Joint Optimization of Transmit Beamforming and Base Station Cache Allocation in Multi-Cell C-RAN

by

Mehran Esmaeili

A thesis submitted to the
School of Graduate and Postdoctoral Studies
in partial fulfillment of the requirements for the degree of

Masters of Applied Science in Electrical and Computer Engineering

The Faculty of Engineering and Applied Science

Department of Electrical and Computer Engineering

University of Ontario Institute of Technology(Ontario Tech University)

Oshawa, Ontario, Canada

November 2021

THESIS EXAMINATION INFORMATION

Submitted by: **Mehran Esmaeili**

Masters of Applied Science in **Electrical and Computer Engineering**

Thesis title: Joint Optimization of Transmit Beamforming and Base Station Cache Allocation in Multi-Cell C-RAN

An oral defense of this thesis took place on November 23, 2021 in front of the following examining committee:

Examining Committee:

Chair of Examining Committee :	Dr. Ying Wang
Research Supervisor :	Dr. Shahram ShahbazPanahi
Research Co-supervisor :	Dr. Min Dong
Examining Committee Member :	Dr. Shahryar Rahnamayan
Thesis Examiner:	Dr. Khalil El-Khatib , Professor, Faculty of Business and Information Technology, Ontario Tech University

The above committee determined that the thesis is acceptable in form and content and that a satisfactory knowledge of the field covered by the thesis was demonstrated by the candidate during an oral examination. A signed copy of the Certificate of Approval is available from the School of Graduate and Postdoctoral Studies.

Abstract

Cloud radio access network (C-RAN) with wireless backhaul has been considered as a possible solution for small cell deployment. Limited capacity of wireless backhaul encourages people to equip the base stations (BSs) with cache storages to mitigate peak-time data traffic. We study a downlink C-RAN consisting of a central processor (CP) and multiple cache-enabled BSs connected to the CP through wireless backhaul links. Through joint optimization of beamforming and cache size allocation, we aim to distribute cache sizes among BSs such that long-term delivery time of files is minimized. Assuming zero-forcing (ZF) beamforming is adopted at CP, cache size allocation problem becomes convex which can be solved by convex optimization techniques. Simulation results show that loss of optimality due to ZF assumption is negligible for large antenna sizes at CP. Moreover, the superiority of our proposed caching scheme over several heuristic ones, including proportional caching, is demonstrated by simulation.

Keywords: cloud radio access networks (C-RAN); beamforming; cash allocation; wireless backhaul; zero-forcing (ZF)

Author's Declaration

I hereby declare that this thesis consists of original work of which I have authored. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I authorize the University of Ontario Institute of Technology (Ontario Tech University) to lend this thesis to other institutions or individuals for the purpose of scholarly research. I further authorize University of Ontario Institute of Technology (Ontario Tech University) to reproduce this thesis by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research. I understand that my thesis will be made electronically available to the public.

_____ Mehran Esameili

Statement of Contributions

I hereby certify that I am the sole author of this thesis and that no part of this thesis has been published or submitted for publication. I have used standard referencing practices to acknowledge ideas, research techniques, or other materials that belong to others. Furthermore, I hereby certify that I am the sole source of the creative works and/or inventive knowledge described in this thesis.

Acknowledgements

Foremost, I would like to express my sincere gratitude to my supervisors, Prof. Shahram ShahbazPnahi and Prof. Min Dong, for their patience, continuous support, enthusiasm, and immense knowledge during my MASc study and research.

I would also like to thank my fellow labmates and friends who helped me a lot in both academia and integrating myself into new culture and country, especially Rosa, Yong, Niloofar, Roozbeh, and Ololade.

My spacial thanks goes to Shadi for being always there for me and her continuous support and understanding throughout my research.

Last but not least, I'm extremely grateful to my family for supporting me spiritually throughout my life and giving me the encouragement I needed throughout these years.

Contents

Thesis Examination Information ii

Abstract iii

Author's Declaration iv

Statement of Contributions v

Acknowledgements vi

Table of Contents vii

List of Figures x

List of Tables xi

List of Abbreviations xii

1 Introduction **1**

1.1 Overview 1

1.2 Motivation and Objective 4

1.3 Thesis Contribution 5

1.4 Thesis Organization 7

1.5 Notation 7

2 Literature Review **9**

2.1 Wireless Caching 9

2.2 Multi-Antenna Beamforming 10
2.2.1 QoS and WRM Problems 11
2.2.2 Rate Region 13
2.2.3 ZF Beamforming 14
2.3 Joint Caching and Beamforming 16
2.3.1 BSs-to-EUs Layer 17
2.3.2 CP-to-BSs Layer 17

3 Joint Optimization of Beamforming and Cache Size Allocation 19

3.1 System Model 19
3.2 Optimization at Beamforming Stage 24
3.2.1 Applying ZF Beamforming 29
3.2.2 Lower Bound 41
3.3 Optimization at Caching Stage 42
3.3.1 Cache Allocation for Files with Equal Popularity 44
3.3.2 Cache Allocation for Files with Different Popularities 45

4 Simulation Results 47

4.1 Comparison between SPZF and MPZF Beamforming Schemes 48
4.2 Cache Allocation for Equi-Popular Files 49
4.3 Cache Allocation for Files with Different Popularities 51

5 Conclusions and Future Work 58

6 Appendices 60

6.1 Proof of Lemma 6.1 60

6.2 Proof of Convexity of $\mathcal{R}_{\mathrm{z}}^{i}$ 62

6.3 Proof of Convexity of $T_{\mathrm{sz}}^{\mathrm{o}}(\mathbf{H}, \boldsymbol{\alpha}(\mathbf{C}, \mathbf{d}))$ 63

Bibliography **65**

List of Figures

3.1 A downlink C-RAN consisting of L cells. The BS of each cell is equipped with a cache unit to pre-store a portion of each file available on the cloud. 20

4.1 Average minimum delivery time versus number of CP's antennas, M. 49

4.2 The CDF of the minimum delivery time under different caching strategies with total cache size $C = 10000$ bits and file size $F = 100$ bits. . 51

4.3 Average minimum delivery time for different caching schemes and the lower bound versus number of antennas M. for $C = 1000$ bits, $K = 100$ e $F = 100$ bits, and $\beta = 1$. 53

4.4 CDF of minimum delivery times under different caching strategies with total cache size $C = 1000$ bits, 100 files with different popularity and file size $F = 100$ bits. 54

4.5 Average minimum delivery time achieved by different caching strategies versus β, for $C = 1000$ bits, $K = 10$, $M = 128$, and $F = 100$ bits. . . 56

4.6 Average minimum delivery time achieved by different caching strategies versus C, for $F = 100$ bits, $M = 128$, and $\beta = 1$. 57

List of Tables

4.1 Simulation Parameters 1 48

4.2 Simulation Parameters 2 50

4.3 Simulation Parameters 3 52

4.4 Comparison of Delivery Time (ns) for Different Caching Strategies . . 55

List of Abbreviations

5G	5th generation
BS	base station
CDF	cumulative distribution function
CP	central processor
C-RAN	cloud radio access network
CSI	channel state information
DPC	dirty paper coding
EU	end-user
IC	interference channel
MIMO	multiple-input multiple-output
MISO	multi-input-single-output
MMF	max-min fair
MPZF	multi-phase ZF
MRT	maximum ratio transmission

MU-MIMO	multi-user MIMO
MU-MISO	multi-user MISO
NP-hard	non-deterministic polynomial-time hard
QoS	quality of service
RAM	random access memory
SIMO	single-input-multiple-output
SINR	signal-to-interference-plus-noise ratio
SNR	signal-to-noise ratio
SPZF	single-phase ZF
WRM	weighted rate maximization
ZF	zero-forcing

Chapter 1

Introduction

1.1 Overview

In traditional wireless communication systems, base stations (BSs) were designed to include all of the analog and digital functionalities required for data transmission over wireless links. As a result, such BSs are made up of antenna towers on top of a large package of hardware components. Such a design brings up several issues among which is the underutilization of the resources allocated to each BS. To be more clear, these BSs has to be powerful enough to meet the requirements of the network in case of heavy data traffic. So, most of the time, they are not functioning at their maximum capacity. Moving toward 5th generation (5G) and beyond, researchers have suggested a new architecture, named as cloud radio access network (C-RAN). In a C-RAN, processing burden of BSs are mitigated by allocating parts of their tasks to the cloud. C-RANs with cache equipped BSs are expected to be a suitable solution to meet the ever-increasing demand for higher data rates, especially in content-centric communications [1–3]. A C-RAN consists of a central processor (CP) and multiple BSs which are connected to the CP through wired or wireless backhaul links. In case of small-cell deployment with a large number of cells, using high speed fiber optic

for the backhaul links may not be practical [4]. Hence, the connection between the CP and BSs will have to be established over the wireless medium. Given the data rate limitations of wireless medium, it is obvious that the backhaul links may not be able to handle data traffic during peak times. This issue together with the fact that content-centric communication is becoming more common these days, have led to the consideration of cache-equipped BSs deployment in each cell [5, 6]. As shown in [7], caching the files at the BSs could significantly reduce network traffic congestion, thereby increasing the system performance in terms of downloading rate.

A cache can be defined as a hidden memory in the form of hardware or software that stores a copy or processed version of some data in order to accelerate retrieval of future requests for that data. In the early days of introduction of cache memory, it was used in corporation with central processor units (CPUs) in order to accelerate CPU requests to random access memories (RAMs). Later, caching proved to be usable in different areas such as web page caching, gaming, etc.

During recent years, researchers have suggested to exploit the benefits of cache storages at the edge of wireless communication networks [8–10]. Indeed, studies have shown that storing popular files at the BSs of a cellular network can significantly decrease both the burden on the CP-to-BSs links and the file downloading time of the end-users (EUs). This is mainly due to the fact that the widespread usage of applications that share multimedia, such as social media, can make different files go viral on the internet. Such files will be requested by a lot of EUs for a couple of hours, days or even weeks. Hence, it is reasonable to store these files at the BSs so that there is no need for the BSs to receive these popular files from CP each time a new request

for them comes from EUs. While decreasing downloading time of EUs obviously results in better user experience, the mitigation of the burden on the CP-to-BSs layer becomes more important when someone considers a lot of small-cells deployed around a CP. In this scenario, the fiber optic links are no longer a practical solution for CP-to-BSs layer, and using the wireless medium between CP and BSs is considered a more viable solution. However, as the capacity of wireless links are considerably lower than fiber optic ones, reducing file downloading traffic is important.

Beamforming with a transmitter equipped with multiple antennas is an important physical layer transmission technique widely used in the area of data transmission. Using beamforming, the transmitter can concentrate the transmission energy towards an specific receiver equipment, providing higher data rate and more reliable connection between the aforementioned devices. To be more specific, beamforming could be defined as the design of beamforming weights, under the transmit power constraint, such that the wireless connection between sender and receiver meets user dependent requirements.

In the literature, beamforming design optimization problems are divided into several types based on their design objectives. Some studies try to maximize the quality of user experience given power constraints, while others try to optimize the cost of the system given quality of service requirements. These related beamforming design optimization problems has been considered for single-cell or multi-cell networks, single group or multi group cells and unicast or multicast beamforming designs.

Most of the research conducted in the area of edge caching has been focused on the performance parameters of the BSs-to-EUs layer without considering the backhaul

limitations [11–15]. Also, existing works considering the CP-to-BSs layer assume fixed cache sizes at the BSs [16–18]. Indeed, these works assume some heuristic cache size allocation schemes, based on which they design BS clustering and beamforming weights in order to optimize delivery time or backhaul cost, given minimum signal-to-interference-plus-noise ratio (SINR) requirements or power budget. To the best of our knowledge, the only works that leverage joint optimization of cache size allocation and beamforming at the CP-to-BSs layer to adapt to the long-term statistics of the wireless channels are [7, 19]. This thesis provides contributions to this novel line of research.

1.2 Motivation and Objective

The cache sizes assigned to different BSs for different files need to be optimized in order to achieve the best possible performance for a limited storage capacity. In practical wireless communication systems, such cache allocation design should consider the long-term statistics of the communication links between CP and BSs. While there are many studies considering the performance of downlink beamforming systems in presence of cache-aided BSs [6, 16–18, 20], few have considered joint optimization of cache allocation and beamforming design [7, 19]. In [7], both downloading rate maximization and downloading time minimization of backhaul links are studied for a C-RAN consisting of cooperating cache-equipped BSs serving a common set of EUs. The authors of [19] consider a multi-cell C-RAN with multiple cooperative BSs at each cell, aiming to maximize the sum of downloading rates of backhaul links.

Motivated by the above, we consider a C-RAN consisting of multiple cells, each of

which has a cache-equipped BS serving its own EUs. Our purpose is to jointly design the optimal cache distribution scheme and optimal beamforming strategy such that long-term expected delivery time of the backhaul links is minimized given power and storage budget limitations.

1.3 Thesis Contribution

In this thesis, we consider a downlink C-RAN consisting of a CP and multiple cache-equipped BSs. We aim to design a CP-to-BSs beamforming strategy and a BSs' cache size allocation scheme such that the average downloading time of the BSs is minimized, given certain constraints on total power budget at the CP and total cache budget at the BSs. Because cache size allocation happens at off-pick times and remains unchanged over many channel realizations, the strategy of caching has to be adapted to the long-term statistics of the backhaul links [7].

The main contribution of our thesis is summarized as it follows.

- Considering the two-stage nature of jointly optimal design of CP beamforming weights and cache size allocation, we formulate an optimization problem for each stage.

- In the first stage, we study the CP beamforming design problem with the purpose of minimizing the downloading time of BSs. We formulate CP beamforming design problem under the most general assumption that the CP beamformer can be time-varying. To the best of our knowledge, such a formulation is novel and has not appeared in the prior work published in this field. Then, we prove rigor-

ously that the CP's time-varying beamfrormers can be assumed to be piecewise constant with a maximum number of L pieces, where L is the number of BSs. Being unable to efficiently find these piecewise constant beamformers, we resort to zero-forcing (ZF) beamforming based on which we propose two problems, multi-phase ZF (MPZF) and single-phase ZF (SPZF). While a suboptimal technique is proposed to solve MPZF problem, we provide a closed-form optimal solution for SPZF problem. We also prove that the minimum downloading time resulted from SPZF is a convex function of the cache sizes at the BSs. During simulation, we show that SPZF actually outperforms MPZF in terms of optimality.

- In the second stage, we study the BSs' cache size allocation problem with the purpose of minimizing average downloading time of BSs over a long period of time. Considering the superiority of SPZF over MPZF, we apply the semi-closed form solution of SPZF problem to the cache size optimization problem which results in a convex optimization problem whose cost function is an expectation of delivery time over random channel realizations and file requests of the EUs. Then, we use sample approximation technique to break this expectation into a summation with limited number of terms. Finally, we are able to accurately and efficiently determine the optimum cache sizes by using convex optimization techniques.

- During simulation section, we first show that the SPZF technique results in a lower downloading time than the MPZF for almost every practical antenna

size at the CP. Moreover, we show that the delivery time resulted from SPZF technique is very close to the lower bound delivery time for reasonably large antenna sizes at the CP. Then, we prove the superiority of our optimal cache size allocation scheme through comparing it with several other intuitive techniques, such as uniform and proportional cache size allocation. We also study the effect of changing different parameters, such as cache budget and antenna size of the CP, on the optimal delivery time resulted from different cache size allocation schemes.

1.4 Thesis Organization

The remainder of this thesis is organized as follows. In Chapter 2 a literature review on wireless caching, multi-antenna beamforming and joint optimization of caching and beamforming is given. Chapter 3 introduces the C-RAN system model, formulates the two stages of our optimization problem consisting of beamforming and cache allocation, and finally, connects the two stages of optimization which results into a convex optimization problem that could efficiently be solved by many techniques. Simulation results are covered in Chapter 4. Conclusion and future works are written in Chapter 5.

1.5 Notation

Throughout this thesis, vectors and matrices are denoted by lower-case bold letters and upper-case bold letters, respectively. The set of real and complex numbers are denoted by $\mathbb{R}$ and $\mathbb{C}$, respectively. We use $\mathbb{R}_+$ and $\mathbb{R}_{++}$ to show the set of non-negative

and positive real numbers, respectively. Strict and non-strict element-wise inequality of vectors are denoted by $\succ$ and $\succcurlyeq$, respectively. Kronecker product is denoted by $\otimes$. The expectation of a function $T(\mathbf{H})$ with respect to a random variable $\mathbf{H}$ is denoted as $\mathbb{E}_{\mathbf{H}}[T(\mathbf{H})]$. Calligraphy letters are used to denote sets. Cardinality of set $\mathcal{R}$ is denoted by $|\mathcal{R}|$.

Chapter 2

Literature Review

2.1 Wireless Caching

Before caching concept finds its applications in wireless communication networks, it has already been widely used by computer networks [21–23]. The main difference between computer networks and wireless communication networks, comes from the fact that wired communication links are reliable and fast, most of the time. But, wireless links usually suffer from fading, interference, attenuation, etc., which make such links weak and unstable. So, generally, it is much complicated to design optimum cache allocation schemes for a system with fast changing wireless channels than a system with fixed rate wired links. Such designs often end up being joint optimization of cache distribution and beamforming design (see 2.3). However, early studies in wireless caching usually assume a heuristic pre-known cache allocation scheme based on which they formulate beamforming design problems to optimize some network related criteria. This section is aimed to provide a review on such works.

The authors of [6] consider a C-RAN with given binary cache status at the BSs and exploit BS clustering and multicast beamforming in order to minimize a customized network cost (i.e., the weighted sum of backhaul and transmit power cost),

given required SINR constraints. They also compare the performance of two heuristic caching strategies, named as popularity-aware caching and random caching. Later, the authors of [18] consider the same problem as [6] with large scale C-RAN. They aim to find an algorithm with linear complexity that could be applied to systems with large number of users and BSs. In [16], the authors consider a coded caching scenario and use a semi-definite relaxation framework aiming to minimize a network cost, defined similar to the one used by [6], given SINR constraints as well as power constraints. The authors of [17] design the delivery phase of a network, consisting of a baseband processing unit, enhanced remote radio heads, and EUs, by maximizing delivery rate for fixed-size caches at the remote radio heads. Optimization of the amount of power allocated to cache-aided small-cell BSs of a multi-cell unicast downlink beamforming system is considered by [20]. Its authors have three purposes including maximization of sum-rate, minimizing average outage probability, and maximization of average user satisfaction.

2.2 Multi-Antenna Beamforming

In this section, we give an overview on the beamforming design problems. At first, we discuss two important class of problems named as quality of service (QoS) and weighted rate maximization (WRM). Then, we provide a review on those papers which work on the problem of finding the rate region of a multi-user downlink beamforming system. At the end, we talk about zero-forcing (ZF) beamforming technique.

2.2.1 QoS and WRM Problems

As it was mentioned, one could group beamforming design problems by looking at the objective function of the optimization problem or the system model based on which the optimization problem is formulated. In terms of design objectives, QoS and WRM problems are two important class of problems. A QoS problem minimizes power usage under SINR constraints. WRM problems aim to maximize a general function of received SINR such as sum-rate, minimum rate or minimum SINR, under power budget constraints. At first, QoS and WRM problems were issued in the systems which consider a single cell with multi-user unicast beamforming design. In this regard, [24] solves both QoS and max-min fair (MMF) problems for the general case of downlink beamforming from a multi-antenna transmitter to multi-antenna receivers, referred to as multiple-input-multiple-output (MIMO) system, via conic optimization techniques. Later, [25] proposes a closed-form solution for QoS problem in a multiuser downlink beamforming scenario with single-antenna receivers. Moreover, [25] discusses the structure of the optimal solution of a general transmit beamforming optimization problem whose goal is to maximize a SINR-dependent utility function under power constraints in a multiuser unicast scenario. The authors of [26] consider a multi-cell unicast beamforming system. They analyse the complexity of a group of optimization problems all of which has to maximize a utility function of SINR given power budget constraints. They prove that such a problem for lots of utility functions including sum-rate function is a non-deterministic polynomial-time (NP-hard) problem. A polynomial time algorithm for the case that utility function is minimum

SINR is proposed.

During past decade, the massive increase in the amount of content centric communication have encouraged researchers to consider multicast beamforming as a promising solution to the high demand for reliable and fast data rates. In multicast beamforming, those users who belong to one group request for the same data. So, the BS could serve all of such users with one beam without making any interference among them. A single-group multicast beamforming is the simplest one among the general multi-group multicast beamforming problems. The authors of [27] , while considering a single-group multicast system, have proposed sub-optimal solutions for both QoS and MMF problems. In continuation to the work done by [27] , [28] showed that both QoS and MMF problems are NP-hard. Moreover, in search of more accurate and efficient ways to solve QoS and MMF problems, [29–34] have proposed other strategies each of which has its own pros and cons.

There exists lots of other studies which have worked on QoS and WRM problems in the context of unicast and multicast beamforming. We did not cover all of these papers in this subsection because such problems are not directly related to what we have done in this thesis. Actually, the purpose of this subsection is to provide an overview on some of the early studies dedicated to beamforming design. This may be helpful in understanding the remaining of this thesis.

2.2.2 Rate Region

Characterization of the rate region is an important topic among multi-user beamforming studies. The transmission rate is calculated as follows

$$\text{Rate} = \log(1 + \text{SINR}), \tag{2.1}$$

where SINR is the signal to interference plus noise ratio of the link. The rate region of a network consisting of a BS and multiple users is all the data rate combinations (vectors) that could be supported by the users while the power budget constraints are satisfied. Among all the rate vectors of the rate region, those ones which are located on the boundary, referred to as Pareto boundary [35], are of special importance. It is because they result in the maximum performance, in terms of data rate, of the communication links. Most of the papers that study the rate region of a wireless communication network aim to characterize this Pareto boundary.

One of the earliest works that aims to find general bounds on the capacity region of interference channels (ICs) is [36]. Several years later, [37] proposes outer bounds on the capacity region of two-user multiple-input-single-output (MISO), single-input-multiple-output (SIMO) and MIMO ICs. The authors of [38] consider an n-user IC between n single-antenna transmitters and n single-antenna receivers. Given a maximum power available at each transmitter, [38] proves that the Pareto boundary of the rate region of such a system is the union of n hyper-surface frontiers of dimension $n-1$. They also determine the conditions upon which these frontier surfaces are convex or concave for the special case of two-user IC. In [39], the rate region of a MISO IC consisting of 2 transmitters and 2 receivers with mutual interference is considered.

The authors formulated a convex optimization problem to characterize the Pareto boundary of the rate region of such a system. They find a closed form solution for this problem. They also show that such a formulation could be used for systems with larger number of users. The authors of [40] have looked at the rate region problem of a two-user MISO IC from a game-theoretic perspective. They have shown that any point on the Pareto boundary could be achieved by a linear combination of maximum ratio transmission (MRT) and ZF beamforming. This way, they have been able to determine the Pareto boundary using one scalar parameter. This has further made them capable of proposing efficient algorithm to find some important operational points including maximum weighted sum-rate, Nash-bargaining, and egalitarian solution. The authors of [41] characterize the Pareto boundary of a two-user MISO IC with a single scalar parameter. They use this parameterization along with iterative polyblock outer approximation algorithm to find maximum sum-rate points. Later, they extend their work to parameterization of multi-user MISO ICs [35]. In [42], the Pareto boundary of a two-user MISO IC is found by a method which is much more computationally efficient compared to the previous works. There exist other works which have proposed distributed algorithms that only use local channel state information (CSI) to compute Pareto boundary of MISO ICs [43–45].

2.2.3 ZF Beamforming

When a transmitter simultaneously use shared wireless medium on the same frequency band to communicate with a number of receivers, inter-symbol interference occurs at the receivers side. Interference decreases SINR which in turn lowers the data rate

that could be supported by a communication link. ZF beamformig is a promising beamforming design technique to avoid inter-symbol interference, specially when the number of antennas at the transmitter is much more than the number of receivers. Although with the help of ZF beamforming we can cancel interference effect at the receivers, it usually comes at the price that the power of desired signal could not reach its maximum. Actually, most of the time, beamforming design problems could be considered as a trade-off between increasing the power of the desired signal at one receiver and decreasing the power of interference at other receivers. The more the power of the desired signal at a specific receiver the more the interference caused by that signal at other receivers. In this sense, we are always looking for an optimum beamformer that the power of the interference signals and the power of desired signal are such that our design criteria are met.

The authors of [46] consider a multi-user MIMO (MU-MIMO) downlink beamforming scenario. They use ZF beamforming in order to propose closed-form low complexity solutions for both sum-rate maximization and QoS problems. Although ZF generally results in loss of optimality, the authors of [46] show that their solutions approach the optimal solution at high signal-to-noise ratio (SNR). In order to improve the performance of simple ZF, [47] introduces a regularized channel inversion technique that could result in higher sum-rates while still being far from sum-capacity at high SNRs. In [47] perturbation is added to regularized channel inversion in order to achieve near-capacity performance. The authors of [48] show that the ZF beamforing strategy could asymptotically achieve the same sum-rate capacity as DPC when number of users goes to infinity. They do so by exploiting the fact that larger number of

users would have orthogonal channels when the number of users goes to infinity. So, they can group such users and transmit data to each group one at a time. MMF and sum-rate maximization problems given per-antenna power constraints in a multi-user MISO (MU-MISO) system are studied in [49]. It shows that in case of per-antenna power constraints, unlike total power constraint, pseudo-inverse is not necessarily the best choice for ZF beamforming. They actually find optimal generalized inverses by using convex optimization techniques which result in better performance than pseudo-inverse ZF beamforming. Optimal ZF beamforming under per-antenna power constraints have also been studied in [50–52] . The authors of [53] compare the ZF beamforming and MRT in terms of spectral-efficiency and energy-efficiency in MU-MISO downlink beamforming system. They show that at high spectral-efficiency and low energy-efficiency regimes ZF outperforms MRT while at low spectral-efficiency and high energy-efficiency the opposite holds.

2.3 Joint Caching and Beamforming

The large body of work which study the benefits of cache deployment in wireless cellular networks, can be classified into two main streams. One stream studies the cache optimization problem by focusing on the BSs to EUs links. In Subsection 2.3.1 we cover those pieces of work which belong to this stream [11–15]. The other stream which explores the potentials of cache deployment by considering CP-to-BSs layer is elaborated in Subsection 2.3.2 [7, 16–19].

2.3.1 BSs-to-EUs Layer

The authors of [11] consider a grid-based wireless network, equipped with caches, in order to find asymptotic laws of joint delivery and replication problem. In [12], the authors aim to optimize the sizes of the caches allocated to certain distribution nodes, called helpers, such that the average downloading time of on-demand video streaming from a BS to the EUs is minimized. The investigation in [13] studies the tradeoff between the BSs' density and cache sizes in the cache-enabled BSs of a small-cell network. The authors of [14], study the performance analysis and optimization of a hybrid caching scheme and corresponding multicasting design in a large-scale cache-enabled heterogeneous wireless network. This hybrid caching scheme stores identical files at all macro-BSs, while pico-BSs cache randomly selected files. The study in [15] investigates coded caching schemes in a small-cell network, where each BS caches a portion of each file as coded packets.

2.3.2 CP-to-BSs Layer

To the best of our knowledge, the studies in [7] and [19] are the only pieces of work which focus on cache optimization problem at the BSs for a downlink C-RAN system. In [7], a single cluster of BSs cooperate with each other in order to deliver a file requested by a single EU at a time. Since each time, only one EU is served, interference is non-existent during file transmission. In [19], the authors consider a multi-cluster C-RAN, where the BSs at each cluster serve their EUs through cooperative transmission. In such a C-RAN, interference exists among CP-to-BSs links in different clusters. For such a network setup, the authors study the BS cache size optimization problem by

focusing on maximizing the long-term expected sum-rate of all clusters. Note that such an approach may not necessarily result in the smallest amount of the average BS downloading time [7].

Chapter 3

Joint Optimization of Beamforming and Cache Size Allocation

In this chapter, we consider the jointly optimal design of cache size allocation and beamforming under power/storage budget constraints. This joint optimization is done in two stages. Beamforming stage comes first with the purpose of minimizing downloading time for given channel realizations. We simplify beamforming optimization problem by presenting multiple lemmas, and we finalize this simplification by proposing two techniques based on ZF beamforming. Finally, using the results of beamforming stage, cache allocation problem is solved by convex optimization techniques with the purpose of minimizing average downloading time over long-term periods.

3.1 System Model

Consider a downlink C-RAN with multiple BSs connected to a CP via a shared wireless back-haul link, as shown in Fig. 3.1. To overcome the data rate limitation of the backhual, each BS uses a local cache with a finite capacity to download and pre-store a portion of each file from the cloud during off-peak time.

We consider a network of L cells, each with a single-antenna BS serving the EUs

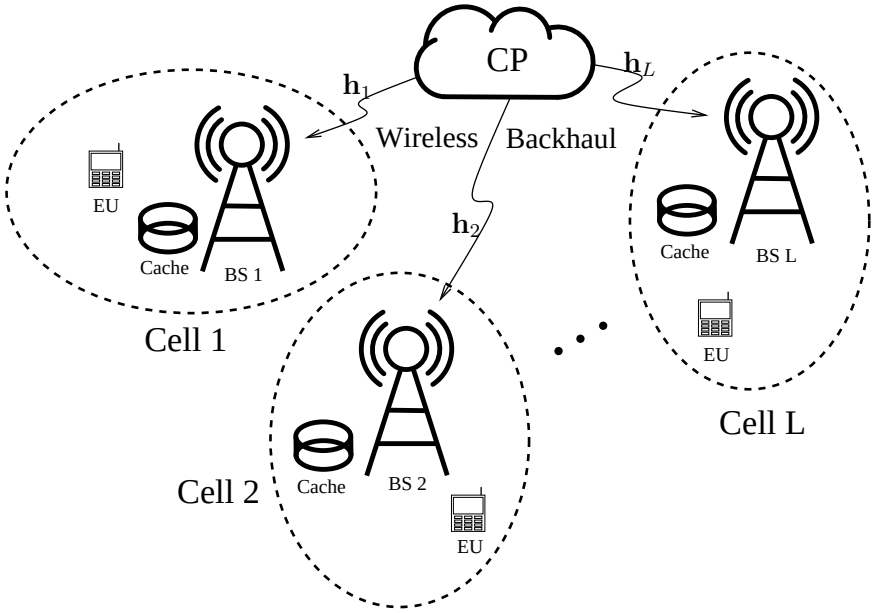

Figure 3.1: A downlink C-RAN consisting of L cells. The BS of each cell is equipped with a cache unit to pre-store a portion of each file available on the cloud.

in its cell. At each transmission instance, referred to as an epoch in the remainder of the paper, each EU requests a file stored in the central data base via its serving BS. If the requested file is completely cached at a BS during off-peak times, that BS does not need to receive anything from the CP during peak times, otherwise the BS needs to receive the remaining part of the file from the CP. We assume that central data base at the CP contains all the files that may be requested by the EUs.

We assume that the CP is equipped with M transmit antennas. We model the wireless channel between the CP and the BSs as block-fading, i.e., the channel remains constant during each epoch and changes over epochs independently. Also, each file transmission from the CP to a BS can be completed during one channel coherence block[1]. The conjugate of the channel vector between the CP and BS l, is denoted as $\mathbf{h}_l \triangleq [h_{1l} \ \cdots \ h_{Ml}]^T \in \mathbb{C}^{M\times 1}$, where h_{ml} is the channel between the m-th antenna

[1]We use coherent block and epoch interchangeably in this paper.

at the CP and BS l, for $l \in \mathcal{L} \triangleq \{1, \ldots, L\}$. The conjugate of the channel matrix between the CP and L BSs is denoted as $\mathbf{H} \triangleq [\mathbf{h}_1 \ \cdots \ \mathbf{h}_L] \in \mathbb{C}^{M \times L}$. We assume that the transmit beamforming weights at the CP may change during each epoch. Hence, at the CP we consider a time-varying beamforming vector, denoted as $\mathbf{w}_l(t) \triangleq [w_{1l}(t) \ \cdots \ w_{Ml}(t)]^T \in \mathbb{C}^{M \times 1}$, to linearly pre-code the message $s_l(t) \in \mathbb{C}$ to BS l, for $t \in [0, T]$, where T is the *delivery time* of that epoch. The delivery time of each epoch is the time duration that it takes to deliver all requested files.

We define $\mathbf{W}(t) \triangleq [\mathbf{w}_1(t) \ \cdots \ \mathbf{w}_L(t)] \in \mathbb{C}^{M \times L}$ as the time-varying beamforming (pre-coding) matrix at the CP. The receiver noise at BS l at time t, denoted as $z_l(t)$, is assumed to be complex Gaussian with zero mean and variance of BN_0, i.e., $z_l(t) \backsim \mathcal{CN}(0, \ BN_0)$, where B is the communication bandwidth and N_0 is the power spectral density of the white noise. Based on these definitions, we write the signal $y_l(t)$ received at BS l at time t as

$$y_l(t) = \mathbf{h}_l^H \mathbf{w}_l(t) s_l(t) + \mathbf{h}_l^H \sum_{j \in \mathcal{L} \setminus \{l\}} \mathbf{w}_j(t) s_j(t) + z_l(t). \tag{3.1}$$

Note that the reason behind defining time-varying beamformers is that once the file requested by a particular base station is delivered, the beamformer corresponding to that base station has to be turned off, i.e., the corresponding beamforming vector has to be set to zero. In this case, all other beamforming weights have to be readjusted to ensure optimal performance as the "interference landscape" has now changed. This fact leads us to consider time-varying beamforming vectors. If the maximum transmit power available at the CP is P, then at any time t, the beamforming vectors must

satisfy the following constraint:

$$\sum_{l\in\mathcal{L}} \|\mathbf{w}_l(t)\|^2 \leq P. \tag{3.2}$$

We assume K files in the central data base at the CP, each file with F bits[2]. The popularity of file k, denoted as ϕ_k, is the probability that file k is requested by any BS, for $k \in \mathcal{K} \triangleq \{1,\ldots,K\}$. The size of the portion of file k, cached at BS l is denoted by c_{lk}[3]. Let $\mathbf{c}_k \triangleq [c_{1k} \ \cdots \ c_{Lk}]^T \in \mathbb{R}^{L\times 1}$ represent the vector of cache sizes used to partially store file k at the L BSs, while $\mathbf{C} \triangleq [\mathbf{c}_1 \ \cdots \ \mathbf{c}_K] \in \mathbb{R}^{L\times K}$ stands for cache allocation matrix of all K files. If C, measured in bits, is the sum of all cache sizes in all BSs, c_{lk}'s must satisfy the following constraints:

$$\sum_{l\in\mathcal{L}}\sum_{k\in\mathcal{K}} c_{lk} \leq C \tag{3.3}$$

$$0 \leq c_{lk} \leq F. \tag{3.4}$$

Based on (3.1), at time t, the SINR of the desired signal at BS l is as

$$\mathrm{SINR}_l(t) = \frac{|\mathbf{h}_l^H \mathbf{w}_l(t)|^2}{\sum\limits_{j\in\mathcal{L}\setminus\{l\}} |\mathbf{h}_l^H \mathbf{w}_j(t)|^2 + BN_0}. \tag{3.5}$$

The achievable rate over of the channel between the CP and BS l is given by

$$r_l(t) = B\log_2(1+\mathrm{SINR}_l(t)). \tag{3.6}$$

Let us define $\mathbf{d} \triangleq [d_1 \ \cdots \ d_L]^T \in \mathcal{K}^{L\times 1}$, where d_l stands for the index of the file requested by BS l in the current epoch. Then, $\alpha_l(\mathbf{C},\mathbf{d}) \triangleq F - c_{l,d_l}$ is the amount of

[2]Similar to the pioneering work of [7], we assume that all files have the same size. This is possible by splitting each file into smaller packets of equal size.

[3]In reality, c_{lk}'s are discrete variables, for the sake of simplicity, we assume that they are continuous. Such an assumption has been widely used in the literature, see [7, 19].

data (measured in bits) to be delivered to the BS l during the current epoch. Since each requested file needs to be completely delivered within an epoch, the delivery time T must satisfy the following constraint[4]:

$$\int_0^T \mathbf{r}(t)dt = \boldsymbol{\alpha}(\mathbf{C}, \mathbf{d}) \ , \quad l \in \mathcal{L}, \tag{3.7}$$

where we use the following definitions are used $\mathbf{r}(t) \triangleq [r_1(t) \ \ \cdots \ \ r_L(t)]^T \in \mathbb{R}^{L\times 1}$ and $\boldsymbol{\alpha}(\mathbf{C}, \mathbf{d}) \triangleq [\alpha_1(\mathbf{C}, \mathbf{d}) \ \ \cdots \ \ \alpha_L(\mathbf{C}, \mathbf{d})]^T \in \mathbb{R}^{L\times 1}$.

Note that in each epoch, the delivery time T depends on the random demand vector $\mathbf{d}$ and on the random channel matrix $\mathbf{H}$. Hence, T is random from one epoch to another. Our goal is to jointly optimize cache sizes $\{c_{lk}\}$ and beamformers $\{\mathbf{w}_l(t)\}$ to minimize the *expected delivery time*, i.e., delivery time averaged over all channel realizations and file demands from the EUs. With this goal, joint caching-beamforming optimization consists of two stages that are performed at two different time scales, i.e., slow and fast time scales. Cache placement is performed in the slow time scale at off-peak time. The goal at this stage is to determine what portion of each file should be stored at the local cache of each BS such that expected delivery time is minimized. The outcome of this stage is optimized cache allocation to each file at each BS. The file delivery is performed at the fast time scale. During this delivery stage, we design beamforming vectors such that the delivery time for given cache sizes (obtained from the first stage) and for a given channel realization and EU demand is minimized. Our beamforming and caching designs must satisfy the total power constraint in (3.2) and the total cache constraints in (3.3) and (3.4), respectively. Based on the

[4]Without loss of generality, we assume that each epoch starts at $t = 0$ and ends at $t = T$, that is in each epoch, the time reference is the start of that epoch.

above discussions, in the fast time scale, the optimal beamforming matrix, denoted as $\mathbf{W}^{\mathrm{o}}(t) \triangleq [\mathbf{w}_1^{\mathrm{o}}(t) \quad \cdots \quad \mathbf{w}_L^{\mathrm{o}}(t)] \in \mathbb{C}^{M\times L}$, and the minimum delivery time, represented by $T^{\mathrm{o}}(\mathbf{H}, \boldsymbol{\alpha}(\mathbf{C}, \mathbf{d}))$, are obtained as the solution to the following optimization problem:

$$\mathcal{P}_1 \ : \ \min_{\mathbf{W}(t), T} \ \ T \tag{3.8a}$$

$$\text{s.t.} \quad \int_0^T \mathbf{r}(t)dt = \boldsymbol{\alpha}(\mathbf{C}, \mathbf{d}) \ , \ \ l \in \mathcal{L} \tag{3.8b}$$

$$\sum_{l\in\mathcal{L}} \|\mathbf{w}_l(t)\|^2 \leq P \ , \ t \in [0 \ T]. \tag{3.8c}$$

Note that $\mathbf{W}^{\mathrm{o}}(t)$ and $T^{\mathrm{o}}(\mathbf{H}, \boldsymbol{\alpha}(\mathbf{C}, \mathbf{d}))$ depend on the channel matrix $\mathbf{H}$ and on the vector $\boldsymbol{\alpha}(\mathbf{C}, \mathbf{d})$, for given $\mathbf{d}$ and $\mathbf{C}$. We can then formulate the optimization problem of expected delivery time in the slow time scale as

$$\mathcal{P}_2 \ : \ \min_{\mathbf{C}} \ \ \mathbb{E}_{\mathbf{H},\mathbf{d}}[T^{\mathrm{o}}(\mathbf{H}, \boldsymbol{\alpha}(\mathbf{C}, \mathbf{d}))] \tag{3.9a}$$

$$\text{s.t.} \quad \sum_{l\in\mathcal{L}} \sum_{k\in\mathcal{K}} c_{lk} \leq C \tag{3.9b}$$

$$0 \leq c_{lk} \leq F \ , \ l \in \mathcal{L} \ , \ k \in \mathcal{K}. \tag{3.9c}$$

In the next section, we focus on tackling $\mathcal{P}_1$ and consider $\mathcal{P}_2$ in Section 3.3.

3.2 Optimization at Beamforming Stage

In this section, we show that without loss of optimality, the rate vector function $\mathbf{r}(t)$ in problem $\mathcal{P}_1$ can be assumed piecewise constant over time in each epoch. This further implies that without loss of optimality, the beamforming vectors $\{\mathbf{w}_l(t)\}_{l=1}^L$ in $\mathcal{P}_1$ can be assumed piecewise constant functions of time t in each epoch. This finding simplifies $\mathcal{P}_1$ and sheds light on the difficulty of solving $\mathcal{P}_1$ and explains why this

optimization problem may not be amenable to a computationally efficient solution, thereby motivating us to resort to suboptimal solutions.

Lemma 3.1. *Without loss of optimality, in $\mathcal{P}_1$, the rate vector function $\mathbf{r}(t)$ and the beamforming vectors $\{\mathbf{w}(t)\}_{l=1}^{L}$ can be assumed piecewise constant over t in each epoch with at most L pieces.*

Proof. Suppose $\mathbf{r}^{\mathrm{o}}(t)$ is the value of $\mathbf{r}(t)$ at the optimum of problem $\mathcal{P}_1$. We now show that $\mathbf{r}^{\mathrm{o}}(t)$ can be a piecewise constant vector function of t with at most L pieces. To show this, we note that the multi-antenna transmission scenario from the CP to the L BSs is similar to that of the multi-user downlink beamforming scenario, where $\mathbf{h}_l$ represents the MISO channel between the transmitter at the CP and BS l. In such a scheme, assuming that receiver (BS) l is served by the beamforming vector $\mathbf{u}_l$, we use $\mathcal{R}$ to represent the set of all achievable rate vectors under a total transmit power budget P, that is

$$\mathcal{R} \triangleq \{\bar{\mathbf{r}} : \sum_{l\in\mathcal{L}} \|\mathbf{u}_l\|^2 \leq P\} \tag{3.10}$$

where $\bar{\mathbf{r}} \triangleq [\bar{r}_1 \ \cdots \ \bar{r}_L]^T$ and

$$\bar{r}_l = B\log_2(1 + \frac{|\mathbf{h}_l^H\mathbf{u}_l|^2}{\sum\limits_{j\in\mathcal{L}\setminus\{l\}} |\mathbf{h}_l^H\mathbf{u}_j|^2 + BN_0}),\ l \in \mathcal{L}. \tag{3.11}$$

Indeed, $\mathcal{R}$ is the achievable rate region in a downlink (transmit) beamforming scheme with L single-antenna receivers and a multi-antenna transmitter, with $\mathbf{u}_l$ being the beamforming vector which serves the l-th receiver, while the total transmitter power

budget is P [35]. We also define $\bar{\mathcal{R}}$ as the convex hull of $\mathcal{R}$, that is

$$\bar{\mathcal{R}} \triangleq \left\{ \bar{\mathbf{r}} \,\middle|\, \bar{\mathbf{r}} = \sum_{j=1}^{J} \tau_j \bar{\mathbf{r}}_j, \bar{\mathbf{r}}_j \in \mathcal{R}, \tau_j \geq 0,\ j = 1, \cdots, J, \sum_{j=1}^{J} \tau_j = 1,\ \forall J \in \mathbb{N} \right\}. \tag{3.12}$$

Note that any point in $\bar{\mathcal{R}}$ can be achieved using time sharing, (i.e., using time-varying beamformers). That is, for any $\bar{\mathbf{r}} \in \bar{\mathcal{R}}$, there exist $\{\bar{\mathbf{r}}_j\}_{j=1}^{J} \subset \mathcal{R}$, such that $\bar{\mathbf{r}} = \sum_{j=1}^{J} \tau_j \bar{\mathbf{r}}_j$. Hence, we can achieve $\bar{\mathbf{r}}$, by using $\bar{\mathbf{r}}_j$ for faction τ_j of the time.

At any time t, the vector $\mathbf{r}^{\mathrm{o}}(t)$ belongs to $\mathcal{R}$ because the l-th entry of $\mathbf{r}^{\mathrm{o}}(t)$ can be obtained from (3.11) by replacing $\mathbf{u}_l$ with $\mathbf{w}_l^{\mathrm{o}}(t)$, which satisfies the power constraint (3.8c). By definition, any convex combination of $\mathbf{r}^{\mathrm{o}}(t)$, $t \in [0, T^{\mathrm{o}}]$ belongs to $\bar{\mathcal{R}}$. Hence, the vector

$$\hat{\mathbf{r}} \triangleq \int_0^{T^{\mathrm{o}}} \frac{1}{T^{\mathrm{o}}} \mathbf{r}^{\mathrm{o}}(t) dt \tag{3.13}$$

belongs to $\bar{\mathcal{R}}$. Note that using (3.8b) and (3.13), we can write $\hat{\mathbf{r}} = \frac{1}{T^{\mathrm{o}}} \boldsymbol{\alpha}(\mathbf{C}, \mathbf{d})$. For notation simplicity, we drop the parameter dependency of $\boldsymbol{\alpha}(\mathbf{C}, \mathbf{d})$ on $\mathbf{C}$ and $\mathbf{d}$ and simply write $\hat{\mathbf{r}} = \frac{1}{T^{\mathrm{o}}} \boldsymbol{\alpha}$.

We now show by contradiction that $\hat{\mathbf{r}}$ is on the Pareto boundary of $\bar{\mathcal{R}}$, denoted as $\mathcal{P}_{\bar{\mathcal{R}}}$. Suppose that $\hat{\mathbf{r}}$ is not on $\mathcal{P}_{\bar{\mathcal{R}}}$. Hence, there exists a real-valued scalar $\theta > 1$, such that $\hat{\mathbf{r}}' = \theta \hat{\mathbf{r}} \succ \hat{\mathbf{r}}$ is a point on $\mathcal{P}_{\bar{\mathcal{R}}}$. Let $\hat{T}$ denote the delivery time corresponding to $\hat{\mathbf{r}}'$, we can then write

$$\boldsymbol{\alpha} = \hat{T} \hat{\mathbf{r}}' = \hat{T} \theta \hat{\mathbf{r}} = \hat{T} \theta \boldsymbol{\alpha} / T^{\mathrm{o}} \tag{3.14}$$

which implies that

$$\hat{T} = \frac{T^{\mathrm{o}}}{\theta} < T^{\mathrm{o}}. \tag{3.15}$$

Because $\hat{\mathbf{r}}'$ is on the Pareto boundary of $\bar{\mathcal{R}}$, as we show in Lemma 6.1 in Appendix 6.1, we can represent $\hat{\mathbf{r}}'$ as a convex combination of at most L vectors in $\mathcal{R}$. That is, there exists $\mathbf{r}_i \in \mathcal{R}$, and $\eta_i \geq 0$, for $i = 1, \ldots, L$, with $\sum_{i=1}^{L} \eta_i = 1$, such that

$$\hat{\mathbf{r}}' = \sum_{i=1}^{L} \eta_i \mathbf{r}_i \,. \tag{3.16}$$

Such a convex combination in (3.16) amounts to time sharing as we show in the sequel: choose the elements of $\mathbf{r}(t)$ to be piecewise constant functions as

$$\mathbf{r}(t) = \mathbf{r}_i, \text{ for } \delta_{i-1} < t \leq \delta_i \tag{3.17}$$

where $\delta_0 = 0$ and $\delta_i = \delta_{i-1} + \hat{T}\eta_i$ for $i = 1, \ldots, L$. By definition, $\mathbf{r}(t) \in \mathcal{R}$, for any t, and we can write

$$\int_0^{\hat{T}} \mathbf{r}(t)dt = \sum_{i=1}^{L} \hat{T}\eta_i \mathbf{r}_i = \hat{T} \sum_{j=1}^{L} \eta_i \mathbf{r}_i = \hat{T}\hat{\mathbf{r}}' = \boldsymbol{\alpha} \tag{3.18}$$

where we used (3.17), (3.16), and (3.14) in the first, the third, and the fourth equalities, respectively. It follows from (3.18) and the fact that $\hat{T} < T^{\text{o}}$ (see (3.15)) that the achievable rate vector $\mathbf{r}(t)$ for $\mathcal{P}_1$, given in (3.17), results in a smaller delivery time. This contradicts the earlier assumption that T^{o} is the shortest delivery time. Hence, $\hat{\mathbf{r}}$ in (3.13), must be on the boundary $\mathcal{P}_{\bar{\mathcal{R}}}$ of $\bar{\mathcal{R}}$, implying that $\hat{\mathbf{r}} = \hat{\mathbf{r}}'$. This leads to the conclusion that the piecewise constant vector $\mathbf{r}(t)$ in (3.17) leads to the shortest delivery time and, thus, is one choice for $\mathbf{r}^{\text{o}}(t)$. Hence, we proved that without loss of optimality, $\mathbf{r}(t)$ in $\mathcal{P}_1$ can be assumed to be a piecewise constant vector function of t with at most L pieces. Consequently, we can assume, without loss of optimality, that the beamforming vectors $\{\mathbf{w}_l(t)\}_{l=1}^{L}$ are piecewise constant vector functions of t. $\square$

Based on Lemma 3.1, we can assume that without loss of optimality, there exist $\{\delta_i\}_{i=0}^{L}$, with $\delta_0 = 0$ and $\delta_L = T$, such that we can write

$$\mathbf{r}(t) = \mathbf{r}_i, \text{ for } \delta_{i-1} < t \leq \delta_i \tag{3.19}$$

$$\mathbf{w}_l(t) = \mathbf{w}_{li}, \text{ for } \delta_{i-1} < t \leq \delta_i. \tag{3.20}$$

Here, $\mathbf{w}_{li}$ and the l-th element of $\mathbf{r}_i$, denoted as r_{li} satisfy the following equation

$$r_{li} = B\log_2(1 + \frac{|\mathbf{h}_l^H \mathbf{w}_{li}|^2}{\sum\limits_{j\in\mathcal{L}\setminus\{l\}} |\mathbf{h}_l^H \mathbf{w}_{ji}|^2 + BN_0}), \quad i, l \in \mathcal{L} \tag{3.21}$$

Using (3.19) and (3.20), we can rewrite the optimization problem $\mathcal{P}_1$ as

$$\mathcal{P}_3 \ : \min_{\mathcal{W},T,\{\delta_i\}} \quad T \tag{3.22a}$$

$$\text{s.t.} \quad \sum_{i\in\mathcal{L}} (\delta_i - \delta_{i-1})\mathbf{r}_i = \boldsymbol{\alpha} \tag{3.22b}$$

$$\sum_{l\in\mathcal{L}} \|\mathbf{w}_{li}\|^2 \leq P \ , \ i \in \mathcal{L} \tag{3.22c}$$

$$0 = \delta_0 \leq \delta_1 \leq \cdots \leq \delta_{L-1} \leq \delta_L = T \tag{3.22d}$$

where $\mathbf{W}_i \triangleq [\mathbf{w}_{1i} \ \cdots \ \mathbf{w}_{Li}] \in \mathbb{C}^{M\times L}$, for $i \in \mathcal{L}$

$\mathcal{W} \triangleq \{\mathbf{W}_1, \ldots, \mathbf{W}_L\}$. We now define $t_i \triangleq \delta_i - \delta_{i-1} \geq 0$, for $i \in \mathcal{L}$. Using the fact that $\sum_{i\in\mathcal{L}} t_i = T$, we rewrite optimization problem $\mathcal{P}_3$ as

$$\mathcal{P}_4 \ : \min_{\mathcal{W},\mathbf{t}} \quad \sum_{i\in\mathcal{L}} t_i \tag{3.23a}$$

$$\text{s.t.} \quad \sum_{i\in\mathcal{L}} t_i\mathbf{r}_i = \boldsymbol{\alpha} \tag{3.23b}$$

$$\sum_{l\in\mathcal{L}} \|\mathbf{w}_{li}\|^2 \leq P \ , \ i \in \mathcal{L} \tag{3.23c}$$

$$t_i \geq 0 \ , \ i \in \mathcal{L} \tag{3.23d}$$

where we define $\mathbf{t} \triangleq [t_1 \quad \cdots \quad t_L]^T \in \mathbb{R}^{L\times 1}$. The optimization problem $\mathcal{P}_4$ has an interesting interpretation: The data delivery in each epoch can be performed

in a maximum of L phases, where in the i-th phase, the CP uses a set of fixed beamformers $\{\mathbf{w}_{li}\}_{l\in\mathcal{L}}$ for transmission to the L BSs. The duration of phase i is t_i, and the transmission rate from the CP to BS l is equal to the l-th element of $\mathbf{r}_i$, i.e., r_{li} in (3.21).

Note that at the optimum of $\mathcal{P}_4$, the rate vectors $\{\mathbf{r}_i\}_{i=1}^{L}$ must be on the boundary of the rate region $\mathcal{R}$. Otherwise, if, at the optimum of $\mathcal{P}_4$, one of the rate vectors, say $\mathbf{r}_{i'}$, is not on the boundary of $\mathcal{R}$, for some $i' \in \mathcal{L}$, then there exists $\kappa > 1$, such that $\kappa\mathbf{r}_{i'}$ belongs to $\mathcal{R}$. Hence, we can replace $\mathbf{r}_{i'}$ and $t_{i'}$ with, $\kappa\mathbf{r}_{i'}$ and $t_{i'}/\kappa$, respectively, while satisfying (3.23c). This reduces the objective fucntion in $\mathcal{P}_4$, thereby contradicting the optimality assumption of $\{\mathbf{r}_i\}i = 1^L$. Hence, $\{\mathbf{r}_i\}_{i=1}^{L}$ must be on the boundary of $\mathcal{R}$. Unfortunately, determining these vectors on the boundary of $\mathcal{R}$, and hence, solving the non-convex optimization problem $\mathcal{P}_4$ do not appear to be computationally affordable. Indeed, finding the beamforming vectors to achieve the boundary of $\mathcal{R}$ is, in general, a challenging problem and has been solved only for the special case of two receivers [39–41]. For general case with more than two receivers users, finding the boundary of $\mathcal{R}$ appears to be an NP-hard problem with highly prohibitive computational complexity. In what follows, we resort to a suboptimal approach to *tackle* optimization problem $\mathcal{P}_1$.

3.2.1 Applying ZF Beamforming

To tackle optimization problem $\mathcal{P}_1$, we consider a delivery scheme which consists of multiple phases in each epoch. At each phase, a unique subset of the beamformers is active, while the rest of the L beamformers are inactive. Thus, each phase is identified

by the subset of active beamformers. For a set of L beamformers, there are a total $2^L - 1$ non-empty subsets, where at least one beamformer is turned on. Hence, we have $2^L - 1$ phases in each epoch. Define $\mathcal{I} \triangleq \{1, \ldots, 2^L - 1\}$. If any $i \in \mathcal{I}$ is used as phase index, and is represented in binary basis as $b_{L-1}b_{L-2}\cdots b_1 b_0$, then in the i-th phase, the l-th beamformer is on (off) if $b_{l-1} = 1$ (0), for $l \in \mathcal{L}$. Let us define $\mathcal{S}_i$ as a subset of BS indices that includes all those BSs whose corresponding bit is 1 in the L-bit binary representation of i. For phase i, let $\mathcal{S}_i$ denote the index set of active beamformers in this phase. Since beamformer l is intended to BS l, effectively, $\mathcal{S}_i$ contains the BS indices that the CP transmits to in phase i.

For any $l \in \mathcal{S}_i$, we define

$$\mathbf{z}_{li} \triangleq \begin{cases} \dfrac{\mathbf{h}_l - \mathbf{H}_{li}(\mathbf{H}_{li}^H \mathbf{H}_{li})^{-1}\mathbf{H}_{li}^H \mathbf{h}_l}{\|\mathbf{h}_l - \mathbf{H}_{li}(\mathbf{H}_{li}^H \mathbf{H}_{li})^{-1}\mathbf{H}_{li}^H \mathbf{h}_l\|_2}, & \text{if } \nu \notin \mathbb{N}, \\ \dfrac{\mathbf{h}_{\nu+1}}{\|\mathbf{h}_{\nu+1}\|}, & \text{if } \nu \in \mathbb{N}, \end{cases} \tag{3.24}$$

where $\nu = \log_2 i$ and matrix

$$\mathbf{H}_{li} \triangleq [\mathbf{h}_{l'}]_{l' \in \mathcal{S}_i \setminus \{l\}}, \text{ for } l \in \mathcal{S}_i, i \in \mathcal{I}\,. \tag{3.25}$$

consists of channel vectors of those BSs, other than BS l, whose corresponding beamformers are active in phase i. Note that when i is a power of 2, $\mathbf{z}_{li}$ is the unit-norm maximum ratio transmission (MRT) beamformer intended to serve BS l; otherwise, $\mathbf{z}_{li}$ is the unit-norm ZF beamforming vector, which ensures $\mathbf{z}_{li}^H \mathbf{h}_{l'} = 0$, for $l' \in \mathcal{S}_i \setminus \{l\}$. Define $\mathcal{T}_i$ as the set of all those time instances where the CP is in phase i. Without loss of generality, we can assume that $\mathcal{T}_i$ is a time interval in $\mathbb{R}^+$ with length t_i, i.e., that is $\mathcal{T}_i \triangleq [a_i,\ a_i + t_i)$, for $i \in \mathcal{I}$, with $t_0 = 0$ and $a_i \triangleq \sum_{j=0}^{i-1} t_j$. We set beamformer

l as

$$\mathbf{w}_l(t) \triangleq \sqrt{p_{li}(t)}\mathbf{z}_{li}, \text{ for } t \in \mathcal{T}_i, \tag{3.26}$$

where $p_{li}(t) \geq 0$ is the power allocated to BS l at time t in phase i. Note that if $l \notin \mathcal{S}_i$, then $p_{li}(t)$ must be zero for $t \in \mathcal{T}_i$, meaning that beamformer l is inactive during phase i. Substituting (3.26) in (3.5), we can obtain the SINR at BS l as

$$\text{SINR}_l(t) = \frac{|\mathbf{h}_l^H \mathbf{w}_l(t)|^2}{BN_0} = p_{li}(t)\gamma_{li}(\mathbf{H}), \text{ for } t \in \mathcal{T}_i. \tag{3.27}$$

where

$$\gamma_{li}(\mathbf{H}) \triangleq \frac{|\mathbf{h}_l^H \mathbf{z}_{li}|^2}{BN_0}, \; l \in \mathcal{S}_i, \; i \in \mathcal{I}. \tag{3.28}$$

Using (3.26) and (3.27), we can rewrite (3.6) as

$$r_l(t) = B\log_2(1 + p_{li}(t)\gamma_{li}(\mathbf{H})), \text{ for } t \in \mathcal{T}_i, \;\; i \in \mathcal{I}. \tag{3.29}$$

In light of (3.29), we can rewrite the constraint (3.8b) as

$$\int_0^T r_l(t)dt = \sum_{i\in\mathcal{I}} \int_{\mathcal{T}_i} B\log_2(1 + p_{li}(t)\gamma_{li}(\mathbf{H}))dt = \alpha_l, \text{ for } l \in \mathcal{L} \tag{3.30}$$

Also, using (3.26), we can rewrite constraint (3.8c) as

$$\sum_{l\in\mathcal{L}} \|\mathbf{w}_l(t)\|^2 = \sum_{l\in\mathcal{L}} p_{li}(t) \leq P, \text{ for } t \in \mathcal{T}_i. \tag{3.31}$$

Based on (3.30) and (3.31), and under the ZF beamforming structure in (3.26), opti-

mization problem $\mathcal{P}_1$, can be rewritten as

$$\mathcal{P}_5 : \min_{\mathbf{P}(t),\mathbf{t}} \quad \sum_{i\in\mathcal{I}} t_i \tag{3.32a}$$

$$\text{s.t.} \quad \sum_{i\in\mathcal{I}} \int_{\mathcal{T}_i} \mathbf{r}_i(t)dt = \boldsymbol{\alpha} \tag{3.32b}$$

$$\sum_{l\in\mathcal{S}_i} p_{li}(t) \le P \ , \ t \in \mathcal{T}_i \ , \ i \in \mathcal{I} \tag{3.32c}$$

$$p_{li}(t) > 0 \ , \ t \in \mathcal{T}_i \ , \ l \in \mathcal{S}_i \ , \ i \in \mathcal{I} \tag{3.32d}$$

$$p_{li}(t) = 0 \ , \ t \in \mathcal{T}_i \ , \ l \notin \mathcal{S}_i \ , \ i \in \mathcal{I} \tag{3.32e}$$

where $\mathbf{t} \triangleq [t_1 \ \ t_2 \ \ \cdots \ \ t_{2^L-1}]^T \in \mathbb{R}^{(2^L-1)\times 1}$ and $\mathbf{r}_i(t) \triangleq [r_{1i}(t) \ \ r_{2i}(t) \ \ \cdots \ \ r_{Li}(t)]^T \in \mathbb{R}^{L\times 1}$, when $t \in \mathcal{T}_i$, $i \in \mathcal{I}$ and $\mathbf{P}(t)$ is an $L \times (2^L - 1)$ matrix whose (l, i)-th element is $p_{li}(t)$.

The following lemma helps us to simplify $\mathcal{P}_5$.

Lemma 3.2. *Without loss of optimality, we can assume that in $\mathcal{P}_5$, the rate vector $\mathbf{r}_i(t)$ and the power $p_{li}(t)$ are constant at each phase.*

Proof. Suppose that at the optimum of $\mathcal{P}_5$, the optimal value of $\mathbf{r}_i(t)$ is denoted as $\mathbf{r}_i^{\text{o}}(t)$. Define the achievable rate region for phase i as

$$\mathcal{R}_{\text{z}}^i \triangleq \left\{ \bar{\mathbf{r}}_i \middle| \sum_{l\in\mathcal{S}_i} \bar{p}_{li} \le P, \ \bar{p}_{li} > 0, l \in \mathcal{S}_i \text{ and } \bar{p}_{li} = 0, l \notin \mathcal{S}_i \right\}$$

where $\bar{\mathbf{r}}_i \triangleq [\bar{r}_{1i} \ \ \cdots \ \ \bar{r}_{Li}]^T$, $\bar{r}_{li} \triangleq B \log_2(1 + \bar{p}_{li}\gamma_{li}(\mathbf{H}))$, with $\gamma_{li}(\mathbf{H})$ given in (3.28). In Appendix 6.2, we show that $\mathcal{R}_{\text{z}}^i$ is in fact a convex set. Denote the optimal value of t_i as t_i^{o}. Since $\mathbf{r}_i^{\text{o}}(t) \in \mathcal{R}_{\text{z}}^i$ for any $t \in \mathcal{T}_i$, vector $\hat{\mathbf{r}}_i$, defined as

$$\hat{\mathbf{r}}_i \triangleq \frac{1}{t_i^{\text{o}}} \int_{\mathcal{T}_i^{\text{o}}} \mathbf{r}_i^{\text{o}}(t)dt \tag{3.33}$$

belongs to $\mathcal{R}_{\mathrm{z}}^{i}$, because $\hat{\mathbf{r}}_i$ is a convex combination of $\{\mathbf{r}_i^{\mathrm{o}}(t)\}_{t\in\mathcal{T}_i}$. Hence, there exist $\{\hat{p}_{li}\}_{l\in\mathcal{L}}$ such that

$$\sum_{l\in\mathcal{S}_i}\hat{p}_{li}\leq P,\ \text{with}\ \ \hat{p}_{li}>0, l\in\mathcal{S}_i\ \text{and}\ \hat{p}_{li}=0, l\notin\mathcal{S}_i \tag{3.34}$$

and

$$\hat{r}_{li}=B\log_2(1+\hat{p}_{li}\gamma_{li}(\mathbf{H}))\ ,\quad \text{for}\ i\in\mathcal{I} \tag{3.35}$$

where $\hat{r}_{li}$ is the l-th element of vector $\hat{\mathbf{r}}_i$. It follows from (3.34) that $\hat{p}_{li}$'s satisfy (3.32c)–(3.32e). Note also that we can use (3.33) to write

$$\sum_{i\in\mathcal{I}}t_i^{\mathrm{o}}\hat{\mathbf{r}}_i=\sum_{i\in\mathcal{I}}\int_{\mathcal{T}_i}\mathbf{r}_i^{\mathrm{o}}(t)dt=\boldsymbol{\alpha} \tag{3.36}$$

which implies, along with (3.35), that $\hat{p}_{li}$'s and t_i^{o}'s satisfy (3.32b). Hence, choosing $p_{li}(t)=\hat{p}_{li}$, $\mathbf{t}=[t_1^{\mathrm{o}}\ \ \cdots t_{2^L-1}^{\mathrm{o}}\ \ ]^T$ will not incur any loss of optimality. Hence the rate vector function $\mathbf{r}_i(t)$, and the power control functions $p_{li}(t)$ can be assumed constant at each phase, without loss of optimality. The proof is complete. □

Based on Lemma 3.2, we can simplify problem $\mathcal{P}_5$ as

$$\mathcal{P}_6:\min_{\mathbf{P},\mathbf{t}}\ \ \sum_{i\in\mathcal{I}}t_i \tag{3.37a}$$

$$\text{subject to}\ \ \sum_{i\in\mathcal{I}}t_i\mathbf{r}_i=\boldsymbol{\alpha} \tag{3.37b}$$

$$\sum_{l\in\mathcal{S}_i}p_{li}\leq P\ ,\ i\in\mathcal{I} \tag{3.37c}$$

$$p_{li}>0\ ,\ l\in\mathcal{S}_i\ ,\ i\in\mathcal{I} \tag{3.37d}$$

$$p_{li}=0\ ,\ l\notin\mathcal{S}_i\ ,\ i\in\mathcal{I} \tag{3.37e}$$

$$t_i\geq 0\ ,\ i\in\mathcal{I} \tag{3.37f}$$

where the l-th element of $\mathbf{r}_i$ is given by

$$r_{li} = B \log_2(1 + p_{li}\gamma_{li}(\mathbf{H})) \ , \quad \text{for } i \in \mathcal{I}, \tag{3.38}$$

and with a slight abuse of notation, $\mathbf{P}$ is an $L \times (2^L - 1)$ matrix whose (l, i)-th, p_{li}, represents the constant power i dedicated to BS l in phase i, for $l \in \mathcal{L},\ i \in \mathcal{I}$. The optimization problem $\mathcal{P}_6$ differs from $\mathcal{P}_4$ in two aspect 1) in $\mathcal{P}_4$, there are only a maximum of L phases, while in $\mathcal{P}_6$, there are a maximum of $2^L - 1$ phases, and 2) $\mathcal{P}_4$ is a beamforming optimization problem with optimization variables being beamforming vectors, while $\mathcal{P}_6$ is a power allocation problem, where given ZF beamforming vectors, power allocation to different phases is optimized.

We now show that the optimal solution to problem $\mathcal{P}_6$ does not need all the $2^L - 1$ phases. In fact, at most L phases with non-zero durations are needed for the optimal solution of $\mathcal{P}_6$. The following lemma articulates this property.

Lemma 3.3. *At the optimum of $\mathcal{P}_6$, out of all the $2^L - 1$ phases, at most L phases with non-zero durations are needed.*

Proof. Let us define

$$\mathcal{R}_{\text{z}} \triangleq \bigcup_{i \in \mathcal{I}} \mathcal{R}_{\text{z}}^i \tag{3.39}$$

where $\mathcal{R}_{\text{z}}^i$ is the achievable rate region for phase i under ZF beamforming and is defined similar to Lemma 3.2. Note that $\mathcal{R}_{\text{z}}$ may not necessarily be convex. Let us also define the convex hull of $\mathcal{R}_{\text{z}}$ as

$$\bar{\mathcal{R}}_{\text{z}} \triangleq \Big\{ \sum_{j=1}^{J} \tau_j \mathbf{r}_j : \mathbf{r}_j \in \mathcal{R}_{\text{z}},\ \tau_j \geq 0,\ j = 1, \cdots J, \\ \sum_{j=1}^{J} \tau_j = 1,\ \forall J \in \mathbb{N} \Big\}. \tag{3.40}$$

The optimal solution to $\mathcal{P}_6$, denoted as $(\mathbf{P}^{\mathrm{o}}, \mathbf{t}^{\mathrm{o}})$ must satisfy the following equations:

$$\sum_{i\in\mathcal{I}} t_i^{\mathrm{o}} \mathbf{r}_i^{\mathrm{o}} = \boldsymbol{\alpha} \tag{3.41a}$$

$$\sum_{l\in\mathcal{S}_i} p_{li}^{\mathrm{o}} \leq P \, , \; i \in \mathcal{I} \tag{3.41b}$$

$$p_{li}^{\mathrm{o}} \geq 0 \, , \; l \in \mathcal{S}_i \, , \; i \in \mathcal{I} \tag{3.41c}$$

$$p_{li}^{\mathrm{o}} = 0 \, , \; l \notin \mathcal{S}_i \, , \; i \in \mathcal{I} \tag{3.41d}$$

$$t_i^{\mathrm{o}} \geq 0 \, , \; i \in \mathcal{I} \, . \tag{3.41e}$$

Here, t_i^{o} and p_{li}^{o} are, respectively, the i-the element of $\mathbf{t}^{\mathrm{o}}$ and the (l,i)-th entry of $\mathbf{P}^{\mathrm{o}}$, while $\mathbf{r}_i^{\mathrm{o}}$ is the value of $\mathbf{r}_i$ at the optimum, for $i \in \mathcal{I}$. Dividing both sides of (3.41a) by the total delivery time under this scheme, i.e., $T_{\mathrm{mz}}^{\mathrm{o}} \triangleq \sum_{i\in\mathcal{I}} t_i^{\mathrm{o}}$ results in

$$\tilde{\mathbf{r}} \triangleq \sum_{i\in\mathcal{I}} \frac{t_i^{\mathrm{o}}}{T_{\mathrm{mz}}^{\mathrm{o}}} \mathbf{r}_i^{\mathrm{o}} = \frac{\boldsymbol{\alpha}}{T_{\mathrm{mz}}^{\mathrm{o}}}. \tag{3.42}$$

In light of (3.42), $\tilde{\mathbf{r}}$ is a convex combination of $|\mathcal{I}|$ points in $\mathcal{R}_{\mathrm{z}}$, and thus, belongs to $\bar{\mathcal{R}}_{\mathrm{z}}$. Moreover, $\tilde{\mathbf{r}}$ must be on the boundary of $\bar{\mathcal{R}}_{\mathrm{z}}$. In order to show this, let us suppose that $\tilde{\mathbf{r}}$ is not on the boundary of $\bar{\mathcal{R}}_{\mathrm{z}}$. Hence, there exists a real-valued scalar $\psi \; > 1$, such that $\tilde{\mathbf{r}}' = \psi\tilde{\mathbf{r}} \succ \tilde{\mathbf{r}}$ is a point on the boundary of $\bar{\mathcal{R}}_{\mathrm{z}}$. If $\tilde{T}$ denotes the delivery time corresponding to $\tilde{\mathbf{r}}'$, we can then write

$$\boldsymbol{\alpha} = \tilde{T}\tilde{\mathbf{r}}' = \tilde{T}\psi\tilde{\mathbf{r}} = \tilde{T}\psi\boldsymbol{\alpha}/T_{\mathrm{mz}}^{\mathrm{o}} \tag{3.43}$$

which implies that

$$\tilde{T} = \frac{T_{\mathrm{mz}}^{\mathrm{o}}}{\psi} < T_{\mathrm{mz}}^{\mathrm{o}}. \tag{3.44}$$

Because $\tilde{\mathbf{r}}'$ is on the Pareto boundary of $\bar{\mathcal{R}}_{\mathrm{z}}$, as we show in Lemma 6.1 in Appendix 6.1, we can represent $\tilde{\mathbf{r}}'$ as a convex combination of at most L vectors that belong to

$\mathcal{R}_{\text{z}} = \bigcup_{i\in\mathcal{I}} \mathcal{R}_{\text{z}}^i$. that is, there exist $\tilde{\mathbf{r}}_j \in \mathcal{R}_{\text{z}}$, for $j = 1, \dots L$, such that

$$\tilde{\mathbf{r}}' = \sum_{j=1}^{L} \zeta_j \tilde{\mathbf{r}}_j \tag{3.45}$$

while $\sum_{j=1}^{L} \zeta_j = 1$ and $\zeta_j \geq 0$, for $j = 1, \dots L$ hold true. Note that different $\tilde{\mathbf{r}}_j$'s may belong to different $\mathcal{R}_{\text{z}}^i$'s. Let $\mathcal{V}$ be the set of indices i for which set $\mathcal{R}_{\text{z}}^i$ include at least one of $\tilde{\mathbf{r}}_j$'s. We now group $\tilde{\mathbf{r}}_j$'s into $V \triangleq |\mathcal{V}| \leq L$ disjoint sets[5], denoted by $\{\tilde{\mathcal{R}}_{\text{z}}^i\}_{i\in\mathcal{V}}$, such that all members of $\tilde{\mathcal{R}}_{\text{z}}^i$ belong to $\mathcal{R}_{\text{z}}^i$, for $i \in \mathcal{V}$. We further define set $\mathcal{L}_i$ as

$$\mathcal{L}^i \triangleq \{j : \tilde{\mathbf{r}}_j \in \tilde{\mathcal{R}}_{\text{z}}^i\}, \text{ for } i \in \mathcal{V}. \tag{3.46}$$

Using (3.46), we can rewrite (3.45) as

$$\tilde{\mathbf{r}}' = \sum_{i\in\mathcal{V}} \sum_{j\in\mathcal{L}^i} \zeta_j \tilde{\mathbf{r}}_j = \sum_{i\in\mathcal{V}} \xi_i \bar{\mathbf{r}}_i \tag{3.47}$$

where we define $\xi_i \triangleq \sum_{j\in\mathcal{L}^i} \zeta_j$ and $\bar{\mathbf{r}}_i \triangleq \sum_{j\in\mathcal{L}^i} \frac{\zeta_j}{\xi_i} \tilde{\mathbf{r}}_j$, for $i \in \mathcal{V}$. Considering the fact that $\zeta_j \geq 0$, it is obvious that for $i \in \mathcal{V}$, $\bar{\mathbf{r}}_i$ is a convex combination of multiple vectors in $\mathcal{R}_{\text{z}}^i$. Since $\mathcal{R}_{\text{z}}^i$ is a convex set (see Appendix 6.2), $\bar{\mathbf{r}}_i$ is indeed a vector in $\mathcal{R}_{\text{z}}^i$, for $i \in \mathcal{V}$. Hence, $\mathbf{r}_i$ is an achievable rate vector for the i-th phase in problem $\mathcal{P}_6$, for $i \in \mathcal{V}$. Multiplying both sides of (3.47) by $\tilde{T}$ results in

$$\boldsymbol{\alpha} = \tilde{T}\tilde{\mathbf{r}}' = \sum_{i\in\mathcal{V}} \tilde{T}\xi_i \bar{\mathbf{r}}_i \tag{3.48}$$

where we have used the first equality in (3.43). Since, for $i \in \mathcal{V}$ $\mathbf{r}_i$ is an achievable rate vector for the i-th phase, there is a matrix $\mathbf{P}$, which achieves $\bar{\mathbf{r}}_i$. Hence, if we

[5]Note that the number of such subsets cannot be greater than the number of the vectors $\tilde{\mathbf{r}}_j$'s, which is L.

choose

$$t_i = \begin{cases} \tilde{T}\xi_i, & \text{if } i \in \mathcal{V} \\ 0, & \text{if } i \notin \mathcal{V}, \end{cases} \tag{3.49}$$

$$\mathbf{r}_i = \begin{cases} \bar{\mathbf{r}}_i, & \text{if } i \in \mathcal{V} \\ 0, & \text{if } i \notin \mathcal{V}, \end{cases} \tag{3.50}$$

such matrix $\mathbf{P}$ along with t_i's as in (3.49), while constituting a feasible point for problem $\mathcal{P}_6$, result in a delivery time equal to

$$\sum_{i\in\mathcal{I}} t_i = \sum_{i\in\mathcal{V}} \tilde{T}\xi_i = \tilde{T}\sum_{i\in\mathcal{V}}\sum_{j\in\mathcal{L}^i} \zeta_j = \tilde{T}\sum_{j=1}^{L}\zeta_j = \tilde{T} < T_{\text{mz}}^{\text{o}}$$

which contradicts the optimality of T_{mz}^{o}. Hence, $\tilde{\mathbf{r}}$, given in (3.42), must be on the boundary of $\bar{\mathcal{R}}_{\text{z}}$, implying that $\tilde{\mathbf{r}} = \tilde{\mathbf{r}}'$. Hence, we proved that without loss of optimality, we can achieve optimal solution of problem $\mathcal{P}_6$ with a vector $\mathbf{t}$ which has L non-zero elements at most. □

Lemma 3.3 implies that optimization problem $\mathcal{P}_6$ has a solution with at most L out of $2^L - 1$ values of t_i's and $\mathbf{r}_i$'s are non-zero. Unfortunately, finding these non-zero values appears to require exponential computational complexity. This means obtaining the globally optimal solution to $\mathcal{P}_6$ is challenging. Hence, in the sequel propose to a suboptimal solution to $\mathcal{P}_6$.

We propose obtain a locally optimal solution to $\mathcal{P}_6$ by using alternate convex search (ACS) technique. First, we claim that without loss of optimality, the constraint in (3.37b) can be replaced with the following constraint:

$$\sum_{i\in\mathcal{I}} t_i\mathbf{r}_i \succcurlyeq \boldsymbol{\alpha}. \tag{3.51}$$

To see this, note that at the optimum, if the l-th element of $\sum_{i\in\mathcal{I}} t_i\mathbf{r}_i$ is larger than the l-th element of $\boldsymbol{\alpha}$, then we can decrease as many non-zero p_{li}'s as needed to

ensure that (3.51) is satisfied with equality. Doing so, we do not violate optimality as the objective function depends only on t_i's. Hence, without loss of optimality we can rewrite $\mathcal{P}_6$, as

$$\mathcal{P}_7 : \min_{\mathbf{P},\mathbf{t}} \quad \sum_{i\in\mathcal{I}} t_i \tag{3.52a}$$

$$\text{subject to} \quad -\sum_{i\in\mathcal{I}} t_i \mathbf{r}_i \preccurlyeq -\boldsymbol{\alpha} \tag{3.52b}$$

$$(3.37c), (3.37d), (3.37e), (3.37f) \tag{3.52c}$$

We observe that $\mathcal{P}_7$ is a biconvex optimization problem. For given feasible p_{li}, this problem is a linear programming problem in $\mathbf{t}$, while for given feasible t_i's, based on (3.38), this problem becomes a convex feasibility problem in $\mathbf{P}$.

Given the biconvexity of $\mathcal{P}_7$, we can directly apply ACS technique to find a locally optimal solution to $\mathcal{P}_7$. We refer to this ACS-based technique as multi-phase ZF (MPZF) method as this technique may require more than one phase. However, the MPZF method suffers from two shortcomings: 1) as will be shown by our numerical examples, this method does not exhibit satisfactory performance when the number of antennas is large, and 2) the minimum delivery time in the objective function of (3.52) may not be a convex function of $\boldsymbol{\alpha}$, and this hinders finding an efficient method to solve the cache allocation problem $\mathcal{P}_2$.

To address the above shortcomings of the MPZF method, we now introduce a low complexity suboptimal solution to $\mathcal{P}_6$ by assuming that the CP completes the delivery of requested files to all BSs in a single phase. More specifically, we assume that the delivery is performed at phase $i = 2^L - 1$ and the remaining phases are nonexistent. This approach would substantially reduce the complexity of our problem. Moreover,

this single-phase assumption makes sense when we consider the fact that the number of antennas at the CP is much more than the number of BSs in a massive MIMO scenario. With this assumption, we replace t_{2^L-1} with T (as other t_i's are assumed to be 0), drop index i from p_{li}'s and $\gamma_{li}(\mathbf{H})$'s, and simplify problem $\mathcal{P}_6$ as

$$\mathcal{P}_8 : \min_{\mathbf{p},T} \quad T \tag{3.53a}$$

$$\text{s.t.} \quad BT\log_2(1+p_l\gamma_l(\mathbf{H})) = \alpha_l \ , \quad l \in \mathcal{L} \tag{3.53b}$$

$$\sum_{l\in\mathcal{L}} p_l \leq P \tag{3.53c}$$

$$p_l \geq 0 \ , \ l \in \mathcal{L} \tag{3.53d}$$

$$T \geq 0 \tag{3.53e}$$

where $\mathbf{p} \triangleq [p_1 \quad \cdots \quad p_L]^T \in \mathbb{R}^{L\times 1}$, and $\gamma_l(\mathbf{H})$ is defined as in (3.28) by setting $i = 2^L - 1$. Hence, from constraint (3.53b), we have

$$p_l = \frac{2^{\frac{\alpha_l}{BT}} - 1}{\gamma_l(\mathbf{H})} \ , \quad \text{for } l \in \mathcal{L}. \tag{3.54}$$

In order to minimize T, the power constraint in (3.53c) has to be met with equality. Thus, Substituting (3.54) into (3.53c) with equality, we conclude that T must satisfy the following equation:

$$\sum_{l\in\mathcal{L}} \frac{2^{\frac{\alpha_l}{BT}} - 1}{\gamma_l(\mathbf{H})} = P. \tag{3.55}$$

Interestingly, equation (3.55) has a unique solution for T. The reason is that α_l's and $\gamma_l(\mathbf{H})$'s, are all positive, and thus, the left side of (3.55) is monotonically decreasing in T. Hence, T can be uniquely obtained from (3.55) using the Newton-Raphson or bisection algorithms. Note that solving (3.55) results in a positive T due to the fact

that negative values t_i makes the left side of (3.55) negative which can not be equal to the positive P. It then follows from (3.54) that $p_l > 0$, for $l \in \mathcal{I}$. Hence, the values found for p_l's and T satisfy (3.53d) and (3.53e), and thus, they are the optimal solution to $\mathcal{P}_8$. We refer to this solution as the single-phase ZF (SPZF) method and denote the optimal value of T as $T^{\text{o}}_{\text{sz}}(\mathbf{H}, \boldsymbol{\alpha}(\mathbf{C}, \mathbf{d}))$, to emphasize its dependency on the channel matrix $\mathbf{H}$ and on the vector $\boldsymbol{\alpha}(\mathbf{C}, \mathbf{d})$, which is a function of the cache allocation matrix $\mathbf{C}$ and the demand vector $\mathbf{d}$. More specifically, $T^{\text{o}}_{\text{sz}}(\mathbf{H}, \boldsymbol{\alpha}(\mathbf{C}, \mathbf{d}))$ is the solution to (3.55) with respect to T.

The following lemma is useful when we perform caching design assuming SPZF is used in the file delivery stage.

Lemma 3.4. *The optimal value $T^{\text{o}}_{sz}(\mathbf{H}, \boldsymbol{\alpha}(\mathbf{C}, \mathbf{d}))$ is a convex function of $\boldsymbol{\alpha}(\mathbf{C}, \mathbf{d})$, and thus, is a convex function of cache allocation matrix $\mathbf{C}$.*

Proof. See Appendix 6.3. □

Note that the SPZF method is a suboptimal method with a single phase. Thus, it provides an upper bound to the optimum objective value of $\mathcal{P}_6$. The MPZF methods is also suboptimal as this method is only guaranteed to attain a local minimum of $\mathcal{P}_6$. Between SPZF and MPZF approaches, it is difficult to know which one offers a better performance. We will show via numerical examples that each of these two approaches can outperform the other one for different ranges of parameters.

3.2.2 Lower Bound

To obtain a lower bound for the optimal value of T in problem $\mathcal{P}_1$, we relax this problem by replacing the constraint in (3.8b) with the following constraint:

$$\int_0^T \boldsymbol{\rho}(t)dt \geq \boldsymbol{\alpha} \tag{3.56}$$

where we define $\boldsymbol{\rho}(t) \triangleq [\rho_1(t) \quad \cdots \quad \rho_L(t)]^T$, $\rho_l(t) = B\log_2(1+\mathrm{SNR}_l(t))$, $\mathrm{SNR}_l(t) \triangleq \frac{|\mathbf{h}_l^H \mathbf{w}_l(t)|^2}{BN_0}$ for $l \in \mathcal{L}$ and $t \in [0\ T]$. In this relaxation, we ignore the interference as we use SNR instead of SINR in the rate expression. Replacing the constraint in (3.8b) in $\mathcal{P}_1$ with (3.56) leads us to the following relaxed optimization problem:

$$\mathcal{P}_1^{\mathrm{lb}} : \min_{\mathbf{W}(t),T} \quad T \tag{3.57a}$$

$$\text{s.t.} \quad \int_0^T \boldsymbol{\rho}(t)dt \succcurlyeq \boldsymbol{\alpha} \tag{3.57b}$$

$$\sum_{l\in\mathcal{L}} \|\mathbf{w}_l(t)\|^2 \leq P\ ,\ t \in [0\ T]\,. \tag{3.57c}$$

The optimal value of the objective function $\mathcal{P}_1^{\mathrm{lb}}$ provides a lower bound for the optimal value of T in $\mathcal{P}_1$. To solve $\mathcal{P}_1^{\mathrm{lb}}$, if we express $\mathbf{w}_l(t)$ as that $\mathbf{w}_l(t) = \sqrt{p_l(t)}\tilde{\mathbf{w}}_l(t)$, with $\|\tilde{\mathbf{w}}_l(t)\|^2 = 1$, one can then assume without loss of optimality that $\tilde{\mathbf{w}}_l(t)$ should be in the direction of $\mathbf{h}_l$, i.e., $\tilde{\mathbf{w}}_l(t) = \mathbf{h}_l/\|\mathbf{h}_l\|$. Hence, we can rewrite $\mathcal{P}_1^{\mathrm{lb}}$ as

$$\mathcal{P}_1^{\mathrm{lb}} : \min_{\mathbf{p}(t),T} \quad T \tag{3.58a}$$

$$\text{s.t.} \quad \int_0^T \boldsymbol{\rho}(t)dt \succcurlyeq \boldsymbol{\alpha} \tag{3.58b}$$

$$\sum_{l\in\mathcal{L}} p_l(t) \leq P\ ,\ t \in [0\ T] \tag{3.58c}$$

$$p_l(t) \geq 0\ ,\ l \in \mathcal{L}\ ,\ t \in [0\ T] \tag{3.58d}$$

where we define: $\mathbf{p}(t) \triangleq [p_1(t) \quad \cdots \quad p_L(t)]^T \in \mathbb{R}^{L\times 1}$, $t \in [0\ T]$ and $\bar{\gamma}_l(\mathbf{H}) \triangleq \frac{\|\mathbf{h}_l\|^2}{BN_0}$, $l \in \mathcal{L}$, and the l-th element of $\boldsymbol{\rho}(t)$ is given by $\rho_l(t) = B\log_2(1+p_l(t)\bar{\gamma}_l(\mathbf{H}))$ for $l \in \mathcal{L}$ and

$t \in [0\ T]$. We can prove that at the optimum, the constraint in (3.58c) is satisfied with equality. Based on this fact and using an approach similar to what we did in Lemma 3.2, we can show that without loss of optimality, $\mathbf{p}(t)$, and consequently, $\boldsymbol{\rho}(t)$, are constant functions for $[0, T]$. Hence, problem $\mathcal{P}_1^{\text{lb}}$ can be rewritten as

$$\mathcal{P}_1^{\text{lb}}: \min_{\mathbf{p},T} \quad T \tag{3.59a}$$

$$\text{s.t} \quad BT \log_2(1 + p_l \bar{\gamma}_l(\mathbf{H})) \geq \alpha_l\ ,\ \ l \in \mathcal{L} \tag{3.59b}$$

$$\sum_{l \in \mathcal{L}} p_l \leq P \tag{3.59c}$$

$$p_l \geq 0\ ,\ l \in \mathcal{L}. \tag{3.59d}$$

where we dropped the time dependency of $\mathbf{p}(t)$ and $p_l(t)$'s and replaced them with $\mathbf{p}$ and p_l, respectively. As problem $\mathcal{P}_1^{\text{lb}}$ is similar to problem $\mathcal{P}_8$, we can solve $\mathcal{P}_1^{\text{lb}}$ using the same approach used for solving $\mathcal{P}_8$. Therefore T is obtained as the solution to the following equation:

$$\sum_{l \in \mathcal{L}} \frac{2^{\frac{\alpha_l}{BT}} - 1}{\bar{\gamma}_l(\mathbf{H})} = P. \tag{3.60}$$

We refer to the optimal value of T obtained by solving (3.60) as $T_{\text{lb}}^{\text{o}}(\mathbf{H}, \boldsymbol{\alpha}(\mathbf{C}, \mathbf{d}))$. Similar to Lemma 3.4, we can show that $T_{\text{lb}}^{\text{o}}(\mathbf{H}, \boldsymbol{\alpha}(\mathbf{C}, \mathbf{d}))$ is a convex function of $\boldsymbol{\alpha}(\mathbf{C}, \mathbf{d})$.

3.3 Optimization at Caching Stage

In this section, we aim to solve problem $\mathcal{P}_2$ in order to find the optimal cache size allocation for different files at different BSs. As finding $T^{\text{o}}(\mathbf{H}, \boldsymbol{\alpha}(\mathbf{C}, \mathbf{d}))$ is not amenable to computationally efficient solution, solving $\mathcal{P}_2$ appears to be of exponential complexity. As such, we replace $T^{\text{o}}(\mathbf{H}, \boldsymbol{\alpha}(\mathbf{C}, \mathbf{d}))$ with $T_{\text{sz}}^{\text{o}}(\mathbf{H}, \boldsymbol{\alpha}(\mathbf{C}, \mathbf{d}))$ and aim to solve

the following minimization problem:

$$\mathcal{P}_9 : \min_{\mathbf{C}} \ \mathbb{E}_{\mathbf{H},\mathbf{d}}[T_{\text{sz}}^{\text{o}}(\mathbf{H}, \boldsymbol{\alpha}(\mathbf{C}, \mathbf{d}))] \tag{3.61a}$$

$$\text{s.t.} \quad \sum_{l\in\mathcal{L}} \sum_{k\in\mathcal{K}} c_{lk} \leq C \tag{3.61b}$$

$$0 \leq c_{lk} \leq F \ , \ l \in \mathcal{L} \ , \ k \in \mathcal{K} \, . \tag{3.61c}$$

Note that optimization problem $\mathcal{P}_9$ is convex. The reason is that as shown in Lemma 3.4, $T_{\text{sz}}^{\text{o}}(\mathbf{H}, \boldsymbol{\alpha}(\mathbf{C}, \mathbf{d}))$ is a convex function of $\mathbf{C}$ for any given value of $\mathbf{H}$ and $\mathbf{d}$, while the expectation is a linear operator which is applied to such convex functions for all possible values of $\mathbf{H}$ and $\mathbf{d}$. Also, the constraints in (3.61b) and (3.61c) are affine and box constraints, respectively. Despite being a convex problem, $\mathcal{P}_9$ is not easy to solve. To explain the reason, we note that the objective function of $\mathcal{P}_9$ can be written as

$$\mathbb{E}_{\mathbf{H},\mathbf{d}}[T_{\text{sz}}^{\text{o}}(\mathbf{H}, \boldsymbol{\alpha}(\mathbf{C}, \mathbf{d}))] = \sum_{\mathbf{d}\in\mathcal{D}} \pi_{\mathbf{d}} \int T_{\text{sz}}^{\text{o}}(\mathbf{H}, \boldsymbol{\alpha}(\mathbf{C}, \mathbf{d})) f(\mathbf{H}) d\mathbf{H} \tag{3.62}$$

where $\pi_{\mathbf{d}} \triangleq \prod_{l\in\mathcal{L}} \phi_{d_l}$, for $\mathbf{d} \in \mathcal{D}$, and $f(\cdot)$ is the pdf of the channel matrix $\mathbf{H}$. The integral in (3.62) does not have a closed form as $T_{\text{sz}}^{\text{o}}(\mathbf{H}, \boldsymbol{\alpha}(\mathbf{C}, \mathbf{d}))$ is not amenable to a closed form but can be obtained numerically as the solution to (3.55). And this is the reason why the convex problem $\mathcal{P}_9$ is not easy to solve. As such, we resort to approximating the integral in (3.62) with sample averaging [54], that is

$$\mathbb{E}_{\mathbf{H},\mathbf{d}}[T_{\text{sz}}^{\text{o}}(\mathbf{H}, \boldsymbol{\alpha}(\mathbf{C}, \mathbf{d}))] = \frac{1}{N} \sum_{\mathbf{d}\in\mathcal{D}} \pi_{\mathbf{d}} \sum_{n=1}^{N} T_{\text{sz}}^{\text{o}}(\mathbf{H}^{(n)}, \boldsymbol{\alpha}(\mathbf{C}, \mathbf{d})) \tag{3.63}$$

where $\mathbf{H}^{(n)}$ is one of the N values of the channel matrix $\mathbf{H}$ that is drawn form the pdf $f(\cdot)$. In what follows, we aim to solve problem $\mathcal{P}_9$ for two different scenarios. In the first scenario, we assume that the files have equal popularities, i.e, $\phi_k = \phi_{k'}, \forall k \in$

$\mathcal{K}, k' \in \mathcal{K}$. In the second scenario, we consider the case where the file popularities are different from each other.

3.3.1 Cache Allocation for Files with Equal Popularity

Assuming that the files have equal popularities results in different values of $\mathbf{d}$ being equi-probable, that is $\pi_{\mathbf{d}} = 1/|\mathcal{D}|$, for all $\mathbf{d} \in \mathcal{D}$. Moreover, as it is intuitively obvious, the cache sizes which have to be allocated to those files have to be equal, i.e, $c_{lk} = c_{lk'} = c_l, \forall k \in \mathcal{K}, \forall k' \in \mathcal{K}, \forall l \in \mathcal{L}$, with c_l being the cache size allocated to any file, at base station l. Hence, we can write $\alpha_l = F - c_{l,d_l} = F - c_l$, which implies that vector $\boldsymbol{\alpha}$, and consequently, $T_{\text{sz}}^{\text{o}}(\mathbf{H}^{(n)}, \boldsymbol{\alpha}(\mathbf{C}, \mathbf{d}))$ does not depend on $\mathbf{d}$ anymore. Indeed, all files will be treated identically. Hence, in this case, the cache matrix $\mathbf{C}$ with size $L \times K$, can be written as $\mathbf{1}_{1\times K} \otimes \mathbf{c}$, where $\mathbf{c} \triangleq [c_1 \ \cdots \ c_L]^T \in \mathbb{R}^{L\times 1}$. Note that c_l is the cache size allocated to each of the K files in the l-th BS. In this scenario, problem $\mathcal{P}_9$ can be rewritten, with a small abuse of notation, as

$$\mathcal{P}_{10}: \min_{\mathbf{c}} \quad \frac{1}{N} \sum_{n\in\mathcal{N}} T_{\text{sz}}^{\text{o}}(\mathbf{H}^{(n)}, \boldsymbol{\alpha}(\mathbf{c}, \mathbf{1})) \tag{3.64a}$$

$$\text{s.t.} \quad \sum_{l\in\mathcal{L}} c_l \leq \frac{C}{K} \tag{3.64b}$$

$$0 \leq c_l \leq F \ , \ l \in \mathcal{L} \tag{3.64c}$$

where we set $\mathbf{d} = \mathbf{1}$ because, as mentioned earlier, the sizes of the cache allocated to different files have to be equal when the files have equal popularities, and hence, $T_{\text{sz}}^{\text{o}}(\mathbf{H}^{(n)}, \boldsymbol{\alpha}(\mathbf{c}, \mathbf{d}))$ does not depend on $\mathbf{d}$ anymore. According to Lemma 3.4, $T_{\text{sz}}^{\text{o}}(\mathbf{H}^{(n)}, \boldsymbol{\alpha}(\mathbf{c}, \mathbf{1})))$ is a convex function of $\mathbf{c}$. Hence, the optimization problem $\mathcal{P}_{10}$ is a convex programming problem, and thus, can be efficiently solved using interior point methods.

We wrap up this subsection by pointing out that in $\mathcal{P}_{10}$ substituting $T^{\mathrm{o}}_{\mathrm{sz}}(\mathbf{H}^{(n)}, \boldsymbol{\alpha}(\mathbf{c}, \mathbf{1}))$ with $T^{\mathrm{o}}_{\mathrm{lb}}(\mathbf{H}^{(n)}, \boldsymbol{\alpha}(\mathbf{c}, \mathbf{1}))$, obtained[6] from (3.60), results in a lower bound on the average delivery time. That is, solving the following optimization problem:

$$\mathcal{P}_{11}: \min_{\mathbf{c}} \quad \frac{1}{N} \sum_{n \in \mathcal{N}} T^{\mathrm{o}}_{\mathrm{lb}}(\mathbf{H}^{(n)}, \boldsymbol{\alpha}(\mathbf{c}, \mathbf{1})) \tag{3.65a}$$

$$\text{s.t.} \quad \sum_{l \in \mathcal{L}} c_l \leq \frac{C}{K} \tag{3.65b}$$

$$0 \leq c_l \leq F \ , \ l \in \mathcal{L}, \tag{3.65c}$$

leads us to a lower bound on the delivery time. Such a lower bound can and will be used in our numerical examples to benchmark the performance of our proposed scheme.

3.3.2 Cache Allocation for Files with Different Popularities

When files have different popularities, the number of terms which contribute to the sample approximate of the objective function in problem $\mathcal{P}_9$ is $N|\mathcal{D}| = NK^L$ (see (3.63)), which can be very large even for moderate values of K and L. This hinders efficiently solving this problem and makes problem $\mathcal{P}_9$ non-tractable. To tackle this issue, we replace the expectation w.r.t $\mathbf{H}$ and $\mathbf{d}$ with its sample approximation and aim to solve the following optimization problem:

$$\mathcal{P}_{12}: \min_{\mathbf{C}} \quad \frac{1}{NN'} \sum_{n'=1}^{N'} \sum_{n=1}^{N} T^{\mathrm{o}}_{\mathrm{sz}}(\mathbf{H}^{(n)}, \mathbf{C}, \mathbf{d}^{(n')}) \tag{3.66a}$$

$$\text{subject to} \quad \sum_{l \in \mathcal{L}} \sum_{k \in \mathcal{K}} c_{lk} \leq C \tag{3.66b}$$

$$0 \leq c_{lk} \leq F \ , \ l \in \mathcal{L} \ , \ k \in \mathcal{K}, \tag{3.66c}$$

[6]Note that with a small abuse of notation, we use $T^{\mathrm{o}}_{\mathrm{lb}}(\mathbf{H}^{(n)}, \boldsymbol{\alpha}(\mathbf{c}, \mathbf{1}))$, instead $T^{\mathrm{o}}_{\mathrm{lb}}(\mathbf{H}^{(n)}, \boldsymbol{\alpha}(\mathbf{C}, \mathbf{1}))$.

where $\mathbf{d}^{(n')}$ is one of the N' values of $\mathbf{d}$ that are drawn from the p.m.f $\pi_{\mathbf{d}}$. The optimization problem $\mathcal{P}_{12}$ is obviously convex and can be solved efficiently using interior point methods. Note that N and N' should be large enough to ensure the sample approximation of the statistical expectation is accurate.

Similar to subsection 3.3.1, a lower bound on the average delivery time could be found by substituting $T^{\mathrm{o}}_{\mathrm{sz}}(\mathbf{H}^{(n)}, \boldsymbol{\alpha}(\mathbf{C}, \mathbf{d}^{(n')}))$ with $T^{\mathrm{o}}_{\mathrm{lb}}(\mathbf{H}^{(n)}, \boldsymbol{\alpha}(\mathbf{C}, \mathbf{d}^{(n')}))$ in problem $\mathcal{P}_{12}$.

Chapter 4

Simulation Results

This chapter is divided into three sections. In section 4.1, we study the beamforming methods introduced in section 3.2 by comparing the performance of SPZF and MPZF for different range of parameters. In section 4.2, we evaluate caching schemes with equally popular files, and in section 4.3, we study scenarios where different files have different popularities. In all of our simulations, we consider a C-RAN model with $L = 5$ BSs which are located at different distances from the CP. The path-loss of the channel between the CP and BS l is modeled as $\lambda_l = 128.1 + 37.6 \log_{10}(d_l)$ dB, where d_l is the distance between the CP and BS l and is measured in kilometers. Moreover, during cache allocation simulations, we generate N and N_{t} sets of Rayleigh faded channel realizations which are used, respectively for training phase (i.e., to solve the cache allocation optimization problem). and for the testing phase (i.e., to evaluate the performance of different caching strategies). In case of having files with different popularities, we generate N' and N'_{t} combinations of file demands for training phase and testing phase, respectively.

Table 4.1: Simulation Parameters 1

Parameters	Values
Number of BSs	5
Backhaul channel bandwidth	20 MHz
Number of antennas at each BS	1
Maximum transmit power P at CP	40 Watts
Antenna gain, G	17 dBi
Background noise	-150 dBm/Hz
Path loss from CP to BS	$128.1 + 37.6 \log_{10}(d)$
Rayleigh small scale fading	0 dB

4.1 Comparison between SPZF and MPZF Beamforming Schemes

In this subsection, given the fact that both SPZF and MPZF provide us with an upper bound on the optimal value of delivery time T in $\mathcal{P}_1$, we compare them to see which one performs better in terms of delivery time. We also find the lower bound on T using the method we described in 3.2.2. For the parameters given in Table 4.1, Fig. 4.1 compares SPZF and MPZF by showing the average delivery time versus number of antennas, M, for 1000 different channel realizations and for 1000 different values for vector $\boldsymbol{\alpha}$. As Fig. 4.1 shows, the MPZF method works better than SPZF technique for small antenna sizes. As number of antennas is increased, downloading time of the SPZF approach improves faster than the MPZF method and approaches the lower bound on the delivery time. Hence, we can claim that the SPZF approach is asymptotically optimal for large number of antenna sizes. Note also that the SPZF

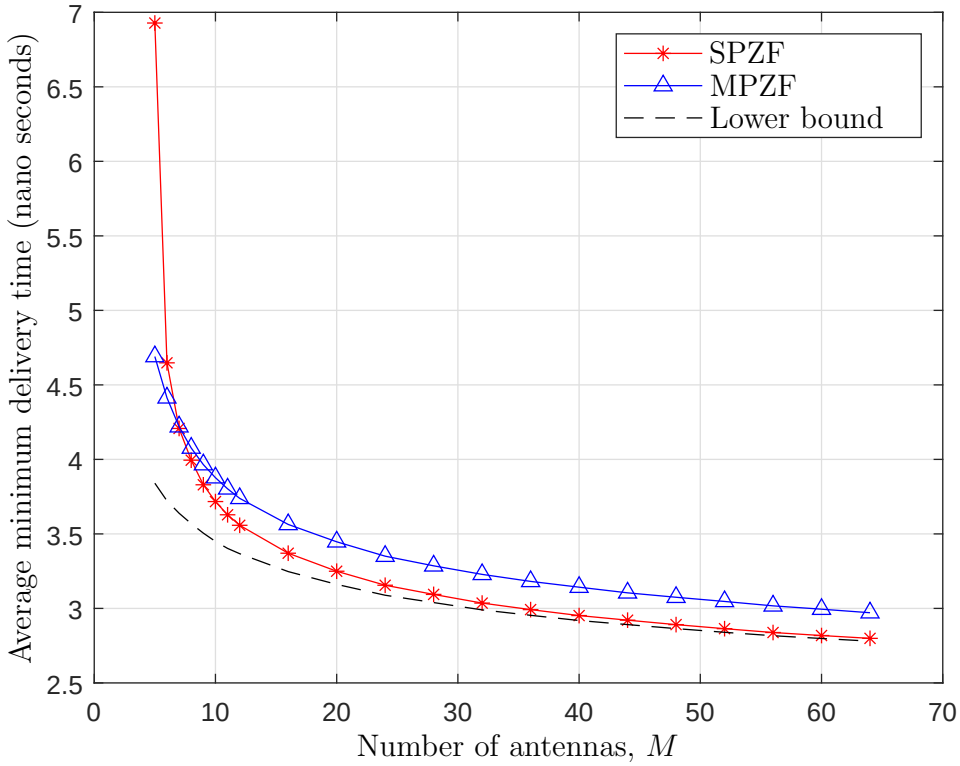

Figure 4.1: Average minimum delivery time versus number of CP's antennas, M.

approach outperforms the MPZF technique for a large range of M. As such, in the rest of our simulations, we rely on the SPZF method for the beamforming stage.

4.2 Cache Allocation for Equi-Popular Files

Relying on the SPZF method for the beamforming stage, in this subsection, we examine the performance of the caching strategy described in subsection 3.3.1 for the case when the files have equal popularities. In addition to the parameters shown in Table 4.1, this subsection relies on the parameters listed in Table 4.2.

We compare our proposed cache allocation scheme with the following strategies:

- No caching, i.e., $c_l = 0$ for all BSs;
- Uniform cache distribution, i.e., $c_l = C/(KL)$ for each BS;

Table 4.2: Simulation Parameters 2

Parameters	Values
Number of antennas at CP	128
Number of files, K	100
File size, F	100 bits
Cache budget, C	10000 bits
Training sample size, N_{tr}	100
Test sample size, N_{te}	1000

- Proportional cache distribution of [7], i.e., c_ls are chosen such that $(F-c_l)/\log_2(1+\frac{PMG}{10^{\frac{\lambda_l}{10}}L\sigma^2})$ is equalized for all l, and $\sum_{l\in\mathcal{L}} c_l = \frac{C}{K}$;
- Lower bound: we find the lower bound on average delivery time by solving problem $\mathcal{P}_{11}$.

In Fig. 4.2, we compare the cumulative distribution functions (CDFs) of the minimum delivery times of 1000 test epoches for 4 different caching schemes and contrast those against the CDF of the lower bound of the delivery times. As Fig. 4.2 shows, our proposed cache allocation scheme performs close to the proportional caching scheme. We also note that there is a large performance gap between the uniform cache distribution method and the proposed scheme (and the proportional caching scheme). As expected, the delivery times of no caching scheme are significantly larger than the corresponding delivery achieved by the proposed scheme.

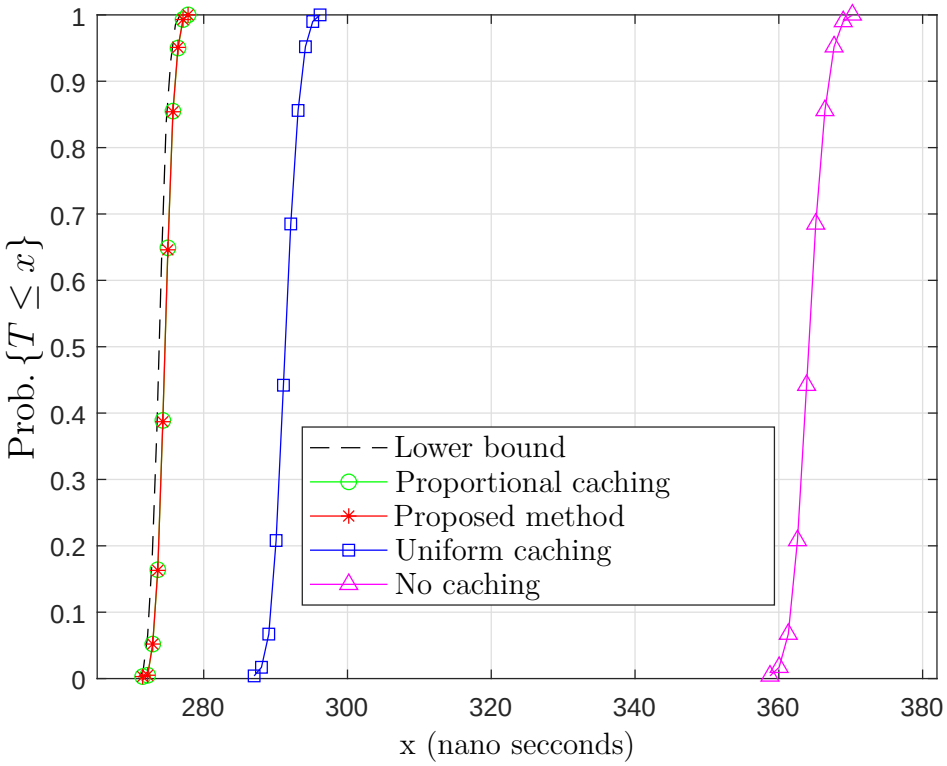

Figure 4.2: The CDF of the minimum delivery time under different caching strategies with total cache size $C = 10000$ bits and file size $F = 100$ bits.

4.3 Cache Allocation for Files with Different Popularities

In this subsection, we discuss the simulation results for a scenario where the files that need to be cached at different BSs, have different popularities. Specifically, we compare different caching schemes, and we study the effect of different simulation parameters (such as number of antennas at the CP M, total available cache size C, and Zipf exponent[1]) on the minimum expected delivery time. We compare our proposed caching schemes with the chacing strategies and the lower bound outlined in Subsection 3.3.2.

[1]The popularity of file k, denoted as ϕ_k, is the probability that file k is requested . This probablity is quite often modeled as Zipf distribution given $\phi_k = \frac{k^{-\beta}}{\sum_{i \in \mathcal{K}} i^{-\beta}}$, for $k \in \mathcal{K}$, where β is referred to as the Zipf exponent.

Table 4.3: Simulation Parameters 3

Parameters	Values
Number of files K	10
File size	100 bits
Cache budget C	1000 bits
Training sample sizes, $N_{\text{tr}}, N'_{\text{tr}}$	100, 100
Test sample sizes, $N_{\text{te}}, N'_{\text{te}}$	1000, 100

We first study the effect of M on the performance of the proposed method. Simulation parameters used in this experiment are listed in Tables 4.1 and 4.3. The Zipf exponent is chosen as $\beta = 1$. As Fig. 4.3 shows, if number of antennas at the CP is increased, the performance of our proposed scheme approaches the lower bound. Indeed for $M = 128$, the average delivery time of the proposed method is less than 0.5% over the lower bound. This implies that our proposed scheme is near optimal for a massive MIMO CP. This figure also shows our proposed solution reduces the average delivery time by almost 3 percent, as compared with the proportional caching scheme, for the given range of M.

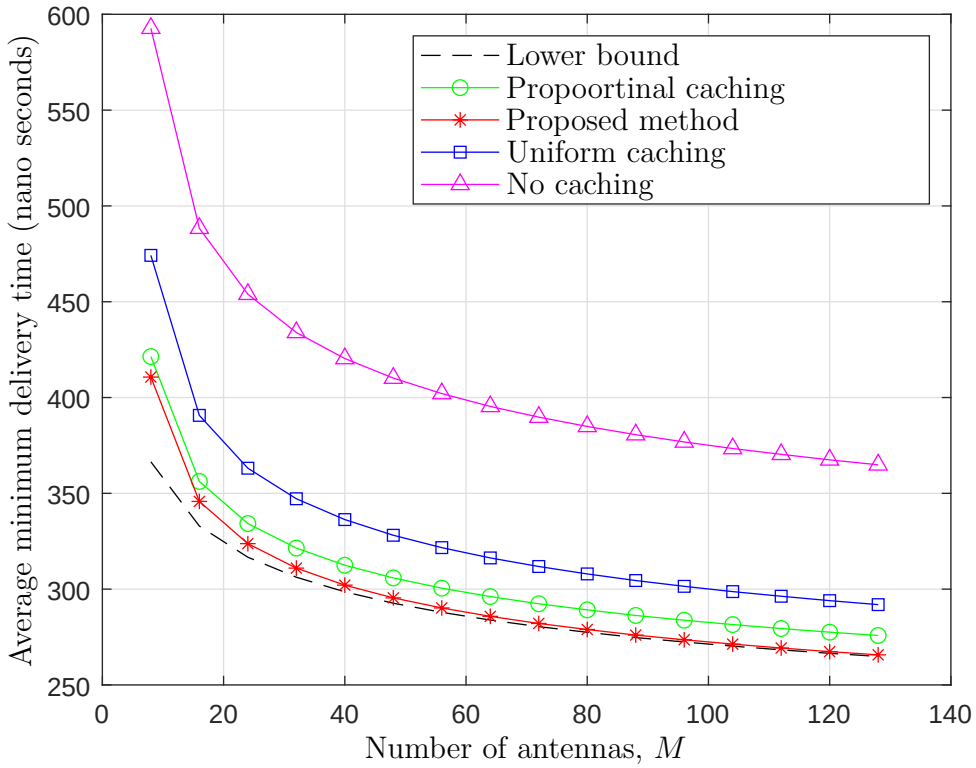

Figure 4.3: Average minimum delivery time for different caching schemes and the lower bound versus number of antennas M. for $C = 1000$ bits, $K = 100$ e $F = 100$ bits, and $\beta = 1$.

We now show an important advantage of our proposed method, as compared to the proportional caching scheme. in Fig. 4.4, assuming $M = 128$, we plot the CDF of the minimum delivery times of 1000 test epoches for the aforementioned four different caching schemes and contrast those against the CDF of the lower bound of the delivery times.

As can be seen from Fig. 4.4, as compared to the proportional caching method, our proposed method results in a lower variance in the minimum delivery times of all epochs. This results in less uncertainty in delivery time, which is a much desired feature in content delivery networks. Fig. 4.4 also shows that both our proposed caching scheme and proportional caching outperform uniform caching by a consider-

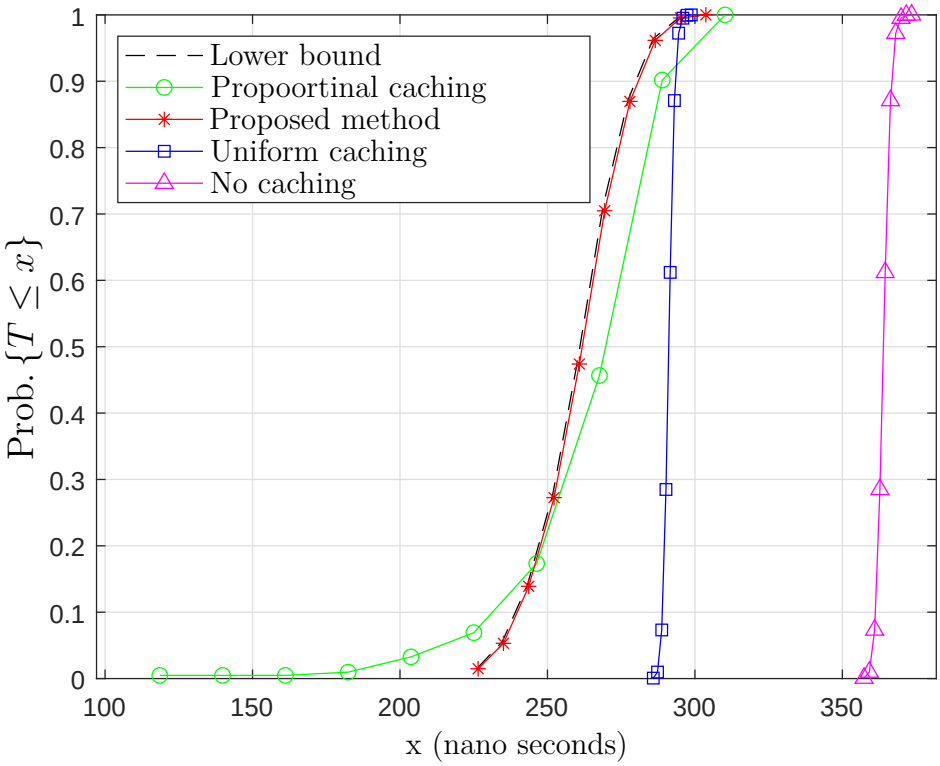

Figure 4.4: CDF of minimum delivery times under different caching strategies with total cache size $C = 1000$ bits, 100 files with different popularity and file size $F = 100$ bits.

able amount. It is worth mentioning that the coefficient of variation of the average minimum delivery time over 1000 different trials is less than 0.5% for the chosen values of N_{tr} and N'_{tr}.

In order to have a better understanding of Fig. 4.4, we have to look at Table 4.4 in which, average delivery time of different caching schemes are compared with each other. Based on Table 4.4, we can say that our optimized caching scheme results in an expected delivery time which is lower than the expected delivery time of the proportional caching scheme by almost 3%. It is also worth mentioning that the performance of our proposed scheme is within 0.5% of the lower bound.

To study the effect of Zipf exponent, β on expected minimum delivery time for

Table 4.4: Comparison of Delivery Time (ns) for Different Caching Strategies

Caching Strategy	Average	Average of lower 20-th percentile	Average of upper 80-th percentile
Lower bound	264.4	242.4	269.9
Optimized	265.3	243.1	270.8
Proportional	274.9	236.5	284.4
Uniform	291.9	289.5	292.5
No caching	364.9	261.9	365.6

different caching schemes, we use the simulation parameters used in the previous experiment except for β, which now ranges from 0 to 2. Fig. 4.5 confirms that the proposed method consistently outperforms the proportional caching method for different values of β while preform very close to the lower bound.

To study the effect of total cache budget, C on expected minimum delivery time, we use the simulation parameters similar to the first example in this subsection except for C which now changes from 0 to LKF. Note that C need not be larger than LKF. Fig. 4.6 shows how the average delivery time changes as we change C for different caching schemes. As we can see in Fig. 4.6, the proportional caching strategy underperforms our proposed solution by 20% for $C = 3000$ bits. This figure also shows that for all shown values of C, the proposed scheme performs close to the lower bound.

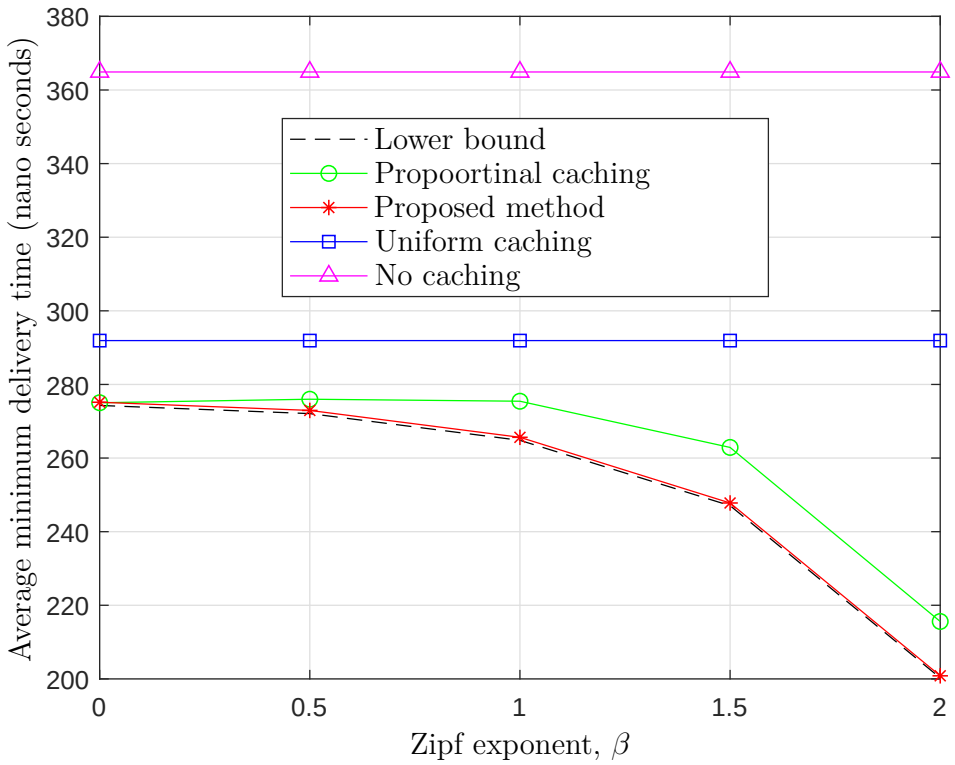

Figure 4.5: Average minimum delivery time achieved by different caching strategies versus β, for $C = 1000$ bits, $K = 10$, $M = 128$, and $F = 100$ bits.

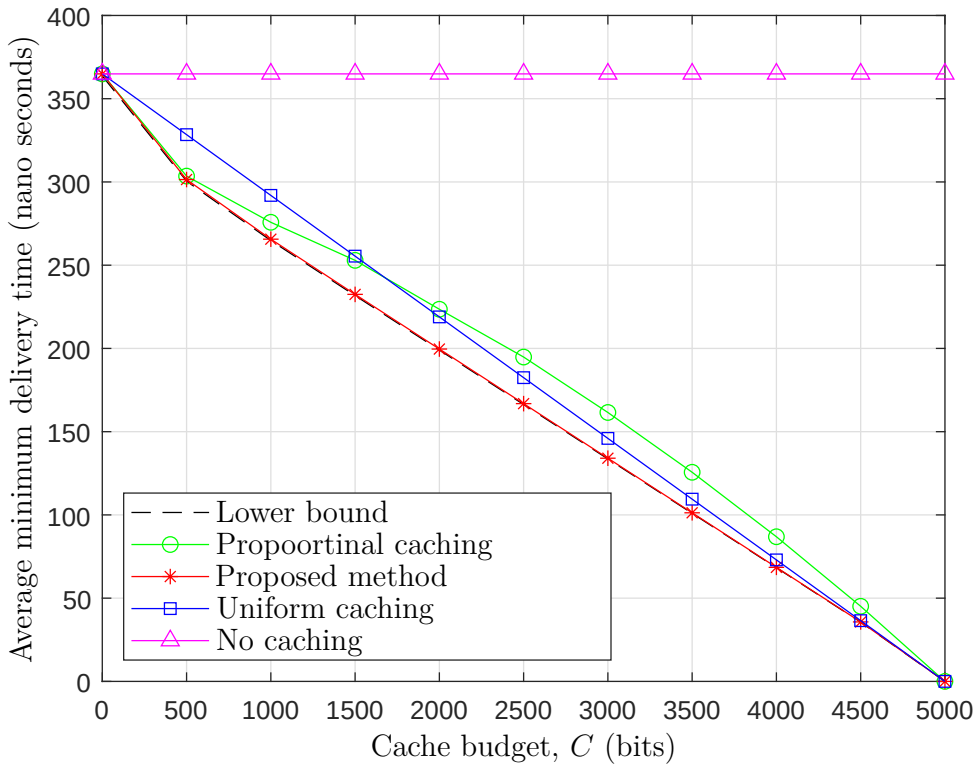

Figure 4.6: Average minimum delivery time achieved by different caching strategies versus C, for $F = 100$ bits, $M = 128$, and $\beta = 1$.

Chapter 5

Conclusions and Future Work

In this thesis, we formulated a two stage cache size allocation problem with the purpose of minimizing the average downloading time of the wireless backhaul of a C-RAN. Using ZF beamforming, we found a semi-closed form solution for the beamforming optimization stage, applying this solution to the cache size allocation problem resulted in a convex optimization problem which could efficiently be solved by many techniques. Through comparing the simulation results of our cache allocation scheme with other heuristic schemes, we verified the excellence of our scheme. Moreover, we proved that loss of optimality due to using ZF beamforming is not considerable by proposing a lower bound on the downloading time, and showing that the gap between this lower bound and the downloading time of our cache allocation scheme is negligible if large antenna sizes are used at CP.

Our work could be extended in several ways. First, we can exploit multi-cast beamforming advantages by grouping multiple BSs whose corresponding EUs require the same content to be delivered. Actually, considering the nature of content-centric communications which implies that the chance of having multiple users requesting the same file is high, adding multi-cast beamforming to our design could result in sensible

improvement of performance. Second, instead of minimization of average downloading time we could study the optimal cache allocation scheme and beamforming strategy for minimization of storage capacity, energy/power usage of the system under QoS constraints. Such a problem becomes important when user experience criteria are imposed on the system. In this case, we have to try our best to minimize the cost of system while satisfying EU requirements. Last but not least, coded caching is a promising technique which has recently been studied by lots of papers. Using coded caching technique in our work has the potential of improving the system performance by considerable amounts.

Chapter 6

Appendices

6.1 Proof of Lemma 6.1

Lemma 6.1. *Any point on the Pareto boundary of convex hull of a set of points in an L dimensional space can be represented as a convex combination of at most L points in that set.*

Proof. Consider a set of points denoted as $\mathcal{A}$ and its convex hull, $\bar{\mathcal{A}}$ in the L dimensional space. Let us denote the Pareto boundary of $\bar{\mathcal{A}}$ as $\mathcal{P}_{\bar{\mathcal{A}}}$. We now prove that any point on $\mathcal{P}_{\bar{\mathcal{A}}}$ could be represented by a convex combination of at most L points in $\mathcal{A}$. To prove this, let us suppose that $\mathbf{a}_0$ is a point on $\mathcal{P}_{\bar{\mathcal{R}}}$, and $\mathcal{Q}$ is the supporting hyperplane of $\bar{\mathcal{A}}$ at $\mathbf{a}_0$ whose equation is

$$\mathbf{q}^T\mathbf{a} = c_{\mathcal{Q}} \tag{6.1}$$

where vector $\mathbf{q}$ is the normal of $\mathcal{Q}$ with an outward direction with respect to $\bar{\mathcal{A}}$. Moreover, we suppose that $\hat{\mathcal{A}} \triangleq \{\mathbf{a}_1, \ldots, \mathbf{a}_{L'}\}$ is a set of rate vectors in $\mathcal{A}$ whose convex combination results in $\mathbf{a}_0$, i.e., there exists a vector $\boldsymbol{\tau} = [\tau_1 \ \cdots \ \tau_{L'}]^T \in \mathbb{R}_{++}^{L'\times 1}$

such that

$$\mathbf{a}_0 = \sum_{l=1}^{L'} \tau_l \mathbf{a}_l \tag{6.2}$$

while

$$\sum_{l=1}^{L'} \tau_l = 1. \tag{6.3}$$

Because $\mathcal{Q}$ is a supporting hyperplane, each one of the rate vectors in set $\hat{\mathcal{A}}$ has to be located either on one side or on the hyperplane of $\mathcal{Q}$. Considering this fact along with the assumption that the direction of $\mathbf{q}$ is outward with respect to $\bar{\mathcal{R}}$ result in

$$\mathbf{q}^T \mathbf{a}_l \leq c_{\mathcal{Q}} \ , \ l = 1, \dots L'. \tag{6.4}$$

Multiplying $\mathbf{q}^T$ by both sides of (6.2) results in

$$c_{\mathcal{Q}} = \sum_{l=1}^{L'} \tau_l (\mathbf{q}^T \mathbf{a}_l) \tag{6.5}$$

in which we have considered the fact that $\mathbf{a}_0$ is a point on $\mathcal{Q}$. Considering equations (6.3), (6.4) and (6.5) clearly indicates that $\mathbf{q}^T \mathbf{a}_l = c_{\mathcal{Q}}$, $l = 1, \dots L'$, i.e., any rate vector of $\hat{\mathcal{A}}$ is located on $\mathcal{Q}$. Hence, all the members of $\hat{\mathcal{A}}$ and $\mathbf{a}_0$ are located on $\mathcal{Q}$. Note that $\mathcal{Q}$ is an $L-1$ dimensional space. In this space, $\mathbf{a}_0$ is a member of the convex hull of $\hat{\mathcal{A}}$. Based on Caratheodory theorem [55], $\mathbf{a}_0$ could be represented by a convex combination of at most $(L-1)+1$ points of $\hat{\mathcal{R}}$. Therefore, either the number of the members of $\hat{\mathcal{A}}$ is less than or equal L, or there is a convex combination of at most L members of $\hat{\mathcal{A}}$ that is equal to $\mathbf{a}_0$. □

6.2 Proof of Convexity of $\mathcal{R}_{\mathbf{z}}^{i}$

Suppose that $\mathbf{r}_i^1 = [r_{1i}^1 \quad \cdots \quad r_{Li}^1]^T$ and $\mathbf{r}_i^2 = [r_{1i}^2 \quad \cdots \quad r_{Li}^2]^T$ are two rate vectors in $\mathcal{R}_{\mathbf{z}}^{i}$ resulting from power vectors $\mathbf{p}_i^1 = [p_{1i}^1 \quad \cdots \quad p_{Li}^1]^T$ and $\mathbf{p}_i^2 = [p_{1i}^2 \quad \cdots \quad p_{Li}^2]^T$, respectively. We now show that for any pair of positive coefficients τ_1 and τ_2 whose summation is 1, $\tau_1\mathbf{r}_i^1 + \tau_2\mathbf{r}_i^2$ is also a member of $\mathcal{R}_{\mathbf{z}}^{i}$. Since $\mathbf{r}_i^1$ and $\mathbf{r}_i^2$ belong to $\mathcal{R}_{\mathbf{z}}^{i}$, we have the following equations:

$$\sum_{l\in\mathcal{S}_i} p_{li}^j \leq P \ , \ j = 1, 2 \tag{6.6}$$

$$r_{li}^j = B\log_2(1 + p_{li}^j\gamma_{li}(\mathbf{H})) \ , \ l \in \mathcal{S}_i \ , \ j = 1, 2 \tag{6.7}$$

$$r_{li}^j = 0 \ , \ l \notin \mathcal{S}_i \ , \ j = 1, 2. \tag{6.8}$$

Now, consider a new power vector $\mathbf{p}_i^3 \triangleq [p_{1i}^3 \quad \cdots \quad p_{Li}^3]^T = \tau_1\mathbf{p}_i^1 + \tau_2\mathbf{p}_i^2$. It is obvious that the power constraints, $\sum_{l\in\mathcal{S}_i} p_{li} \leq P$, $p_{li} > 0$, $l \in \mathcal{S}_i$ and $p_{li} = 0$, $l \notin \mathcal{S}_i$ are all satisfied for $\mathbf{p}_i^3$. This means that the rate vector $\mathbf{r}_i^3$ resulting from $\mathbf{p}_i^3$ is a member of $\mathcal{R}_{\mathbf{z}}^{i}$. The elements of vector $\mathbf{r}_i^3 \triangleq [r_{1i}^3 \quad \cdots \quad r_{Li}^3]^T$ are as follows

$$r_{li}^3 = B\log_2(1 + (\tau_1 p_{li}^1 + \tau_2 p_{li}^2)\gamma_{li}(\mathbf{H})) \ , \ l \in \mathcal{S}_i \tag{6.9}$$

$$r_{li}^3 = 0 \ , \ l \notin \mathcal{S}_i. \tag{6.10}$$

Because $B\log_2(1 + p_{li}\gamma_{li}(\mathbf{H}))$ is a concave function of p_{li}, We have the following inequality:

$$r_{li}^3 \geq \tau_1 r_{li}^1 + \tau_2 r_{li}^2 \ , \ l \in \mathcal{S}_i. \tag{6.11}$$

Based on (6.8) and (6.10), the following equality also holds:

$$r_{li}^3 = \tau_1 r_{li}^1 + \tau_2 r_{li}^2 = 0 \ , \ l \notin \mathcal{S}_i. \tag{6.12}$$

Based on (6.11) and (6.12), it is obvious that we can find a power vector $\mathbf{p}'_i \preccurlyeq \mathbf{p}^3_i$ whose corresponding rate vector $\mathbf{r}'_i$ satisfies the following equation:

$$\mathbf{r}'_i = \tau_1 \mathbf{r}^1_i + \tau_2 \mathbf{r}^2_i. \tag{6.13}$$

Because $\mathbf{p}'_i \preccurlyeq \mathbf{p}^3_i$ and $\mathbf{p}^3_i$ satisfies the power constraint, $\mathbf{p}'_i$ also satisfies the power constraint. Hence $\mathbf{r}'_i$ is in $\mathcal{R}^i_z$.

6.3 Proof of Convexity of $T^{\mathrm{o}}_{\mathbf{sz}}(\mathbf{H}, \boldsymbol{\alpha}(\mathbf{C}, \mathbf{d}))$

To show that the optimal value of the objective function of problem $\mathcal{P}_8$, denoted as $T^{\mathrm{o}}_{\mathrm{sz}}(\mathbf{H}, \boldsymbol{\alpha}(\mathbf{C}, \mathbf{d}))$ is a convex function of $\boldsymbol{\alpha}(\mathbf{C}, \mathbf{d})$, let us suppose the optimal value of problem $\mathcal{P}_8$ for vectors $\boldsymbol{\alpha}_1 = \boldsymbol{\alpha}(\mathbf{C}_1, \mathbf{d}_1) = [\alpha_{11}(\mathbf{C}_1, \mathbf{d}_1) \quad \cdots \quad \alpha_{L1}(\mathbf{C}_1, \mathbf{d}_1)]^T$ and $\boldsymbol{\alpha}_2 = \boldsymbol{\alpha}(\mathbf{C}_2, \mathbf{d}_2) = [\alpha_{12}(\mathbf{C}_2, \mathbf{d}_2) \quad \cdots \quad \alpha_{L2}(\mathbf{C}_2, \mathbf{d}_2)]^T$ are $T_1 = T^{\mathrm{o}}_{\mathrm{sz}}(\mathbf{H}, \boldsymbol{\alpha}_1)$ and $T_2 = T^{\mathrm{o}}_{\mathrm{sz}}(\mathbf{H}, \boldsymbol{\alpha}_2)$, respectively. Moreover, $\mathbf{p}^1 = [p^1_{1i} \quad \cdots \quad p^1_{Li}]^T \in \mathbb{R}^{L\times 1}$ and $\mathbf{p}^2 = [p^2_{1i} \quad \cdots \quad p^2_{Li}]^T \in \mathbb{R}^{L\times 1}$ are considered as the optimal power vectors obtained by solving problem $\mathcal{P}_8$, when vector $\boldsymbol{\alpha}$ is equal to $\boldsymbol{\alpha}_1$ and $\boldsymbol{\alpha}_2$, respectively. We can thus write

$$\sum_{l\in\mathcal{L}} p^j_{li} \leq P \ , \ j = 1, 2 \tag{6.14}$$

$$T_j \mathbf{r}^j = \boldsymbol{\alpha}_j \ , \ j = 1, 2 \tag{6.15}$$

where $\mathbf{r}^j \triangleq [r^j_1 \quad \cdots \quad r^j_L]^T \in \mathbb{R}^{L\times 1}$ and $r^j_l = B\log_2(1 + p^j_{li}\gamma_{li}(\mathbf{H}))$, $j = 1, 2$, $l \in \mathcal{L}$. Based on (6.14), $\mathbf{r}^1$ and $\mathbf{r}^2$ belong to $\mathcal{R}^i_z$. Based on (6.15), we can write

$$(\tau_1 T_1 + \tau_2 T_2)\frac{\tau_1 T_1 \mathbf{r}^1 + \tau_2 T_2 \mathbf{r}^2}{\tau_1 T_1 + \tau_2 T_2} = \tau_1 \boldsymbol{\alpha}_1 + \tau_2 \boldsymbol{\alpha}_2 \tag{6.16}$$

where $\tau_1 + \tau_2 = 1$ and $\tau_j \geq 0$, $j = 1, 2$. In Appendix 6.2, we showed that $\mathcal{R}_{\text{z}}^i$ is a convex set. Let us consider the rate vector $\mathbf{r}^3 = \frac{\tau_1 T_1 \mathbf{r}^1 + \tau_2 T_2 \mathbf{r}^2}{\tau_1 T_1 + \tau_2 T_2}$ which is indeed a convex combination of $\mathbf{r}^1$ and $\mathbf{r}^2$. Since $\mathcal{R}_{\text{z}}^i$ is a convex set, $\mathbf{r}^3$ also has to belong to $\mathcal{R}_{\text{z}}^i$. Let us suppose $\mathbf{p}^3 = [p_{1i}^3 \ \cdots \ p_{Li}^3]^T \in \mathbb{R}^{L \times 1}$ is the power vector that results in rate vector $\mathbf{r}^3$. As such, $\mathbf{p}^3$ satisfies the power constraint $\sum_{l \in \mathcal{L}} p_{li}^3 \leq P$ while, based on (6.16) , its corresponding rate vector, $\mathbf{r}^3$, satisfies the following equation:

$$(\tau_1 T_1 + \tau_2 T_2)\mathbf{r}^3 = \tau_1 \boldsymbol{\alpha}_1 + \tau_2 \boldsymbol{\alpha}_2. \tag{6.17}$$

Using (6.17) along with the fact that $\mathbf{p}_3$ satisfies the power constraint $\sum_{l \in \mathcal{L}} p_{li}^3 \leq P$, we deduct that if we set $\boldsymbol{\alpha} = \tau_1 \boldsymbol{\alpha}_1 + \tau_2 \boldsymbol{\alpha}_2$ in problem $\mathcal{P}_8$, there is at least one feasible solution to problem $\mathcal{P}_8$ whose corresponding optimal delivery time is equal to $\tau_1 T_1 + \tau_2 T_2$. Hence, $T_{\text{sz}}^{\text{o}}(\mathbf{H}, \boldsymbol{\alpha}_3)$ has to satisfy

$$T_{\text{sz}}^{\text{o}}(\mathbf{H}, \boldsymbol{\alpha}_3) \leq \tau_1 T_1 + \tau_2 T_2 \tag{6.18}$$

which implies that $T_{\text{sz}}^{\text{o}}(\mathbf{H}, \boldsymbol{\alpha}(\mathbf{C}, \mathbf{d}))$ is a convex function of $\boldsymbol{\alpha}(\mathbf{C}, \mathbf{d})$ in problem $\mathcal{P}_8$.

Using an approach similar to what we did above, we can show that $T_{\text{lb}}^{\text{o}}(\mathbf{H}, \boldsymbol{\alpha}(\mathbf{C}, \mathbf{d}))$, (i.e., the optimal value of T for $\mathcal{P}_{11}$ which is obtained by solving (3.60)) is also a convex function of $\boldsymbol{\alpha}(\mathbf{C}, \mathbf{d})$.

Bibliography

[1] P. Rost, C. J. Bernardos, A. D. Domenico, M. D. Girolamo, M. Lalam, A. Maeder, D. Sabella, and D. Wubben, "Cloud technologies for flexible 5G radio access networks," *IEEE Commun. Mag.*, vol. 52, no. 5, pp. 68–76, 2014.

[2] O. Simeone, A. Maeder, M. Peng, O. Sahin, and W. Yu, "Cloud radio access network: Virtualizing wireless access for dense heterogeneous systems," *Journal of Communications and Networks*, vol. 18, no. 2, pp. 135–149, 2016.

[3] T. Q. S. Quek, M. Peng, O. Simeone, and W. Yu, *Cloud Radio Access Networks: Principles, Technologies, and Applications*. Cambridge University Press, 2017.

[4] B. Dai and W. Yu, "Energy efficiency of downlink transmission strategies for cloud radio access networks," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 4, pp. 1037–1050, 2016.

[5] H. Liu, Z. Chen, X. Tian, X. Wang, and M. Tao, "On content-centric wireless delivery networks," *IEEE Wireless Commun. Mag.*, vol. 21, no. 6, pp. 118–125, 2014.

[6] M. Tao, E. Chen, H. Zhou, and W. Yu, "Content-centric sparse multicast beamforming for cache-enabled cloud RAN," *IEEE Trans. Wireless Commun.*, vol. 15, no. 9, pp. 6118–6131, 2016.

[7] B. Dai, Y.-F. Liu, and W. Yu, "Optimized base-station cache allocation for cloud radio access network with multicast backhaul," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 8, pp. 1737–1750, 2018.

[8] M. A. Maddah-Ali and U. Niesen, "Fundamental limits of caching," *IEEE Trans. Inform. Theory*, vol. 60, no. 5, pp. 2856–2867, 2014.

[9] F. Xu, M. Tao, and K. Liu, "Fundamental tradeoff between storage and latency in cache-aided wireless interference networks," *IEEE Trans. Inform. Theory*, vol. 63, no. 11, pp. 7464–7491, 2017.

[10] A. Sengupta, R. Tandon, and O. Simeone, "Cache aided wireless networks: Tradeoffs between storage and latency," in *Proc. of Annual Conf. on Information Sciences and Systems (CISS)*, 2016, pp. 320–325.

[11] S. Gitzenis, G. S. Paschos, and L. Tassiulas, "Asymptotic laws for joint content replication and delivery in wireless networks," *IEEE Trans. Inform. Theory*, vol. 59, no. 5, pp. 2760–2776, 2013.

[12] K. Shanmugam, N. Golrezaei, A. G. Dimakis, A. F. Molisch, and G. Caire, "Femtocaching: Wireless content delivery through distributed caching helpers," *IEEE Trans. Inform. Theory*, vol. 59, no. 12, pp. 8402–8413, 2013.

[13] E. Bastug, M. Bennis, and M. Debbah, "Cache-enabled small cell networks: Modeling and tradeoffs," in *2014 11th International Symposium on Wireless Communications Systems (ISWCS)*, 2014, pp. 649–653.

[14] Y. Cui, D. Jiang, and Y. Wu, "Analysis and optimization of caching and multicasting in large-scale cache-enabled wireless networks," *IEEE Trans. Wireless Commun.*, vol. 15, no. 7, pp. 5101–5112, 2016.

[15] X. Xu and M. Tao, "Modeling, analysis, and optimization of coded caching in small-cell networks," *IEEE Trans. Commun.*, vol. 65, no. 8, pp. 3415–3428, 2017.

[16] Y. Ugur, Z. H. Awan, and A. Sezgin, "Cloud radio access networks with coded caching," in *WSA 2016; 20th International ITG Workshop on Smart Antennas*, 2016, pp. 1–5.

[17] S.-H. Park, O. Simeone, and S. Shamai Shitz, "Joint optimization of cloud and edge processing for fog radio access networks," *IEEE Trans. Wireless Commun.*, vol. 15, no. 11, pp. 7621–7632, 2016.

[18] Y. Li, M. Xia, and Y.-C. Wu, "First-order algorithm for content-centric sparse multicast beamforming in large-scale C-RAN," *IEEE Trans. Wireless Commun.*, vol. 17, no. 9, pp. 5959–5974, 2018.

[19] Y. Li, M. Xia, and Y. Wu, "Caching at base stations with multi-cluster multicast wireless backhaul via accelerated first-order algorithms," *IEEE Trans. Wireless Commun.*, vol. 19, no. 5, pp. 2920–2933, 2020.

[20] F. Cheng, Y. Yu, Z. Zhao, N. Zhao, Y. Chen, and H. Lin, "Power allocation for cache-aided small-cell networks with limited backhaul," *IEEE Access*, vol. 5, pp. 1272–1283, 2017.

[21] S. Borst, V. Gupta, and A. Walid, "Distributed caching algorithms for content distribution networks," in *2010 Proceedings IEEE INFOCOM*, 2010, pp. 1–9.

[22] P. Krishnan, D. Raz, and Y. Shavitt, "The cache location problem," *IEEE/ACM Transactions on Networking*, vol. 8, no. 5, pp. 568–582, 2000.

[23] M. Korupolu and M. Dahlin, "Coordinated placement and replacement for large-scale distributed caches," *IEEE Transactions on Knowledge and Data Engineering*, vol. 14, no. 6, pp. 1317–1329, 2002.

[24] A. Wiesel, Y. Eldar, and S. Shamai, "Linear precoding via conic optimization for fixed MIMO receivers," *IEEE Trans. Signal Process.*, vol. 54, no. 1, pp. 161–176, 2006.

[25] E. Bjornson, M. Bengtsson, and B. Ottersten, "Optimal multiuser transmit beamforming: A difficult problem with a simple solution structure [lecture notes]," *IEEE Signal Processing Mag.*, vol. 31, no. 4, pp. 142–148, 2014.

[26] Y.-F. Liu, Y.-H. Dai, and Z.-Q. Luo, "Coordinated beamforming for MISO interference channel: Complexity analysis and efficient algorithms," *IEEE Trans. Signal Process.*, vol. 59, no. 3, pp. 1142–1157, 2011.

[27] N. Sidiropoulos, T. Davidson, and Z.-Q. Luo, "Transmit beamforming for physical-layer multicasting," *IEEE Trans. Signal Process.*, vol. 54, no. 6, pp. 2239–2251, 2006.

[28] Z.-Q. Luo, N. D. Sidiropoulos, P. Tseng, and S. Zhang, "Approximation bounds for quadratic optimization with homogeneous quadratic constraints," *SIAM Journal on optimization*, vol. 18, no. 1, pp. 1–28, 2007.

[29] S. X. Wu, W.-K. Ma, and A. M.-C. So, "Physical-layer multicasting by stochastic transmit beamforming and alamouti space-time coding," *IEEE Trans. Signal Process.*, vol. 61, no. 17, pp. 4230–4245, 2013.

[30] A. Abdelkader, A. B. Gershman, and N. D. Sidiropoulos, "Multiple-antenna multicasting using channel orthogonalization and local refinement," *IEEE Trans. Signal Process.*, vol. 58, no. 7, pp. 3922–3927, 2010.

[31] L.-N. Tran, M. F. Hanif, and M. Juntti, "A conic quadratic programming approach to physical layer multicasting for large-scale antenna arrays," *IEEE Signal Processing Lett.*, vol. 21, no. 1, pp. 114–117, 2014.

[32] B. Gopalakrishnan and N. D. Sidiropoulos, "High performance adaptive algorithms for single-group multicast beamforming," *IEEE Trans. Signal Process.*, vol. 63, no. 16, pp. 4373–4384, 2015.

[33] A. Konar and N. D. Sidiropoulos, "Fast approximation algorithms for a class of non-convex QCQP problems using first-order methods," *IEEE Trans. Signal Process.*, vol. 65, no. 13, pp. 3494–3509, 2017.

[34] R. Jiang, H. Liu, and A. M.-C. So, "LPA-SD: An efficient first-order method for single-group multicast beamforming," in *Proc. IEEE Workshop on Signal Processing advances in Wireless Commun.(SPAWC)*, 2018, pp. 1–5.

[35] E. A. Jorswieck, E. G. Larsson, and D. Danev, "Complete characterization of the pareto boundary for the MISO interference channel," *IEEE Trans. Signal Process.*, vol. 56, no. 10, pp. 5292–5296, 2008.

[36] A. Carleial, "Interference channels," *IEEE Trans. Inform. Theory*, vol. 24, no. 1, pp. 60–70, 1978.

[37] S. Vishwanath and S. Jafar, "On the capacity of vector gaussian interference channels," in *Information Theory Workshop*, 2004, pp. 365–369.

[38] M. Charafeddine, A. Sezgin, and A. Paulraj, "Rate region frontiers for n-user interference channel with interference as noise," *ArXiv*, vol. abs/1008.3437, 2010.

[39] X. Shang and B. Chen, "Achievable rate region for downlink beamforming in the presence of interference," in *2007 Conference Record of the Forty-First Asilomar Conference on Signals, Systems and Computers*, 2007, pp. 1684–1688.

[40] E. A. Jorswieck and E. G. Larsson, "The MISO interference channel from a game-theoretic perspective: A combination of selfishness and altruism achieves pareto optimality," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Process. (ICASSP)*, 2008, pp. 5364–5367.

[41] ——, "Linear precoding in multiple antenna broadcast channels: Efficient computation of the achievable rate region," in *2008 International ITG Workshop on Smart Antennas*, 2008, pp. 21–28.

[42] J. Lindblom, E. Karipidis, and E. G. Larsson, "Closed-form parameterization of the pareto boundary for the two-user MISO interference channel," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Process. (ICASSP)*, 2011, pp. 3372–3375.

[43] J. Qiu, R. Zhang, Z.-Q. Luo, and S. Cui, "Optimal distributed beamforming for MISO interference channels," *IEEE Trans. Signal Process.*, vol. 59, no. 11, pp. 5638–5643, 2011.

[44] C. Li, C. He, and L. Jiang, "Distributed beamforming design of the pareto boundary for MISO interference channels," in *2014 IEEE/CIC International Conference on Communications in China (ICCC)*, 2014, pp. 748–752.

[45] J. Li, D. Wang, P. Zhu, L. Tang, and X. You, "Closed-form solutions to the pareto boundary and optimal distributed strategy for the two-user MISO interference channel," in *2012 International Conference on Wireless Communications and Signal Processing (WCSP)*, 2012, pp. 1–6.

[46] Q. Spencer, A. Swindlehurst, and M. Haardt, "Zero-forcing methods for downlink spatial multiplexing in multiuser MIMO channels," *IEEE Trans. Signal Process.*, vol. 52, no. 2, pp. 461–471, 2004.

[47] C. Peel, B. Hochwald, and A. Swindlehurst, "A vector-perturbation technique for near-capacity multiantenna multiuser communication-part i: channel inversion and regularization," *IEEE Trans. Commun.*, vol. 53, no. 1, pp. 195–202, 2005.

[48] T. Yoo and A. Goldsmith, "Optimality of zero-forcing beamforming with multiuser diversity," in *Proc. IEEE Int. Conf. Communications (ICC)*, vol. 1, 2005, pp. 542–546 Vol. 1.

[49] A. Wiesel, Y. C. Eldar, and S. Shamai, "Zero-forcing precoding and generalized inverses," *IEEE Trans. Signal Process.*, vol. 56, no. 9, pp. 4409–4418, 2008.

[50] F. Boccardi and H. Huang, "Optimum power allocation for the MIMO-BC zero-forcing precoder with per-antenna power constraints," in *2006 40th Annual Conference on Information Sciences and Systems*, 2006, pp. 504–504.

[51] K. Karakayali, R. Yates, G. Foschini, and R. Valenzuela, "Optimum zero-forcing beamforming with per-antenna power constraints," in *Proc. IEEE Int. Symp. on Infor. Theory (ISIT)*, 2007, pp. 101–105.

[52] S.-R. Lee, J.-S. Kim, S.-H. Moon, H.-B. Kong, and I. Lee, "Zero-forcing beamforming in multiuser MISO downlink systems under per-antenna power constraint and equal-rate metric," *IEEE Trans. Wireless Commun.*, vol. 12, no. 1, pp. 228–236, 2013.

[53] H. Yang and T. L. Marzetta, "Performance of conjugate and zero-forcing beamforming in large-scale antenna systems," *IEEE J. Sel. Areas Commun.*, vol. 31, no. 2, pp. 172–179, 2013.

[54] J. R. Birge and F. Louveaux, *Introduction to Stochastic Programming*. Springer, 2011.

[55] G. E. Danninger-Uchida, *Carathéodory theorem*. Boston, MA: Springer US, 2001, pp. 236–237. [Online]. Available: https://doi.org/10.1007/0-306-48332-7_51