must be performed without precoding. The derived power allocation scheme is remarkably simple, i.e., each cell transmits with full power to only one user. Note that joint transmission is still possible, when two cells select the same user for data transmission. Moreover, we prove that, in this scenario, the joint transmission case happens with higher probability when the maximum transmit power is high, or the two users are in the overlapped cell-edge area.
A.1 Introduction

Transmit power allocation is an effective way for increasing the sum rate of wireless communication systems. It has been proved in [1] that water-filling power allocation is optimal when maximizing the sum rate under a sum power constraint, assuming different data signals are transmitted through orthogonal channels. Considering a broadcast channel that allows different user data to share the same channel, the optimal solution is then achieved by assigning the total power to the only one user with the best channel gain, namely greedy power allocation (GPA) [2]. Due to the neglecting of interference and the sum power constraint, the results in [1, 2] are not readily applicable for a two-cell power allocation analysis. By assuming per-cell power constraints, it has been proved in [3] that the maximum sum rate for a two-cell system is achieved by binary power control (BPC), i.e., each cell either transmits with full power or does not transmit. However, the formulation of the optimization problem in [3] is restricted to the condition that only one user can be served in each cell at a time slot.

Coordinated multi-point (CoMP) joint transmission has been considered as a promising technique to increase the sum rate and cell-edge performance. If both user data and channel state information (CSI) are shared by coordinated base stations (BSs), multiple BSs can jointly provide data transmission to a user and thereby improve the received signal quality. Major setbacks of CoMP joint transmission are, however, the large CSI feedback overhead, the capacity and latency constraints of backhaul links, and the imperfect synchronization between coordinated BSs [4]. A tradeoff between the system performance and the required amount of CSI feedback and backhaul exchange has been pointed out [4, 5]. This tradeoff is one of the reasons for restricting the use of joint transmission to a limited number of cells of the system [6, 7]. In [6], considering a two-cell single user CoMP system, a joint water-filling power allocation (Jo-WF) is proved to be optimal for maximizing the capacity of a frequency-selective channel. Since different data are assumed to be transmitted through orthogonal channels, the interference does not exist, which simplifies the optimization problem in [6]. Taking interference into account, BPC is proved to be optimal for maximizing the sum rate with joint transmission [7]. However, similar to [3], it is restricted that at most one user is allowed to be served in each cell at a time slot.

In this paper, the power allocation problem is addressed for the downlink of a two-cell two-user joint transmission system, in a scenario where the carrier phases between two BSs are un-synchronized. As we motivate in Section II, in some applications phase noise and frequency drift will lead to that phase synchronization between BSs is very difficult to achieve. For such applications the studied scenario is the most realistic; for other applications, it can be considered as a worst-case scenario. The objective is to maximize the sum rate under per-cell power constraints. The optimal solution for this scenario is shown to be simple, i.e., each cell transmits with full power to only one user. Note that joint transmission is still possible, when two cells select the same user for data transmission. Moreover, we show that dynamic switching between CoMP and Non-CoMP transmission, depending on the channel condition, is the optimal way under a per-cell power-limited condition for the considered scenario.
A.2 System Model

We consider the downlink of a two-cell two-user CoMP system, where data symbols of the two users are jointly transmitted from the two cells at the same time using the same spectral resource (see Figure A.1). In principle, if infinite cooperation between the two cells is enabled, the scenario is effectively a network MIMO broadcast channel, where all communication links (including the interfering ones) are exploited to coherently transmit user data via joint precoding [4]. However, from a practical point of view, a big challenge for the real-world implementation of CoMP joint transmission is the issue of synchronization between coordinated BSs. The carrier phases between coordinated BSs are difficult to be synchronized mainly due to the effect of carrier frequency offset, or/and phase noise from local oscillators in each BS [8, 9]. BSs with un-synchronized oscillators will result in independent time-continuous phase shifts imposed on the links between BSs and UEs. Here, we study a worst case scenario where the phase shift of each link (arising from the oscillator) varies much faster than the channel fading [9], and the phase difference between any pair of links shows a random uniform distribution between $-\pi$ to $\pi$. Note that the performance of joint precoding highly relies on reasonably accurate phase information of the signals that will be received. In the considered scenario, due to the random uniform phasing effect, precoding is not considered when performing joint transmission\(^1\).

For any given time slot, let $h_{im}$ denote the channel from cell $i$ to user $m$, which is assumed to be constant over one time slot, with $i = 1, 2$ and $m = 1, 2$. Let $\theta_{im}$ denote the phase shift imposed on $h_{im}$, which varies over this time slot with $(\theta_{im} - \theta_{jk}) \sim U(-\pi, \pi)$, for any $(i, m) \neq (j, k)$. The data symbols, denoted by $x_1$ and $x_2$, are jointly transmitted from both cells to UE 1 and UE 2 respectively, without precoding. The power allocation across users in cell $i$ is defined by $P_i = [P_{i1}, P_{i2}]$. Each link from cell $i$ to user $m$ contains both the desired data symbols $x_m$ with transmit power $P_{im}$ and the unwanted interference data symbols $x_k$ with power $P_{ik}, k \neq m$. Assume that $x_1$ and $x_2$ are independent Gaussian distributed random variables with zero mean and unit variance. Then, the signal-to-

\(^1\)The achieved sum rate is a lower bound for a precoded system with perfect synchronization.
interference-plus-noise ratio (SINR) of user \( m \) can be written as \(^2\)

\[
\gamma_m = \frac{E_{\theta,x} \left( \left| \sum_{i=1}^{2} \sqrt{P_{im} h_{im} e^{j\theta_{im}}} x_{m} \right|^{2} \right)}{E_{\theta,x} \left( \left| \sum_{i=1, k \neq m}^{2} \sqrt{P_{ik} h_{im} e^{j\theta_{im}} x_{k}} \right|^{2} \right) + \sigma^2} = \frac{\sum_{i=1}^{2} P_{im} G_{im}}{\sum_{i=1, k \neq m}^{2} P_{ik} G_{im} + \sigma^2}, \quad m = 1, 2,
\]

where \( G_{im} = |h_{im}|^2 \) and \( \sigma^2 \) is the variance of the independent zero-mean additive white Gaussian noise. By treating interference as noise, the expected sum rate of the two users normalized by the bandwidth can be expressed as

\[
R = \log_2 ((1 + \gamma_1)(1 + \gamma_2)).
\]

### A.3 Optimal Transmit Power Allocation

The objective is to find an optimal power allocation vector \( \mathbf{P}_i^* \) for each cell \( i \) that maximizes the sum rate \( R \) subject to per-cell power constraints. Since the logarithm is a monotonically increasing function, the equivalent problem is

\[
\max_{\mathbf{P}_1, \mathbf{P}_2} J(\mathbf{P}_1, \mathbf{P}_2) = \left( 1 + \frac{\sum_{i=1}^{2} P_{i1} G_{i1}}{\sum_{i=1}^{2} P_{i2} G_{i1} + \sigma^2} \right) \times \left( 1 + \frac{\sum_{i=1}^{2} P_{i2} G_{i2}}{\sum_{i=1}^{2} P_{i1} G_{i2} + \sigma^2} \right)
\]

such that \( \sum_{m=1}^{2} P_{im} \leq P_{i,\text{max}}, \; P_{im} \geq 0, \; i = 1, 2 \).

**Lemma 1.** Let \( \mathbf{P}_i^* \) be the optimal transmit power allocation vector, \( i = 1, 2 \). Then there is at least one cell index \( i \), such that the power constraint for this cell is satisfied with equality.

**Proof:** Let \( \mathcal{F} \) denote the feasible set of (3). Consider a pair of transmit power allocation vectors \( [\mathbf{P}_1, \mathbf{P}_2] \in \mathcal{F} \) such that for each \( i = 1, 2 \), \( \sum_{j=1}^{2} P_{im} < P_{i,\text{max}} \). Then, it is possible to find a factor \( \alpha \), with \( \alpha = \min_i (P_{i,\text{max}}/\sum_{m=1}^{2} P_{im}) > 1 \), such that at least one of the per-cell power constraints is satisfied with equality, and

\[
J(\alpha \mathbf{P}_1, \alpha \mathbf{P}_2) = \left( 1 + \frac{\sum_{i=1}^{2} P_{i1} G_{i1}}{\sum_{i=1}^{2} P_{i2} G_{i1} + \sigma^2/\alpha} \right) \times \left( 1 + \frac{\sum_{i=1}^{2} P_{i2} G_{i2}}{\sum_{i=1}^{2} P_{i1} G_{i2} + \sigma^2/\alpha} \right) > J(\mathbf{P}_1, \mathbf{P}_2).
\]

Thus, for any power allocation vector \( [\mathbf{P}_1, \mathbf{P}_2] \) satisfying \( \sum_{m=1}^{2} P_{im} < P_{i,\text{max}}, \; i = 1, 2 \), it is possible to find a power allocation vector \( [\alpha \mathbf{P}_1, \alpha \mathbf{P}_2] \) that achieves a larger sum rate, when one of the per-cell power constraints is satisfied with equality, i.e., \( \sum_{m=1}^{2} P_{im} = P_{i,\text{max}}, \; i = 1 \) or 2.

\(^2\)Note that only the amplitude information of the channel coefficients are needed, which reduces both feedback and backhaul overhead.
Lemma 2. Let the power constraint for cell $i$ be satisfied with equality, $i = 1$ or 2. Then cell $i$ transmits to only one user with its maximum transmit power.

Proof: Without loss of generality, we assume that the power constraint for cell 1 is satisfied with equality, i.e., $P_{11} + P_{12} = P_{1,\text{max}}$. Thus, for any given feasible transmit power allocation vector $P_2$ of cell 2, the objective function in (3) can be simplified as a function of $P_{11}$

$$J(P_{11}) = \frac{(P_{1,\text{max}} G_{11} + (P_{21} + P_{22}) G_{21} + \sigma^2)}{(P_{1,\text{max}} - P_{11}) G_{11} + P_{22} G_{21} + \sigma^2)} \times \frac{(P_{1,\text{max}} G_{12} + (P_{21} + P_{22}) G_{22} + \sigma^2)}{P_{11} G_{12} + P_{21} G_{22} + \sigma^2}$$

$$= \frac{-P_{11}^2 + BP_{11} + C}{A},$$

where

$$A = (P_{1,\text{max}} G_{11} + (P_{21} + P_{22}) G_{21} + \sigma^2) \times (P_{1,\text{max}} G_{12} + (P_{21} + P_{22}) G_{22} + \sigma^2) / G_{11} G_{12} > 0,$$

$$B = P_{1,\text{max}} + (P_{22} G_{21} + \sigma^2) / G_{11} - (P_{21} G_{22} + \sigma^2) / G_{12},$$

$$C = (P_{1,\text{max}} G_{11} + P_{22} G_{21} + \sigma^2) (P_{21} G_{22} + \sigma^2) / G_{11} G_{12}.$$

Since $A > 0$ and independent of $P_{11}$, the solution of maximizing $J(P_{11})$ is the solution for minimizing $-P_{11}^2 + BP_{11} + C$. Note that $0 \leq P_{11} \leq P_{1,\text{max}}$. If $B \leq 0$, $-P_{11}^2 + BP_{11} + C$ is a monotonically decreasing function of $P_{11}$. Else if $B \geq 2P_{1,\text{max}}$, $-P_{11}^2 + BP_{11} + C$ is an increasing function of $P_{11}$. Otherwise, $-P_{11}^2 + BP_{11} + C$ is a concave function of $P_{11}$. In either case, the maximum value of $J(P_{11})$ is obtained at the boundary point of $P_{11}$, i.e., 0 or $P_{1,\text{max}}$. Since $P_{11} + P_{12} = P_{1,\text{max}}$, the optimal transmit power allocation vector $P_1^* = \{P_{11}^*, P_{12}^*\}$ for cell 1 is $[P_{1,\text{max}}, 0]$ or $[0, P_{1,\text{max}}]$. Note that $P_1^*$ is optimal for any given feasible power allocation vector $P_2$. Hence, $P_1^*$ is the optimal solution for (3), given $P_{11} + P_{12} = P_{1,\text{max}}$. Thus, cell 1 transmits to only one user with its maximum transmit power.

Based on Lemmas 1 and 2, we present the following theorem for the optimal power allocation solution.

Theorem 1. For the two-cell two-user CoMP joint transmission case, considering a random uniform phasing effect due to un-synchronized oscillators, the sum rate maximizing power allocation under per-cell power constraints is that each cell should transmit to only one user with its maximum transmit power. Mathematically, let $\triangle F_i$ denote two sets of the corner points of the feasible domain for cell $i$, with $\triangle F_i = \{[P_{1,\text{max}}, 0], [0, P_{1,\text{max}}]\}$, $i = 1, 2$. Then,

$$[P_1^*, P_2^*] = \arg\max_{P_1 \in \triangle F_1, P_2 \in \triangle F_2} J(P_1, P_2).$$

Proof: Based on Lemma 1 and Lemma 2, at least one cell will transmit with full power to only one user. Without loss of generality, we assume that cell 1 transmits with full
power to user 1, i.e., $P_1^* = [P_{1,\text{max}}, 0]$. Then, (3) becomes

$$\max_{P_2} J(P_2) = \left(1 + \frac{P_{\text{max}} G_{11} + P_2 G_{21}}{P_{22} G_{21} + \sigma^2}\right) \times \left(1 + \frac{P_{\text{max}} G_{12} + P_2 G_{22}}{P_{22} G_{22} + \sigma^2}\right)$$

s.t. $P_{1} + P_{22} \leq P_{2,\text{max}}$, $P_{21} \geq 0$, $P_{22} \geq 0$.

In order to find $P_2^*$, similar to the one in [3], by calculating the derivative of $J(P_2)$ with respect to $P_{21}$, we have

$$f(P_2) = \frac{\partial J(P_2)}{\partial P_{21}} = \frac{D(P_{21} + E)^2 + F}{I},$$

where

$$D = G_{21} G_{22} > 0,$$

$$E = \left(\frac{P_{\text{max}} G_{12} + \sigma^2}{G_{22}}\right) > 0,$$

$$F = P_{22} G_{22} G_{21} \left(P_{\text{max}} G_{12} + \sigma^2\right) - P_{22} G_{22} \left(P_{\text{max}} G_{11} + P_{22} G_{21} + \sigma^2\right),$$

$$I = \left(P_{\text{max}} G_{12} + P_{22} G_{22} + \sigma^2\right)^2 \left(P_{22} G_{21} + \sigma^2\right) > 0.$$ 

Since $I > 0$, the solution of $f(P_2) = 0$ comes from the solution of $D(P_{21} + E)^2 + F = 0$. Note that $0 \leq P_{21} \leq P_{2,\text{max}}$. Hence, if $F \geq -D E^2$, $J(P_2)$ is an increasing function of $P_{21}$. Else if $F \leq -D(P_{2,\text{max}} + E)^2$, $J(P_2)$ is a decreasing function of $P_{21}$. Otherwise, $D(P_{21} + E)^2 + F$ has one zero for $P_{21}$ and changes from negative to positive when increasing $P_{21}$, i.e., there is a minimum point for $J(P_2)$. In either case, the maximum value of $J(P_2)$ is obtained at the boundary point of $P_{21}$. Using the same method, the above analysis also holds for $P_{22}$. Hence, $P_2^*$ is found in the set of boundary points of the feasible domain, that is,

$$P_2^* \in \{P_{21} + P_{22} = P_{2,\text{max}}, P_{21} \geq 0, P_{22} \geq 0\}$$

$$\cup \{P_{21} = 0, 0 \leq P_{22} \leq P_{2,\text{max}}\}$$

$$\cup \{P_{22} = 0, 0 \leq P_{21} \leq P_{2,\text{max}}\}. \quad (9)$$

If $P_{21}^* = 0$, the objective function in (7) becomes to maximize $J(P_{22}) = \left(1 + \frac{P_{\text{max}} G_{11}}{P_{22} G_{21} + \sigma^2}\right)$ subject to $0 \leq P_{22} \leq P_{2,\text{max}}$, where the maximum value is obtained when $P_{22}^* = P_{2,\text{max}}$. Else if $P_{22}^* = 0$, $J(P_2)$ turns out to be a monotonically increasing function of $P_{21}$ in the domain of $0 \leq P_{21} \leq P_{2,\text{max}}$, with the optimal solution achieved by $P_{21}^* = P_{2,\text{max}}$. Otherwise, $P_{21}^* + P_{22}^* = P_{2,\text{max}}$ with $P_{21} \geq 0$, $P_{22} \geq 0$, and similar to Lemma 2, the solution of (7) is $P_2^* = [P_{2,\text{max}}, 0]$ or $[0, P_{2,\text{max}}]$. In conclusion, assuming $P_1^* = [P_{1,\text{max}}, 0]$, the optimal solution of (3) is found by $P_2^* = [P_{2,\text{max}}, 0]$ or $[0, P_{2,\text{max}}]$. Due to the symmetry, we can conclude that the optimal solution for the considered worst case scenario is derived by (6). That is, depending on the noise and channel gains, each cell should transmit to only one user with its maximum transmit power. Note that joint transmission is still possible, when two cells select the same user for data transmission, e.g., $P_1^* = [P_{1,\text{max}}, 0]$ and $P_2^* = [P_{2,\text{max}}, 0]$. □
A.4 Joint Transmission Analysis

According to (6), the optimal power allocation falls into one of the following four cases:

1. Two cells jointly transmit data to user 1 with full power;
2. Two cells jointly transmit data to user 2 with full power;
3. Cell $i$ transmits data to user $i$ with full power, $i = 1, 2$;
4. Cell $i$ transmits data to user $m$ with full power, $i \neq m$, $i = 1, 2$ and $m = 1, 2$.

In this section, we address the following questions: Under what kind of conditions will each case be the optimal solution? For what kind of system will joint transmission (case 1 and case 2) happen with high probability?

Assume that the two cells have the same maximum power constraint, $P_{\text{max}}$. Let $R(n)$ denote the sum rate of case $n$, with $n = 1, 2, 3, 4$. Based on (2), we have

$$R(1) = \log_2 (1 + \beta (G_{11} + G_{21})), \quad R(2) = \log_2 (1 + \beta (G_{12} + G_{22})), \quad R(3) = \log_2 \left( \frac{G_{11}}{G_{21} + 1/\beta} \frac{G_{22}}{G_{12} + 1/\beta} \right), \quad R(4) = \log_2 \left( \frac{G_{21}}{G_{11} + 1/\beta} \frac{G_{12}}{G_{22} + 1/\beta} \right),$$

where $\beta = P_{\text{max}} / \sigma^2$. Let $\mathcal{G}(k)$ denote the feasible set of $R(k) = \max_n R(n)$, with $n = 1, 2, 3, 4$. Then, we have

$$\mathcal{G}(1) = \{G_x > G_y, G_x > -c_2, G_y < c_1\}$$
$$\mathcal{G}(2) = \{G_x < G_y, G_x < c_1, G_y > -c_2\}$$
$$\mathcal{G}(3) = \{G_x > c_1, G_y > c_1\}$$
$$\mathcal{G}(4) = \{G_x < -c_2, G_y < -c_2\}$$

where $G_x = G_{11} - G_{12}$, $G_y = G_{22} - G_{21}$, $c_1 = \beta G_{12} G_{21} > 0$, $c_2 = \beta G_{11} G_{22} > 0$. Hence, the first question can be answered by (10), which is illustrated in Figure A.2 (a).

Note that $\mathcal{G}(1) \cup \mathcal{G}(2)$ is the joint transmission region, which increases with $c_1$ or $c_2$. Since $c_1$ and $c_2$ monotonically increase with $P_{\text{max}}$, the joint transmission region increases with $P_{\text{max}}$. On the other hand, assume that $P_{\text{max}}$ is fixed. Since $c_1 > 0$, the channel power gains in region $\mathcal{G}(3)$ should satisfy $G_{11} > G_{12}$ and $G_{22} > G_{21}$, i.e., user $i$ is closer to the BS of cell $i$, $i = 1, 2$. Let $d_i$ denote the distance between user $i$ to its closer BS. When the two users move from the cell-center areas towards the overlapped cell-edge area, i.e., $d_i$ increases, it would be general that $G_{11}$ and $G_{22}$ decrease while $G_{12}$ and $G_{21}$ increase, leading to $G_x$ and $G_y$ decreasing while $c_1$ increasing. Thus, according to (10), the region of $\mathcal{G}(3)$ decreases when $d_i$ increases. For the region of $\mathcal{G}(4)$, the channel power gains should satisfy $G_{11} < G_{12}$ and $G_{22} < G_{21}$. In this case, the closer BS of user $i$ is cell $j$, with $i \neq j$, $i = 1, 2$ and $j = 1, 2$. Similarly, we can prove that the region of $\mathcal{G}(4)$ decreases.

---

3For cases 3) and 4), our system model turns out to be an interference channel, as each user can only be served by one cell.
when $d_i$ increases. Since $G(1) \cup G(2)$ is the complement set of $G(3) \cup G(4)$, the joint transmission region increases with $d_i$. Thus, we can conclude that the joint transmission region increases when the transmit power increases, or when the two users move from cell-center areas towards the overlapped cell-edge area (see Figure A.2).

### A.5 Numerical Results

The performance of the proposed power allocation scheme is studied by Monte-Carlo simulation, over 10,000 independent realizations of users’ locations. Based on the system model described in Figure A.1, we consider a two-cell two-user system with a radius of $R = 500$ m in each cell. The path loss is $L(d) = 128.1 + 37.6 \log_{10}(d)$ in dB. The shadowing standard deviation is 8 dB, and the fast fading is Rayleigh distributed. The AWGN power is -105 dBW. Assume that the two cells have the same maximum power constraint, $P_{\text{max}}$.

Figure A.3 (a) shows that the joint transmission probability increases with $P_{\text{max}}$ when users are located in the overlapped cell-edge area. In Figure A.3 (b), the joint transmission probability with different $d_i$ is plotted when $P_{\text{max}} = 20$ W. Two users are dropped along the dashed line in Figure A.1 symmetrically. We can see that the joint transmission probability increases with $d_i$, in agreement with the conclusion made in section IV.

Figure A.4 shows the sum rate with different $P_{\text{max}}$ when users are located in the overlapped cell-edge area. Zero-forcing joint precoding (ZFJP) is shown as a benchmark assuming perfect synchronization [10]. Considering imperfect synchronization, the proposed power allocation scheme (proposed-PA) is compared with the equal power allocation (EPA), GPA [2], BPC [3], and Jo-WF [6]. The proposed-PA scheme achieves the best performance in the considered worst case scenario. Moreover, the crossing point in Figure A.4 indicates that, when the transmit power is low, the Jo-WF scheme is superior to BPC by allowing joint transmission for one user. However, notice that interference is not taken into account in the Jo-WF scheme. Thus, when the transmit power is high (the interference becomes serious), BPC outperforms Jo-WF.
Figure A.3: (a) Joint transmission probability vs. $P_{\text{max}}$, (b) Joint transmission probability vs. distance ratio

Figure A.4: Average sum rate vs. maximum transmit power constraint per cell
A.6 Conclusions

We address the downlink power allocation problem for a two-cell two-user joint transmission system, where synchronization between base stations is extremely difficult. The derived solution has a simple feature, i.e., each cell transmits with full power to only one user. Numerical results show that the proposed scheme obtains significant gains over equal power allocation, binary power control, greedy power allocation and joint waterfilling schemes. Extension to systems with more than two cells and two users is our ongoing work. Early results indicate that the proposed scheme is still optimal in the targeted scenario when applying a low SINR approximation.

References


Paper B

Resource Allocation for Clustered Network MIMO OFDMA Systems

Jingya Li, Carmen Botella, and Tommy Svensson

Published in EURASIP Journal on Wireless Communications and Networking 2012
The layout has been revised.
Typographical adjustments have been made.
Abstract

In this paper, we address the resource allocation problem for the downlink of a large network MIMO OFDMA system with 3-sector base stations (BSs). The system is statically divided into a number of disjoint clusters of sectors. A two-step resource allocation scheme is proposed involving the inter-cluster and the intra-cluster levels. As a first step or inter-cluster level, two cooperative frequency reuse approaches are designed to mitigate the inter-cluster interference. A user partition method is proposed to divide the users of each cluster into cluster-edge and cluster-center users. To balance the cell-edge and the cell-average performance, a fairness jug function is introduced to determine the frequency partition of the cooperative frequency reuse approaches. Then, as a second step or intra-cluster level, a utility-based joint scheduling and power allocation algorithm is proposed for each cluster, to maximize the sum utility of all users in the cluster under per-sector power constraints. Zero-forcing joint transmission is used across multiple sectors within the same cluster. Simulation results show that the proposed scheme can efficiently reduce the inter-cluster interference and provide considerable performance improvement in terms of both the cell-edge and cell-average user data rate. The proposed two-step resource allocation scheme can be implemented independently in each cluster without inter-cluster information exchange, which is an attractive property for practical systems, since it reduces both the network signaling overhead and the computational complexity.
B.1 Introduction

Driven by the demands to support data applications at higher throughput and spectral efficiency, orthogonal frequency division multiplexing (OFDM) based multiple access is being considered as a promising technique for the future wireless networks. OFDMA has been adopted as the downlink access technology of 3rd generation partnership project (3GPP) long term evolution (LTE) and LTE-Advanced standards [1]. Based on the OFDM technique, OFDMA inherits the immunity to intra-cell interference. However, the inter-cell interference is still a major issue. In fact, a frequency reuse factor being equal to one causes serious inter-cell interference to users in the cell-edge areas, leading to poor cell-edge performance. Viable inter-cell interference mitigation approaches are reviewed in [2], including the use of power control, fractional frequency reuse, opportunistic spectrum access, intra and inter-cell interference cancellation, and multiple input multiple output (MIMO) techniques.

Recently, coordinated multi-point transmission/reception (CoMP) has been proposed in 3GPP LTE-Advanced as a key technique to increase the system spectrum efficiency as well as the cell-edge performance [1]. In the case of CoMP joint transmission, both data and channel state information (CSI) of the users in CoMP mode can be shared by coordinated multiple cells, which can act as a single and distributed antenna array. Data to a user can be simultaneously transmitted from the coordinated cells to improve the received signal quality. Hence, the inter-cell interference is reduced by exploiting the signals transmitted from other cells to assist the transmission rather than treating them as interference. Notice that this technique is also referred as network coordination or network MIMO [3].

In a global network MIMO system, without any feedback, backhaul and synchronization constraints, the inter-cell interference can be completely eliminated. However, from a practical point of view, a major setback of global coordination is the large amount of feedback needed from the users and the large signaling overhead required for the inter-cell information exchange. An interesting tradeoff between the system performance and the required amount of CSI feedback and backhaul exchange has been pointed out [4–9]. This tradeoff is one of the reasons for restricting the use of network MIMO techniques to a limited number of cells or areas of the system. The system is typically divided into clusters of cells, and the joint transmission is implemented within the cells included in each cluster. The cluster formation can be static [10–14], or dynamic [15–17]. The static cluster formation specifies a predefined set of clusters of cells which do not change in time, whereas the dynamic clustering approaches form the clusters based on the varying channel conditions that users experience to different cells. Note that a coordinated cluster also causes inter-cluster interference to the users in the neighboring clusters, especially to users in the cluster-edge area. Therefore, the design of efficient inter-cluster interference coordination strategies and radio resource management algorithms is of great interest in the field of clustered network MIMO systems.

Previous studies about resource allocation in network MIMO OFDMA systems have mainly focused on the global network MIMO case [18–20], or on the single cluster case without considering inter-cluster interference [21–23]. Recently, research has shifted towards the limited coordination case with different cluster formation models, that is, resource allocation with dynamic clustering [24, 25] or resource allocation with static clustering [14]. In [24], an inter-cluster interference canceler performing linear processing on
the downlink transmission signals is proposed for multi-user MIMO distributed antenna systems. However, a central unit (CU) is needed to collect the global CSI of all users in the system and to calculate the transmission weight vectors for each cluster. Notice that this centralized framework requires an enormous amount of feedback and backhaul overhead. In [25], a centralized multi-cell network scheduling algorithm is proposed to minimize the inter-cluster interference by performing clustering from the user's point of view, which also needs a CU for global network scheduling. For a system with a large number of cells and a large number of users, a high computational burden will be caused in the CU. In [14], the authors instead consider a more realistic system model, where the network is divided into a number of disjoint static clusters, and limited inter-cluster coordination is used to pre-cancel interference for the users at the edge of neighboring clusters. In this approach, perfect CSI is available at the cluster side for both the cluster users and edge users in the neighboring clusters. Hence, each cluster can help the edge users in the neighboring clusters by taking these users into account when designing the precoding matrices. However, with a large number of users in each cluster, in order to serve a cluster-edge user, all the neighboring clusters need to provide a given number of degrees of freedom by dropping some scheduled users of their own, which leaves fewer degrees of freedom to serve their own users. In addition, a joint scheduling across clusters is needed for the whole network, which requires inter-cluster communication and increases the complexity of the resource allocation design.

Fractional frequency reuse is a promising technique for inter-cell interference mitigation. Instead of using spatial degrees of freedom to suppress the inter-cell interference, it restricts the available frequency resources of different cells through a predefined frequency reuse rule or through appropriate power control. In [26], the division of frequency resources is investigated for the uplink of a linear network MIMO system. Since the system is considered to be uniformly clustered in a linear grid, the inter-cluster interference can be completely eliminated with simple reuse strategies, e.g., half of the available frequency resources are assigned to each cluster with different resources assigned to adjacent clusters. In [27], the authors consider a more realistic scenario and employ appropriate power control in frequency such that adjacent clusters generate different interference levels in different subchannels. However, the power control problem is formulated in a centralized way such that a CU is needed for the network to solve the optimization problem for all the clusters. In [28, 29], two frequency reuse schemes were proposed for a multi-cell OFDMA system, supporting non-coherent joint transmission to cell-edge users by user-centric dynamic clustering. Due to the user-centric nature of the clustering, these approaches also required a joint scheduling across cells for the whole network.

In this paper, we address the resource allocation problem for the downlink of a clustered network MIMO OFDMA system with 3-sector base stations (BSs). Each sector has one directional antenna, and it is associated with a directional cell area. The whole system is statically divided into disjoint clusters of sectors. Due to practical issues (e.g., synchronization constraints, feedback constraints, backhaul network constraints and the system complexity), inter-cluster information exchange may not be feasible in realistic cellular systems. Targeting practical scenarios, radio resource allocation is independently performed in each cluster without inter-cluster communication. Zero-forcing beamforming is considered as the coherent joint transmission scheme within each cluster, which allows multiple users to share the same subchannel in each time slot by choosing proper beamforming weights. A two-step resource allocation scheme is proposed, which involves
both inter-cluster and intra-cluster levels:

- As a first step or inter-cluster level of resource allocation, two novel cooperative frequency reuse approaches (CFR-1 and CFR-2) are proposed to mitigate the inter-cluster interference. A user partition method based on the long term channel gain is introduced to divide the users of each cluster into cluster-edge and cluster-center users. Frequency subchannels in each cluster are separated into two orthogonal sets, that is, cluster-edge and cluster-center frequency sets. The inter-cluster interference is reduced by mapping different groups of cluster-edge users to different subchannels of the cluster-edge frequency set in a cooperative way. We also show that the frequency partition (the size of the cluster-center frequency set) and the user partition threshold are the parameters that can be optimized to balance the cell-edge and cell-average performance.

- As a second step or intra-cluster level of resource allocation, a sub-optimal utility-based joint scheduling and power allocation algorithm is proposed for each cluster with low complexity. Assume that perfect CSI is available at the cluster side for the users within this cluster. The algorithm jointly determines the set of users scheduled on each subchannel, and the power allocation across subchannels. The objective is to maximize the sum utility of all users in the cluster subject to per-sector power constraints.

The main contributions of our scheme are listed as follows:

- Frequency reuse approach performed in the first step (inter-cluster level) can effectively reduce the inter-cluster interference for cluster-edge users. Moreover, the user partition and the frequency partition are performed at the first step. In this way, only a subset of users is mapped to each subchannel, leading to a significant reduction of both the feedback requirements and the computational complexity in the second step of the proposed scheme.

- The proposed two-step resource allocation scheme can be implemented in different time scales, i.e., the inter-cluster interference mitigation would be more static than the intra-cluster scheduling and power allocation performed in the second step. Moreover, radio resource allocation is independently performed in each cluster without inter-cluster information exchange. Therefore, no inter-cluster coordination links are needed, which is an attractive property for the deployment of practical systems.

The proposed resource allocation scheme is compared with the universal frequency reuse scheme and the inter-cluster interference pre-cancellation strategy proposed in [14]. Simulation results demonstrate that a significant improvement on both the cell-edge and the cell-average performance can be obtained by the proposed scheme, with a much lower computational complexity.

The rest of the paper is organized as follows: Section B.2 describes the system model and introduces the problem formulation for the downlink of a clustered network MIMO OFDMA system. In Section B.3, two cooperative frequency reuse approaches are proposed for mitigating inter-cluster interference. Then, for the intra-cluster level, a joint scheduling and power allocation algorithm is proposed for each cluster in Section B.4. Section B.5
B.2 System Model and Problem Formulation

Consider the downlink of a network MIMO OFDMA system with multiple three-sectoring BSs. Each sector has one directional antenna, and it is associated with a directional cell area. The three antennas of each BS are located at the same site. Each user is equipped with one receive antenna and assigned to a serving sector that is selected based on long-term channel gain, i.e., pathloss and shadow fading. The whole system is statically divided into a number of disjoint clusters, where each cluster consists of three neighboring sectors belonging to different BSs. An example of a coordinated cluster is illustrated in Figure B.1. Assume that all the clusters have the same number of sectors $B$. Each sector has the same $N$ subchannels, that is, there are $N$ subchannels available for joint transmission per cluster. Every cluster is working independently in the system. Without loss of generality, we consider that a given cluster is denoted by $c$ with $K_c$ users. Each of the sectors in cluster $c$ is denoted by $(c,b)$, where $b \in \{1,\ldots,B\}$. Concentrating on an arbitrary time slot, the sectors within cluster $c$ provide joint transmission for the scheduled users (denoted by $S(c,n)$) on each subchannel $n$ based on the available CSI. The received signal of the scheduled user $k$ on subchannel $n$ in cluster $c$ is given as

$$y_{k,n}^c = \sum_{b=1}^{B} h_{k,n}^{(c,b)} w_{k,n}^{(c,b)} x_{k,n}^c + \sum_{b=1}^{B} h_{k,n}^{(c,b)} \sum_{i \in S(c,n), i \neq k} w_{i,n}^{(c,b)} x_{i,n}^c$$
$$+ \sum_{\hat{c} \neq c}^{B} \sum_{b=1}^{B} h_{k,n}^{(\hat{c},\hat{b})} \sum_{j \in S(\hat{c},n)} w_{j,n}^{(\hat{c},\hat{b})} x_{j,n}^{\hat{c}} + z_{k,n},$$

(1)
where $h^{(c,b)}_{k,n}$ denotes the complex channel response between sector $(c,b)$ and user $k$ on subchannel $n$, consisting of path loss, shadow fading, and small-scale fading. $w^{(c,b)}_{k,n}$ is the beamforming weight for user $k$ on subchannel $n$ with respect to sector $(c,b)$. $x_{k,n}^{c}$ denotes the data symbol for user $k$ on subchannel $n$, which is transmitted from all the sectors inside cluster $c$. $z_{k,n}$ is the additive white Gaussian noise at user $k$ on subchannel $n$ with zero mean and variance $\sigma^2$.

Let $h_{k,n}^{c} = [h^{(c,1)}_{k,n}, \ldots, h^{(c,B)}_{k,n}]$ and $w_{k,n}^{c} = [w^{(c,1)}_{k,n}, \ldots, w^{(c,B)}_{k,n}]^T$ denote the channel vector and the beamforming vector from all sectors in cluster $c$ to user $k$ on subchannel $n$, respectively. The maximum number of users that can be supported on subchannel $n$ in cluster $c$ is bounded by the total number of transmit antennas of cluster $c$, i.e., $|S(c,n)| \leq B$. In this paper, zero-forcing joint transmission is used to eliminate the intra-cluster interference. The beamforming matrix is defined as the pseudo-inverse of the channel matrix. Thus, we have

$$h_{k,n}^{c}w_{i,n}^{c} = \begin{cases} 0, & i \in S(c,n), \ i \neq k; \\ 1, & i \in S(c,n), \ i = k. \end{cases} \quad \text{(2)}$$

Then, the received signal $y_{k,n}^{c}$ becomes

$$y_{k,n}^{c} = x_{k,n}^{c} + \sum_{c' \neq c} \sum_{b=1}^{B} h^{(c',b)}_{k,n} \sum_{j \in S(c,n)} w^{(c',b)}_{j,n} x_{j,n}^{c'} + z_{k,n}. \quad \text{(3)}$$

Denote $p_{k,n}^{c} = x_{k,n}^{c}x_{k,n}^{c*}$ as the symbol power allocated to user $k$ on subchannel $n$ across the $B$ sectors in cluster $c$. Then, the signal to interference plus noise ratio (SINR) of user $k$ on subchannel $n$ is derived as

$$\gamma_{k,n} = \frac{p_{k,n}^{c}}{\sum_{c' \neq c} \sum_{j \in S(c,n)} \|h_{k,n}^{c}w_{j,n}^{c'}\|^2 p_{j,n}^{c'} + \sigma^2}. \quad \text{(4)}$$

Finally, based on the Shannon theorem, the achievable transmission rate of user $k$ on subchannel $n$ can be expressed as

$$R_{k,n} = W \log_2(1 + \beta \gamma_{k,n}), \quad \text{(5)}$$

where $W$ is the bandwidth of each subchannel, and $\beta$ is the SINR gap, which is a constant related to the target bit error rate (BER) given as $\beta = -1.5/\ln(\text{BER})$ using M-QAM modulation [30]. Then, the instantaneous data rate for user $k$ at a given time slot becomes

$$R_k = \sum_{n=1}^{N} R_{k,n} = W \sum_{n=1}^{N} \log_2(1 + \beta \gamma_{k,n}). \quad \text{(6)}$$

The transmit power of sector $(c,b)$ on subchannel $n$ is given by

$$p_{n}^{(c,b)} = \sum_{k \in S(c,n)} \|w_{k,n}^{(c,b)}\|^2 p_{k,n}^{c}. \quad \text{(7)}$$
and the total transmit power of sector \((c, b)\) is

\[
P_{(c,b)} = \sum_{n=1}^{N} P_{n}^{(c,b)} = \sum_{n=1}^{N} \sum_{k \in S_{c,n}} \| w_{k,n}^{(c,b)} \|^{2} P_{k,n}^{c}.
\] (8)

In this paper, the maximum available transmit power at each sector is restricted to a \(P_{\max}^{(c,b)}\) value, that is, \(P_{(c,b)} \leq P_{\max}^{(c,b)}\) for sector \((c, b)\).

Targeting practical scenarios, radio resource allocation is independently performed in each cluster. The objective is to maximize the sum utility of all users in the cluster under per-sector power constraints. For any given time slot, the coordinated sectors within each cluster can jointly determine 1) the set of users scheduled on each subchannel, and 2) the symbol power allocated to each scheduled user.

Concentrate on one arbitrary cluster \(c\) with \(K_{c}\) users. Let \(U_{k}(\cdot)\) denote the utility function of user \(k\), which is assumed to be continuously differentiable, non-decreasing and concave to balance the efficiency and fairness of the system performance. Let \(P^{c} = [p_{k,n}^{c}]\) denote the \(K_{c} \times N\) sized symbol power allocation matrix in a scheduling interval, and \(S^{c} = [S_{c,n}]\) denote the selected user sets on each subchannel. Then, the objective function of maximizing the sum utility of all users under per-sector power constraints can be formulated as

\[
\max U (P^{c}, S^{c}) = \sum_{k=1}^{K_{c}} U_{k} (\overline{R}_{k}(t)) .
\] (9)

The average data rate of user \(k\) at time slot \(t\) is updated using an exponentially low-pass time window as [31]

\[
\overline{R}_{k}(t) = (1 - \rho) \overline{R}_{k}(t-1) + \rho R_{k}(t),
\] (10)

where \(\rho = (T_{s}/T_{w})\), \(T_{s}\) is the slot length, and \(T_{w}\) is the length of the window. \(R_{k}(t)\) is the instantaneous data rate for user \(k\) at a time slot \(t\) and can be derived by (6). Then, based on the first-order Taylor expansion, the objective function in (9) can be rewritten as

\[
\max U (P^{c}, S^{c}) = \sum_{k=1}^{K_{c}} U_{k}^{'} (\overline{R}_{k}(t-1)) \overline{R}_{k}(t),
\] (11)

which can be interpreted as maximizing the weighted sum rate, since \(U_{k}^{'} (\overline{R}_{k}(t-1))\) is fixed at time slot \(t\). From now on, \(\mu_{k}\) is used to represent \(U_{k}^{'} (\overline{R}_{k}(t-1))\). It should be pointed out that the first order Taylor expansion approximation is sub-optimal. However, this approximation relaxes the original complex optimization problem to a weighted sum rate maximization problem, which greatly simplifies the algorithm design. The weights are adaptively controlled by the marginal utility with respect to the current average rates. Specifically, as has been analyzed in [31], if the utility function is defined as a natural logarithm of the user’s average data rate at the current time slot, the objective becomes to maintain proportional fairness among users. Therefore, the utility-based algorithm presented by the first order Taylor expansion approximation can be treated as a general framework for allocating multi-user shared resources.

Substituting (6) and (8) into (11), the optimization problem under per-sector power
constraints can be expressed as

\[
\max U (P^c, S^c) = \sum_{n=1}^{N} \sum_{k \in S(n)} \mu_k W \log_2 (1 + \beta \gamma_{k,n})
\]

s.t. \( \sum_{n=1}^{N} \sum_{k \in S(n)} \|w_{k,n}^b\|^2 \leq P_{\max}^{(c,b)}, b \in \{1, \ldots, B\} \).

(12)

Note that the overall resource allocation problem is a non-convex combinatorial optimization problem. In addition, to solve the overall optimization problem, a significant amount of CSI feedback and information exchange between clusters is needed. Thus, computing its optimal solution within one step would require global network coordination, which is not realistic for implementation in real systems. Therefore, we propose a two-step resource allocation scheme to get closer to a practical implementation. In the following, we focus on a system with \( B = 3 \). The proposed scheme can be easily extended to the \( B > 3 \) case, e.g., multiple neighboring clusters can be grouped together to form a new bigger cluster.

**B.3 Inter-Cluster Interference Mitigation**

As shown in Section B.2, the intra-cluster interference can be completely eliminated by joint transmission as long as \( |S(c,n)| \leq B \). However, since the neighboring clusters are also using the same \( N \) subchannels, a cluster of coordinated sectors still causes inter-cluster interference to the users in the neighboring clusters, especially to the users in the cluster-edge area.

In this section, based on the idea of static fractional frequency reuse, two cooperative frequency reuse approaches are proposed to mitigate the inter-cluster interference. These frequency reuse approaches will be considered as the first step or inter-cluster level of the proposed resource allocation scheme.

**B.3.1 Cooperative Frequency Reuse Scheme 1**

From a cluster-specific point of view, users in each cluster can be divided into two classes, that is, cluster-edge users (CEU) and cluster-center users (CCU). In [14], the authors propose a user partition method based on user locations, and determine an inter-cluster coordination area by a predefined coordination distance. However, this distance parameter based user partition could be a hard decision for a real implementation. In addition, the effect of shadow fading on the users is ignored. In [28, 29], the user partition is instead based on the long term channel gain, which is more suitable for a practical use. Since the clusters are overlapping in [28, 29], the partition is performed from a cell-specific point of view, where users in each cell are divided into cell-center users and cell-edge users. In this paper, we propose a cluster-specific user partition approach based on the long term channel gain, which is defined as follows.

**Definition:** User partition threshold, \( \Delta l \), is the threshold used for classifying CEU and CCU. User \( k \) in sector \((c,b)\) estimates and feeds back to its serving sector the long term channel gains from its serving sector and from four candidate neighboring sectors, that
is, the two neighboring sectors within the same cluster $c$, and the other two neighboring sectors within the BS where its serving sector belongs to. In the example of Figure 1, the measurement set for the UE consists of its serving sector, the two neighboring sectors belonging to BS1 and the two neighboring sectors belonging to its coordinated cluster (the shadowed area). After obtaining these values, cluster $c$ finds out the weakest long term channel gain within the cluster (denoted by $l_{in}^k$ in dB) and the strongest long term channel gain from the two candidate neighboring sectors outside the cluster (denoted by $l_{out}^k$ in dB).

Note that $l_{in}^k$ reflects the weakest link within the cluster, which is the dominant link that affects the performance gain provided by intra-cluster zero-forcing joint transmission [32]. $l_{out}^k$ reflects the strongest interference link outside the cluster. If $l_{out}^k - l_{in}^k \geq \Delta l$, inter-cluster interference would compromise the intra-cluster joint transmission gain, i.e., inter-cluster interference would be a big challenge for user $k$. Hence, user $k$ is considered as a CEU if $l_{out}^k - l_{in}^k \geq \Delta l$; otherwise, it is regarded as a CCU. The threshold value can be predefined by each cluster or by the network (as employed in the handoff algorithm for practical wireless networks), and it can be a parameter to optimize according to the network design objective. Note that the measurements required from the users are based on the long term channel gain, which can be obtained from the ones used for the handoff process [33]. Hence, there is no measurements and feedback increase from the users by using this user partition method. One approach to further reduce the feedback would be to obtain $l_{in}^k$ and $l_{out}^k$ at the user side. Hence, these values or $l_{out}^k - l_{in}^k$ could be instead fed back for user partition.

The $N$ frequency subchannels are divided into two orthogonal sets, $G$ and $F$, where $F$ is further divided into three orthogonal subsets, marked by $f_i$, with $i = \{1, 2, 3\}$. Subchannels in set $G$ are used for CCU with frequency reuse factor of one for each cluster, while subchannels in set $F$ are used for CEU with frequency reuse factor of $1/3$ for each cluster. This cooperative frequency reuse scheme, named as CFR-1, is shown in Figure B.2. Note that neighboring clusters adopt orthogonal subchannels for the CEU. Hence, the inter-cluster interference can be significantly reduced. The frequency partition (the size of the cluster-center frequency set $G$) can be a parameter to optimize, and it is treated by system level simulation in Section B.5.
B.3.2 Cooperative Frequency Reuse Scheme 2

In CFR-1, the frequency reuse factor for CEU is $1/3$, that is, only one third of the sub-channels in set $F$ is available for CEU in each cluster. In this subsection, we propose a second cooperative frequency reuse scheme, named as CFR-2, where the frequency reuse factor for CEU is $2/3$ in each cluster.

Assume that every three neighboring clusters are grouped together and respectively marked as Cluster 1, Cluster 2 and Cluster 3 (see Figure B.3). Given the marker of each cluster (Cluster 1, 2 or 3), CEU in each cluster are further divided into two types according to their dominant interference clusters. The dominant interference cluster of user $k$ is defined as the one that the neighboring sector with the strongest long term channel gain ($l^\text{out}_k$) belongs to. The subchannels in set $F$ are further divided into three orthogonal subsets, marked by $f_i$, with $i = \{1, 2, 3\}$. Then, frequency subset $f_i$ is assigned for the CEU whose dominant interference cluster is cluster $i$. Based on the above defined frequency reuse rule, two subsets of $F$ are available for CEU within each cluster, where different type of CEU use the subchannels belonging to different subsets. As illustrated in Figure B.3, the subchannels used for CEU in neighboring clusters in CFR-2 are not orthogonal any more. However, for each CEU, the inter-cluster interference coming from its dominant interference cluster can be eliminated. In the example of Figure B.3, the subchannels belonging to subset $f_2$ can be used for CEU1, when the dominant interference cluster of CEU1 is cluster 2. Since $f_2$ is not available in cluster 2 according to the frequency reuse rule of CFR-2, there will be no inter-cluster interference introduced by cluster 2 for CEU1.

B.4 Joint Scheduling and Power Allocation

In the inter-cluster interference mitigation or first step, users in each cluster are divided into two groups (CEU/CCU) and mapped to different frequency sets. In this section, a utility-based joint scheduling and power allocation algorithm is proposed for each cluster to solve (12), which is considered as the intra-cluster level or the second step of the
proposed resource allocation scheme.

In realistic cellular systems, inter-cluster information exchange may not be feasible due to practical issues, e.g., synchronization constraints, feedback constraints, backhaul network constraints and the system complexity. Targeting practical scenarios, joint scheduling and power allocation is proposed to be independently performed in each cluster without inter-cluster communication. Assume that each cluster only has perfect CSI of the users inside this cluster, and the user data are shared by these sectors error-free and without delay. With joint transmission among the sectors within the cluster, the intra-cluster interference can be completely eliminated as long as \(|S(c,n)| \leq B\). Note that the inter-cluster interference is reduced by frequency reuse in the first step of the resource allocation scheme. In order to avoid inter-cluster information exchange and the interdependency issues among clusters, the remaining inter-cluster interference is not considered in this section in the SINR expression, i.e., the signal to noise ratio (SNR) of user \(k\) on subchannel \(n\) in cluster \(c\) based on the zero forcing joint transmission given in (4) is derived as \(p_{k,n}/\sigma^2\). Without loss of generality, we suppress the cluster index, and concentrate on one arbitrary cluster. \(b\) is used to represent \((c,b)\) to denote each of the sectors in cluster \(c\).

Equation (12) is a non-convex combinatorial optimization problem, thus, computing its globally optimal solution may not be feasible in practice. Here, we propose a suboptimal low complexity joint scheduling and power control algorithm for practical implementations. Assume \(P_{\text{max},b} = P_{\text{max}}\) for all sectors, and that the total transmit power \(P_{\text{max}}\) in each sector is equally pre-allocated to all the available subchannels. Then, the per-sector power constraints are reduced to per-subchannel power constraints with a constraint value \(P_{\text{max}}/N\), and the joint scheduling and power allocation problems are decoupled on each subchannel. On each subchannel \(n\), equal user power allocation [3] is adopted for the users scheduled in set \(S(n)\), that is, each selected user is allocated with the same symbol power given by

\[
p_{k,n} = p_n = \frac{P_{\text{max}}}{N \max_b \sum_{k \in S(n)} \|w_{k,n}^b\|^2}, \quad k \in S(n).
\]  

(14)

It should be pointed out that equal user power allocation [3] is suboptimal, since it typically results in only one sector meeting the maximum transmitted power pre-allocated to subchannel \(n\), and the remaining \(B-1\) sectors transmit below the \(P_{\text{max}}/N\) value. Actually, for each subchannel \(n\), the relaxed power allocation problem is a convex problem
for the scheduled users in set $S(n)$. Similar to reference [19], the optimal solution can be obtained based on waterfilling distribution via standard optimization techniques. However, as mentioned in [19], the computational complexity for obtaining the optimal value is still high. For simplicity, equal user power allocation is adopted in this paper.

Since we consider only SNR, i.e., $\gamma_{k,n} = p_{k,n}/\sigma^2$. Thus,

$$\gamma_{k,n} = \gamma_n = p_n/\sigma^2, \ k \in S(n).$$

(15)

Then, the sum utility of the user set $S(n)$ on subchannel $n$ can be calculated by

$$U_n(S(n)) = \sum_{k \in S(n)} \mu_k W \log_2(1 + \beta \gamma_n).$$

(16)

Hence, under the above assumptions, the sub-optimal solution becomes an exhaustive search. For each subchannel $n$, the coordinated sectors search all possible user sets $S(n)$ in the cluster. The chosen user set $S^*(n)$ will be the one that achieves the highest sum utility on subchannel $n$. Since the maximum number of users that can be supported on a subchannel is bounded by the total number of transmit antennas of the cluster, i.e., $|S(n)| \leq B$, the number of feasible selected user sets for each subchannel is $K^B$, with $K = K^c$ denotes the number of users in cluster $c$. Therefore, the complexity is $O(N \times K^B)$, which is prohibitively high. However, after the user partition and frequency partition at the first step shown in Section B.3, only a subset of users is mapped to each subchannel. Therefore, the number of the feasible user sets for exhaustive search on each subchannel is reduced. The complexity is then reduced to $O\left(\sum_{i=1}^3 |f_i| \times K_i^B + |G| \times K_g^B\right)$, where $K_i$ is the number of CEU that are mapped to the frequency subset $f_i$, and $K_g$ is the number of CCU that are mapped to the frequency subset $G$, with $\sum_{i=1}^3 |f_i| + |G| = N$ and $\sum_{i=1}^3 K_i + K_g = K$. The proposed two-step resource allocation algorithm is summarized in Algorithm 1.

---

**Algorithm 1** Proposed two-step resource allocation scheme

**Step 1 Inter-cluster interference mitigation**
1: In each cluster, divide the users into CCU and CEU.
2: Map CCU to the subchannels in frequency set $G$.
3: Map CEU to the subchannels in frequency subset $f_i$ according to the predefined cooperative frequency reuse scheme (CFR-1 or CFR-2).

**Step 2 Joint scheduling and power allocation**
1: For each subchannel $n$, find all the users mapped to it.
2: For each feasible user set $S(n)$ on subchannel $n$ ($|S(n)| \leq B$), derive the beamforming matrix by zero-forcing joint transmission.
3: Calculate the sum utility $U_n(S(n))$ of user set $S(n)$, based on (14), (15) and (16).
4: Find the optimal user set $S^*(n)$ with the maximum sum utility $U_n(S^*(n))$, and derive the corresponding transmit power $p^*_n$ for each user in $S^*(n)$ by (14).

---

**B.5 Simulation Results**

In this section, the performance of the proposed scheme is studied by system level simulations. We consider a network MIMO OFDMA system with 19 clusters. A wrap-around
B.5 Simulation Results

Table B.1: Performance comparison for different schemes

<table>
<thead>
<tr>
<th></th>
<th>IPC</th>
<th>UFR</th>
<th>CFR-2</th>
<th>CFR-1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fairness Jug Index (Mbps)</td>
<td>1.60</td>
<td>2.96</td>
<td>3.32</td>
<td>6.90</td>
</tr>
<tr>
<td>Complexity per cluster</td>
<td>984,200</td>
<td>984,150</td>
<td>33,891</td>
<td>67,271</td>
</tr>
<tr>
<td>Inter-cluster communication requirement</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
</tbody>
</table>

technique is adopted to avoid the boundary effect, which causes the clusters in the boundary of the cellular network to receive less interference. Each cluster consists of $B = 3$ neighboring sectors with one transmit antenna each. The antenna gain pattern measured in dB is given based on [1]

$$A(\varphi) = -\min \left( \frac{\varphi}{\varphi_{3dB}}, A_m \right),$$

where $\varphi$ is the angle that the user forms with the sector boresight $\varphi \in [-180^\circ, 180^\circ]$. $\varphi_{3dB} = 70^\circ$ is the angle associated with the half power beamwidth and $A_m = 20$ dB is the maximum attenuation for the sidelobe. The per-sector power constraint, $P_{max}$, is 43 dBm. The cell radius is 500 m. A typical urban multipath channel model [34] is used with path loss $L(d) = 128.1 + 37.6 \log_{10}(d)$ in dB, with $d$ in km. The number of subchannels, $N$, is 50, with each subchannel consists of 12 contiguous subcarriers. The subcarrier spacing is 15 kHz. $K = 27$ users are uniformly dropped in each cluster with one receive antenna for each user. A full-buffer traffic model is assumed for each user. The natural logarithm function is used as the users’ utility function, $U_k(\cdot)$. The throughput filter window length, $T_w$, is set to 100 time slots. The user partition threshold, $\Delta_l$, is set to -2 dB, which will be shown as a proper choice to balance the cell-edge and cell-average data rate. This value of $\Delta_l$ results in 16 CEU and 11 CCU in average for each cluster (over 100 different user locations). The average number of CCU and CEU per cluster ($K_i = 16, K_g = 11$) will be used for calculating the complexity of the proposed Scheme-1 and Scheme-2, as shown in Table B.1. Let Scheme-1 denote the proposed resource scheme adopting CFR-1 as the first step, while Scheme-2 denotes the proposed resource allocation scheme using CFR-2 as the first step. The simulation parameters are summarized in Table B.2.

B.5.1 Frequency Partition and User Partition

In this article, we consider the following performance metrics defined in 3GPP [1]:

- Throughput cumulative distribution function (CDF) or user average data rate CDF, which is the CDF of the average data rate including all the users in the system.
- Cell-average user data rate, which is the 50% point of the user average data rate CDF, denoted by $R_{ave}$.
- Cell-edge user data rate, which is the 5% point of the user average data rate CDF, denoted by $R_{edge}$.

First, we consider the effect of frequency partition on the performance of Scheme-1 and Scheme-2. Assume the set of subchannels used for CEU, denoted by $F$, is equally divided into three subsets, that is $|f_1| = |f_2| = |f_3| = |F|/3$, with $|F| + |G| = 50$. The subchannels in set $G$ are used for CCU.
Table B.2: Simulation parameters

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of clusters</td>
<td>19</td>
</tr>
<tr>
<td>Number of sectors per cluster, $B$</td>
<td>3</td>
</tr>
<tr>
<td>Number of transmit antennas per sector</td>
<td>1</td>
</tr>
<tr>
<td>Maximum transmit power per sector, $P_{\text{max}}$</td>
<td>43 dBm</td>
</tr>
<tr>
<td>Number of users per cluster, $K^c$</td>
<td>27</td>
</tr>
<tr>
<td>Number of receive antennas per user</td>
<td>1</td>
</tr>
<tr>
<td>Number of subchannels, $N$</td>
<td>50</td>
</tr>
<tr>
<td>Subchannel bandwidth</td>
<td>180 kHz</td>
</tr>
<tr>
<td>Cell radius</td>
<td>500 m</td>
</tr>
<tr>
<td>Multipath channel model</td>
<td>Typical Urban [34]</td>
</tr>
<tr>
<td>Pathloss (dB)</td>
<td>$L(d) = 128.1 + 37.6\log_{10}(d)$, $d$ in km</td>
</tr>
<tr>
<td>Traffic model</td>
<td>Full Buffer</td>
</tr>
<tr>
<td>User utility function, $U(\cdot)$</td>
<td>$\ln(\cdot)$</td>
</tr>
<tr>
<td>Throughput filter window length, $T_w$</td>
<td>100</td>
</tr>
<tr>
<td>User partition threshold, $\Delta l$</td>
<td>-2 dB</td>
</tr>
</tbody>
</table>

Considering different sizes of the cluster-center frequency set $G$, Figure B.4 shows the CDF of the user average data rate including all CEU in the system and the CDF curves including all CCU in the system for Scheme-1. While the corresponding CDF curves for Scheme-2 are plotted in Figure B.5. It can be seen that there is a tradeoff when choosing $|G|$. Actually, if $|G|$ is large, more subchannels will be allocated to CCU, leading to CCU’s performance increasing with CEU’s performance decreasing for both of these two schemes. In Scheme-1, the inter-cluster interference is significantly reduced for CEU due to the use of CFR-1. The cluster-center areas are still interference limited, since the frequency reuse factor is one for these areas. Hence, we can observe from Figure B.4 that CEU achieve higher date rates compared with CCU in Scheme-1 in most cases, e.g., when $|G| < 29$. In Scheme-2, the inter-cluster interference is handled by CFR-2, and CEU still suffer the interference from some neighboring clusters. Therefore, the CEU’s data rates are lower compared with that of CCU in Scheme-2, as shown in Figure B.5.

In order to have a further understanding of the two proposed schemes with respect to the system performance metrics $R_{\text{edge}}$ and $R_{\text{avg}}$, the user average data rate CDF curves including all users in the system with respect to different values of $|G|$ are plotted in Figure B.6 and Figure B.7 for Scheme-1 and Scheme-2, respectively. Notice that $R_{\text{edge}}$ is defined based on the user average data rate (the 5% point of the user average data rate CDF including all the users in the system) according to 3GPP [1], instead of the user category (CEU). Hence, $R_{\text{edge}}$ is not equivalent to the average date rate of CEU. For example, as shown in Figure B.4, the users with lower average data rate actually come from CCU in Scheme-1. While for Scheme-2, the users with lower average data rate come from CEU as shown in Figure B.5. The average data rate of CCU increases as the number of frequency bands allocated to the CCU ($|G|$) increases. Hence, it is reasonable to see from Figure B.6 and Figure B.7 that $R_{\text{edge}}$ increases as $|G|$ increases in Scheme-1, while $R_{\text{edge}}$ decreases as $|G|$ increases in Scheme-2.

In Figure B.6, there are plateau regions at CDF values of 0.4 and 0.93. The plateau
Figure B.4: CDF of user average data rate in Scheme-1

Figure B.5: CDF of user average data rate in Scheme-2
Figure B.6: CDF of user average data rate in Scheme-1

Figure B.7: CDF of user average data rate in Scheme-2
region at 0.4 is actually the point separating the CCU and CEU in the Scheme-1, while the plateau region at the CDF value around 0.93 comes from the CEU (see Figure B.4). Based on the user partition method proposed in this paper, the CEU group consists of both the users that are close to the neighboring clusters and have poor channel qualities (sector-edge users), and the users that are close to their serving sector and have good channel qualities (sector-center users). The CDF value around 0.93 is actually the point separating the sector-edge users and the sector-center users within the CEU group.

The above results are obtained by setting the user partition threshold, $\Delta l$, to be -2 dB. Besides of frequency partition, the user partition would also affect the value of $R_{\text{edge}}$ and $R_{\text{ave}}$. In order to balance the cell-edge and cell-average performance, a utility function is defined in [14] to evaluate the effect of coordination distance on both $R_{\text{edge}}$ and $R_{\text{ave}}$. In this paper, we map the distance parameter to the size of the cluster-center frequency set $|G|$ and the user partition threshold $\Delta l$. Then, a fairness jug function with respect to the value of $|G|$ and $\Delta l$ is defined as

$$J (|G|, \Delta l) = \alpha R_{\text{edge}} (|G|, \Delta l) + (1 - \alpha) R_{\text{ave}} (|G|, \Delta l),$$

where $\alpha \in [0,1]$ is a fairness factor reflecting the design objective. When the aim is to improve the cell-edge date rate, we choose a larger value of $\alpha$. If the objective is the sum rate, a smaller value of $\alpha$ is picked. As an example, $\alpha = 2/3$ is selected in our simulation, which means we target the cell-edge performance. Recall that a user $k$ is considered as a CEU if $l_{\text{out}}^k - l_{\text{in}}^k \geq \Delta l$; otherwise, it is a CCU. Therefore, the number of CEU decreases as the value of $\Delta l$ increases. For example, in our simulations, the percentage of CEU is around 22.2% with $\Delta l = -6$ dB, and 66.7% with $\Delta l = 6$ dB.

The values of Fairness Jug Index with respect to different $|G|$ and $\Delta l$ are shown in Figure B.8 and Figure B.9 for Scheme-1 and Scheme-2, respectively. For each scheme, the user average data rate CDF including all users in the system is first plotted with respect to each pair of $(|G|, \Delta l)$. Then, $R_{\text{edge}} (|G|, \Delta l)$ and $R_{\text{ave}} (|G|, \Delta l)$ can be obtained from the corresponding CDF curve, according to the definition in 3GPP. The Fairness Jug Index, $J (|G|, \Delta l)$, is finally derived by substituting $R_{\text{edge}} (|G|, \Delta l)$ and $R_{\text{ave}} (|G|, \Delta l)$ into (18). It can be seen that the maximum values for both Scheme-1 and Scheme-2 are achieved when $\Delta l = -2$ dB. While the optimal frequency partition is achieved with $|G| = 29$ and $|F| = 7$ for Scheme-1, $|G| = 17$ and $|F| = 11$ for Scheme-2. Compared with Scheme-1, Scheme-2 needs more subchannels for CEU. As has been mentioned above, Scheme-1 performs CFR-1 for inter-cluster interference mitigation, where neighboring clusters use orthogonal subchannels for CEU. Hence, the inter-cluster interference is significantly reduced in Scheme-1, leading to a large performance improvement for CEU (see Figure B.4). In Scheme-2, CEU still suffer the interference from some neighboring clusters. Hence, more cluster-edge subchannels are needed in Scheme-2 to increase the user data rate of CEU (see Figure B.5).

### B.5.2 Performance Analysis

In this simulation, based on the results from Figure B.8 and Figure B.9, we choose $|G| = 29$ for Scheme-1 and $|G| = 17$ for Scheme-2. $\Delta l$ is set to be -2 dB. Besides the proposed Scheme-1 and Scheme-2, the following two schemes are considered as reference schemes for performance comparison.
Figure B.8: Fairness jug Index in Scheme-1 with different $|G|$ and $\Delta l$

Figure B.9: Fairness jug Index in Scheme-2 with different $|G|$ and $\Delta l$
• Universal frequency reuse (UFR), where all subchannels are available for each cluster and the proposed joint scheduling and power allocation is adopted independently within each cluster irrespective of the user category (CEU/CCU).

• Inter-cluster interference pre-cancellation (IPC), where each cluster helps the neighboring CEU by taking these users into account when designing precoding matrices. In this scheme, the two-step joint scheduling approach described in [14] is adopted as follows.

  – In a first step (intra-cluster level), each cluster performs joint scheduling and power allocation within its own cluster (the same approach as in the UFR scheme).

  – In a second step (inter-cluster level), IPC is independently performed on each subchannel: CEU scheduled on the corresponding subchannel inform the neighboring helper clusters. Then, each cluster deals with the requests from CEU in the neighboring clusters in a sequential way, and it selects to always help those CEU by randomly dropping some of its own users scheduled on the same subchannel. After this re-scheduling process, each cluster redesigns the transmit power for the scheduled users using (14).

In Figure B.10, the CDF of the average SINR of all users in the system is evaluated for the two proposed schemes. For each user, the SINR is calculated per trial, and then averaged over 100 trials. In each trial, the SINR of each user is obtained by averaging the SINR values over all subchannels that are assigned to it, considering the remaining interference power from one ring of 6 neighboring clusters. Note that the wrap-around technique is adopted to avoid the boundary effect for the users in the boundary clusters (the clusters in the boundary of the cellular deployment). Compared with the UFR Scheme, it is shown that the average SINR performance of the proposed schemes is significantly improved due to the inter-cluster interference reduction. Similar to the analysis for Figure B.6, the plateau region at 0.4 for Scheme-1 in this figure is actually the point separating the SINR values of CCU and CEU. Due to the use of CFR-1 in Scheme-1, the inter-cluster interference reduction of CEU is quite significant, resulting in a better SINR performance for CEU compared with that of CCU in Scheme-1.

Figure B.11 and Figure B.12 show the CDF of the user average data rate and the CDF of the user average utility for the considered resource allocation schemes. Table B.1 gives the performance comparison of the different schemes in terms of Fairness Jug Index, computational complexity and inter-cluster communication requirements. We can see that:

• Although inter-cluster interference is treated in the IPC scheme, its performance is even worse than the UFR scheme, which does not take inter-cluster interference into account. The poor performance of the IPC scheme was caused by the following two main reasons.

  – Limited degrees of freedom for each cluster to help the neighboring CEU while serving its own users. In the IPC scheme, in order to serve a neighboring CEU by interference pre-cancellation, the cluster needs to provide a certain number of degrees of freedom by dropping some scheduled users of its own, which leaves
Figure B.10: CDF of user average SINR

Figure B.11: CDF of user average data rate
fewer degrees of freedom to serve its own users. As explained in lemma 2 of [14], for a cluster with $B$ coordinated sectors and $k_e$ neighboring CEU to help, the maximum number of users that can be supported simultaneously by joint transmission in this cluster is bounded by $K_{max}^h \leq \lfloor (BN_t) / N_r \rfloor - k_e$, where $N_t$ denotes the number of transmit antennas per sector and $N_r$ denotes the number of receive antennas per user. Therefore, to help a neighboring CEU, the total number of users that the IPC scheme can support is reduced. In our system model, $N_t = N_r = 1$. Hence, the maximum number of users that can be served within a cluster is determined by the cluster size $B$. With a larger $B$, there would be spare degrees of freedom left for each cluster to help neighboring CEU. However, note that due to path loss, the CEU do not benefit from far away sectors’ transmission. As shown in [14], there is a diminishing gain with the increase of the cluster size. Therefore, the performance of Scheme-1 and Scheme-2 with respect to the IPC scheme might improve for a larger $B$.

- A smart global scheduler is required for jointly scheduling users across multiple clusters. Note that the two-step joint scheduling approach in [14] aggressively protects the neighboring CEU by randomly dropping some scheduled users of its own, which leaves fewer degrees of freedom to serve its own users. Since the dropped users in each cluster are randomly picked out, some of its own scheduled CEU might also be dropped out in step 2, resulting in the performance degradation of CEU. In addition, for each cluster, the neighboring CEU come from six neighboring clusters. Hence, the chance that a cluster receives requests from neighboring CEU is very high. In an extreme case for a subchannel, where the number of scheduled neighboring CEU of a cluster is large (larger or equal to $B$), the two-step joint scheduling strategy would force the cluster to drop all its scheduled users on this subchannel to aggressively protect the neighboring CEU, leading to both CEU and CCU performance degradation. Hence, in order to improve the performance of IPC scheme, a smart global scheduler is
required for joint user scheduling across multiple clusters. However, this global optimization requires inter-cluster communication and increases the complexity of the resource allocation design.

- In Figure B.11, compared with UFR, the cell-edge user data rate of the proposed Scheme-2 is improved by 20%, while the cell-average user data rate in the Scheme-2 is improved by 5%. The proposed Scheme-1 achieves a much more significant performance improvement compared to the UFR scheme, with about 180% increase of cell-edge user data rate and 90% increase of cell-average user data rate. Note that although the average data rate and the average utility performance of Scheme-2 is close to that of UFR scheme, the complexity of Scheme-2 is much lower compared with UFR scheme, which can be observed in Table B.1.

- As shown in Table B.1, the computational complexity of the proposed two schemes is much lower than the UFR scheme and the IPC scheme, with Scheme-2 achieving the lowest complexity. Since the inter-cluster interference mitigation approaches (CFR-1 and CFR-2) are adopted in the first step of Scheme-1 and Scheme-2, only a subset of users is mapped to each subchannel. Hence, the number of feasible user sets for exhaustive search in the second step for Scheme-1 and Scheme-2 is significantly reduced. With $\Delta I = -2$ dB, the average numbers of CEU and CCU per cluster are $K_i = 16$ and $K_g = 11$ respectively in our simulation. As explained in Section B.4, the complexity per cluster for Scheme-1 and Scheme-2 is $O \left( \sum_{i=1}^{3} |f_i| \times K_i^B + |G| \times K_g^B \right)$. The number of the feasible selected user sets for each subchannel in the UFR scheme is $K^B$. Therefore, the complexity for the UFR scheme is $N \times K^B$. In the IPC scheme, each cluster performs joint scheduling and power allocation within its own cluster based on UFR in a first step, then each cluster needs to perform one more time user selection according the requests from CEU in the neighboring clusters. Hence, the complexity for IPC is $N \times (K^B + 1)$. Note that the complexity of the four schemes considered in this paper increases exponentially with the cluster size $B$.

### B.6 Conclusions

The resource allocation problem has been considered for the downlink of a clustered network MIMO OFDMA system. A two-step resource allocation scheme with inter-cluster interference mitigation and intra-cluster joint scheduling and power allocation has been presented. In particular, the main task of managing the inter-cluster interference is accomplished by two cooperative frequency reuse approaches at the first step of the proposed resource allocation scheme. A user partition method based on the long term channel gain is introduced to divide the users of each cluster into cluster-edge and cluster-center users. Frequency subchannels in each cluster are separated into cluster-edge and cluster-center frequency sets. The inter-cluster interference is reduced by mapping different groups of cluster-edge users to different subchannels of the cluster-edge frequency set in a cooperative way. We have shown that there is a tradeoff between the cell-edge and cell-average performance while choosing the frequency partition and the user partition, i.e., the size of the cluster-center frequency set and the user partition threshold. As the second step, a sub-optimal utility-based joint scheduling and power allocation algorithm is proposed for
each cluster as the intra-cluster level of resource allocation. The objective is to maximize the sum utility of all users within the cluster under per-sector power constraints. Note that in realistic scenarios, the two-step approach can be implemented in different time scales, i.e., the inter-cluster interference mitigation would be more static than the intra-cluster scheduling and power allocation performed in the second step. It has been demonstrated by simulation results that our proposed resource allocation scheme can provide a considerable performance improvement in terms of the cell-edge user data rate, the cell-average user data rate and the user average utility. In addition, the proposed two-step resource allocation scheme can be implemented independently in each cluster without inter-cluster communication, which is an attractive property for practical systems, since it reduces both the network signaling overhead and the computational complexity.

In this article, we have assumed that each sector has one transmit antenna and each user has one receive antenna. Zero-forcing beamforming is used as the network MIMO joint transmission scheme. In future work, multiple antennas at both the sector side and the user side will be investigated. Frequency reuse combined with interference precancellation techniques will be studied for managing the inter-cluster interference. More efficient joint scheduling and power control algorithms with advanced multi-antenna joint transmission methods will be considered. The results in this article assume that perfect CSI is available at the cluster side for the users within the cluster. Investigation of imperfect CSI is of practical importance and it is a topic of our future work.

References


[34] 3GPP TR 36.942 V9.1.0, Evolved Universal Terrestrial Radio Access (E-UTRA); Radio Frequency (RF) system scenarios, 2010
Performance Evaluation of Coordinated Multi-Point Transmission Schemes with Predicted CSI

Jingya Li, Agisilaos Papadogiannis, Rikke Apelfröd, Tommy Svensson, and Mikael Sternad

Published in Proceedings of IEEE Personal, Indoor, and Mobile Radio Communications (PIMRC’12)
pp. 1-6.
Sydney, Australia, Sept. 2012
©2012 IEEE
The layout has been revised.
Typographical adjustments have been made.
Abstract

Coordinated multi-point (CoMP) transmission is considered as an efficient technique to improve cell-edge performance as well as system spectrum efficiency. In CoMP-enabled systems, a cluster of coordinated base stations (BSs) are typically assumed to be connected to a control unit (CU) via backhaul links, and the provided performance gain relies heavily on the quality of the channel state information (CSI) available at the CU side. In this paper, we consider the downlink of a CoMP cluster and compare three different CoMP transmission schemes: zero-forcing coherent joint transmission, non-coherent joint transmission and coordinated scheduling. Moreover, for each of the analyzed schemes, the performance in terms of average sum rate of the CoMP cluster is studied with predicted CSI, considering the effects of the feedback and backhaul latency, as well as the user mobility. Compared to zero-forcing coherent joint transmission, we show that non-coherent joint transmission and coordinated scheduling are more robust to channel uncertainty. In addition, depending on the latency, user mobility and user locations, different schemes would achieve the highest average sum rate performance. Hence, a system could switch between the transmission schemes to improve the sum rate.
C.1 Introduction

Recently, Coordinated multi-point (CoMP) transmission has been considered as a promising technique to mitigate inter-cell interference (ICI) and improve spectrum efficiency in wireless communication systems [1]. CoMP transmission techniques can be divided into two main categories [2]:

- Joint transmission, where data to a single user is simultaneously transmitted from multiple BSs. The ICI is then reduced by using the signals transmitted from other BSs to assist the transmission instead of acting as interference.
- Coordinated scheduling and/or coordinated beamforming, where data to a user is transmitted from one BS. However, scheduling and beamforming are coordinated to control ICI.

In CoMP-enabled systems, a cluster of coordinated BSs are typically assumed to be connected to a control unit (CU) via backhaul links [3]-[6]. In frequency division duplex (FDD) systems, each user within the cluster needs to estimate and predict the channel state information (CSI) from all coordinated BSs, and then to feed it back to its serving BS. In a second step, each coordinated BS forwards this information via backhaul links to the CU. Based on the available predicted CSI, the CU designs the joint transmission and/or user scheduling and/or coordinated beamforming scheme. It then forwards these decisions via backhaul links to each coordinated BS. For the joint transmission approach, the user data also needs to be shared by all coordinated BSs or by a subset of BSs through backhaul links.

Depending on the restrictions of the feedback and backhaul links and the amount of information to be shared among coordinated BSs, the entire transmission loop within the CoMP cluster would introduce different degrees of latency, resulting in outdated CSI measurements [6]. In addition, the quality of the predicted CSI would also be affected by the mobility of the served users [7]. Note that the design of CoMP transmission schemes relies heavily on the quality of the CSI available at the CU.

In this paper, three CoMP transmission schemes that require different degrees of BS coordination are compared:

- Coherent joint transmission. In this approach, data symbols of all users within the CoMP cluster are available for all coordinated BSs. A linear precoding based on zero-forcing is performed for mapping the data symbols of all users to the transmit antenna of each BS [8].
- Non-coherent joint transmission. The BSs within the CoMP cluster are divided into user-specific cooperative BS sets. The data symbol of each user is non-coherently transmitted from a subset of BSs, i.e., its cooperative BS set, without joint phase adjustment [9].
- Coordinated scheduling. In this scheme, data to a single user is transmitted from its serving BS. However, scheduling decisions are jointly made at the CU to control ICI [10].

The performance in terms of sum rate is here evaluated for each CoMP transmission scheme. As a baseline, the performance of a traditional single cell transmission scheme without BS coordination is also given. We show that depending on the feedback and backhaul latency, user mobility, as well as user location, a system could switch between the transmission schemes to improve the sum rate.

Notation: Here, $(\cdot)^H$, $(\cdot)^T$ and $(\cdot)^{-1}$ denote the conjugate transpose, transpose and matrix inversion operations, respectively. The notation $1_{[m \times n]}$ and $0_{[m \times n]}$ represent the
matrix with \( m \) rows and \( n \) columns filled with ones and zeros, respectively. \(|\mathcal{M}|\) denotes the cardinality of the set \( \mathcal{M} \).

## C.2 System Model

We consider the downlink of a CoMP cluster, where \( N \) single-antenna BSs are connected via backhaul links to a CU. \( M \) single-antenna users are grouped together using a particular resource slot, e.g., a time slot or a subchannel. The \( N \) BSs are assumed to have the same maximum power constraint \( P_{\text{max}} \) and to share the same resource slot. Let \( \mathbf{x} = [x_1, \ldots, x_N]^T \) denote the signal vector transmitted from all \( N \) BSs, with \( x_n^H x_n \leq P_{\text{max}} \) for all \( n \in \{1, \ldots, N\} \). The received signal at user \( m \in \{1, \ldots, M\} \) can then be expressed as

\[
y_m = \mathbf{h}_m^H \mathbf{x} + n_m, \tag{1}
\]

where \( \mathbf{h}_m = [h_{m1}, \ldots, h_{mN}] \) denotes the channel vector between user \( m \) and all \( N \) BSs. Above, \( n_m \) is the sum of the thermal noise and the uncoordinated out-of-cluster interference, modeled as independent complex additive Gaussian noise with zero mean and covariance \( \sigma^2 \).

We assume that the system works in FDD mode. Each user \( m \) needs to predict the channel vector \( \mathbf{h}_m \), and feed back the predicted channel vector \( \hat{\mathbf{h}}_m \) to the CU via uplink control channels. Hence, the predicted channel matrix of the system available at the CU is

\[
\hat{\mathbf{H}} = [\hat{\mathbf{h}}_1^T, \ldots, \hat{\mathbf{h}}_M^T]^T \in \mathbb{C}^{M \times N},
\]

which will be used for the CoMP transmission scheme design.

In this paper, similar to [11], we assume that the predicted channel vector \( \hat{\mathbf{h}}_m(t|t-\Delta t) \) of each user \( m \) at time slot \( t \) is predicted by Kalman predictors using pilot measurements available up to time slot \( t - \Delta t \). The prediction horizon \( \Delta t \) corresponds to the delay between the channel observation and the data transmission, including the implementation of user scheduling and precoding, and the feedback and backhaul latency. For optimal filters, such as the Kalman predictor, the prediction error \( \Delta \mathbf{h}_m(t) \) is uncorrelated with the prediction. This is achieved in the simulations by first modeling the predicted channel and the prediction error as i.i.d. complex circular symmetric Gaussian variables with variances \( \sigma^2_h \) and \( \sigma^2_{\Delta h} \) respectively. Second, the true channel is calculated as

\[
\mathbf{h}_m(t) = \hat{\mathbf{h}}_m(t|t-\Delta t) + \Delta \mathbf{h}_m(t). \tag{2}
\]

Given a variance of the true channel \( \sigma^2_h \), the variance of the channel prediction error, \( \sigma^2_{\Delta h} \), can be extracted from the covariance matrix of (19) in the appendix. Then, \( \sigma^2_h \) can be found from \( \sigma^2_h = E[\mathbf{h}_m \mathbf{h}_m^H] = \sigma^2_h + \sigma^2_{\Delta h} \).

As can be seen in the appendix, the prediction performance depends on the pilot SNR (through (18)), on the prediction horizon (through (17)) and on the fading statistics including the shape of the Doppler spectrum and the maximum Doppler frequency, i.e. the user velocity \( v \) (through the poles of (16)). These factors were thoroughly investigated in [12]. In this paper, we assume a flat Doppler spectrum (as the one in figure 6.8 of [12]), a carrier frequency of \( f_c = 2 \text{ GHz} \), and a fading channel modeled as a fourth order Auto Regressive (AR) model. For \( \rho = 4 \) adjacent flat fading pilot bearing subcarriers, with a spacing in time of 0.64 ms, that are predicted jointly, the calculated prediction performance is then presented in Figure C.1. Here the prediction performance is given in terms of Normalized Mean Squared Error (NMSE) \( \frac{\sigma^2_{\Delta h}}{\sigma^2_h} \), for different user velocities and pilot SNR, as a function of the prediction horizon. Clearly, an increased velocity leads to a decreased predictability.
C.3 CoMP Transmission Schemes

With the predicted channel matrix $\mathbf{\hat{H}}$, three different CoMP transmission schemes are considered within a CoMP cluster. The objective is to maximize the sum rate of the cluster, under per-BS power constraints.

### C.3.1 Coherent Joint Transmission

Assume that the data symbols of all the $M$ users within the cluster are shared among the $N$ coordinated BSs. A linear precoding approach, zero-forcing, is considered as the coherent joint transmission scheme in this section. Note that with linear precoding among $N$ single-antenna BSs, at most $N$ single-antenna users can be served on the same resource slot without inter-user interference.

Let $\mathcal{M}$ denote the set of scheduled users in a given resource slot, with $\mathcal{M} \subseteq \{1, \ldots, M\}$ and $|\mathcal{M}| \leq N$. Let $\mathbf{b} \in \mathbb{C}^{\lvert \mathcal{M} \rvert}$ denote the data symbols of the selected users in set $\mathcal{M}$. A precoding matrix $\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_{\lvert \mathcal{M} \rvert}] \in \mathbb{C}^{N \times \lvert \mathcal{M} \rvert}$ is designed for mapping the data symbol vector $\mathbf{b}$ into the transmit signal vector $\mathbf{x}$, that is,

$$\mathbf{x} = \mathbf{Wb}. \quad (3)$$

The $m^{th}$ column of $\mathbf{W}$, $\mathbf{w}_m = [w_{1m}, \ldots, w_{Nm}]^T$, is the precoding vector for user $m$ in the set $\mathcal{M}$. Substituting (3) into (1), the received signal of user $m$ can be rewritten as

$$y_m = \mathbf{h}_m \mathbf{w}_m b_m + \sum_{i \in \mathcal{M}, i \neq m} \mathbf{h}_m \mathbf{w}_i b_i + n_m. \quad (4)$$

Let $p_m = b_m b_m^H$ denote the symbol power allocated to user $m$ across the $N$ BSs. The true signal to interference plus noise ratio (SINR) of user $m$ is then given by

$$\gamma_m = \frac{\|\mathbf{h}_m \mathbf{w}_m\|^2 p_m}{\sum_{i \in \mathcal{M}, i \neq m} \|\mathbf{h}_m \mathbf{w}_i\|^2 p_i + \sigma^2}. \quad (5)$$
Thus, the true sum rate of the cluster can be expressed as
\[ C = \sum_{m \in \mathcal{M}} \log_2(1 + \gamma_m). \] (6)

Let \( \hat{H}(\mathcal{M}) \in \mathbb{C}^{\lvert \mathcal{M} \rvert \times N} \) denote the predicted channel submatrix related to the set of scheduled users. Using zero-forcing precoding, the precoding matrix is obtained as the pseudo-inverse of the predicted channel matrix,
\[ W = \hat{H}(\mathcal{M})^H (\hat{H}(\mathcal{M})\hat{H}(\mathcal{M})^H)^{-1}. \] (7)

Based on \( \hat{H} \), the CU needs to design the scheduled user set \( \mathcal{M} \) and the power allocation vector \( p = [p_1, \ldots, p_{\lvert \mathcal{M} \rvert}] \), so as to maximize the sum rate under per-BS power constraints. The optimization problem for the CU can be formulated as
\[
\begin{align*}
\max_{\mathcal{M}, p} & \sum_{m \in \mathcal{M}} \log_2(1 + \hat{\gamma}_m) \\
\text{s.t.} & 1) \sum_{m \in \mathcal{M}} \|w_{nm}\|^2 p_m \leq P_{\text{max}}, n \in \{1, \ldots, N\}, \\
2) & p_m > 0, m \in \mathcal{M}, \\
3) & \mathcal{M} \subseteq \{1, \ldots, M\}, \lvert \mathcal{M} \rvert \leq N.
\end{align*}
\] (8)

Note that only \( \hat{H} \) is available at the CU. Hence, joint user scheduling and power allocation is designed based on the predicted SINR \( \hat{\gamma}_m = p_m/\sigma^2 \), which is derived from (5) by using \( \hat{h}_m \) instead of the true channel vector \( h_m \).

If the scheduled user set is predetermined and feasible, i.e. if \( \mathcal{M} \) is fixed and satisfies the constraint 3), then, the above problem becomes a joint power allocation problem. This problem is convex, since the objective function is a concave function of \( p \) and the remaining constraints 1) and 2) are linear. Therefore, the optimal solution with respect to a given \( \mathcal{M} \) can be obtained by numerical convex optimization (CVX) [13]. By solving the joint power allocation problem for every possible user set, the optimal \( \mathcal{M}^* \) and \( p^* \) can then be obtained. The true sum rate can then be derived by substituting \( \mathcal{M}^* \) and \( p^* \) into (6). In the following, this zero-forcing joint transmission scheme with optimal power allocation is denoted as ZF-OPA.

In order to reduce the complexity, a sub-optimal equal power allocation is considered [8]. In this case, for any given user set, \( \mathcal{M} \), the power allocation vector is derived as
\[ p = \left\{ \frac{P_{\text{max}}}{\sum_{m \in \mathcal{M}} \|w_{nm}\|^2} \right\} \mathbf{1}_{\lvert \mathcal{M} \rvert \times 1}. \] (9)

We refer this zero-forcing joint transmission scheme with equal power allocation as ZF-EPA in this paper.

### C.3.2 Non-coherent Joint Transmission

In this scheme, joint transmission is non-coherently performed without phase adjustment. Hence, this non-coherent joint transmission scheme might be more robust to channel uncertainty than coherent joint transmission.

Let \( S = [s_{nm}] \) denote a user selection indicator matrix of size \( N \times M \). If BS \( n \) transmits data to user \( m \), \( s_{nm} = 1 \); otherwise, \( s_{nm} = 0 \). Assume that a BS transmits data non-coherently to at most one user in any given resource slot. Then, at most one single element
in each row of $S$ is non-zero. Hence, the $N$ BSs within a cluster are grouped into several subclusters, forming a group of user-specific cooperative BS sets (CBS). Denote $CBS_m$ as the CBS of user $m$, with $CBS_m = \{ n | s_{nm} = 1, \forall n \in \{1, ..., N\} \}$ consisting of the BSs that provide data transmission to user $m$. Note that user $m$ only receives its data from the BSs included in $CBS_m$. Hence, the amount of user data that needs to be exchanged via backhaul links between BSs is reduced. The data symbol of user $m$ is transmitted non-coherently from the BSs in $CBS_m$ without phase adjustment, i.e., $x_n = l_n b_m$ for $\forall n \in CBS_m$, where $l_n \in \mathbb{R}$. Therefore, ICI cannot be mitigated by cancellation. The received signal of user $m$ is given by

$$y_m = \sum_{i \in CBS_m} h_{mi} x_i + \sum_{j \in CBS_m} h_{mj} x_j + n_m, \tag{10}$$

where $\overline{CBS_m}$ is the complement set of $CBS_m$. Denote $P_n = x_n^H x_n$ as the transmit power of BS $n$. The true SINR for user $m$ is given as

$$\gamma_m = \frac{\| \sum_{i \in CBS_m} h_{mi} \sqrt{P_i} \|^2}{\| \sum_{j \in CBS_m} h_{mj} \sqrt{P_j} \|^2 + \sigma^2}. \tag{11}$$

Thus, the true sum rate can be calculated by

$$C = \sum_{m=1}^M \log_2(1 + \gamma_m). \tag{12}$$

With the objective of maximizing the sum rate of the cluster, the CU needs to design the user selection indicator matrix $S$ and the power allocation vector $P = [P_1, ..., P_n]$ based on the predicted channel matrix $\hat{H}$. The optimization problem under per-BS power constraints can be formulated as

$$\max_{S, P} \sum_{m=1}^M \log_2(1 + \hat{\gamma}_m) \tag{13}$$

s.t. 1) $0_{[N \times 1]} \preceq P \preceq P_{\text{max}} 1_{[N \times 1]}$, 2) $s_{nm} \in \{0, 1\}$,

3) $\sum_{m=1}^M s_{nm} \leq 1, n \in \{1, ..., N\}$,

where $\hat{\gamma}_m$ is derived from (11) by using the predicted $\hat{h}_m$ instead of the true channel vector $h_m$.

The optimization problem (13) is a non-convex problem. Based on [9], a suboptimal binary power control (BPC) is considered for power allocation, i.e., $P_n = 0$ or $P_{\text{max}}$ for $\forall n \in \{1, ..., N\}$. Then, the relaxed problem becomes an exhaustive binary search. The CU searches all the possible values of the user selection indicator matrix $S$ and all feasible boundary point sets for binary power control. The chosen matrix $S^*$ and transmit power vector $P^*$ will be the ones that achieve the highest $\sum_{m=1}^M \log_2(1 + \hat{\gamma}_m)$. The corresponding true sum rate can then be obtained by substituting $S^*$ and $P^*$ into (12). In this paper, the non-coherent joint transmission scheme with BPC is named as NCJT-BPC.

### C.3.3 Coordinated Scheduling

In the considered coordinated scheduling scheme, data to a single user is only transmitted from its serving BS, which is selected based on the long term channel quality measurements, including pathloss and shadow fading. Hence, user data exchange between BSs
is not needed. Similar to NCJT-BPC, it is assumed that a BS can transmit data to at most one user in any given resource slot. The received signal of the selected user $m$ to be served by BS $n$ can be expressed as

$$y_m = h_{mn}x_n + \sum_{j \neq n} h_{mj}x_j + n_m.$$  

(14)

Recall that $P_n = x_n^H x_n$ denotes the transmit power of BS $n$, with $P_n \leq P_{\max}$. Then, the true SINR for user $m$ is given as

$$\gamma_m = \frac{\|h_{mn}\|^2 P_n}{\sum_{j \neq n} \|h_{mj}\|^2 P_j + \sigma^2}.$$  

(15)

Thus, the true sum rate can be calculated by (12).

User scheduling and power allocation decisions are jointly made at the CU to control ICI. With the predicted channel matrix $\hat{\mathbf{H}}$, the CU designs the user selection indicator matrix $\mathbf{S}$ and the power allocation vector $\mathbf{P} = [P_1, \ldots, P_n]$, in order to maximize the sum rate subject to per-BS power constraints. The optimization problem can be formulated similar to (13). However, the predicted SINR ($\hat{\gamma}_m$) is instead derived from (15) with the predicted $\hat{h}_{mn}$. Binary power control, which is shown to be a very efficient suboptimal power allocation solution [10], is performed in this scheme. Then, similar to NCJT-BPC, the suboptimal $\mathbf{S}$ and $\mathbf{P}$ can be derived by an exhaustive binary search. In this paper, the coordinated scheduling scheme with binary power control is named as CS-BPC.

C.4 Simulation Results

As depicted in Figure C.2, we consider the downlink of a CoMP cluster with $N = 3$ neighboring sectors. $M = 3$ single-antenna users are grouped together using a particular resource slot\(^1\). The cluster radius $R$ is 500 m. The path loss model is $PL(d) = 128.1 + 37.6 \log_{10}(d)$ in dB, with $d$ given in km. Long-term shadowing is log-normally distributed with zero mean and standard deviation 8 dB. The system SNR is set to 18 dB, which is defined as the received SNR at the boundary of the cell, assuming full power transmission $P_{\max}$ from the BS, accounting only for pathloss gain $PL(R)$ and ignoring shadowing and fast fading [14].

Assume that sector $n$ is the serving sector of user $n$, with $n = \{1, 2, 3\}$. Each user is moving from the cluster center to the sector center of its serving sector along the dashed line of Figure C.2. The performance in terms of cluster sum rate is studied for different CoMP transmission schemes (ZF-OPA, ZF-EPA, NCJT-BPC, CS-BPC), with respect to different sets of user starting locations. For each set of user starting locations, the sum rate is averaged over 1000 independent shadow fading realizations. Single cell transmission without BS coordination, denoted as SC, is used as baseline.

C.4.1 Sum Rate Performance with Perfect CSI

Let $d$ be the distance between a user and the center of its serving sector as shown in Figure C.2. Assume that perfect CSI is available at CU. In Figure C.3, the average sum rate of

\(^1\)Note that there is no constraint on the number of users within the cluster for all the CoMP transmission schemes considered in this paper. However, based on the system model, $M=3$ is already a full load scenario when focusing on one resource slot. Adding more users will just provide multi-user diversity gain for all schemes.
Figure C.2: A CoMP cluster of 3 neighboring sectors (the shadowed area).

each transmission scheme is plotted versus the normalized distance \((d/R)\). Compared with the SC scheme, the considered CoMP transmission schemes provide a significant average sum rate gain, especially for the users located at the cluster center areas or cell-edge areas (the users with large values of \(d\)).

Note that ZF-OPA and ZF-EPA achieve superior performance compared to the NCJT-BPC scheme. That is because, with perfect CSI at the CU, zero-forcing precoding performed in ZF-OPA and ZF-EPA can completely remove the ICI for all users within the CoMP cluster. In the NCJT-BPC scheme, a single user receives data symbols from a subset of BSs, hence, the BSs outside its cooperative BS set \((CBS_m)\) would still introduce ICI. In addition, the data symbols of each user are transmitted without phase adjustment, which would also result in performance degradation.

The CS-BPC scheme has the worst performance among the considered CoMP transmission schemes, as multi-BS joint transmission is not supported \(^2\). However, ICI is controlled in the CS-BPC scheme by coordinating the user scheduling and power allocation decisions of the BSs within the cluster. Hence, compared with the SC scheme, CS-BPC can still provide a large performance gain for the users located in the cluster center areas, where ICI is high.

Note that with the objective of maximizing sum rate, some users may be excluded from transmission. Figure C.4 shows the probability of serving different number of users for each CoMP transmission scheme versus different normalized distance. We can see that the probability of serving all the \(M = 3\) users decreases for all the considered CoMP transmission schemes as the normalized distance increases, i.e., when the users move towards the cluster center area. Compared with NCJT-BPC and CS-BPC for the users located at cluster center \((d/R = 1)\), ZF-OPA and ZF-EPA can achieve much higher probability of serving all the 3 users, as ICI cancellation is provided via zero-forcing. For CS-BPC, where ICI can only be reduced via coordinated scheduling, the probability of serving only one user at the cluster center is very high, i.e., 80%. The probability of serving 3 users,

\(^2\)The difference in performance between NCJT-BPC and CS-BPC is influenced by the speed and accuracy of cell selection algorithm.
2 users and 1 user for NCJT-BPC at \( d/R = 1 \) is 7.5%, 43.9% and 48.6% respectively.

### C.4.2 Sum Rate Performance with Predicted CSI

Due to practical issues (e.g., feedback and backhaul constraints, user mobility), only imperfect CSI is available at the CU, which affects the performance of CoMP transmission schemes. In this subsection, the performance of the considered CoMP transmission schemes is evaluated with predicted CSI. The effects of feedback and backhaul latency \( \Delta t \), and the user mobility \( v \) are considered. Channel prediction accuracy is obtained under the assumptions of Figure C.1.

First, we investigate the effect of the feedback and backhaul latency on the average sum rate versus the normalized distance \( (d/R) \) for \( \Delta t = 10.2, 20.4, \) and 30.6 ms respectively, when the user speeds \( (v) \) are set to 5 km/h. Figure C.5 shows that the average sum rate of the considered CoMP transmission schemes decreases over all cluster area as \( \Delta t \) increases. Compared with the achieved performance under perfect CSI (see Figure C.3), the average sum rate of the sector center users when \( \Delta t = 30.6 \) ms is decreased approximately by 30.2%, 24.1%, 1.5% and 1.2% for ZF-OPA, ZF-EPA, NCJT-BPC and CS-BPC respectively. For the cluster center users with \( \Delta t = 30.6 \) ms, the average sum rate of ZF-OPA and ZF-EPA dramatically decreases to 53.4% and 49.5%; while the performance loss due to imperfect CSI for NCJT-BPC and CS-BPC is 10.8% and 6.0% respectively. Hence, NCJT-BPC and CS-BPC are more robust to the effect of delay.

When the value of delay is relatively small, e.g., \( \Delta t = 10.2 \) ms, ZF-OPA and ZF-EPA still achieve better performance compared with other transmission schemes. However, when the delay increases, e.g., \( \Delta t = 20.4 \) ms, the average sum rate of NCJT-BPC begin to converge to that achieved by ZF-EPA. When \( \Delta t = 30.6 \) ms, NCJT-BPC outperforms ZF-OPA for \( d/R > 0.4 \), with ZF-EPA falling below all other CoMP transmission schemes for \( d/R < 0.9 \). Note that in a realistic CoMP system, backhaul links can be implemented via high-latency X2 interfaces [6]. Considering the feedback latency and the data sharing among coordinated BSs, the total latency may be greater than 30.6 ms. Therefore, for high-latency backhaul links, NCJT-BPC is a better choice for CoMP transmission design.
Figure C.4: The probability of serving different number of users vs. different normalized distance, $d/R$. Perfect CSI is assumed to be available at the CU.

Figure C.5: Average sum rate vs. different normalized distance, $d/R$. $v = 5$ km/h. $\Delta t = 10.2, 20.4, 30.6$ ms.
Figure C.6 shows the effect of user mobility on the performance of the considered transmission schemes. The feedback and backhaul latency ($\Delta t$) is set to 5.1 ms. The average sum rate of each scheme is plotted versus normalized distance for $v = 10, 30, 50 \text{ km/h}$ respectively. We can see that the performance of ZF-OPA and ZF-EPA significantly decreases as the user velocity ($v$) increases. The NCJT-BPC and CS-BPC schemes are more robust to the effect of user mobility.

Note that ZF-OPA, which achieves the best performance with perfect CSI, falls below NCJT-BPC for most distances when $v = 30 \text{ km/h}$. When $v = 50 \text{ km/h}$, where the channel uncertainty becomes higher, NCJT-BPC and CS-BPC converge to the SC scheme for $d/R < 0.4$, with ZF-OPA and ZF-EPA falling even below the SC scheme in the sector center area ($d/R < 0.6$). Hence, for the high mobility users located in the sector center area, a system would choose the SC scheme for data transmission. However, for $d/R > 0.4$, NCJT-BPC and CS-BPC can still achieve significant performance improvement compared with the SC scheme, e.g., the average sum rate gain provided by NCJT-BPC and CS-BPC at $d/R = 1$ is 168.3% and 119.9% respectively. Therefore, the system could switch to NCJT-BPC to increase the sum rate when the high mobility users are located at the cluster center area.

C.5 Conclusions

In this contribution, three CoMP transmission schemes, zero-forcing coherent joint transmission, non-coherent joint transmission and coordinated scheduling, have been compared for a CoMP cluster under predicted CSI. The effects of feedback and backhaul latency, as well as user mobility are studied. The considered performance metric is the average sum rate.
rate of a CoMP cluster, under per-BS power constraints. It has been shown that non-coherent joint transmission and coordinated scheduling are more robust to the channel uncertainty, while the performance of zero-forcing joint transmission heavily relies on the quality of CSI available at the control unit. Therefore, depending on the feedback and backhaul latency, user mobility and user locations, a system could switch between the transmission schemes to improve sum rate. For example, with low feedback and backhaul latency, zero-forcing joint transmission can be selected to serve the low mobility users. For the high mobility users located at the sector center area, a system would choose traditional single cell transmission, and then switch to the non-coherent joint transmission scheme when the high mobility users move to the cluster center area.

The CoMP transmission schemes in this paper are designed with the objective of maximizing sum rate. Hence, all users in the cluster are not always served in a particular resource slot. In future work, the user fairness will be taken into account. In addition, distributed CoMP network frameworks will be considered and compared with this centralized framework.

C.6 Appendix

We here assume Kalman predictors, located at the $m = 1, \ldots, M$ users, with perfect knowledge of the channel statistics over time, modeled by an Auto Regressive (AR) model

$$z_m(t + 1) = Az(t) + Be_m(t), \quad h_m(t) = Cz(t). \quad (16)$$

Here, $z_m(t)$ is the state vector, $e_m(t)$ is the process noise and $A$, $B$ and $C$ are the state space matrices on diagonal form.

To improve prediction performance, at the price of higher computational complexity, a number of $\rho > 1$ adjacent pilot bearing subcarriers can be predicted jointly. Then every $h_{m,n}$ in $h_m$ is a $\rho$ sized vector including the channels of the pilot bearing subcarriers. We here assume that these subcarriers are flat fading, which is reasonable for an OFDM system when $\rho$ is kept low and pilots are not too sparse in frequency.

Through the Kalman equations (see e.g. equations (3.4.33)-(3.4.39) of [12]) and the pilot measurements up to time $t$ we gain an estimate of the state variable vector $\hat{z}_m(t|t)$ and also its covariance matrix $P(t|t)$, at time $t$. The later can be used to iteratively calculate the covariance matrix for the predicted state variable vector $\hat{z}_m(t + \Delta t|t)$ at time $t + \Delta t$ through

$$P(t + k|t) = AP(t + k - 1|t)A^H + BQB^H, \quad (17)$$

for $k = 1, \ldots, \Delta t$. Here, $Q$ is the covariance matrix of the process noise in (16) and given by $Q = R_h \otimes \left( C \left( B1B^H \otimes (1 - aa^H) \right) C^H \right)$ [12], where $\otimes$ denotes element wise division, and $a$ is a vector with the eigenvalues of $A$ (i.e., the poles of the system). For flat fading channels and orthogonal pilots, the covariance matrix $R_h = E[hh^H]$ is a block diagonal matrix

$$R_h = \text{diag}\{p_{m1}\sigma^2_{h_{1,1}}/\sigma^2, \ldots, p_{mN}\sigma^2_{h_{N,N}}/\sigma^2\}, \quad (18)$$

where $\frac{p_{mn}\sigma^2_{h_{m,n}}}{\sigma^2}$ is the SNR of the measured pilot from BS $n$. Since the channel is predicted as $\hat{h}_m(t|t - \Delta t) = C\hat{z}_m(t + \Delta t|t)$, the covariance of the channel prediction error can then
be calculated through \( R_{\Delta h} = CP(t + \Delta t|t)C^H \). Assuming that the predictions of the channels from different BS are uncorrelated we get that, for flat fading channel,

\[
R_{\Delta h} = \text{diag}\left\{ \frac{\sigma^2_{\Delta h,1}}{\sigma^2_1[l\times l]}, \ldots, \frac{\sigma^2_{\Delta h,N}}{\sigma^2_1[l\times l]} \right\}.
\]

(19)

References


Paper D

I/Q Imbalance in Two-Way AF Relaying

Jingya Li, Michail Matthaiou, and Tommy Svensson

Published in
IEEE Transactions on Communications,
©2014 IEEE
The layout has been revised.
Typographical adjustments have been made.
Abstract

We analyze the performance of dual-hop, two-way amplify-and-forward relaying in the presence of in-phase and quadrature-phase imbalance (IQI) at the relay node. In particular, two power allocation schemes, namely, fixed power allocation and instantaneous power allocation, are proposed to improve the system reliability and robustness against IQI under a total transmit power constraint. For each proposed scheme, the outage probability is investigated over independent, non-identically distributed Nakagami-$m$ fading channels, and exact closed-form expressions and bounds are derived. Our theoretical analysis indicates that without IQI compensation, IQI can create fundamental performance limits on two-way relaying. However, these limits can be avoided by performing IQI compensation at source nodes. Compared to the equal power allocation scheme, our numerical results show that the two proposed power allocation schemes can significantly improve the outage performance, thus reducing the IQI effects, especially when the total power budget is large.
D.1 Introduction

Relaying-assisted transmission has been considered as a promising technique to improve the system reliability and ensure high quality of service for future wireless communication networks [1, 2]. In standard unidirectional or one-way relaying, the spectral efficiency is inherently low since the transmission from the source to the destination occupies two phases (i.e., time slots). For this reason, bidirectional or two-way relaying has recently received significant research attention. Since two-way relaying allows two sources to exchange data through a relay simultaneously within two phases, it can improve the spectral efficiency [3–5]. In this context, amplify-and-forward (AF) and decode-and-forward are the two most popular relaying protocols. Recall that the former is more cost effective and has lower implementation complexity, since it only amplifies the received signal without performing any decoding. For these reasons, the focus of this paper is on two-way AF relaying.

Most works in the area of AF relaying assume that the transceiver hardware is perfect [3–9]. In practice, though, due to cost constraints, direct-conversion based architectures are typically deployed at the transceivers of AF relays for frequency translation, which translate the signal directly from/to the radio frequency (RF) band to/from the baseband without an intermediate frequency (IF) stage [10, 11]. This architecture enables low-cost monolithic integration of the RF front-ends, thanks to the elimination of external IF and image rejection filters [12]. However, due to the limited accuracy of the analog hardware and the up/down conversion operations at the relay node, direct-conversion based relaying systems are intimately affected by several RF impairments, e.g., direct current voltage offsets through self-mixing, flicker noise, and in-phase and quadrature-phase imbalance (IQI) [12–17]. This paper will focus on the latter impairment, which refers to the phase and amplitude mismatch between the in-phase (I) and quadrature (Q) signals at the transceivers. In general, such imbalance creates an additional image-signal, leading to significant performance loss especially in high-rate systems [13].

Since the hardware of the low-cost relay nodes is most likely to be of poor quality, relays are more prone to IQI. However, to the best of our knowledge, up to now very few works have investigated the impact of IQI on relaying systems. In particular, an analytical expression for the average symbol error probability (SEP) of one-way relaying was derived in [18], where IQI was considered only at the destination node. In [19], the authors considered IQI at both the source and the destination nodes in multi-hop one-way AF relaying, and evaluated the average SEP over Nakagami-\(m\) fading channels. The effects of IQI at the relay node were investigated in [20] for orthogonal frequency-division multiplexing one-way AF relaying, where it was shown that under moderate levels of IQI, AF transmission performs even worse than the direct transmission. In [21], analytical expressions for the outage probability (OP) and ergodic capacity were derived for one-way AF relaying in the presence of IQI at the relay node. The literature on two-way relaying impaired by IQI seems to be even more scarce. To the best of our knowledge, the only relevant study is [22], which considered IQI at the two source nodes of a two-way AF relaying system. The results of [22] include integral-based bounds on the average SEP, along with the development of a baseband compensation algorithm. Note that the work in [22] did not consider IQI effects at the relay node and all theoretical results were limited to the Rayleigh fading case. Most importantly, [22] did not work out any power allocation schemes to mitigate IQI.
Motivated by the above discussion, we hereafter characterize the performance of dual-hop two-way AF relaying systems, where IQI affects both the transmitter (TX) and receiver (RX) front-ends of the relay node. The contributions of this paper can be summarized as follows:

- Two power allocation schemes, i.e., fixed power allocation (FPA) and instantaneous power allocation (IPA), are proposed to improve the system reliability, thereby mitigating the IQI effects in two-way relaying. New, exact expressions for the OP are derived for the FPA scheme, considering IQI at the relay node. Moreover, tractable bounds on the OP are obtained for the IPA scheme with perfect I/Q matching at the relay node. Contrary to [22], all these results are given in closed-form and can be efficiently evaluated. Our analysis considers independent, non-identically distributed Nakagami-\(m\) fading for the two source-to-relay channels. Recall that this fading model, which has been extensively used in the performance analysis of wireless communication systems [23], imposes several mathematical challenges that are addressed in our manipulations henceforth.

- In order to gain more insights into the impact of the IQI, we elaborate on the asymptotically high SNR regime for the FPA scheme. Our analysis demonstrates that, when the relay node has the same TX and RX IQI, the fraction of the total power allocated to the relay should increase if the IQI at the relay increases. In addition, for the asymmetric channel case (i.e., different average powers of the source-to-relay channels), the source associated with the weakest link should transmit more power compared to the source associated with the strongest link. We further show that these observations also apply for the IPA scheme.

- Both analytical and numerical results indicate that without IQI compensation, IQI can create fundamental performance limits on two-way relaying. More specifically, when the target signal-to-interference-plus-noise ratio (SINR) value is larger than the inverse of the joint image-leakage ratio, as defined in Section D.2, the relaying system will always be in outage for the case without IQI compensation. However, we show that this SINR ceiling effect can be avoided by performing IQI compensation at source nodes. Finally, we demonstrate that with IQI compensation, IPA can reduce the OP by up to 4dB compared to equal power allocation (EPA).

The remainder of this paper is organized as follows: Section D.2 introduces the system and signal model. In Section D.3, we pursue an OP analysis by assuming that the transmit powers are fixed in a long time scale. Then, the transmit powers are optimized to minimize the maximum OP of the two sources. In Section D.4, an instantaneous power allocation scheme is formulated and tractable bounds on the corresponding OP are deduced. Section D.5 provides a set of numerical results to corroborate our theoretical analysis. Finally, the main results of the paper are summarized in Section D.6.

### D.2 System and Signal Model

We consider a two-way AF relaying system where two source nodes, \(S_1\) and \(S_2\), communicate with each other through a single relay node. All nodes are equipped with a single antenna, and transmission at all nodes is constrained to the half-duplex mode, i.e., the
nodes cannot transmit and receive at the same time. The data transmission is carried out in two phases, as depicted in Fig. D.1. In phase 1, $S_1$ and $S_2$ transmit simultaneously their information to the relay. In phase 2, the relay amplifies the received signal and broadcasts it to both source nodes. The RF front-ends of the source nodes are assumed to be perfect. In this paper, we focus on the impact of the IQI at the relay node, since it normally deploys lower-quality hardware. In addition, we consider a general scenario where the relay node suffers from both RX and TX IQI.\(^1\)

### D.2.1 IQI Model

In general, IQI is modeled as the phase and/or amplitude imbalance between the I and Q signal paths at the transceivers. Here, we consider an asymmetrical IQI model, where the I branch is assumed to be ideal and the errors are modeled in the Q branch. Note that this is the prevalent model used in the open literature [13, 18, 22]. In the case of TX IQI, the baseband representation of the up-converted TX signal can be expressed as

$$\hat{x} = G_1 x + G_2^* x^*$$ \hspace{1cm} (1)

where $x$ is the baseband TX signal under perfect TX I/Q matching, and $G_1$, $G_2$ are given by

$$G_1 \triangleq \left(1 + g_T e^{j\phi_T}\right)/2 \quad \text{and} \quad G_2 \triangleq \left(1 - g_T e^{-j\phi_T}\right)/2$$ \hspace{1cm} (2)

where $g_T$ and $\phi_T$ model the TX amplitude and phase mismatch, respectively. Regarding the RX IQI, the down-conversion of the RF RX signal is

$$\hat{y} = K_1 y + K_2 y^*$$ \hspace{1cm} (3)

where $y$ denotes the down-converted baseband RX signal under perfect RX I/Q matching. The coefficients $K_1$ and $K_2$ are given by

$$K_1 \triangleq \left(1 + g_R e^{-j\phi_R}\right)/2 \quad \text{and} \quad K_2 \triangleq \left(1 - g_R e^{j\phi_R}\right)/2$$ \hspace{1cm} (4)

where $g_R$ and $\phi_R$ denote the RX amplitude and phase mismatch, respectively. The $x^*$ and $y^*$ terms in (1) and (3) are often referred to as the mirror signals introduced by IQI [13]. It is noted that for perfect I/Q matching, these imbalance parameters become $g_T = g_R = 1$ and $\phi_T = \phi_R = 0$; thus, in this case, we will have $G_1 = K_1 = 1$ and $G_2 = K_2 = 0$.

\(^1\)In general, IQI is caused by analog imperfections in the I and Q arms. Specifically, the main sources of IQI include: gain errors of the mixers, phase errors between the local oscillator signals during up- and down-conversions, unbalanced low pass filters and unbalanced digital-to-analog convertors at the TX front-end, and, finally, mismatched analog-to-digital convertors at the RX front-end [12–15]. As such, IQI levels can be different in the TX and RX front-ends of the relay node [20].
D.2.2 End-to-end SNR

Let $h_1$ and $h_2$ denote the channel coefficients for the $S_1$-to-relay link and the $S_2$-to-relay link, respectively. The amplitudes $g_1 \triangleq |h_1|$ and $g_2 \triangleq |h_2|$ are modeled as independent, non-identical Nakagami-$m$ random variables with fading parameters $m_1, m_2 \geq 0.5$, and average powers $\Omega_1 \triangleq \mathbb{E}\{|h_1|^2\}$, $\Omega_2 \triangleq \mathbb{E}\{|h_2|^2\}$, respectively. Here, the operator $\mathbb{E}\{\cdot\}$ stands for expectation. The complex Gaussian receiver noises at $S_1$, $S_2$, and the relay are denoted by $n_1 \sim \mathcal{CN}(0, N_1)$, $n_2 \sim \mathcal{CN}(0, N_2)$, and $n_r \sim \mathcal{CN}(0, N_r)$, respectively. For brevity in interpretation and without significant loss of generality of the obtained insights, it is assumed that all noise powers are $N_1 = N_2 = N_r = 1$mW. We will now use the relationships (1) and (3) to derive the end-to-end SINR for each source, considering the two-phase, two-way AF relaying protocol. We assume that the channels between the sources and the relay are reciprocal, and remain constant during these two phases [3–10].

In phase 1, $S_1$ and $S_2$ transmit simultaneously their information to the relay. Under RX IQI, the baseband RX signal after down-conversion at the relay node, $y_r$, is

$$y_r = K_1(h_1x_1 + h_2x_2 + n_r) + K_2(h_1x_1 + h_2x_2 + n_r)^*$$

where $x_i \in \mathbb{C}$ is the transmitted signal from $S_i$, with average transmit power $\mathbb{E}\{|x_i|^2\} = P_i$ for $i = 1, 2$. In phase 2, the relay amplifies the received signal at baseband with an amplification factor $G$, up-converts it to RF, and then broadcasts it to both sources. With TX IQI at the relay, the baseband RX signal at $S_i$ is given by

$$y_i = h_i(G_1(Gy_r) + G_2(Gy_r)^*) + n_i.$$  

Substituting (5) into (6), we can write the received signals at $S_i$ as

$$y_i = GAh_1^2x_i + GB|h_1|^2x_i^* + GAh_1h_jx_j + GBh_1h_j^*x_j^* + GAh_in_r + GBh_in_r^* + n_i$$

where

$$A \triangleq G_1K_1 + G_2^2K_2^* \text{ and } B \triangleq G_1K_2 + G_2^2K_1^*.$$  

We assume that each source node $S_i$ has perfect instantaneous information of the channel between itself and the relay, $h_i$, $i = 1, 2$. This information is then forwarded to the relay. Thus, the variable amplification factor $G$ can be selected at the relay node as

$$G = \sqrt{\frac{P_r}{D(\rho_1P_1 + \rho_2P_2 + 1)}}$$

where $P_r$ is the transmit power of the relay node. Also, $\rho_i \triangleq |h_i|^2$ for $i = 1, 2$, and

$$D \triangleq (|G_1|^2 + |G_2|^2)(|K_1|^2 + |K_2|^2).$$

The IQI parameters ($A$, $B$ and $D$) and the gain factor $G$ are broadcasted from the relay to both sources. Recall that the channel $h_i$ is known at $S_i$, $i = 1, 2$. Therefore, each source

\footnote{Note that the proposed power allocation schemes and the pursued performance analysis can be easily extended to the case where all nodes (i.e., $S_1$, $S_2$ and the relay) have different noise powers.}
can cancel the corresponding self-interference terms, i.e., \( GA_i x_i + GB |h_i|^2 x_i \) for \( S_i \), such that
\[
\hat{y}_i = GA_i h_j x_j + GB h_i h_j^* x_j^* + GA_i n_r + GB h_i n_i^* + n_i.\tag{11}
\]
Without IQI compensation, the received SINRs at \( S_i \) can be obtained from (11) as
\[
\gamma_{i,\text{IQI}} = \frac{\rho_i \rho_j P_j}{\kappa \rho_i \rho_j P_j + (1 + \kappa) \rho_i + \frac{1}{|A|^2 \sigma^2}}\tag{12}
\]
where the ratio \( \kappa \triangleq |B|^2 / |A|^2 \) is referred to as the joint image-leakage ratio \([24, 25]\).

From (12), we observe that for any given non-zero channel gains \( \rho_1 \) and \( \rho_2 \), \( \gamma_{i,\text{IQI}} \to \frac{1}{\kappa} \) as \( P_1, P_2 \) and \( P_r \) grow large for \( i = 1, 2 \). Thus, there exists a fundamental ceiling, \( \frac{1}{\kappa} \), on the end-to-end SNR for both source nodes in the high SNR regime. This also implies that, without IQI compensation, outage will always happen if the target SINR threshold is above the inverse of the joint image rejection ratio.

In order to circumvent this fundamental limitation, we can perform IQI compensation at each source \( S_i \) by augmenting the signal \( \hat{y}_i \) with its conjugate, such that
\[
\begin{bmatrix}
\hat{y}_i \\
\hat{y}_i^*
\end{bmatrix} = G \begin{bmatrix}
A h_i & B h_i \\
B^* h_i^* & A^* h_i^*
\end{bmatrix} \begin{bmatrix}
h_j x_j \\
h_j^* x_j^*
\end{bmatrix} + \begin{bmatrix}
\tilde{n}_i \\
\tilde{n}_i^*
\end{bmatrix}\tag{13}
\]
where \( \tilde{n}_i \triangleq GA_i n_r + GB h_i n_i^* + n_i \). The transmitted data can be recovered by applying the filter matrix \( G \begin{bmatrix}
A h_i & B h_i \\
B^* h_i^* & A^* h_i^*
\end{bmatrix}^{-1} \) on the received data vector \( \begin{bmatrix}
\hat{y}_i \\
\hat{y}_i^*
\end{bmatrix} \), yielding
\[
\begin{bmatrix}
\hat{y}_i \\
\hat{y}_i^*
\end{bmatrix} = \begin{bmatrix}
h_j x_j \\
h_j^* x_j^*
\end{bmatrix}
+ \frac{1}{G (|A h_i|^2 + |B h_i|^2)} \begin{bmatrix}
A^* h_i^* & -B h_i \\
-B^* h_i^* & A h_i
\end{bmatrix} \begin{bmatrix}
\tilde{n}_i \\
\tilde{n}_i^*
\end{bmatrix}.\tag{14}
\]
Thus, the received SNRs at \( S_i \) after IQI compensation can be obtained as
\[
\gamma_i = \frac{\rho_i \rho_j P_j}{\rho_i + \frac{1}{|A|^2 \sigma^2}}\tag{15}
\]
where
\[
\alpha \triangleq \frac{|A|^2 - |B|^2}{(|A|^2 + |B|^2) D}.\tag{16}
\]
Note that this IQI parameter, \( \alpha > 0 \), will play a key role in our subsequent OP analysis and power allocation design. Comparing (15) to (12), we see that the SINR ceiling effect can be avoided by performing IQI compensation at source nodes. Clearly, for the case of ideal hardware at the relay node, we have \( A = D = 1 \) and \( B = 0 \), thus, \( \kappa = 0 \) and \( \alpha = 1 \).

**Remark 1:** For one-way relaying systems as considered in [21], it can be shown that the end-to-end SNR after IQI compensation at the destination node has the same form of (15),

---

3This zero-forcing (ZF) based IQI compensation is selected due to analytical tractability. The development of more sophisticated compensation schemes is beyond the scope of this paper. Note that in practice, we have \( \kappa < 1 \), i.e., \( |B| < |A| \). Thus, the filter matrix \( \frac{1}{G} \begin{bmatrix}
A h_i & B h_i \\
B^* h_i^* & A^* h_i^*
\end{bmatrix}^{-1} \) always exists.
D.3 Fixed Power Allocation and Performance Analysis

except from the fact that the amplification factor $G$ is reduced to $G = \sqrt{P_r / (P_j P_r + 1)}$, where $S_i$ and $S_j$ are considered as the destination and the source, respectively. Note that $\gamma_i$ in the form of (15) increases as $G$ decreases. Thus, in IQI-impaired relaying systems, one-way relaying provides higher received SNR compared to two-way relaying when considering IQI compensation at the receive node(s). However, it should be pointed out that this SNR gain is achieved at the cost of the system spectral efficiency, since one-way relaying allows only one node to receive data within two phases.

Utilizing the SNR expression in (15), in the following, we propose two robust power allocation schemes, in which the transmit power values are optimized in different time scales to improve the system reliability.

D.3 Fixed Power Allocation and Performance Analysis

In this section, we consider an FPA scheme, where the transmit powers from all nodes ($S_1$, $S_2$ and the relay) are assumed to be fixed in a long time scale, during which the channel statistics do not change. The maximum OP of the two sources is minimized under a total power consumption constraint, $P_{\text{max}}$, which represents an important design factor as for the case of green communications [26–29].

D.3.1 Exact Outage Probability Analysis and Fixed Power Allocation

Recall that the channel amplitude $g_i$ follows a Nakagami-$m_i$ distribution with fading parameters $m_i$ and average powers $\Omega_i$ for $i = 1, 2$. Therefore, $\rho_i \triangleq g_i^2$ is a Gamma random variable with shape parameter $m_i$ and scale parameter $\Omega_i / m_i$. The corresponding probability distribution function (PDF) and cumulative distribution function (CDF) for $|h_i|$ and $\rho_i$ can be found in [23, Eq. (2.20) and (2.21)], respectively. In the considered two-way relaying system, an outage event occurs at $S_i$ if its instantaneous received SNR, $\gamma_i$, falls below a certain threshold $\gamma_{\text{th}}$. Therefore, the OP at $S_i$ can be calculated as

$$P_{\text{out},i}(\gamma_{\text{th}}) = \Pr \{ \gamma_i \leq \gamma_{\text{th}} \}$$

for $i = 1, 2$, where $\Pr \{ \cdot \}$ denotes probability. Substituting (9) into (15), the received SNR at $S_i$ can be written in a general form as

$$\gamma_i = \frac{a_i \rho_1 \rho_2}{c_i \rho_1 + d_i \rho_2 + 1}$$

for $i = 1, 2$, with positive values $a_i$, $c_i$ and $d_i$, given as

$$\begin{cases} a_1 = \alpha P_2 P_r, & c_1 = \alpha P_r + P_1, & d_1 = P_2, \\ a_2 = \alpha P_1 P_r, & c_2 = P_1, & d_2 = \alpha P_r + P_2. \end{cases}$$

Based on the received SNR expression in (18), we present the following exact closed-form OP as a function of $P_1$, $P_2$ and $P_r$. 
Proposition 1: For Nakagami-\(m\) fading channels with FPA and integer \(m_i\), the OP at \(S_i\) in the presence of joint IQI at the relay node is

\[
P_{\text{out},i}(\gamma_{\text{th}}) = 1 - \frac{2}{\Gamma(m_1)} \left( \frac{m_1}{\Omega_1} \right)^{m_1} \exp \left( -\frac{m_1 \tilde{\gamma}_i}{\Omega_1} - \frac{c_i \tilde{\gamma}_i m_2}{d_i \Omega_2} \right) \times \sum_{k=0}^{m_2-1} \sum_{l=0}^{m_1-1} \sum_{j=0}^{1} \frac{1}{k!} \left( \frac{c_i m_2}{d_i \Omega_2} \right)^k \left( \frac{m_1 - 1}{j} \right) \times \tilde{\gamma}_i^{k+m_1-l} \left( \frac{\tilde{\gamma}_i + \frac{1}{c_i}}{d_i \Omega_1 \Omega_2} \right)^{l+j-k+1} \times K_{l+j-k+1} \left( 2 \sqrt{\frac{(c_i \tilde{\gamma}_i + 1) \tilde{\gamma}_i m_1 m_2}{d_i \Omega_1 \Omega_2}} \right)
\]

where \(\tilde{\gamma}_i \triangleq \frac{d_i}{a_i} \gamma_{\text{th}}\) with the parameters \(a_i\), \(c_i\) and \(d_i\) given in (19) for \(i = 1, 2\). Also, \(\Gamma(x) = \int_0^\infty t^{x-1} \exp(-t) \, dt\) is the Gamma function, and \(K_{v}(\cdot)\) is the \(v\)-th order modified Bessel function of the second kind.

Proof: See Appendix D.7.1.

The FPA problem can now be formulated as

\[
\min_{P_1,P_2,P_r} \max (P_{\text{out},1}(\gamma_{\text{th}}), P_{\text{out},2}(\gamma_{\text{th}})) \\
\text{s.t. } P_1 + P_2 + P_r \leq P_{\text{max}}.
\]

Corollary 1: Let \(P_{1,\text{opt}}, P_{2,\text{opt}}\) and \(P_{r,\text{opt}}\) be an optimal solution for the problem in (21); then, \(P_{1,\text{opt}} + P_{2,\text{opt}} + P_{r,\text{opt}} = P_{\text{max}}\).

Proof: Assume that there exists an optimal solution \((P_1, P_2, P_r)\) such that \(P_1 + P_2 + P_r < P_{\text{tot}}\). Then, it is possible to find a feasible solution \((\eta P_1, \eta P_2, \eta P_r)\) with \(\eta \triangleq \frac{P_{\text{tot}}}{P_1 + P_2 + P_r} > 1\), such that the power constraint is satisfied with equality. From (18), the corresponding \(\gamma_i\) at \(S_i\) is

\[
\gamma_i(\eta P_1, \eta P_2, \eta P_r) = \frac{a_i \rho_1 \rho_2}{\eta} + \frac{d_i \rho_1 + d_i \rho_2}{\eta} > \gamma_i(P_1, P_2, P_r)
\]

for \(i = 1, 2\). Since both SNRs are improved, from (17), the OPs of both sources are reduced. Thus, the maximum OP of the two sources is reduced, which contradicts the assumption that \((P_1, P_2, P_r)\) is an optimal solution. Therefore, the maximum OP is minimized when the inequality constraint in (21) is satisfied with equality.

Corollary 2: At the optimal point for the problem in (21), the OPs of \(S_1\) and \(S_2\) satisfy \(P_{\text{out},1}(\gamma_{\text{th}}) = P_{\text{out},2}(\gamma_{\text{th}})\).

Proof: See Appendix D.7.2.
Corollary 1 implies that the system should always use the total transmit power in order to minimize the maximum OP of the two sources. Corollary 2 claims that at the optimal solution, the two sources should have the same outage probability. From Corollaries 1 and 2, the optimization problem (21) can be transformed to

\[
\begin{align*}
\min_{P_1, P_2, P_r} & \quad P_{\text{out},1}(\gamma_{th}) \\
\text{s.t.} & \quad P_1 + P_2 + P_r = P_{\text{max}}, \\
& \quad P_{\text{out},1}(\gamma_{th}) = P_{\text{out},2}(\gamma_{th}).
\end{align*}
\]  

(23)

Unfortunately, the optimal solution of (23) cannot be expressed in closed-form. However, with the exact OP expression in (20), we can easily obtain the optimal power values numerically.

**Remark 2:** Due to power amplifiers’ hardware limitations, in some cases, individual power constraints need to be considered. For such cases, let \(P_{1,\text{max}}, P_{2,\text{max}}, \) and \(P_{r,\text{max}}\) denote the maximum transmit power of \(S_1, S_2\) and the relay, respectively. Then, by following the same methodology used in Corollaries 1 and 2, we can show that at the optimal solution, the following properties hold: 1) the relay should always transmit with its maximum power, i.e., \(P_{r,\text{opt}} = P_{r,\text{max}}\); 2) at least one of the source nodes will transmit with its maximum power, i.e., either \(P_{1,\text{opt}} = P_{1,\text{max}}\) or \(P_{2,\text{opt}} = P_{2,\text{max}}\); 3) the two source nodes have the same OP, i.e., \(P_{\text{out},1}(\gamma_{th}) = P_{\text{out},2}(\gamma_{th})\). Based on the above observations, the optimal solution of FPA under individual power constraints can be also easily obtained.

### D.3.2 Asymptotic Outage Probability Analysis and Fixed Power Allocation

In order to gain some insights into the impact of IQI on FPA, in this subsection, we elaborate on the asymptotically high SNR regime. Without loss of generality, let us assume that

\[
P_1 = \lambda P_{\text{max}}, \quad P_2 = \mu P_{\text{max}} \quad \text{and} \quad P_r = (1 - \lambda - \mu) P_{\text{max}}
\]  

(24)

with \(\lambda \geq 0, \mu \geq 0\) and \(\lambda + \mu \leq 1\). Let \(P_{\text{out},i}(\gamma_{th})\) denote the asymptotic OP at \(S_i\) for \(i = 1, 2\), with respect to a certain SNR threshold \(\gamma_{th}\). As the total transmit power \(P_{\text{max}}\) grows large, \(\tilde{\gamma}_i\) in (20) tends to zero for \(i = 1, 2\).

To pursue our asymptotic analysis, we first invoke the properties of Bessel function, \(K_v(x) = K_v(x)\) and that

\[
K_v(x) \rightarrow \begin{cases} 
-\ln(x/2) - C, & v = 0 \\
\Gamma(v) \left(\frac{2}{x}\right)^v, & v > 0,
\end{cases}\quad \text{as } x \to 0
\]

where \(C = 0.5772...\) is the Euler-Mascheroni constant \([30, \text{Eq. (9.73)}]\). Let \(\xi\) denote the item that contains \(\tilde{\gamma}_i\) in (20) such that

\[
\xi \triangleq \tilde{\gamma}_i^{k+m_1-j-1} \left(\frac{\tilde{\gamma}_i + 1}{c_i}\right)^{k-l} \left(\frac{\left(c_i \tilde{\gamma}_i + 1\right) \tilde{\gamma}_i m_2 \Omega_1}{d_i m_1 \Omega_2}\right)^{t_j+k+1} \\
\times K_{t_j-k+1} \left(2 \sqrt{\frac{\left(c_i \tilde{\gamma}_i + 1\right) \tilde{\gamma}_i m_1 m_2}{d_i \Omega_1 \Omega_2}}\right).
\]
It is easy to show that $\xi$ is non-zero as $\tilde{\gamma}_i \to 0$, if and only if $k = 0, m_1 - j - 1 = 0$. Hence, in the high SNR regime, the OP expression in (20) can be approximated by

$$P_{\text{out}, i}^\infty (\gamma_{\text{th}}) = 1 - \exp \left( - \frac{m_1 + c_i m_2}{d_i \Omega_2} \tilde{\gamma}_i \right)$$

for $i = 1, 2$. Therefore, in the high SNR regime, the optimization problem (23) can be approximated by

$$\min_{\lambda, \mu} P_{\text{out}, 1}^\infty (\gamma_{\text{th}}) \text{ s.t. } \lambda \geq 0, \mu \geq 0, \lambda + \mu \leq 1, P_{\text{out}, 1}^\infty (\gamma_{\text{th}}) = P_{\text{out}, 2}^\infty (\gamma_{\text{th}}).$$

(26)

Define $\beta \triangleq \frac{m_2}{\Omega_2} / \frac{m_1}{\Omega_1}$. Plugging (19) into (24), combining it with the equality constraint in (26), and after some simple algebra, the latter becomes

$$(\mu - \beta \lambda) \alpha (1 - \lambda - \mu) + (\mu - \lambda) (\mu + \beta \lambda) = 0.$$  

(27)

Now, we present the following insightful results for the FPA scheme in the high SNR regime:

**Corollary 3**: At the optimal point for the problem in (26), the power ratio between the two sources, $P_{1, \text{opt}}/P_{2, \text{opt}}$, satisfies $P_{1, \text{opt}}/P_{2, \text{opt}} < 1$ for $\beta > 1$, $P_{1, \text{opt}}/P_{2, \text{opt}} > 1$ for $\beta < 1$, and $P_{1, \text{opt}}/P_{2, \text{opt}} = 1$ for $\beta = 1$.

**Proof**: Note that $\alpha (1 - \lambda - \mu) > 0$ and $(\mu + \beta \lambda) > 0$. Therefore, in order to satisfy (27), we should have $(\mu - \beta \lambda) (\mu - \lambda) < 0$ or $\mu - \beta \lambda = \mu - \lambda = 0$. If $\beta > 1$, then we get $\mu - \beta \lambda < 0$ and $\mu - \lambda > 0$, thus, $P_{1, \text{opt}}/P_{2, \text{opt}} = \lambda / \mu < 1$. Similarly, it can be shown that if $\beta < 1$, then $P_{1, \text{opt}}/P_{2, \text{opt}} < 1$; while $P_{1, \text{opt}}/P_{2, \text{opt}} = 1$ if $\beta = 1$. $\blacksquare$

Corollary 3 implies that the ratio of the optimal transmit powers from $S_1$ and $S_2$ depends on the channel fading parameters $m_i$ and $\Omega_i$. When the channels $h_1$ and $h_2$ have the same fading parameter, i.e., $m_1 = m_2$, the source associated with the weakest channel, i.e., smaller $\Omega_i$, should transmit more power compared to the source associated with the strongest channel.

For the special case of i.i.d. Nakagami-$m$ fading channels ($m_1 = m_2$ and $\Omega_1 = \Omega_2$), we have $\beta = 1$. From Corollary 3, at the optimal point for problem (26), we have $P_{1, \text{opt}}/P_{2, \text{opt}} = 1$. Thus, $\mu_{\text{opt}} = \lambda_{\text{opt}}$, and the problem (26) can be simplified to

$$\min_{\lambda} \frac{2}{1 - 2\lambda} + \frac{\alpha}{\lambda} \text{ s.t. } 0 \leq 2\lambda \leq 1.$$  

(28)
By taking the first derivative of the objective function in (28) with respect to $\lambda$ and setting it to zero, we get $\lambda_{\text{opt}} = \frac{\sqrt{\alpha}}{2(\sqrt{\alpha}+1)}$. Substituting $\lambda_{\text{opt}}$ and $\mu_{\text{opt}}$ into (24), the optimal transmit powers can be obtained as

$$P_{1,\text{opt}} = P_{2,\text{opt}} = \frac{\sqrt{\alpha}}{2(\sqrt{\alpha}+1)} P_{\text{max}},$$

and $P_{r,\text{opt}} = \frac{1}{\sqrt{\alpha}+1} P_{\text{max}}$. (29)

From (29), we see that $P_{1,\text{opt}}$ and $P_{2,\text{opt}}$ are monotonically increasing functions of $\alpha$, while $P_{r,\text{opt}}$ is a monotonically decreasing function of $\alpha$. Plugging (2), (4) and (8) into (16), $\alpha$ can be replaced by

$$\alpha = \frac{8 (g_T g_R \cos \phi_T \cos \phi_R)^2}{(1 + g_T^2)(1 + g_R^2)(1 + g_T^2 g_R^2 + 2 g_T g_R \sin \phi_T \sin \phi_R)}$$

(30)

based on which, we present the following result:

**Corollary 4:** Suppose $g_T = g_R = g$, $\phi_T = \phi_R = \phi$, with $\Delta g \triangleq |1 - g|$ and $|\phi| < \frac{\pi}{2}$. The optimal transmit power at the relay node, $P_{r,\text{opt}}$ in (26), increases as the IQI at the relay node increases, i.e., as $\Delta g$ and/or $|\phi|$ increase.

**Proof:** See Appendix D.7.3.

Corollary 4 claims that, for i.i.d. Nakagami-$m$ fading channels and at high SNRs, when the relay node has the same TX and RX IQI, the power allocated to the relay should increase as the IQI at the relay node increases. This result is expected based on the SNR expression after IQI compensation. From (15), we see that by applying the inverse matrix on the received data, the IQI-induced interference can be completely cancelled. However, at the same time, this IQI compensation scheme enhances the noise power, since $\alpha$ decreases as the IQI level increases - a phenomenon known as noise enhancement for ZF-type of transformations [31]. Therefore, in order to improve the received SNR, we need to increase the transmit power from the relay node, thus, to increase $D_\alpha^2$. Interestingly, this observation stands in contrast to the case of no IQI compensation [32], in which the optimal transmit power from the relay should decrease with the IQI level.

### D.4 Instantaneous Power Allocation and Performance Analysis

The FPA scheme requires that each source node has knowledge of the statistical characteristics of $h_1$ and $h_2$.\(^4\) In this section, we assume that instantaneous channel state information of $h_1$ and $h_2$ is available at both sources. Then, the transmit powers can be optimized for each instantaneous channel realization to improve the system reliability and minimize the impact of IQI.

\(^4\)Note that, regardless of the power allocation scheme, local instantaneous channel state information is required, that is, each source node $S_i$ needs to have the instantaneous information of the channel between itself and the relay, i.e., $h_{i\rightarrow r}$. This local instantaneous information is needed for performing self-interference cancellation, as was discussed in Section D.2.
D.4.1 Instantaneous Power Allocation

We consider an IPA problem, which maximizes the minimum SNR of the two sources. As before, we assume that the system has a maximum total power constraint $P_{\text{max}}$. From (9), the transmit power of the relay node, $P_r$, can be written as $P_r = G^2D(\rho_1 P_1 + \rho_2 P_2 + 1)$. Therefore, the total transmit power, $P_{\text{tot}}$, can be calculated as

$$P_{\text{tot}} = P_1 + P_2 + G^2D(\rho_1 P_1 + \rho_2 P_2 + 1). \quad (31)$$

For each instantaneous channel realization, the optimization problem can, then, be formulated as

$$\max_{P_1, P_2, G^2} \min(\gamma_1, \gamma_2) \quad \text{s.t.} \quad P_1 + P_2 + G^2D(\rho_1 P_1 + \rho_2 P_2 + 1) \leq P_{\text{max}}, \quad (32)$$

Similar to Corollaries 1 and 2, it can be shown that the minimum SNR is maximized when the inequality constraint in (32) is satisfied with equality. Moreover, at the optimal point, it is required that $\gamma_1 = \gamma_2$. Hence, the IPA problem (32) is equivalent to

$$\max_{P_1, P_2, G^2} \gamma_1 \quad \text{s.t.} \quad P_1 + P_2 + G^2D(\rho_1 P_1 + \rho_2 P_2 + 1) = P_{\text{max}}, \quad \gamma_1 = \gamma_2. \quad (33)$$

Plugging (15) into the two equality constraints in (33), we have

$$P_1 = \frac{1 + \alpha G^2D\rho_2}{1 + \alpha G^2D\rho_1} P_2 \quad \text{and} \quad P_2 = \frac{(1 + \alpha G^2D\rho_1) \mathcal{J}}{G^2} \quad (34)$$

where

$$\mathcal{J} \triangleq \frac{(P_{\text{max}} - G^2D) G^2}{(1 + \alpha G^2D\rho_1) (1 + G^2D\rho_2) + (1 + \alpha G^2D\rho_2) (1 + G^2D\rho_1)}. \quad (35)$$

Then, (33) becomes $\max_{G^2} (\alpha D\rho_1 \rho_2 \mathcal{J})$, which can be simplified to $\max_{G^2} \mathcal{J}$. By calculating the first derivative of $\mathcal{J}$ with respect to $G^2$ and setting it to zero, the optimal $G^2$ can be derived as

$$G^2_{\text{opt}} = \frac{P_{\text{max}}}{D\left(1 + K\right)} \quad (36)$$

where

$$K \triangleq \sqrt{1 + \frac{(\alpha + 1)}{2} (\rho_1 + \rho_2) P_{\text{max}} + \alpha \rho_1 \rho_2 P^2_{\text{max}}}. \quad (37)$$

Substituting (36) into (34), the optimal $P_1$, $P_2$ and $P_r$ can be obtained as

$$P_{1,\text{opt}} = \frac{K (K + 1 + \alpha \rho_2 P_{\text{max}})}{\mathcal{I}} P_{\text{max}} \quad (38)$$
$$P_{2,\text{opt}} = \frac{K (K + 1 + \alpha \rho_1 P_{\text{max}})}{\mathcal{I}} P_{\text{max}} \quad (39)$$
$$P_{r,\text{opt}} = \left(1 - \frac{K (2K + 2 + \alpha (\rho_1 + \rho_2) P_{\text{max}})}{\mathcal{I}}\right) P_{\text{max}} \quad (40)$$
where $I \triangleq (K + 1 + \alpha \rho_1 P_{\text{max}})(K + 1 + \rho_2 P_{\text{max}}) + (K + 1 + \alpha \rho_2 P_{\text{max}})(K + 1 + \rho_1 P_{\text{max}})$. Plugging (36) and (39) into (15), the optimal SNRs of two sources are obtained as $\gamma_1 = \gamma_2 = \gamma$, where

$$\gamma \triangleq \frac{\alpha K \rho_1 \rho_2 P_{\text{max}}^2}{I}. \quad (41)$$

From (38) and (39), we observe that similar to the FPA scheme, the source associated with the weakest channel, i.e., smaller $\rho_i$, should transmit more power compared to the source associated with the strongest channel. By calculating the first derivative of $P_{r,\text{opt}}$ with respect to $\alpha$, it can be proved that the optimal instantaneous transmit power at the relay node, $P_{r,\text{opt}}$ in (40), is a monotonically decreasing function of $\alpha$. Moreover, it can be shown that the optimal SNR, $\gamma$, is an increasing function of $\alpha$. Thus, similar to Corollary 4, we can show that when the relay node has the same TX and RX IQI, as the IQI at the relay node increases, the optimal SNR decreases and the instantaneous power allocated to the relay node should increase.

**Remark 3:** When individual power constraints are considered, it can be shown that at the optimal point, the first two properties for the FPA scheme (shown in Remark 2) also hold for the IPA scheme. The third property for the IPA scheme becomes: 3) the two source nodes have the same SNR, i.e., $\gamma_1 = \gamma_2$. Based on these properties, the optimal powers of the IPA scheme under individual power constraints can be also easily obtained.

**Perfect I/Q Matching Case**

Since a general performance analysis is challenging, if not impossible, we now focus on the perfect I/Q matching case. Recall that $\alpha = 1$ and $D = 1$ for the case of perfect I/Q matching. Substituting $\alpha = 1$ into (37), we get $\mathcal{K} = \sqrt{(\rho_1 P_{\text{max}} + 1)(\rho_2 P_{\text{max}} + 1)}$. Thus, the optimal power allocation solution reduces to [7]

$$P_{1,\text{opt}} = \frac{P_{\text{max}} \sqrt{\rho_2 P_{\text{max}} + 1}}{2 \left( \sqrt{\rho_1 P_{\text{max}} + 1} + \sqrt{\rho_2 P_{\text{max}} + 1} \right)} \quad (42)$$

$$P_{2,\text{opt}} = \frac{P_{\text{max}} \sqrt{\rho_1 P_{\text{max}} + 1}}{2 \left( \sqrt{\rho_1 P_{\text{max}} + 1} + \sqrt{\rho_2 P_{\text{max}} + 1} \right)} \quad (43)$$

$$P_{r,\text{opt}} = \frac{P_{\text{max}}}{2} \quad (44)$$

and the optimal SNR becomes

$$\gamma = \frac{\rho_1 \rho_2 P_{\text{max}}^2}{2 \left( \sqrt{\rho_1 P_{\text{max}} + 1} + \sqrt{\rho_2 P_{\text{max}} + 1} \right)^2}. \quad (45)$$

From (44), we see that, for the perfect I/Q matching case, the relay and the source nodes should equally share the total transmit power, i.e., $P_{r,\text{opt}} = P_{1,\text{opt}} + P_{2,\text{opt}} = \frac{P_{\text{max}}}{2}$. From (30), it can be shown that $\alpha \leq 1$ with the equality achieved if and only if $g_T = g_R = 1$ and $\phi_T = \phi_R = 0$. Thus, for the IQI cases, we will always have $\alpha < 1$. Note that the optimal solution in (42)-(44) is obtained when $\alpha = 1$. Recall also that $P_{r,\text{opt}}$ is a monotonically decreasing function of $\alpha$. Therefore, when IQI exists at the relay node, at the optimal point, the power allocated to the relay should always be larger than half of the total transmit power.

In the following, by using the SNR expression in (45), we analyze the OP of two-way relaying, considering perfect I/Q matching at the relay node with optimal IPA.
D.4.2 Outage Probability Analysis

Note that $\gamma_1 = \gamma_2 = \gamma$ for the optimal IPA. Thus, $P_{\text{out},1}(\gamma_{\text{th}}) = P_{\text{out},2}(\gamma_{\text{th}})$. The exact OP for the optimal IPA is difficult to obtain. For this reason, we seek to derive tight upper and lower bounds on the OP. We begin our analysis with some new lower bounds.

Lower Bounds

From (45), the received SNR can be upper bounded by $\gamma \leq \gamma_{\text{upp}1}$, where

$$\gamma_{\text{upp}1} \triangleq \frac{\rho_1 \rho_2 P_{\text{max}}^2}{2 \left( \sqrt{\rho_1 P_{\text{max}}} + \sqrt{\rho_2 P_{\text{max}}} \right)^2} = \frac{P_{\text{max}}}{2 \left( \frac{1}{\sqrt{\rho_1}} + \frac{1}{\sqrt{\rho_2}} \right)^2}. \tag{46}$$

Note that (46) can be rewritten as $\gamma_{\text{upp}1} = \frac{1}{2} X^2$ with $X \triangleq \sqrt[4]{\rho_1 \rho_2 P_{\text{max}}}$. Utilizing the fact that $\min(\sqrt{\rho_1 P_{\text{max}}}, \sqrt{\rho_2 P_{\text{max}}})$ is a tight upper bound of $X$ when $\sqrt{\rho_1 P_{\text{max}}}$ and $\sqrt{\rho_2 P_{\text{max}}}$ grow large, we present the following lower bound on the OP, which becomes exact in the high SNR regime.

Proposition 2: For Nakagami-$m$ fading channels with IPA, the OP with perfect I/Q matching at the relay node is lower bounded as $P_{\text{out}}(\gamma_{\text{th}}) \geq P_{\text{out},\text{low}1}(\gamma_{\text{th}})$, where

$$P_{\text{out},\text{low}1}(\gamma_{\text{th}}) = 1 - 2 \prod_{i=1}^{2} \left( 1 - \frac{\gamma \left( m_i, \frac{m_i}{\tilde{\gamma} \Omega_i} \right)}{\Gamma(m_i)} \right). \tag{47}$$

Here, $\tilde{\gamma} \triangleq \frac{P_{\text{max}}}{x_{\text{th}}}$ and $\gamma(s, x) = \int_0^x t^{s-1} \exp(-t) \, dt$ is the lower incomplete Gamma function.

Proof: See Appendix D.7.4.

Continuing on the same path and according to the geometric mean-harmonic mean inequality, $\gamma_{\text{upp}1}$ in (46) can be further upper bounded as

$$\gamma_{\text{upp}1} \leq \frac{P_{\text{max}} \sqrt{\rho_1 \rho_2}}{8} = \frac{P_{\text{max}} g_1 g_2}{8}. \tag{48}$$

Now, we provide an alternative lower bound on the OP, which is tight at low and moderate SNRs.

Proposition 3: For Nakagami-$m$ fading channels with IPA and integer $m_2$, the OP with perfect I/Q matching at the relay node is lower bounded as $P_{\text{out}}(\gamma_{\text{th}}) \geq P_{\text{out},\text{low}2}(\gamma_{\text{th}})$, where

$$P_{\text{out},\text{low}2}(\gamma_{\text{th}}) = 1 - \frac{2}{\Gamma(m_1)} \left( \frac{m_1}{\Omega_1} \right)^{m_1} \sum_{k=0}^{m_2-1} \frac{1}{k!} \left( \frac{16 m_2}{\tilde{\gamma}^2 \Omega_2} \right)^k \times \frac{16 m_2 \Omega_1}{\tilde{\gamma}^2 \Omega_2 m_1} \left( \frac{8}{\tilde{\gamma}} \right)^k \left( \frac{m_1 m_2}{\Omega_1 \Omega_2} \right). \tag{49}$$

Proof: See Appendix D.7.5.
Note that Proposition 2 applies for any arbitrary fading parameters, whereas Proposition 3 is valid when \(m_2\) is a positive integer. For the special case of Rayleigh fading, \(m_1 = m_2 = 1\), we present the following lower bound, which remains tight over the entire SNR regime.

**Proposition 4**: For Rayleigh fading channels with IPA, the OP with perfect I/Q matching at the relay node is lower bounded as \(P_{\text{out}}(\gamma_{\text{th}}) \geq P_{\text{out,low3}}(\gamma_{\text{th}})\), where

\[
P_{\text{out,low3}}(\gamma_{\text{th}}) = 1 - \sum_{k=0}^{N} \exp\left(-\frac{1}{\tilde{\gamma}_1} \left(1 + \frac{1}{(k + 1) \Delta t}\right)^2\right) \\
\times \left(\exp\left(-\frac{(k \Delta t + 1)^2}{\tilde{\gamma}_2}\right) - \exp\left(-\frac{((k + 1) \Delta t + 1)^2}{\tilde{\gamma}_2}\right)\right) \\
- \exp\left(-\frac{1}{\tilde{\gamma}_1} - \frac{(N + 1) \Delta t + 1)^2}{\tilde{\gamma}_2}\right).
\] (50)

Here, \(N\) is an arbitrary positive integer and the interval \(\Delta t\) is an arbitrary positive value.

**Proof**: See Appendix D.7.6.

The choice of \(N\) and \(\Delta t\) affects the tightness of the lower bound, as it will be shown by our numerical results. In general, a larger \(N\) combined with a smaller \(\Delta t\) will provide a tighter lower bound.

**Upper Bounds**

As a next step, we propose new upper bounds on the OP for the IPA case.

From (45), the received SNR \(\gamma\) can be lower bounded by \(\gamma \geq \gamma_{\text{low1}}\), where

\[
\gamma_{\text{low1}} = \frac{P_{\text{max}}^2 \rho_1 \rho_2}{4(2 + \rho_1 P_{\text{max}} + \rho_2 P_{\text{max}})}
\] (51)

where we have used the inequality that \((\sqrt{\rho_1 P_{\text{max}}} + 1 + \sqrt{\rho_2 P_{\text{max}}} + 1)^2 \leq 2(\rho_1 P_{\text{max}} + \rho_1 P_{\text{max}} + 2)\). Based on (51), and utilizing the same methodology used for deducing Proposition 1, we now present the following upper bound on the OP, which is tight at low and moderate SNRs.

**Proposition 5**: For Nakagami-\(m\) fading channels with IPA and integer \(m_i\), the OP with perfect I/Q matching at the relay node is upper bounded as \(P_{\text{out}}(\gamma_{\text{th}}) \leq P_{\text{out,upp1}}(\gamma_{\text{th}})\).
where

\[
P_{\text{out,upp}1}(\gamma_{\text{th}}) = 1 - \frac{2}{(m_1 - 1)!} \left( \frac{m_1}{\Omega_1} \right)^{m_1} \exp \left( -\frac{m_1 \hat{\gamma} - m_2 \hat{\gamma}}{\Omega_1 - \Omega_2} \right) \times \sum_{k=0}^{m_2-1} \sum_{l=0}^{m_1-1} \sum_{j=0}^{1} \frac{1}{k!} \left( \frac{m_2}{\Omega_2} \right)^k \left( \frac{k}{l} \right) \left( \frac{m_1 - 1}{j} \right) \times \hat{\gamma}^{k+m_1-j-1} \left( \hat{\gamma} + \frac{2}{P_{\text{max}}} \right)^{k-l} \times \left( \frac{(P_{\text{max}} \hat{\gamma} + 2) \hat{\gamma} m_1 \Omega_2}{P_{\text{max}} m_1 \Omega_2} \right)^{\frac{i+j+1}{2}} \times K_{l+j-k+1} \left( 2 \sqrt{\frac{(P_{\text{max}} \hat{\gamma} + 2) \hat{\gamma} m_1 m_2}{P_{\text{max}} \Omega_1 \Omega_2}} \right)
\]

(52)

with \( \hat{\gamma} \triangleq \sqrt{\frac{2 \gamma_{\text{th}}}{P_{\text{max}}}} \).

As before, for the special case of Rayleigh fading \((m_1 = m_2 = 1)\), we present an alternative upper bound, which remains tight over the entire SNR regime.

**Proposition 6:** For Rayleigh fading channels with IPA, the OP with perfect I/Q matching at the relay node is upper bounded as \( P_{\text{out}}(\gamma_{\text{th}}) \leq P_{\text{out,upp}2}(\gamma_{\text{th}}) \), where

\[
P_{\text{out,upp}2}(\gamma_{\text{th}}) = 1 - \exp \left( -\frac{\hat{\gamma}^2}{\Omega_2} \left( 1 + \frac{\hat{\gamma}}{P_{\text{max}}} \right)^2 - \frac{(N + 1) \Delta t + \hat{\gamma}}{\Omega_1} \right) - \sum_{k=0}^{N} \exp \left( -\frac{\hat{\gamma}^2}{\Omega_2} \left( 1 + \frac{\hat{\gamma}}{P_{\text{max}}} \right)^2 - \frac{(k \Delta t + \hat{\gamma})}{\Omega_1} \right) + \sum_{k=0}^{N} \exp \left( -\frac{\hat{\gamma}^2}{\Omega_2} \left( 1 + \frac{\hat{\gamma}}{P_{\text{max}}} \right)^2 - \frac{(k \Delta t + \hat{\gamma})}{\Omega_1} \right)
\]

(53)

with \( \hat{\gamma} \triangleq \sqrt{\frac{2 \gamma_{\text{th}}}{P_{\text{max}}}} \). As in (50), \( N \) is an arbitrary positive integer and the interval \( \Delta t \) is an arbitrary positive value.

**Proof:** See Appendix D.7.7.

\[ \blacksquare \]

## D.5 Numerical Results

In this section, we present a set of numerical results to evaluate the performance of the proposed power allocation schemes and to verify our analytical results. The noise power is assumed to be 1mW.

### D.5.1 Performance of FPA

Figure D.2 shows the maximum OP of the two sources, i.e., \( \max(P_{\text{out,1}}(\gamma_{\text{th}}), P_{\text{out,2}}(\gamma_{\text{th}})) \), versus the phase mismatch parameter, \( \phi \). We assume a symmetric IQI case with \( \phi_T = \)
φ_R = φ and g_T = g_R = 1. The asymptotic OP of the FPA scheme, obtained by solving (26), is compared against the exact FPA solution of (23). We see that the OP increases as the phase imbalance of the relay node increases. As anticipated, the asymptotic OP based FPA scheme performs worse than the exact optimal FPA scheme. However, the performance loss becomes negligible when the average channel gain difference between the S_1-to-relay link and the S_2-to-relay link, |Ω_2 − Ω_1|, is small. This is mainly because that, when |Ω_2 − Ω_1| is large, the optimal transmit power of the source associated with the strongest channel is much smaller compared to the transmit powers of the other two nodes, as will be shown in Fig. D.4. Thus, it is difficult to satisfy the asymptotic assumption for both sources, that is ˜γ_i in (20) tends to zero for i = 1, 2. In this case, the optimal powers, which are obtained from the asymptotic OPs of the two sources, can result in performance loss. However, the asymptotic power allocation in (26) becomes more accurate, as |Ω_2 − Ω_1| decreases. In fact, when Ω_1 = Ω_2 = 1, the asymptotic OP-based FPA scheme performs almost the same as the optimal FPA scheme, with much less computational burden. In Fig. D.3, the maximum OP is plotted as a function of the maximum total transmit power, P_max. We see that the outage probability of the asymptotic OP-based FPA scheme remains remarkably close to the one achieved by the optimal FPA scheme over the entire range of P_max.

Figure D.4 plots the transmit powers of the optimal FPA scheme used in Fig. D.2 for Ω_1 = Ω_2 = 1, and Ω_1 = 0.1, Ω_2 = 1. As anticipated, the optimal transmit power at the relay node, P_{r, opt}, increases as the phase imbalance, φ, increases. Moreover, for the considered asymmetric channel case, the source associated with the weakest channel, i.e., S_1, transmits more power compared to the source associated with the strongest channel.

**Figure D.2:** Maximum outage probability vs. the phase mismatch parameter φ. The channel fading parameters are m_1 = m_2 = 2. The SNR threshold is γ_{th} = 1. The maximum total transmit power is P_{max} = 30dBm.
**Figure D.3:** Maximum outage probability vs. the maximum total transmit power, $P_{\text{max}}$. The channel fading parameters are $m_1 = m_2 = 2$. The SNR threshold is $\gamma_{\text{th}} = 1$. The IQI parameters are $g_T = g_R = 1$, $\phi_T = \phi_R = \pi/30$.

**Figure D.4:** Optimal transmit powers $P_{1,\text{opt}}$, $P_{2,\text{opt}}$ and $P_{r,\text{opt}}$ for FPA vs. the phase mismatch parameter $\phi$. The channel fading parameters are $m_1 = m_2 = 2$. The SNR threshold is $\gamma_{\text{th}} = 1$. The maximum total transmit power is $P_{\text{max}} = 30\text{dBm}$. 
**D.5 Numerical Results**

**D.5.2 Performance of IPA**

Figure D.5 shows the optimal power values in (38)-(40) for the IPA scheme as a function of the phase mismatch parameter, \( \phi \), for two different channel realizations, i.e., \( \rho_1 = \rho_2 = 1 \) and \( \rho_1 = 0.1, \rho_2 = 1 \). As before, we assume a symmetric IQI case with \( \phi_T = \phi_R = \phi \) and \( g_T = g_R = 1 \). Similar to the FPA scheme, we see that the optimal transmit power at the relay node, \( P_{r,\text{opt}} \), increases as \( \phi \) increases. For the considered asymmetric channel case, the source associated with the weakest link, \( S_1 \), transmits with a higher power. Recall that the optimal solution for the IPA scheme is achieved when \( P_{1,\text{opt}} + P_{2,\text{opt}} + P_{r,\text{opt}} = P_{\text{max}} \).

Let \( \vartheta \triangleq \frac{P_{1,\text{opt}}}{P_{2,\text{opt}}} \). From (38) and (39), we can show that \( \vartheta \) decreases as \( \phi \) increases, if \( \rho_1 < \rho_2 \). Thus, as shown in Fig. D.5, the optimal transmit power from the source associated with the weakest link, i.e., \( P_{1,\text{opt}} = \frac{\vartheta}{1+\vartheta} (P_{\text{max}} - P_{r,\text{opt}}) \), decreases as \( \vartheta \) increases. On the other hand, the optimal transmit power at the source associated with the strongest link, i.e., \( P_{2,\text{opt}} = \frac{1}{1+\vartheta} (P_{\text{max}} - P_{r,\text{opt}}) \), is only slightly affected by the phase imbalance. As \( \phi \) grows large, the value of \( \vartheta \) converges to 1, thus, the optimal powers of both sources tend to the same value. Moreover, as anticipated, for the special case of perfect I/Q matching, i.e., when \( \phi_T = \phi_R = 0 \) and \( g_T = g_R = 1 \), the optimal value of \( P_{r,\text{opt}} \) is always \( P_{\text{max}}/2 \), which is independent of the channel fading characteristics.

Figure D.6 investigates the analytical bounds derived in Propositions 2, 3 and 5 for the OP versus \( P_{\text{max}} \), considering perfect I/Q matching at the relay node. As anticipated, we see that the OP decreases as \( m \) increases, since the channel conditions become better when \( m \) becomes large. Moreover, \( P_{\text{out,low1}}(\gamma_{\text{th}}) \) matches very well with the numerical results for very large \( P_{\text{max}} \), while \( P_{\text{out,low2}}(\gamma_{\text{th}}) \) and \( P_{\text{out,upp1}}(\gamma_{\text{th}}) \) are tight for small and moderate \( P_{\text{max}} \). Generally speaking, in the performance evaluation, the maximum of the two lower bounds, i.e., \( \max(P_{\text{out,low1}}(\gamma_{\text{th}}), P_{\text{out,low2}}(\gamma_{\text{th}})) \) can be selected to provide a new lower
Figure D.6: Outage probability vs. $P_{\text{max}}$. The channel fading parameters are $m_1 = m_2 = m$, $\Omega_1 = \Omega_2 = 0.5$. The SNR threshold is $\gamma_{\text{th}} = 1$. The IQI parameters are $g_T = g_R = 1$ and $\phi_T = \phi_R = 0$.

bound which is tight over the entire SNR regime. In Fig. D.7, we examine $P_{\text{out,low3}}(\gamma_{\text{th}})$ and $P_{\text{out,upp2}}(\gamma_{\text{th}})$ derived in Propositions 4 and 6 respectively for the Rayleigh fading case.

As mentioned before, a larger $N$ combined with a smaller $\Delta t$ will provide tighter lower and upper bounds. We also observe that, for the lower bound, the improvement brought by further increasing $N$ beyond 10 becomes negligible. Regarding the upper bound, the minimum value of $N$ that provides satisfactory tightness is $N = 50$.

### D.5.3 Comparison Between FPA and IPA

#### Channel Knowledge Requirements

Regardless of the power allocation scheme, each source node $S_i$ requires local instantaneous channel state information (i.e., $h_i$), in order to perform successful self-interference cancellation. Also, the relay requires knowledge of both $h_1$ and $h_2$ in order to design the amplification factor $G$. In order to perform optimal power allocation, for the FPA scheme, each source node $S_i$ also requires the statistical characteristics of the channel between the other source and the relay (i.e., $m_j$ and $\Omega_j$, where $j = \frac{i}{2}$ with $i = 1, 2$); on the other hand, the IPA scheme requires each source node $S_i$ to have instantaneous information of the channel between the other source and the relay (i.e., $h_j$). Therefore, compared to the IPA scheme, FPA requires less channel knowledge.

#### Power Control Signaling Overhead

The transmit powers, $P_1$ and $P_2$, of the FPA scheme are selected at $S_1$ and $S_2$, respectively, based on statistical channel information. Thus, in the FPA scheme, the transmit powers
D.5 Numerical Results

Figure D.7: Outage probability vs. $P_{\text{max}}$. The channel fading parameters are $m_1 = m_2 = 1$ and $\Omega_1 = \Omega_2 = 1$. The SNR threshold is $\gamma_{\text{th}} = 5$. The IQI parameters are $g_T = g_R = 1$ and $\phi_T = \phi_R = 0$.

are fixed in a long time scale, during which the channel statistics do not change. On the other hand, in the IPA scheme, the transmit powers are optimized and updated in each instantaneous channel realization. Thus, compared to the IPA scheme, FPA also requires less power control signaling overhead.

Outage Probability

Finally, we compare the outage performance of the proposed optimal FPA and IPA schemes with the EPA scheme, where $P_1 = P_2 = P_r = P_{\text{max}}/3$. Figure D.8 shows the maximum OP versus $P_{\text{max}}$ of these three schemes. Compared to the EPA scheme, the proposed FPA and IPA schemes can significantly improve the maximum OP, thus improving the system reliability, especially when the total power budget $P_{\text{max}}$ is large. Specifically, for a given $\text{OP} = 10^{-2}$, the FPA and IPA schemes can deliver 2dB and 4dB power savings respectively, compared to the EPA scheme for the symmetric channel case (i.e., $\Omega_2 = \Omega_1 = 1$). We also observe that the IPA scheme outperforms the FPA scheme, especially at moderate and high $P_{\text{max}}$, while the performance gain increases as the average channel gain difference between the two links, $|\Omega_2 - \Omega_1|$, decreases. This implies that IPA is particularly effective for the symmetric channel case, when the total power budget $P_{\text{max}}$ is large. For the considered asymmetric channel cases, it is better to select the FPA scheme, which slightly increases the minimum OP, however, requires less channel knowledge and overhead signaling at the two source nodes.

Figure D.9 depicts the maximum OP versus the target SINR $\gamma_{\text{th}}$. We compare the OP performance of the case with IQI compensation against the case without IQI compensation for different power allocation schemes. The IQI parameters are $g_T = g_R = 1.1$, $\phi_T = \phi_R = \frac{\pi}{8}$. The corresponding value of the joint image-leakage ratio, $\kappa$, is 0.15. Recall that
Figure D.8: Maximum outage probability vs. $P_{\text{max}}$. The channel fading parameters are $m_1 = m_2 = 1$. The SNR threshold is $\gamma_{\text{th}} = 1$. The IQI parameters are $g_T = g_R = 1.1$ and $\phi_T = \phi_R = \pi/30$.

Figure D.9: Maximum outage probability vs. the SINR threshold $\gamma_{\text{th}}$ for different IQI parameters. The channel fading parameters are $m_1 = m_2 = 5$, $\Omega_1 = 0.5$ and $\Omega_2 = 1$. The maximum transmit power is $P_{\text{max}} = 25$dBm.
\textbf{D.6 Conclusions}

We analyzed the outage performance of a dual-hop two-way AF relaying in the presence of I/Q imbalance at the relay node. Two power allocation schemes, i.e., FPA and IPA, were proposed to improve the system reliability under a total transmit power constraint. In addition, a closed-form outage probability expression for the FPA scheme with I/Q imbalance was obtained. Compared to the equal power allocation scheme, it was shown that the proposed two schemes can significantly improve the outage performance, thus reducing the I/Q imbalance effects, especially when the total power budget is large. We also observed that IPA is particularly effective for the symmetric channel case. On the other hand, for the asymmetric channel case, it is better to select the FPA scheme to mitigate I/Q imbalance and keep the signaling overhead as low as possible.

\textbf{D.7 Appendices}

\textbf{D.7.1 Proof of Proposition 1}

Let \( p_{\rho_i}(x) \) and \( F_{\rho_i}(x) \) denote the PDF and CDF of \( \rho_i, \ i = 1, 2 \), respectively. Plugging (18) into (17), the OP can be written as

\[
P_{\text{out}, i}(\gamma_{\text{th}}) = \Pr \left\{ \frac{a_i \rho_1 \rho_2}{c_i \rho_1 + d_i \rho_2 + 1} \leq \gamma_{\text{th}} \right\}.
\]

(54)
Following the same methodology used in [35–37] (and most recently adopted in [21, 33]), (54) can be expressed as

\[
P_{\text{out},i}(\gamma_{\text{th}}) = \int_{0}^{\tilde{\gamma}_i} 1 \cdot p_{\rho_1}(\rho_1)d\rho_1 + \int_{\tilde{\gamma}_i}^{\infty} F_{\rho_2}(\tilde{\gamma}_i (c_i \rho_1 + 1) \over (d_i (\rho_1 - \tilde{\gamma}_i))) p_{\rho_1}(\rho_1)d\rho_1
\]

\[
= 1 - \frac{(m_1 \Omega_1)}{\Gamma(m_1)} \sum_{k=0}^{m_2-1} \left( \frac{1}{k!} \left( \frac{c_i m_2}{d_i \Omega_2} \right)^k \right) I_k
\]

where from (55) to (56), we have used the binomial expansion of \( F_{\rho_2} (\cdot) \) for integer \( m_2 \) [23, Eq. (2.21)]. Also,

\[
I_k \triangleq \int_{\tilde{\gamma}_i}^{\infty} \left( \frac{\rho_1 + \frac{1}{c_i}}{\rho_1 - \tilde{\gamma}_i} \right)^k \rho_1^{m_1-1}
\]

\[
\times \exp \left( - \frac{m_1 \Omega_1}{(c_i m_2 \Omega_2 d_i \rho_1 (\rho_1 - \tilde{\gamma}_i))} \right) d\rho_1.
\]

By making the change of variables \( x \rightarrow \rho_1 - \tilde{\gamma}_i \) and after some simple algebra, (57) becomes

\[
I_k = \exp \left( - \left( \frac{m_1 \Omega_1}{c_i m_2 \Omega_2} \tilde{\gamma}_i \right) \right)
\]

\[
\times \int_{0}^{\infty} \left( x + \frac{1}{c_i} \right)^k \left( x + \tilde{\gamma}_i \right)^{m_1-1} x^{-k}
\]

\[
\times \exp \left( - \frac{m_1 \Omega_1}{\Omega_1} x - \frac{(c_i \tilde{\gamma}_i + 1) \tilde{\gamma}_i m_2 1}{d_i \Omega_2} \right) dx.
\]

Using the definition of binomial coefficients for integer \( m_1 \), (58) can be rewritten as

\[
I_k = \exp \left( - \left( \frac{m_1 \Omega_1}{c_i m_2 \Omega_2} \tilde{\gamma}_i \right) \right)
\]

\[
\times \sum_{l=0}^{k} \sum_{j=0}^{m_1-1} \binom{k}{l} \binom{m_1-1}{j} \tilde{\gamma}_i^{m_1-1-j} \left( \frac{1}{c_i} \right)^{k-l} I_{lj}
\]

where

\[
I_{lj} \triangleq \int_{0}^{\infty} x^{l+j-k} \exp \left( - \frac{m_1 \Omega_1}{\Omega_1} x - \frac{(c_i \tilde{\gamma}_i + 1) \tilde{\gamma}_i m_2 1}{d_i \Omega_2} \right) dx.
\]

Utilizing [30, Eq. (3.471.9)], \( I_{lj} \) can be obtained as

\[
I_{lj} = 2 \left( \frac{(c_i \tilde{\gamma}_i + 1) \tilde{\gamma}_i m_2 \Omega_1}{d_i m_1 \Omega_2} \right)^{l+j-k+1}
\]

\[
\times K_{l+j-k+1} \left( 2 \sqrt{\frac{(c_i \tilde{\gamma}_i + 1) \tilde{\gamma}_i m_1 m_2}{d_i \Omega_1 \Omega_2}} \right).
\]

By substituting (61) into (59) and combining it with (56), the result in (20) is obtained.
D.7.2 Proof of Corollary 2

Without loss of generality, let us assume that there exists an optimal solution \((P_1, P_2, P_r)\) such that \(P_{\text{out},1}(\gamma_{\text{th}}, (P_1, P_2, P_r)) > P_{\text{out},2}(\gamma_{\text{th}}, (P_1, P_2, P_r))\). Based on Corollary 1, we have that \(P_1 + P_2 + P_r = P_{\text{tot}}\). From (18), we observe that decreasing \(P_1\) will increase \(\gamma_1\), while it will decrease \(\gamma_2\). Therefore, \(P_{\text{out},1}(\gamma_{\text{th}})\) decreases as \(P_1\) decreases, while \(P_{\text{out},2}(\gamma_{\text{th}})\) increases as \(P_1\) decreases. Thus, it is possible to find a feasible solution \((\tau P_1, P_2, P_r)\) with \(\tau < 1\) which satisfies that \(P_{\text{out},2}(\gamma_{\text{th}}, (\tau P_1, P_2, P_r)) = P_{\text{out},1}(\gamma_{\text{th}}, (P_1, P_2, P_r))\) and \(P_{\text{out},1}(\gamma_{\text{th}}, (\tau P_1, P_2, P_r)) < P_{\text{out},1}(\gamma_{\text{th}}, (P_1, P_2, P_r))\). Hence, replacing \(P_1\) by \(\tau P_1\) will result in the same value of the worst OP, i.e., \(P_{\text{out},1}(\gamma_{\text{th}}, (P_1, P_2, P_r))\). Therefore, \((\tau P_1, P_2, P_r)\) is also an optimal solution. However, since \(\tau < 1\), \(\tau P_1 + P_2 + P_r < P_{\text{tot}}\). This contradicts the conclusion made in Corollary 1 that the worst OP is minimized when the inequality constraint in (21) is satisfied with equality.

D.7.3 Proof of Corollary 4

Substituting \(g_T = g_R = g\) and \(\phi_T = \phi_R = \phi\) into (30), and after some simple algebra, we get

\[
\alpha = \frac{8 \cos^4 \phi}{\left(\frac{1}{g} + g\right)^4 - 2 \left(\frac{1}{g} + g\right)^2 (1 - \sin^2 \phi)}
\]  

(62)

from which, we see that \(\alpha\) decreases as \(|\phi|\) increases for \(|\phi| < \frac{\pi}{2}\). The first derivative of \(\alpha\) with respect to \(g\) can be derived as

\[
\frac{\partial \alpha}{\partial g} = f(g) \left(\frac{1}{g^2} - 1\right) \begin{cases} 
\leq 0, & g \geq 1, \\
geq 0, & 0 < g \leq 1
\end{cases}
\]  

(63)

where \(f(g) = \frac{32 \cos^4 \phi \left(\frac{1}{g} + g\right) \left(\left(\frac{1}{g} + g\right)^2 - (1 - \sin^2 \phi)\right)}{\left(\left(\frac{1}{g} + g\right)^2 - 2 \left(\frac{1}{g} + g\right)^2 (1 - \sin^2 \phi)\right)^2} > 0\). Thus, \(\alpha\) increases as \(\Delta g \triangleq |1 - g|\) increases. Recall that \(P_{r,\text{opt}}\) is a monotonically decreasing function of \(\alpha\). Hence, the statement in Corollary 4 is proved.

D.7.4 Proof of Proposition 2

Note that \(X \leq \min \left(\sqrt{\rho_1 P_{\text{max}}}, \sqrt{\rho_2 P_{\text{max}}}\right)\). Hence,

\[
\gamma_{\text{upp},1} \leq \left(\frac{\min \left(\sqrt{\rho_1 P_{\text{max}}}, \sqrt{\rho_2 P_{\text{max}}}\right)}{2}\right)^2.
\]  

(64)
Therefore, $P_{\text{out}}(\gamma_{th})$ can be lower bounded by $P_{\text{out}}(\gamma_{th}) \geq P_{\text{out,low1}}(\gamma_{th})$, where

$$P_{\text{out,low1}}(\gamma_{th}) \triangleq \Pr \left\{ \frac{\left(\min\left(\sqrt{\rho_1 P_{\text{max}}}, \sqrt{\rho_2 P_{\text{max}}}\right)\right)^2}{2} \leq \gamma_{th} \right\}$$

$$= F_{\min\left(\sqrt{\rho_1 P_{\text{max}}}, \sqrt{\rho_2 P_{\text{max}}}\right)}\left(\sqrt{2\gamma_{th}}\right)$$

$$= 1 - \Pr \left\{ \min\left(\sqrt{\rho_1 P_{\text{max}}}, \sqrt{\rho_2 P_{\text{max}}}\right) > \sqrt{2\gamma_{th}} \right\}$$

$$\overset{(a)}{=} 1 - \Pr \left\{ \rho_1 > \frac{2\gamma_{th}}{P_{\text{max}}} \right\} \Pr \left\{ \rho_2 > \frac{2\gamma_{th}}{P_{\text{max}}} \right\}$$

$$= 1 - \prod_{i=1}^{2} F_{\rho_i}\left(\frac{1}{\tilde{\gamma}}\right)$$

(65)

where $(a)$ follows from the fact that $\rho_1$ and $\rho_2$ are independent of each other. The desired result is obtained by substituting the CDF of $\rho_1$ and $\rho_2$ into (65).

### D.7.5 Proof of Proposition 3

Let $p_{g_i}(x)$ and $F_{g_i}(x)$ denote the PDF and CDF of $g_i$, $i = 1, 2$, respectively. From (48), $P_{\text{out}}(\gamma_{th})$ can be lower bounded by $P_{\text{out}}(\gamma_{th}) \geq P_{\text{out,low2}}(\gamma_{th})$, where

$$P_{\text{out,low2}}(\gamma_{th}) \triangleq \Pr \left\{ \frac{P_{\text{max}} g_1 g_2}{8} \leq \gamma_{th} \right\} = \Pr \left\{ g_1 g_2 \leq \frac{4}{\tilde{\gamma}} \right\}$$

(66)

with

$$\Pr \left\{ g_1 g_2 \leq \frac{4}{\tilde{\gamma}} \right\} = \int_{0}^{\infty} p_{g_1}(x) F_{g_2}\left(\frac{4}{\tilde{\gamma}x}\right) dx$$

(67)

$$= 1 - 2 \left( \frac{m_1}{\Omega_1} \right)^{m_1} \sum_{k=0}^{m_2-1} \frac{1}{k!} \left( \frac{16 m_2}{\Omega_2^2 \tilde{\gamma}^2} \right)^k J_k$$

(68)

where from (67) to (68) we have used the binomial expansion for integer $m_2$. Also,

$$J_k \triangleq \int_{0}^{\infty} x^{2m_1-2k-1} \exp \left( -\frac{m_1}{\Omega_1} x^2 - \frac{16 m_2}{\Omega_2^2 \tilde{\gamma}^2} \frac{1}{x^2} \right) dx$$

$$= \left( \frac{16 m_2 \Omega_1}{\tilde{\gamma}^2 \Omega_2 m_1} \right)^{\frac{m_1-1}{2}} K_{m_1-k} \left( \frac{8 \tilde{\gamma}}{\gamma} \sqrt{\frac{m_1 m_2}{\Omega_1 \Omega_2}} \right).$$

(69)

Here, we have used [30, Eq. (3.478.4)] to evaluate the integral in (69). By substituting (69) into (68), the result in (49) is obtained.
D.7.6 Proof of Proposition 4

From (46), $P_{\text{out}}(\gamma_{\text{th}})$ can be lower bounded by

$$P_{\text{out}}(\gamma_{\text{th}}) \geq \Pr \left\{ \frac{P_{\text{max}}}{2 \left( \frac{1}{g_1} + \frac{1}{g_2} \right)^2} \leq \gamma_{\text{th}} \right\}$$

$$= \Pr \left\{ \frac{1}{g_1} + \frac{1}{g_2} \geq \sqrt{\gamma} \right\},$$  \hspace{1cm} (70)

with

$$\Pr \left\{ \frac{1}{g_1} + \frac{1}{g_2} \geq \sqrt{\gamma} \right\} = 1 - \frac{2 \tilde{I}_1}{\bar{\gamma} \Omega_2}$$  \hspace{1cm} (71)

where $\tilde{I}_1 \triangleq \tilde{\gamma} \int_0^{\infty} x \exp \left( -\frac{1}{\tilde{\gamma} \Omega_2} x^2 - \frac{1}{\Omega_2} x^2 \right) dx$. By making the change of variables $y \rightarrow \sqrt{\gamma} x$, we get

$$\tilde{I}_1 \overset{(a)}{=} \int_1^{\infty} y \exp \left( -\frac{1}{\tilde{\gamma} \Omega_2} y^2 \right) f_1 (y) \, dy$$

$$\overset{(b)}{=} \sum_{k=0}^{N} \int_{k\Delta t+1}^{(k+1)\Delta t+1} y \exp \left( -\frac{1}{\tilde{\gamma} \Omega_2} y^2 \right) f_1 (y) \, dy$$

$$+ \int_{(N+1)\Delta t+1}^{\infty} y \exp \left( -\frac{1}{\tilde{\gamma} \Omega_2} y^2 \right) f_1 (y) \, dy$$

where from (a) to (b) we have partitioned the integration region into $N+1$ segments with an equal width $\Delta t$ for the first $N$ segments, and defined $f_1 (y) \triangleq \exp \left( -\frac{1}{\tilde{\gamma} \Omega_1} \left( 1 + \frac{1}{y^{1/2}} \right)^2 \right)$. Note that $f_1 (y)$ is an increasing function of $y$ for $y > 1$. Hence, we can upper bound $\tilde{I}_1$ as

$$\tilde{I}_1 \leq \sum_{k=0}^{N} \exp \left( -\frac{1}{\tilde{\gamma} \Omega_1} \left( 1 + \frac{1}{(k+1) \Delta t} \right)^2 \right)$$

$$\times \int_{k\Delta t+1}^{(k+1)\Delta t+1} y \exp \left( -\frac{y^2}{\tilde{\gamma} \Omega_2} \right) dy$$

$$+ \exp \left( -\frac{1}{\tilde{\gamma} \Omega_1} \right) \int_{(N+1)\Delta t+1}^{\infty} y \exp \left( -\frac{y^2}{\tilde{\gamma} \Omega_2} \right) dy.$$  \hspace{1cm} (72)

By evaluating the integrals in (72) with the aid of [30, Eq. (3.381.9)], then substituting the result into (71) and combining with (70), we can readily obtain (50).

D.7.7 Proof of Proposition 6

From (45), and noting that $\sqrt{\rho_i P_{\text{max}} + 1} \leq \sqrt{\rho_i P_{\text{max}} + 1} = g_i \sqrt{P_{\text{max}} + 1}$, the end-to-end SINR, $\gamma$, can be lower bounded as

$$\gamma \geq \frac{g_1^2 g_2^2 P_{\text{max}}^2}{8 \left( \sqrt{P_{\text{max}} + 1} \right)^2}.$$  \hspace{1cm} (73)
based on which, the OP can be upper bounded as
\[
P_{\text{out}}(\gamma_{\text{th}}) \leq \Pr \left\{ g_1^2 g_2^2 \leq \frac{2 \gamma_{\text{th}}}{P_{\text{max}}} \left( g_1 + g_2 + \frac{2}{\sqrt{P_{\text{max}}}} \right)^2 \right\}
\]

\[
= \Pr \left\{ g_1 g_2 \leq \frac{\gamma}{g_1 + g_2 + \frac{2}{\sqrt{P_{\text{max}}}}} \right\}
\]

(74)

with
\[
\Pr \left\{ g_1 g_2 \leq \frac{\gamma}{g_1 + g_2 + \frac{2}{\sqrt{P_{\text{max}}}}} \right\}
\]

\[
= \int_0^\gamma 1 \cdot p_{g_1}(x) dx + \int_\gamma^\infty \frac{\gamma}{x - \gamma} \cdot p_{g_1}(x) dx
\]

\[
= 1 - \frac{2}{\Omega_1} I_2.
\]

(75)

Applying once more the methodology of Appendix D.7.6, we get
\[
I_2 \triangleq \int_\gamma^\infty x \exp \left( -\frac{1}{\Omega_1} x^2 \right) f_2(x) dx
\]

\[
= \sum_{k=0}^N \int_{k \Delta t + \gamma}^{(k+1) \Delta t + \gamma} x \exp \left( -\frac{1}{\Omega_1} x^2 \right) f_2(x) dx
\]

\[
+ \int_{(N+1) \Delta t + \gamma}^\infty x \exp \left( -\frac{1}{\Omega_1} x^2 \right) f_2(x) dx
\]

with
\[
f_2(x) \triangleq \exp \left( -\frac{\gamma^2}{\Omega_2} \left( 1 + \frac{\gamma + 2}{x - \gamma} \right)^2 \right).
\]

Note that \( f_2(x) \) is an increasing function of \( x \) for \( x > \gamma \). Hence, we get
\[
I_2 \geq \sum_{k=0}^N \exp \left( -\frac{\gamma^2}{\Omega_2} \left( 1 + \frac{\gamma + 2}{k \Delta t} \right)^2 \right)
\]

\[
\times \int_{k \Delta t + \gamma}^{(k+1) \Delta t + \gamma} x \exp \left( -\frac{1}{\Omega_1} x^2 \right) dx
\]

\[
+ \exp \left( -\frac{\gamma^2}{\Omega_2} \left( 1 + \frac{\gamma + 2}{(N+1) \Delta t} \right)^2 \right)
\]

\[
\times \int_{(N+1) \Delta t + \gamma}^\infty x \exp \left( -\frac{1}{\Omega_1} x^2 \right) dx.
\]

(76)

By evaluating the integrals in (72) with the aid of [30, Eq. (3.381.9)], then, substituting the result into (75) and combining it with (74), we can obtain (53).
References


Performance Analysis and Cooperation Mode Switch in HARQ-based Relaying

Jingya Li, Behrooz Makki, and Tommy Svensson

Published in Proceedings of IEEE Global Communications Conference (GLOBECOM’14)
pp. 1-6.
Austin, Texas, USA, Dec. 2014
©2014 IEEE
The layout has been revised.
Typographical adjustments have been made.
Abstract

We study the optimal, in terms of power-limited outage probability (OP), placement of the relay and investigate the effect of relay placement on the optimal cooperation mode of the source and the relay nodes. Using hybrid automatic repeat request (HARQ) based relaying techniques, general expressions for the OP and the average transmit power are derived. The results are then particularized to the repetition time diversity (RTD) protocol. The analytical expressions are used to find the transmit powers minimizing the power-limited OP. Our results demonstrate that adaptive power allocation reduces the OP significantly. For instance, consider a Rayleigh fading channel, an OP of $10^{-3}$ and a maximum of 2 RTD-based retransmissions. Then, compared to equal power allocation, the required transmission signal-to-noise ratio (SNR) is reduced by 5 dB, if adaptive power allocation is utilized. Moreover, depending on the relay positions and the total power budget, the system should switch between the single-node transmission mode and the joint transmission mode, in order to minimize the outage probability.
E.1 Introduction

The raised user expectations of quality of service and the rapid growth of the data traffic impose a great challenge for the design of future wireless communication networks. One widely acknowledged cost- and energy-efficient solution is to densely deploy low-power access points coexisting with the traditional macro base stations, a concept named heterogeneous dense networks [1]. This approach offloads traffic from macro base stations and reduces the average distance between users and transmitters, leading to higher system reliability and, potentially, energy efficiency.

The performance of heterogeneous dense networks highly depends on the cooperation schemes between the access points and the base station, as well as the placement of the access points in the network. In this paper, we investigate the above two problems by considering a relay-assisted cooperative system using HARQ, in which the relay nodes can be interpreted as the access points, and the source node can be mapped to a base station. Note that HARQ-based relaying only requires statistical knowledge of the channel and adaptive feedback bits, and it does not need instantaneous channel state information (CSI) at transmitters, which is challenging to obtain in dense networks.

For HARQ-based relaying techniques, relay-assisted cooperative transmission can be divided into two categories, namely, Single-Node Transmission (SNT) and multi-node Joint Transmission (JT). In the SNT mode, only one node (either the source or the relay) is active in each retransmission round. In the JT mode, once the relay decodes the data correctly, the source and the relay use, e.g., distributed space-time coding to provide joint retransmission to the destination. The performance of HARQ-based relaying was investigated in [2–7] and in [8–13], considering the SNT mode and the JT mode, respectively. However, it is not clear which cooperation mode is better. Moreover, most existing works assume that all nodes have fixed powers for each (re)transmission round [4–13]. Retransmission power adjustment was investigated in [2] to improve the energy efficiency for basic HARQ schemes. However, the results in [2] are limited to the SNT mode.

Motivated by the above discussion, we hereafter study a cooperation mode switch in HARQ-based relaying using adaptive power allocation. OP is minimized under a total power constraint, which represents an important design factor as for the case of future green communications. The contributions of this paper are as follows:

- General expressions for the OP and the average total transmit power are derived, which apply to different HARQ-based relay networks with arbitrary transmit powers. The results are then particularized to the RTD protocol.

- Retransmission power adjustment is used for minimizing the power-limited OP. We demonstrate that, compared to equal power allocation, with an OP of $10^{-3}$ and a maximum of 2 retransmissions, the average SNR is reduced by 5 dB in RTD-based relaying if adaptive power allocation is utilized. Depending on the channel conditions and the total power budget, switching between the SNT mode and the JT mode is the optimal way for minimizing the OP.

- The optimal placement of the relay and the effect of the relay position on the optimal cooperation mode are investigated. We show that, in order for minimizing the power-limited OP the optimal relay position with equal power allocation is
closer to the source. On the contrary, when performing adaptive power allocation, the optimal relay position is closer to the destination.

E.2 System Model

We consider a relay-assisted cooperative communication system comprising one source (S), relay (R) and destination node (D), as illustrated in Fig. E.1. All nodes are equipped with a single antenna. Let $h_{sr}$, $h_{rd}$ and $h_{sd}$ denote the channel coefficients of the S-to-R link, the R-to-D link, and the S-to-D link, respectively. Correspondingly, we define the channel gains as $g_\vartheta \triangleq |h_\vartheta|^2$ with $\vartheta = \{sr, rd, sd\}$. It is assumed that the channels are quasi-static, that is, the channel coefficients remain constant within a transmission block of $L_c$ channel uses, generally determined by the channel coherence time, and then change independently from one block to another according to their probability density functions (PDF).

An HARQ-based retransmission protocol is utilized with a maximum of $M$ retransmission rounds, i.e., each codeword is (re)transmitted a maximum of $M + 1$ times. Targeting for low-mobility users, the block length $L_c$ is assumed to be long such that all retransmission rounds occur in a single fading block. Define a packet as the subcodeword transmitted within a transmission block. Thus, the fading coefficients do not change during a packet period, and change independently from one packet to another. Moreover, the channel coefficient of each link is assumed to be perfectly known by its corresponding receiver. However, the transmitter of each link has no instantaneous CSI, except for the HARQ feedback bits. It should be pointed out that, while we study the quasi-static model, it is not a necessary assumption and as shown in [14], there are mappings between the performance of HARQ protocols in quasi-static and fast-fading conditions.

In each transmission block, S starts transmitting data to both R and D. If the data is successfully decoded by D, an acknowledgment (ACK) is fed back from D to both S and R, then the retransmissions stop. Otherwise, D feeds back a negative-acknowledgment (NACK). R becomes active and feeds back an ACK to S, as soon as it decodes the source message before D. Once R is active, S and R provide joint data transmission to D, and the data transmission continues until the data is correctly decoded by D or the maximum number of retransmissions is reached. Based on the feedback bits (ACK/NACK signals) from R and D, the nodes S and R adapt their transmit powers so as to improve the OP. Note that, although the cooperation protocol allows joint transmission by S and R, it may be optimal to keep the S off when the R decodes the data correctly.
E.3 Problem Formulation

The following notation is used throughout the paper.

- $Q$ denotes the total number of information nats transmitted in each packet.
- $l$ represents the subcodeword length in each (re)transmission round. Thus, $r \triangleq Q/l$ (in nats-per-channel-use) gives the initial codeword rate.
- $P_{s1}^i$ and $P_{s2}^i$ denote the transmit power from $S$ at the $i$-th round when $R$ is inactive and active, respectively.
- $P_r^i$ is the transmit power from $R$ at the $i$-th round.
- $\Pr\{A_m\}$ denotes the probability of the event that the data is successfully decoded by $D$ at the end of the $m$-th (re)transmission round while it was not decodable before. In this case, the subcodewords may have been sent by $S$ only or by $S$ and $R$ jointly.
- $\Pr\{A_{m,n}\}$ denotes the probability of the event that $D$ successfully decodes the data at the end of the $m$-th (re)transmission round (and not before), when the codeword is only sent by $S$, i.e. $R$ is inactive in rounds $1, \ldots, m$. $\Pr\{A_{m,n}\}$ is the probability of the event that $D$ successfully decodes the data at the end of the $m$-th (re)transmission round (and not before), when $R$ is active in rounds $n+1, \ldots m$, with $n < m$.
- $\Pr\{B_n\}$ is the probability of the event that the data is successfully decoded by $R$ at the end of the $n$-th (re)transmission round while it was not decodable before.
- $\Pr\{S_m\}$ is the probability of the event that data transmission stops at the end of the $m$-th (re)transmission round. In this case, either the maximum (re)transmission number, $M + 1$, is reached, or the data is decoded by $D$.

According to the definitions,

$$
\Pr\{A_m\} = \Pr\{A_{m,m}\} \left(1 - \sum_{n=1}^{m-1} \Pr\{B_n\}\right)
+ \sum_{n=1}^{m-1} \Pr\{A_{m,n}\} \Pr\{B_n\},
$$

(1)

$$
\Pr\{S_m\} = \begin{cases} 
\Pr\{A_m\}, & m = 1, \ldots, M; \\
1 - \sum_{i=1}^{M} \Pr\{A_i\}, & m = M + 1.
\end{cases}
$$

(2)

Our objective is to minimize the OP subject to an average total transmit power constraint $\phi_{\text{tot}}$. The target optimization problem can be written as

$$
\min_{\{P_{s1}, P_{s2}, P_r\}} \psi, \quad \text{s.t. } \phi \leq \phi_{\text{tot}}.
$$

(3)

where $\psi$ denotes the OP and $\phi$ is the average total transmit power from $S$ and $R$. In order to formulate the optimization problem, we first derive general expressions for $\psi$ and $\phi$. 
**E.3 Problem Formulation**

*E.3.1 Outage Probability*

The OP $\psi$ is defined as the probability that D cannot decode the data until the (re)transmission is stopped, that is,

$$\psi = 1 - \sum_{m=1}^{M+1} \Pr \{ A_m \}. \quad (4)$$

Substituting (1) into (4), we get

$$\psi = 1 - \sum_{m=1}^{M+1} \Pr \{ A_{m,m} \} - \sum_{m=1}^{M+1} \sum_{n=1}^{m-1} (\Pr \{ A_{m,n} \} - \Pr \{ A_{m,m} \}) \Pr \{ B_n \}. \quad (5)$$

*E.3.2 Average Transmit Power*

The average total transmit power is defined as [15]

$$\phi \triangleq (\bar{\mathcal{E}}^s + \bar{\mathcal{E}}^r) / \bar{L} \quad (6)$$

where $\bar{\mathcal{E}}^s$ and $\bar{\mathcal{E}}^r$ are the average energy of S and R, averaged over many packet transmissions, respectively. Also, $\bar{L}$ is the expected number of channel uses within each packet transmission period. If data transmission stops at the $m$-th round, the total number of channel uses is $ml$. Therefore, the expected number of channel uses in each packet period is

$$\bar{L} = \left( \sum_{m=1}^{M+1} \Pr \{ S_m \} m \right) l = \left( M + 1 - \sum_{m=1}^{M} (M + 1 - m) \Pr \{ A_m \} \right) l. \quad (7)$$

Next, we derive the expression for $\bar{\mathcal{E}}^r$. If R correctly decodes the data at the end of the $n$-th round and data transmission stops at the end of the $m$-th round with $n < m$ (an event denoted by $B_n & S_m$), the total energy consumed by R is $\mathcal{E}^r = \sum_{i=n+1}^{m} P_i l$. Hence, the consumed energy of the R is a random variable given by

$$\mathcal{E}^r = \sum_{i=n+1}^{m} P_i l, \text{ if } B_n & S_m, \quad m = 1, \ldots, M + 1, \quad n < m. \quad (8)$$

According to the definition,

$$\Pr \{ B_n & S_m \}$$

is given by

$$= \begin{cases} \Pr \{ A_{m,n} \} \Pr \{ B_n \}, & m = 1, \ldots, M; \\ 1 - \sum_{i=1}^{n} \Pr \{ A_{i,i} \} - \sum_{j=n+1}^{M} \Pr \{ A_{j,n} \} \Pr \{ B_n \}, & m = M + 1. \end{cases} \quad (9)$$
Therefore, the expected energy consumed by R is

\[
\bar{E}^r = \sum_{m=1}^{M} \sum_{i=1}^{m-1} \Pr\{A_{m,n}\} \Pr\{B_n\} \sum_{i=n+1}^{M} P_i^s l \\
+ \sum_{n=1}^{M} \left( 1 - \sum_{i=1}^{n} \Pr\{A_{i,i}\} - \sum_{j=n+1}^{M} \Pr\{A_{j,n}\} \right) \Pr\{B_n\} \\
\times \sum_{i=n+1}^{M+1} P_i^s l. \tag{10}
\]

The same procedure is applied to derive \( \bar{E}^s \). If data transmission stops at the end of the \( m \)-th round while the codeword is only sent by S, the total energy consumed by S is \( \bar{E}^s = \sum_{i=1}^{m} P_i^s l \) for \( m = 1, \ldots, M + 1 \). If data transmission stops at the end of the \( m \)-th round while the codeword is jointly transmitted by S and R in rounds \( n + 1, \ldots, m \), the total energy consumed by S is \( \bar{E}^s = \sum_{i=1}^{n} P_i^s l + \sum_{i=n+1}^{m} P_i^s l \) for \( n < m \) and \( m = 2, \ldots, M + 1 \). Thus, the consumed energy of the S is a random variable given by

\[
\bar{E}^s = \begin{cases} 
\sum_{i=1}^{m} P_i^s l, & \text{if } B_1 \& \ldots \& B_{m-1} \& S_m \\
\sum_{i=1}^{m} P_i^s l + \sum_{i=n+1}^{m} P_i^s l, & \text{if } B_n \& S_m 
\end{cases}
\]

where

\[
\Pr\{B_1 \& \ldots \& B_{m-1} \& S_m\} = \\
\begin{cases} 
\Pr\{A_{m,m}\} \left( 1 - \sum_{n=1}^{m-1} \Pr\{B_n\} \right), & m = 1, \ldots, M \\
\left( 1 - \sum_{i=1}^{M} \Pr\{A_{i,i}\}\right) \left( 1 - \sum_{n=1}^{M} \Pr\{B_n\} \right), & m = M + 1.
\end{cases}
\]

Therefore, the expected energy consumed by S is

\[
\bar{E}^s = \sum_{m=1}^{M} \Pr\{A_{m,m}\} \left( 1 - \sum_{n=1}^{m-1} \Pr\{B_n\} \right) \sum_{i=1}^{m} P_i^s l \\
+ \sum_{m=1}^{M} \sum_{n=1}^{m-1} \Pr\{A_{m,n}\} \Pr\{B_n\} \left( \sum_{i=1}^{n} P_i^s l + \sum_{i=n+1}^{M} P_i^s l \right) l \\
+ \left( 1 - \sum_{m=1}^{M} \Pr\{A_{m,m}\}\right) \left( 1 - \sum_{n=1}^{M} \Pr\{B_n\} \right) \sum_{i=1}^{M+1} P_i^s l \\
+ \sum_{n=1}^{M} \left( 1 - \sum_{i=1}^{n} \Pr\{A_{i,i}\} - \sum_{j=n+1}^{M} \Pr\{A_{j,n}\} \right) \Pr\{B_n\} \\
\times \left( \sum_{i=1}^{n} P_i^s l + \sum_{i=n+1}^{M+1} P_i^s l \right) l. \tag{11}
\]

Substituting (7), (10) and (11) into (6), and noting that \( \psi \) can be expressed by (5), the average total transmit power is obtained as

\[
\phi = \frac{\zeta}{M + 1 - \sum_{m=1}^{M} (M + 1 - m) \Pr\{A_m\}}. \tag{12}
\]
where

\[ \varsigma = \sum_{m=1}^{M} \Pr \{ A_{m,m} \} \left( 1 - \sum_{n=1}^{m-1} \Pr \{ B_n \} \right) \sum_{i=1}^{m} P_{i1}^{s} \]

\[ + \sum_{m=1}^{M} \sum_{n=1}^{m-1} \Pr \{ A_{m,n} \} \Pr \{ B_n \} \left( \sum_{i=1}^{n} P_{i1}^{s} + \sum_{i=n+1}^{m} (P_{i2}^{s} + P_i^{r}) \right) \]

\[ + \left( 1 - \sum_{m=1}^{M} \Pr \{ A_{m,m} \} \right) \left( 1 - \sum_{n=1}^{M} \Pr \{ B_n \} \right) \sum_{i=1}^{M+1} P_{i1}^{s} \]

\[ + \sum_{n=1}^{M} \left( 1 - \sum_{i=1}^{n} \Pr \{ A_{i,i} \} - \sum_{j=n+1}^{M} \Pr \{ A_{j,n} \} \right) \Pr \{ B_n \} \]

\[ \times \left( \sum_{i=1}^{n} P_{i1}^{s} + \sum_{i=n+1}^{M+1} (P_{i2}^{s} + P_i^{r}) \right). \]

It should be pointed out that the expressions for the OP and the average total transmit power, given in (5) and (12) respectively, are general for all HARQ protocols. The difference between different HARQ protocols and fading models is in the probability terms \( \Pr \{ A_{m,m} \} \), \( \Pr \{ A_{m,n} \} \) and \( \Pr \{ B_n \} \). In the following, as an example, we particularize the general problem formulation (3) to the RTD protocol. The same method can be applied for other HARQ protocols such as incremental redundancy and basic HARQ.

**E.4 Performance Analysis for the RTD Protocol**

From (3), it can be seen that, in order to formulate the optimization problems for the classical RTD protocol, we need to represent the probabilities \( \Pr \{ A_{m,m} \} \), \( \Pr \{ A_{m,n} \} \) and \( \Pr \{ B_n \} \) as functions of \( P_{s1}^{s} \), \( P_{s2}^{s} \) and \( P_i^{r} \) with \( i = 1, \ldots, M+1 \). We consider independent Rayleigh fading channels, where the amplitudes of the communication links (i.e., \( |h_{sr}| \), \( |h_{rd}| \) and \( |h_{sd}| \)) are independent non-identical Rayleigh distributed random variables. Thus, the channel gains (i.e., \( g_\vartheta \) with \( \vartheta = \{ sr, rd, sd \} \)) follow the exponential distribution with PDF \( f_{g_\vartheta}(x) = \lambda_\vartheta e^{-\lambda_\vartheta x} \) for \( x > 0 \). The corresponding cumulative distribution function (CDF) is \( F_{g_\vartheta}(x) = 1 - e^{-\lambda_\vartheta x} \) for \( x > 0 \). Here, \( \lambda_\vartheta = \frac{1}{\bar{g}_\vartheta} \) for \( \vartheta = \{sr, rd, sd\} \), where \( \bar{g}_\vartheta \) is the expected channel gain determined by the pathloss and shadowing between the corresponding nodes. The noise powers at R and D are set to 1.

In the RTD protocol, the same subcodeword is (re)transmitted in each round, and the received signals are combined by the receiver (R and D) using maximum ratio combining. Therefore, the received SNR at R after the \( n \)-th (re)transmission round increases to

\[ \gamma_n^{r} = g_{sr} \sum_{i=1}^{n} P_{i1}^{s}, \quad (13) \]

and the data rate reduces to \( \frac{r}{n} \), where \( r \) is the initial codeword rate defined in Section E.2. The data is successfully decoded by R at the end of the \( n \)-th round (and not before) if

\[ \frac{r}{n-1} > \frac{1}{n-1} \log \left( 1 + \gamma_{n-1}^{r} \right) \quad \text{and} \quad \frac{r}{n} \leq \frac{1}{n} \log \left( 1 + \gamma_{n}^{r} \right). \]

Hence, the probability that R decodes
the data at the end of round $n$, $\Pr \{B_n\}$, is

$$
\Pr \{B_n\} = \Pr \left\{ \log \left( 1 + g_{sr} \sum_{i=1}^{n-1} P_{i1}^s \right) < r \leq \log \left( 1 + g_{sr} \sum_{i=1}^{n} P_{i1}^s \right) \right\}
$$

$$
= \Pr \left\{ \frac{e^r - 1}{\sum_{i=1}^{n-1} P_{i1}^s} \leq g_{sr} < \frac{e^r - 1}{\sum_{i=1}^{n} P_{i1}^s} \right\}
$$

$$
= e^{-\lambda_{sd} \frac{e^r - 1}{\sum_{i=1}^{n} r_{i1}}} - e^{-\lambda_{sd} \frac{e^r - 1}{\sum_{i=1}^{n} r_{i1}}}.
$$

The received SNR at D after the $m$-th (re)transmission round depends on whether the data is only sent by S or not. If R does not decode the data before the $m$-th round, the received SNR is

$$
\gamma_{d,m,m} = g_{sd} \sum_{i=1}^{m} P_{i1}^s.
$$

On the other hand, if R decodes the data at the end of $n$-th round, $n < m$, then S and R will provide joint transmission to D in rounds $n+1, \ldots, m$. Thus, the received SNR increases to

$$
\gamma_{d,m,n} = g_{sd} \left( \sum_{i=1}^{n} P_{i1}^s + \sum_{i=n+1}^{m} P_{i2}^r \right) + g_{rd} \sum_{i=n+1}^{m} P_{i}^r.
$$

Therefore, the probabilities $\Pr \{A_{m,m}\}$ and $\Pr \{A_{m,n}\}$ are found as

$$
\Pr \{A_{m,m}\} = \Pr \left\{ \log \left( 1 + g_{sd} \sum_{i=1}^{m-1} P_{i1}^s \right) < r \leq \log \left( 1 + g_{sd} \sum_{i=1}^{m} P_{i1}^s \right) \right\}
$$

$$
= \Pr \left\{ \frac{e^r - 1}{\sum_{i=1}^{m-1} P_{i1}^s} \leq g_{sd} < \frac{e^r - 1}{\sum_{i=1}^{m} P_{i1}^s} \right\}
$$

$$
= e^{-\lambda_{sd} \frac{e^r - 1}{\sum_{i=1}^{m} r_{i1}}} - e^{-\lambda_{sd} \frac{e^r - 1}{\sum_{i=1}^{m} r_{i1}}}.
$$

and

$$
\Pr \{A_{m,n}\} = \Pr \left\{ \log \left( 1 + g_{sd} \left( \sum_{i=1}^{n} P_{i1}^s + \sum_{i=n+1}^{m-1} P_{i2}^r \right) + g_{rd} \sum_{i=n+1}^{m-1} P_{i}^r \right) < r \leq \log \left( 1 + g_{sd} \left( \sum_{i=1}^{n} P_{i1}^s + \sum_{i=n+1}^{m} P_{i2}^r \right) + g_{rd} \sum_{i=n+1}^{m} P_{i}^r \right) \right\}
$$

$$
= e^{-\lambda_{sd} \frac{e^r - 1}{\sum_{i=n+1}^{m} r_{i1}}} - e^{-\lambda_{sd} \frac{e^r - 1}{\sum_{i=n+1}^{m} r_{i1}}} + K_1 + K_2
$$

(18)
E.5 Simulation Results and Discussions

In this section, we evaluate the performance of the RTD-based relaying with adaptive power allocation under power-limited conditions. The optimization problem is nonconvex in general and difficult to solve [9]. However, our analytical expressions obtained in Sections E.3 and E.4 make it possible to find the (sub)optimal power allocation solutions and evaluate the system performance numerically. As the problem is nonconvex, we cannot guarantee that our results are globally optimal. For this reason, we have checked our results with different methods such as the iterative algorithm of [16], and the “fminsearch” and “fmincon” functions of MATLAB. In all cases, the same results are obtained which is an indication of a reliable result. The performance of the optimal adaptive power allocation schemes are compared with an equal power allocation scheme where $P^a_{i1} = P^a_{i2} = P^a_i$ for $\forall i$.

E.5.1 On the Impact of Adaptive Power Allocation

Outage Probability Analysis

The OP of the RTD-based relaying system is plotted in Fig. E.2 as a function of the average total transmit power for $M = 1$ and $M = 2$ retransmissions, i.e., 2 and 3 (re)transmissions. We see that, compared to the equal power allocation scheme, adaptive power allocation reduces the OP considerably when the total transmit power budget is large. Moreover, for a given average total transmit power, increasing the number of retransmissions $M$ leads to substantial outage performance gain for the adaptive power allocation scheme. For instance, consider an OP of $10^{-3}$ and a maximum of 2 RTD-based retransmissions. Then, compared to equal power allocation, the required transmission SNR is reduced by 5 dB, if adaptive power allocation is utilized. This is because with more number of retransmissions R gets more involved and the good characteristics of R-to-D link are properly exploited.

respectively, where

$$K_1 =\begin{cases} \frac{-\lambda_{rd}(e^{\lambda_{rd}}-1)\sum_{m=1}^{m-1} p^r_{d} e^{-\frac{\lambda_{rd}}{2}}}{\sum_{m=1}^{m-1} p^r_{d} e^{-\frac{\lambda_{rd}}{2}}}, & \text{if } \frac{\lambda_{rd}}{\lambda_{ad}} = C_1 \\ \frac{-\lambda_{rd}(e^{\lambda_{rd}}-1)\sum_{m=1}^{m-1} p^r_{d} e^{-\frac{\lambda_{rd}}{2}}}{\sum_{m=1}^{m-1} p^r_{d} e^{-\frac{\lambda_{rd}}{2}} - e^{\frac{\lambda_{rd}}{2}}}, & \text{otherwise} \end{cases}$$

$$K_2 =\begin{cases} \frac{-\lambda_{ad}(e^{\lambda_{ad}}-1)\sum_{m=1}^{m-1} p^r_{d} e^{-\frac{\lambda_{ad}}{2}}}{\sum_{m=1}^{m-1} p^r_{d} e^{-\frac{\lambda_{ad}}{2}}}, & \text{if } \frac{\lambda_{ad}}{\lambda_{rd}} = C_2 \\ \frac{-\lambda_{ad}(e^{\lambda_{ad}}-1)\sum_{m=1}^{m-1} p^r_{d} e^{-\frac{\lambda_{ad}}{2}}}{\sum_{m=1}^{m-1} p^r_{d} e^{-\frac{\lambda_{ad}}{2}} - e^{\frac{\lambda_{ad}}{2}}}, & \text{otherwise} \end{cases}$$

with $C_1 \triangleq \frac{\sum_{m=1}^{m-1} p^r_{d}}{\sum_{m=1}^{m-1} p^r_{d} e^{-\frac{\lambda_{ad}}{2}}} \sum_{m=1}^{m-1} p^r_{d} e^{-\frac{\lambda_{rd}}{2}}$, $C_2 \triangleq \frac{\sum_{m=1}^{m-1} p^r_{d}}{\sum_{m=1}^{m-1} p^r_{d} e^{-\frac{\lambda_{rd}}{2}}} \sum_{m=1}^{m-1} p^r_{d} e^{-\frac{\lambda_{ad}}{2}}$.

Substituting the probability terms $\Pr\{A_{m,m}\}$, $\Pr\{A_{m,n}\}$ and $\Pr\{B_n\}$ into (5) and (12), closed-form expressions for the OP and the average total transmit power, i.e., the optimization parameters of problem 3, are derived for the considered RTD protocol, as functions of $P^a_{i1}$, $P^a_{i2}$ and $P^a_i$, $i = 1, \ldots, M + 1$. Then, from (3), the optimization problem can be formulated.
Cooperation Mode Switch

Figure E.3 investigates the cooperation mode switch with adaptive power allocation for the RTD-based relaying. The optimal power values are plotted as a function of $\phi_{\text{tot}}$. The expected channel gains are set to $\bar{g}_{sd} = 0.5$ and $\bar{g}_{sr} = \bar{g}_{rd} = 1$. From Fig. E.3 we see that $P_{22}^w = 0$ when $\phi_{\text{tot}}$ is small. When $\phi_{\text{tot}} > 6$dBm, $P_{22}^w$ becomes larger than zero and it increases as $\phi_{\text{tot}}$ increases. This implies that when the total transmit power budget is small, the SNT mode is optimal for the RTD-based relaying. As the total transmit power becomes large, the RTD-based relaying system should switch to the JT mode in order to minimize the OP. Moreover, as anticipated, when R is active, the transmit power from the R node is always larger than the transmit power from the S, i.e., $P_{r}^2 > P_{s}^2$. This is because, with the considered values of $\lambda_{sr}$, $\lambda_{rd}$ and $\lambda_{sd}$, the R-to-D channel has a better average channel characteristics compared to the S-to-D channel, i.e., $\bar{g}_{rd} > \bar{g}_{sd}$. This point is summarized in Theorem 1 as follows.

**Theorem 1:** Under an average total transmit power constraint, there exist a threshold $\phi_{\text{th}}$, such that the SNT mode is optimal in terms of OP if $\phi_{\text{tot}} < \phi_{\text{th}}$. On the other hand, when $\phi_{\text{tot}} \geq \phi_{\text{th}}$, the JT mode becomes the optimal cooperation mode in terms of outage-limited average power.

**Proof:** The proof follows from the same concept as in the water-filling technique where the better channels receive more power. Without loss of generality, we assume that $\bar{g}_{rd} > \bar{g}_{sd}$, i.e., the R-to-D channel has a better average channel characteristics than the S-to-D channel. Then, once R decodes the data, it is always better to allocate more transmit power to R (i.e., the node which is more likely to see a better channel to the D). When the total power budget $\phi_{\text{tot}}$ is small, the water-level is low. Thus, all the transmit power will be allocated to R, while the transmit power from S (i.e., the node associated to the weaker channel) is zero. As the total power budget $\phi_{\text{tot}}$ increases, the water-level
increases. Therefore, both R and S become active and the relay network exploits the cooperative diversity. In this case, the JT mode becomes optimal in order to achieve better diversity gain.

E.5.2 On the Impact of the R Position

We consider a setup where R is located on the line between S and D. The position of R is identified by the ratio $\delta \triangleq \frac{d_{sr}}{d_{sd}}$, where $d_{sr}$ and $d_{sd}$ denote the distances between S and R, and S and D, respectively. The expected channel gains are modeled by $\bar{g}_\vartheta = \alpha_\vartheta (d_\vartheta)^{-\beta_\vartheta}$ for $\vartheta = \{sr, rd, sd\}$, where $\beta_\vartheta$ is the path loss exponent of the corresponding link, and $\alpha_\vartheta$ is a parameter independent of $d_\vartheta$ and contains parameters such as the transmit antenna gain, receive antenna gain and shadowing. Here, we set $\beta_{sr} = \beta_{sd} = \beta_{rd} = 3$, $\alpha_{sr} = \alpha_{sd} = \alpha_{rd}$ and $\bar{g}_{sd} = 1$. We consider the RTD-based relaying with $M = 1$ and $r = 1$. The (re)transmission powers for each node are optimized to minimize the OP $P_{out}$, subject to an average total power constraint $\phi_{tot} = 10$dBm.

Outage Probability Analysis

Figure E.4 illustrates the OP as a function of the distance ratio. It can be seen that, compared to equal power allocation, adaptive power allocation can reduce the OP substantially, especially when the distance ratio is large, i.e., when R is placed closer to D. We also observe that there exists an optimal relay position that minimizes the OP. The optimal relay position with equal power allocation is closer to S. While when performing adaptive power allocation, the optimal relay position is closer to D. This observation is intuitive, since as the distance ratio $\delta$ increases, i.e., when the R moves towards the D, the R-to-D link becomes much better than the S-to-D link on average. Thus, once the
R decodes the data, the system should switch to the SNT mode, i.e., the S should be switched off while all power should be allocated to the R (see Fig. E.4). For the equal power allocation scheme, the transmit powers allocated to the S and R are always the same, thus, the optimal relay position should be closer to S, in order to gain more benefits from the S-to-D link.

**Cooperation Mode Switch-E**

Finally, we investigate the effects of R position on the cooperation mode selection. Figure E.5 plots the optimal transmit powers, $P_{s2}^h$ and $P_r^h$, of the adaptive power allocation scheme and the power values of the equal power allocation scheme used in Fig. E.4. We see that when $\delta \geq 0.8$, $P_{s2}^h$ becomes zero, i.e., S stops retransmission when R is active. This implies that when the average channel gain difference between R-to-D link and S-to-D link, $\bar{g}_{rd} - \bar{g}_{sd}$, is small, the JT mode is optimal. However, as $\bar{g}_{rd} - \bar{g}_{sd}$ becomes large, the system should switch to the SNT mode in order to minimize the OP.

**E.6 Conclusion**

In this paper, cooperation mode switch was investigated in relay-HARQ networks by using adaptive power allocation. In particular, analytical expressions for the OP and the average total transmit power have been derived for the RTD protocol. Retransmission power adjustment is used for minimizing the OP of the network, subject to an average total transmit power constraint. We showed that adaptive power allocation can significantly reduce the OP. Moreover, it is always better to select the SNT mode when the total power budget is low, or when the average channel gain difference between the R-to-D link and the S-to-D link is large. However, as the total transmit power increases or the average channel gain difference decreases, the system should switch to the JT mode. Finally, the
optimal cooperation mode is remarkably affected by the relay position.

References


Figure E.5: Optimal transmit powers, $P_{22}^S$ and $P_2^R$, vs. the distance ratio $\delta$. 


Joint Precoding and Load Balancing Optimization for Energy-Efficient Heterogeneous Networks

Jingya Li, Emil Björnson, Tommy Svensson, Thomas Eriksson and Mérouane Debbah

The layout has been revised.
Typographical adjustments have been made.
Abstract

This paper considers a downlink heterogeneous network, where different types of multi-antenna base stations (BSs) communicate with a number of single-antenna users. Multiple BSs can serve the users by spatial multiflow transmission techniques. Assuming imperfect channel state information at both BSs and users, the precoding, load balancing, and BS operation mode are jointly optimized for improving the network energy efficiency. We minimize the weighted total power consumption while satisfying quality of service constraints at the users. This problem is non-convex, but we prove that for each BS mode combination, the considered problem has a hidden convexity structure. Thus, the optimal solution is obtained by an exhaustive search over all possible BS mode combinations. Furthermore, by iterative convex approximations of the non-convex objective function, a heuristic algorithm is proposed to obtain a suboptimal solution with low complexity. We show that although multi-cell joint transmission is allowed, in most cases, it is optimal for each user to be served by a single BS. The optimal BS association condition is parametrized, which reveals how it is impacted by different system parameters. Simulation results indicate that putting a BS into sleep mode by proper load balancing is an important solution for energy savings.
F.1 Introduction

The rapid growth of data traffic in wireless networks impose great challenges on future wireless communication systems [1–3], in particular on improving the spectral efficiency as well as the energy efficiency. At the same time, the users are expecting that future networks will provide a uniform quality of service (QoS) over the coverage area. In many challenging scenarios, e.g., in shopping malls, dense urban environments, or during the occurrence of traffic jams, the users are non-uniformly distributed over the network [4].

One widely acknowledged cost- and energy-efficient approach to tackle these challenges is the concept of heterogeneous dense networks, where the traditional macro base stations (BSs) are complemented with a dense deployment of low-cost and low-power BSs [5–7]. By adding such a large number of small cells, the corresponding low-power BSs can offload traffic from the macro BSs, reduce the average distance between users and transmitters, and thereby improve the data rates and/or reduce the average transmit power. Since the data traffic load fluctuates greatly over the day [8], both macro and small cells might be needed at peak hours while there is an opportunity to turn off some BSs when there is little traffic in the corresponding coverage areas. Load balancing is the technique that maps the current traffic load to the available transmission resources, i.e., associates users with BSs. Mathematically speaking, the network would like to find the BS association that maximizes some performance metric, under the condition that the QoS requirements of all users are fulfilled.

Different from the traditional cellular networks, the densely deployed BSs will be heterogeneous in the number of antennas, transmit power, backhaul capacity and reliability, coverage area, etc. Moreover, the channel state information (CSI) at each BS is likely to be different and imperfect. In this complex scenario, a major research problem is to design low-complexity and robust coordinated multi-BS transmission schemes that minimize the total power consumption, while satisfying the QoS expectations of the users.

The total power consumption of the network can be modeled with a circuit part that depends on the transceiver hardware and a dynamic part that is a function of the transmitted signal power [9–12]. Adding more low-power BSs can reduce the dynamic power consumption due to the shorter propagation distances, but require more hardware; thus, it will increase the circuit power part. Note that the circuit power consumption also depends on the operational mode of each BS, i.e., whether the BS is active or in sleep mode. It has been shown that, putting a BS into sleep mode when there is nothing to transmit or receive is an important solution for energy savings [10]. Therefore, to actually improve the overall power efficiency of a heterogeneous network, the cooperation scheme, the BS operational modes, and the load balancing must be properly and jointly optimized.

Simulation-based studies for load balancing in heterogeneous networks have been performed within 3GPP, and several biased-received-power based criteria were proposed to control the number of users associated with the low-power BSs [5, 6]. Moreover, load balancing was analyzed in [13–22] for systems where the BSs are distributed according to stochastic point processes. Using stochastic geometry tools, these works have compared how different BS association rules (e.g., the nearest-BS based, the highest-received-power based, the maximum signal-to-interference-and-noise ratio (SINR) based, and the biased-SINR based cell selection) affect the downlink SINR distribution [14–16] and the average achievable rate [16–20]. We note that the results in [13–18, 20, 21] are limited to BSs with single antennas, while contemporary and future networks use multiple antennas for
downlink precoding. The papers [22] and [19] consider the practically important case of multi-antenna BSs, but these results are restricted to single-cell zero-forcing precoding with perfect CSI; in contrast, imperfect CSI and inter-cell interference coordination are essential properties of future heterogeneous networks. Moreover, shadowing has a great impact on the system performance of heterogeneous networks, but was not considered in [13, 14, 16, 17, 19–22], probably due to mathematical intractability.

The precoding design is of paramount importance in multi-antenna cellular networks, since it determines the achievable array gains and interference suppression [23]. Joint precoding and load balancing was studied in [24] for a homogeneous network, where all BSs are turned on and there is no explicit power constraints. In [25] and [26], the authors investigated joint load balancing and power control in heterogeneous networks with single-antenna BSs, where different algorithms were proposed to maximize the minimum rate subject to per-BS power constraints. Considering heterogeneous networks with multi-antenna BSs, joint load balancing and precoding algorithms were designed in [27–29] to maximize various system utilities. In [30], downlink linear precoding problems were studied jointly with BS selection. The objective was to either minimize the total transmit power or maximize the sum rate performance. The results in [30] show that by imposing certain sparsity patterns in the precoding vectors, the number of active BSs in the network can be effectively reduced. With the objective of improving network energy efficiency, radio resource optimization was studied in [31] for the downlink of an orthogonal frequency-division multiplexing (OFDM) system. In particular, the power allocation, sub-carrier allocation and the number of activated transmit antennas were jointly optimized for maximization of the energy efficiency of data transmission (bit/Joule delivered to the users). However, the work in [31] did not optimize the precoding vectors and the results were limited to a single-cell scenario. In [32], using a stochastic geometry based model, the energy efficiency of both multi-cell homogeneous and heterogeneous networks was analyzed by considering active and sleep modes for macro BSs with fixed power control. Since both BSs and users are assumed to have a single antenna in [32], precoding design was not considered.

Joint precoding and load balancing design problem is typically a mixed-integer non-linear programming problem, for which finding the global optimum is challenging [33]. Inspired by the compressive sensing literature, the reweighted $l_1$-norm technique has been adopted in [34–36], where different heuristic algorithms were proposed for solving joint precoding and BS clustering design problems. In [37, 38], group sparse optimization has been used to improve the energy efficiency of cloud radio access networks, where the weighted mixed $l_1/l_p$-norm minimization is used to induce group sparsity on the beamforming. The BSs are switched off based on the obtained group sparsity patterns. Note that in [34–38] the algorithms are designed based on the assumption of perfect CSI at both BSs and users. In this paper, we study joint precoding and load balancing optimization for energy efficient heterogeneous networks with imperfect CSI. The goal is to minimize the weighted total power consumption while satisfying QoS constraints at the users and transmit power constraints at the BSs. Although it is practically convenient and desirable to associate each user with only one BS per time-frequency resource block, our system model allows for serving users by multiple BSs. The paper investigates the following important system design questions: 1) Which and how many BSs should each user be associated with? 2) How should the precoding matrices be selected when having imperfect CSI? 3) How can we decide on the operational mode (active or sleep) for each BS? The contributions of
this paper can be summarized as follows:

- We formulate the joint load balancing and precoding as a non-convex optimization problem. We show that for a given combination of BS modes, the considered optimization problem can be reformulated as a convex semi-definite problem. Thus, we obtain the global optimal solution by an exhaustive search over all possible BS mode combinations. The obtained global optimal solution serves as an upper bound for any other suboptimal precoding and load balancing solutions, e.g., the strategies proposed in [5, 6, 13–22].

- We derive the structure of the optimal solution, by investigating the structure of the dual problem. Our result verifies the intuition that, in most cases, it is optimal for each user to be served by a single BS. However, there are also occasions when multi-BS association is beneficial. Moreover, we show that the load balancing rules previously considered in [5, 6, 13–22] are not optimal when minimizing the total power consumption under per-BS transmit power constraints and per-user QoS constraints. The optimal BS association rule consists of comparing weighted channel norms, where the weighting matrix depends on channel uncertainty, power constraints, and QoS constraints.

- We propose an efficient iterative algorithm that resolves the non-convexity of the original optimization problem by iterative convex approximations of the power consumption functions. Each iteration solves a convex problem with a modified objective function. This convex objective function is updated in each iteration such that most of the BSs with small transmit powers in the solution are driven to sleep mode. We show that the idea behind the proposed algorithm is very similar to the reweighted $l_1$-norm minimization based methods used in [34–36].

- Numerical results are provided to show how putting BSs into sleep mode by proper load balancing is a key to energy savings in heterogeneous networks. The BS activation probability is shown to depend on the target QoS requirements, as well as the ratio between the circuit power consumed in the active mode and that consumed in the sleep mode.

The remainder of this paper is organized as follows: Section F.2 introduces the system and signal model. In Section F.3, we analyze the optimal precoding and load balancing design. In Section F.4, an iterative heuristic algorithm is proposed to obtain a suboptimal solution with low complexity. Section F.5 provides a set of numerical results to illustrate our analytical results and the proposed algorithms. Finally, the main results of the paper are summarized in Section F.6.

Notation: we use upper-case bold face letters, such as $\mathbf{E}$, for matrices and lower-case bold face letters, such as $\mathbf{h}$, for vectors. $\mathbf{W} \succeq 0$ represents that the matrix $\mathbf{W}$ is positive semidefinite. $|\mathcal{C}|$ denotes the cardinality of a set $\mathcal{C}$. The operator $\mathbb{E}\{\cdot\}$ stands for expectation. The notation $\sim$ denotes “distributed as”, $\triangleq$ is used to mark definitions, $\|\cdot\|$ represents the Euclidean norm, and $\text{Tr}(\cdot)$ is the matrix trace.
We consider the downlink of a heterogeneous network consisting of $M$ base stations (BSs) and $K$ single-antenna users, as illustrated in Fig. F.1. The heterogeneity lies in the assumption that the $M$ BSs are different in terms of the number of transmit antennas, the power consumption characteristics, the channel propagation model, and the CSI quality. BSs with the same characteristics can be said to belong to the same tier or category (e.g., macro or small BS), but we stress that our system model supports anything from 1 to $M$ tiers. The users are not pre-associated with any particular cell and are randomly distributed in the network coverage area.

BS $v$ is assumed to have $N_v$ antennas. The channel from BS $v$ to user $k$ is assumed to be flat-fading, and denoted by $h_{k,v}^H \in \mathbb{C}^{1 \times N_v}$ for $v = 1, \ldots, M$ and $k = 1, \ldots, K$. In practice, these channels are imperfectly known at the BSs. This is modeled as $h_{k,v} = \hat{h}_{k,v} + e_{k,v}$, where $\hat{h}_{k,v}$ is the known estimate of $h_{k,v}$ at BS $v$. The error vector $e_{k,v} \sim \mathcal{CN}(0, \mathbf{E}_{k,v})$ is assumed to have zero-mean and a covariance matrix $\mathbf{E}_{k,v} \in \mathbb{C}^{N_v \times N_v}$. This is, for example, a good model of time-division duplex (TDD) systems where the channels are Rayleigh fading, $h_{k,v} \sim \mathcal{CN}(0, g_{k,v} \mathbf{I}_{N_v})$, and the BS uses uplink pilot signals for channel estimation. If the minimum mean-squared error (MMSE) channel estimator is used [39–42], then estimation errors are Gaussian distributed and the error covariance becomes

$$
\mathbf{E}_{k,v} = \frac{g_{k,v}}{1 + \gamma_{k,v}^p} \mathbf{I}_{N_v}
$$

where $\gamma_{k,v}^p = \frac{p \sigma_{k,v}^2}{\sigma_k^2}$ denotes the pilot SNR, $p$ is the total pilot power and $\sigma_k^2$ is the noise power. The users also need to acquire CSI, but only for the precoded channels; this is further discussed in Section F.2.2.

The received signal at user $k$ is

$$
y_k = \sum_{v=1}^{M} h_{k,v}^H x_v + n_k
$$
where \( x_v \in \mathbb{C}^{N_v \times 1} \) is the transmitted signal from BS \( v \) and \( n_k \sim \mathcal{CN}(0, \sigma_k^2) \) is the independent additive receiver noise at user \( k \).

A main goal of this paper is to determine the optimal association between users and BSs. It makes practical sense to only associate one BS with each user, but we will not make this limiting assumption at this point since we simply do not know if it is optimal. Instead, we assume that all BSs are able to transmit to all users at the same time-frequency resource block, and then our analysis will tell which and how many BSs that each user should be associated with. Motivated by the fact that tight phase synchronization between BSs is extremely difficult to achieve in practice, only linear spatial multiflow transmission is allowed [43]. This is a scheme for multiple access that allows each user to receive different parallel data streams from multiple BSs. These streams are detected sequentially at the user, based on conventional successive interference cancellation techniques [44]. Define \( V \triangleq \{1, 2, \ldots, M\} \) as the set of all BSs in the network, and let \( V_k \subseteq V \) denote the set of BSs that provide data transmission to user \( k \). Then, the set of users associated with BS \( v \) can be represented by \( U_v = \{k|v \in V_k\} \). Let \( s_{k,v} \sim \mathcal{CN}(0, 1) \) be the coded independent information symbols for user \( k \), transmitted from BS \( v \). Then, the desired signals for user \( k \) transmitted by BS \( v \) is \( \mathbf{w}_{k,v}s_{k,v} \), where \( \mathbf{w}_{k,v} \in \mathbb{C}^{N_v \times 1} \) is the linear precoding vector for user \( k \) at BS \( v \). The aggregated transmitted signal from BS \( v \) is

\[
\mathbf{x}_v = \sum_{k \in U_v} \mathbf{w}_{k,v}s_{k,v}. \tag{3}
\]

### F.2.1 Power Consumption Model

From (3), the expected transmit power from BS \( v \) can be calculated as

\[
P_{\text{trans},v} = \sum_{k \in U_v} \|\mathbf{w}_{k,v}\|^2 \mathbb{E}\{|s_{k,v}|^2\} = \sum_{k \in U_v} \|\mathbf{w}_{k,v}\|^2. \tag{4}
\]

In this paper, we adopt the linear approximated power consumption model proposed in [10, Eq. (4-3)] for 10 MHz bandwidth, where the total consumed power of BS \( v \), for \( v \in V \), is

\[
P_v = \begin{cases} N_vP_{\text{active},v} + \Delta_vP_{\text{trans},v}, & 0 < P_{\text{trans},v} \leq P_{v,\text{max}} \\ N_vP_{\text{sleep},v}, & P_{\text{trans},v} = 0, \end{cases} \tag{5}
\]

where \( P_{\text{active},v} \) is the hardware power consumption at BS \( v \) at the minimum non-zero transmit power, \( P_{\text{sleep},v} \) denotes the sleep mode power consumption of BS \( v \) with \( P_{\text{sleep},v} \leq P_{\text{active},v} \). Note that \( P_{\text{sleep},v} > 0 \) in the sleep mode (due to the DC-DC power supply, mains supply, active cooling, maintaining backhaul connections, and enabling fast turn on control signaling) [9, 10]. Here, \( P_{v,\text{max}} \) is the peak transmit power constraint for BS \( v \). The scaling factor, \( \Delta_v \geq 1 \), models the inefficiency of the power amplifier; that is, how much extra power that is consumed at BSs when the transmitted power is \( P_{\text{trans},v} \). Some example values of \( P_{\text{active},v}, P_{\text{sleep},v}, P_{v,\text{max}} \) and \( \Delta_v \) for different BS types can be found in [10, Table 8], and some of these are also given in Table F.1.

### F.2.2 Aggregated Received SINR

Each user might receive multiple information symbols, thus we need an aggregated performance measure for each user. The natural choice is the sum spectral efficiency of the
user when successive interference cancellation is applied. The power consumption at the user side might depend on how many symbols that the user receives, but this paper has an operator perspective where only the power consumptions at BSs is considered—this is the dominating factor in the downlink.

**Lemma 1** Assume that user $k$ knows the effective precoded channels $w_{H,k,v}^H \hat{h}_{k,v}$ (for all $l$ and $v$). Then, an achievable ergodic sum spectral efficiency of user $k$ is $R_k = \mathbb{E}\{\log_2 (1 + \gamma_k)\}$ where the expectation is with respect to the aggregated instantaneous SINR

$$\gamma_k = \frac{\sum_{v \in \mathcal{V}_k} |\hat{h}_{k,v}^H w_{k,v}|^2}{I_k + E_k + \sigma_k^2}$$

with

$$I_k \triangleq \sum_{v \in \mathcal{V}} \sum_{l \in \mathcal{U}_v, l \neq k} w_{l,v}^H (\hat{h}_{k,v} \hat{h}_{k,v}^H + E_{k,v}) w_{l,v}$$

being the co-user interference and

$$E_k \triangleq \sum_{v \in \mathcal{V}_k} w_{k,v}^H E_{k,v} w_{k,v}$$

is the effective estimation errors on the channels related this user.

**Proof:** The achievable sum spectral efficiency is obtained, similarly to [44, 45], by decoding the Gaussian information sequences from the different BSs in a sequential manner, using conventional successive interference cancellation. Since the users only know the effective channels $w_{H,k,v}^H \hat{h}_{k,v}$ and not the true channels $w_{l,v}^H h_{k,v}$, the channel uncertainty is handled by computing a lower bound on the mutual information, using the approach from [46] where all signals that are uncorrelated with $w_{k,v}^H \hat{h}_{k,v}$ are treated as Gaussian noise (which is the worst case in terms of mutual information).

This lemma provides an achievable lower bound on the capacity, since the latter is unknown under imperfect CSI. We note that Lemma 1 assumes that the users know the effective precoded channels. In practice, the users can estimate these effective channels using downlink pilots, and get estimates of $w_{l,v}^H \hat{h}_{k,v}$ that are at least as accurate $w_{l,v}^H \hat{h}_{k,v}$. Hence, it might be possible to achieve higher spectral efficiencies than in Lemma 1. Nevertheless, the aggregated SINR in (6) is the most convenient one for precoding design, since the BSs can only utilize their own CSI in the optimization.

### F.2.3 Problem Formulation

The focus of this paper is on the joint design of load balancing (i.e., the UE association in $\mathcal{U}_v$) and precoding vectors ($w_{k,v}$) for $v = 1, \ldots, M$ and $k = 1, \ldots, K$, which is an optimization that takes place at every channel realization. To this end, the goal is to minimize the weighted total power consumption (for any given channel realization) while satisfying a set of SINR constraints (or, equivalently, spectral efficiency constraints) for

1\footnote{This is the total power consumption.}
each user and a set of transmit power constraints for each BS. These constraints are referred to as the QoS constraints. With (4), (5) and (6) in hand, the optimization problem can be formulated as

\[
\min_{\{U_v, \{w_{k,v}\}\}} \sum_{v=1}^{M} a_v P_v \\
\text{subject to } \gamma_k \geq \Gamma_k, \forall k \\
P_{\text{trans},v} \leq P_{v,\max}, \forall v
\]  

(9)

where \( \Gamma_k > 0 \) is the target SINR value for user \( k \). By satisfying this QoS target for every channel realization, the ergodic spectral efficiency is \( R_k \geq \log_2 (1 + \Gamma_k) \). The weights \( a_v > 0 \) are used to balance the power consumptions of different BSs. For the rest of the paper, we assume that the problem (9) has at least one feasible solution, which is reasonable in dense networks with an over-provisioning of access points. In practice, if no feasible solution exists, the SINR constraints have to be relaxed either by decreasing the target SINRs or by removing users [24].

**F.3 Optimal Precoding and Load Balancing**

In this section, we solve the optimization problem in (9). As a first step, we show that the set variables \( U_v \) can be eliminated by optimizing over all precoding vectors.

**Lemma 2** The original problem (9) is equivalent to

\[
\min_{\{w_{k,v}\}} \sum_{v=1}^{M} a_v P_v \\
\text{subject to } \gamma_k \geq \Gamma_k, \forall k \\
\sum_{k=1}^{K} \|w_{k,v}\|^2 \leq P_{v,\max}, \forall v
\]

(10)

where \( P_v \) can be rewritten as

\[
P_v = \begin{cases} 
N_v P_{\text{active},v} + \Delta_v \sum_{k=1}^{K} \|w_{k,v}\|^2, & 0 < P_{\text{trans},v} \leq P_{v,\max} \\
N_v P_{\text{sleep},v}, & P_{\text{trans},v} = 0
\end{cases}
\]

(11)

and \( \gamma_k \) is reformulated as

\[
\gamma_k = \frac{\sum_{v=1}^{M} \left| \hat{h}_{k,v}^H w_{k,v} \right|^2}{I_k + E_k + \sigma_k^2}
\]

(12)

with \( I_k \) rewritten as

\[
I_k = \sum_{v=1}^{M} \sum_{l=1}^{K} w_{l,v}^H \left( \hat{h}_{k,v} \hat{h}_{k,v}^H + E_{k,v} \right) w_{l,v}
\]

(13)

and \( E_k \) replaced by

\[
E_k = \sum_{v=1}^{M} w_{k,v}^H E_{k,v} w_{k,v},
\]

(14)
Proof: Note that if BS $j$ does not serve a particular user $k$ (i.e., $k \not\in U_j$ and $j \not\in V_j$), then all terms that would have contained $w_{k,j}$ in the SINR of (6) and the transmit power (4) are missing. This is equivalent to setting $w_{k,j} = 0$ and adding said terms (which then are zero). Hence, the sets $U_v$ and $V_k$ are fully determined by checking which of the precoding vectors are zero:

$$U_v = \{ k | w_{k,v} \neq 0, k \in \{1, \ldots, K\} \},$$

$$V_k = \{ v | w_{k,v} \neq 0, v \in V \}.$$  \hspace{1cm} (15)

The sets $U_v$ can therefore be removed as optimization variables from (9), if we add the missing terms in (4) and (6). The corresponding equivalent problem is the one stated in this lemma.

This lemma shows that we do not need to optimize the BS association sets $U_v$ since these are implicitly determined by checking which precoding vectors that are non-zero. Note that although the expressions for $P_v$, $\gamma_k$, $I_k$, and $E_k$ in Lemma 2 are different from the expressions in Section F.2, the values are identical for every selection of precoding vectors $\{w_{k,v}\}$. As will be shown later, even if all BSs are allowed to transmit to all users at the same time-frequency resource block, in most cases, at the optimal point, each user $k$ will be connected to only one BS.

The optimization problem (10) is not convex. In particular, the power consumption function in (11), which is in the form of a fixed transaction cost function, leads to a hard combinatorial problem [47]. Moreover, the SINR constraints of (10) do not have a standard convex form. In the following, we first show that, for each combination of BS modes (active or sleep), problem (10) can be reformulated as a convex problem. Then, the global optimum can be found by solving this convex problem for all $2^M$ combinations of modes.

Define $w_k \triangleq [w_{k,1}^T, w_{k,2}^T, \ldots, w_{k,M}^T] \in \mathbb{C}^{(\sum_{v=1}^M N_v) \times 1}$ as the aggregated precoding vector for user $k$ from all BSs. We notice that the received SINR, $\gamma_k$ in (12), can be expressed as

$$\gamma_k = \frac{w_k^H \hat{R}_k w_k}{\sum_{j=1}^K w_j^H (\hat{R}_j + E_j) w_j + w_k^H E_k w_k + \sigma_k^2},$$

using the block-diagonal matrices

$$E_k \triangleq \text{diag} \left( E_{k,1}, E_{k,2}, \ldots, E_{k,M} \right)$$

$$\hat{R}_k \triangleq \begin{bmatrix}
\hat{R}_{k,1} & 0 & \cdots & 0 \\
0 & \hat{R}_{k,2} & \ddots & \vdots \\
\vdots & 0 & \ddots & 0 \\
0 & \ddots & 0 & \hat{R}_{k,M} 
\end{bmatrix}$$

with the diagonal blocks $\hat{R}_{k,v} \triangleq \hat{h}_{k,v} \hat{h}_{k,v}^H \in \mathbb{C}^{N_v \times N_v}$ for $v = 1, \ldots, M$.

Similarly, the power constraints in (10) are written in terms of $w_k$ as $\sum_{k=1}^K w_k^H Q_v w_k$, where

$$Q_v \triangleq \text{diag} \left( Q_{1,v}, Q_{2,v}, \ldots, Q_{M,v} \right)$$

(19)
With this notation, the optimization problem (10) looks like a classical precoding optimization problem of the type in [48], but with the important difference that \( \hat{R}_k \) has rank \( M \) and not rank 1 as in the case with one BS per user. Hence, we cannot use the second-order cone techniques from [48], but the following semi-definite relaxation approach.\(^2\) Semi-definite relaxation means that the optimization variables changed to \( W_k \triangleq w_k w_k^H \succeq 0 \) instead of \( w_k \). This would require an additional rank constraint, rank (\( W_k \)) = 1, \( \forall k \), but this one is dropped as a relaxation.

**Lemma 3** Let \( z_v \) be the BS mode indicator for \( v \in \mathcal{V} \): \( z_v = 1 \) if BS \( v \) is active, and \( z_v = 0 \) if BS \( v \) is in sleep mode. Consider the following semi-definite relaxation of (10) for fixed BS modes:

\[
\begin{align*}
\text{minimize} & \quad \sum_{k=1}^K \text{Tr}(A W_k) + J(z) \\
\text{subject to} & \quad \text{Tr}(\hat{R}_k W_k) - \Gamma_k \sum_{l=1, l \neq k}^K \text{Tr}\left((\hat{R}_k + E_k) W_l\right) \\
& \quad - \Gamma_k \text{Tr}(E_k W_k) \geq \Gamma_k \sigma_k^2, \forall k \\
& \quad \sum_{k=1}^K \text{Tr}(Q_v W_k) \leq z_v P_v, \max, \forall v
\end{align*}
\]

(21)

where \( z_v \in \{0, 1\}, \forall v \) and

\[
\begin{align*}
A & \triangleq \text{diag}(a_1 \Delta_1 I_{N_1}, a_2 \Delta_2 I_{N_2}, \ldots, a_M \Delta_M I_{N_M}) \\
J(z) & = \sum_v a_v N_v \left(P_{\text{active},v} z_v + P_{\text{sleep},v} (1 - z_v)\right).
\end{align*}
\]

(22) (23)

The problem (21) is a convex semi-definite program and it always has a rank one solution, if the problem is feasible.

**Proof:** For any fixed combination of BS modes \( z = [z_1, \ldots, z_M] \), \( J(z) \) in (23) is fixed. Then, the problem (21) is on the form of (P2) in [49]. Based on [49, Theorem 1], this type of optimization problems always has optimal solutions with rank one if it is feasible.

\[\blacksquare\]

Based on this lemma, we solve the original precoding and load balancing problem as follows.

**Theorem 1** The global optimum to (9) is obtained by solving (21) for each of the \( 2^M \) mode combinations (\( z_v = 0 \) or \( z_v = 1 \) for each \( v \)) and selecting the solution that provides the lowest weighted total power consumption.
To summarize, Lemma 3 shows that semi-definite relaxation is tight for the problem at hand. For each fixed mode $z$, we can solve (21) using standard convex optimization software, such as CVX [50] or YALMIP [51]. By doing this for all $2^M$ mode combinations, the global optimum to (9) is obtained. We stress that (9) optimizes the precoding, load balancing (i.e., BS association), and BS modes jointly. When the number of BSs ($M$) is large, the exhaustive search can be improved by using branch and bound methods to reduce the computational complexity [47]. The global optimum to (9) is a benchmark for any suboptimal heuristic load-balancing and precoding algorithms; for example, the ones proposed in [5, 6, 13–15, 17, 19, 20, 22].

F.3.1 Structure of the Optimal Load Balancing

Theorem 1 shows how to solve the joint precoding and load balancing optimization problem (9) using convex optimization techniques. Although it provides the truly optimal solution, it brings little insight on the structure of the optimal load balancing. In the following, we will analyze the dual problem of (21) and thereby shed light on the optimal BS association.

Recall from Lemma 3 that (21) is a semi-definite optimization problem. This means that it is a convex problem that satisfies Slater’s constraint qualification, which implies strong duality [52, Sec. 5.2.3]. The dual problem has the same optimal objective value as the original problem. Define $\mathcal{A} \triangleq \{v|z_v = 1, v \in V\}$ as the set of active BSs, and $\mathcal{S} \triangleq \{v|z_v = 0, v \in V\}$ as the set of BSs in the sleep mode. The Lagrangian of (21) is

$$\mathcal{L} (\{W_k, \lambda_k, \mu_i, \nu_j\})$$

$$= J (z) + \sum_{k=1}^{K} \text{Tr} (AW_k) - \sum_{k=1}^{K} \lambda_k \text{Tr} (\hat{R}_k W_k)$$

$$+ \sum_{k=1}^{K} \lambda_k \Gamma_k \sum_{l=1}^{K} \text{Tr} \left( (\hat{R}_k + E_k) W_l \right)$$

$$+ \sum_{k=1}^{K} \lambda_k \Gamma_k \left( \text{Tr} (E_k W_k) + \sigma_k^2 \right)$$

$$+ \sum_{i \in \mathcal{A}} \mu_i \left( \sum_{k=1}^{K} \text{Tr} (Q_i W_k) - P_{i,\text{max}} \right)$$

$$+ \sum_{j \in \mathcal{S}} \nu_j \sum_{k=1}^{K} \text{Tr} (Q_j W_k)$$

(24)

where $\lambda_k, \mu_i, \nu_j \geq 0$ are the Lagrange multipliers associated to the $k$-th user’s SINR constraint, the power constraint for BS $i$ in set $\mathcal{A}$, and the power constraint for BS $j$ in set $\mathcal{S}$, respectively. The dual problem to (21) is an unconstrained maximization of the dual function, defined as

$$g (\{\lambda_k, \mu_i, \nu_j\}) = \minimize_{\{W_k\}} \mathcal{L} (\{W_k, \lambda_k, \mu_i, \nu_j\}).$$

(25)
Define
\[ B_k \triangleq A + \lambda_k \Gamma_k E_k + \sum_{l=1}^{K} \lambda_l \Gamma_l (\hat{R}_l + E_l) + \sum_{i \in S} \mu_i Q_i + \sum_{j \in S} \nu_j Q_j, \] (26)
which is a block-diagonal matrix the \( \nu \)-th block is
\[ B_{k,\nu} \triangleq a_v \Delta_v I_{N_v} + \lambda_k \Gamma_k E_{k,\nu} + \sum_{l=1}^{K} \lambda_l \Gamma_l (\hat{R}_{l,\nu} + E_{l,\nu}) \]
\[ + \sum_{i \in A} \mu_i Q_{i,\nu} + \sum_{j \in S} \nu_j Q_{j,\nu}. \] (27)

From (24), it is easy to show that \( g(\{\lambda_k, \mu_i, \nu_j\}) = J(z) + \sum_{k=1}^{K} \lambda_k \Gamma_k \sigma_k^2 - \sum_{i \in S} \mu_i P_{i,\text{max}}, \) if \( B_k - \lambda_k \hat{R}_k \succeq 0 \) for all \( k = 1, \ldots, K; \) otherwise, \( g(\{\lambda_k, \mu_i, \nu_j\}) = -\infty. \) Hence, the dual problem of (21) becomes
\[
\begin{align*}
\text{maximize} & \quad J(z) + \sum_{k=1}^{K} \lambda_k \Gamma_k \sigma_k^2 - \sum_{i \in S} \mu_i P_{i,\text{max}} \\
\text{subject to} & \quad B_k - \lambda_k \hat{R}_k \succeq 0, \forall k. 
\end{align*}
\] (28)

**Lemma 4** Let \( \{\lambda^*_k, \mu^*_i\} \) denote the optimal Lagrange multipliers to (28), and let \( B^*_{k,\nu} \) be the value of \( B_{k,\nu} \) in (27) for these multipliers. The optimal precoding vectors are
\[
w^*_{k,\nu} = \begin{cases} \alpha_{k,\nu} (B^*_{k,\nu})^{-1} \hat{h}_{k,\nu}, & \text{if } \lambda^*_k = \frac{1}{\hat{h}^H_{k,\nu}(B^*_{k,\nu})^{-1} \hat{h}_{k,\nu}}, \\ 0, & \text{otherwise}, \end{cases}
\] (29)
where \( \alpha_{k,\nu} \geq 0 \) is a scaling factor.

**Proof:** Since for any fixed \( z, \) strong duality holds for (21) and the solution has rank one as \( W^*_k = w^*_k(w^*_k)^H, \) the optimal \( w^*_k \) can be calculated by setting the first-order derivative of the Lagrangian in (24) with respect to \( w_k \) to zero; that is,
\[
\frac{\partial L(\{W_k, \lambda^*_k, \mu^*_i, \nu^*_j\})}{\partial w_k} |_{w_k^*} = 0
\] (30)
from which we have the condition
\[
B^*_{k,\nu} w^*_{k,\nu} = \lambda^*_k \hat{h}^H_{k,\nu} \hat{h}^H_{k,\nu} w^*_{k,\nu}, \quad \forall \nu.
\] (31)
Hence,
\[
w^*_{k,\nu} = \alpha_{k,\nu} (B^*_{k,\nu})^{-1} \hat{h}_{k,\nu}
\] (32)
for all \( k \) and \( \nu, \) where \( \alpha_{k,\nu} \triangleq \lambda^*_k \hat{h}^H_{k,\nu} w^*_{k,\nu} \) is a scalar. If we now multiply (30) by \((w^*_k)^H\) from the left, we obtain the equivalent condition
\[
2(w^*_k)^H (B^*_k - \lambda^*_k \hat{R}_k) w^*_k = 0 \iff
\]
\[
2(w^*_{k,\nu})^H (B^*_{k,\nu} - \lambda^*_k \hat{h}^H_{k,\nu}) w^*_{k,\nu} = 0
\] (33)
By plugging (32) into (33), we obtain the condition

$$2 \alpha_{k,v} \left( \hat{h}_{Hk,v} \left( B_{k,v}^* \right)^{-1} \hat{h}_{k,v} - \lambda_k \left( \hat{h}_{Hk,v} \left( B_{k,v}^* \right)^{-1} \hat{h}_{k,v} \right)^2 \right) = 0$$  \hspace{1cm} (34)$$

which is satisfied when either \( \lambda_k \hat{h}_{k,v} \hat{h}_{Hk,v}^* w_{k,v}^* = 1 \) or \( \alpha_{k,v} = 0 \). These two cases correspond to the two cases in (29).

Lemma 4 gives the structure of the optimal precoding vectors. In particular, it helps us to understand the optimal BS association (i.e., which precoding vectors \( w_{k,v}^* \) that are non zero).

**Theorem 2** The optimal BS association for user \( k \) falls into one of the following two cases:

1. It is only served by one BS \( v^* \), with \( v^* = \arg \max_v \left( \hat{h}_{Hk,v} \left( B_{k,v}^* \right)^{-1} \hat{h}_{k,v} \right) \), that is, \( \mathcal{V}_k = \{ v^* \} \);

2. It is served by a set of BSs \( \mathcal{V}_k = \left\{ v^* \mid v^* = \arg \max_v \left( \hat{h}_{Hk,v} \left( B_{k,v}^* \right)^{-1} \hat{h}_{k,v} \right) \right\} \) where \( |\mathcal{V}_k| > 1 \).

**Proof:** We know from (29) in Lemma 4 that user \( k \) is associated with BSs \( v \) only if

$$\lambda_k^* = \frac{1}{\hat{h}_{Hk,v} \left( B_{k,v}^* \right)^{-1} \hat{h}_{k,v}} \hspace{1cm} (35)$$

Dual feasibility requires that \( B_k - \lambda_k \hat{R}_k \succeq 0 \) for all \( k \), or equivalently that \( u_{k,v}^H \left( B_{k,v}^* - \lambda_k^* \hat{R}_{k,v} \right) u_{k,v} \geq 0 \) for all vectors \( u_{k,v} \). By selecting \( u_{k,v} = \left( B_{k,v}^* \right)^{-1} \hat{h}_{k,v} \), this conditions becomes

$$\left( \hat{h}_{Hk,v} \left( B_{k,v}^* \right)^{-1} \hat{h}_{k,v} - \lambda_k^* \left( \hat{h}_{Hk,v} \left( B_{k,v}^* \right)^{-1} \hat{h}_{k,v} \right)^2 \right) \geq 0$$

$$\Rightarrow \lambda_k^* \leq \frac{1}{\hat{h}_{Hk,v} \left( B_{k,v}^* \right)^{-1} \hat{h}_{k,v}}. \hspace{1cm} (36)$$

Hence, the equality in (35) can only be achieved for the BSs that have the largest value on \( \hat{h}_{Hk,v} \left( B_{k,v}^* \right)^{-1} \hat{h}_{k,v} \). This can be one or multiple BSs, as reflected by the theorem.

Theorem 2 proves that single-BS association is optimal in most cases, although our system model supports spatial multifold transmission from multiple BSs (a similar result was obtained in [18] in for single-antenna BSs). The optimal BS association for user \( k \) is the one with the largest value of \( \hat{h}_{Hk,v} \left( B_{k,v}^* \right)^{-1} \hat{h}_{k,v} \). We notice that \( B_{k,v}^* \) in (27) is the weighted sum of several terms; the spatial directions of interfering channels, the noise variance, the channel uncertainty, and the matrices from the power constraints. These terms are weighted by the different Lagrange multipliers, which means that the QoS and power constraints that are hard to satisfy will have a large impact on \( B_{k,v}^* \) and vice versa.

The BS association rule is based on the norm of the channel \( ||\hat{h}_{Hk,v}||^2 \) from BS \( v \), which is
then weighted through $B_{k,v}^*$. The weighing will punish BSs with smaller power budget, lower estimation quality, and/or many users with high QoS targets.

As seen from Case 2 in Theorem 2, it may happen that multiple BSs are associated with a certain user. This occurs when the most appropriate BS does not have the power resources to satisfy the QoS target, thus another BS needs to help out. This result stands in contrast to [24] where single-BS association always occurs since there are no power constraints. The probability of multi-BS association is evaluated in Section F.5.

The optimal BS association rule is clearly a complicated function of the channel quality, estimation quality, power constraints, and QoS constraints. This stands in contrast to heuristic association rules (e.g., the nearest-BS based, the highest-received-power based, the max-SINR based, the biased-received-power based and the biased-SINR based load balancing criteria), which are generally not optimal in terms of maximizing the energy efficiency under per-BS transmit power constraints and per-user QoS constraints. These heuristic association rules have been studied under various conditions (different from our system model); see for example [5, 6, 13–15, 19–22, 32]. Hopefully, these heuristics can evolve in future works, based on insights on the optimal BS association from Theorem 2.

### F.4 Iterative Heuristic Algorithm Design

In this section, we tackle the non-convex problem (9) by iterative convex approximations of the power consumption functions. In particular, each iteration solves a problem with a modified objective function, which is convex. This convex objective function is updated in each iteration such that most of the BSs with small transmit powers in the solution are driven to sleep mode. The proposed algorithm will find a suboptimum to the original problem in (9).

Note that $0 \leq P_{\text{trans},v} \leq P_{v,\text{max}}$ for each BS $v, v \in \mathcal{V}$. Thus, the total consumed power of BS $v$, $P_v$ in (5), can be relaxed with its convex envelope, $P_v^{\text{c.e.}}$ over the interval $[0, P_{v,\text{max}}]$, where

$$P_v^{\text{c.e.}}(P_{\text{trans},v}) \triangleq N_v P_{\text{sleep},v} + \Delta'_{v} P_{\text{trans},v}$$

(37)

with

$$\Delta'_{v} \triangleq N_v \frac{(P_{\text{active},v} - P_{\text{sleep},v})}{P_{v,\text{max}}} + \Delta_v$$

(38)

which is the largest convex function smaller than or equal to $P_v$ over the interval. Replacing $P_v$ with $P_v^{\text{c.e.}}$, problem (10) and (9) are relaxed to

$$\begin{align*}
\text{minimize} & \quad \sum_{v=1}^{M} a_v P_v^{\text{c.e.}} \\
\text{subject to} & \quad \gamma_k \geq \Gamma_k, \forall k \\
& \quad \sum_{k=1}^{K} \|w_{k,v}\|^2 \leq P_{v,\text{max}}, \forall v.
\end{align*}$$

(39)

The idea, which is based on replacing an indicator function of a bounded variable with its convex envelope, is often referred to as the $l_1$-norm relaxation, where sparse solutions can be obtained. The relaxed problem (39) can be reformulated as a convex optimization problem.
minimize \( \sum_{k=1}^{K} \text{Tr} \left( \mathbf{A}' W_k \right) + \sum_{v=1}^{M} a_v N_v P_{\text{sleep},v} \)

subject to \( \text{Tr} \left( \hat{\mathbf{R}}_k W_k \right) - \Gamma_k \sum_{l=1}^{K} \text{Tr} \left( (\hat{\mathbf{R}}_k + \mathbf{E}_k) W_l \right) \)

\[ - \Gamma_k \text{Tr} (\mathbf{E}_k W_k) \geq \Gamma_k \sigma_k^2, \forall k \]

\[ \sum_{k=1}^{K} \text{Tr} \left( \mathbf{Q}_v W_k \right) \leq P_{v,\text{max}}, \forall v \]  \( (40) \)

where \( \mathbf{A}' \) is a modified block diagonal matrix of \( \mathbf{A} \), with \( \Delta_v \) replaced by \( \Delta'_v \) for each block \( v \). Note that based on Lemma 3, the rank-one constraints are dropped without loss of optimality. Compared to the original problem (10), the relaxed problem (40) has the same feasible set, but a modified objective function. The optimal value of (40) is a lower bound on the optimal value of the original problem (10).

The proposed iterative heuristic algorithm is as follows:

1. \( i := 0 \); Initialize \( W_k^{(0)} \) for \( k = 1, \ldots, K \) by solving the relaxed convex problem (40).

2. \( i := i+1 \); Obtain the transmit power of each BS \( v \) as \( P_{\text{trans},v}^{(i-1)} = \sum_{k=1}^{K} \text{Tr} \left( \mathbf{Q}_v W_k^{(i-1)} \right) \).

Define \( \hat{P}_v^{(i)} (\text{trans},v) \triangleq N_v P_{\text{sleep},v} + \Delta_v^{(i)} P_{\text{trans},v} \), where

\[ \Delta_v^{(i)} \triangleq \frac{N_v (P_{\text{active},v} - P_{\text{sleep},v})}{P_{\text{trans},v}^{(i-1)} + \delta} + \Delta_v. \]  \( (41) \)

Solve the modified optimization problem

\minimize_{\{W_k \succeq 0\}} \sum_{k=1}^{K} \text{Tr} \left( \mathbf{A}^{(i)} W_k \right) + \sum_{v=1}^{M} a_v N_v P_{\text{sleep},v} \)

subject to \( \text{Tr} \left( \hat{\mathbf{R}}_k W_k \right) - \Gamma_k \sum_{l=1}^{K} \text{Tr} \left( (\hat{\mathbf{R}}_k + \mathbf{E}_k) W_l \right) \)

\[ - \Gamma_k \text{Tr} (\mathbf{E}_k W_k) \geq \Gamma_k \sigma_k^2, \forall k \]

\[ \sum_{k=1}^{K} \text{Tr} \left( \mathbf{Q}_v W_k \right) \leq P_{v,\text{max}}, \forall v \]  \( (42) \)

where \( \mathbf{A}^{(i)} \) is the modified block diagonal matrix of \( \mathbf{A} \), with \( \Delta_v \) replaced by \( \Delta_v^{(i)} \) for each block \( v \).

3. Let \( W_k^{(i)} \) be the solution to this problem.

4. If \( P_{\text{trans},v}^{(i-1)} \) and \( P_{\text{trans},v}^{(i)} \) are approximately\(^3\) equal for each \( v \), return \( W_k := W_k^{(i)} \).

Otherwise, go back to step 2).

\(^3\)There are many different ways to define “approximately equal”, such as \( \max_v \left| P_{\text{trans},v}^{(i-1)} - P_{\text{trans},v}^{(i)} \right| \leq \varepsilon \) and \( \sum_{v=1}^{M} \left| P_{\text{trans},v}^{(i-1)} - P_{\text{trans},v}^{(i)} \right| \leq \varepsilon \). The latter is used as a stopping criterion in our simulation with \( \varepsilon = 10^{-6} \).
Note that $\delta$ in (41) is a non-negative small value, which can be interpreted as a soft threshold for deciding when a BS is set to sleep mode. Define $P^*_{\text{trans},v} \triangleq \sum_{k=1}^{K} \text{Tr} (Q_v W^*_k)$. Thus, for $P^*_{\text{trans},v} \gg \delta$, we have $\hat{P}_v (P^*_{\text{trans},v}) \triangleq N_v P_{\text{sleep},v} + \left( \frac{N_v (P_{\text{active},v} - P_{\text{sleep},v})}{P^*_{\text{trans},v} + \delta} + \Delta_v \right) P^*_{\text{trans},v} \approx N_v P_{\text{active},v} + \Delta_v P^*_{\text{trans},v} = P_v (P^*_{\text{trans},v})$, and BS $v$ is in the active mode; while for $P^*_{\text{trans},v} = 0$, $\hat{P}_v (P^*_{\text{trans},v}) \triangleq N_v P_{\text{sleep},v}$ and BS $v$ is under the sleep mode.

For each iteration as shown in step 2), when $P_{\text{trans},v}^{(i-1)}$ is small, the modified $\Delta_v^{(i)}$ in (41) becomes large, i.e., the derivative of the power consumption function $\hat{P}_v (P_{\text{trans},v})$ increases. Therefore, the modified optimization problem (42) will push the smalls $P_{\text{trans},v}^{(i-1)}$ to zero; that is, the BSs with small transmit powers in the solution to the previous problem are driven to sleep mode. This leads to sparse solutions of $W^*_k$.

**Lemma 5** The proposed iterative heuristic algorithm always converges.

**Proof:** The modified objective function of problem (42) is on the form of the objective function in [47, Eq. (21)], which always gives convergence; that is, with $0 \leq P_{\text{trans},v} \leq P_{v,\text{max}}$ a convex, compact set, and $\delta > 0$, we can show that $P_{\text{trans},v}^{(i)} - P_{\text{trans},v}^{(i-1)} \to 0$ for $v = 1, \ldots, M$. A proof of convergence for this type of heuristic algorithms is given in [47, Appendix B].

Note that, upon convergence, the partial derivative with respect to $P_{\text{trans},v}$ of the function minimized in the last iteration is given by

$$\frac{N_v (P_{\text{active},v} - P_{\text{sleep},v})}{P^*_{\text{trans},v} + \delta} + \Delta_v,$$

which is equal to the derivative of the function

$$f (P_{\text{trans},v}) = \sum_{v=1}^{M} \alpha_v \log (P_{\text{trans},v} + \delta) + \sum_{v=1}^{M} \Delta_v P_{\text{trans},v}$$

at $P_{\text{trans},v} = P^*_{\text{trans},v}$, where $\alpha_v \triangleq N_v (P_{\text{active},v} - P_{\text{sleep},v})$. From the equality of the first-order conditions for optimality, we see that the iterative procedure finds a local minimum of $f (P_{\text{trans},v})$. The log-sum function $\sum_{v=1}^{M} \alpha_v \log (P_{\text{trans},v} + \delta)$ is used as a smooth surrogate for the circuit power consumption part of the objective function. Therefore, our proposed heuristic algorithm is very similar to the weighted $l_1$-norm minimization methods, where the weighting factors are chosen based on the log-sum surrogate function of the $l_0$-norm [53].

**F.5 Numerical Results**

Numerical results are presented in this section to illustrate our analytical results and the proposed algorithms. The purpose of this section is not to provide a large-system analysis, but to compare the heuristic algorithm from Section F.4 with the optimal solution from Theorem 1, for which the complexity of mode selection grows quickly with the number of BSs.
The propagation environment is a simplified version of the dense urban information society model (TC2) used in the METIS project [54], as illustrated in Fig. F.2. The model consists of four square-shaped buildings of dimensions $120 \times 120$ m, each with 6 floors. A macro BS (MBS) is complemented with 4 small cell BSs (SBSs). The MBS has 4 transmit antennas, and the SBSs have 2 transmit antennas each. Load-balancing is particularly important in the lightly loaded cases that occur during the majority of the day [8], because then there is an opportunity to turn off BSs and associate users with other BSs than the closest one. Hence, in most of the simulations, we consider five users that are randomly and uniformly dropped in the network, whereof 4 users are indoors and 1 user is outdoors in every user drop. The system bandwidth is 10 MHz. Here, we adopt the indoor and outdoor propagation models, PS#1 - PS#4, identified in METIS. More details regarding network deployment and propagation modes can be found in [54, Table 3.7 and Section 8.1]. We assume independent Rayleigh small-scale fading. The MMSE channel estimation errors are calculated based on (1) with the total pilot power $p = P_{v,\text{max}}/2$. Table F.1 shows the power model parameters and is based on [10, Table 6 and Table 8].

Three different joint precoding and load balancing schemes are compared in the scenario depicted in Fig. F.2. We name these three schemes as “Optimal”, “Heuristic” and “All Active” respectively. The “Optimal” scheme obtains the global optimal solution as described in Theorem 1, by an exhaustive search over all $2^5$ possible BS mode combinations. The “Heuristic” scheme follows the algorithm proposed in Section F.4, and the value of the soft threshold $\delta$ is set to $10^{-4}$. The “All Active” scheme is used as our performance baseline, which solves the optimization problem (9) by assuming that all BSs are active, i.e., the BS mode indicator $z_v = 1$ for all BSs $v \in \mathcal{V}$. For each scheme, the performance is averaged over 1000 independent user drops that provide feasible solutions for our optimization problem (9). For each user drop, the algorithms are evaluated over 50 independent channel realizations. The weights $a_v$ are set to 1 for all BSs.

Define the dynamic part of the total power consumption as the total RF power (i.e., $\sum_{v=1}^{M} a_v \Delta_v P_{\text{trans},v}$), and the remaining part of the total power consumption as the circuit power (i.e., $\sum_v a_v N_v P_{\text{active},v} z_v + \sum_v a_v N_v P_{\text{sleep},v} (1 - z_v)$). Figs. F.3 and F.4 demonstrate the total RF power and the total power consumption as a function of target spectral
Table F.1: Power model parameters for different BS types.

<table>
<thead>
<tr>
<th>BS type</th>
<th>$N_v$</th>
<th>$P_{v,\text{max}}$ [W]</th>
<th>$P_{\text{sleep},v}$ [W]</th>
<th>$P_{\text{active},v}$ [W]</th>
<th>$\Delta_v$</th>
</tr>
</thead>
<tbody>
<tr>
<td>MBS</td>
<td>4</td>
<td>39.8</td>
<td>75.0</td>
<td>130.0</td>
<td>4.7</td>
</tr>
<tr>
<td>SBS</td>
<td>2</td>
<td>6.3</td>
<td>39.0</td>
<td>56.0</td>
<td>2.6</td>
</tr>
</tbody>
</table>

Figure F.3: Total RF power (the dynamic part $\sum_{v=1}^{M} a_v \Delta_v P_{\text{trans},v}$) vs. target spectral efficiency per user ($R_k$). 

Efficiency per user, respectively. As expected, the total power consumption and the RF power increase as the target spectral efficiency increases. Fig. F.3 shows that the RF power for the “All Active” scheme is less than that of the “Heuristic” and “Optimal” schemes. This is expected since all BSs are active in the “All Active” scheme, whereas for the “Heuristic” and “Optimal” schemes, some BSs are put into sleep mode. With more BSs being active, the “All Active” scheme provides better energy-focusing and less propagation losses between the users and the transmitters, and will therefore reduce the total RF power. However, as can be seen from Fig. F.4, compared to the “All Active” scheme, the “Heuristic” and “Optimal” schemes can substantially reduce the total power consumption, especially when the target QoS is small. This is because the circuit power consumption under the sleep mode is much lower compared to the one under the active mode, i.e., $P_{\text{sleep},v} \ll P_{\text{active},v}$. For the “All Active” scheme, the increase in the circuit part from the extra power consumed by activating BSs clearly outweighs the decrease in the dynamic part. This implies that putting a BS into sleep mode by proper load balancing is an important solution for energy savings in heterogeneous networks.

Fig. F.5 plots the cumulative distribution function (CDF) of the total power consumption for the considered three schemes. The target spectral efficiency per user $R_k$ is 4 bit/s/Hz. We observe that compared to the “All Active” scheme, 20% of the total power consumption can be saved by the “Optimal” scheme with 70% probability and by the “Heuristic” scheme with 55% probability. For some user drops, the energy consumption can be reduced by 30% for both the “Optimal” and “Heuristic” schemes.

Fig. F.6 demonstrates the BS activation probability versus the target spectral efficiency per user. Here, the activation probability of the SBS is averaged over the probabilities of
Figure F.4: Total power consumption ($\sum_{v=1}^{M} a_v P_v$) vs. target spectral efficiency per user ($R_k$).

Figure F.5: The CDF of total power consumption.
the four SBSs depicted in Fig. F.2. We see that for the “All Active” scheme, the activation probabilities of the MBS and SBS is always one, since all BSs are always active in this scheme. Moreover, as anticipated, for both the “Heuristic” and “Optimal” schemes, the BS activation probabilities of the MBS and SBS increase as the target spectral efficiency per user increases. This is because in order to satisfy the raised QoS expectations of all users, the probability that a BS becomes active should increase so as to provide better energy-focusing and less propagation losses. Over the considered range of target spectral efficiency per user, the “Optimal” scheme has lower activation probability for the MBS and higher activation probability for the SBS as compared to the “Heuristic” scheme. Note that the circuit power consumed under the active mode $P_{\text{active},v}$ for the MBS is much higher than that of the SBSs. Thus, as shown in Fig. F.4, the “Optimal” scheme results in better energy saving as compared to the “Heuristic” scheme.

Figs. F.7-F.9 investigate the impact of the ratio $\eta \equiv P_{\text{sleep},v}/P_{\text{active},v}$ on the overall energy efficiency for different schemes. The values of $P_{\text{active},v}$ are fixed to 130W and 56W for the MBS and SBSs respectively. The target spectral efficiency $R_k$ is fixed to 3 bit/s/Hz. In Figs. F.7 and F.8, the total RF power and the total power consumption are plotted as a function of the ratio $\eta$, respectively. It can be seen from Fig. F.7 that the RF power of the “Optimal” and “Heuristic” schemes decreases as the ratio $\eta$ (or equivalently $P_{\text{sleep},v}$) increases, especially when the ratio $\eta$ is large (close to 1). This is because it is better to turn on more BSs, to reduce the RF power, when the difference between the active and sleep modes decreases. The BS activation probability increases more for the “Optimal” scheme, compared to the “Heuristic” scheme. Hence, we observe that the total RF power reduces more significantly for the “Optimal” scheme. From Fig. F.8, we see that the total power consumption increases almost linearly as $\eta$ increases. This is mainly due to the increase of $P_{\text{sleep},v}$.

Although the system allows all BSs to transmit to all users simultaneously at the same time-frequency resource block, Fig. F.9 shows that the probability that a user is served by multiple BSs is less than 4.2% for all the considered schemes over the entire range of $\eta$ when the target spectral efficiency $R_k$ is fixed to 3 bit/s/Hz. Not shown
Here, the joint transmission probability has also been evaluated over different targets of spectral efficiency, i.e., for $R_k = \{1, 2, 3, 4, 5\}$ bits/s/Hz, while the ratio $\eta$ is fixed according to Table F.1. For these cases, simulation shows that the probability of multi-BS joint transmission is less than 4% over the considered range of $R_k$. Fig. F.10 shows the joint transmission probability as a function of the number of users, for a target spectral efficiency of 1 bit/s/Hz. The probability increases with the number of users, since it is harder to satisfy the QoS targets, but it is still in the range of a few percentages. These observations are in line with Theorem 2. From Fig. F.9, we also observe that, for the “Optimal” and “Heuristic” schemes, the joint transmission probability increases as the ratio $\eta$ increases. This is expected since by increasing $\eta$, the BS activation probability increases. Thus, the joint transmission probability also increases. Compared to the “Heuristic” algorithm, the “Optimal” scheme has a lower BS activation probabilities, and therefore it also has a lower joint transmission probability.

## F.6 Conclusions

This paper analyzed the energy efficiency in heterogeneous networks. More specifically, the downlink precoding vectors, load balancing (i.e., user-BS association), and BS operational modes were jointly optimized to minimize the weighted total power consumption. In order to verify how many BSs that should serve a user at the optimal load balancing solution, each user can be served by multiple BSs using spatial multifold transmission. We proved that the optimal BS association rule consists of comparing weighted channel norms, where the weighting matrices depend on channel uncertainty, power constraints and QoS constraints. Moreover we proved that, in most cases, it is optimal for each user to be served by a single BS. Multiple BSs only serve a user when the primary BS does not have the power resources to deliver the full QoS, in which case neighboring BSs can cooperate in order to provide the full QoS. An iterative heuristic algorithm was proposed to find a suboptimal solution with relatively low complexity and it achieves good performance in relation to the optimal scheme. Our numerical results showed that the total
Figure F.8: Total power consumption vs. $\eta = P_{\text{sleep},v}/P_{\text{active},v}$. The target spectral efficiency per user $R_k$ is 3 bit/s/Hz.

Figure F.9: Joint transmission probability vs. $\eta = P_{\text{sleep},v}/P_{\text{active},v}$. The target spectral efficiency per user $R_k$ is 3 bit/s/Hz.
Figure F.10: Joint transmission probability vs. the number of users. The target spectral efficiency per user $R_k$ is 1 bit/s/Hz.

Power consumption can be greatly reduced by putting a BS into sleep mode using proper load balancing.

References


