Optimizing Resource Allocation with Energy Efficiency and Backhaul Challenges

Jialing Liao

A dissertation submitted in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

of

University College London.

Department of Electrical and Electronic Engineering

University College London

December 19, 2018

I, Jialing Liao, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the work.

Abstract

To meet the requirements of future wireless mobile communication which aims to increase the data rates, coverage and reliability while reducing energy consumption and latency, and also deal with the explosive mobile traffic growth which imposes high demands on backhaul for massive content delivery, developing green communication and reducing the backhaul requirements have become two significant trends. One of the promising techniques to provide green communication is wireless power transfer (WPT) which facilitates energy-efficient architectures, e.g. simultaneous wireless information and power transfer (SWIPT). Edge caching, on the other side, brings content closer to the users by storing popular content in caches installed at the network edge to reduce peak-time traffic, backhaul cost and latency. In this thesis, we focus on the resource allocation technology for emerging network architectures, i.e. the SWIPT-enabled multiple-antenna systems and cache-enabled cellular systems, to tackle the challenges of limited resources such as insufficient energy supply and backhaul capacity. We start with the joint design of beamforming and power transfer ratios for SWIPT in MISO broadcast channels and MIMO relay systems, respectively, aiming for maximizing the energy efficiency subject to both the Quality of Service (QoS) constraints and energy harvesting constraints. Then move to the content placement optimization for cache-enabled heterogeneous small cell networks so as to minimize the backhaul requirements. In particular, we enable multicast content delivery and cooperative content sharing utilizing maximum distance separable (MDS) codes to provide further caching gains. Both analysis and simulation results are provided throughout the thesis to demonstrate the benefits of the proposed algorithms over the state-of-the-art methods.

Impact Statement

This thesis mainly contributes in developing resource allocation technology for wireless mobile networks under the challenges of limited resources such as insufficient energy supply and backhaul capacity. In particular, we have studied the joint design of beamforming and power transfer ratios for SWIPT in MISO broadcast channels and MIMO relay systems, and optimized the content placement for cache-enabled heterogeneous small cell networks taking the advantages of coded caching, multicast content delivery and cooperative content sharing.

The significance of the proposed resource allocation techniques in academia mainly comes from the following aspects: (i) responding to the requirements of future wireless mobile communication for increasing the data rates, coverage and reliability while reducing energy consumption and latency; (ii) developing green communication and improving energy efficiency; (iii) dealing with the high demands on backhaul for massive content delivery imposed by the explosive mobile traffic growth. SWIPT has been recognized as an important mechanism for battery-limited mobile communications. The trade-off between information decoding and energy harvesting, the combination with multi-antenna technologies, and the imperfect CSI scenario, which we have discussed, are all essential issues to guarantee the performance of the SWIPT systems. As one of the key techniques enabling fog radio access network (F-RAN), Internet of Things (IoT), edge caching provides an effective means to facilitate caching, computing and communication (3C) services to provide more flexible and intelligent connection. Our research provides some prior results for edge caching in cellular networks, and also opens up a series of new directions for further research: mobility-aware caching, content popularity predic-

tion and evolution, privacy preservation, joint caching and transmission design and the combination of emerging networks and technologies.

From the perspective of industry, resource allocation technology for wireless mobile networks also plays an important role. Although the SWIPT technology has not been widely applied to industry due to some technical reasons, e.g. the limited energy harvesting rate, the continuous efforts and attempt offer potentials for the application of SWIPT technology in industry. Edge caching, on the other hand, has attracted lots of attention in information technology companies. For instance, Google has launched the Google Global Cache (GCC) system which aims to serve locally requests for YouTube content while reducing backhaul costs. Netflix has focused on studying spatio-temporal demands, popularity prediction, and caching mechanisms. Hierarchical caching mechanism has been utilized by Facebook for delivering pictures to users. Thanks to the efforts of many companies, 3GPP Standards have mentioned implementing edge caching in wireless networks. And we believe that the standardization efforts will in turn motivate further research in edge caching in wireless networks in both academic and industry.

Acknowledgements

Firstly, I would like to express my sincere gratitude and appreciation to my supervisor, Prof. Kai-Kit Wong, for granting me the opportunity to pursue my PhD degree at UCL, guiding me through my first steps in research and sharing his knowledge and experience. I would also like to thank Dr. Miguel Rodrigues and Dr. Xu Zhu for serving as my thesis committee members. Moreover, I thank all the group members and the lab mates in both Boston House and MPEB 708 for creating a friendly and inspirational atmosphere for research.

Many thanks to all of my friends met in London, Wenting, Jie, James, Victoria, Willie and Qun, for sharing the joy and sadness altogether so that I have had lots of joyful times and also found the courage to get through bad moments. I also want to express my gratitude to my friends in China for their warm greetings now and then.

Last but not least, my deepest thanks go to my sister for always being on my side and giving me strengths every time when I face difficulties. Only you can understand how I get here, and I would like to dedicate this work to you for your unconditional encouragement and support.

Contents

List of Figures

List of Abbreviations

5G	Fifth-Generation
QoS	Quality of Service
EH	Energy Harvesting
RF	Radio Frequency
WPT	Wireless Power Transfer
BS	Base Station
SWIPT	Simultaneous Wireless Information and Power Transfer
CR	Cognitive Radio
MIMO	Multiple-Input Multiple-Output
D2D	Device-to-Device
NOMA	Non-Orthogonal Multiple Access
mmWave	Millimeter Wave
HetNet	Heterogeneous Network
CoMP	Coordinated Multipoint
CSI	Channel State Information
IoT	Internet of Things

MTC	Machine Type Communication
CAPEX	Capital Expenditure
OPEX	Operational Expenditure
UT	User Terminal
SINR	Signal-to-Interference-Plus-Noise Ratio
TS	Time Switching
MDS	Maximum Distance Separable
MINLP	Mixed Integer Nonlinear Program
UA	User Attrition
PS	Power Splitting
AS	Antenna Switching
SS	Spatial Switching
OFDM	Orthogonal Frequency-Division Multiplexing
AN	Artificial Noise
SDP	Semi-Definite Programming
SOCP	Second Order Cone Programming
SCA	Successive Convex Approximation
CDN	Content Distribution Networking
ICN	Information Centric Networking
LFU	Least Frequently Used
LRU	Least Recently Used

F-RAN	Fog Random Access Network
C-RAN	Cloud-Radio Access Network
BBU	Baseband Unit
RRH	Remote Radio Head
SBS	Small Cell Base Station
MBS	Macro Base Station
PPP	Poisson Point Processes
PCP	Poisson Cluster Process
SNR	Signal-to-Noise Ratio
MS	Mobile Station
ID	Information Decoder
ER	Energy Receiver
SDR	Semi-Definite Relaxation
AWGN	Additive White Gaussian Noise
NAF	Naive Amplify-and-Forward
MILP	Mixed Integer Linear Program
eICIC	Enhanced Inter-cell Interference Coordination Technique
SIC	Successive Interference Cancellation

List of Symbols

$\mathbf{I}$	an Identity Matrix with Appropriate Dimension
$\mathbf{0}$	an All-Zero Matrix with Appropriate Dimension
$\mathbf{S}^{-1}$	the Inverse of the Square Full-Rank Matrix $\mathbf{S}$
$\mathbf{t}r(\mathbf{S})$	the Trace of the Square Matrix $\mathbf{S}$
$\mathbf{S} \succeq 0$	$\mathbf{S}$ is Positive Semi-Definite
$\mathbf{S} \succ 0$	$\mathbf{S}$ is Positive Definite
$\mathbf{S}^{H}$	the Hermitian of Matrix $\mathbf{S}$
$\mathbf{S}^{T}$	the Transpose of Matrix $\mathbf{S}$
$\mathbf{R}ank(\mathbf{S})$	the Rank of Matrix $\mathbf{S}$
$\mathbb{C}^{x \times y}$	the Space of $x \times y$ Complex Matrices
$\mathbb{C}^{x}$	the $x \times 1$ Complex Vector Space
$\mathbb{R}$	the Real Number Space
$\|\|\mathbf{x}\|\|_2$	the Euclidean Norm of a Complex Vector $\mathbf{x}$

Chapter 1

Introduction

1.1 Motivation

1.1.1 Simultaneous Wireless Information and Power Transfer

The target of the fifth generation (5G) wireless communication is to enhance wireless connectivity by increasing the data rates, coverage, bandwidth and reliability while reducing energy consumption and latency [1]. In response to two of the major requirements of 5G systems, developing green communication and improving energy efficiency have become the trends. One of the promising techniques to provide green communication while maintaining the required Quality of Service (QoS) is energy harvesting (EH), which works by converting ambient energy source such as sound, heat and radio frequency (RF) signals into electricity. These energy sources can provide more flexible, portable power supply compared to batteries. Initially, natural energy sources were considered for EH in wireless networks [2,3]. However, the energy efficiency was not satisfactory because the natural sources in ambient environments are always irregular and unpredictable. Moreover, the performance depends heavily on the environments [4]. Wireless power transfer (WPT) avoids these problems by enabling the nodes to utilize electromagnetic radiation to get charged [5]. In this case, both the ambient signals and the specified power sources, e.g. base stations (BSs), can be used to provide energy. Recently, WPT has been more frequently considered for near-field than far-field, with the existence of several challenges for implementing short distance WPT such as distance limitations, main-

tenance of field strengths, high cost, and resonant inductive tuning [6]. Moreover, the essential demand of enhancing communication distance motivates the research on far-field WPT techniques. Another significant trend for WPT is to merge WPT into wireless communication networks for better resource utilization. Therefore, simultaneous wireless information and power transfer (SWIPT), a technology that can transfer information and power simultaneously to the users, was first introduced in [5] from the perspective of information theory.

Recently, SWIPT has been recognized as one of the key technologies to achieve green communication and attracted tremendous attentions. As mentioned, SWIPT can achieve joint energy and information transmission in the era of 5G communication, and also be integrated with many modern communications networks and technologies, such as multi-carrier systems [7], cognitive radio (CR) [8], full-duplex communications [9], multiple-input multiple-output (MIMO) [10], device-to-device communication (D2D) [11], symbol level precoding [12], cooperative relaying [13], non-orthogonal multiple access (NOMA) [14], millimeter wave (mmWave) communications [15], heterogeneous networks (HetNets) [16], coordinated multipoint (CoMP) systems [17], smart grid [18] and sensor networks [19]. Despite the variations in different application scenarios, in general, SWIPT can provide notable gains in numerous perspectives, such as spectral efficiency, power consumption, interference management and latency by enabling simultaneous power and information transmission [4, 5, 20]. Nevertheless, challenges appear in order to balance between the performances of information transmission and power transfer as the power transfer process destroys information transmission. For example, more sufficient power from the energy harvesting can be ensured by increasing the transmit power. However, that increases interference as well as the susceptibility to eavesdropping which destroys the effective and secure information transmission, and hence gives rise to the research on secure SWIPT transmission aiming at enhancing the throughput and security of the SWIPT systems [21]. Long-distance information and power transfer is another significant issue for SWIPT systems considering the multi-path fading effects due to wireless propagation. To this end,

multiple antenna technology and cooperative communication technology have been merged into SWIPT systems where joint optimization of the transmit beamforming and power transfer ratios is required. However, the joint information and energy scheduling problems are always noncovex and difficult to handle. Since it is unlikely to have perfect channel state information (CSI) at the transmitters in reality, additional challenges arise when imperfect CSI is assumed, which leaves lots of space for further research in this field. Besides the conventional issues for implementing SWIPT in wireless networks which requires careful investigation, there are a series of directions possible for future research such as resource scheduling aspect, CSI feedback strategies, information theoretic framework, hardware impairments, channel coding techniques, internet of things (IoT), machine type communications (MTC), and satellite communication [22].

1.1.2 Edge Caching

The second part of the thesis is focused on the resource allocation technology aimed at reducing backhaul traffic, in particular, which is driven by the explosion of traffic stemming from healthcare, machine-to-machine communication, connected vehicles, social media, smart metering, IoTs and other new applications. By the end of 2016, mobile data traffic has reached 7.2 exabytes per month, and is expected to continue growing exponentially to reach 49 exabytes per month by the end of 2021 (seeing Fig 1.1) [23]. There will be changes in users' consuming habits as well due to the big boost of data rate. In particular, video will make up 82% of the total traffic by 2021. These high traffic requirements motivate the era of 5G networks to provide high spectral efficiency, dense deployment, new spectrum, green networking by facilitating a number of promising techniques, such as ultra-wideband communication, massive MIMO communication, mmWave communication, and HetNets [24, 25]. However, all these techniques rely on expensive backhaul links between the core network and BSs (or among BSs) [26–29], which justifies a need for reducing backhaul traffic. Moreover, the sheer volume and dimensionality of large-scale data sets in mobile traffic streaming brings a fundamental challenge of big data analytics and decision making, and thereby requires more decentralized

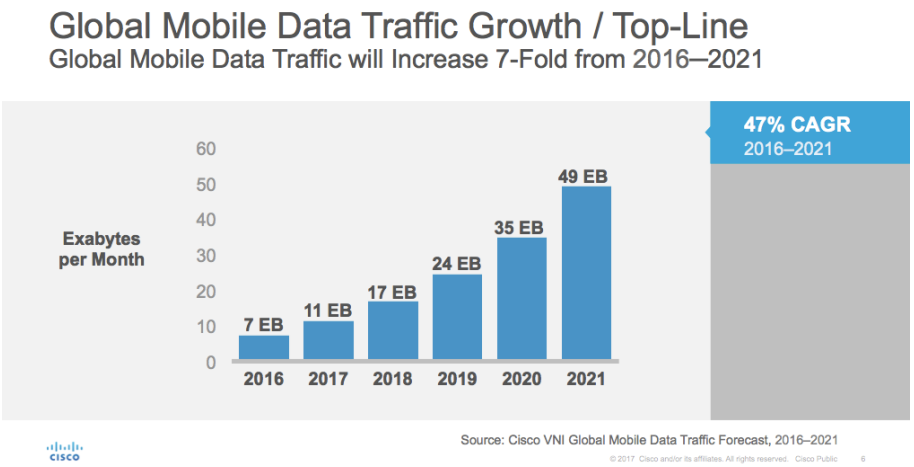


Figure 1.1: Global mobile data traffic growth by 2021 forecast by Cisco

and flexible network architectures with predictive resource management leveraging recent advances in context-awareness, storage and fog computing.

Besides the exposure of mobile data traffic, the backhaul traffic itself has some notable features which bring potential bottlenecks in terms of successful content delivery, energy efficiency, and latency.

Firstly, when a user makes a request, the serving base station needs to fetch the corresponding content from the core network via backhaul links, in a typical cellular network. As a result, there exists a huge amount of data traffic at the core network in peak time, which causes long delay and impose high requirement. In this case, if we equip the BSs with cache memories, the BSs can save popular content in their local storage. In so doing, they can serve the users directly using the cached content rather than fetching the content via backhaul repeatedly. Ideally, if the BSs can predict the user requests and update the cached data in advance [30, 31], the users can be served by the BSs directly without using the backhaul links, which guarantees timely data transmission in peak hours. Secondly, duplicate data transmission actually makes up a large portion of the total backhaul traffic due to the users' eagerness for current popular content (including the hottest audios, videos, and webpages). As a result, the same content needs to be repeatedly sent though the backhaul links to a number of users over and over again. If this type of duplicate

data transmission can be avoided, the backhaul traffic can be significantly reduced. And edge caching actually solves this issue effectively.

In addition, installing storage costs less than improving the backhaul capacity. For instance, a piece of memory storage of 2-3 TB only costs about 100 US dollars [32] while equipping backhaul links in network is usually quite expensive with a capital expenditure (CAPEX) including equipment and infrastructure expenses and operational expenditure (OPEX) breakdown [33]. Due to the advantages mentioned above, edge caching has played a crucial role in 5G networks as an effective approach to reducing backhaul traffic by storing popular content in caches equipped at the network edge. In doing so, the cached content can be delivered to users from local caches rapidly instead of being downloaded from the core network via backhaul, which helps reduce the peak-time traffic and latency in communicating with the core network when users make requests.

Currently, caching has been considered in BS or/and user terminals (UTs) to release the backhaul requirements while enhancing the overall performance of the networks from different perspectives [34–36, 50–52]. However, there are still many crucial issues that require to be better looked into despite of the progresses that have been made in the existing attempts. The first issue is about cache content placement, i.e. where and what to cache. Though early studies have shown that caching the most popular items at each of the small cells without coverage overlapping provides the highest local caching gains, for more complicated but efficient topologies, e.g. the small cell networks facilitated with multicast content delivery or/and content sharing, the content placement requires to be redesigned based on the coordination and cooperation among different cells instead of being optimized in terms of each cell separately. Moreover, the performance comparison between base station caching and D2D caching, as well as the combination of the two caching mechanisms, has not been well studied. Third, the impacts of the heterogeneity of the networks and content characteristics on the caching decision need further clarification. Third, how the transmission characteristics of the wireless networks (e.g. the broadcast nature of the wireless medium) affect the design of caching approaches,

i.e. the joint design of caching and transmission policies is another hot issue which opens up many potential research directions such as (i) multicast-aware caching for serving multiple concurrent requests, (ii) cooperative caching for content sharing among the caches and the associated interference management, and routing issues due to simultaneous transmission, (iii) mobility-aware caching considering the fact that the users may move from one cell to another during data transferring, and (iv) content popularity prediction and evolution, privacy preservation, and the combination of emerging networks and technologies.

In the following section, we present the outline and contributions of our work towards tackling the challenges mentioned above.

1.2 Outline of the Thesis

This thesis consists of two main parts dealing with SWIPT and edge caching, respectively. We first study the joint optimization of beamforming and power transfer ratios for SWIPT enabled systems in Chapter 3 and Chapter 4. Then we move to cache enabled small cell networks, where the cache content placement is carefully designed in Chapter 5 and Chapter 6. In particular, the results in Chapter 6 are derived for cache enabled small cells with heterogeneous settings and take the advantages of multicast content delivery and content sharing. Below, we briefly discuss the contribution of our work in each chapter.

Chapter 2: Background. This chapter provides a more profound introduction on SWIPT technology and edge caching technology. The main design aspects in the two fields and the essential issues are discussed followed by a more detailed investigation on the beamforming design for SWIPT enable systems and content placement and delivery strategies for cache enable networks, respectively.

Chapter 3: Robust Beamforming for SWIPT Broadcast Channels. In this chapter, we study MISO broadcast system for SWIPT using receiver power splitting and aim to optimize jointly the beamforming vectors and the power splitting ratios for minimizing the transmit power subject to the individual signal-to-interference-plus-noise ratio (SINR) and the energy-harvesting constraints assuming imper-

fect CSI. We propose a reverse convex non-smooth optimization algorithm, which provides the near-optimal rank-one solution, compared to semi-definite relaxation (SDR) guided iterations in literature.

Chapter 4: Beamforming for SWIPT MIMO Relaying. In this chapter, we consider SWIPT for a two-hop MIMO relay system where the relay is powered by harvesting energy from the source via time switching (TS) to finish information forwarding. We aim to maximize the rate of the system subject to the power constraints at both the source and relay nodes. In the first scenario where the source covariance matrix is an identity matrix, we present the joint-optimal solution for relaying matrix and the TS ratio in closed form. An iterative scheme is then proposed for jointly optimizing the source and relaying matrices and the TS ratio.

Chapter 5: Optimizing Cache Placement for Heterogeneous Small Cell Networks. In this chapter, we optimize the cache content placement for a typical cache-enabled small cell network with heterogeneous file and cache sizes. In particular, multicast content delivery is adopted to reduce the backhaul rate exploiting the independence among maximum distance separable (MDS) coded packets. We estimate the possible joint user requests using the file popularity information and aim at minimizing the long-term average backhaul load subject to the cache capacity constraints. The problem is reformulated into a mixed integer nonlinear program (MINLP) and solved with existing solvers after linearization.

Chapter 6: Coding, Multicast and Cooperation for Cache-Enabled Heterogeneous Small Cell Networks. This chapter considered the design of content caching and sharing for cache-enabled heterogeneous small cell networks using MDS codes under heterogeneous file and network settings. We first presented the multicast-aware caching and the cooperative caching schemes, for minimizing the long-term average backhaul load or the user attrition (UA) cost subject to the overall cache capacity constraint, and obtained the optimal content placement in both cases via convexification. A compound caching scheme, referred to as multicast-aware cooperative caching, was then proposed exploiting the independence of MDS coded packets to further reduce the backhaul requirements. In this case, a greedy

algorithm can be used for small scale networks while for large scale networks a multicast-aware in-cluster cooperative caching algorithm was developed. The advantages of storing coded packets over the uncoded fragments in all the scenarios as well as the benefits of utilizing multicast-aware caching and/or cooperative caching over common caching schemes have been analyzed.

Chapter 7: Conclusion and Future work. This chapter summarizes the main contributions of this thesis and introduces several potential research directions for future work.

1.3 Publications

The results, the ideas and figures are included in the following publications:

- J. Liao, M. R. A. Khandaker and K. K. Wong, "Robust Power-Splitting SWIPT Beamforming for Broadcast Channels," *IEEE Commun. Lett.*, vol. 20, no. 1, pp. 181–184, Jan. 2016.

- J. Liao, M. R. A. Khandaker and K. K. Wong, "Energy harvesting enabled MIMO relaying through power splitting," in *Proc. IEEE 17th Intl. Workshop on Signal Processing Advances in Wireless Commun. (SPAWC)*, Edinburgh, UK, July 2016, pp. 1–5.

- J. Liao, K. K. Wong, M. R. A. Khandaker, Z. Zheng, "Optimizing Cache Placement for Heterogeneous Small Cell Networks," *IEEE Commun. Lett.*, vol. 21, no. 1, pp. 120–123, Jan. 2017.

- J. Liao, K. K. Wong, Y. Zhang, Z. Zheng and K. Yang, "Coding, Multicast and Cooperation for Cache-Enabled Heterogeneous Small Cell Networks," *IEEE Trans. Wireless Commun.*, vol. 16, no. 10, pp. 6838–6853, Oct. 2017.

- J. Liao, K. K. Wong, Y. Zhang, Z. Zheng and K. Yang, "MDS Coded Cooperative Caching for Heterogeneous Small Cell Networks," in *Proc. Global Commun. Conf. (GLOBECOM)*, Singapore, Dec. 2017, pp. 1–7.

Chapter 2

Background

This chapter aims to provide the reader with the information and reference necessary to better understand the study in the following several chapters of this dissertation. General information about wireless background, and receiver architectures and beamforming for SWIPT will be presented firstly, and then the focus will shift onto the profound introduction about edge caching from various perspectives, such as the development, taxonomy, and caching strategy of cache-enabled wireless networks.

2.1 Wireless Fading Channels

Channel is the physical medium over which signal is transmitted from the sender to the receiver. One of the distinct features of wireless channels is fading, which refers to the random attenuation in the signal strength and the random phase shift of the received signal, as a result of the radio wave propagation in the environment. In general, two types of fading effects characterize wireless fading channel, namely large scale fading and small scale fading. The former refers to path loss characterized by distance and shadowing caused by prominent terrain contours such as hills, forests, and tall buildings [37]. Small-scale fading occurs as a result of multipath propagation, which mainly results from a combination of effects such as diffraction, reflection and refraction. Frequency selectivity is another important characteristic of wireless fading channels [38]. When all frequency components of the transmitted signal experience the same fading magnitude, the fading is said to be frequency

flat. This occurs when the coherence bandwidth of the channel B_c is larger than the signal bandwidth B_s, i.e. $B_s < B_c$. On the contrary, if the frequency components are affected by different amplitude gains and phase shifts, it is called frequency selective, which occurs when $B_s > B_c$.

Recently, there have been a range of statistical models which characterize wireless fading channels with applicable trade-off between accuracy and complexity. In the following, we introduce several typical channel models which will be considered in this thesis.

2.1.1 Wireless Channel models

A. Rayleigh Fading

Rayleigh fading model is often used for multipath fading channels with no direct line-of-sight (LOS) path. The probability density function (PDF) of the channel fading amplitude is given by [39]

$$p(\alpha) = \frac{2\alpha}{\Omega} \exp\left(-\frac{\alpha^2}{\Omega}\right), \alpha \geq 0, \tag{2.1}$$

where α is the channel fading amplitude, and $\Omega = \mathbb{E}\{\alpha^2\}$ denotes the mean value.

B. Rician Fading

Rician fading is frequently used to characterize propagation paths with one strong direct LOS component and many random weaker components. Here the channel fading amplitude α follows the distribution

$$p(\alpha) = \frac{2(1+n^2)e^{-n^2}\alpha}{\Omega} \exp\left(-\frac{(1+n^2)\alpha^2}{\Omega}\right) I_0\left(2n\alpha\sqrt{\frac{1+n^2}{\Omega}}\right), \alpha \geq 0, \tag{2.2}$$

where $I_0(*)$ is the Bessel function of the first kind [40], and n denotes the fading parameter ranging from 0 to 1. The Rician factor refereed to as K, which reflects the connection between the power of the LOS component and the power of the Rayleigh component, can then be derived using $K = n^2$. Particularly when $K = 0$, there is no LOS component and the Rician PDF is reduced to the Rayleigh PDF.

2.1.2 Channel State Information

The channel state information (CSI) of a communication link is defined as the known channel properties describing the propagation of a signal between the transmitter and the receiver, and reflecting the combined effect of wireless transmission, such as scattering, fading, and power decay with distance [41]. As it is crucial to adapt transmissions to current channel conditions to reconstruct the useful information with minimum distortion at the receiver, CSI needs to be estimated at the receiver and fed back to the transmitter. And the process of CSI acquisition is referred to as channel estimation.

There are two major types of channel estimation methods, the data-aided method and the blind estimation method. The data-aided method is also referred to as training sequence (or pilot sequence), where a known sequence is transmitted and the channel coefficient will be estimated by removing the known data at the received signal. If there is no noise at the received signal, the estimation will be perfect. There are two typical ways to insert the pilot sequence, block-type pilot arrangement and comber-type pilot arrangement [42]. According to the evaluation criteria, the channel estimation methods can be divided into least-square (LS) estimation and the minimum mean square error (MMSE) estimation. The former aims to minimize the sum of squared errors while the later focuses on the mean square error (MSE) between estimated and actual received signal [43]. Generally speaking, MMSE estimation outperforms LS estimation by shortening the required pilot sequence and reducing the estimation error. However, MMSE estimation requires the channel correlation and noise correlation information in advance, and therefore can be seen as the Bayesian counterpart to LS estimation. In contrast, blind estimation only depends on the received data without any training symbols [44]. Data-aided estimation usually achieves better accuracy than blind estimation at the expense of a higher overhead, e.g., more bandwidth.

2.1.3 Imperfect CSI Models

In practice, CSI needs to be estimated in the presence of noise, and therefore the acquired CSI is destined to be imperfect. For this reason, channel errors are some-

times considered in the design of more robust communication systems. Channel errors are often modeled using stochastic or deterministic (worst-case) models. A popular stochastic model utilizes Gaussian random variables to model the CSI uncertainty, while the deterministic model usually assumes that the error is modeled by an ellipsoid-bounded uncertainty region [45]. Compared with the perfect CSI case, the imperfect CSI case is always more challenging for optimizing the resource allocation for wireless communication. Here, we discuss about the typical types of channel uncertainties.

A. Ellipsoidal Channel Vector Uncertainty

If the CSI is estimated in the form of channel vector (as is typical in a multi-dimensional channel such as in a multiple antenna system), then the channel at the receiver, denoted as the vector $\mathbf{h}$, can be modeled as [46]

$$\mathbf{h} \triangleq \hat{\mathbf{h}} + \Delta\mathbf{h}, \tag{2.3}$$

where $\Delta\mathbf{h}$ is the CSI estimate while $\Delta\mathbf{h}$ is the CSI error. Depending on the estimation methods or feedback schemes, the channel errors follow specific random distributions. Without loss of generality, we consider the case that $\Delta\mathbf{h}$ is subject to colored noise and bounded by an ellipsoid, i.e.,

$$\mathscr{H} = \{\Delta\mathbf{h} : \Delta\mathbf{h}^{\dagger}\mathbf{C}\Delta\mathbf{h} \leq 1\}, \tag{2.4}$$

where $\mathbf{C} \succeq 0$ determines the quality of CSI and is assumed known at the receiver.

B. Ellipsoidal Channel Covariance Uncertainty

Compared with the channel vector, the second-order statistics of the channel changes more slowly. Thus, the estimated CSIs in the form of channel covariances are more practical. In this case, we model channel covariance uncertainties as [47]

$$\mathbf{H} \triangleq \mathbf{h}\mathbf{h}^{\dagger} = \widehat{\mathbf{H}} + \Delta\mathbf{H}, \tag{2.5}$$

where $\widehat{\mathbf{H}}$ denotes the CSI estimate at the receiver while $\Delta\mathbf{H}$ corresponds to the CSI errors. Similarly, we consider the more practical scenario where the channel covariance matrices are estimated in the presence of colored noises, and $\Delta\mathbf{H}$ is bounded by

ellipsoidal regions. In addition, $\Delta\mathbf{H}$ should be set to guarantee the positive semidefiniteness properties of the matrix $\widehat{\mathbf{H}}+\Delta\mathbf{H}$, i.e.,

$$\widetilde{\mathscr{H}} = \{\Delta\mathbf{H} : \Delta\mathbf{H} = \Delta\mathbf{H}^{\dagger}, \widehat{\mathbf{H}}+\Delta\mathbf{H} \succeq 0, \mathrm{T}r(\Delta\mathbf{H}^{\dagger}\mathbf{C}\Delta\mathbf{H}) \leq 1\}, \tag{2.6}$$

where the parameters $\mathbf{C} \succeq 0$ are known a priori.

C. Stochastic CSI Errors

The CSI error may also be modeled as Gaussian random variables following a known distribution, i.e., [48]

$$\mathbf{h} \triangleq \hat{\mathbf{h}} + \Delta\mathbf{h}, \tag{2.7}$$

where CSI error $\Delta\mathbf{h}$ is modeled as zero-mean Gaussian random variables with covariances C, i.e., $\Delta\mathbf{h} \sim \mathscr{N}(0,C)$. As $\Delta\mathbf{h}$ is unbounded, meeting deterministic constraints in all time would be impossible and therefore probabilistic conditions occur that guarantee the deterministic conditions being satisfied with high probability.

2.2 Performance Measures

There are many performance metrics that are commonly used in evaluating the performance of resource allocation schemes in wireless communication systems. The most relevant ones that are used throughout this thesis are signal-to-noise ratio (SNR), signal-to-interference-and-noise ratio (SINR), channel capacity, hit ratio and backhaul load. In this section, a brief introduction of these performance metrics is now presented.

2.2.1 SNR and SINR

The output SNR is defined as the ratio between the power of signal component in the output and the power of the noise component in the output. Let P be the power of the transmit symbol defined as $P = \mathbb{E}(ss^*)$, σ^2 be the noise level, and h be the fading channel, the SNR r can be expressed as

$$r = \frac{P|h|^2}{\sigma^2}. \tag{2.8}$$

Due to the fading effects of the wireless channels, the average SNR is considered, which is defined by $\tilde{r} = \mathbb{E}_{|h|}(r)$.

The SNR has high impact on the quality of data detection, and is therefore frequently used as the performance metric in both network optimization and performance analysis.

The definition of SINR is similar to the SNR except that the impact of co-channel interferences is considered in addition to the impact of the noise. In wireless communication systems, interference frequently appears as the result of frequency reuse among neighboring cells, which makes the SINR more practical than the SNR. For instance, assuming that there is only one strong interference channel of which the fading coefficient is defined as g, the received signal can then be expressed as

$$y = hs + gx + n, \tag{2.9}$$

where s, x, n are the transmitted signal, the interference signal and the noise signal, respectively. The power of the interference signal is defined as $P_i = \mathbb{E}(xx^*)$. Considering co-channel interference, the SINR is defined as the ratio between the power of the desired signal and the sum power of the interference and noise, i.e.,

$$\widetilde{r} = \frac{P|h|^2}{P_i|g|^2 + \sigma^2}. \tag{2.10}$$

2.2.2 Channel Capacity

Firstly proposed by Claude Shannon in [49], channel capacity refers to the maximum rate at which information can be reliably transmitted over a communication channel. Channel capacity is defined as the maximum of the mutual information between the transmitter and the receiver. The instantaneous channel capacity is given by

$$C = \log_2\left(1 + \frac{P|h|^2}{\sigma^2}\right). \tag{2.11}$$

Depending on the situations, channel capacity can be divided into two types, i.e., ergodic capacity and outage capacity, according to the property of the fading channel h. The basic assumption for ergodic capacity is that the transmission time is long enough to present the long-term ergodic properties of the fading process. In this case, the ergodic capacity can be written as $C_e = \mathbb{E}_{|h|}\{C\}$. However, the assumption on ergodicity is not always achievable in practical delay-constrained

communication systems. On the contrary, when there is no significant channel variability in the transmission process, the actual transmitted rate may exceed the instantaneous channel capacity at a non-negligible probability. In this case, it is better to consider $q\%$ outage capacity C_o, which is defined as the channel capacity which can be achieved by $(100-q)\%$ of the channel realizations and guarantees reliable services, i.e., $\Pr\{C \leq C_o\} \leq q\%$.

2.2.3 Hitting Rate

Hitting rate is one of the most frequently used metrics in measuring the performance of cache content placement scheme. In deterministic caching, it is defined as the ratio between the number of cached files that are requested and the overall number of files stored in the storage. For instance, a hitting rate higher than 90% demonstrates that most of the requests are satisfied by the cache. On the other hand, in probabilistic caching, the hitting ratio is the probability for the required file being stored in the cache. An efficient caching scheme maximizes the cache hits while minimizing the cache misses, yielding higher hitting rate, lower latency, and better storage utilization. Although hitting rate directly shows the cache miss, it fails to reflect the impact of the caching scheme on the performance of wireless communication system [50].

2.2.4 Network Delay

Network delay is defined as the response time from the time when the file is requested until delivery [51]. From the perspective of the users, it is necessary to minimize the delay of being served which is critical to user's experience especially for delay-sensitive content services. As fetching content from local storage is much faster than delivering content from the core network, the network delay can be reduced by taking full advantage of storage space.

2.2.5 Backhaul Load

In cache enabled wireless networks, backhaul load refers to the amount of content that requires to be delivered form the core network to the BSs via backhaul. As the user demands are usually assumed to be unknown and random following particular

distributions, e.g., Zipf's distribution, the backhaul load is usually measured in long term average in terms of all possible user demands [52]. In general, higher backhaul load results in higher peak-time traffic and network delay, and places higher demand on backhaul capacity.

2.3 More on SWIPT Systems

2.3.1 Receiver Architectures for SWIPT Systems

In wireless communication systems, RF signals can convey both information and energy simultaneously, and hence RF-based SWIPT has become a promising energy harvesting technology where the terminals can not only access wireless data but also harvest energy from RF signals simultaneously. Though early fundamental studies on SWIPT have assumed lossless information and power transfer with the same signal [53], it is difficult to achieve in reality as the power transfer affects the performance of information transmission directly. For instance, enhancing transmit power can increase harvested energy but also impose high interference. Moreover, the RF signal acts as a dual-purpose carrier for conveying information and energy to the receivers simultaneously. However, the huge gap between the power sensitivity for energy harvesting receivers (-10 dBm) and information decoding receivers (-60 dBm), which is referred to as the near-far issue, becomes a barrier to implement SWIPT technology and requires reconsideration of receiver architecture design [10].

In general, there are two types of receivers in SWIPT systems, the separated and collocated receivers. Fig. 2.1 presents the different receivers architectures.

A. Separated Receivers

For separated receivers, the receivers are either responsible for energy harvesting (EH) or information decoding (ID). For instance, a location based receiver scheduling was proposed in [54], where receivers located near or far away from the transmitter were assigned for energy harvesting or information decoding, respectively. This technique can be easily implemented.

B. Collocated Receiver

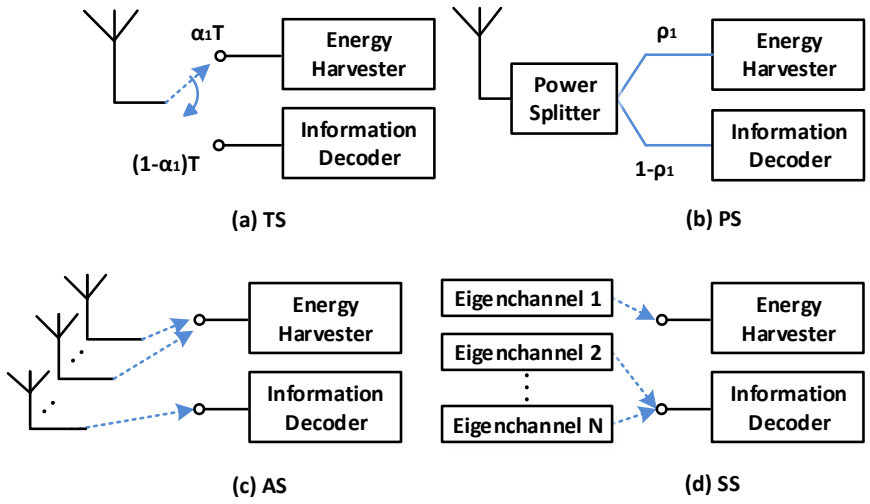


Figure 2.1: Typical SWIPT receiver structures. α_1 denotes the TS factor, ρ_1 denotes the PS factor, and T denotes the transmission block duration.

The collocated receiver can coordinates between the energy harvesting and the information decoding processes. To achieve SWIPT for systems with collocated receivers, the received signal requires to be split for information decoding and energy harvesting, respectively. The signal splitting strategies involve different domains, i.e. time, power, antenna, space domains [4].

- **Time Switching (TS):** In consideration of TS, each time slot is partitioned into two orthogonal slots. The receiver works alternatively as information decoder and energy harvester by a certain ratio referred to as a TS ratio [55].

- **Power Splitting (PS):** As an opposite, the PS technique enables SWIPT by splitting the received signal into two parts of different power levels according to a PS ratio, one of which is converted to base-band for information decoding while the other is sent to the rectenna circuit for energy harvesting [56]. Compared with TS, the PS technique requires higher complexity in receiver design, but achieves information and power transfer in the same time slot which makes it more suitable for delay-sensitive networks. Compared with other techniques, it achieves the best trade-off between information rate and

harvested energy.

- **Antenna Switching (AS):** In this architecture, the receiver always has multiple antennas a subset of which is assigned to decode the information, while the remaining are responsible for energy harvesting [57]. Antenna switch is an easy-to-implement and low-complexity protocol.

- **Spatial Switching (SS):** In this architecture, spatial domain is mainly on the spatial degrees of freedom of the channel rather than the antenna elements. Utilizing the singular value decomposition (SVD) of the MIMO channel, the SS technique transforms the communication link into parallel eigenchannels conveying either information or energy [58].

2.3.2 The Application of Beamforming for SWIPT Systems

Beamforming has been widely applied in wireless communications to focus the signal in a narrow direction by generating pencil beams. In wireless networks, the usage of beamforming can bring lots of advantages. By focusing the signal towards the intended receiver, and therefore reducing the multi-path attenuation and interference, beamforming enhances SINR, data rate, security level and spectral efficiency, and therefore helps accommodate more users and provide larger coverage area [59]. However, there are still a couple of challenges in beamforming design for wireless networks. For instance, the usage of antenna arrays and signal processing modules places higher requirements on money and power in deployment. Moreover, the SCI acquisition and signal processing needed in learning the channel status and deriving optimal beamforming solution, result in higher complexity and overhead.

As mentioned previously, the power sensitivity of the EH process is quite low compared with the ID process [10]. And the EH process is quite sensible to signal decay caused by the properties of wireless propagation, such as scattering, reflection, and fading. To deal with this problem, multi-antenna techniques, e.g., beamforming, have been widely applied in SWIPT enabled communication networks to guarantee efficient information and power transfer [20]. The advantages of beamforming in wireless communication mentioned above make beamforming a

significant transmission scheme in SWIPT systems.

And the main target for beamforming design in SWIPT systems is to derive the optimal beamforming solution that achieves desired energy-information transmission tradeoff. Recently, many studies have focused on the beamforming design in SWIPT systems. For instance, an optimal beamforming solution was derived for MISO SWIPT with two and three PS-based receivers in [60]. Physical layer security is another important issue worth careful investigation in beamforming design for SWIPT systems [59]. Another challenge will be channel conditions acquirement, because the selection and updating of transmit beams and transmit scheme depend heavily on channel conditions as mentioned in the previous section. Compared with the perfect CSI case, the imperfect CSI case is always more challenging for beamforming design in SWIPT systems. However, it is quite difficult to guarantee perfect CSI for SWIPT systems in reality, as the CSI acquirement of the ID receiver will be heavily interfered by the interference of both information and power signals while the EH receiver, even worse, usually has no special circuit to feedback the CSI to the transmitter. Robust beamforming is consequently an indispensable process for resource allocation for SWIPT systems. What is more, the energy constraints, which always conflict with the information transmission efficiency, bring extra difficulties to the beamforming design, compared with the conventional networks.

2.3.3 Taxonomy of Beamforming for SWIPT Systems

Recently, there have been many papers focusing on beamfoming design for SWIPT systems which differ from each other in a variety of design aspects, such as the network topology, performance metric, and mathematical tool. These works focused on joint data transmission and power transfer factor design for SWIPT systems in terms of enhancing the energy or spectral efficiency assuming either perfect CSI or imperfect CSI with bounded or stochastic uncertainties at the transmitters. Due to the broadcasting nature of wireless channels, increasing transmit power can not only boost the harvested energy but also increase the susceptibility to eavesdroppers, which shows synergy between communication, energy and security, a significant issue for SWIPT systems [21, 61, 62].

A. Network Topology

The emergence and development of SWIPT has opened up numerous new opportunities including the SWIPT works for orthogonal frequency-division multiplexing (OFDM) systems [64], frequency-selective channels [65] and multiuser scenarios, such as the relay channel [13], the interference channel [66], the multicasting channel [67], and the broadcast channels [56] and [68]. Instead of considering the essential issues in implementing SWIPT for conventional networks, the research of SWIPT has recently been extended to involve emerging network topologies and techniques to achieve the targets of the 5G cellular communication systems, such as full-duplex communications, MIMO, D2D, symbol level precoding, NOMA, mmWave communications, HetNets, CoMP systems, smart grid, and sensor networks [9–19].

B. Performance Metric and Constraints

The beamforming design problem in SWIPT systems has been formulated from different perspectives, such as minimizing total transmit power or outage probability [56], maximizing system capacity or throughput [63], and deriving rate-energy trade-off [10]. The possible constraints in these cases may differ according to the design objectives. There are several general types of constraints, such as power consumption constraint, energy harvesting constraint and the SINR constraint [10, 56, 63]. To control the power consumption, the total transmit power should not surpass a given threshold. On the other hand, to ensure enough power from the energy harvesting, we need to make sure that the power received at the EH receiver is higher than a threshold. To meet quality of service requirements, the SINR at the ID receiver should be large enough to guarantee accurate information decoding at the ID receiver. However, those constraints are always conflicting, and also not jointly convex, which turns the beamforming design problems non-convex optimization problems. As an example, [10] characterized the rate-energy regions for MIMO broadcast systems for SWIPT with separated and co-located information and energy receivers.

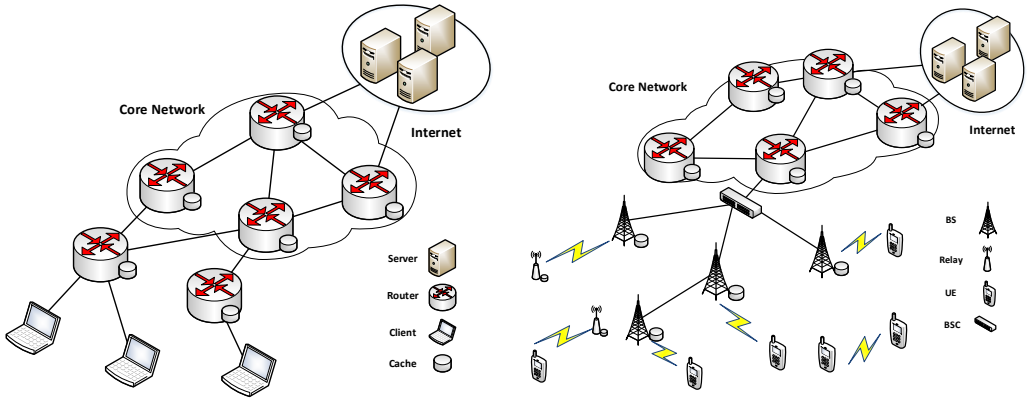
D. Optimisation Strategy

The most frequently used optimization strategies for beamforming design in SWIPT systems are Semi-Definite Programming (SDP) [69], Second Order Cone Programming (SOCP) [70], sub-optimal algorithms like block coordinate descent (BCD) method [71], or convex relaxation methods like, Successive Convex Approximation (SCA) [72]. Meanwhile, reformulation is a good way to convexify the original unconvex problems by strategies such as introducing new variables and matrix transformation. For robust beamforming design with ellipsoidal channel vector uncertainty, S-Procedure is usually used for reformulation. Taking the SWIPT MIMO system with perfect CSI as an example [10], the transmit power minimization problem subject to the SINR and harvested energy constraints is a minimum convex SDP because the objective function is linear and the constraints are defined by a finite number of convex sets [73]. However, the number of constraints becomes infinite when it comes to the imperfect CSI scenario. To handle the infinitely many inequalities, S-Procedure is an efficient way to turn the constraints with channel uncertainty into linear matrix inequalities (LMIs) [45]. In particular, for MISO systems with beamforming vector defined as $\mathbf{b}$, there is always uncovexity in beamforming design problems due to the presence of $(\mathbf{b}\mathbf{b}^H)$ in essential performance metrics such as transmitted power or SINR. By replacing $(\mathbf{b}\mathbf{b}^H)$ with a new matrix $\mathbf{W}$ which satisfies $\mathbf{W} \triangleq \mathbf{b}\mathbf{b}^H$, the original problem can again be reformulated into an SDP. However, this type of reformulation is not completely equivalent as we are not guaranteed to be able to derive the optimal vector $\mathbf{b}$ by decomposing the obtained $\mathbf{W}$, unless the rank of $\mathbf{W}$ equals to 1. To deal with this rank-one issue, we need to prove that R$ank(\mathbf{W}) = 1$ holds true, or alternatively resort to suboptimal solution utilizing randomization [74].

2.4 More on Wireless Edge Caching

2.4.1 The Development of Edge Caching

As one of the emerging techniques to deal with the explosive wireless traffic growth, caching techniques have been widely applied in wired networks for web caching firstly emerged in the early 1990s [75], content distribution networking (CDN) since



(a) Cache enabled wired network.

(b) Cache enabled cellular network.

Figure 2.2: Architectures of typical wired and wireless networks with caching.

the late 1990s and 2000s [76, 77] and more recently, information centric networking (ICN) [78], with the popular content changing from web pages and images, to videos generated by servers or clients, and the cache memories installed at clients, proxy servers, and routers, respectively. In particular, ICN achieves efficient content placement based on content popularity and network parameters [78]. Fig. 2.2a presents a typical cache-enabled wired network consisting of clients, routers, and servers. When the clients send requests to the routers, if the routers store the requested content, the clients can be served by the routers directly without contacting the core network, which reduces the latency in the network [79].

The concept of caching has recently been introduced to the physical layer to reduce peak-time traffic, latency as well as the requirement for expensive high capacity backhaul links [34]. Similarly, the main idea of caching in cellular networks is to store popular content at the network edge, either at the BSs or/and UTs to bring the content closer to the users. Fig. 2.2b provides a typical cache-enabled cellular network consisting of a wired core network, BSs and backhaul links connected to the internet. If the requested content is stored in the user itself, neighboring users, or serving BSs, the user can fetch the content therein. Otherwise, the requested content should be fetched from the core network via backhaul. Though implementing caching techniques in wireless networks share a lot of resemblance with that

in wired networks, the wired caching strategies in upper layer cannot be directly applied to cellular networks due to the unique structures and transmission features of cellular networks [80]. Firstly, wireless users usually have smaller cache capacities than the wired clients. Moreover, challenges appear due to node mobility and dynamic network topology in wireless networks. For instance, wired clients are usually fixed while users in wireless networks may move from one cell to another. And wireless channels are more uncertain and complicated compared with wired channels. The unique barriers of wireless transmission, such as co-channel interference and limited spectrum, make it even more challenging to design efficient caching strategies for wireless networks.

For cache-enabled cellular networks, one needs to address, e.g., where to cache, what to cache, the corresponding delivery and transmission policy, and so on [35]. Regarding the first question, caching can take place at the BSs or UTs. By caching at the BSs, we can reduce the traffic in backhaul and improve the energy and spectral efficiencies, while caching at the UTs adds cooperation gain and improves network scalability, facilitating D2D links.

Also, cache content placement addresses what to cache. As far as content updating is concerned, caching schemes can be divided into adaptive caching and proactive caching. Adaptive caching, a.k.a. pull-based caching, works in a reactive manner by storing content in the caches on demand. In this scheme, caching decision is performed only after users have made their requests so that online algorithms, such as the least frequently used (LFU) and least recently used (LRU), can be used. As a result, the cached content in each cell is updated every time a new round of requests is made by the users. By contrast, proactive caching is a push-based approach which proactively estimates user demand patterns and performs content placement before the users make requests. Popular schemes include common uniform placement, popularity based placement, probabilistic placement [50], partition-based placement [36] and other offline schemes. When caching contents, we can either store the entire files or fragments of the files based on file splitting to ensure diversity of the cached contents in the case that the cache capacity is rel-

atively limited compared to the average file size. For this reason, network codes, such as MDS codes have been utilized to construct file pieces in order to improve storage utilization. In optimizing content placement, the objective is usually on one of the followings: the hit ratio [50], latency [51], backhaul load [52], service cost, spectral/energy efficiency, and so on.

Recently, considerable research has been done on physical layer caching. Inspired by the primary research [81] where caching schemes were designed in presence of routing, cooperative, and physical layers in 2009, an comprehensive information-theoretic study was first given in [52] in 2010 for a homogeneous system with a single content server and several users served with a shared link. Subsequently in, e.g., [82–89], more complex network topologies with heterogeneous network settings have been studied for, respectively, nonuniform file popularity, file sizes and cache sizes, random requests, secure delivery, interference channel, D2D networks, and recently fog random access networks (F-RANs). In 2012, a typical femto-cell BSs (helpers) assisted wireless distributed caching network was proposed in [90] aimed at minimizing the network delay by jointly designing the content placement and the cooperation among the helpers. More recently, wireless caching techniques have been extended into diverse network topologies, e.g. cache-enabled macro-cellular networks [91, 92], HetNets [93, 94], D2D cellular networks [95,96], cloud-radio access networks (C-RANs) and fog-radio access networks (F-RANs) [97, 98]. Most of these papers focused on the typical issues of designing the content placement and content delivery algorithms.

Another hot topic for cache-enabled networks is the beamforming design which focuses more on the transmission aspects in presence of caching. In [97], instantaneous beamforming and BS activation for C-RAN were addressed, while [99] considered the joint design of data assignment and beamforming for a cooperative multi-cell network both assuming a given cache content placement in a short-term time scale. In [100], beamforming and cache content placement were jointly optimized utilizing a mixed time-scale stochastic optimization scheme. In addition, performance analysis of cache-enabled wireless networks has also been extensively

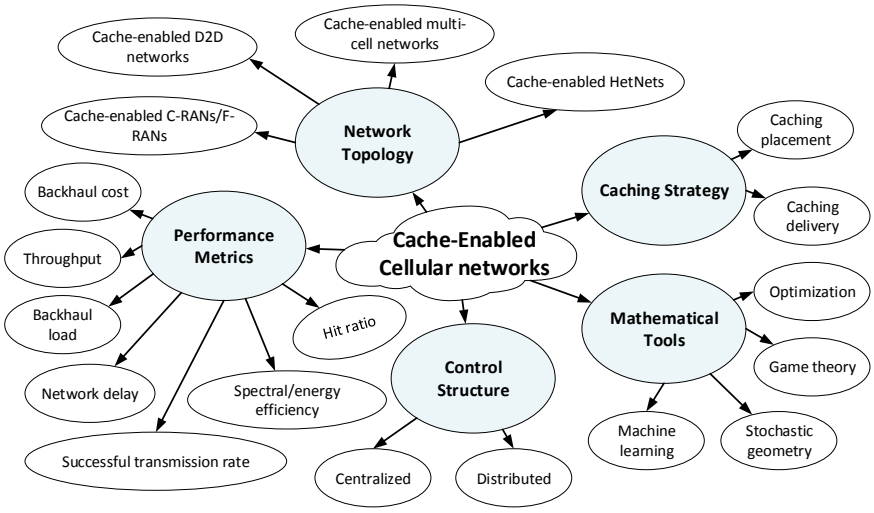


Figure 2.3: Taxonomy of cache-enabled wireless networks.

conducted in the literature, e.g., [101–104]. To summarize, those results largely analyzed cache-enabled small-cell networks using stochastic geometry to model the stochastic properties of channel fading and interference. However, the results either ignored the spatial diversity of the cached content and disabled the coordination and cooperation aspect among different cells [101, 102] or even ignored the file popularity information altogether [103, 104].

Apart from the above, emerging topics in physical layer caching also include multicast-aware caching [105], hierarchical caching [106], mobility-aware caching [107, 108], cooperative caching, and caching architecture design in fog-RAN [109]. Moreover, learning, matching and online algorithms have also be used to solve physical layer caching problems in [110–112].

2.4.2 Taxonomy of Cache-Enabled Wireless Networks

According to the network topology, caching model, performance metric, control structure and mathematics tool, the existing research on physical layer caching can be classified into different categories. Next, we briefly discuss the caching techniques in cellular networks from different design aspects (seeing Fig. 2.3) [113].

A. Network Topology

There are four typical types of cache-enabled cellular networks: cache-enabled

macro-cellular networks, cache-enabled HetNets, cache-enabled D2D networks, and cache-enabled CRANs/F-RANs. One of the most notable differences among them is the cache location which ranges from the BSs, the SBSs, the UEs, the baseband units (BBUs), and remote radio heads (RRHs), respectively. Moreover, reducing the requirements on backhaul is the major target in the first two types of networks, while edge caching is expected to facilitate collaboration and improve the spectral efficiency and energy efficiency additionally in cache-enabled D2D networks and C-RANs/F-RANs. In the following, we briefly discuss the different networks, respectively.

- **Cache-Enabled Multi-Cell Networks:** In this type of networks [91], each BS is equipped with memory storage to store popular content. And it is always assumed that no coverage overlap exists among the BSs.

- **Cache-Enabled HetNets:** Cache-enabled HetNets integrates a variety of technologies, and cell layers, i.e. macro-cells, small cells (femtocells, picocells), and relay nodes, in order to meet the higher requirements on coverage, capacity, and latency. In this case, coverage overlap and spectrum sharing occurs among the macro-cells, and other nodes, which causes concerns on interference and coordination aspects [93]. The associated network topology can be grid-based or random spatial. For the former, the macro base station (MBS) is central-located with a number of small cell base Stations (SBSs) being deployed inside the cell. In the random spatial topology, the MBSs and SBSs are randomly deployed in each cell based on stochastic geometry [114]. Since both the MBSs and SBSs can have memory storage, the multiple-level hierarchical caching and cooperative caching can be considered to further improve the caching gains.

- **Cache-Enabled C-RANs/F-RANs:** C-RAN is a centralized, cloud computing-based architecture while F-RAN takes the advantages of both edge caching and centralized processing and creates a new distributed, edge computing-based radio access network [116]. In both of the two structures,

there are a number of BBUs with caches clustered as the BBU pool, the RRHs connected to the BBU pool through fronthaul links, and the UEs. In C-RAN, the BBU pool carries memory and offers collaborative processing, while the edge nodes, RRHs, take charges of edge caching and signal processing in F-RAN, which reduces the latency and transmission power by moving major transmission tasks from the BBU pool from RRHs [117].

- **Cache-Enabled D2D Networks:** In cache-enabled D2D networks, the devices are enabled to have memory storage and communicate directly with nearby devices without contacting the BSs [95, 115]. In this case, adjacent devices usually form a cluster, if the requested file has not been stored in local cache, it can fetch the file from any user in the same cluster that caches the file. Compared with the base station caching, D2D caching has some unique challenges due to the properties of the network topology, such as smaller cache capacity, more limited transmitted power and coverage, denser communication links, and higher user mobility.

B. Caching Strategy

As mentioned, the caching process always has two phases: the content placement phase and the content delivery phase. The former decides which files should be stored in the caches, and occurs during off-peak hours, e.g. night time. The later usually happens right after the users make requests in peak traffic periods. In this phase, the requested files need to be sent to the users from the caches or the core network. Note that the content placement phase, except for the online caching case, happens in a relative long-term time scale compared with the content delivery phase which occurs instantaneously, and therefore lots of papers have assumed that the content placement has been fixed when designing the content delivery strategies, while authors in [100] proposed a new strategy for joint optimization of the two phases utilizing mixed time-scale stochastic optimization.

- **Caching Placement Strategy:** As mentioned previously, there are a variety of content placement schemes which can be grouped into several types according to different aspects, e.g. file partition, placement updating, caching

model, performance metric and etc. For instance, we can distinguish the content placement schemes among uncoded caching or coded caching, proactive caching or reactive caching, deterministic caching or probabilistic caching, and so on. For the uncoded caching, complete files or uncoded segments of the files will be stored in the caches. For the coded placement, each file is partitioned, processed by particular codes [118,119], and then cached, in order to improve the utilization of storage. Most of the literature considers proactive content placement strategy, e.g. popularity based placement and probabilistic placement, which is an offline caching strategy deciding the content placement by actively predicting the file popularity. Oppositely, adaptive content placement strategy, e.g. the least frequently used (LFU), least recently used (LRU), and other online algorithms [121], determines how to update the content in the caches according to the users' demands. The basic idea of the two typical replacement strategies is removing the least recently requested or least frequently used content from the caches, and then replacing them with more popular content, respectively [122].

- **Caching Delivery Strategy:** The target of caching delivery strategy design is to find a most effective way to transmit the requested content to the users. To do so, we need to decide where and how to fetch the requested content, which requires carefully optimization of user association, transmission method (e.g. unicast, or multicast), power allocation, channel allocation, and other required transmission parameters, according to different performance metrics and physical transmission conditions. In particular, multicast transmission plays an important role in content delivery by utilizing a single multicast transmission to serve temporal-spatial requests for the same content, instead of sending multiple unicast transmissions each corresponding to an individual request [123]. Coded multicast is another effective way to satisfy different requests concurrently with multicast utilizing linear coding [52].

C. Control Structure

- **Centralized:** There always exists an entity working as a central controller,

which collects useful information for the entire network, e.g., the content popularity, user demands, and channel information, and then uses it to decide optimum caching strategies [94,124]. The centralized structure guarantees the global optimality at the expense of solving large-scale optimization problem.

- **Distributed:** The nodes decide the caching strategies only based their own information so that local optimality is achieved in a distributed structure. Since the information about other nodes does not need to be considered, the size of the caching problem is much smaller compared with centralized algorithms.

D. Mathematical Tools

- **Optimization:** In general, the caching optimization problems usually aim to maximize/minimize a performance metric, e.g. hit ratio, successful transmission rate, backhaul cost, or network delay, under the storage capacity constraints and QoS requirements by designing the placement and delivery strategies.

- **Game Theory:** Game theory is a mathematical tool modeling the conflict and cooperation among competing players, and has been widely used for resource allocation. In cache-enabled wireless networks, the entities, e.g. the SBSs, and UEs, need both competition and cooperation due to limited storage capacity, transmit power, or backhaul resource, and therefore they can be considered as the players competing to maximize their own utilities [125, 126].

- **Stochastic Geometry:** Stochastic geometry is a typical tool to model random network topology, where independent homogeneous Poisson point processes (PPPs) or Poisson cluster process (PCP) are used to model the heterogeneity of the MBS and SBS locations [114]. Moreover, this model takes the interference aspect among different links into account, which makes it more practical and suitable for performance analysis and optimization for cache-enabled wireless networks with random topology or high mobility.

- **Machine Learning:** Though it has been frequently assumed that the BS or

UT has knowledge of the file popularity when designing caching strategies in literature, the nodes usually have no prior information about that in practice. Machine learning provides a solution to this problem by collecting the users' request history and using that to predict the file popularity [31, 127].

2.4.3 Content Placement and Delivery Strategies

In this section, we briefly introduce several typical models in designing the content placement and delivery strategies for cache enabled cellular networks.

A. Femto-Caching

In [94], a wireless system was considered where UTs communicated with a set of distributed cache-enabled helpers. The authors minimized the expected downloading time by optimizing the content placement at the helpers in both the uncoded caching case and coded caching case. For uncoded caching, the users can communicate with multiple helpers and therefore may have conflicting interests on optimum content placement in the shared helpers because of different transmission speeds on those links. The cache content allocation became a combinatorial distributed caching problem requiring sophisticated algorithms to solve it. To this end, they modeled the network as a bipartite graph, as shown in Fig. 2.4, with the weight of each edge indicating the associated transmission speed of the link between the connected user and helper. Now that the uncoded caching problem became a monotone submodular function maximization with matroid constraints and NP-hard to get the optimal solution, they resorted to the greedy algorithm by caching the file that brought the greatest caching gain in each iteration [32].

Opposite to the uncoded caching case, where the content assignment matrix $\mathbf{X}$ was binary, that in the coded caching case satisfied $\mathbf{X} \in [0,1]^{N\times H}$. N and H denoted the number of files and the number of helpers, respectively. The cache content allocation became convex, and can be linearized by introducing new variables and then solved in a distributed manner. Even if we need to deal with thousands of files in large-scale networks, approximation methods are available to solve the problem, which makes the coded caching scheme gain better applicability than the uncoded caching schemes. In addition, the coded caching scheme makes better use of the

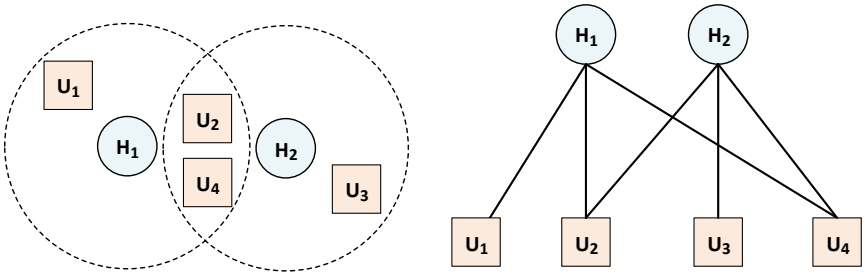


Figure 2.4: An example of a bipartite graph indicating the connectivity between the UTs and the helpers.

content diversity and therefore improves the memory utilization.

B. Coded Caching (Coded-Multicasting)

Coded caching was firstly proposed by Mohammad Ali Maddah-Ali and Urs Niesen in [52] where a new information-theoretic model was formulated for the caching problem. They investigated both the content placement and delivery phases so that the demands of different users can be satisfied with a single coded multicast transmission in the delivery phase to achieve a global caching gain. In their model, a single server network with K users was considered, where all the users were connected to the server through a shared, error-free link. The server can access all the N files ($N \leq K$) while each user can store at most M files in its local cache and request a single file at a time. They also assumed simplest settings, such as uniform file popularity, file size and cache size. The aim was to minimize the rate, i.e., the load of the shared link in the delivery phase, in worst case.

To illuminate the theory of coded multicast, here we introduce an example of a typical shared network consisting of a single server, two users and two files, depicted in Fig. 2.5. We can then easily derive that when cache size $M = 0$ and $M = 2$, the multicast rate equals to 2 and 0, respectively. More generally, when $M = 1$, both of the files, denoted as A and B, are equally split into two subfiles, i.e., $A = (A1, A2)$ and $B = (B1, B2)$. We let each user cache one disjoint subfile of each file, e.g. $Z1 = (A1, B1)$ and $Z2 = (A2, B2)$, and consider the worst-case scenario when each user requests different files, for example that user one requests file A and

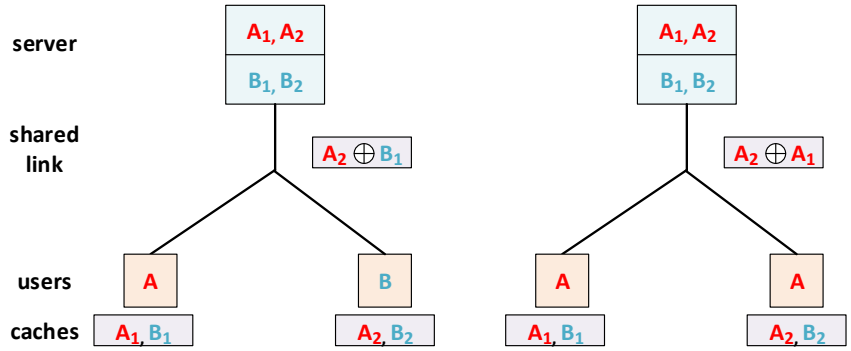


Figure 2.5: An example of coded caching strategy for two files (A,B), two users and cache size $M = 1$ with two typical user requests.

user two requests file B. Normally, we can deliver $A2$ to user one and $B1$ to user two. In coded multicast, we reduce the backhaul load by making use of both of the cached packages and the delivered packages to decode the missing subfiles. That is due to the fact that each user may has part of the file that the other user needs. In this case, we utilize bitwise XOR denoted as $\oplus$ and let the server multicast $A2 \oplus B1$ to both of the users. Now that user one has cached $B1$, it can easily recover $A2$ from $A2 \oplus B1$. And user two can recover $B1$ using cached content $A2$ as well. In so doing, the rate drops from 2 subfiles to 1 subfiles. The same logic can be used to deal with all other possible requests and larger networks. Through the derivation, the rates for uncoded caching and coded caching, denoted by $R(M)$ and $R^*(M)$, can be written as

$$R(M) = K \cdot (1 - M/N), \tag{2.12}$$

$$R^*(M) = K \cdot (1 - M/N) \cdot \frac{1}{1 + KM/N}, \tag{2.13}$$

where K is the rate without caching. Thus, coded caching brings an global caching gain of $\frac{1}{1+KM/N}$ in addition to the local caching gain $(1 - M/N)$.

Note that in this case, the placement phase must be carefully designed in order to guarantee simultaneous multicasting opportunities for all possible requests to achieve the global caching gain. It is more challenging for complex network topologies with heterogeneous network settings, [82–89]. In addition, efforts have

also been done to extend the results of the centralized, offline caching scheme to decentralized coded caching [120], and online coded caching [121]. The former no longer relies on the server to know and control the placement and delivery phases while the later takes the evolution of content popularity into account and proposes a content replacement protocol.

C. Probabilistic Caching

Probabilistic caching was firstly used for cache-enabled HetNets in [50]. In this model, the probability distributions for cache content placement at different BSs are the same. Now that define the file library as $\{c_1, c_2, \ldots, c_N\}$ with equal file size, and the cache size in each BS as M (normalized by the file size), the probability for file j being stored at BS i can be written as

$$b_j = \mathbb{P}(c_j \in \Xi^i), \tag{2.14}$$

where Ξ^i is the exact content distribution of BS i, and $\{b_j\}$ satisfy

$$\sum_j b_j \leq M, \tag{2.15}$$

$$0 \leq b_j \leq 1, \forall j. \tag{2.16}$$

The target of this model is to maximize the total hit probability of the typical user which can be expressed as follows:

$$f(b_1, \ldots, b_N) = 1 - \sum_j a_j \sum_m p_m (1 - b_j)^m, \tag{2.17}$$

with m being the number of BSs that covers the user, and p_m being the probability for the considered user being covered by m BSs. a_j denotes the content popularity for file j, i.e., the probability for file j being requested.

Once the probability vector $\{b_j\}$ has been designed, there is an easy way for the BSs to implement the content placement accordingly. As shown in the example of Fig. 2.6 which assumes equality in (2.15), the cache memory is divided into M continuous chunks of unit length. Then we fill the memory chunks with the probability values of the N files, $\{b_j\}$, one by one. If the current chunk has been occupied, the remaining part of the current probability fills the chunk that follows. We then

uniformly pick a random number within $[0,1]$ and draw a vertical line which crosses the memory chunks. As $b_j \in [0,1]$ and the space between two adjacent intersections always equals to an unit length, we can always guarantee that the intersections appear at M distinct contents. Moreover, the probability for one of the intersections appearing at file j exactly equals to b_j as depicted in Fig. 2.6. For instance, file 2 is cached only when the selected random falls in $(b_1, b_2+b_1]$ when $(b_1+b_2 \leq 1)$, or $[0, b_1+b_2-1] \cup (b_1, 1]$ when $(b_1+b_2 > 1)$. In both of the cases, the total probability of caching this file is b_2.

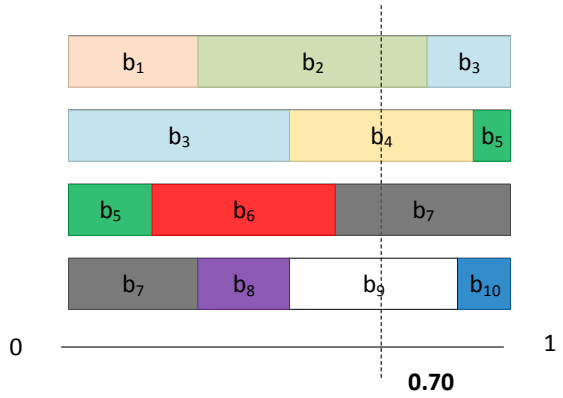


Figure 2.6: An example of the probabilistic placement policy when $N = 10$ and $M = 4$. Drawn uniformly a random number (0.7), the vertical line intersects with the memory chunks at $\{c2, c4, c7, c9\}$, respectively, i.e. the four files will be cached.

Compared with deterministic caching strategies, the random-caching-based methods, e.g. probabilistic caching, are more suitable for the optimization and analysis of random topologies, e.g., HetNets, or high mobility scenarios. However, the associated caching problem may become too complicated to develop any non-iterative methods.

Chapter 3

Robust Beamforming for SWIPT Broadcast Channels

3.1 Overview

To combat the effects of multipath fading in energy harvesting enabled networks, e.g. decaying the power transfer efficiency as well as the spectral efficiency and hence hindering long-distance SWIPT, multiple antenna techniques are adopted at transmitters and/or receivers to provide spatial energy and information diversity gains. In this case, the beamforming and energy transfer ratios need to be jointly designed in order to achieve trade-offs between the wireless information transmission and energy transfer, e.g. ensuring particular signal to noise ratio (SNR) while satisfying least harvested energy threshold as discussed in [10]. In particular, the broadcast channel is a typical multiuser network of great interests, where the base station (BS) communicates with several mobile stations (MSs). Using SWIPT, each MS can be an information decoder (ID) as well as energy receiver (ER), either by time-switching or power splitting technologies. In this chapter, we consider MISO broadcast system for SWIPT using receiver power splitting and aim to optimize jointly the beamforming vectors and the power splitting ratios for minimizing the transmit power of the base station subject to the individual SINR and the energy-harvesting constraints at the MSs. However, the CSI is assumed imperfect but has a deterministic uncertainty region. Unlike existing attempts that resort to iterations

guided by semi-definite relaxation (SDR), we propose a reverse convex non-smooth optimization algorithm, which provides the near-optimal rank-one solution.

3.2 Related Work

As mentioned, the SWIPT systems with the communication nodes adopting multi-antenna techniques have received considerable attentions so as to facilitate long-distance wireless energy transfer. As opposite to the MIMO channels, the beamforming design in the multiple-input single-output (MISO) channels always meets a crucial rank-one issue of proving the tightness of SDR which makes it challenging to obtain the optimal solution. Instead, approximation strategies, e.g. randomization are used to get the suboptimal solution. [10] and [54] provided pioneer works on the beamforming design for MISO broadcast SWIPT systems with single or multiple separate information and energy receivers. Recently, the joint optimization problem of power splitting ratios and beamforming was studied in [56] for MISO broadcast SWIPT systems with multiple co-located information and energy receivers assuming perfect CSI at the BS. Later in [68], the results were extended to cope with the case of imperfect CSI, via a highly complex *suboptimal* two-step optimization process, which relies on alternatively solving SDR problems with a K-dimensional search (where K denotes the number of users in the broadcast system). The greedy searching algorithm not only imposed high computing load, but also lacked careful clarification on the rank-one issue. In this chapter, we revisit the problem in [68] which aims to minimize the transmit power of the BS subject to the SINR and the energy harvesting constraints at the MSs, assuming the availability of imperfect CSI at the BS, for the MISO SWIPT broadcast system. For MIMO SWIPT broadcasting in [10], there is no rank-one issue and this means that existing results will not be optimal for MISO SWIPT systems. The contributions of our proposed approach over [68] are twofold: (i) significant complexity reduction and (ii) near-optimality. In particular, we present a feasible SDR-guided randomization approach for the joint optimization of transmit beamforming and receive power splitting factors. In contrast to [68], the SDR-based solution is *non-iterative* but only provides an upper-

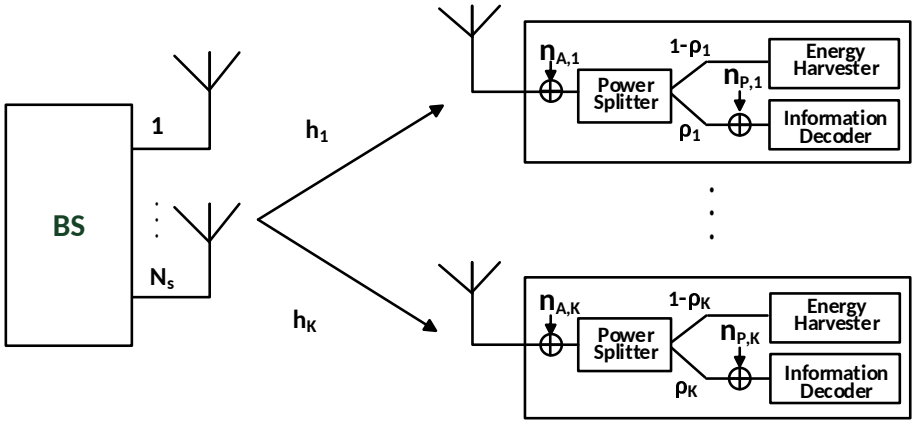


Figure 3.1: A MISO SWIPT broadcast system with power splitters.

bound performance after rescaling. Hence, we propose a reverse convex constraint based penalty function method which guarantees a rank-one and near-optimal solution.

3.3 System Model

Consider a K-user MISO broadcast system as illustrated in Fig. 3.1, where the BS, with N_{s} antennas, communicates with K single-antenna MSs. Each MS acts simultaneously as an ID and an ER via power splitting. With transmit beamforming at the BS, the received signal at the kth MS can be written as

$$y_k = \mathbf{h}_k^H \sum_{i=1}^{K} \mathbf{b}_i s_i + n_{\mathrm{A},k}, \text{ for } k = 1,\ldots,K, \tag{3.1}$$

where $\mathbf{b}_i$ and s_i denote the transmit beamforming vector and the data symbol for the ith MS, respectively, $\mathbf{h}_k$ is the channel vector between the BS and the kth MS, $n_{\mathrm{A},k}$ is the antenna noise at the kth MS, and $(\cdot)^H$ is the Hermitian operation.

With a power splitter at the kth MS, suppose that we have the power splitting ratio $\rho_k \in [0,1]$. Then the signal split to the ID of the kth receiver is given by

$$y_{\mathrm{I},k} = \sqrt{\rho_k}\left(\mathbf{h}_k^H \sum_{k=1}^{K} \mathbf{b}_i s_i + n_{\mathrm{A},k}\right) + n_{\mathrm{P},k}, \tag{3.2}$$

where $n_{\mathrm{P},k}$ denotes the additive noise at the ID of the kth MS. Meanwhile, the signal split to the energy harvester of the kth MS can be expressed as

$$y_{\mathrm{E},k} = \sqrt{1-\rho_k}\left(\mathbf{h}_k^H \sum_{i=1}^{K} \mathbf{b}_i s_i + n_{\mathrm{A},k}\right). \tag{3.3}$$

As such, the SINR of the ID at the kth MS is given by

$$SINR_k = \frac{\rho_k \mathbf{h}_k^H \mathbf{b}_k \mathbf{b}_k^H \mathbf{h}_k}{\rho_k \sigma_{\mathrm{A},k}^2 + \sigma_{\mathrm{P},k}^2 + \rho_k \mathbf{h}_k^H \left(\sum_{\substack{i=1 \\ i \neq k}}^{K} \mathbf{b}_i \mathbf{b}_i^H \right) \mathbf{h}_k}, \tag{3.4}$$

and the power harvested at the kth MS is written as

$$\mathrm{E}_k = \xi_k (1 - \rho_k) \left(\mathbf{h}_k^H \left(\sum_{i=1}^{K} \mathbf{b}_i \mathbf{b}_i^H \right) \mathbf{h}_k + \sigma_{\mathrm{A},k}^2 \right), \tag{3.5}$$

where $\xi_k \in (0, 1]$ is the energy conversion efficiency of the energy harvester, and $\mathbb{E}\{|s_i|^2\} = 1$ has been assumed.

To guarantee desirable QoS in communication and also enough power in the energy harvesting, the beamforming design for MISO broadcast SWIPT should satisfy the SINR constraints and energy harvesting constraints which are written as

$$SINR_k = \frac{\rho_k \mathbf{h}_k^H \mathbf{b}_k \mathbf{b}_k^H \mathbf{h}_k}{\rho_k \sigma_{\mathrm{A},k}^2 + \sigma_{\mathrm{P},k}^2 + \rho_k \mathbf{h}_k^H \left(\sum_{\substack{i=1 \\ i \neq k}}^{K} \mathbf{b}_i \mathbf{b}_i^H \right) \mathbf{h}_k} \geq \gamma_k, \tag{3.6}$$

and

$$\mathrm{E}_k = \xi_k (1 - \rho_k) \left(\mathbf{h}_k^H \left(\sum_{i=1}^{K} \mathbf{b}_i \mathbf{b}_i^H \right) \mathbf{h}_k + \sigma_{\mathrm{A},k}^2 \right) \geq \eta_k, \tag{3.7}$$

where $\gamma_k > 0$ and $\eta_k > 0$ are the given SINR and energy harvesting thresholds at the kth MS, respectively.

3.4 Robust Optimization

Here we model the channel by

$$\mathbf{h}_k = \hat{\mathbf{h}}_k + \Delta \mathbf{h}_k, \text{ for } k = 1, \ldots, K, \tag{3.8}$$

where $\mathbf{h}_k$ is the actual channel vector, but $\hat{\mathbf{h}}_k$ denotes the CSI estimate with an error vector $\Delta \mathbf{h}_k$, which satisfies

$$\|\Delta \mathbf{h}_k\|_2 = \|\mathbf{h}_k - \hat{\mathbf{h}}_k\|_2 \leq \varepsilon_k, \text{ for } \varepsilon_k \geq 0, \tag{3.9}$$

where $\{\varepsilon_k\}$ denotes the given threshold of the channel uncertainty reflecting the quality of the estimates.

We aim to minimize the BS transmit power subject to the SINR and the energy-harvesting constraints at the MSs as

$$\min_{\substack{\{\mathbf{b}_k\}\\ \{0<\rho_k<1\}}} \sum_{k=1}^{K} \mathbf{b}_k^H \mathbf{b}_k \ \ s.t. \tag{3.10a}$$

$$\min_{\Delta\mathbf{h}_k} \frac{|\mathbf{b}_k^H(\mathbf{h}_k+\Delta\mathbf{h}_k)|^2}{\gamma_k} - \sum_{\substack{i=1\\ i\neq k}}^{K} |\mathbf{b}_i^H(\mathbf{h}_k+\Delta\mathbf{h}_k)|^2 \geq \sigma_{\mathrm{A},k}^2 + \frac{\sigma_{\mathrm{P},k}^2}{\rho_k}, \forall k, \tag{3.10b}$$

$$\min_{\Delta\mathbf{h}_k} \sum_{i=1}^{K} |\mathbf{b}_i^H(\mathbf{h}_k+\Delta\mathbf{h}_k)|^2 \geq \frac{\eta_k}{\xi_k(1-\rho_k)} - \sigma_{\mathrm{A},k}^2, \forall k, \tag{3.10c}$$

$$\|\Delta\mathbf{h}_k\|_2 \leq \varepsilon_k, \forall k. \tag{3.10d}$$

Due to imperfect CSI, however, our problem is not convex and has infinitely many constraints as opposed to that in [56].

Lemma 3.1 *(S-Procedure): Let* $f_i(\boldsymbol{x}) = \boldsymbol{x}^H \boldsymbol{A}_i \boldsymbol{x} + 2\boldsymbol{b}_i^H \boldsymbol{x} + c_i, i = 1,2$ *where* $\boldsymbol{A}_i \in \mathbb{C}^{n\times n}, \boldsymbol{b}_i \in \mathbb{C}^n$ *and* $c_i \in \mathbb{R}$. *The implication* $f_1(\boldsymbol{x}) \leq 0 \Rightarrow f_2(\boldsymbol{x}) \leq 0$ *holds if and only if there exists* $\mu \geq 0$ *satisfying [45, 73]*

$$\mu \begin{bmatrix} \boldsymbol{A}_1 & \boldsymbol{b}_1 \\ \boldsymbol{b}_1^H & c_1 \end{bmatrix} - \begin{bmatrix} \mathbf{A}_2 & \boldsymbol{b}_2 \\ \mathbf{b}_2^H & c_2 \end{bmatrix} \succeq \mathbf{0}. \tag{3.11}$$

Now, we define $\mathbf{W}_k \triangleq \mathbf{b}_k \mathbf{b}_k^H$ and substitute (3.10d) and (3.10b) into *Lemma 1*. Then for any arbitrary k, we obtain that $f_1^k(\Delta\mathbf{h}_k) = \Delta\mathbf{h}_k^H \Delta\mathbf{h}_k - \varepsilon_k$ with $\mathbf{A}_1^k = \mathbf{I}_{N_s}, \mathbf{b}_1^k = 0, c_1^k = -\varepsilon_k$. In the same way, we can derive that $f_2^k(\Delta\mathbf{h}_k) = \Delta\mathbf{h}_k^H \mathbf{A}_2^k \Delta\mathbf{h}_k + \mathbf{b}_2^k \Delta\mathbf{h}_k + c_2^k$ with

$$\mathbf{A}_2^k = -\frac{\mathbf{W}_k}{\gamma_k} + \sum_{\substack{i=1\\ i\neq k}}^{K} \mathbf{W}_i \tag{3.12}$$

$$\mathbf{b}_2^k = \mathbf{A}_2^k \mathbf{h}_k, \tag{3.13}$$

$$c_2^k = \sigma_{\mathrm{A},k}^2 + \frac{\sigma_{\mathrm{P},k}^2}{\rho_k} + \mathbf{h}_k^H \mathbf{A}_2^k \mathbf{h}_k, \tag{3.14}$$

where the superscript k indicates the reformulation is done in terms of user k. We will then get a group of positive semi-definite matrices named $\{\mathbf{\Gamma}_k\}$ (see (3.15)), associated with parameters $\{\mu_k\}$. Similarly, if we apply S-Procedure to (3.10d) and

(3.10c), then we will have another group of semi-definite matrices $\{\mathbf{\Upsilon}_k\}$ (see (3.16)), associated with parameters $\{\lambda_k\}$. To be brief, we let $\mathbf{\Pi}_k = \frac{\mathbf{W}_k}{\gamma_k} - \sum_{\substack{i=1 \\ i\neq k}}^{K} \mathbf{W}_i, \mathbf{\Theta}_k = \sum_{i=1}^{K} \mathbf{W}_i, \phi_k = \sigma_{\mathrm{A},k}^2 + \frac{\sigma_{\mathrm{P},k}^2}{\rho_k} - \mu_k \varepsilon_k$, and $\varphi_k = \frac{\eta_k}{\xi_k(1-\rho_k)} + \sigma_{\mathrm{A},k}^2 - \lambda_k \varepsilon_k$.

$$\mathbf{\Gamma}_k(\{\mathbf{W}_k\}, \rho_k, \mu_k) = \begin{bmatrix} \mu_k \mathbf{I}_{N_s} + \mathbf{\Pi}_k & \mathbf{\Pi}_k \mathbf{h}_k \\ \mathbf{h}_k^H \mathbf{\Pi}_k & \mathbf{h}_k^H \mathbf{\Pi}_k \mathbf{h}_k - \phi_k \end{bmatrix} \succeq \mathbf{0}, \tag{3.15}$$

$$\mathbf{\Upsilon}_k(\{\mathbf{W}_k\}, \rho_k, \lambda_k) = \begin{bmatrix} \lambda_k \mathbf{I}_{N_s} + \mathbf{\Theta}_k & \mathbf{\Theta}_k \mathbf{h}_k \\ \mathbf{h}_k^H \mathbf{\Theta}_k & \mathbf{h}_k^H \mathbf{\Theta}_k \mathbf{h}_k - \varphi_k \end{bmatrix} \succeq \mathbf{0}. \tag{3.16}$$

Based on the S-Procedure, (3.10) becomes

$$\min_{\substack{\{\mathbf{W}_k\},\{\rho_k\}, \\ \{\mu_k\},\{\lambda_k\}}} \sum_{k=1}^{K} tr(\mathbf{W}_k) \quad \text{s.}t. \tag{3.17a}$$

$$\mathbf{\Gamma}_k(\{\mathbf{W}_k\}, \rho_k, \mu_k) \succeq \mathbf{0}, \forall k, \tag{3.17b}$$

$$\mathbf{\Upsilon}_k(\{\mathbf{W}_k\}, \rho_k, \lambda_k) \succeq \mathbf{0}, \forall k, \tag{3.17c}$$

$$\mathbf{W}_k \succeq \mathbf{0}, 0 < \rho_k < 1, \forall k, \tag{3.17d}$$

$$\mu_k \geq 0, \lambda_k \geq 0, \forall k, \tag{3.17e}$$

$$\mathrm{R}ank(\mathbf{W}_k) = 1, \forall k. \tag{3.17f}$$

In terms of $\{\rho_k\}$, ϕ_k and φ_k are both convex as the corresponding second derivatives are positive. Ignoring the rank-one constraint, problem (3.17) will be convex but cannot be solved by optimization packages CVX [73] due to the coupling of $\frac{1}{\rho_k}$ and $\frac{1}{1-\rho_k}$ in $\mathbf{\Gamma}_k$ and $\mathbf{\Upsilon}_k$. This was why [68] resorted to iterative suboptimal approaches. Here, we propose to solve the problem by introducing a group of new variables, q_k and $\tilde{q}_k$ to get a definitely convex problem after rank relaxation

which can be processed by existing solvers:

$$\min_{\substack{\{\mathbf{W}_k\},\{\rho_k\},\{q_k\},\\ \{\tilde{q}_k\},\{\mu_k\},\{\lambda_k\}}} \sum_{k=1}^{K} tr(\mathbf{W}_k) \quad \text{s.}t. \tag{3.18a}$$

$$\tilde{\mathbf{\Gamma}}_k(\{\mathbf{W}_k\}, q_k, \mu_k) \succeq \mathbf{0}, \forall k, \tag{3.18b}$$

$$\tilde{\mathbf{\Upsilon}}_k(\{\mathbf{W}_k\}, \tilde{q}_k, \lambda_k) \succeq \mathbf{0}, \forall k, \tag{3.18c}$$

$$\mathbf{W}_k \succeq \mathbf{0}, 0 < \rho_k < 1, \forall k, \tag{3.18d}$$

$$q_k \geq \frac{1}{\rho_k}, \tilde{q}_k \geq \frac{1}{1-\rho_k}, \forall k, \tag{3.18e}$$

$$\mu_k \geq 0, \lambda_k \geq 0, \forall k. \tag{3.18f}$$

$$\mathrm{R}ank(\mathbf{W}_k) = 1, \forall k. \tag{3.18g}$$

where $\tilde{\mathbf{\Gamma}}_k$ and $\tilde{\mathbf{\Upsilon}}_k$ are similar to those of $\mathbf{\Gamma}_k$ and $\mathbf{\Upsilon}_k$ except that we change $\frac{1}{\rho_k}$ and $\frac{1}{1-\rho_k}$ to q_k and $\tilde{q}_k$, respectively.

Proposition 3.1 *Regardless of the new variables q_k and $\tilde{q}_k$, problems (3.17) and (3.18) are equivalent. The optimal solution to either of the two problems should also be optimal for the other one.*

Proof 3.1 *Please refer to* Appendix A.

However, the rank-one constraint makes both (3.17) and (3.18) non-convex. To tackle this, SDR with randomization is used.

3.4.1 SDR Guided Randomization

In particular, the rank constraint is first dropped to obtain a suboptimal solution. Then the randomization technique is used to generate the feasible solutions to (3.17). Assuming that the solution of SDR is $\mathbf{W}_k^*, \forall k$, with the eigenvalue decomposition defined as $\mathbf{W}_k^* = \mathbf{U}\mathbf{\Sigma}\mathbf{U}^H$, the feasible beamforming vector of (3.17) under randomization can then be given by

$$\mathbf{b}_k = \mathbf{U}\mathbf{\Sigma}^{\frac{1}{2}}\mathbf{v}. \tag{3.19}$$

Here $\mathbf{U}$ is unitary and $\mathbf{\Sigma}$ is diagonal with eigenvalue arranged in decreasing order, and $\mathbf{v}$ is a vector of complex circularly symmetric uncorrelated Gaussian random

variables with zero-mean and unit-variance. However, some of the constraints in (3.17) may be violated after randomization, and one needs to rescale the beamforming vector $\mathbf{b}_k$ with an appropriate factor α_k to meet the constraints. Thus we have

$$\widetilde{\mathbf{b}}_k = \alpha_k \mathbf{b}_k, \forall k. \tag{3.20}$$

Then we reformulate the problem as follows and rely on CVX to derive the optimal scaling factors:

$$\min_{\{\beta_k\}} \sum_{k=1}^{K} tr(\widetilde{\mathbf{W}}_k) \ \ s.t. \tag{3.21a}$$

$$\tilde{\mathbf{\Gamma}}_k(\{\widetilde{\mathbf{W}}_k\}, q_k^*, \mu_k^*) \succeq \mathbf{0}, \forall k, \tag{3.21b}$$

$$\tilde{\mathbf{\Upsilon}}_k(\{\widetilde{\mathbf{W}}_k\}, \tilde{q}_k^*, \lambda_k^*) \succeq \mathbf{0}, \forall k, \tag{3.21c}$$

$$\widetilde{\mathbf{W}}_k \succeq \mathbf{0}, \forall k, \tag{3.21d}$$

where $\widetilde{\mathbf{W}}_k = \beta_k \widetilde{\mathbf{b}}_k \widetilde{\mathbf{b}}_k^H, \beta_k = \alpha_k^2$, and $q_k^*, \mu_k^*, \tilde{q}_k^*, \lambda_k^*$ are the corresponding solution by the SDP approach. With the optimal scaling factors, we can easily generate $\mathbf{b}_k$ using (3.19) and (3.20). The downside is that randomization always offers worse performance due to the relaxation involved. As a remedy, in the following, we solve the problem by expressing the rank-one constraint (3.18g) as a single reverse convex constraint which is then incorporated into the objective function as a penalty function. The resulting problem belongs to the class of concave programming with a nonsmooth objective.

3.4.2 Penalty Function Method

Since $\mathbf{W}_k, \forall k$ is always positive semi-definite, we then have $tr(\mathbf{W}_k) \geq \lambda_{max}(\mathbf{W}_k)$ where $\lambda_{max}(\mathbf{W}_k)$ is the maximum eigenvalue of $\mathbf{W}_k$. In this case, if $tr(\mathbf{W}_k) \leq \lambda_{max}(\mathbf{W}_k)$ also holds, it will be easy to prove that $tr(\mathbf{W}_k) = \lambda_{max}(\mathbf{W}_k)$. That is to say, $\mathbf{W}_k$ has only one non-zero eigenvalue. Then we will have $Rank(\mathbf{W}_k) = 1, \forall k$. Thus, the rank-one constraints (3.18g) can be expressed by the following constraint [130]:

$$\sum_{k=1}^{K} (tr(\mathbf{W}_k) - \lambda_{max}(\mathbf{W}_k)) \leq 0. \tag{3.22}$$

Substitute (3.22) into problem (3.18), and we can then obtain

$$\min_{\substack{\{\mathbf{W}_k\},\{\rho_k\},\{q_k\},\\ \{\tilde{q}_k\},\{\mu_k\},\{\lambda_k\}}} \sum_{k=1}^{K} tr(\mathbf{W}_k) \quad \text{s.}t. \tag{3.23a}$$

$$(3.18\text{b})\text{–}(3.18\text{f}),\ (3.22). \tag{3.23b}$$

Note that the function $\lambda_{max}(\mathbf{X})$ is proved to be convex on the set of Hermitian matrices [129]. Therefore, we can easily derive that $\sum_{k=1}^{K}(tr(\mathbf{W}_k)-\lambda_{max}(\mathbf{W}_k))$ is actually a concave function of $\mathbf{W}_k$, which is to say that (3.22) is a reverse convex constraint [128]. Consequently, problem (3.23) is now a convex program with additional reverse convex constraint, a typical type of nonconvex global optimization [128, 130].

It is worth pointing out that when $tr(\mathbf{W}_k)-\lambda_{max}(\mathbf{W}_k)$ is small enough, we will have $\mathbf{W}_k \approx \lambda_{max}(\mathbf{W}_k)\mathbf{w}_{k,max}\mathbf{w}_{k,max}^H$, where $\mathbf{w}_{k,max}$ denotes the unit-norm eigenvector corresponding to the maximum eigenvalue $\lambda_{max}(\mathbf{W}_k)$ (i.e., $\left\|\mathbf{w}_{k,max}\right\|=1$). Then the optimal beamforming vector can be obtained as

$$\mathbf{b}_k = \lambda_{max}(\mathbf{W}_k)^{\frac{1}{2}}\mathbf{w}_{k,max}, \tag{3.24}$$

satisfying the rank-one constraints (3.18g). Our aim is therefore to make $\sum_{k=1}^{K}(tr(\mathbf{W}_k)-\lambda_{max}(\mathbf{W}_k))$ as small as possible. Thus we consider the alternative formulation to (3.18):

$$\min_{\substack{\{\mathbf{W}_k\},\{\rho_k\},\{q_k\},\\ \{\tilde{q}_k\},\{\mu_k\},\{\lambda_k\}}} \quad \tau \tag{3.25a}$$

$$\text{s.}t. \quad (3.18\text{b})\text{–}(3.18\text{f}). \tag{3.25b}$$

where $\tau \triangleq \sum_{k=1}^{K}(tr(\mathbf{W}_k)+\kappa(tr(\mathbf{W}_k)-\lambda_{max}(\mathbf{W}_k)))$ and $\kappa>0$ is a constant. If the weight κ is chosen to be large enough, then the difference $tr(\mathbf{W}_k)-\lambda_{max}(\mathbf{W}_k)$ will be minimized. Clearly, the objective of (3.25) is to minimize both $\sum_{k=1}^{K}tr(\mathbf{W}_k)$ and $\sum_{k=1}^{K}(tr(\mathbf{W}_k)-\lambda_{max}(\mathbf{W}_k))$.

Lemma 3.2 *Let* $\mathbf{X}$ *and* $\mathbf{Y}$ *be positive semidefinite matrices. Using the fact that a sub-gradient of* $\lambda_{max}(\mathbf{Y})$ *is* $\mathbf{y}_{max}\mathbf{y}_{max}^H$*, we always have* $\lambda_{max}(\mathbf{X})-\lambda_{max}(\mathbf{Y}) \geq \mathbf{y}_{max}^H(\mathbf{X}-\mathbf{Y})\mathbf{y}_{max}$*. Meanwhile,* $\lambda_{max}(\mathbf{Y})$ *and* $\mathbf{y}_{max}$ *denote the maximum eigenvalue and corresponding eigenvector of* $\mathbf{Y}$*, respectively.*

According to *Lemma 2*, given some feasible $\mathbf{W}_k^{(n)}$ of problem (3.25), we obtain

$$\begin{aligned}\mathrm{tr}(\mathbf{W}_k^{(n+1)}) + \kappa\Big[\mathrm{tr}(\mathbf{W}_k^{(n+1)}) - \lambda_{max}(\mathbf{W}_k^{(n)}) \\ - (\mathbf{w}_{k,max}^{(n)})^H \left(\mathbf{W}_k^{(n+1)} - \mathbf{W}_k^{(n)}\right) \mathbf{w}_{k,max}^{(n)}\Big] \\ \leq \mathrm{tr}(\mathbf{W}_k^{(n)}) + \kappa\left(\mathrm{tr}(\mathbf{W}_k^{(n)}) - \lambda_{max}(\mathbf{W}_k^{(n)})\right), \end{aligned} \tag{3.26}$$

where the superscript n denotes the n-th iteration. Accordingly, the following SDP problem gives an optimal solution $\mathbf{W}_k^{(n+1)}$ that is better than $\mathbf{W}_k^{(n)}$ of problem (3.25):

$$\begin{aligned}\min_{\substack{\{\mathbf{W}_k\},\{\rho_k\},\{q_k\},\\ \{\tilde{q}_k\},\{\mu_k\},\{\lambda_k\}}} \sum_{k=1}^{K} \mathrm{tr}(\mathbf{W}_k^{(n+1)}) + \kappa\Big[\mathrm{tr}(\mathbf{W}_k^{(n+1)}) \\ -\lambda_{max}(\mathbf{W}_k^{(n)}) - (\mathbf{w}_{k,max}^{(n)})^H \left(\mathbf{W}_k^{(n+1)} \\ -\mathbf{W}_k^{(n)}\right)\mathbf{w}_{k,max}^{(n)}\Big] \end{aligned} \tag{3.27a}$$

$$\text{s.}t.\ \ (3.18b)\text{–}(3.18f). \tag{3.27b}$$

Now, (3.27) can be further simplified to

$$\min_{\substack{\{\mathbf{W}_k\},\{\rho_k\},\{q_k\},\\ \{\tilde{q}_k\},\{\mu_k\},\{\lambda_k\}}} \sum_{k=1}^{K} \mathrm{tr}\left(\mathbf{W}_k\right) + \kappa\left[\mathrm{tr}(\mathbf{W}_k) - (\mathbf{w}_{k,max}^{(n)})^H \mathbf{W}_k \times \mathbf{w}_{k,max}^{(n)}\right] \tag{3.28a}$$

$$\text{s.}t.\ \ (3.18b)\text{–}(3.18f). \tag{3.28b}$$

Due to the initial condition $\mathrm{tr}(\mathbf{W}_k^{(0)}) - \lambda_{max}(\mathbf{W}_k^{(0)}) = 0$, at some n, we will have $\mathrm{tr}(\mathbf{W}_k^{(n)}) - \lambda_{max}(\mathbf{W}_k^{(n)}) = 0$.

The proposed nonsmooth iterative algorithm to resolve the rank-one beamforming problem is summarized in Algorithm 1. $\{\zeta_k\}$ is the iteration terminating threshold.

3.4.3 Complexity Analysis

Firstly, the *suboptimal* two-step optimization process in [68] is highly complex, because it requires alternatively solving SDR problems with a K-dimensional search. That is to say we need to solve $\frac{K}{\Delta\rho}$ SDP problems (where $\Delta\rho$ denotes step length), which makes the two-step optimization much more computationally demanded than the proposed methods. The complexity for a single SDP is

Algorithm 1 A nonsmooth iterative algorithm

1. **Initialization**

Choose a proper value of $\kappa > 0$ and a feasible solution $(\mathbf{W}_k^{(0)}, \rho_k^{(0)}), \forall k$, of (3.28).

Set $n = 0$.

2. **Repeat**

a. Solve problem (3.28) to obtain $\mathbf{W}_k^{(n+1)}$, and $\rho_k^{(n+1)}, \forall k$.

b. **if** $\mathbf{W}_k^{(n+1)} = \mathbf{W}_k^{(n)}$ then

set $\kappa = 2\kappa$

end if

c. $n = n+1$

Until $tr(\mathbf{W}_k^{(n)}) \approx \lambda_{max}(\mathbf{W}_k^{(n)})$

3. **Reset** $\mathbf{W}_k^{(0)} = \mathbf{W}_k^{(n)}, \rho_k^{(0)} = \rho_k^{(n)}, n = 0$.

4. **Repeat**

a. Solve problem (3.28) to obtain $\mathbf{W}_k^{(n+1)}$ and $\rho_k^{(n+1)}$.

b. $n = n+1$

Until $tr(\mathbf{W}_k^{(n)}) \approx \lambda_{max}(\mathbf{W}_k^{(n-1)})$, i.e. $\left| tr(\mathbf{W}_k^{(n)}) - \lambda_{max}(\mathbf{W}_k^{(n-1)}) \right| \leq \zeta_k$

5. **Calculate $\mathbf{b}_k$ according to (3.24)**.

$O(n_{\text{sdp}}^{0.5}(m_{\text{sdp}}n_{\text{sdp}}^3 + m_{\text{sdp}}^2 n_{\text{sdp}}^2 + m_{\text{sdp}}^3))$. Meanwhile, n_{sdp} denotes the dimension of the positive semidefinite cone and m_{sdp} is the number of constraints. The sum complexity would be $O(\frac{K}{\Delta\rho} n_{\text{sdp}}^{0.5}(m_{\text{sdp}}n_{\text{sdp}}^3 + m_{\text{sdp}}^2 n_{\text{sdp}}^2 + m_{\text{sdp}}^3))$. In our considered robust beamforming design, it holds true that $n_{\text{sdp}} = N_s + K + 1$ and $m_{\text{sdp}} = K + 1$. Our proposed SDP method, which is based on reformulation, only requires to solve a single SDP without iteration. The SDP guided randomization comprises of two SDPs, to derive the optimal $\{\mathbf{W}_k\}$ and the optimal scaling factors $\{\beta_k\}$, respectively. The complexity of those SDPs is all approximate to that of the ones in [68] according to the structures of the problems. Though each iteration requires an SDP solver, it has been noticed that the PenFun method only requires tens of iterations to converge, which makes it competitive in terms of computational complexity in addition to the performance advantages. In addition, in the perfect CSI case [56], where full knowledge of CSI is assumed, the rank-one guaranteed optimal solution is achieved

by solving a single SDP.

3.5 Simulation Results

In this section, the performance of the proposed methods is investigated via simulations. We considered $\gamma_k = \gamma, \eta_k = \eta$, $\xi_k = 0.5$, $\sigma_{\mathrm{A},k}^2 = 10^{-8}$, ζ_k, and $\sigma_{\mathrm{P},k}^2 = 10^{-6}, \varepsilon_k = 0.001, \forall k$. In each realization, the frequency-flat channels are generated according to Rician fading channel modeling. The channel vector is modeled as

$$\mathbf{h}_k = \frac{1}{\sqrt{d_k^{m_k}}} \left(\sqrt{\frac{K_R}{1+K_R}} \mathbf{h}_k^{\mathrm{L}OS} + \sqrt{\frac{1}{1+K_R}} \mathbf{h}_k^{\mathrm{N}LOS} \right), \tag{3.29}$$

where $\mathbf{h}_k^{LOS} = 10^{-2}[1,\, e^{j\theta_k},\, e^{j2\theta_k}, \ldots,\, e^{j(N_s-1)\theta_k}]^T$ with $\theta_k = -\pi \sin\phi_k$, $\phi_k \in [-\pi, \pi]$ is randomly generated and the Rician ratio $K_R = 5\mathrm{d}B$, $d_k(=1.5)$ and $m_k(=2.7)$ denote the BS to MS distances and the path loss exponents, respectively, with reference to [13], and $\mathbf{h}_k^{\mathrm{N}LOS}$ is an independent zero-mean complex Gaussian random variable with variance of 10^{-2}. The final results are obtained by averaging over 1000 Monte Carlo simulation runs. To demonstrate the advantages of the proposed methods, the randomization approach, the penalty function method (PenFun), the SDP method, and the optimal performance with perfect CSI [56] are all compared.

Fig. 3.2 shows the performance in terms of the transmit power versus SINR targets (γ) and fixed harvested power threshold $\eta = 10$ dBm with $K = 2$ and $K = 4$, respectively. Here we set $N_s = 4$. As can be observed, the minimum transmit power rises with the increase of the number of MSs. Also in both cases, the randomization approach shows an upper-bound performance compared with the other methods due to randomization. PenFun also performs nearly as the SDP method and is also quite close to the perfect CSI case which demonstrates that the proposed PenFun method not only guarantees a rank-one solution but also yields the global optimal solution. Moreover, the gap between PenFun and the randomization approach is narrowed when increasing the SINR threshold while that between the PenFun method and the perfect CSI case follows a reverse trend.

Next, we compare the performance of the methods mentioned above versus the energy harvesting threshold η with targeted SINR fixed at $\gamma = 10$ dB, $K = 4$ and

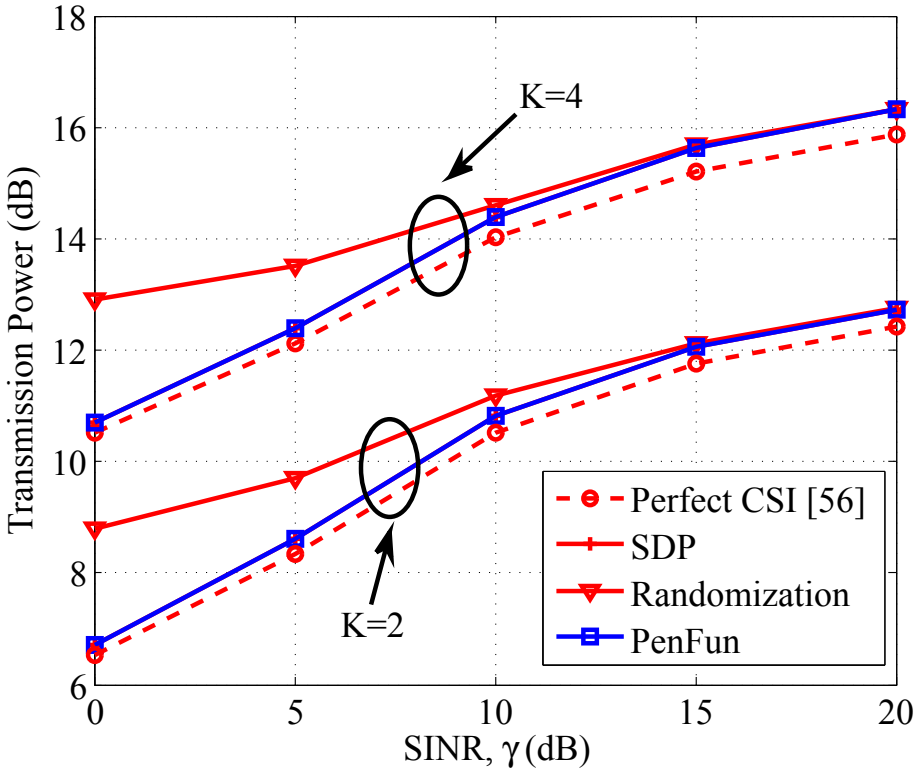


Figure 3.2: The BS transmit power versus the SINR γ.

$N_s = 4$ or 8 in Fig. 3.3. Similarly, in this figure the PenFun method outperforms the randomization approach and shows comparable performance to the SDP method, and the perfect CSI case. In addition, we provide two groups of data with different numbers of transmission antennas N_s to discuss how the number of transmission antennas affects the performance. As can be seen from the figure, increasing the number of antennas at BS can reduce the minimum demanded transmit power to some degree.

Finally, we discuss the impact of the channel uncertainty threshold ε on the performance of the proposed robust beamforming schemes. Here, we compare the performance of the methods mentioned above versus ε with targeted SINR, energy harvesting threshold, number of users, number of transmission antennas fixed at $\gamma =$ 10 dB, $\eta = 10$ dBm, $K = 4$, $N_s = 4$, and the distance $d = 1.5$ or 0.5, as can be seen in Fig. 3.4. In this figure, the transmission power rises with the increase of the channel uncertainty threshold ε. In all the cases, the transmission power for the PenFun method is lower than the one for the randomization approach, and approximate to

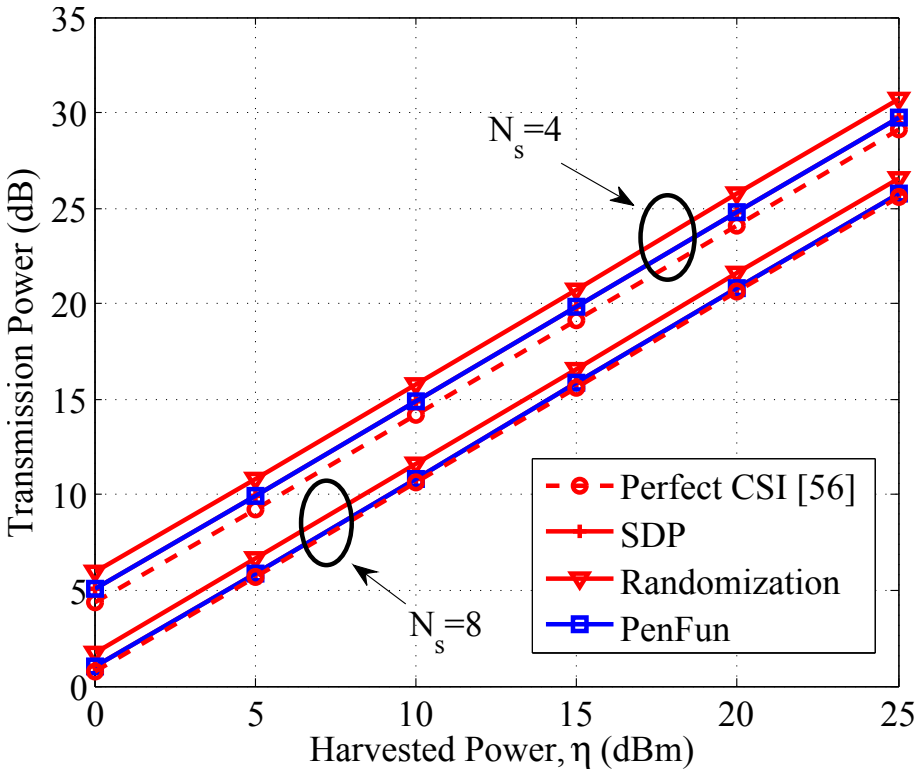


Figure 3.3: Transmission power versus harvested power η.

the one for the SDP method. In the case of small ε, the proposed PenFun method performs close to the perfect CSI case. When the channel estimation error goes up, the performance gaps between the perfect CSI case and the other three schemes gradually increase as expected.

3.6 Summary

In this chapter, a MISO SWIPT broadcast system was investigated where a multiple-antenna transmitter communicated with multiple users each with single antenna while acting simultaneously as an ID and an ER via power splitting. The joint-optimal transmit beamforming and power-splitting ratio with imperfect CSI was obtained using a penalty function based method. In particular, we have shown that the penalty function method yields a more reliable and better solution than the randomization method. Moreover, the penalty function method performs nearly as the SDP method and is also quite close to the perfect CSI case which demonstrates that the proposed PenFun method not only guarantees a rank-one solution but also yields

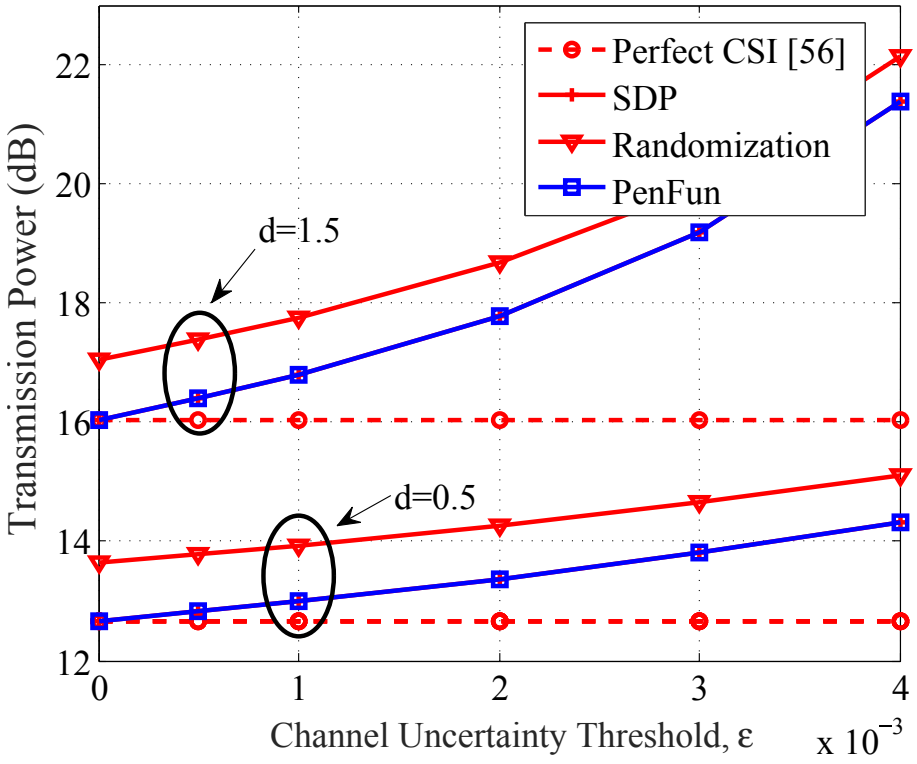


Figure 3.4: Transmission power versus channel uncertainty threshold ε.

the global optimal solution. Note that flat-fading channels have been assumed for resource allocation in SWIPT systems in this thesis, and the extension to the more complex frequency-selective channels would be useful and deserves further investigation. It has been noticed that the frequency-diversity gain can also be exploited to further improve the energy transfer efficiency, by transmitting more power over the sub-band with higher channel gain, for SWIPT over frequency-selective channels. Theoretically, energy transfer efficiency is maximized by transmitting at the frequency with the strongest channel frequency response [53]. However, it is imposed to split the transmission power over multiple strong sub-bands to satisfy various regulations in practice, such as the power spectral density constraint. Therefore, it is necessary to extend the research into wireless power transfer in single and multi-antenna frequency-selective channels under a general SWIPT setup [7, 131].

While we focused on energy efficiency optimization for power splitting based SWIPT MISO multi-user networks here, rate maximization was studied for time switching based SWIPT MIMO relaying in the next chapter.

Chapter 4

Beamforming for SWIPT MIMO Relaying

4.1 Overview

Combining multiple-input multiple-output (MIMO) antenna and relaying is a promising means to enhance both coverage and performance of wireless communications networks and hence has received considerable attentions [132–135]. The optimal transmission policies have been intensively investigated in order to achieve the best performance of the cooperative communication networks. On the other hand, energy harvesting has emerged as an attractive component for relaying and cooperative communication as the battery storage of the relay nodes is usually limited. In consideration of SWIPT, the cooperative techniques of the multi-antenna relay networks require to be revisited to finish information transmission while satisfying the demands on harvested energy. In this chapter, we consider SWIPT for a MIMO relay system. The relay is powered by harvesting energy from the source via time switching (TS) and utilizes the harvested energy to forward the information signal. Our aim is to maximize the rate of the system subject to the power constraints at both the source and relay nodes. In the first scenario in which the source covariance matrix is an identity matrix, we present the joint-optimal solution for relaying and the TS ratio in closed form. An iterative scheme is then proposed for jointly optimizing the source and relaying matrices and the TS ratio.

4.2 Related Work

As a crucial issue for multi-antenna relay networks, the optimal source, relay, and receive matrices have been explored from all kinds of perspectives. In [132], the optimal structure of the relay processing matrix to maximize the rate assuming unitary source precoding was presented. Joint source and relay optimization was considered in [133, 134]. Recently, more complex applications such as robust beamforming design with imperfect CSI were studied [135]. Moreover, research has been carried out to study SWIPT for multi-antenna relay networks [136–138]. For complexity reasons, preference was given to the time-switching (TS) mechanism over power splitting for energy harvesting, e.g., [137, 138]. Unfortunately, the existing approaches failed to provide a joint energy transfer ratio and precoding matrices design for a generic MIMO relay system. For example, [136, 137] depended on either semi-definite relaxation (SDR) and existing solvers or iterative approaches while in [138] a closed form solution was given for a MISO relay system rather than general MIMO network. Closed-form solutions as well as the structures of the source covariance and relay beamforming matrices are not well understood. To this end, we consider the rate maximization problem for the SWIPT MIMO relay system. The fixed source covariance matrix case and the joint source, relay, and TS ratio optimization case are both investigated. Unlike the existing attempts which rely on SDR, we give the structures of the optimal relay beamforming matrix and the source covariance matrix and propose a closed-form solution and an iterative solution for the two cases, respectively.

4.3 System Model

We consider a two-hop MIMO relay network with an energy harvesting relay node and assume that the source, relay and destination nodes are all equipped with multiple antennas, as shown in Fig. 4.1. The numbers of antennas are M, L, N, respectively. The relay harvests energy from the source and uses it to forward the information. Here we assume that the direct link between the source and destination is negligible with perfect CSI known at all nodes.

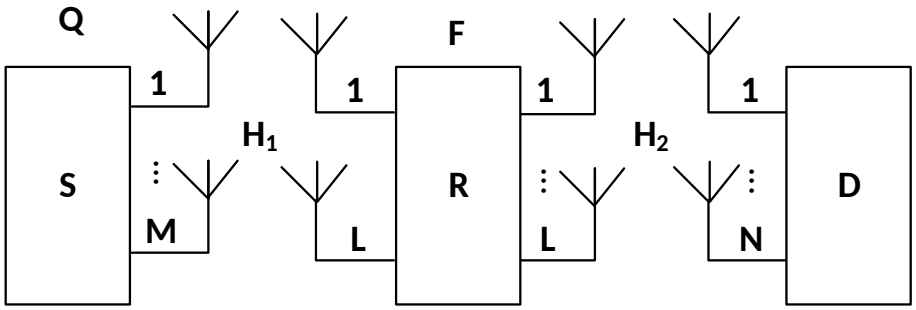


Figure 4.1: A SWIPT enabled relay system.

The TS-based relaying involves three phases, as shown in Fig. 4.2, with T being the block length, and ε denoting the TS ratio. In the first phase, the channel and source covariance matrices are defined as $\widetilde{\mathbf{H}}_1$ and $\widetilde{\mathbf{Q}}$, respectively. $\widetilde{\mathbf{s}}$ is the source symbol vector. In the information transmit phases, we will use $\mathbf{s}$ to denote the source symbol vector, P to denote maximum transmit power, $\mathbf{H}_1$ and $\mathbf{H}_2$ to represent the channel matrices between the source and relay, and the relay and the destination, respectively, $\mathbf{Q}$ to denote the source covariance matrix and $\mathbf{F}$ the relay beamforming matrix. We will also consider the additive white Gaussian noises (AWGNs), $\mathbf{n}_1$ and $\mathbf{n}_2$, at the relay and destination nodes with variances σ_1^2 and σ_2^2, respectively. In this chapter, $T = 1$ is assumed. The superscript H is the Hermitian operator and $\mathbf{I}$ is an identity matrix.

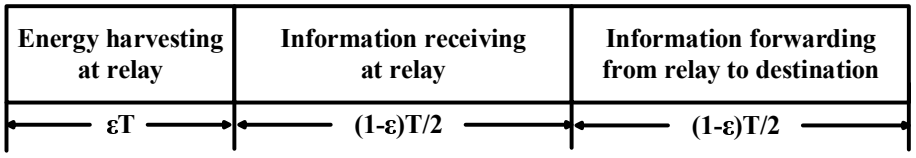


Figure 4.2: The framework of the proposed TS relaying.

The harvested power at the relay can be expressed as

$$\mathrm{tr}(\mathbf{y}_e\mathbf{y}_e^H) = \mathrm{tr}(\widetilde{\mathbf{H}}_1\widetilde{\mathbf{Q}}\widetilde{\mathbf{H}}_1^H + \sigma_1^2\mathbf{I}_D), \tag{4.1}$$

where $\mathbf{y}_e = \widetilde{\mathbf{H}}_1\widetilde{\mathbf{s}} + \mathbf{n}_1$ is the received signal at the relay in the energy harvesting phase. Then in the information transmission phase, the received signal at the relay can be written as

$$\mathbf{y}_r = \mathbf{H}_1\mathbf{s} + \mathbf{n}_1. \tag{4.2}$$

In the last time phase, the received signal at the destination is

$$\mathbf{y}_d = \mathbf{H}_2\mathbf{F}\mathbf{H}_1\mathbf{s} + \mathbf{H}_2\mathbf{F}\mathbf{n}_1 + \mathbf{n}_2. \tag{4.3}$$

4.4 Relay and TS Ratio Only Design

Here, we fix the source covariance matrix in the information transmit phase as $\mathbf{Q} = \mathbb{E}(\mathbf{s}\mathbf{s}^H) = \frac{P}{D}\mathbf{I}_D$, where D ($D \leq \min(M,L,N)$) is the number of data streams.

To proceed, we firstly consider the direct link between the source node and the destination node without relaying. In this scenario, the received signal can be written as

$$\mathbf{y}_0 = \mathbf{H}_0\mathbf{s} + \mathbf{n}_0, \tag{4.4}$$

where $\mathbf{H}_0$ denotes the channel matrix of the direct link. The noise is modeled with complex circular white Gaussian, i.e., $\mathbf{n}_0 \to \mathcal{N}(0, \sigma_0^2)$. The achievable rate, which is refereed to as the instantaneous capacity corresponding to a packet duration, can then be written as [139]

$$C_0 = \frac{1-\varepsilon}{2}\log_2 \det\left(\mathbf{I}_D + \rho_0 \mathbf{H}_0^H \mathbf{H}_0\right), \tag{4.5}$$

where det represents the determinant of a matrix, and $\rho_0 \triangleq \frac{P}{D\sigma_0^2}$ is the normalized SNR. And the term $\frac{1-\varepsilon}{2}$ results from the time switching EH strategy.

Comparing the two scenarios with and without relaying, the equivalent covariance matrix of the equivalent total noise term (with interference) in (4.2) for the relay scenario refereed to as $\mathbf{R}$ satisfies

$$\mathbf{R} = \sigma_2^2\left(\mathbf{I}_D + \frac{\sigma_1^2}{\sigma_2^2}\mathbf{H}_2\mathbf{F}\mathbf{F}^H\mathbf{H}_2^H\right). \tag{4.6}$$

The noise whitening matrix can then be defined as $\mathbf{R}^{-1/2}$, and the equivalent channel matrix can be written as $\widetilde{\mathbf{H}} \triangleq \mathbf{R}^{-1/2}\mathbf{H}_2\mathbf{F}\mathbf{H}_1$.

Substituting the noise whitening matrix $\mathbf{R}^{-1/2}$ and equivalent channel matrix $\widetilde{\mathbf{H}}$ into (4.5), the achievable rate can be written as

$$C = \frac{1-\varepsilon}{2}\log_2\det\left(\mathbf{I}_D + \frac{P}{D}\widetilde{\mathbf{H}}^H\widetilde{\mathbf{H}}\right)$$
$$= \frac{1-\varepsilon}{2}\log_2\det\left(\mathbf{I}_D + \frac{P}{D}(\mathbf{R}^{-1/2}\mathbf{H}_2\mathbf{F}\mathbf{H}_1)^H(\mathbf{R}^{-1/2}\mathbf{H}_2\mathbf{F}\mathbf{H}_1)\right)$$
$$= \frac{1-\varepsilon}{2}\log_2\det\left(\mathbf{I}_D + \frac{P}{D}\mathbf{H}_1^H(\mathbf{H}_2\mathbf{F})^H\mathbf{R}^{-1}(\mathbf{H}_2\mathbf{F})\mathbf{H}_1\right) \quad (4.7)$$

Substituting the expression of $\mathbf{R}$ in (4.6) into (4.7), the achievable rate can be further reformulated into [132]

$$C = \frac{1-\varepsilon}{2}\log_2\det\left(\mathbf{I}_D + \frac{P}{D\sigma_1^2}\mathbf{H}_1^H\underbrace{(\frac{\sigma_1}{\sigma_2}\mathbf{H}_2\mathbf{F})^H\left(\mathbf{I}_D + (\frac{\sigma_1}{\sigma_2}\mathbf{H}_2\mathbf{F})(\frac{\sigma_1}{\sigma_2}\mathbf{H}_2\mathbf{F})^H\right)^{-1}(\frac{\sigma_1}{\sigma_2}\mathbf{H}_2\mathbf{F})}\mathbf{H}_1\right)$$
$$= \frac{1-\varepsilon}{2}\log_2\det\left(\mathbf{I}_D + \frac{P}{D\sigma_1^2}\mathbf{H}_1^H\underbrace{\left(\mathbf{I}_D - \left(\mathbf{I}_D + (\frac{\sigma_1}{\sigma_2}\mathbf{H}_2\mathbf{F})^H(\frac{\sigma_1}{\sigma_2}\mathbf{H}_2\mathbf{F})\right)^{-1}\right)}\mathbf{H}_1\right). \quad (4.8)$$

Meanwhile, the equivalence of the underbraced terms can be verified utilizing straightforward proof or matrix inverse lemma. Moreover, we let $\mathbf{S} = \mathbf{I}_D + \frac{\sigma_1^2}{\sigma_2^2}\mathbf{F}^H\mathbf{H}_2^H\mathbf{H}_2\mathbf{F}$, and the SNR at the relay $\rho_1 \triangleq \frac{P}{D\sigma_1^2}$, (4.8) can be rewritten as [132]

$$C = \frac{1-\varepsilon}{2}\log_2\det\left(\mathbf{I}_D + \rho_1\mathbf{H}_1\mathbf{H}_1^H - \rho_1\mathbf{H}_1\mathbf{H}_1^H\mathbf{S}^{-1}\right). \quad (4.9)$$

Therein, we utilize the important property that $\det(\mathbf{I}+\mathbf{AB}) = \det(\mathbf{I}+\mathbf{BA})$ for any complex conjugate symmetric matrices $\mathbf{A}$ and $\mathbf{B}$.

The transmitted signal at the relay

$$\mathbf{x}_r = \mathbf{F}\mathbf{H}_1\mathbf{s} + \mathbf{F}\mathbf{n}_1 \quad (4.10)$$

will have to satisfy the harvested power constraint

$$\frac{1-\varepsilon}{2}\sigma_1^2\mathrm{tr}(\mathbf{F}(\mathbf{I}_D + \rho_1\mathbf{H}_1\mathbf{H}_1^H)\mathbf{F}^H) \le \varepsilon\eta\,\mathrm{tr}(\widetilde{\mathbf{H}}_1\widetilde{\mathbf{Q}}\widetilde{\mathbf{H}}_1^H + \sigma_1^2\mathbf{I}_D) \quad (4.11)$$

where $0 \le \eta \le 1$ is the energy conversion efficiency.

Note that in the energy harvesting phase, the objective is to maximize the harvested power subject to the transmit power constraint, solution to which is given in [10] as described below.

Lemma 4.1 *Let the singular value decomposition (SVD) of* $\widetilde{\mathbf{H}}_1$ *be* $\widetilde{\mathbf{H}}_1 = \mathbf{U}\Gamma^{\frac{1}{2}}\mathbf{V}^H$ *where* $\mathbf{U}$ *and* $\mathbf{V}$ *are unitary while the diagonal elements of* Γ*,* $g_1, g_2, ..., g_D$*, are arranged in a descending order.* v_1 *is defined as the first column of* $\mathbf{V}$*.* P_0 *denotes the transmit power threshold in the energy harvesting phase. The optimal solution to the optimization problem*

$$\max_{\widetilde{\mathbf{Q}}} \quad \text{tr}(\widetilde{\mathbf{H}}_1\widetilde{\mathbf{Q}}\widetilde{\mathbf{H}}_1^H + \sigma_1^2\mathbf{I}_D) \tag{4.12}$$

$$\textit{s.t.} \qquad \text{tr}(\widetilde{\mathbf{Q}}) \leq P_0, \widetilde{\mathbf{Q}} \succeq 0. \tag{4.13}$$

is $\widetilde{\mathbf{Q}} = P_0 v_1 v_1^H$ *and the corresponding maximum harvested power is given by* $g_1 P_0 + \sigma_1^2 D$*.*

Proof 4.1 *See [10, Proposition 2.1].*

Let us now define $\mathbf{G} = \frac{\sigma_1}{\sigma_2}\mathbf{F}$ and formulate the following optimization problem

$$\max_{\mathbf{G},\varepsilon} \; C \;\text{ s.t. }\; \frac{1-\varepsilon}{2}\text{tr}(\mathbf{G}(\mathbf{I}_D + \rho_1\mathbf{H}_1\mathbf{H}_1^H)\mathbf{G}^H) \leq \frac{\varepsilon\eta}{\sigma_2^2}(g_1P_0 + \sigma_1^2 D). \tag{4.14}$$

Note that for fixed ε, problem (4.14) becomes technically identical to the one considered in [132]. Moreover, it can be proved that the presence of ε does not change the optimal structure of the relay processing matrix. Hence, we have the relay processing matrix given by

$$\mathbf{F} = \mathbf{V}_2\mathbf{\Lambda}_F\mathbf{U}_1^H, \tag{4.15}$$

where $\mathbf{\Lambda}_F$ denotes a diagonal matrix. $\mathbf{V}_2$ and $\mathbf{U}_1$ come from the SVDs of the matrices:

$$\mathbf{H}_1 = \mathbf{U}_1\mathbf{\Sigma}_1\mathbf{V}_1^H, \tag{4.16}$$

$$\mathbf{H}_2 = \mathbf{U}_2\mathbf{\Sigma}_2\mathbf{V}_2^H. \tag{4.17}$$

Now, let $\mathbf{G} = \mathbf{V}_2\mathbf{X}^{\frac{1}{2}}(\mathbf{I} + \rho_1\mathbf{\Lambda}_1)^{-\frac{1}{2}}\mathbf{U}_1^H$, where $\mathbf{X}$ is a diagonal matrix with $\mathbf{X} = \text{diag}(x_1, x_2, \ldots, x_D)$. In addition, we let $\mathbf{\Lambda_1} = \mathbf{\Sigma}_1\mathbf{\Sigma}_1^H$ and $\mathbf{\Lambda_2} = \mathbf{\Sigma}_2^H\mathbf{\Sigma}_2$ with the

diagonal vectors $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_D]$ and $\boldsymbol{\beta} = [\beta_1, \ldots, \beta_D]$, respectively. The optimization problem (4.14) can then be rewritten as

$$\max_{\{x_k \geq 0\}, 0 \leq \varepsilon \leq 1} f(\{x_k\}, \varepsilon) \text{ s.t.} \tag{4.18a}$$

$$g(\{x_k\}, \varepsilon) \triangleq \frac{\varepsilon \eta}{\sigma_2^2}(g_1 P_0 + \sigma_1^2 D) - \frac{1-\varepsilon}{2} \sum_{k=1}^{D} x_k \geq 0, \tag{4.18b}$$

in which $f(\{x_k\}, \varepsilon) \triangleq \frac{1-\varepsilon}{2}[\sum_{k=1}^{D} \log_2(1+\rho_1 \alpha_k) + \sum_{k=1}^{D} \log_2(\frac{1+\beta_k x_k}{1+\rho_1 \alpha_k + \beta_k x_k})]$. Considering the Lagrangian, we have

$$\min_{\substack{\{x_k \geq 0\}, 0 \leq \varepsilon \leq 1 \\ \nu \geq 0, \{\lambda_k \geq 0\}}} \mathscr{L} \triangleq f(\{x_k\}, \varepsilon) + \nu g(\{x_k\}, \varepsilon) + \sum_{k=1}^{D} \lambda_k x_k. \tag{4.19}$$

Based on the Karush-Kuhn-Tucker (KKT) conditions, we have

$$\nu g(\{x_k\}, \varepsilon) = 0, \tag{4.20a}$$

$$\lambda_k x_k = 0, \forall k, \tag{4.20b}$$

$$\nabla_{x_k} \mathscr{L} = 0, \forall k, \tag{4.20c}$$

$$\nabla_{\varepsilon} \mathscr{L} = 0. \tag{4.20d}$$

Since $\lambda_k \geq 0$, $k = 1, ..., D$, using (4.20c) we can show that

$$\nu \geq \frac{1}{\ln 2} \frac{\left(\frac{\rho_1 \alpha_k}{\beta_k}\right)}{\left(x_k + \frac{1}{\beta_k}\right)\left(x_k + \frac{\rho_1 \alpha_k + 1}{\beta_k}\right)}, \forall k. \tag{4.21}$$

Let $h_k(x_k) = \frac{1}{\ln 2} \frac{\frac{\rho_1 \alpha_k}{\beta_k}}{(x_k + \frac{1}{\beta_k})(x_k + \frac{\rho_1 \alpha_k + 1}{\beta_k})}$. It is obvious that h_k decreases when x_k rises from 0 to ∞, i.e., $h_k(0) = \frac{1}{\ln 2} \frac{\rho_1 \alpha_k \beta_k}{1+\rho_1 \alpha_k}$, $h_k(\infty) \to 0$. Then according to (4.20b), we know that

$$x_k \left[\nu - \frac{1}{\ln 2} \frac{\frac{\rho_1 \alpha_k}{\beta_k}}{(x_k + \frac{1}{\beta_k})(x_k + \frac{\rho_1 \alpha_k + 1}{\beta_k})} \right] = 0, \forall k. \tag{4.22}$$

If $\nu \geq \frac{1}{\ln 2} \frac{\rho_1 \alpha_k \beta_k}{1+\rho_1 \alpha_k}$, we know that $x_k = 0$. Otherwise, if $0 < \nu < \frac{1}{\ln 2} \frac{\rho_1 \alpha_k \beta_k}{1+\rho_1 \alpha_k}$, then we will have $x_k > 0$ and $\nu = \frac{1}{\ln 2} \frac{\frac{\rho_1 \alpha_k}{\beta_k}}{(x_k + \frac{1}{\beta_k})(x_k + \frac{\rho_1 \alpha_k + 1}{\beta_k})}$. Then according to [132], the optimal x_k can be written as

$$x_k = \frac{1}{2\beta_k} \left(\sqrt{\rho_1^2 \alpha_k^2 + \frac{4}{\ln 2} \rho_1 \alpha_k \beta_k \mu} - \rho_1 \alpha_k - 2 \right)^+, \tag{4.23}$$

where $(a)^+ = \max\{0, a\}$ and $\mu = \frac{1}{v}$ can be obtained by substituting (4.23) into (4.20d) such that

$$l(\mu) \triangleq -\frac{1}{2}\sum_{k=1}^{D}\log_2\left[(1+\rho_1\alpha_k)\left(\frac{1+\beta_k x_k}{1+\rho_1\alpha_k+\beta_k x_k}\right)\right] + \frac{1}{\mu}\left[\frac{\eta}{\sigma_2^2}(g_1P_0+\sigma_1^2 D)+\frac{1}{2}\sum_{k=1}^{D}x_k\right] = 0. \tag{4.24}$$

Obviously, $l(\mu)$ decreases when $\mu \geq \max_k \ln 2\frac{1+\rho_1\alpha_k}{\rho_1\alpha_k\beta_k}$. Moreover, when $\mu \in [\min_k \ln 2\frac{1+\rho_1\alpha_k}{\rho_1\alpha_k\beta_k}, \max_k \ln 2\frac{1+\rho_1\alpha_k}{\rho_1\alpha_k\beta_k}]$, x_k either maintain at 0 or increases with μ which makes $l(\mu)$ decreases within this interval. To be exact, we have

$$l(\infty) \to -\frac{1}{2}\sum_{k=1}^{D}\log_2(1+\rho_1\alpha_k) < 0, \tag{4.25}$$

and

$$l\left(\min_k \ln 2\frac{1+\rho_1\alpha_k}{\rho_1\alpha_k\beta_k}\right) = \left(\max_k \frac{1}{\ln 2}\frac{\rho_1\alpha_k\beta_k}{1+\rho_1\alpha_k}\right)\frac{\eta}{\sigma_2^2}(g_1P_0+\sigma_1^2 D) > 0. \tag{4.26}$$

In contrast, when $\mu \in (0, \min_k \ln 2\frac{1+\rho_1\alpha_k}{\rho_1\alpha_k\beta_k}]$, $x_k = 0, \forall k$ and

$$l(u) = \frac{1}{\mu}\frac{\eta}{\sigma_2^2}(g_1P_0+\sigma_1^2 D) > 0. \tag{4.27}$$

As a consequence, we can always find an optimal value $\mu^* \in (\min_k \ln 2\frac{1+\rho_1\alpha_k}{\rho_1\alpha_k\beta_k}, \infty)$ which makes $l(\mu^*) = 0$ and this can be done by root-finding approaches such as bisection. When μ^* and $x_k, \forall k$, are known, the optimal TS ratio can easily be calculated according to (4.20a), which is given by

$$\varepsilon^* = \frac{\sum_{k=1}^{D} x_k^*}{\frac{2\eta}{\sigma_2^2}(g_1P_0+\sigma_1^2 D)+\sum_{k=1}^{D} x_k^*}. \tag{4.28}$$

Regarding the complexity of the algorithm, the computations mainly result from the singular value decompositions of channel matrices. The complexity of root-finding is rather low.

4.5 Joint Source, Relay and TS Ratio Design

In this section, we consider a more general but challenging scenario with any available $\mathbf{Q}$. Based on (4.7), the achievable rate in this case is given by

$$\begin{aligned} C &= \frac{1-\varepsilon}{2}\log_2\det\left(\mathbf{I}_D+\widetilde{\mathbf{H}}^H\widetilde{\mathbf{H}}\mathbf{Q}\right) \\ &= \frac{1-\varepsilon}{2}\log_2\det\left(\mathbf{I}_D+\mathbf{H}_1^H(\mathbf{H}_2\mathbf{F})^H\mathbf{R}^{-1}(\mathbf{H}_2\mathbf{F})\mathbf{H}_1\mathbf{Q}\right) \\ &= \frac{1-\varepsilon}{2}\log_2\det\left(\mathbf{I}_D+\mathbf{R}^{-1}(\mathbf{H}_2\mathbf{F})\mathbf{H}_1\mathbf{Q}\mathbf{H}_1^H(\mathbf{H}_2\mathbf{F})^H\right) \\ &= \frac{1-\varepsilon}{2}\log_2\det\left(\mathbf{I}_D+\frac{\mathbf{H}_2\mathbf{F}\mathbf{H}_1\mathbf{Q}\mathbf{H}_1^H\mathbf{F}^H\mathbf{H}_2^H}{\sigma_2^2\mathbf{I}_D+\sigma_1^2\mathbf{H}_2\mathbf{F}\mathbf{F}^H\mathbf{H}_2^H}\right). \end{aligned} \tag{4.29}$$

Note that to derive the results in (4.29), we utilize the expression of $\mathbf{R}$ in (4.6) and also the property that $\det(\mathbf{I}+\mathbf{AB})=\det(\mathbf{I}+\mathbf{BA})$ for any complex conjugate symmetric matrices $\mathbf{A}$ and $\mathbf{B}$. And it is obvious that when $\mathbf{Q}$ is fixed at $\mathbf{Q}=\frac{P}{D}$, the expression of achievable rate becomes the same as that in the previous section, i.e. (4.7).

Then the optimization problem of interest becomes

$$\begin{aligned} \max_{\substack{\mathbf{F},\varepsilon \\ \mathrm{tr}(\mathbf{Q})\le P}} \; C \;\; \text{s.t.} \;\; & \frac{1-\varepsilon}{2}\mathrm{tr}(\sigma_1^2\mathbf{F}\mathbf{F}^H+\mathbf{F}\mathbf{H}_1\mathbf{Q}\mathbf{H}_1^H\mathbf{F}^H) \\ & \le \varepsilon\eta\,\mathrm{tr}(\widetilde{\mathbf{H}}_1\widetilde{\mathbf{Q}}\widetilde{\mathbf{H}}_1^H+\sigma_1^2\mathbf{I}_D). \end{aligned} \tag{4.30}$$

By introducing an equivalent channel $\widehat{\mathbf{H}}_1=\mathbf{H}_1\mathbf{Q}^{\frac{1}{2}}$, the optimization problem becomes similar to the previous fixed source covariance matrix case. Therefore, we have $\widehat{\mathbf{F}}=\mathbf{V}_2\widehat{\mathbf{\Sigma}}_F\widehat{\mathbf{U}}_1^H$ where $\widehat{\mathbf{\Sigma}}_F$ is diagonal, and $\widehat{\mathbf{U}}_1$ and $\mathbf{V}_2$ come from the SVDs of $\widehat{\mathbf{H}}_1=\widehat{\mathbf{U}}_1\widehat{\mathbf{\Sigma}}_1\widehat{\mathbf{V}}_1^H$, and $\mathbf{H}_2$ given in (4.17).

As can be observed, both the objective function and the energy harvesting constraint have nothing to do with $\widehat{\mathbf{U}}_1$ which indicates that any available $\widehat{\mathbf{H}}_1$ with the same $\widehat{\mathbf{\Sigma}}_1$ acts equally in terms of the rate and the energy harvesting constraint. That is to say, the optimal $\mathbf{Q}$ must require the least transmit power. Considering the fact that the presence of the TS ratio ε will not change the structures of the source covariance and relay processing matrices, we provide the optimal structures of the source covariance and relay processing matrices below.

Lemma 4.2 *The optimal solution of the optimization problem (4.30) has the following structures*

$$\mathbf{F} = \mathbf{V}_2 \mathbf{\Sigma}_F \mathbf{U}_1^H, \tag{4.31}$$

$$\mathbf{Q} = \mathbf{V}_1 \mathbf{\Lambda}_Q \mathbf{V}_1^H, \tag{4.32}$$

where $\mathbf{\Sigma}_F, \mathbf{\Lambda}_Q$ *are diagonal matrices, and the unitary matrices* $\mathbf{U}_1, \mathbf{V}_1, \mathbf{U}_2, \mathbf{V}_2$ *have been defined in (4.16) and (4.17).*

Proof 4.2 *See [133, Theorem 1].*

Then we let $\mathbf{\Lambda_Q} = \text{diag}(q_1, q_2, \ldots, q_D)$, and $\mathbf{\Lambda_F} = \mathbf{\Sigma}_F^2 = \text{diag}(f_1, f_2, \ldots, f_D)$. Substituting (4.31) and (4.32) into (4.30) and introducing a set of new variables $d_k = f_k(\alpha_k q_k + \boldsymbol{\sigma}_1^2), \forall k$, the optimization problem (4.30) can be rewritten as

$$\max_{0 \le \varepsilon \le 1, \{d_k\}, \{q_k\}} \quad \tilde{f}(\varepsilon, \{d_k\}, \{q_k\}) \tag{4.33a}$$

$$\text{s.t.} \quad \sum_{k=1}^{D} q_k \le P, \tag{4.33b}$$

$$\tilde{g}(\varepsilon, \{d_k\}, \{q_k\}) \ge 0, \tag{4.33c}$$

where we have defined

$$\tilde{f}(\varepsilon, \{d_k\}, \{q_k\}) \triangleq \frac{1-\varepsilon}{2} \sum_{k=1}^{D} \log_2 \frac{\left(1 + \frac{\alpha_k}{\sigma_1^2} q_k\right)\left(1 + \frac{\beta_k}{\sigma_2^2} d_k\right)}{1 + \frac{\alpha_k}{\sigma_1^2} q_k + \frac{\beta_k}{\sigma_2^2} d_k}, \tag{4.34}$$

$$\tilde{g}(\varepsilon, \{d_k\}, \{q_k\}) \triangleq \varepsilon \eta (g_1 P_0 + \sigma_1^2 D) - \frac{1-\varepsilon}{2} \sum_{k=1}^{D} d_k. \tag{4.35}$$

Note that (4.33) involves only scalar variables in contrast to matrix variables in (4.30). But the problem is still non-convex and a closed-form solution is difficult to obtain. In the following, we develop an alternating optimization based iterative algorithm which can be proved to converge at least to a local optimal solution. Since the subproblems are convex, close-form solutions are derived by solving Lagrangian dual problems. To proceed, we let $\boldsymbol{q} = [q_1, q_2, \ldots, q_D]^T$, and $\boldsymbol{d} = [d_1, d_2, \ldots, d_D]^T$. When either $\boldsymbol{q}$ or $\boldsymbol{d}$ is fixed, the corresponding problem to update the other one becomes equivalent to the relay and TS ratio only design problem in the previous

section. As mentioned previously, the complexity falls into the computations of SVD. Finally, an iterative procedure can be designed by optimizing $\boldsymbol{q}$ and $\boldsymbol{d}$ alternately.

4.5.1 Optimization with Fixed $\boldsymbol{q}$

We first fix $\boldsymbol{q}$ and search for the optimal $\boldsymbol{d}$ and ε with the given $\boldsymbol{q}$. Considering the Lagrangian of the problem, we have the following dual problem:

$$\max_{\substack{0\le\varepsilon\le 1,\{d_k\ge 0\}\\ \nu_1\ge 0,\{\lambda_k\ge 0\}}} \mathscr{L} \triangleq \tilde{f}(\varepsilon,\{d_k\})+\nu_1\tilde{g}(\varepsilon,\{d_k\})+\sum_{k=1}^{D}\lambda_k d_k. \tag{4.36}$$

Based on the KKT conditions, we have

$$\nu_1\tilde{g}(\varepsilon,\{d_k\})=0, \tag{4.37a}$$

$$\lambda_k d_k = 0, \forall k, \tag{4.37b}$$

$$\nabla_{d_k}\mathscr{L}=0, \forall k, \tag{4.37c}$$

$$\nabla_{\varepsilon}\mathscr{L}=0. \tag{4.37d}$$

Comparing with the fixed source covariance matrix case, we notice that d_k here is equivalent to $\sigma_2^2 x_k$. As a result, we have

$$d_k=\frac{\sigma_2^2}{2\beta_k}\left(\sqrt{\left(\frac{q_k}{\sigma_1^2}\alpha_k\right)^2+\frac{4q_k\alpha_k\beta_k\mu_1}{\sigma_1^2\sigma_2^2\ln 2}}-\frac{q_k}{\sigma_1^2}\alpha_k-2\right)^+. \tag{4.38}$$

Meanwhile, $\mu_1=\frac{1}{\nu_1}$ is decided by (4.37d). Substituting (4.38) into (4.37d), we have

$$\begin{aligned}\widetilde{l}(\mu_1)\triangleq &-\frac{1}{2}\left[\sum_{k=1}^{D}\log_2\frac{\left(1+\frac{\alpha_k}{\sigma_1^2}q_k\right)\left(1+\frac{\beta_k}{\sigma_2^2}d_k\right)}{1+\frac{\alpha_k}{\sigma_1^2}q_k+\frac{\beta_k}{\sigma_2^2}d_k}\right]\\ &+\frac{1}{\mu_1}\left[\eta(g_1P_0+\sigma_1^2D)+\frac{1}{2}\sum_{k=1}^{D}d_k\right]=0.\end{aligned} \tag{4.39}$$

Using (4.39), we can obtain the optimal μ_1 by a bisection search and then use it calculate $\boldsymbol{d}$. Then according to (4.37a), we have

$$\varepsilon^*=\frac{\sum_{k=1}^{D}d_k}{2\eta(g_1P_0+\sigma_1^2D)+\sum_{k=1}^{D}d_k}. \tag{4.40}$$

4.5.2 Optimization with Fixed $\boldsymbol{d}$ and ε

Since the transmit power constraint at relay (4.33c) is independent on $\mathbf{q}$, and thus can be satisfied with the obtained $\boldsymbol{d}$ and ε when fixing $\boldsymbol{q}$, here we do not need to consider it any longer. Referring to the Lagrangian of (4.33), we have

$$\max_{\substack{\{q_k\geq 0\},\\ \nu_2\geq 0,\{\lambda_k\geq 0\}}} \mathscr{L} \triangleq \tilde{f}(\{q_k\}) + \nu_2\left(P - \sum_{k=1}^{D} q_k\right) + \sum_{k=1}^{D}\lambda_k q_k. \tag{4.41}$$

Deriving the relevant KKT conditions again, we obtain

$$\nu_2\left(P - \sum_{k=1}^{D} q_k\right) = 0, \tag{4.42a}$$

$$\lambda_k q_k = 0, \forall k, \tag{4.42b}$$

$$\nabla_{q_k}\mathscr{L} = 0, \forall k. \tag{4.42c}$$

Then according to (4.42c), we have

$$\frac{1-\varepsilon}{2}\frac{\alpha_k}{\ln 2\sigma_1^2}\left(\frac{1}{1+\frac{\alpha_k}{\sigma_1^2}q_k} - \frac{1}{1+\frac{\alpha_k}{\sigma_1^2}q_k+\frac{\beta_k}{\sigma_2^2}d_k}\right) - \nu_2 + \lambda_k = 0. \tag{4.43}$$

Thus from (4.43), we have

$$q_k = \frac{\sigma_1^2}{2\alpha_k}\left[\sqrt{\left(\frac{\beta_k}{\sigma_2^2}d_k\right)^2 + \frac{2(1-\varepsilon)\beta_k d_k \alpha_k \mu_2}{\sigma_1^2\sigma_2^2\ln 2}} - \frac{\beta_k}{\sigma_2^2}d_k - 2\right]^+, \tag{4.44}$$

where $\mu_2 = \frac{1}{\nu_2}$. Substituting (4.44) into (4.42a), the optimal ν_2 can easily be derived through root finding methods.

4.5.3 Iterative Optimization

The iteration to solve (4.30) is given in Algorithm 1. Meanwhile ζ is the iteration terminating threshold.

4.6 Simulation Results

This section investigates the performance of the proposed schemes for the MIMO relay system. The results of the naive amplify-and-forward (NAF) algorithm are also provided for comparison. In the NAF algorithm, we use the ε derived with

Algorithm 2 Iteration Framework for TS Relaying

1. **Initialization: Let q satisfying (4.33b)**
2. **Calculate optimal d and ε with fixed q using (4.38) and (4.40)**
3. **Re-optimize q with the obtained d and ε using (4.44)**
4. **Return to Step 2 until convergence, i.e. $|q^* - q| \leq \zeta * |q|$**

the uniform source precoding scheme and let $\mathbf{Q} = \frac{P}{D}\mathbf{I}$ and $\mathbf{F} = \sqrt{\chi}\mathbf{I}$ where χ is the scalar that makes the constraint (4.11) satisfied. Here we assume that $N = M = L$ and $\zeta = 10^{-3}$. Both of $\mathbf{H_1}$ and $\mathbf{H_2}$ are modeled as flat Rician fading channels with a series of independent zero-mean complex Gaussian random variables with variance of -10dB.

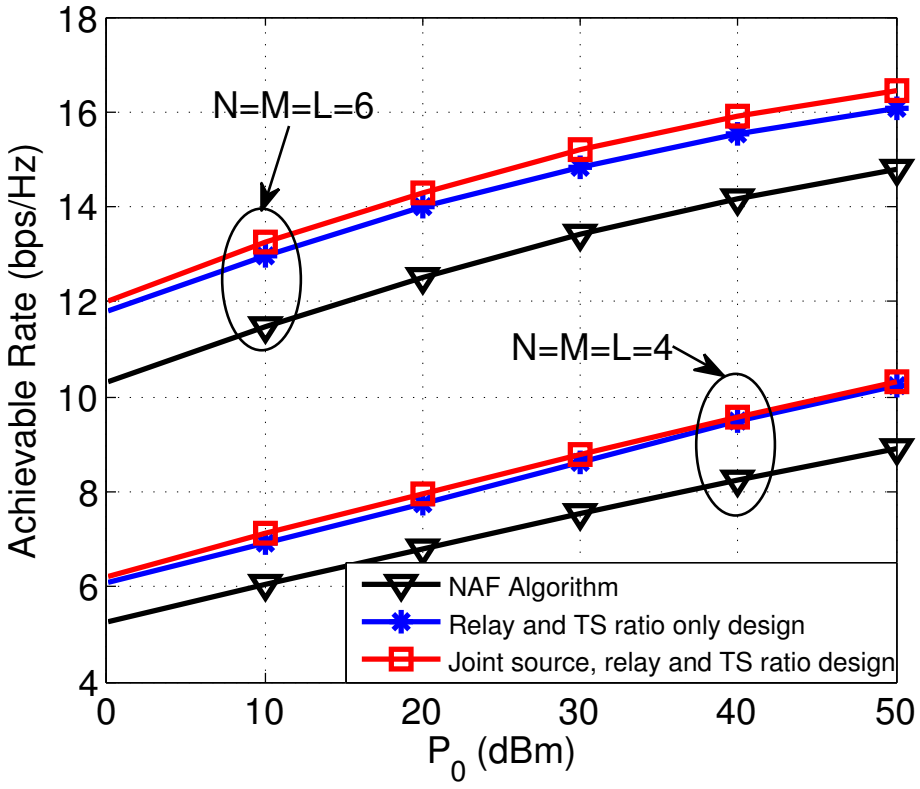


Figure 4.3: Rate results against different P_0.

Fig. 4.3 plots the achievable rate of the proposed TS relaying schemes against various P_0 with $P = 1$. The considered values of P_0 range from 0 dBm to 50 dBm. The numbers of antennas are all set to be 4 and 6, respectively. As is expected, the NAF algorithm falls far behind the proposed schemes. The joint source, relay and

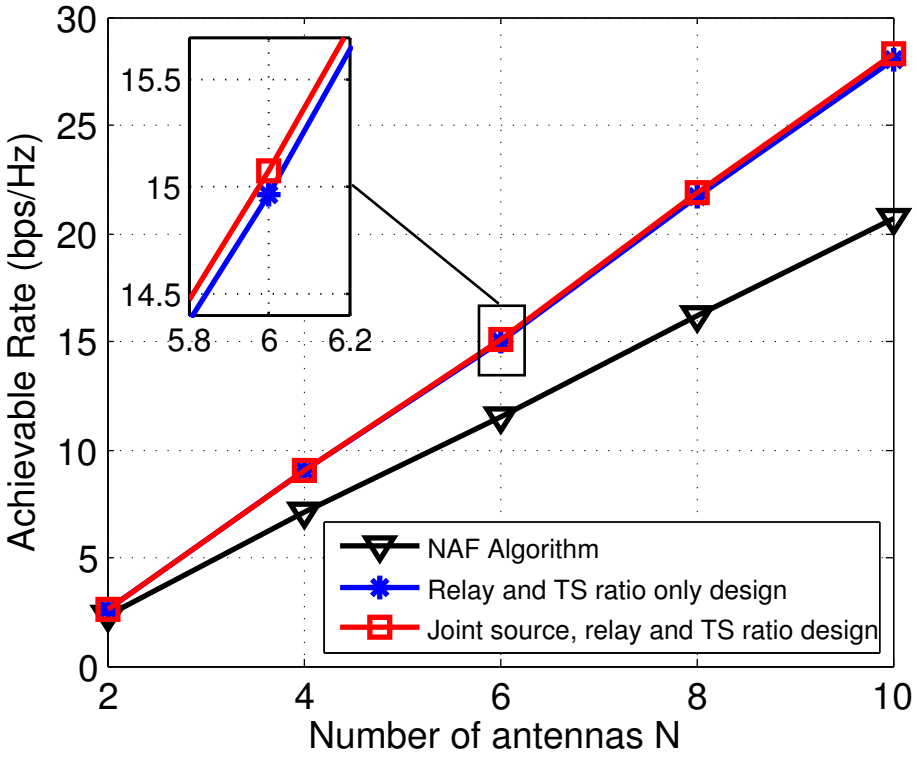


Figure 4.4: Rate results against the number of antennas N.

TS ratio optimization outperforms the relay and TS ratio only optimization with their gap increasing slightly as the numbers of antennas increase. In all cases, the achievable rate increases as either P_0 or the numbers of antennas increase.

Fig. 4.4 then presents the rate results for different numbers of antennas with $P_0 = P = 1$. It is clear that the instantaneous capacities of the proposed schemes are much better than that of the NAF algorithm and the joint source, relay and TS ratio optimization shows performance gain over the uniform source precoding scheme. Notably, if numbers of the antennas increase, the performance gaps among the three schemes also increase, which agrees with the results in Fig. 4.3.

4.7 Summary

This chapter studied the rate maximization of a MIMO relay network with a time switching based energy harvesting relay node. We started with the fixed source covariance matrix scenario assuming uniform source precoding and then considered joint optimization with the source covariance. Closed-form solution as well as an

iterative scheme were proposed, respectively, for the two cases. Simulations demonstrate that joint optimization of the source, relay and TS ratio yields rate gain over the relay and TS ratio only optimization. Now that the a series of crucial issues in beamforming for SWIPT systems have been investigated, such as the power splitting and time switching based receiver design, the combination of multi-antenna techniques, robust beamforming, and also the application in multi-user networks and relay systems, with the performance metrics ranging from energy efficiency to data rate. In the chapters that follow, we will continue to investigate the resource allocation techniques for wireless edge caching.

Chapter 5

Optimizing Cache Placement for Heterogeneous Small Cell Networks

5.1 Overview

To achieve the targets of the 5G cellular communication systems, a new round of exploration on communication technology has begun. To name just a few, massive MIMO antennas, millimeter wave, D2D communication, small cell networks, etc., have recently attracted considerable attentions. Although these techniques are anticipated to contribute massively to 5G, challenges arise due to high demands on backhaul for massive content delivery imposed by the explosive mobile traffic growth. An effective solution to tackle this is to cache popular files at the network edge before users request them. By doing so, contents are brought closer to users and presumably, the peak-time traffic at core network, latency, and backhaul cost can be much reduced. In this chapter, a typical cache-enabled small cell network with heterogeneous file and cache sizes is considered with maximum distance separable (MDS) codes used for content restructuring. In particular, multicast content delivery is adopted to reduce the backhaul rate exploiting the independence among MDS coded packets. Unlike the online settings in literature which assume perfect user request information, we estimate the possible joint user requests using the file popularity information and aim at minimizing the long-term average backhaul load subject to the cache capacity constraints by optimizing the content placement in

all the cells jointly. The problem is reformulated into a mixed integer nonlinear program (MINLP) and solved with existing solver after linearization. Mathematical analysis and simulation results are provided to demonstrate the advantages of exploiting MDS codes and multicast content delivery in terms of reducing the backhaul requirements for cache-enabled small cell networks.

5.2 Related Work

As mentioned above, we unlock the potential of multicast-aware content delivery to reduce the backhaul requirements for cache-enabled small cell networks taking the advantage of the MDS codes. The works related to ours are [123, 143] which focused on optimizing the content placement for cache enabled small-cell networks. In [143], MDS coded caching was considered with homogeneous network settings, i.e. identical file sizes, homogenous cache sizes and file popularity for all the cells which gave rise to the assumption of identical content placement in all the cells. And the backhaul load minimization was performed in terms of any single user with cache misses of different users served with separate unicast transmissions via backhaul. As opposite to [143], [123] considered uncoded multicast-aware caching in delay tolerant networks assuming that the consecutive requests for the same file within a multicast period can be served by a single multicast transmission. Here the number of requests associated to a particular file and small cell base station (SBS) was modeled with the Poisson probability distribution determined by the length of the multicast period and a unique parameter which was given directly without clarification on the relation with the information of file popularity and the served users. Both of the proposed linear relaxation based scheme and the heuristic scheme worked in a greedy manner towards all kinds of joint user demands in the cells as well as the possible content placement and hence could not be used in the coded caching scenario. In this chapter, we aim to obtain the optimal (offline) cache content placement for minimizing the long-term average backhaul rate subject to cache capacity constraints for small cell networks. Unlike [143] considering an unlikely setting of identical content placement in all caches with homogeneous

settings, we consider a much more practical scenario with heterogeneous file and cache sizes and in this case the content placement in different caches will not always be the same. Hence, it is no longer available for the MBS to deliver the uncached content via a shared link. However, it is obviously not a good idea to use unicast between the MBS and SBSs. To tackle this problem, we utilize the independence among MDS coded packets and unlock the potential of multicast content delivery to reduce the backhaul rate. A near-optimal solution is obtained using a specific solver for mixed integer linear program (MILP) after a series of reformulations.

5.3 System Model

In this section, the network model with caching policy, as well as the content characteristics which involve the structure of the network coding and the file popularity profiles are presented.

5.3.1 Network Model

We consider a small cell network comprising a single MBS, and K small cells each consisting of a single SBS and I_k users among which each SBS can only answer to the requests of a maximum of $I(I \geq I_k, \forall k)$ users at the same time (seeing Fig. 5.1). The requests of the remaining users are served by the MBS. It is assumed that there is no coverage overlapping amongst all the SBSs which operate in subchannels disjoint with the MBS. Moreover, enhanced inter-cell interference coordination techniques (eICIC) or/and orthogonal spectra are utilized by the neighboring SBSs [152, 153]. We also assume that the MBS has access to all the files defined as $\mathscr{F} \triangleq \{f_1, f_2, \ldots, f_N\}$ with distinct file sizes $\mathbf{s} = [s_1, s_2, \ldots, s_N]$. The users located outside of the small cells can only be served by the MBS and hence are ignored when considering the backhaul rate from the MBS to the SBSs. Note that it is also assumed that each user is able to request one file at one time slot. Instead of assuming identical cache size in all SBSs which is difficult to satisfy in practice, here we consider that the SBSs have heterogeneous cache sizes. We let $M_k (M_k \leq \sum_{j=1}^{N} s_j)$ be the cache size in SBS k. By caching part of the files in the SBSs before users requesting, we are able to bring the content closer to the users and hence reduce

the peak data rate, latency and backhaul rate, giving rise to the so-called local gain described in [52]. In the following, we describe the caching policy briefly.

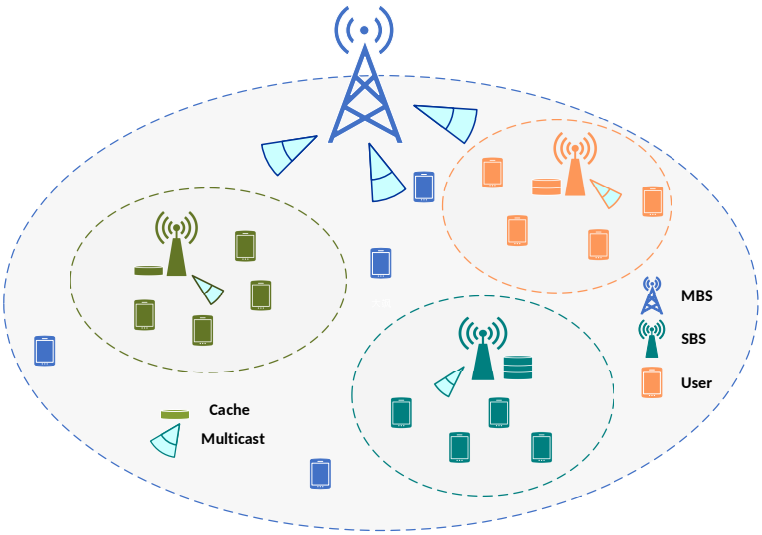


Figure 5.1: Multicast-aware cache enabled heterogeneous small cell networks.

Since each cell is allocated with limited cache capacity to store popular content, the SBSs push the cached packets to the users when requested while the uncached parts are delivered to the SBSs via the backhaul from the MBS. Taking advantages of the independence among the MDS codes, we adopt *multicasting* between MBS and SBSs to reduce the backhaul rate. In this case, the least amount of coded packets to be delivered via backhaul in order to rebuild the requested file is determined by the user request profile as well as the cache content placement in all the cells jointly. Based on the file popularity information, we aim to obtain the optimal content placement in order to minimize the average backhaul load in terms of all the possible user request profiles under cache capacity constraints.

5.3.2 MDS Coding

MDS codes are employed to construct pieces of a file that can be put back together to recover the file. They are particularly suitable for our settings of multicast-aware caching in which the cached content in different cells needs to be coordinated. Compared to the case of storing uncoded fragments, MDS codes bring a unique benefit that the coded packets are all independent from each other so that a certain num-

ber of randomly drawn packets will be sufficient to recover the file. This allows us to use only the number of packets stored in each cell, instead of the details of the packets, to derive the backhaul load, simplifying the analysis.

Considering MDS codes parametrized by (l_j, n_j), file j is equally cut into n_j fragments and then coded into l_j independent packets any n_j of which can rebuild the file. The flowchart presenting the MDS coding process is given in Fig. 5.2.

We assume that the SBS in cell k caches $m_{k,j}$ coded packets of file j and let $\mathbf{m}^j = [m_{1,j}, m_{2,j}, \ldots, m_{K,j}]$ be the content placement vector for file j. To recover the requested file with minimum redundancy, file j is coded into $l_j = \sum_{k=1}^{K} m_{k,j} + n_j - \min_{k=1}^{K} m_{k,j}$ packets to ensure that the uncached packets delivered from the MBS are different from all the cached packets, even in an extreme case that the SBSs store totally different packets.

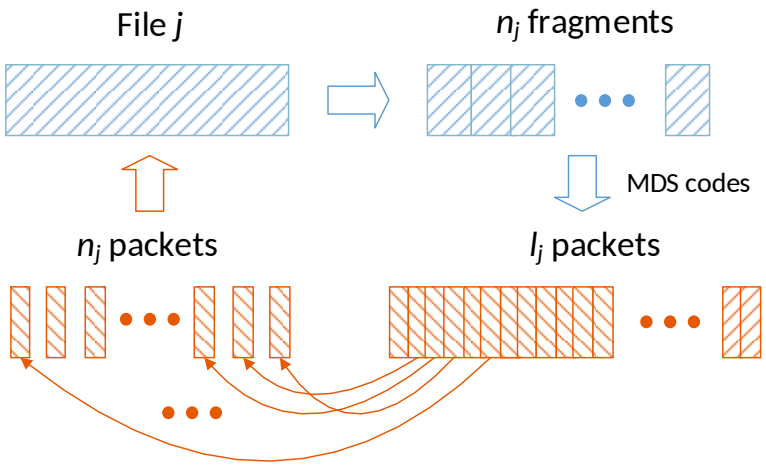


Figure 5.2: The flowchart of the MDS coding process

5.3.3 File Popularity Profile

Without loss of generality, here we assume that the file popularity in all the cells obey Zipf's distribution. Assuming that the popularity of the files is arranged in a descending order according to the Zipf's law, the frequency for file j to be requested by each user can be written as [154]

$$p_j = \frac{(1/j^{\gamma})}{\sum_{i=1}^{N} (1/i^{\gamma})}, \ \forall j, \tag{5.1}$$

where γ is the skewness reflecting the concentration of the popularity distribution. A higher γ means a more concentrated popularity distribution. Hence, the probability of file j not being requested by the users in the cell is

$$\alpha_j = (1 - p_j)^I, \ \forall j. \tag{5.2}$$

Thus, the probability for file j being requested by any of the users in the cell will be $1 - \alpha_j$.

5.4 Content Placement Optimization

As mentioned, the cache content placement is optimized aiming to minimize the average backhaul load in terms of all possible user request profiles which means that the content placement should be carefully designed to satisfy different requests at all the cells simultaneously with a single multicast transmission instead of multiple unicast transmissions to each SBS separately at the least backhaul rate. Unlike in literature where it was usually assumed that we had the knowledge of the actual requests for all the cells, here we analyze all possible request profiles and their probabilities to appear using the learned file popularity. Note that in consideration of multicast transmission at the MBS, the coordination between the requests in different cells counts a lot and hence joint user request profiles in all the cells are focused rather than the user request profile in individual cell. Consequently, we let Π_j be the collection of all the possible user request profiles and $\pi_j \in \Pi_j$ denote a particular user request profile for file j in all cells. Given any user request profile π_j, $\mathscr{K}_{\pi_j}$ is used to denote the set of the cells where file j is required by the served users. In case that file j is requested in all the cells except cell K, we let $\pi_j = [1,1,\ldots,1,0]_{1\times K}$ where 1 means that file j is requested by users in the considered cell while 0 states that none of the users in the cell requests the file. And then it follows that $\mathscr{K}_{\pi_j} = \{1,2,\ldots,K-1\}$ for the mentioned π_j. The joint user request profile for all the files simultaneously can be written as $\{\pi_1,\ldots,\pi_N\}$. For each file j, if there are $t(\leq K)$ cells where the served users request file j, the corresponding file request profile π_j and the cell set $\mathscr{K}_{\pi_j}$ may have $\binom{K}{t}$ possible combinations. In this way, the total number of different π_j and $\mathscr{K}_{\pi_j}$ will be as high as 2^K.

Our aim is to minimize the long-term average backhaul load, i.e., the volume of the file packets needed to be delivered via backhaul using multicasting, subject to the cache capacity constraints in terms of all possible user request profiles by optimizing the cache content placement. The average backhaul rate is obtained by taking expectation of the instantaneous backhaul rate with respect to the joint probability of user request profile for all the files $\{\pi_1,\ldots,\pi_N\}$. Mathematically, the problem can be written as

$$\min_{\{m_{k,j}\}} \sum_{\{\pi_1,\ldots,\pi_N\}} \sum_{j=1}^{N} \left(1-\min_{k\in\mathscr{K}_{\pi_j}} \frac{m_{k,j}}{n_j}\right) s_j P_r(\{\pi_1,\ldots,\pi_N\}) \tag{5.3a}$$

$$\text{s.t.} \quad \sum_{j=1}^{N} \frac{m_{k,j}}{n_j} s_j \leq M_k, \ \forall k, \tag{5.3b}$$

$$0 \leq m_{k,j} \leq n_j, \ \forall k, j, \tag{5.3c}$$

where $P_r(\{\pi_1,\ldots,\pi_N\})$ shows the joint probability that a certain user request profile, i.e. $\{\pi_1,\ldots,\pi_N\}$ appears, and s_j denotes the size for file j. Considering the number of the cells and the size of the file profile, the analysis and calculation of the joint probability would be rather complex. To proceed, reformulation has been given to simplify the expression of the average backhaul rate in the following lemma.

Lemma 5.1 *Based on the fact that the backhaul load for a particular file j only relies on π_j regardless of $\{\pi_i\}_{i\neq j}$, the average backhaul rate in (5.3a) can be rewritten as*

$$C_{\text{multicast}}^{\text{MDS}} = \sum_{j=1}^{N} \sum_{\pi_j\in\Pi_j} \left(1-\min_{k\in\mathscr{K}_{\pi_j}} \frac{m_{k,j}}{n_j}\right) s_j P_r(\pi_j), \tag{5.4}$$

where $P_r(\pi_j)$ shows the probability that a certain user request profile π_j appears.

Proof 5.1 *See* Appendix B.

To show the advantages of storing MDS coded packets over storing the uncoded segments directly in our settings, we assume that the SBS in cell k stores m_j^k different fragments randomly drawn among the n_j fragments. In this case, all fragments for file j except the ones that have been stored in all of the cells requesting the

particular file have to be sent from the MBS via backhaul using multicast. Consequently, which fragments are stored in the caches is also needed to be learned to determine the backhaul rate except for the numbers of packets stored in the cells. Let $\mathscr{M}_j$ show the detail of the fragments of file j stored in the caches and $d\left(\mathscr{M}_j, \mathscr{K}_{\pi_j}\right)$ denote the number of same fragments stored in all the cells requesting file j. The backhaul rate is given by

$$C_{\text{multicast}}^{\text{Uncoded}} = \sum_{j=1}^{N} \sum_{\pi_j \in \Pi_j} \left(1 - \frac{d\left(\mathscr{M}_j, \mathscr{K}_{\pi_j}\right)}{n_j}\right) s_j P_r(\pi_j). \tag{5.5}$$

Due to the fact that the number of same fragments stored in all the cells requesting file j is always less than or equal to the minimum number of the fragments stored in those cells, i.e. $d\left(\mathscr{M}_j, \mathscr{K}_{\pi_j}\right) \leq \min_{k \in \mathscr{K}_{\pi_j}} m_{k,j}$, it is proved that the utility of MDS codes helps reduce the average backhaul rate. Specially, if the uncoded segments are assumed to be randomly drawn among the n_j fragments *equiprobably*, the probability of each segment of file j being stored in all the cells requesting the file would be

$$\rho_j = \prod_{k \in \mathscr{K}_{\pi_j}} \frac{\binom{N-1}{m_{k,j}-1}}{\binom{N}{m_{k,j}}} = \prod_{k \in \mathscr{K}_{\pi_j}} \frac{m_{k,j}}{N}. \tag{5.6}$$

Because of $\frac{m_{k,j}}{N} \leq 1, \forall k \in \mathscr{K}_{\pi_j}$, it holds true that $\rho_j \leq \min_{k \in \mathscr{K}_{\pi_j}} \frac{m_{k,j}}{N}$. In this case, the expectation of d in terms of different $\mathscr{M}_j$ with given $\mathbf{m}^j$ and π_j is given by $\overline{d}\left(\mathbf{m}^j, \mathscr{K}_{\pi_j}\right) = N\rho_j \leq \min_{k \in \mathscr{K}_{\pi_j}} m_{k,j}$. The same conclusion can be drawn.

Although $P_r(\pi_j)$ can be calculated using (5.2) in (5.4), it would be difficult to fully list all possible user request profiles and analyze the objective function correspondingly as mentioned in the beginning of this section. However, if we know the relationships among the values of all the elements in $\mathbf{m}^j$, a closed-form expression of the objective function can be obtained in the following lemma.

Lemma 5.2 *Let $r_{k,j}$ denote the rank of the value of $m_{k,j}$ in $\mathbf{m}^j$. For instance, $r_{k,j} = 1$ means $m_{k,j}$ is the smallest while $r_{k,j} = K$ states that $m_{k,j}$ is the largest in $\mathbf{m}^j$. The objective function (5.3a) can then be rewritten as*

$$C_{\text{multicast}}^{\text{MDS}} = \sum_{j=1}^{N} \sum_{k=1}^{K} \left(1 - \frac{m_{k,j}}{n_j}\right) s_j (1 - \alpha_j) \alpha_j^{(r_{k,j}-1)}. \tag{5.7}$$

Proof 5.2 *Firstly, we divide the possible user request profiles for each file, e.g., π_j into $K+1$ types defined as $\{\pi_j^0, \pi_j^1, \pi_j^2, \ldots, \pi_j^K\}$ according to the different values of the associated backhaul load for file j, i.e., $\{0, 1-\frac{m_{1,j}}{n_j}, 1-\frac{m_{2,j}}{n_j}, \ldots, 1-\frac{m_{K,j}}{n_j}\}$, respectively. Note that when file j is not requested by any of the cells, the backhaul is not needed. If cell k stores the least number of packets of file j among all the cells requesting file j, i.e., $\min_{t\in\mathcal{K}_{\pi_j}} \frac{m_{t,j}}{n_j} = \frac{m_{k,j}}{n_j}$, then the associated user request profile π_j^k will imply that file j is requested by cell k and that probably some cells have cached more packets of file j but there will not be any cell t satisfying $r_{t,j} < r_{k,j}$, i.e., $m_{t,j} \leq m_{k,j}$. Hence, we have $P_r(\pi_j^k) = (1-\alpha_j)\alpha_j^{(r_{k,j}-1)}$. Finally, after summing up all types of user request profiles $\{\pi_j^k\}$ for all files, the average backhaul rate can be written as (5.7).*

As a comparison, in the typical unicast case without coverage overlap among the SBSs, the backhaul rates for storing uncoded fragments directly or coded packets would be [143]

$$C_{\text{unicast}} = \sum_{j=1}^{N}\sum_{k=1}^{K}\left(1-\frac{m_{k,j}}{n_j}\right)s_j(1-\alpha_j). \tag{5.8}$$

Note that after using multicast, additional multipliers $0 < \alpha_j^{(r_{k,j}-1)} \leq 1, \forall k, \forall j$ appear and hence bring a global gain. That is to say, $C_{\text{multicast}}^{\text{MDS}}$ is always smaller than C_{unicast}.

Substituting (5.7) into (5.3), the problem of interest becomes

$$\min_{\{m_{k,j}\}} \sum_{j=1}^{N}\sum_{k=1}^{K}\left(1-\frac{m_{k,j}}{n_j}\right)s_j(1-\alpha_j)\alpha_j^{(r_{k,j}-1)} \tag{5.9a}$$

$$\text{s.t. (5.3b) and (5.3c).} \tag{5.9b}$$

Note that we can separate the files into an arbitrary number of fragments. We define $q_{k,j} \triangleq \frac{m_{k,j}}{n_j}$ as the cached percentage of file j in SBS k. Accordingly, we let $\mathbf{q}^j =$

$[q_{1,j}, q_{2,j}, \ldots, q_{K,j}]$ and their ranks remain the same. Then (5.9) is rewritten as

$$\min_{\{q_{k,j}\}} \quad \sum_{j=1}^{N} \sum_{k=1}^{K} \left(1 - q_{k,j}\right) s_j (1-\alpha_j) \alpha_j^{(r_{k,j}-1)} \tag{5.10a}$$

$$\text{s.t.} \quad \sum_{j=1}^{N} q_{k,j} s_j \leq M_k, \forall k, \tag{5.10b}$$

$$0 \leq q_{k,j} \leq 1, \ \forall k, j. \tag{5.10c}$$

In (5.10), $\{q_{k,j}\}$ are to be optimized and hence unknown before the problem is solved. It is impossible to predict the ranks $\{r_{k,j}\}$ which depend on the values of $\{q_{k,j}\}$. To tackle this problem, we firstly sort $\mathbf{q}^j, \forall j$ in an ascending order and define the sorted variables as $\mathbf{g}^j = [g_{1,j}, g_{2,j}, \ldots, g_{K,j}]$ with $r_{k,j} = k$ in $\mathbf{g}^j, \forall j$. Problem (5.10) is then expressed as

$$\min_{\{q_{k,j}\},\{g_{k,j}\}} \quad \sum_{j=1}^{N} \sum_{k=1}^{K} \left(1 - g_{k,j}\right) s_j (1-\alpha_j) \alpha_j^{(k-1)} \tag{5.11a}$$

$$\text{s.t.} \quad \mathbf{g}^j = \text{sort}(\mathbf{q}^j), \forall j, \tag{5.11b}$$

$$\text{(5.10b)–(5.10c)}. \tag{5.11c}$$

Nevertheless, sorting the variables to be optimized is definitely unconvex. The challenge then becomes the problem of finding a way to demonstrate the relationships between $\mathbf{q}^j$ and $\mathbf{g}^j$ in order to satisfy the cache capacity constraints.

Lemma 5.3 *By introducing a new matrix* $\mathbf{X} = [x_{t,j}^k]_{K \times N \times K}$ *with* $x_{t,j}^k \in \{0,1\}$ *such that*

$$q_{k,j} = \sum_{t=1}^{K} g_{t,j} x_{t,j}^k, \tag{5.12}$$

we can rewrite problem (5.11) as

$$\min_{\{g_{k,j}\},\{x^k_{t,j}\}} \quad \sum_{j=1}^{N}\sum_{k=1}^{K}\left(1-g_{k,j}\right)s_j(1-\alpha_j)\alpha_j^{(k-1)} \tag{5.13a}$$

$$\text{s.t.} \quad \sum_{j=1}^{N}\sum_{t=1}^{K} g_{t,j}x^k_{t,j}s_j \le M_k, \forall k, \tag{5.13b}$$

$$\sum_{t=1}^{K} x^k_{t,j} = 1, \forall k, j, \tag{5.13c}$$

$$\sum_{k=1}^{K} x^k_{t,j} = 1, \forall t, j, \tag{5.13d}$$

$$0 \le g_{k,j} \le 1, \forall k, j, \tag{5.13e}$$

$$g_{k,j} \le g_{k+1,j}, \forall k < K, \text{ and } \forall j, \tag{5.13f}$$

$$x^k_{t,j} \in \{0,1\}, \forall t, j, k. \tag{5.13g}$$

Proof 5.3 *See* Appendix C.

Clearly, (5.13) is a mixed integer nonlinear program (MINLP) and hence cannot be solved directly. Therefore, we resort to linearizing the products of the variables in (5.13b) to make (5.13) an MILP which can be solved by well-known solvers such as Gurobi [155]. To proceed, the following lemma is necessary.

Lemma 5.4 *Let z be the product of a binary x and a continuous variable $y(0 \le y \le \widetilde{y})$. We can linearize the equation $z = xy$ by adding the following constraints* equivalently

$$z \le x\widetilde{y}, \tag{5.14}$$

$$z \ge y - (1-x)\widetilde{y}, \tag{5.15}$$

$$z \le y, \tag{5.16}$$

$$z \ge 0. \tag{5.17}$$

Proof 5.4 *See* Appendix D.

According to *Lemma 5.4*, we can easily replace the products in constraint (5.13b) with a new group of variables defined as $\mathbf{Z} = [z^k_{t,j}]_{K\times N\times K}$. Then (5.13)

can be rewritten as

$$\min_{\{g_{k,j}\},\mathbf{X},\mathbf{Z}} \quad \sum_{j=1}^{N}\sum_{k=1}^{K}\left(1-g_{k,j}\right)s_j(1-\alpha_j)\alpha_j^{(k-1)} \tag{5.18a}$$

$$\text{s.t.} \quad \sum_{j=1}^{N}\sum_{t=1}^{K} z_{t,j}^{k} s_j \leq M_k, \forall k, \tag{5.18b}$$

$$z_{t,j}^{k} \leq x_{t,j}^{k}, \forall t, j, k, \tag{5.18c}$$

$$z_{t,j}^{k} \geq g_{k,j} - (1 - x_{t,j}^{k}), \forall t, j, k, \tag{5.18d}$$

$$z_{t,j}^{k} \leq g_{k,j}, \forall t, j, k, \tag{5.18e}$$

$$z_{t,j}^{k} \geq 0, \forall t, j, k, \tag{5.18f}$$

$$(5.13c)\text{–}(5.13g). \tag{5.18g}$$

Based on the equivalence between the obtained linear constrains and the assumption of $z_{t,j}^{k} = g_{t,j}x_{t,j}^{k}$ with the basic settings of $\{g_{t,j}\}$ and $\{x_{t,j}^{k}\}$, it is apparent that the obtained solution to problem (5.18) will also be the solution to (5.13).

5.5 Simulation Results

Here, the performance of the proposed scheme is studied via simulations. For comparison, results for the uniform and popularity based content placement schemes in [123, 143] are also provided. The uniform scheme assumes that all files are equally cached in each SBS with $q_{k,j} = M_k/(Ns_j), \forall k, j$. In the popularity based scheme, the files are put into each cache one by one according to their popularity (from high to low) until the cache is fully occupied. To show the benefit of multicast, the unicast scenario in (5.8) is studied as well. In the following, we consider a small cell network with $K = 3$ cells and the cache sizes $M = [\frac{M_0}{3} + \Delta m, \frac{M_0}{3}, \frac{M_0}{3} - \Delta m]$ where M_0 is the total cache size while Δm is the cache size differentiation. We consider $N = 10$ files with their sizes randomly chosen uniformly between 1 and 5 independently. Unless otherwise specified, we set $M_0 = 20, \Delta m = 3, I = 10, \gamma = 1$. Note that the backhaul rates have been scaled with the total file size. The results in the multicast and unicast scenarios are presented using the left and right axes, respectively.

Fig. 5.3 studies the average backhaul rates against different total cache sizes.

As we can see, the backhaul rates decrease with the increase of total cache size in all cases. Also, as expected, the multicast based schemes outperform the unicast based scheme. The proposed scheme, which considers the heterogeneity of the cache and file sizes as well as the file popularity, apparently reduces the backhaul load among all.

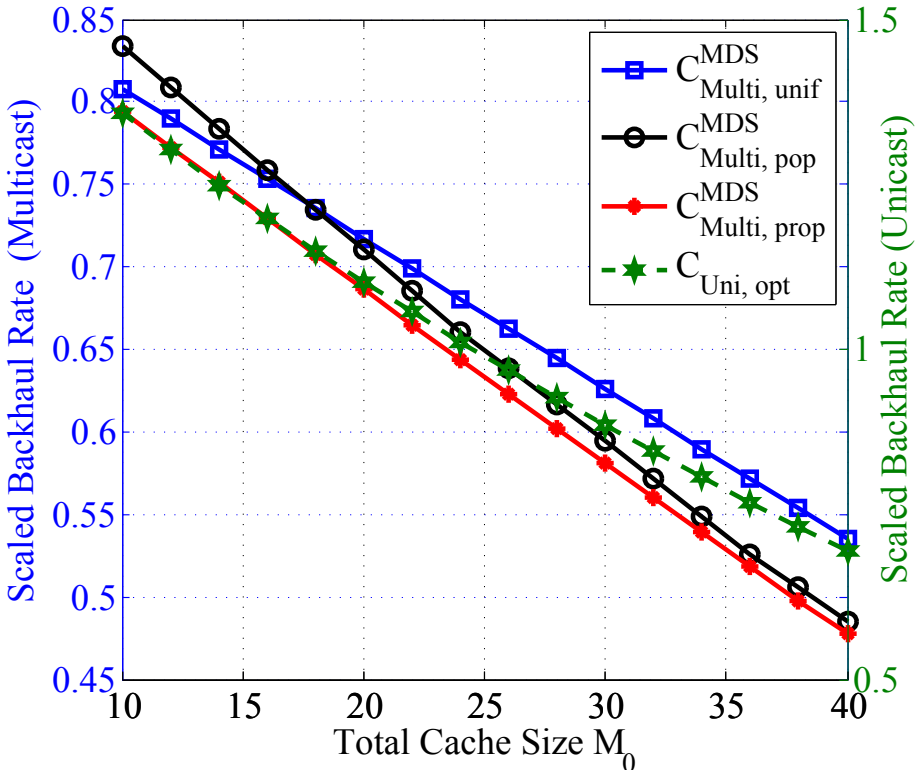


Figure 5.3: The backhaul rates versus the total cache size M_0.

Fig. 5.4 explores the impact of the cache size differentiation Δm on the backhaul rates with fixed M_0. Similar to Fig. 5.3, the proposed scheme shows the best performance. In addition, the backhaul rate reduction of the proposed scheme over the popularity based scheme increases drastically when improving Δm which illustrates the significance of the proposed scheme in small cell networks with heterogeneous cache sizes.

Fig. 5.5 shows the impact of the number of served users I in each cell on the backhaul rate. In the unicast scenario, the backhaul rate rises rapidly with the increase of I. Conversely, the backhaul rates of the multicast based schemes increase slightly when I reaches a certain degree, which greatly increases the maximum

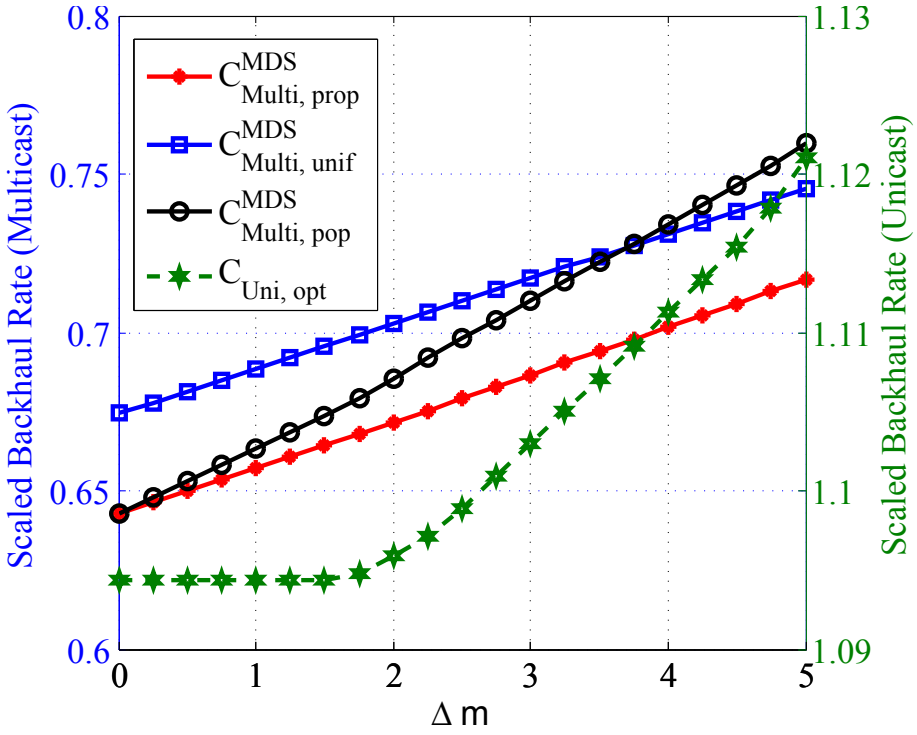


Figure 5.4: The backhaul rates versus Δm.

number of potential served users.

Finally, the impact of the skewness γ is studied in Fig. 5.6. Again, the multicast based schemes show better performance than the unicast based scheme and the proposed scheme yields the lowest backhaul rate. Moreover, the performance gain of the proposed scheme to the popularity based scheme changes little when increasing γ while that to the uniform scheme, which ignores the file popularity, rises more obviously.

5.6 Summary

In this chapter, the optimization of cache content placement was investigated for MDS coded caching enabled small cell networks with heterogeneous file and cache sizes. To minimize the average backhaul rate, multicast transmission was adopted. The cache content placement optimization is initially formulated as a nonconvex problem, then reformulated into a typical MILP, and finally solved by optimization tools. Results showed that the proposed scheme using MILP outperforms the existing schemes in terms of backhaul requirements. In the next chapter, the research

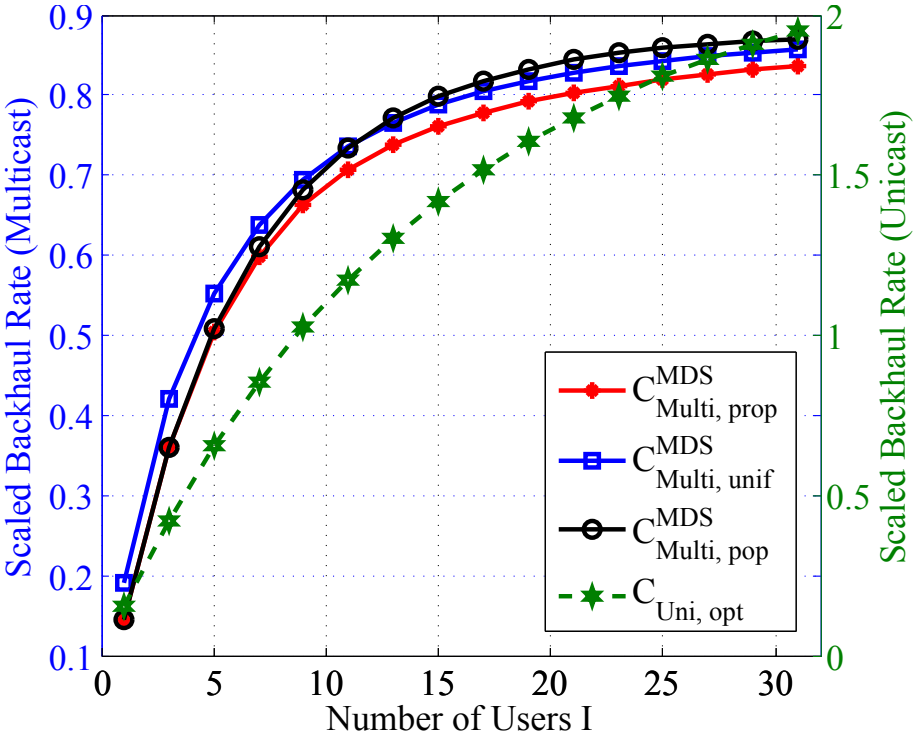


Figure 5.5: The backhaul rates versus I.

will be extended into more practical scenarios with distinct number of users and content popularity in each cells, and provide a more profound study on improving caching gains by making the best use of multicast based content delivery and cooperative content sharing.

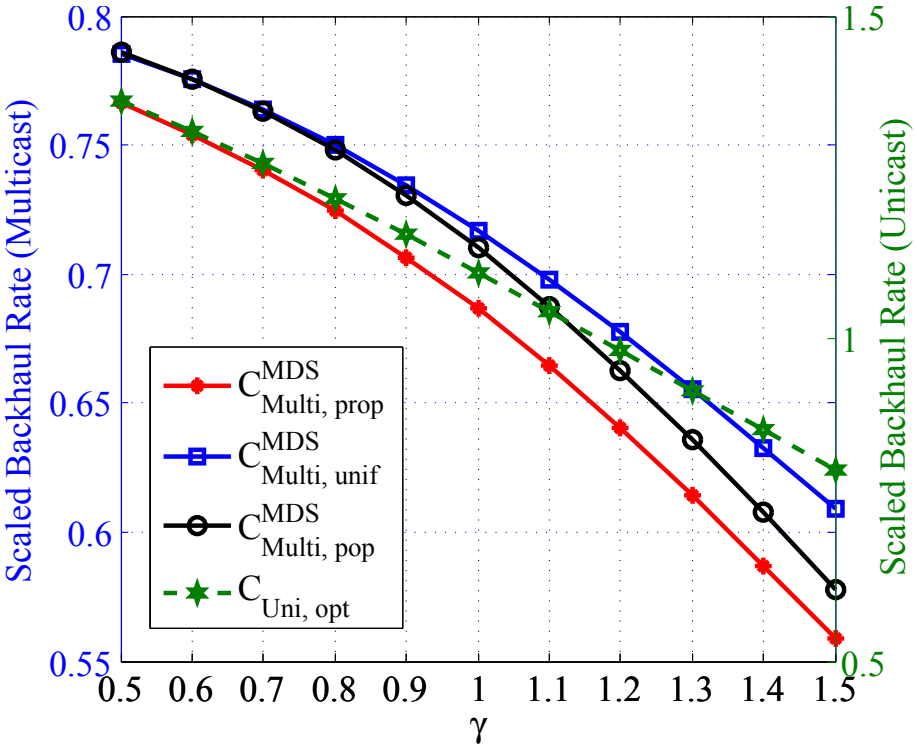


Figure 5.6: The backhaul rates versus skewness γ.

Chapter 6

Coding, Multicast and Cooperation for Cache-Enabled Heterogeneous Small Cell Networks

6.1 Overview

Caching at the wireless edge is a promising approach to deal with massive content delivery in heterogeneous wireless networks which have high demands on backhaul. In this chapter, a typical cache-enabled small cell network under heterogeneous file and network settings is considered using MDS codes for content restructuring. Unlike those in the literature considering online settings with the assumption of perfect user request information, we estimate the joint user requests using the file popularity information and aim to minimize the *long-term average* backhaul load for fetching content from external storage subject to the overall cache capacity constraint by optimizing the content placement in all the cells jointly. Both multicast-aware caching and cooperative caching schemes with optimal content placement are proposed. In order to combine the advantages of multicast content delivery and cooperative content sharing, a compound caching technique, which is referred to as multicast-aware cooperative caching, is then developed. For this technique, a greedy approach and a multicast-aware in-cluster cooperative approach are proposed for the small scale networks and large scale networks, respectively. Mathematical analysis and sim-

ulation results are presented to illustrate the advantages of MDS codes, multicast and cooperation in terms of reducing the backhaul requirements for cache-enabled small cell networks.

6.2 Related Work

Of relevance to our work are [105, 123, 140–149] where they focused on the optimization of content placement for cache-enabled small-cell networks. Firstly, [140] studied the optimal caching and user association strategy for a small cell network with a macro cell and multiple cache-enabled SBSs which was similar to ours. However, multicast transmission and collaboration at the BSs were not considered with also the limits of storing entire files and homogenous file popularity.

By using file partitioning and network coding, storing subfiles in the caches instead of storing entire files has been well recognized as an effective way to improve content diversity. The optimal uncoded and coded data allocation strategies with the minimum expected costs were studied in [141], where only one single file was considered ignoring the diversity of the required file library in practice. In [142], both the analysis and optimization were extended to the multiple files scenario with two partition-based caching designs studied for a large scale successive interference cancellation (SIC)-enabled wireless network. In [143], MDS coded caching was considered with homogeneous network settings, i.e., same file sizes, cache sizes and file popularity for all the cells, which gave rise to identical content placement in all cells. Any cache miss was dealt with by separate costly unicast transmissions via the backhaul.

In addition to the studies on caching strategies using multiple unicast transmissions to serve the requests mentioned above, multicasting transmission at BSs to serve the requests for the same file simultaneously has been explored to support massive content delivery over wireless networks. In [144], joint throughput-optimal caching and scheduling algorithms were developed to maximize the service rates with both elastic and inelastic requests. For inelastic services, optimal mutlicasting scheduling was discussed while unicast communication was assumed for elastic re-

quests. In another work [123], the authors studied uncoded multicast-aware caching in delay tolerant networks, with the assumption that consecutive requests for the same file within a multicast period can be served by a single multicast transmission. Although heterogeneous settings were assumed, no extra challenges were brought in this case since the discrete optimization problem was solved in a rather heuristic and exhaustive manner with all the possible joint user request profiles fully listed and calculated which limits its usage in large scale networks and coded caching scenarios. In the previous chapter [145], although coded multicast-aware caching was proposed, the research was limited to the partly heterogeneous settings of distinct cache and file sizes but homogeneous file popularity and numbers of users in all the cells. Beisdes caching design, [105, 146] offer performance analysis towards caching and multicasting for single-tier and multi-tier HetNets, respectively. Although they provide some content diversity, the assumptions made in [105, 146] greatly limit the full usage of this diversity. For instance, the file library for the BSs in the same tier to cache from is actually the same while those for BSs in different tiers are mutually exclusive. The identical caching in the macro-tier, the random caching design with the same probability distribution in the pico-tier as well as the uncoded caching limitation of storing entire files altogether lead to this issue. Another main difference is that they focused more on multicast transmission between caches and users while we also exploit the multicast opportunities for delivering the uncached content.

While the works mentioned above are offline schemes with limited cache sizes, an online cooperative caching scheme with infinite cache capacity was presented in [147]. In this case, the energy consumption for content updating in the caches was considered which can be ignored in offline schemes in a long-term time scale. Due to the fact that the previous content placement and the current user demands were given and the caching policies for different files were mutually independent, the formulated problem was actually linear and therefore could be easily solved. Subsequently in [148], the study was extended to the joint design of caching, routing and interference management with perfect user request information.

Finally in [149], an in-network cooperative caching scheme was proposed assuming that the cooperative SBSs were connected to the same service gateway to share cached content. It was assumed that the costs for fetching content from any of the cooperative SBSs were identical and so did the costs for fetching content from the content provider to the SBSs. In that effort, a cooperative caching utility maximization problem was decomposed into a number of sub-problems in different network domains and addressed by a decentralized heuristic scheme with the strong assumption of knowing the actual file demands of each user. Furthermore, the scheme is suboptimal, and the heterogeneity of the locations of the SBSs and file popularity in different cells were not well addressed.

Considering the heterogeneity of cache-enabled small-cell networks, such as distinct file popularity, file sizes, cache sizes, coverages and locations of different SBSs, not only requires redesign of content placement but also cache size allocation amongst the SBSs, as mentioned in [150, 151]. In this setup, cache size allocation and content placement in different cells will generally not be the same. Considering also the fact that file sizes may be large compared to the limited cache size in practice, files are usually split into fragments. Nevertheless, note that all of the above-mentioned works considered whole file caching except [143, 145, 148]. When the fragments are randomly selected and stored in the caches without coding, both the number of fragments in each cell and which fragments that are stored (i.e., the degree of content duplication amongst the cells), determine the backhaul load. As a result, it would be very difficult for the MBS to deliver the uncached content via a shared link to all the cells and unicast content delivery is therefore commonly used between the MBS and SBSs at the expense of high backhaul cost [140–144, 147–149]. On the other hand, cache content overlap among different cells would restrain cooperative caching from being effective.

In this chapter, our aim is to unleash the potential of multicast-aware caching and cooperative caching by taking advantages of the inherent independence amongst the MDS coded packets for minimizing the average bakchaul rate. In summary, this chapter has made the following major contributions:

- We develop offline caching schemes optimizing the long-term average performance of the cache-enabled network by estimating all possible joint user requests in different cells simultaneously without the knowledge of the actual user requests assumed in [147–149]. Furthermore, unlike [123], we classify the large number of possible user request profiles into several types according to their values of the associated backhaul load and therefore reduce the computational complexity in terms of user request uncertainty in the analysis of multicast-aware caching. Moreover, a multicast-aware in-cluster cooperative approach is proposed suitable for large-scale networks.

- Unlike the homogenous settings considered in [143, 145, 149], the heterogeneity of the parameters that affects the design of cache management and cooperative policy is all considered with the coordination among different SBSs and files. Also, cache size allocation is optimized subject to an overall cache capacity budget rather than uniform or an arbitrarily given heterogeneous allocation in literature.

- Furthermore, we derive the performance gains of storing coded packets over uncoded fragments in the caches and quantify the advantages of multicast-aware and cooperative caching over common caching schemes via mathematical analysis or/and simulation results. Benefited from the independence of the MDS coded packets, we combine the merits of multicast-aware caching and cooperative caching to greatly reduce the backhaul load.

6.3 System Model

6.3.1 Network Model

A small cell network is considered which comprises a single MBS, and K non-overlapping small cells each consisting of a single SBS and I_k users, for the kth cell. Let $\mathscr{K} \triangleq \{1,\ldots,K\}$ denote the set of SBSs which operate in disjoint subchannels with the MBS in order to remove the impact of interference. Besides, any interference among neighboring SBSs is assumed eliminated by techniques such as eICIC

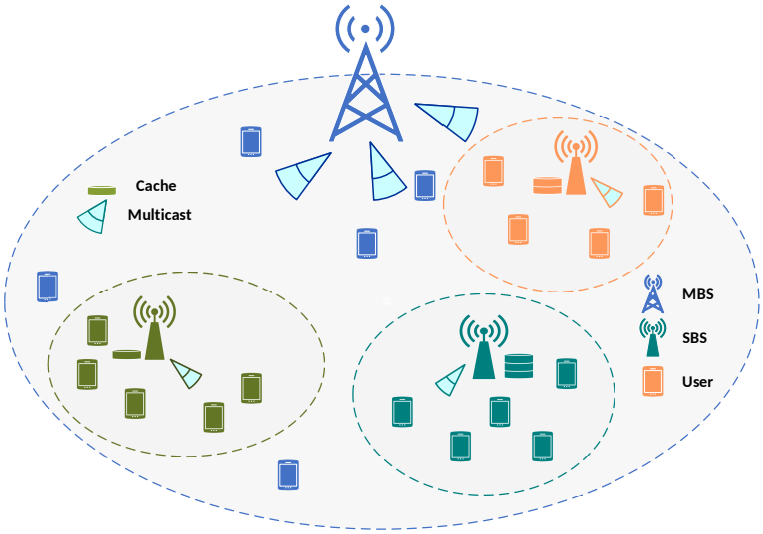
or/and orthogonal multiple access [152, 153]. We assume that the MBS has access to all files in the set $\mathscr{F} \triangleq \{f_1, f_2, \ldots, f_N\}$ with respective file sizes $\mathbf{s} \triangleq [s_1, s_2, \ldots, s_N]$ while the SBSs have limited cache capacities that are subject to a network-wide total cache capacity budget M. We let M_k denote the cache capacity for SBS k, with $M_k \leq \sum_{j=1}^{N} s_j$. SBSs can push the cached packets to the users when requested while the uncached parts have to be delivered to the SBSs via backhaul from the MBS (or cooperative SBSs in the case of cooperative caching). Note that the users located outside of any small cells are ignored when considering the backhaul requirements, as they can only be served by the MBS.

- **Multicast-aware caching:** If this approach is used, SBSs will fetch the uncached content from the MBS via backhaul using multicast, see Fig. 6.1a. Based on the file popularity information, we obtain the optimal content placement to minimize the average backhaul load for all possible user request profiles with the overall cache capacity budget.

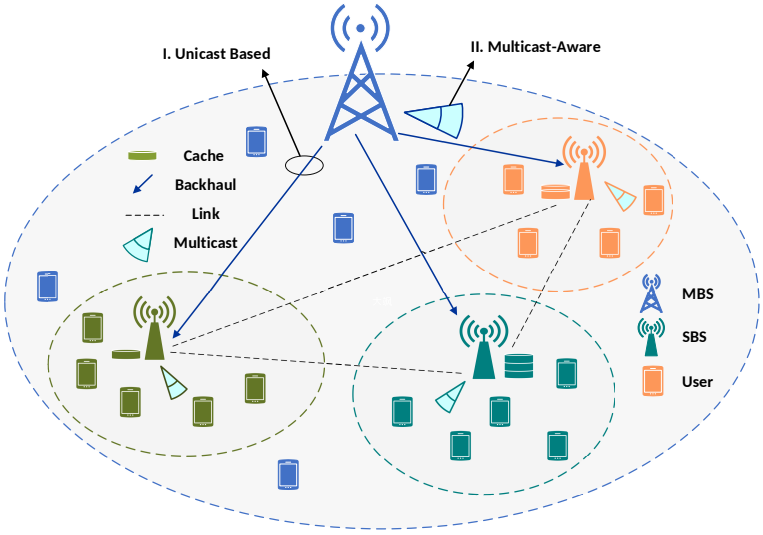
- **Cooperative caching:** As shown in Fig. 6.1b, neighboring SBSs can be connected to each other via high-capacity links to share their cached content in different cells collaboratively. In this scheme, the uncached content can be fetched from not only the MBS via backhaul but also the cooperative SBSs via the fronthaul links. Considering the different costs for fetching content from the MBS and the neighboring SBSs, we adopt the concept of user attrition (UA) cost introduced in [147], which denotes the overall cost for fetching content from an external storage, to evaluate the performance of the cooperative caching scheme. Cache content placement and the policy for SBS cooperation are to be jointly optimized to minimize the UA cost. Unless stated otherwise, this scheme uses unicast for content delivery.

- **Multicast-aware cooperative caching:** In this approach, multicast-based content delivery and content sharing amongst neighboring SBSs are combined with the aid of MDS codes. In contrast to conventional cooperative caching, multicasting is applied by the MBS to deliver content to the SBSs

requesting the same file simultaneously, see Case II of Fig. 6.1b.



(a) Multicast-aware caching



(b) Cooperative caching

Figure 6.1: Cache-enabled heterogeneous small-cell networks.

6.3.2 MDS Coding

As described in the previous chapter, MDS codes can bring benefits to both the content placement phase and the content delivery phase, in particular for the applications in multicast-aware caching and cooperative-caching which take the coordination and cooperation among different cells into account in the design of content

placement and delivery. As the principles and features of the MDS coding have been presented in the previous chapter, they are omitted here for the sake of brevity.

Without loss of generality, we still parametrize MDS codes by (l_j, n_j) such that file j is cut into n_j fragments and then coded into l_j independent packets by MDS. Any n_j packets can rebuild the entire file. Considering that the kth SBS caches $m_{k,j}$ coded packets of file j, we let $\mathbf{m}^j \triangleq [m_{1,j}, m_{2,j}, \ldots, m_{K,j}]$ be the content placement vector for file j. For multicast-aware caching, to ensure that the uncached packets delivered from the MBS are totally different from the ones cached in local servers, file j should be coded into at least

$$l_j = \underbrace{\sum_{k=1}^{K} m_{k,j}}_{\text{unique packets cached in SBSs}} + \underbrace{n_j - \min_{k\in\{1,\ldots,K\}} m_{k,j}}_{\text{unique packets delivered via backhaul}} \text{ packets.}$$

For unicast and multicast-aware cooperative caching scenarios, the total number of packets has to be at least

$$l_j = \underbrace{\sum_{k=1}^{K} m_{k,j}}_{\text{unique packets cached in SBSs}} + \underbrace{n_j - \min_{k\in\{1,\ldots,K\}} \sum_{t=1}^{K} x_{k,j}^{t}}_{\text{unique packets delivered via backhaul}} ,$$

where $x_{k,j}^{t}$ denotes the number of packets delivered from SBS t to SBS k to serve the requests for file j so that there is no content overlap in both content sharing process amongst the cooperative SBSs and content delivery phase at the MBS.

6.3.3 File Popularity Profile

Note that users in different cells may have different preferences towards the files. The most popular file in one cell may receive least attentions from another cell. It is thus better to consider local file popularity in each cell rather than the global popularity in the entire network which is often the case in the literature. Without loss of generality, here we assume that the file popularity in each cell obeys Zipf's distribution but with unique skewness parameter and popularity rank. According to the Zipf's law, the frequency for file j to be requested by each user in cell k can then

be written as [154]

$$p_{k,j} = \frac{\left(1/\lambda_{k,j}^{\gamma_k}\right)}{\sum_{i=1}^{N}(1/i^{\gamma_k})},\ \forall k, j, \tag{6.1}$$

where γ_k is the skewness in cell k reflecting the concentration of the popularity distribution and $\lambda_{k,j}$ denotes the rank of the popularity of file j in cell k. For instance, $\lambda_{k,j} = 1$ means file j is the most popular file in cell k. Hence, the probability of file j not being requested by the users in cell k is

$$\alpha_{k,j} = (1 - p_{k,j})^{I_k},\ \forall k, j. \tag{6.2}$$

Thus, the probability for file j being requested by at least one of the users in cell k will be $1 - \alpha_{k,j}$.

6.4 Multicast-Aware Caching

The aim is to minimize the average backhaul load for all possible user request profiles, meaning that content placement should be done to satisfy different requests for all the cells simultaneously with a single multicast transmission instead of multiple unicast transmissions to each SBS separately.

6.4.1 Problem Formulation

Different from the literature where the knowledge of the actual requests from the cells was usually assumed, we analyze all possible request profiles and their probabilities using the learned file popularity. Here, the joint user request profile in all the cells is focused rather than the user request profiles in individual cells. We let Π_j denote the collection of all the possible user request profiles and $\pi_j \in \Pi_j$ denote a particular user request profile for file j in all cells. Given any user request profile π_j, $\mathscr{K}_{\pi_j}$ is used to denote the set of the cells where file j is required by the served users. In case that file j is requested in all the cells except cell K, we have $\pi_j = [1,1,\ldots,1,0]_{1\times K}$ where 1 means that file j is requested by users in the considered cell while 0 states that none of the users in the cell requests the file. Therefore, it follows that $\mathscr{K}_{\pi_j} = \{1,2,\ldots,K-1\}$ for the mentioned π_j. The joint user request profile for all the files simultaneously can be written as $\{\pi_1,\ldots,\pi_N\}$. For each file j, if there are $t(\leq K)$ cells where the served users request file j, the corresponding

file request profile π_j and the cell set $\mathscr{K}_{\pi_j}$ may have $\binom{K}{t}$ possible combinations. In this way, we evaluate that the total number of different π_j and $\mathscr{K}_{\pi_j}$ will be as high as 2^K.

The average backhaul load is defined as the average volume of the file packets requiring to be fetched from the MBS via backhaul with a single multicast transmission in terms of all possible user request profiles. Our objective is to minimize the average backhaul load subject to the overall cache capacity constraint. Mathematically, that is,

$$\min_{\{m_{k,j}\}} \sum_{\{\pi_1,\ldots,\pi_N\}} \sum_{j=1}^{N} \left(1 - \min_{k\in\mathscr{K}_{\pi_j}} \frac{m_{k,j}}{n_j}\right) s_j P_r(\{\pi_1,\ldots,\pi_N\}) \tag{6.3a}$$

$$\text{s.t.} \sum_{k=1}^{K}\sum_{j=1}^{N} \frac{m_{k,j}}{n_j} s_j \le M, \tag{6.3b}$$

$$0 \le m_{k,j} \le n_j,\ \forall k,j, \tag{6.3c}$$

where $P_r(\{\pi_1,\ldots,\pi_N\})$ denotes the joint probability that a certain user request profile for all the files, i.e., $\{\pi_1,\ldots,\pi_N\}$ appears. Since there are multiple cells, users and also requested files, the required analysis and calculation of the joint probabilities would be rather complex. To this end, the following lemma is used to simplify the objective function in (6.3a).

Lemma 6.1 *Based on the fact that the backhaul load for a particular file j only relies on π_j regardless of $\{\pi_i\}_{i\neq j}$, the average backhaul rate in (6.3a) can be rewritten as*

$$R_{\text{multicast}}^{\text{MDS}} = \sum_{j=1}^{N} \sum_{\pi_j\in\Pi_j} \left(1 - \min_{k\in\mathscr{K}_{\pi_j}} \frac{m_{k,j}}{n_j}\right) s_j P_r(\pi_j). \tag{6.4}$$

where $P_r(\pi_j)$ is the probability that π_j appears.

Proof 6.1 *See [145, Appendix A].*

The following lemma exploits the relationships among the elements in $\mathbf{m}^j$ to express $R_{\text{multicast}}^{\text{MDS}}$ in closed form. Let $r_{k,j}$ be the rank of the value of $m_{k,j}$ among those of all the elements in $\mathbf{m}^j$. For instance, $r_{k,j}=1$ means $m_{k,j}$ is the smallest in $\mathbf{m}^j$ while $r_{k,j}=K$ states that $m_{k,j}$ is the largest.

Lemma 6.2 *The backhaul load in (6.3a) can be rewritten as*

$$R_{\text{multicast}}^{\text{MDS}}=\sum_{j=1}^{N}\sum_{k=1}^{K}\left(1-\frac{m_{k,j}}{n_j}\right)s_j(1-\alpha_{k,j})\prod_{t\in\mathscr{T}_{k,j}}\alpha_{t,j}, \tag{6.5}$$

in which $\mathscr{T}_{k,j}$ denotes the collection of cells storing no more packets of file j than cell k, i.e., $\mathscr{T}_{k,j}=\{t|r_{t,j}<r_{k,j}\}$.

Proof 6.2 *See* Appendix E.

6.4.2 Comparison

As a comparison, in the typical unicast case, the backhaul rate for storing uncoded fragments directly or the MDS coded packets would have been given by

$$R_{\text{unicast}}=\sum_{j=1}^{N}\sum_{k=1}^{K}\left(1-\frac{m_{k,j}}{n_j}\right)s_j(1-\alpha_{k,j}). \tag{6.6}$$

It can be observed in (6.5) and (6.6) that additional multipliers $0<\prod_{t\in\mathscr{T}_{k,j}}\alpha_{t,j}\leq 1,\forall k,\forall j$ appear after using multicast transmission at the MBS in the content delivery phase, and hence bring a global gain, i.e., $R_{\text{multicast}}^{\text{MDS}}<R_{\text{unicast}}$ [52]. On the other hand, it is worth pointing out that storing MDS coded packets has advantages over uncoded segments in the case of multicast-aware caching for minimizing the average backhaul rate. We assume that cell k stores $m_{k,j}$ different fragments randomly drawn from the n_j fragments *equiprobably*, and all fragments except the ones stored in all the cells requesting the particular file have to be sent from the MBS. Therefore,

$$R_{\text{multicast}}^{\text{uncoded}}=\sum_{j=1}^{N}\sum_{\pi_j\in\Pi_j}\left(1-\rho_{\pi_j}\right)s_jP_r(\pi_j), \tag{6.7}$$

where ρ_{π_j} denotes the probability of a certain fragment of file j being stored in all the cells requesting the file given by

$$\rho_{\pi_j}=\prod_{k\in\mathscr{K}_{\pi_j}}\frac{\binom{n_j-1}{m_{k,j}-1}}{\binom{n_j}{m_{k,j}}}=\prod_{k\in\mathscr{K}_{\pi_j}}\frac{m_{k,j}}{n_j}. \tag{6.8}$$

Since $\frac{m_{k,j}}{n_j}\leq 1,\forall k$, it holds true that $\rho_{\pi_j}\leq\min_{k\in\mathscr{K}_{\pi_j}}\frac{m_{k,j}}{n_j}$. Thus, we derive that $R_{\text{multicast}}^{\text{MDS}}\leq R_{\text{multicast}}^{\text{Uncoded}}$. A rigorous proof has been provided in our previous work [154].

6.4.3 Optimization

Defining $q_{k,j} \triangleq \frac{m_{k,j}}{n_j}$ and using (6.5), (6.3) can be recast into

$$\min_{\{q_{k,j}\}} \sum_{j=1}^{N}\sum_{k=1}^{K}\left(1-q_{k,j}\right)s_j(1-\alpha_{k,j})\prod_{t\in\mathscr{T}_{k,j}}\alpha_{t,j} \tag{6.9a}$$

$$\text{s.t.} \sum_{k=1}^{K}\sum_{j=1}^{N} q_{k,j}s_j \le M, \tag{6.9b}$$

$$0 \le q_{k,j} \le 1,\ \forall k,j. \tag{6.9c}$$

Unfortunately, before $\{q_{k,j}\}$ are obtained, it is impossible to know the ranks $\{r_{k,j}\}$, or $\mathscr{T}_{k,j}$. To tackle this, we sort the elements of $\mathbf{q}^j, \forall j$ in an ascending order and define the sorted variables as $\mathbf{g}^j \triangleq [g_{1,j},\ldots,g_{K,j}], \forall j$. To illustrate the relationships between $\mathbf{q}^j$ and $\mathbf{g}^j$, a new matrix $\mathbf{Y} \triangleq [y^k_{t,j}]_{K\times N\times K}$ with $y^k_{t,j} \in \{0,1\}$ is defined such that

$$q_{k,j} = \sum_{t=1}^{K} g_{t,j}y^k_{t,j}. \tag{6.10}$$

If $q_{k,j}$ is the tth lowest in $\mathbf{q}^j$, i.e., $r_{k,j}=t$, we let $y^k_{t,j}=1$ and $y^k_{\bar{t},j}=0, \forall \bar{t}\neq t$. Note that the ranks are assumed to be unique integers even if there are several elements of $\mathbf{q}^j$ equal to each other. The characteristics of $\{y^k_{t,j}\}$ are concluded in the following constraints (6.11e)–(6.11g). Now, (6.9) becomes

$$\min_{\{g_{t,j}\},\{y^k_{t,j}\}} \sum_{j=1}^{N}\sum_{t=1}^{K}\left(1-g_{t,j}\right)s_j\varphi_{t,j} \tag{6.11a}$$

$$\text{s.t.} \sum_{k=1}^{K}\sum_{j=1}^{N}\sum_{t=1}^{K} g_{t,j}y^k_{t,j}s_j \le M, \tag{6.11b}$$

$$g_{t,j} \le g_{t+1,j}, \forall t<K, \text{ and } \forall j, \tag{6.11c}$$

$$0 \le g_{t,j} \le 1, \forall t,j, \tag{6.11d}$$

$$\sum_{t=1}^{K} y^k_{t,j} = 1, \forall k,j, \tag{6.11e}$$

$$\sum_{k=1}^{K} y^k_{t,j} = 1, \forall t,j, \tag{6.11f}$$

$$y^k_{t,j} \in \{0,1\}, \forall t,j,k, \tag{6.11g}$$

where $\varphi_{t,j}$ is the probability that $100g_{t,j}\%$ of file j requires delivery from the MBS via backhaul. Define a new group of variables $\{\sigma_t\}$ satisfying $q_{\sigma_t,j} = g_{t,j}$ as the

indices mapping $g_{t,j}$ to $q_{\sigma_t,j}$. For instance, $\sigma_t = 1$ states that $q_{1,j}$ ranks the tth in $\mathbf{q}^j$, i.e., $q_{1,j} = g_{t,j}$. And we can then obtain the expression of $\varphi_{t,j}$ given by $\varphi_{t,j} = (1-\alpha_{\sigma_t,j})\prod_{v=1}^{t-1}\alpha_{\sigma_v,j}$ based on (9a) and the definition of $\mathbf{g}^j$. Utilizing (10), it holds true that $y_{t,j}^{\sigma_t} = 1$. Hence, $\varphi_{t,j}$ can be further rewritten as

$$\varphi_{t,j} = \left[\sum_{k=1}^{K}\left(1-\alpha_{k,j}\right)y_{t,j}^{k}\right]\prod_{v=1}^{t-1}\left[\sum_{k=1}^{K}(\alpha_{k,j}y_{v,j}^{k})\right], \forall t > 1 \tag{6.12}$$

with $\varphi_{1,j} = \sum_{k=1}^{K}\left(1-\alpha_{k,j}\right)y_{1,j}^{k}$.

Due to the coupling among the variables in the constraints as well as the objective function, (6.11) is a mixed integer nonlinear program (MINLP) and is difficult to deal with. The expression of $\varphi_{t,j}$ also makes it too complex to be linearized. As such, reformulation is done here to simplify the constraints.

Lemma 6.3 *Based on the characteristics of $\{y_{t,j}^k\}$, the overall cache capacity constraint in (6.11b) can be re-expressed as $\sum_{t=1}^{K}\sum_{j=1}^{N} g_{t,j}s_j \le M$. Hence, (6.11) can be rewritten as*

$$\min_{\{g_{k,j}\},\{y_{t,j}^k\}} \quad \sum_{j=1}^{N}\sum_{t=1}^{K}\left(1-g_{t,j}\right)s_j\varphi_{t,j} \tag{6.13a}$$

$$\text{s.t.} \quad \sum_{t=1}^{K}\sum_{j=1}^{N} g_{t,j}s_j \le M, \tag{6.13b}$$

$$\textit{(6.11c)–(6.11g)}, \tag{6.13c}$$

with the optimal allocated cache sizes given by

$$M_k = \sum_{j=1}^{N}\sum_{t=1}^{K} g_{t,j}y_{t,j}^k s_j, \forall k. \tag{6.14}$$

Proof 6.3 *According to (6.10), it can be easily proved that (6.14) holds. Then utilizing the constraint (6.11f), we obtain*

$$\begin{aligned}\sum_{k=1}^{K} M_k &= \sum_{k=1}^{K}\sum_{j=1}^{N}\sum_{t=1}^{K} g_{t,j}y_{t,j}^k s_j, \\ &= \sum_{j=1}^{N}\sum_{t=1}^{K} g_{t,j}\left(\sum_{k=1}^{K} y_{t,j}^k\right)s_j = \sum_{j=1}^{N}\sum_{t=1}^{K} g_{t,j}s_j.\end{aligned} \tag{6.15}$$

Hence, we get (6.13b), which completes the proof.

After utilizing *Lemma 6.3*, $\{g_{t,j}\}$ and $\mathbf{Y}$ are now decoupled in the constraints of (6.13). To proceed, we firstly fix $\{g_{t,j}\}$ and optimize $\mathbf{Y}$. The problem of interest is given by

$$\mathscr{P}(\{g_{t,j}\}): \quad \min_{\{y_{t,j}^k\}} \sum_{j=1}^{N}\sum_{t=1}^{K}\left(1-g_{t,j}\right)s_j\varphi_{t,j} \tag{6.16a}$$

$$\text{s.t.} \quad \text{(6.11e)–(6.11g)}, \tag{6.16b}$$

with $\{g_{t,j}\}$ satisfying (6.11c)–(6.11d) and (6.13b). Obviously, $\{y_{t,j}^k\}$ are independent with each other in different files in problem (6.16). As a result, we can separate the problem into a number of sub-problems with regard to different file j, e.g.,

$$\mathscr{P}_j(\{g_{t,j}\}): \quad \min_{\{y_{t,j}^k\}} \sum_{t=1}^{K}\left(1-g_{t,j}\right)\varphi_{t,j} \tag{6.17a}$$

$$\text{s.t.} \quad \sum_{t=1}^{K} y_{t,j}^k = 1, \forall k, \tag{6.17b}$$

$$\sum_{k=1}^{K} y_{t,j}^k = 1, \forall t, \tag{6.17c}$$

$$y_{t,j}^k \in \{0,1\}, \forall t,k. \tag{6.17d}$$

The coupling and complexity of $\varphi_{t,j}$ makes it intractable to find the optimal $\{y_{t,j}^k\}$ even when $\{g_{t,j}\}$ are given. To tackle this problem, we analyze the impact of $\{y_{t,j}^k\}$ on the objective function based on the characteristics of $\{g_{t,j}\}$ and $\{y_{t,j}^k\}$, and infer the relations among $\{y_{t,j}^k\}$ and the probabilities $\{\alpha^j\}$. For illustrative purposes, we let $\alpha^j \triangleq [\alpha_{1,j}, \alpha_{2,j}, \ldots, \alpha_{K,j}]$, rearrange the elements in α^j in a descending order and define the new vector as $\beta^j \triangleq [\beta_{1,j}, \beta_{2,j}, \ldots, \beta_{K,j}]$. Let $\{\theta_k\}$ reflect the one-to-one correspondence between the elements of β^j and α^j satisfying $\beta_{k,j} = \alpha_{\theta_k,j}, \forall k$. Meanwhile, α^j, β^j, and $\{\theta_k\}$ are all known. The result is given in the following lemma.

Lemma 6.4 *The optimal probability $\varphi_{t,j}^*$ would be $\varphi_{t,j}^* = \left(1-\beta_{t,j}\right)\prod_{v=1}^{t-1}\beta_{v,j}$. Accordingly, the optimal $\{y_{t,j}^k\}$ to problem (6.17) are given by*

$$y_{t,j}^k = \begin{cases} 1, & \text{if } k = \theta_t, \\ 0, & \text{otherwise}. \end{cases} \tag{6.18}$$

Proof 6.4 *See* Appendix F.

Since *Lemma 6.4* holds true for all the files, (6.13) becomes

$$\min_{\{g_{t,j}\}} \sum_{j=1}^{N}\sum_{t=1}^{K}\left(1-g_{t,j}\right)s_j\left(1-\beta_{t,j}\right)\prod_{v=1}^{t-1}\beta_{v,j} \tag{6.19a}$$

$$\text{s.t. (6.13b)–(6.13c)}, \tag{6.19b}$$

which is convex and hence can be easily solved by well known solvers, e.g., CVX [73]. Then substituting (6.18) into (6.14), the optimal cache capacities in each cell can be rewritten as

$$M_k = \sum_{j=1}^{N} g_{t,j}s_j|_{\theta(t)=k}, \forall k, \tag{6.20}$$

with the optimal content placement given by

$$q_{k,j} = g_{t,j}|_{\theta(t)=k}, \forall k, j. \tag{6.21}$$

In the proposed multicast-aware caching scheme, we classify the large number of possible user request profiles into several types according to the values of the associated backhaul load. By doing so, we reduce the computational complexity in terms of user request uncertainty massively from $\mathcal{O}(N^K)$ to $\mathcal{O}(KN)$ to obtain the optimal solution.

6.5 Cooperative Caching

In this section, we consider that the SBSs can fetch content from the neighboring SBSs via some high capacity links and study the optimal cooperative caching policy among the SBSs. Note that the independence amongst the MDS coded packets cached in all the cells almost surely guarantees that the shared contents are always non-overlapping.

6.5.1 Problem Formulation

Cooperative caching consists of three phases:

(i) the content placement phase,

(ii) the content sharing phase among the SBSs, and

(iii) the content delivery phase from the MBS via backhaul.

Note that in the content delivery phase, we assume that unicast is used by the MBS to sent uncached content to the SBSs.

Since backhaul load is unable to provide sufficient insight about the impact of cooperative content sharing on reducing the backhaul requirements, here we utilize user attrition (UA) cost, i.e., the overall cost for fetching content from an external storage, to evaluate the performance of the cooperative caching schemes. To further eliminate the redundancy, we assume that the SBSs can selectively deliver part of the packets from their own caches to the requested SBS rather than the whole of the cached packets. The amounts of shared content among the cooperative SBSs are defined as $\mathbf{X} = \{x^t_{k,j}\}_{K\times N\times K}$ where $x^t_{k,j}$ denotes the number of packets delivered from SBS t to SBS k for file j. Thus, we let f^t_k be the associated unit cost when SBS k fetches unit data (e.g., per MB) from SBS t and f^M_k be the cost for delivering unit data to SBS k from MBS.

The UA costs are modeled as the products of the data loads of the BSs and the associated unit costs [147]. Furthermore, it is assumed that the unit costs are proportional to the square of the minimum distances between the associated BSs with the unit cost coefficients defined as f_0 and f^M_0, respectively, according to [123, 141, 147]. Note that $\{f^t_k\}$ must satisfy the triangle inequality, i.e., $f^t_k \leq f^t_l + f^l_k$, and the cost for fetching content from local storage can be ignored, i.e., $f^k_k = 0, \forall k$. Moreover, the UA costs for fetching content from the MBS via backhaul are usually higher than those caused by the cooperation between the SBSs due to proximity.

Instead of focusing on the backhaul load, our objective here is to minimize the average UA cost, i.e., the cost of fetching content from external storage, subject to a given overall cache capacity constraint by optimizing the cache content placement and cooperation policy jointly. In this case, the expected UA cost defined as $C^{\text{MDS}}_{\text{coop}}$ can be written as

$$C^{\text{MDS}}_{\text{coop}} = \sum_{j=1}^{N}\sum_{k=1}^{K}\left[\left(1-\min\left(1,\sum_{t=1}^{K}\frac{x^t_{k,j}}{n_j}\right)\right)f^M_k + \sum_{t=1}^{K}\frac{x^t_{k,j}}{n_j}f^t_k\right]s_j(1-\alpha_{k,j}). \quad (6.22)$$

Hence, the problem of interest is given by

$$\min_{\{m_{k,j}\},\{x^t_{k,j}\}} \quad C^{\text{MDS}}_{\text{coop}} \tag{6.23a}$$

$$\text{s.t.} \quad \sum_{k=1}^{K}\sum_{j=1}^{N} \frac{m_{k,j}}{n_j} s_j \leq M, \tag{6.23b}$$

$$0 \leq m_{k,j} \leq n_j, \ \forall k, j, \tag{6.23c}$$

$$0 \leq x^t_{k,j} \leq m_{t,j}, \ \forall k, j, t, \tag{6.23d}$$

where the cache size allocation problem is merged into the optimization of the content placement as mentioned in *Lemma 6.3*. Apparently, $x^k_{k,j} = m_{k,j}, \forall k, j$ holds true in (6.23).

6.5.2 Comparison

The significance of adopting MDS codes is to avoid content overlap among the fragments stored in different caches, hence reducing the average UA cost. Suppose that SBS k stores $m_{k,j}$ different fragments randomly drawn among the n_j fragments and $x^k_{t,j}$ of the $m_{k,j}$ fragments are randomly selected to be sent to SBS t. It is difficult to ensure that the fragments from the neighboring cells are always mutually exclusive. Thus, both the number of fragments stored in local cache and sent to other cells and which fragments being cached and shared contribute in deciding the backhaul rate and the average UA cost.

Lemma 6.5 *Given any cooperative caching policy satisfying constraints (6.23b)–(6.23d), the UA cost in the coded scenario is always lower than the associated cost in the uncoded scenario defined as* $C^{\text{uncoded}}_{\text{coop}}$*, i.e.,* $C^{\text{MDS}}_{\text{coop}} \leq C^{\text{uncoded}}_{\text{coop}}$.

Proof 6.5 *See* Appendix G.

6.5.3 Optimization

We can tackle (6.23) by proving that the optimal cooperative caching policy always satisfies $\sum_{t=1}^{K} \frac{x^t_{k,j}}{n_j} \leq 1, \forall k, j$. Letting $(\{\tilde{x}^t_{k,j}\}, \{\tilde{m}_{k,j}\})$ be the optimal solution to (6.23) with at least a group of (k^*, j^*) satisfying $\sum_{t=1}^{K} \frac{\tilde{x}^t_{k^*,j^*}}{n_j} > 1$, we can always find some

$(\{x^t_{k,j}\},\{\tilde{m}_{k,j}\})$ with $x^t_{k,j}=\tilde{x}^t_{k,j},\forall(k,j,t)\neq(k^*,j^*,t)$ and $\sum_{t=1}^{K}\frac{\tilde{x}^t_{k^*,j^*}}{n_j}=1$ which satisfy all the constraints in (6.23) while demanding the same cost from backhaul but a lower cost from content sharing among the cooperative SBSs. Consequently, the average UA cost is given by

$$C^{\text{MDS}}_{\text{coop}}=\sum_{j=1}^{N}\sum_{k=1}^{K}\left[\left(1-\sum_{t=1}^{K}z^t_{k,j}\right)f^M_k+\sum_{t=1}^{K}z^t_{k,j}f^t_k\right]\times s_j(1-\alpha_{k,j}), \tag{6.24}$$

where we let $q_{k,j}=\frac{m_{k,j}}{n_j}$ and $z^t_{k,j}=\frac{x^t_{k,j}}{n_j}$. Problem (6.23) can then be rewritten as

$$\min_{\{q_{k,j}\},\{z^t_{k,j}\}} \quad (6.24) \tag{6.25a}$$

$$\text{s.t.} \quad \sum_{k=1}^{K}\sum_{j=1}^{N}q_{k,j}s_j\leq M, \tag{6.25b}$$

$$0\leq q_{k,j}\leq 1,\ \forall k,j, \tag{6.25c}$$

$$\sum_{t=1}^{K}z^t_{k,j}\leq 1,\ \forall k,j, \tag{6.25d}$$

$$0\leq z^t_{k,j}\leq q_{t,j},\ \forall k,j,t, \tag{6.25e}$$

which is linear and can easily be solved using, e.g., CVX.

For comparison, the average UA cost in the unicast based non-cooperative caching scenario is given by

$$C^{\text{unicast}}_{\text{noncoop}}=\sum_{j=1}^{N}\sum_{k=1}^{K}\left(1-q_{k,j}\right)f^M_k s_j(1-\alpha_{k,j}). \tag{6.26}$$

As $f^t_k\leq f^M_k$ and $z^k_{k,j}=q_{k,j},\forall k,t,j$, we have

$$C^{\text{MDS}}_{\text{coop}}\leq\sum_{j=1}^{N}\sum_{k=1}^{K}\left(1-\sum_{t=1}^{K}z^t_{k,j}+\sum_{t\neq k}z^t_{k,j}\right)\times f^M_k s_j(1-\alpha_{k,j})\leq C^{\text{unicast}}_{\text{noncoop}}. \tag{6.27}$$

6.6 Multicast-Aware Cooperative Caching

In this section, a compound caching policy named multicast-aware cooperative caching is proposed to take the advantages of both multicasting at the MBS and collaboration among the SBSs. Global optimal caching scheme is proposed for small scale networks followed by the multicast-aware in-cluster cooperative caching scheme developed particularly for the large scale networks.

6.6.1 Small Scale Networks

Lemma 6.6 *In case of multicast-aware cooperative caching, the UA cost can be written as*

$$C_{\text{mult,coop}}^{\text{MDS}} = \sum_{j=1}^{N} \left[\sum_{\pi_j \in \Pi_j} \left(1 - \min_{k \in \mathscr{K}_{\pi_j}} \sum_{t=1}^{K} z_{k,j}^t \right) \max_{k \in \mathscr{K}_{\pi_j}} f_k^M \times P_r(\pi_j) + \sum_{k=1}^{K} \sum_{t=1}^{K} z_{k,j}^t f_k^t (1 - \alpha_{k,j}) \right] s_j. \quad (6.28)$$

Proof 6.6 *See* Appendix H.

The average UA cost minimization problem is

$$\min_{\{q_{k,j}\},\{z_{k,j}^t\}} C_{\text{mult,coop}}^{\text{MDS}} \text{ s.t. (6.25b)–(6.25e).} \quad (6.29)$$

We recognize that similar content in different cells is preferred for multicast-aware caching while for cooperative caching the cached content in different cells should be mutually exclusive. The use of MDS codes strikes a balance in the combination. It is worth pointing out that multicast-aware cooperative caching brings additional multicast gain in most cases in terms of minimizing the long term average UA cost considering the large numbers of BSs, files, and user request profiles while unicast content delivery might only be preferred in rare extreme cases, e.g., when only a few cells with steeply graded unit costs require the same file. To eliminate the impact of these special cases, a new group of binary variable can be introduced to identify which content delivery strategy is preferred for each user request profile in the case of small scale networks.

Lemma 6.7 *Given any mutlicast-aware cooperative caching policy* $(\{q_{k,j}\},\{z_{k,j}^t\})$ *satisfying the constraints in (6.29), the UA cost in the coded scenario is always much lower than that in the uncoded case, i.e.,* $C_{\text{mult,coop}}^{\text{MDS}} \leq C_{\text{mult,coop}}^{\text{uncoded}}$.

Proof 6.7 *See* Appendix I.

To solve (6.29), we resort to a greedy algorithm by listing all possible user request profiles for each file. Furthermore, a number of new variables and constraints need to be added to linearize the function $\min(\cdot)$. That is, for any user

request profile π_j, we introduce a new variable ξ_{π_j} subject to the constraints, i.e., $(0 \leq \xi_{\pi_j} \leq \sum_{t=1}^{K} z_{k,j}^t, \forall k \in \mathscr{K}_{\pi_j})$, to replace $\min_{k\in\mathscr{K}_{\pi_j}} \sum_{t=1}^{K} z_{k,j}^t$ in (6.28). Since (6.29) can be linearized, general solvers can be employed to solve it for small-scale networks. However, in practical scenarios with dozens of BSs and thousands of files, the greedy approach is not viable.

6.6.2 Large Scale Networks

In order to reduce the complexity in large scale networks, we propose a multicast-aware in-cluster cooperative caching scheme by decomposing a macro cell into a series of annular regions $\{\mathscr{C}^u, \forall u \in [1,U]\}$ with their radii between $R_u \pm \Delta R_u (\Delta R_u \ll R_u)$. In each annulus, the neighboring SBSs form a number of disjoint clusters defined as $\{\mathscr{S}_1^u, \mathscr{S}_2^u, \ldots, \mathscr{S}_{L_u}^u\}$ where L_u is the number of clusters in the uth annulus. Let $|\mathscr{S}_l^u|$ denote the number of SBSs in cluster $\mathscr{S}_l^u$. It is assumed that the SBSs in the same cluster $\mathscr{S}_l^u$ can share content over high capacity links with a cost $f_l^u = f_0 \overline{d}_l^u$ where $\overline{d}_l^u$ is the average of the squares of the distances among the cooperative SBSs. The cost for retrieving content from the MBS is $f_u^M = f_0^M R_u^2$ where R_u is the radius for the uth annulus. The UA cost in cluster $\mathscr{S}_l^u$ is given by

$$C_l^u = \sum_{j=1}^{N} \left[\sum_{\pi_{l,j}^u \in \Pi_{l,j}^u} \left(1 - \min_{k\in\mathscr{K}_{\pi_{l,j}^u}} \sum_{t\in\mathscr{S}_l^u} z_{k,j}^t \right) f_u^M P_r(\pi_{l,j}^u) + \sum_{k\in\mathscr{S}_l^u} \sum_{t\in\mathscr{S}_l^u \setminus k} z_{k,j}^t f_l^u (1-\alpha_{k,j}) \right] s_j. \quad (6.30)$$

Therefore, this scheme solves

$$\min_{\{q_{k,j}\},\{z_{t,j}^k\}} \sum_u \sum_l C_l^u \quad (6.31a)$$

$$\text{s.t.} \sum_{t\in\mathscr{S}_l^u} z_{k,j}^t \leq 1, \ \forall k \in \mathscr{S}_l^u, \forall j, \forall l, \forall u, \quad (6.31b)$$

$$0 \leq z_{k,j}^t \leq q_{t,j}, \forall t, k \in \mathscr{S}_l^u, \forall j, \forall l, \forall u, \quad (6.31c)$$

$$0 \leq q_{k,j} \leq 1, \forall k \in \mathscr{S}_l^u, \forall j, \forall l, \forall u, \quad (6.31d)$$

$$\sum_u \sum_l \sum_j \sum_{k\in\mathscr{S}_l^u} q_{k,j} s_j \leq M. \quad (6.31e)$$

For the sake of mathematical tractability, we decompose the problem into a number of sub-problems each minimizing the UA cost for a cluster. In this case, we let $q_{k,j} = q^u_{l,j}, \forall k \in \mathscr{S}^u_l, \forall j, l, u$ and the sub-problem for cluster $\mathscr{S}^u_l$ is given by

$$\mathscr{P}(\{q^u_{l,j}\}) : \min_{\{z^k_{t,j}\}} \; C^u_l \tag{6.32a}$$

$$\text{s.t.} \sum_{t \in \mathscr{S}^u_l} z^t_{k,j} \leq 1, \; \forall k \in \mathscr{S}^u_l, \forall j, \tag{6.32b}$$

$$0 \leq z^t_{k,j} \leq q^u_{l,j}, \; \forall t, k \in \mathscr{S}^u_l, \forall j. \tag{6.32c}$$

Because the cost for fetching content from local cache can be ignored, it holds true that $z^k_{k,j} = q^u_{l,j}, \forall k \in \mathscr{S}^u_l$. For any given cache composition satisfying the constraints (6.31b)–(6.31e), we find it important to understand the volume of content that is needed to be fetched from the MBS via backhaul. Let $D^u_l = \sum_{j=1}^{N} \sum_{k \in \mathscr{S}^u_l} q^u_{l,j} s_j (1 - \alpha_{k,j})$. Given cache composition, D^u_l is always constant and hence can be ignored. The objective function can then be further reformulated into

$$\widetilde{C}^u_l = C^u_l + D^u_l = \sum_{j=1}^{N} \left[\sum_{\pi^u_{l,j} \in \Pi^u_{l,j}} \left(1 - \min_{k \in \mathscr{K}_{\pi^u_{l,j}}} \lambda_{k,j} \right) \right.$$
$$\left. \times f^M_u P_r(\pi^u_{l,j}) + \sum_{k \in \mathscr{S}^u_l} \lambda_{k,j} f^u_l (1 - \alpha_{k,j}) \right] s_j. \tag{6.33}$$

where $\lambda_{k,j} = \sum_{t \in \mathscr{S}^u_l} z^t_{k,j}$ denotes the percentage of file j accessible to SBS k within the cluster and is subject to

$$0 \leq \lambda_{k,j} \leq 1, \; \forall k \in \mathscr{S}^u_l, \forall j, \tag{6.34}$$

$$q^u_{l,j} \leq \lambda_{k,j} \leq |\mathscr{S}^u_l| q^u_{l,j}, \; \forall t, k \in \mathscr{S}^u_l, \forall j. \tag{6.35}$$

Note that with the assumption of homogeneous content placement in the SBSs in the same cluster, this gives the overall percentage of a certain file j SBS k gets access to, i.e., $\lambda_{k,j}$. In the following, we focus on obtaining the optimal values of $\{\lambda_{k,j}\}$. Similar to the multicast-aware caching scenario, the objective function can

be rewritten as

$$\widetilde{C}_l^u = \sum_{j=1}^{N} \sum_{k \in \mathscr{S}_l^u} \Bigg[\left(1-\lambda_{k,j}\right) f_u^M (1-\alpha_{k,j}) \prod_{t \in \mathscr{T}_{k,j}} \alpha_{t,j} + \lambda_{k,j} f_l^u (1-\alpha_{k,j}) \Big] s_j, \quad (6.36)$$

where $\mathscr{T}_{k,j}$ is the set of cells satisfying $\mathscr{T}_{k,j} = \{t | r_{t,j} < r_{k,j}\}$ as in *Lemma 6.2*. In this case, we manage to obtain the actual relation amongst $\lambda_{k,j}, \forall k \in \mathscr{S}_l^u$ in the following lemma.

Lemma 6.8 *Given any homogeneous cache decomposition in cluster $\mathscr{S}_l^u$, it holds true that the optimal percentages for file j accessible to the SBSs within the cluster either at local cache or from the cooperative SBSs are always the same regardless of the distinct probabilities for file j being requested by users in different cells, i.e., $\lambda_{k,j} = \lambda_{t,j}, \forall k,t \in \mathscr{S}_l^u$.*

Proof 6.8 *See* Appendix J.

According to *Lemma 6.8*, we let $\lambda_{k,j} = \lambda_{l,j}^u, \forall k \in \mathscr{S}_l^u$. The associated UA cost in (6.30) can be rewritten as

$$C_l^u = \sum_j \left(1-\lambda_{l,j}^u\right) f_u^M \omega_{l,j}^u s_j + \sum_{j=1} \sum_{k \in \mathscr{S}_l^u} \left(\lambda_{k,j} - q_{l,j}^u\right) \times f_l^u (1-\alpha_{k,j}) s_j, \quad (6.37)$$

where $\omega_{l,j}^u$ is the probability for file j being requested by any of the users served by the SBSs in the cluster $\mathscr{S}_l^u$ given by

$$\omega_{l,j}^u = 1 - \prod_{k \in \mathscr{S}_l^u} \alpha_{k,j}, \forall j,l,u. \quad (6.38)$$

Therefore, (6.31) can then be recast into

$$\min_{\{q_{l,j}^u\},\{\lambda_{l,j}^u\}} \quad \sum_u \sum_l C_l^u \quad (6.39a)$$

$$\text{s.t.} \quad 0 \le \lambda_{l,j}^u \le 1, \ \forall j,\forall l,\forall u, \quad (6.39b)$$

$$q_{l,j}^u \le \lambda_{l,j}^u \le |\mathscr{S}_l^u| q_{l,j}^u, \ \forall j,\forall l,\forall u. \quad (6.39c)$$

$$0 \le q_{l,j}^u \le 1, \forall j,\forall l,\forall u, \quad (6.39d)$$

$$\sum_u \sum_l \sum_j q_{l,j}^u s_j \le M. \quad (6.39e)$$

The problem is now linear with smaller sets of variables and constraints and can be solved by well-known solvers.

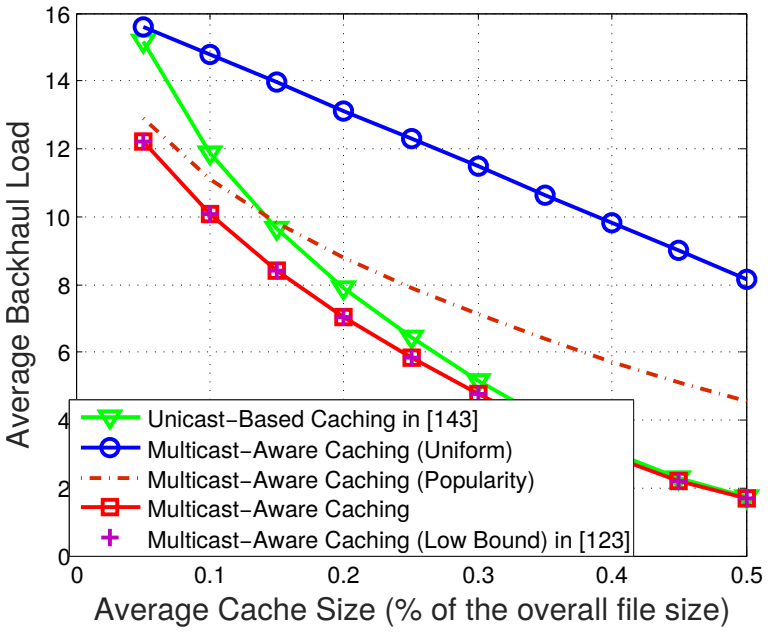
6.7 Simulation Results

Here, we evaluate the performances of the proposed coded caching schemes in terms of the average backhaul load as well as the UA cost via computer simulations. A typical small cell network with $K = 10$ cells and $N = 100$ files is considered for the evaluation of multicast-aware caching scheme and the overall cooperative caching schemes while a large scale network with $K = 28, N = 1000$ is considered for in-cluster cooperative caching schemes. The MBS is located at the center of the macro cell with radius $R = 400$km while the SBSs are randomly deployed uniformly within the cell without coverage overlapping. To show clearly the capabilities for the SBSs to accommodate the files, the overall cache capacity budget is presented as the average cache size for each SBS scaled by the overall file size given by $\rho = M/K/\sum_j s_j$. Unless otherwise specified, we set $\rho = 0.25$ for multicast-aware caching and in-cluster caching schemes while $\rho = 0.05$ is assumed for overall cooperative caching schemes to ensure the participation of backhaul in content delivery. The file sizes are randomly chosen uniformly within $[0, 500]$MB. The skewness parameters $\{\gamma_k\}$ are selected randomly within $[0, 2]$ while the popularity ranks of the files in each cell are generated randomly. Also, the number of users in each cell is set to be ranged within $[0, 10]$, respectively. For cooperative caching, the neighboring SBSs are linked when the distances between them are less than a given threshold. Here, we consider that two SBSs can share content in their caches when the cost for retrieving content from the other SBS is lower than that of fetching content from the MBS. The unit cost coefficients for the two routes for fetching content from external storage are set as $f_0^M = 2$ and $f_0 = 1$.

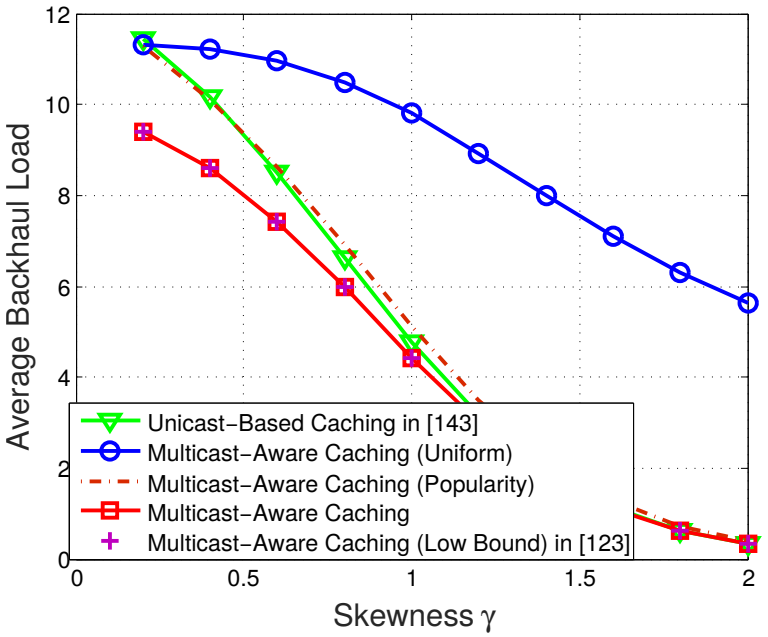
Below describes all the considered schemes.

- **Unicast-Based Caching (Non-Cooperative Caching)**: This is the unicast-based non-cooperative caching scheme with optimal cache management [143].

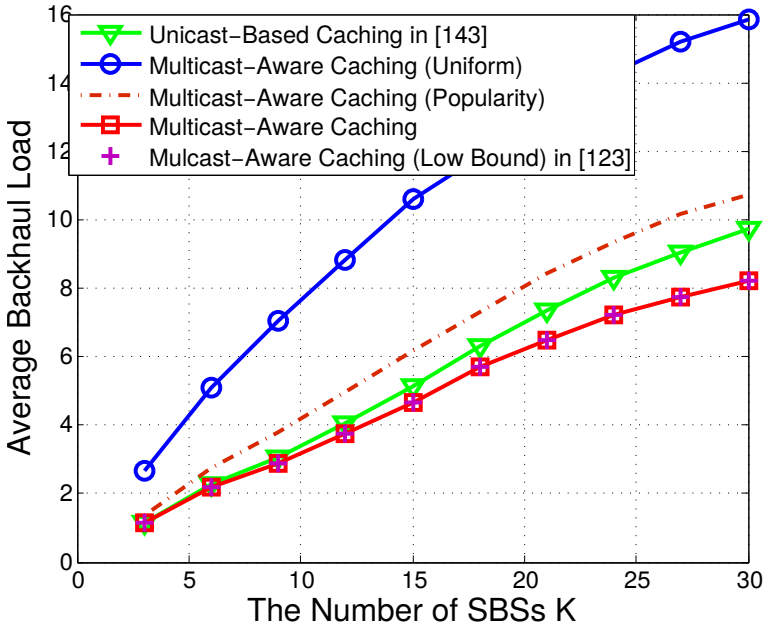
- **Multicast-Aware Caching (Uniform)**: This scheme performs multicast-aware caching with uniform cache size allocation and content placement.
- **Multicast-Aware Caching (Popularity)**: This is same as above except with popularity based content placement.
- **Multicast-Aware Caching**: This refers to our *proposed* multicast-aware caching scheme with optimal cache content placement.
- **Multicast-Aware Caching (Low Bound)**: This refers to the method with optimal cache size allocation and content placement of the linear relaxed multicast aware uncoded caching problem in [123]. Notice that this is practically impossible and only serves as a lower bound.
- **Cooperative Caching**: This refers to our *proposed* unicast-based cooperative caching scheme with optimal cache management and cooperation policy.
- **Multicast-Aware Cooperative Caching (Uniform)**: This is the multicast-aware cooperative caching scheme that uses uniform cache size allocation and content placement.
- **Multicast-Aware Cooperative Caching (Popularity)**: Same as above except with popularity content placement.
- **Multicast-Aware Cooperative Caching**: This refers to our *proposed* multicast-aware cooperative caching with optimal cache management and cooperation policy.
- **In-Cluster Cooperative Caching**: This scheme is similar to cooperative caching except that cooperation is enabled among the SBSs in the same clusters.
- **Multicast-Aware In-Cluster Cooperative Caching**: This scheme is similar to multicast-aware cooperative caching except that multicasting and cooperation are enabled among the SBSs in the same clusters.



(a) Impact of overall cache size M.



(b) Impact of skewness γ.



(a) Impact of number of SBSs K.

Figure 6.2: The average backhaul rate of the proposed multicast-aware caching scheme versus the unicast based caching scheme and the multicast-aware caching schemes.

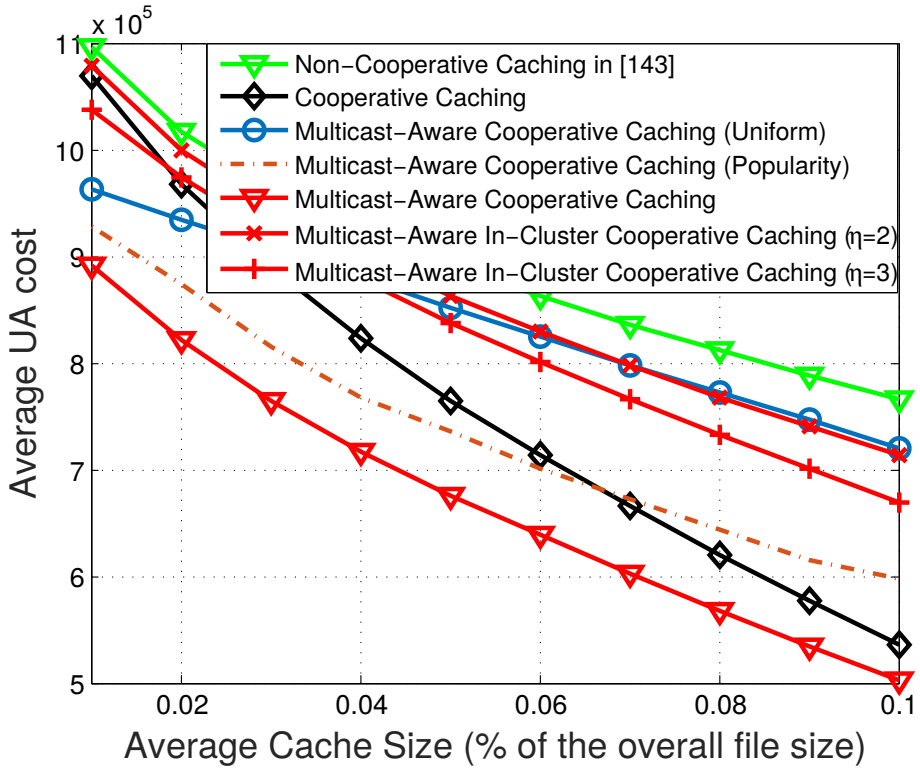
6.7.1 Multicast-aware caching

Results in Fig. 6.2 are provided for the proposed multicast-aware caching scheme, with different content placements, and compared with the uniform based caching scheme. Moreover, the impacts of different parameters and file profile are investigated. As can be seen in Fig. 6.2a, the increase of overall cache size budget leads to a decrease in backhaul rates in all the cases. Also, the proposed multicast-aware caching scheme with optimal content placement, which reaches the low bound of the multicast aware uncoded caching scheme in [123] using linear relaxation and optimal cache management at much lower commuting complexity, shows apparent advantages over the unicast based scheme as expected while the multicast-aware caching schemes with uniform and popularity based content placement show worse performances due to the naive cache management, confirming the significance of multicast transmission in content delivery as well as the centralized cache management in heterogeneous small cell networks. Similar results can be observed in

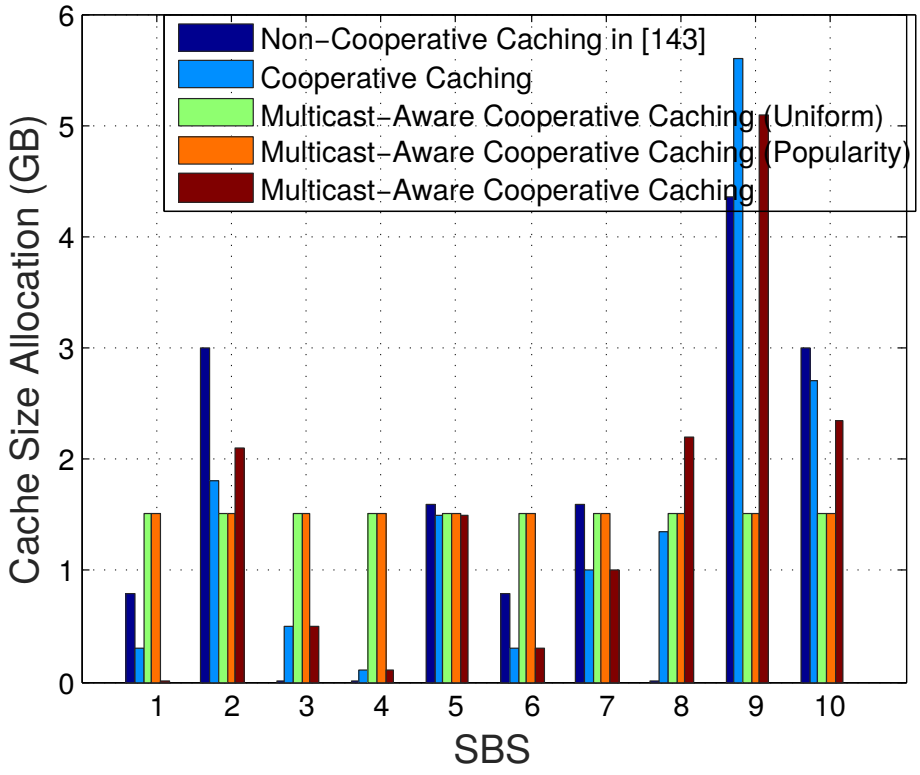
Fig. 6.2b against the skewness parameter of the Zipf's distribution, with $\gamma = \gamma_k, \forall k$ and distinct popularity ranks for the files in different cells. The impact of the number of BSs on the backhaul rate is shown in Fig. 6.2a where the gain improves in denser networks. Again, the multicast-aware scheme outperforms other caching schemes.

6.7.2 Cooperative caching (unicast and multicast)

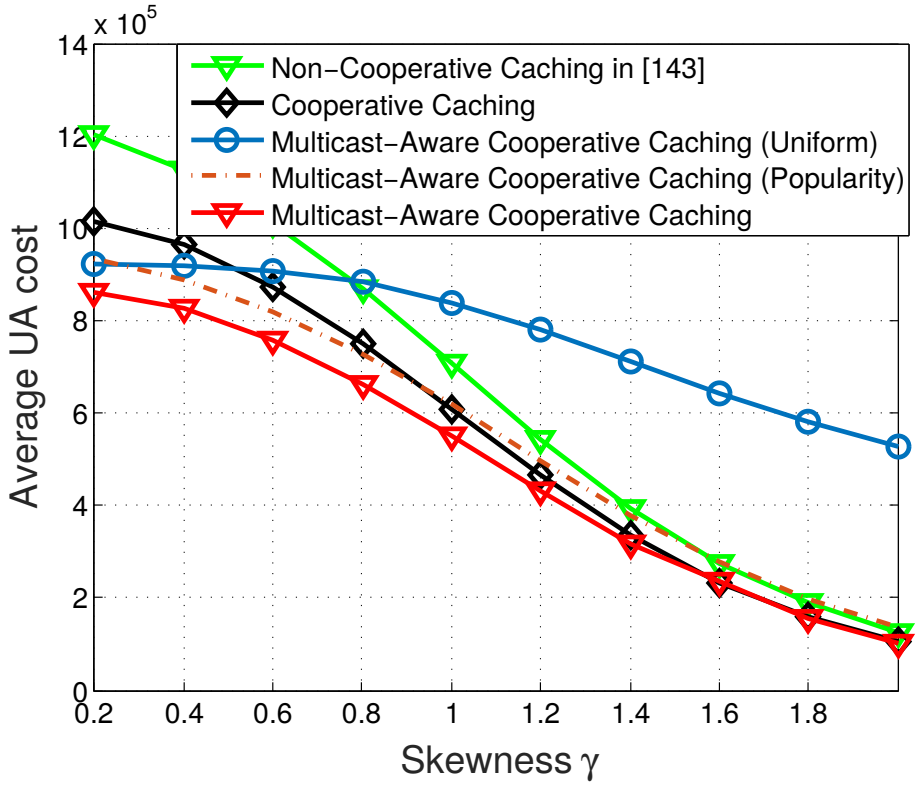
Results in Fig. 6.3 compare the performance of the proposed cooperative caching schemes with that of the non-cooperative scheme in terms of the average UA cost. As can be observed, the proposed multicast-aware cooperative caching scheme shows the best performances followed by the unicast based cooperative caching scheme while the non-cooperative caching scheme yields the worst performance in all the cases. In addition, the multicast-aware cooperative caching schemes using common content placement demand higher UA costs compared with the proposed optimal multicast-aware caching scheme as expected. As we see in Fig. 6.3a, the UA costs decrease with the overall cache size in all cases. Apparently, the utility of cooperation in caching and multicast-aware caching reduce the average UA cost in the network dramatically. For comparison, we also present the results of multicast-aware in-cluster cooperative caching scheme with the maximum cluster size, i.e., the maximum number of SBSs in the clusters defined as η, equal to 2 and 3, respectively. Though the in-cluster caching scheme causes certain performance loss compared with the overall cooperative caching schemes, it largely reduces the computational complexity which makes it suitable for large-scale networks where overall cooperative caching schemes are unviable. Moreover, we can see in the figure that the performance gap can be narrowed by increasing the maximum cluster size η. Fig. 6.3b presents the cache size allocation among the SBSs using different caching schemes when $\rho = 0.05$. Results show that the optimal cache sizes for different cells are always heterogeneous as opposed to the assumption of uniform cache size allocation in many caching networks. Similar conclusions on the impacts of the skewness and the number of users to the non-cooperative case mentioned above can be drawn from Figs. 6.3a and 6.3b. Next, Figs. 6.3a and 6.3b investigate the impacts



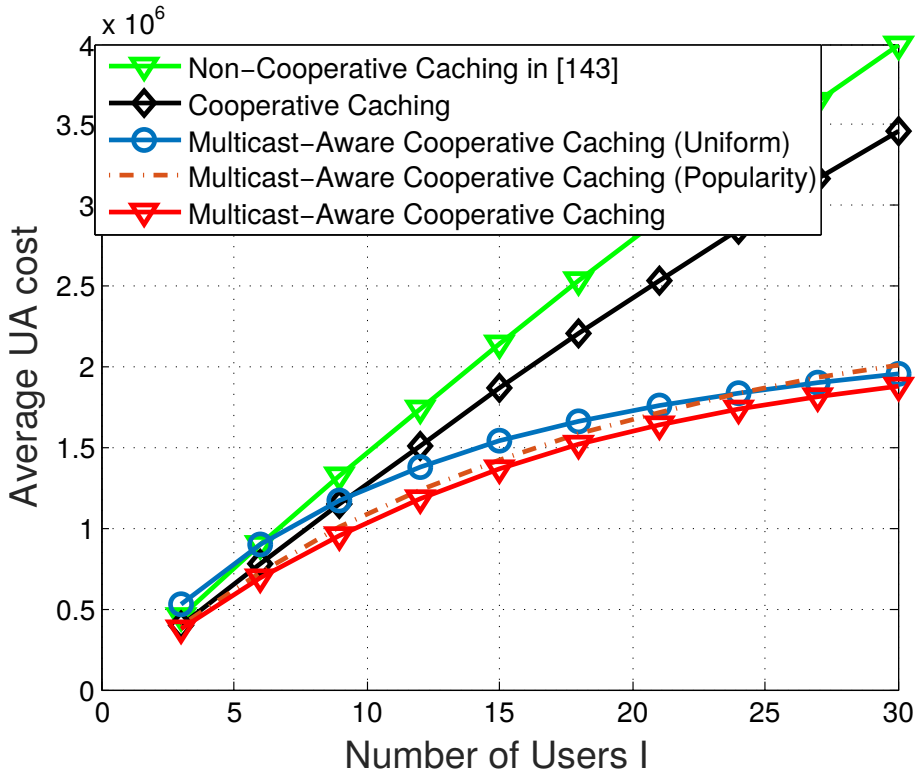
(a) Impact of overall cache size M.



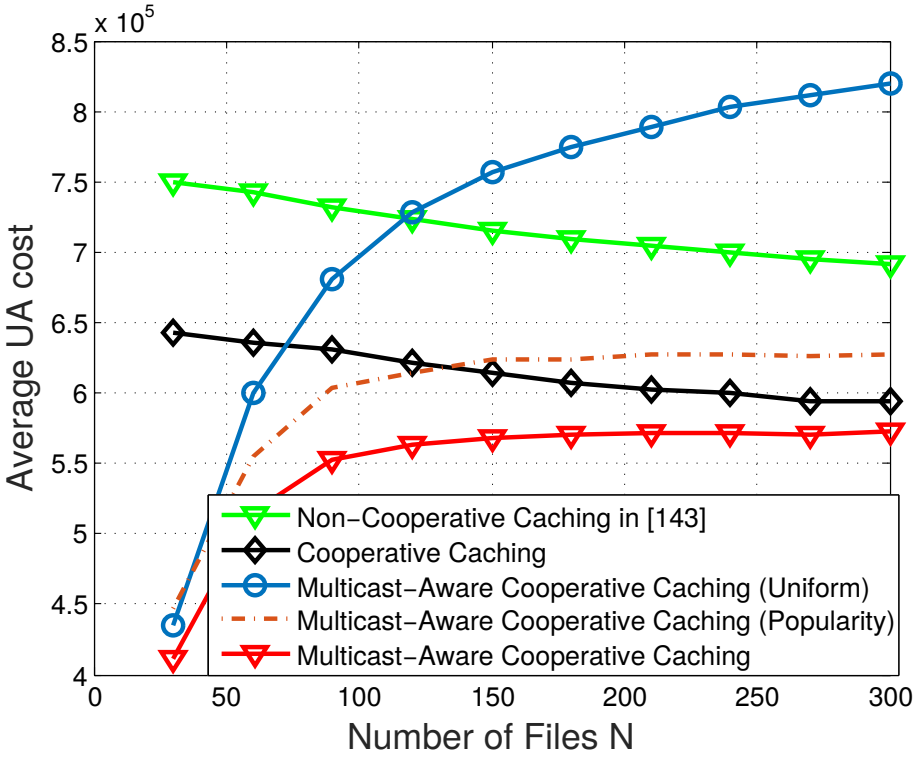
(b) Allocated cache sizes ($\rho = 0.05$).



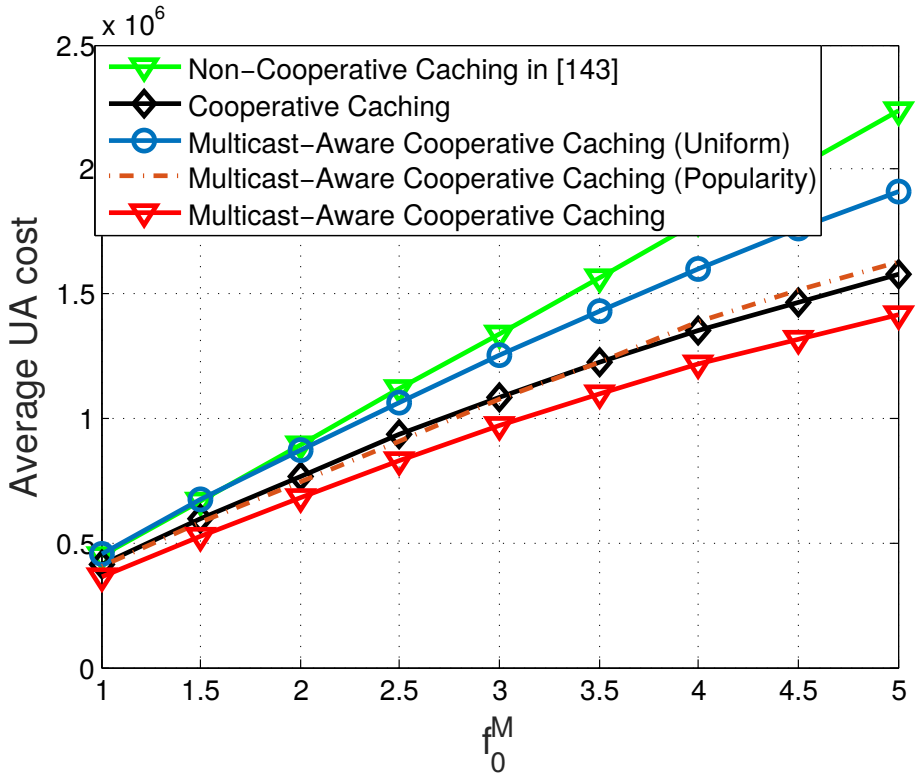
(a) Impact of skewness γ.



(b) Impact of number of users I.



(a) Impact of number of files N.



(b) Impact of cost coefficient f_0^M.

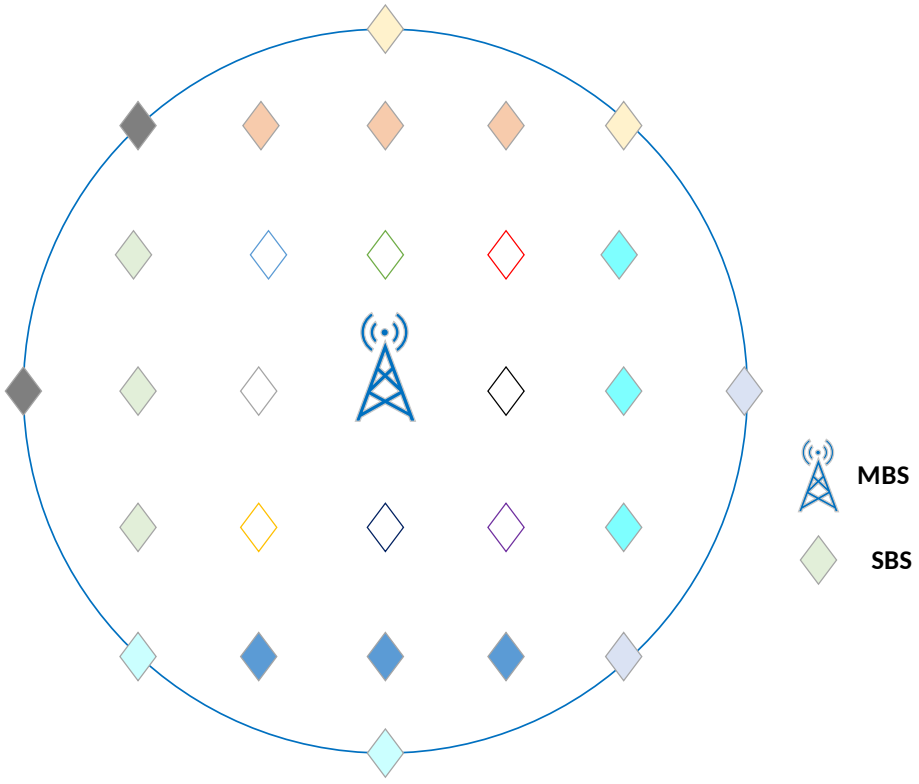
Figure 6.3: The average UA cost of the proposed cooperative caching schemes versus the non-cooperative scheme.

of the number of files and the cost coefficient f_0^M. As we can see in Fig. 6.3a,the UA cost reduction of the proposed multicast-aware cooperative caching scheme decreases with the number of files when $\rho = 0.05$ and $s_j = 250\text{MB}, \forall j$ to unicast based cooperative scheme. Finally, the impact of the ratio between the unit cost coefficients is studied in Fig. 6.3b where $f_0 = 1$ but f_0^M varies. Apparently, the UA cost of the non-cooperative caching scheme is proportional to f_0^M while the cooperative schemes have much better tolerance towards the increase of f_0^M for fetching content via backhaul.

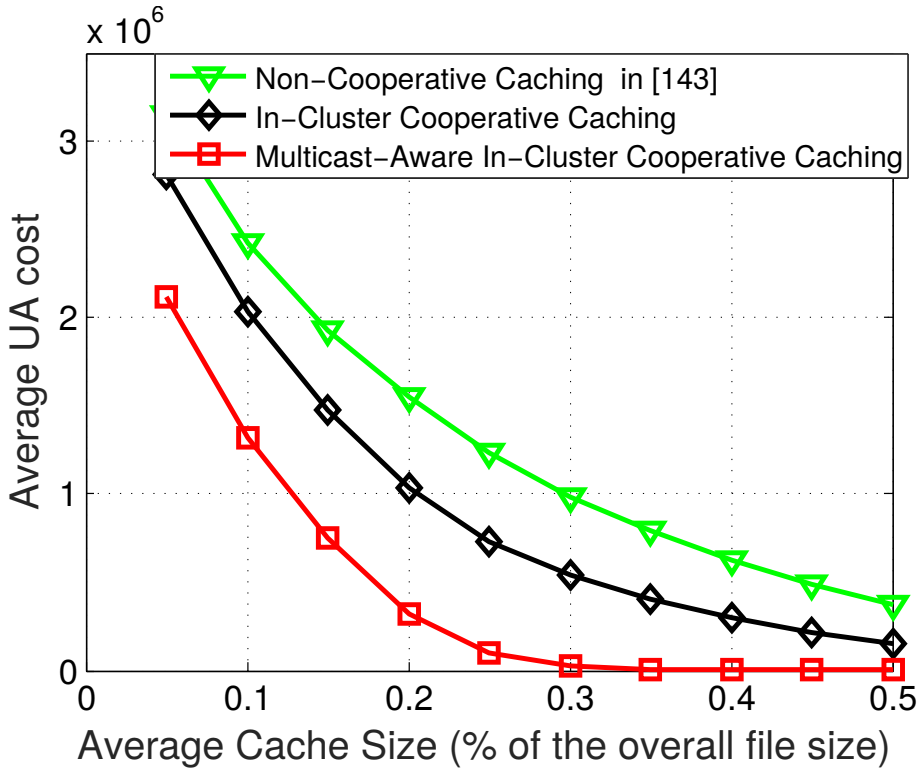
6.7.3 Multicast-aware and in-cluster cooperative caching

Now, a large-scale small cell network with $K = 28$ cells and $N = 1000$ files is considered where the greedy algorithm for multicast-aware cooperative caching scenario is no longer efficient due to high computational complexity and hence in-cluster cooperative caching schemes are considered. Here we assume typical grid deployment of the SBSs as depicted in Fig. 6.4a. The MBS is located at the center of the macro cell with radius $R = 400\text{km}$ and the distance between any two of the neighboring SBSs is fixed at $d = R/3$. The SBSs are divided into 4 annuli based on the distances and then the neighboring SBSs in each annulus are allocated into a number of disjoint clusters where the SBSs in the same color form a cluster. Unless stated otherwise, same parameters as before are used.

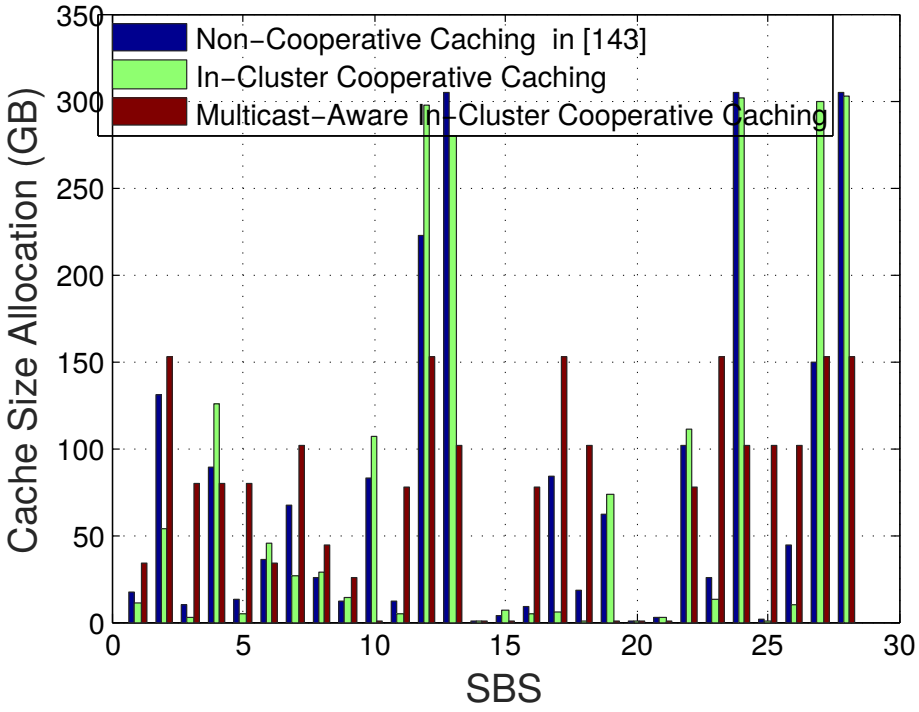
Results for the multicast-aware in-cluster caching scheme are provided in Fig. 6.4. We see that the multicast-aware in-cluster cooperative caching scheme achieves the best UA cost performance followed by the in-cluster cooperative caching scheme while the non-cooperative caching scheme gives the highest UA cost. Compared with that in small scale networks, the UA cost reduction becomes more obvious. The reason may be that the network topologies are different and denser which gives rise to larger number of clusters and the average cluster size than those in the previous scenarios.



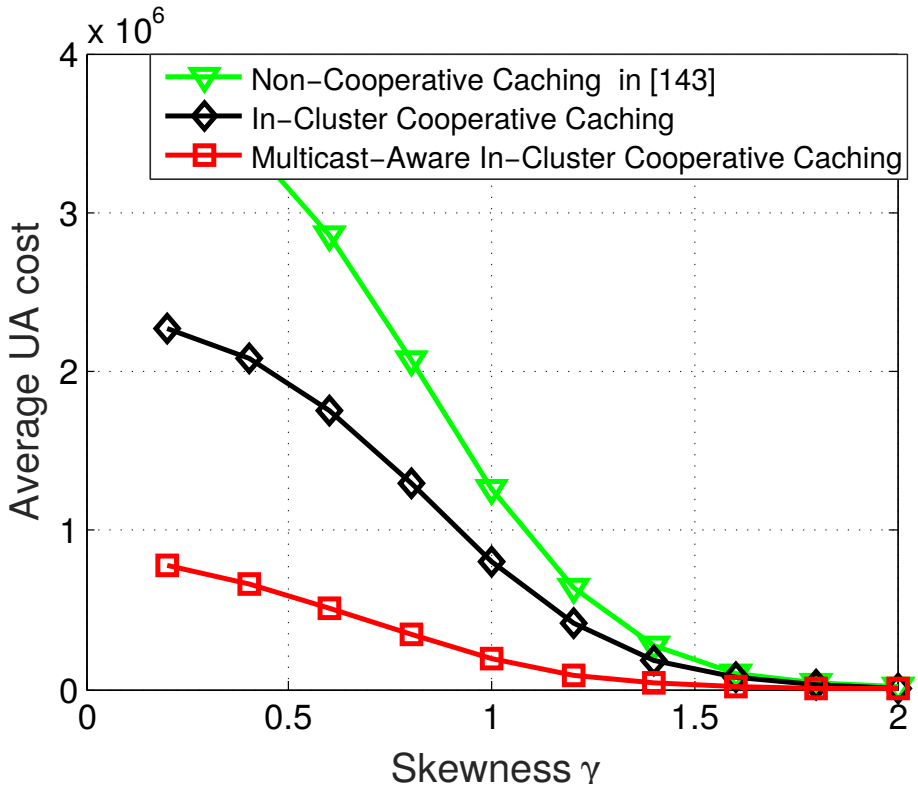
(a) Base Station Deployment.



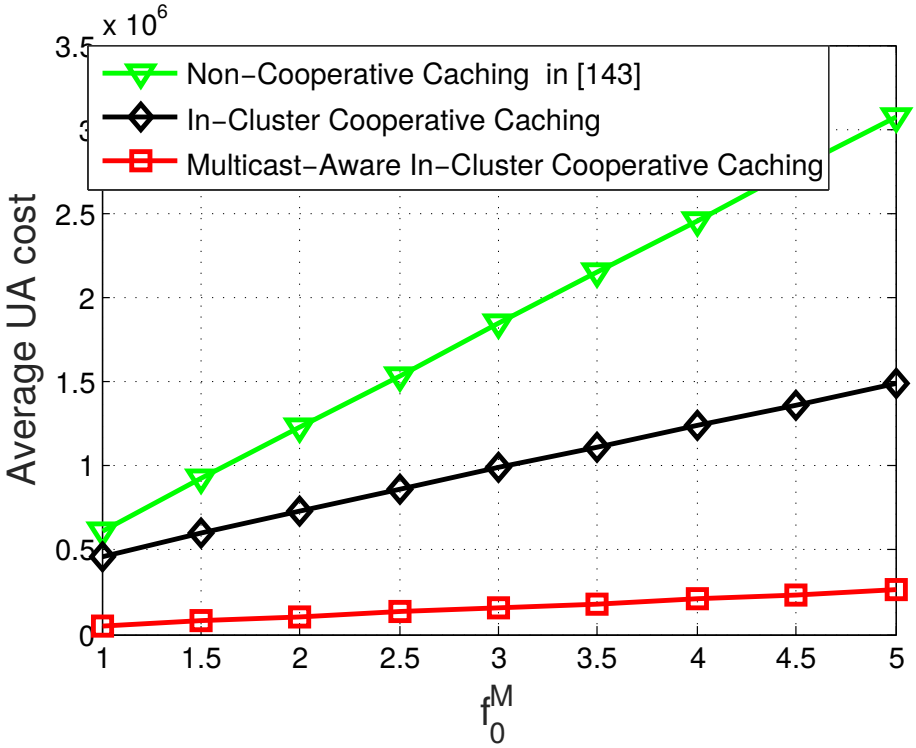
(b) Impact of overall cache size M.



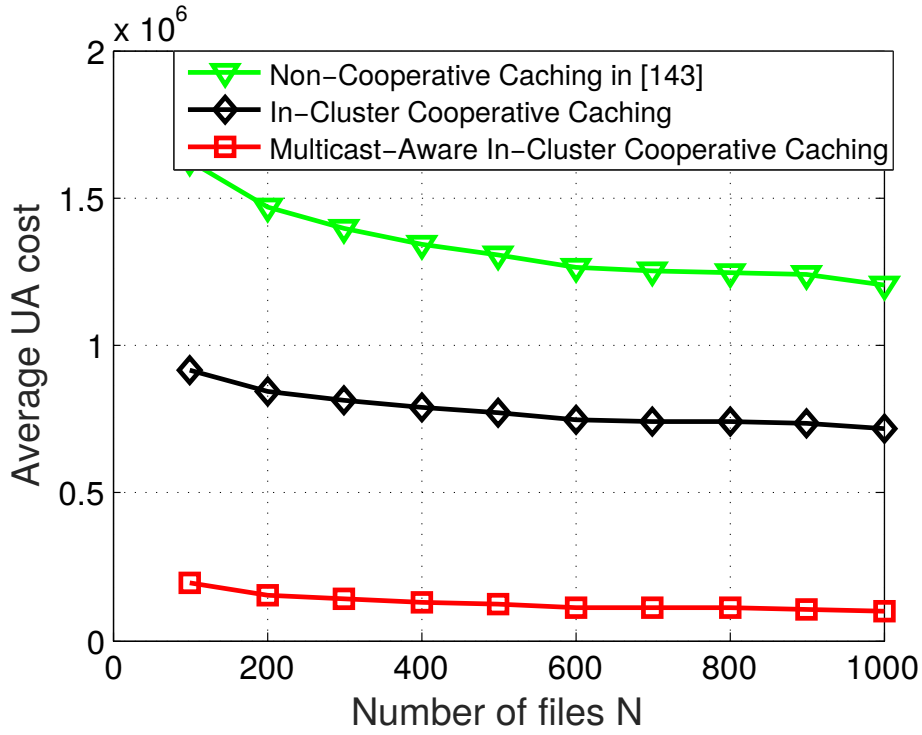
(a) Allocated cache sizes ($\rho = 0.25$).



(b) Impact of skewness γ.



(a) Impact of cost coefficient f_0^M.



(b) Impact of number of files N.

Figure 6.4: The average UA cost of the proposed multicast-aware in-cluster cooperative caching scheme versus in-cluster cooperative caching scheme and non-cooperative caching scheme.

6.8 Summary

In this chapter, we considered the design of content caching and sharing for cache-enabled heterogeneous small cell networks using MDS codes under heterogeneous file and network settings. We first presented two coded caching schemes, dubbed as the multicast-aware caching and the cooperative caching schemes, for minimizing the long-term average backhaul load or the UA cost subject to the overall cache capacity constraint. In both cases, we have obtained the optimal content placement by reformulating the original problems into convex ones. A compound caching scheme, referred to as multicast-aware cooperative caching, was then proposed exploiting the independence of MDS coded packets to further reduce the backhaul requirements. In this case, a greedy algorithm can be used for small scale networks while for large scale networks a multicast-aware in-cluster cooperative caching algorithm was developed. The advantages of storing coded packets over the uncoded fragments in all the scenarios as well as the benefits of utilizing multicast-aware caching and/or cooperative caching over common caching schemes have been analyzed.

Chapter 7

Conclusion and Future Work

7.1 Conclusion

In this thesis, we studied the resource allocation strategies in both SWIPT systems and cache-enabled networks aiming at overcoming the bottlenecks on energy supply and backhaul capacity. Below, we summarize the main contributions of this thesis.

In Chapter 3, we considered the joint design of the transmit beamforming and power-splitting ratio aiming to minimizing the transmit power subject to the individual SINR and the harvested energy constraints for a MISO SWIPT broadcast system with imperfect CSI. In comparison with the existing method of two-step optimization scheme by iteratively updating the transmit beamforming and power-splitting ratio, we firstly proposed an SDR guided randomization algorithm, which is *non-iterative* but only provides an upper-bound performance after rescaling. To further improve the performance, we proposed a reverse convex constraint based penalty function method which guarantees a rank-one and near-optimal solution. The simulation results showed that the penalty function method not only yields a better solution than the randomization method, but also performs nearly as the SDP method and is quite close to the perfect CSI case, which demonstrates that the proposed PenFun method not only guarantees a rank-one solution but also yields the global optimal solution.

In Chapter 4, the emphasis is shifted into time-switching based EH and MIMO techniques. We studied the joint source, relay matrices and time switching ratio de-

sign for the rate maximization of a MIMO relay network with an energy harvesting relay node. The communication process was divided into three phases, first energy harvesting phase, then information transmission from the source node to the relay node, and finally the information forwarding form the relay node to the destination node. To efficiently finish the entire communication process, the power constraints at the source node and the relay node need to be satisfied. We started with the fixed source covariance matrix scenario assuming uniform source precoding and then considered joint optimization with the source covariance. Closed-form solution as well as an iterative scheme were proposed, respectively, for the two cases, which provided promising principles for designing SWIPT for multi-hop MIMO relay systems.

Chapter 5 presented an optimal cache content placement strategy for small cell networks with heterogeneous file and cache sizes. To minimize the average backhaul rate subject to the cache capacity constraints, multicast was adopted instead of multiple unicast transmissions in the content delivery phase. In particular, the multicast content delivery is facilitated by utilizing the characteristics of the MDS codes, e.g. the independence among the MDS coded packages. The problem is formulated as a nonconvex problem and finally reformulated into a MILP solved by optimization tool. The analysis and simulation results showed the advantages of storing the coded packets over storing uncoded fragments as well as utilizing multicast content delivery over the existing schemes in terms of minimizing the backhaul rate. The impacts of the parameters in both the network and content aspects have also been carefully investigated.

In Chapter 6, we investigated the cache content placement for cache-enabled heterogeneous small cell networks using MDS codes under heterogeneous file and network settings, such as heterogeneous file and cache sizes, distinct numbers of users. In particular, local content popularity is considered instead of the global content popularity. Taking the advantages of multicast content delivery and content sharing among adjacent BSs, we presented two coded caching schemes, dubbed as the multicast-aware caching and the cooperative caching schemes, receptively,

for minimizing the long-term average backhaul load or the UA cost subject to the overall cache capacity constraint. By reformulating the original problems into convex ones, we have derived the optimal content placement in both cases. To further reduce the backhaul requirements, we proposed a compound caching scheme, referred to as multicast-aware cooperative caching, for which a greedy algorithm and a multicast-aware in-cluster cooperative caching algorithm were developed for small scale and large scale networks, respectively. Through analysis, we demonstrated the performance gains of utilizing MDS codes, multicast-aware caching, and cooperative caching for coded caching, content delivery and content sharing accordingly.

7.2 Future Work

The goals for future work include further research on SWIPT and edge caching. In the following, I highlight several research directions which I would like to explore.

7.2.1 Energy Harvesting enabled UAVs

Unmanned aerial vehicles (UAVs) have been widely adopted in military and civilian applications, due to the features of high manoeuvrability and affordability [156]. An important tendency for the evolution of UAVs is towards increasingly smaller size. However, the short endurance of small size UAVs becomes a bottleneck as they are too small and lightweight to carry enough fuel or batteries. Energy harvesting is one of the effective way to deal with this problem by providing energy supply without increasing payloads. On the other hand, UAVs are particularly suitable for energy harvesting due to the high flexible deployment and manoeuvrability which makes them easily implemented even in remote or dangerous environments with poor infrastructures. In addition, the communication distance between an UAV and a energy-constrained IoT device is relatively limited, which is good for maintaining a satisfactory WPT efficiency. And in order to direct the power signal to the UAV, beamforming techniques are needed for energy harvesting enabled UAV systems. At the same time, the acquisition of channel information and locations of the UAVs also brings potential research topics such as robust beamforming, security and privacy. In terms of the UAVs, the deployment and trajectory design demands careful

investigation as well in order to obtain desired energy efficiency for EH within the considered coverage [157].

7.2.2 Energy Harvesting enabled IoTs

As one of the promising services for the 5G communication, IoT facilitates enormous number of devices and enable them to efficiently connect and communicate with each other without direct human interference. Aiming to provide self-sustainable and long lifetime communications, the implementation of IoT requires to adopt energy harvesting enabled devices and utilize energy harvesting transmission techniques to improve energy efficiency [158]. In this case, cooperative beamforming, robust beamforming and physical layer security, become crucial issues requiring to be carefully dealt with. Moreover, IoT facilitates edge intelligence, such as edge caching and edge computing, the implementation of energy harvesting technique in existence of edge caching and computing would also be a significant aspect in making full of the IoT systems [159].

7.2.3 Energy Harvesting enabled Satellite Communication

The application of energy harvesting techniques in satellite communication has attracted considerable attentions. Most of the satellites are solar-powered, which means that solar based energy harvesting can be utilized to provide power supply to the satellites, and then continue to facilitate both wireless power transmission and information transmission to ground stations (low power consumed base stations such as drones) using the harvested energy via microwave beams [160]. This strategy unlocks the potential of applying SWIPT in satellite communication, and can also be combined with the UAV techniques, e.g. exploring the development and implementation of satellite assisted EH enabled drones or base stations, in order to further remove the barriers caused by the battery limited communication.

7.2.4 Content Popularity Estimation and Evolution

For proactive caching approaches, the caching decision is made based on the prelearned content popularity information and we always assume that such knowledge is perfect. However, the accuracy of file popularity information actually depends on

content popularity estimation via machine learning tools. Therefore, it is important to boost the accuracy of content popularity estimation which affects implementing caching strategies. Motivated from this, a fundamental study on content popularity learning for cache-enabled networks has been conducted in [110] where learning-aided caching strategies were proposed with the file popularity dynamically learned by observing user requests. Another issue is about content popularity evolution due to the facts that new files may be generated and become the most popular objects and the popularity ranks of current files also tend to shift over time. In deed, Markov chain model has been utilized to track such evolution. However, the study discussing the impacts that this kind of evolution has on caching decisions is actually rare. Finally, it is crucial to consider local file popularity rather that global file popularity for multi-cell systems as the users in different cells may have different preferences towards the files, e.g., the most popular file in one cell may receive least attentions from the users in another cell.

7.2.5 Privacy-Aware Caching

Privacy-Aware Caching is aimed to introduce the privacy issue into content-centric networks and investigate the measurement of information leakage in content delivery. The privacy concerns mainly come from two aspects, the untrusted base stations which can easily infer users preference by answering to the users queries, and the learning process of the content popularity which requires collecting the user request history. These give rise to the research on caching-forward protocol design as well as the privacy-aware machine learning to control information leakage in both the content delivery phase and the content popularity learning phase. Privacy has been recognized as one of the open issues in IoT, big data, and other applications of machine learning, and therefore conveys huge research potential in both fields of networking and data science. Machine learning, is now one of the hottest issue, not only for predicting the content popularity in cache-enabled networks, but also for understanding the wireless communication from a brand-new perspective.

7.2.6 Joint Transmission and Caching Designs

In literature, those focusing on caching policy design usually ignore the impacts of the physical layer parameters while others focus more on the transmission aspects for more realistic network models but assuming either most popular files are cached or the cache placement is given in the considered time slot. A fundamental work for joint transmission and caching design has been provided in [100] where a two time-scale joint optimization of MIMO precoding and cache control was proposed for cache-enabled opportunistic cooperative MIMO (CoMP) to minimize the transmit power. In the short-term time scale, the precoding matrices were optimized based on the instantaneous channel state information and cache states subject to the rate constraint. In the long-term time scale, the cache content placement was designed using the user requests information as well as the precoding matrices subject to the cache capacity constraint. Due to the coupling between caching and transmission, we deem that the joint designs are important and worthy of further investigations.

7.2.7 Mobility-Aware Caching

The mobility of users is also an important aspect that affects the performance of caching approaches and hence requires further investigations for cache-enabled networks. As the users may move rapidly from one cell to another before the data delivery is finished, the requests of a user can be served by multiple BSs sequentially. Therefore, the caching decision must be performed taking into account the predictions about user mobility patterns and also the cooperation and coordination among different caches, which adds more challenges to caching problems besides the uncertainties in terms of content popularity and user demand patterns. Discrete-time Markov chain model has been adopted to analyze the impacts of mobility of users on caching management in [107, 108].

Appendix A

Proof of *Proposition3.1*

Suppose $(\mathbf{W}_k^*, \rho_k^*, \mu_k^*, \lambda_k^*)$ be the optimal solution of (3.17). Letting $q_k^* = \frac{1}{\rho_k^*}$, and $\tilde{q}_k^* = \frac{1}{1-\rho_k^*}$, it is easy to see that $(\mathbf{W}_k^*, \rho_k^*, q_k^*, \tilde{q}_k^*, \mu_k^*, \lambda_k^*)$ also satisfies the constraints in (3.18). Oppositely, if $(\mathbf{W}_k^*, \rho_k^*, q_k^*, \tilde{q}_k^*, \mu_k^*, \lambda_k^*)$ is the optimal solution for (3.18), then $\mathbf{\Gamma}_k$ and $\mathbf{\Upsilon}_k$ will both be positive semi-definite (PSD) due to the fact that $\mathbf{\Gamma}_k - \tilde{\mathbf{\Gamma}}_k \succeq \mathbf{0}, \mathbf{\Upsilon}_k - \tilde{\mathbf{\Upsilon}}_k \succeq \mathbf{0}$. Also, the objective function is not directly related to $q_k^*, \tilde{q}_k^*$ such that we can solve (3.18) with CVX instead of (3.17).

Appendix B

Proof of *Lemma 5.1*

In (5.3a), the instantaneous backhaul rates for all kinds of possible user request profiles $\{\pi_1,\ldots,\pi_N\}$ are summed up to obtain the average backhaul rate while that for a particular user request profile is composed of the associated backhaul rates for all the files. Equivalently, the average backhaul rate can also be calculated by summing up the average backhaul rate for each file in terms of all kinds of possible user request profiles. Mathematically, we able able to rewrite (5.3a) as

$$C_{\text{multicast}}^{\text{MDS}}=\sum_{j=1}^{N}\sum_{\{\pi_1,\ldots,\pi_N\}}\left(1-\min_{k\in\mathscr{K}_{\pi_j}}\frac{m_{k,j}}{n_j}\right)s_jP_r(\{\pi_1,\ldots,\pi_N\}). \tag{B.1}$$

For a particular file j, the volume of packets to be sent via backhaul is subject to the content placement $\mathbf{m}^j$ and the associated user request profile π_j regardless of the user request profiles for other files $\{\pi_i\}_{i\neq j}$. That is to say any user request profile $\{\pi_1,\ldots,\pi_N\}$ with the same π_j would yield the same backhaul rate for file j. Consequently, when calculating the backhaul rate for a file, we can only consider different user request profiles for the certain file and ignore the user request profiles for other files. Hence, (B.1) can be further reformulated into

$$C_{\text{multicast}}^{\text{MDS}}=\sum_{j=1}^{N}\sum_{\pi_j}\left(1-\min_{k\in\mathscr{K}_{\pi_j}}\frac{m_{k,j}}{n_j}\right)s_jP_r(\pi_j), \tag{B.2}$$

which is the same with (5.4) obtained by considering the user request profile for each file and then summing up the backhaul rate for all the files. The equivalence between (5.3a) and (5.4) is hence proved.

Appendix C

Proof of *Lemma 5.3*

Since the solution of (5.11) always satisfies $\widetilde{\mathbf{g}}^j = \text{sort}(\widetilde{\mathbf{q}}^j), \forall j$, it can be easily proved that $\widetilde{q}_{k,j} = \widetilde{g}_{\widetilde{r}_{k,j},j}$ where $\widetilde{r}_{k,j}$ is the rank of $\widetilde{q}_{k,j}$ in $\widetilde{\mathbf{q}}^j$. Note that the ranks must be unique integers. Hence, if we let $\widetilde{x}^k_{t,j} = 1|_{\widetilde{r}_{t,j}=k}$ and otherwise $\widetilde{x}^k_{t,j} = 1$, we will then get $\widetilde{q}_{k,j} = \sum_{t=1}^K \widetilde{g}_{t,j}\widetilde{x}^k_{t,j}$, which satisfy all the constraints in (5.13). Hence, $\{\widetilde{g}_{k,j}\}$ and $\{\widetilde{x}^k_{t,j}\}$ are the solution of (5.13). Oppositely, if $\{\widetilde{g}_{k,j}\}$ and $\{\widetilde{x}^k_{t,j}\}$ are known to be the solution to (5.13), it is easy to prove that $\{\widetilde{g}_{k,j}\}$ are the solution to (5.11) and then use them to recover $\{\widetilde{q}_{k,j}\}$, i.e., $\widetilde{q}_{k,j} = \sum_{t=1}^K \widetilde{g}_{t,j}\widetilde{x}^k_{t,j}$. In this case, the rank of $\widetilde{q}_{k,j}$ in $\widetilde{\mathbf{q}}^j$ is $\widetilde{r}_{k,j} = t|_{\widetilde{x}^k_{t,j}=1}$. The equivalence is therefore proved.

Appendix D

Proof of *Lemma 5.4*

Firstly, we prove that any (x, y, z) with $z = xy$ can satisfy constraints (5.14)–(5.17). Based on the definition that $0 \leq y \leq \widetilde{y}$ and $x \in \{0, 1\}$, we know that z is monotonically increasing with both x and y. Thus, (5.14), (5.16) and (5.17) always hold. Also, when $x = 0$, we get $y - \widetilde{y} \leq 0$ and $z = 0$ in (5.15). Similarly, when $x = 1$, we can prove (5.15). Now suppose that (x, y, z) satisfies (5.14)–(5.17) and we prove that $z = xy$ by contradiction. Assume that there is a z satisfying $z > xy$. According to (5.16), we then get $z \leq y$ which indicates that $x = 0$ and hence $z > 0$. This contradicts with (5.14). The assumption cannot be true. Similarly, we can prove that once $z < xy$ happens, x must be equal to 1 and $z < y$ in order to satisfy (5.17). This in turn violates (5.15) which requires $z \geq y$. Consequently, $z = xy$ always holds in this case. *Lemma 5.4* is then proved.

Appendix E

Proof of *Lemma 6.2*

Firstly, we divide the possible user request profiles for each file, e.g., π_j into $K+1$ types defined as $\{\pi_j^0, \pi_j^1, \pi_j^2, \ldots, \pi_j^K\}$ according to the different values of the associated backhaul load (in percentage) for file j, i.e., $\{0, 1-\frac{m_{1,j}}{n_j}, 1-\frac{m_{2,j}}{n_j}, \ldots, 1-\frac{m_{K,j}}{n_j}\}$, respectively. Note that π_j^0 states that file j is not requested by users in any of the cells, and hence backhaul is no longer needed in this case. If cell k stores the least number of packets of file j among all the cells requesting file j, i.e., $\min_{t\in\mathscr{K}_{\pi_j}} \frac{m_{t,j}}{n_j} = \frac{m_{k,j}}{n_j}$, then the associated user request profile π_j^k will imply that file j is requested by cell k and that there will not be any cell t satisfying $r_{t,j} < r_{k,j}$. Considering the definition of $\mathscr{T}_{k,j}$, we obtain that $P_r(\pi_j^k) = (1-\alpha_{k,j})\prod_{t\in\mathscr{T}_{k,j}} \alpha_{t,j}$. Summing up all types of user request profiles $\{\pi_j^k\}$ for all files, the average backhaul rate can be written as (6.5) which ends the proof of the lemma.

Appendix F

Proof of *Lemma 6.4*

Here pairwise comparison is used to tackle the problem caused by the uncertain relation of $\{\alpha_{\vartheta_v,j}\}$. Firstly, we utilize a simple example to help better clarify this lemma.

Example 1. Let $K = 3$. Then it follows that $\alpha^j = [\alpha_{1,j}, \alpha_{2,j}, \alpha_{3,j}]$. Now, assume that for any given j, the only three nonzero elements of $\{y^t_{k,j}\}$ are given by $y^{\theta_1}_{1,j} = 1, y^{\theta_2}_{2,j} = 1, y^{\theta_3}_{3,j} = 1$. Then we let $\varphi_{t,j} = \left(1-\alpha_{\theta_t,j}\right)\prod_{v=1}^{t-1}(\alpha_{\theta_v,j}), \forall t$ using (6.12). As such, the objective function can be rewritten as

$$R^{\mathrm{MDS}}_{\mathrm{multicast}} = \left(1-g_{1,j}\right)\left(1-\alpha_{\theta_1,j}\right) + \left(1-g_{2,j}\right)\left(1-\alpha_{\theta_2,j}\right) \times \alpha_{\theta_1,j} + \left(1-g_{3,j}\right)\left(1-\alpha_{\theta_3,j}\right)\alpha_{\theta_1,j}\alpha_{\theta_2,j}. \quad \text{(F.1)}$$

Now we prove that the optimal $\{y^t_{k,j}\}$ must ensure that $\alpha_{\theta_1,j} \geq \alpha_{\theta_2,j} \geq \alpha_{\theta_3,j}$ by contradiction. Assume $\alpha_{\theta_2,j} < \alpha_{\theta_3,j}$ and calculate $R^{\mathrm{MDS}}_{\mathrm{multicast}}$ using (F.1). Then we exchange the values of $\alpha_{\theta_2,j}$ and $\alpha_{\theta_3,j}$ and recalculate the objective function. The difference between the former and the later objective function can be given by

$$\Delta R^{\mathrm{MDS}}_{\mathrm{multicast}} = \left(g_{3,j} - g_{2,j}\right)\alpha_{\theta_1,j}\left(\alpha_{\theta_3,j} - \alpha_{\theta_2,j}\right). \quad \text{(F.2)}$$

Considering $g_{2,j} \leq g_{3,j}$ and $\alpha_{\theta_2,j} < \alpha_{\theta_3,j}$, we prove that $\Delta R^{\mathrm{MDS}}_{\mathrm{multicast}} \geq 0$. That is to say, for any $\alpha_{\theta_2,j} < \alpha_{\theta_3,j}$, we can always obtain a smaller or at least equal objective function by exchanging $\alpha_{\theta_2,j}$ and $\alpha_{\theta_3,j}$. Hence, $\alpha_{\theta_2,j} \geq \alpha_{\theta_3,j}$ is essential to minimize the backhaul load. In the same way, we can prove that $\alpha_{\theta_1,j} \geq \alpha_{\theta_2,j}$. Consequently, $\alpha_{\theta_1,j} \geq \alpha_{\theta_2,j} \geq \alpha_{\theta_3,j}$ is proved. The same conclusion can easily be extended to

the K cell scenario which indicates that $\alpha_{\theta_1,j} \geq \alpha_{\theta_2,j} \geq \cdots \alpha_{\theta_K,j}$. The rigorous mathematical proof is presented below.

We let $\phi_\nu^j, \forall \nu = 2,3,\ldots,K$ be the summation of the items in $R_{\text{multicast}}^{\text{MDS}}$ that involves $\alpha_{\vartheta_{\nu-1},j}$ and $\alpha_{\vartheta_\nu,j}$, given by

$$\phi_\nu^j = \sum_{k=\nu-1}^{\nu} \left(1-g_{k,j}\right)\left(1-\alpha_{\vartheta_k,j}\right)\prod_{t=1}^{k-1}\alpha_{\vartheta_t,j}s_j. \tag{F.3}$$

Since $\alpha_{\vartheta_t,j}, t = 1,2,\ldots,\nu-2$ are interchangeable in ϕ_ν^j, the relation among them will not affect the value of ϕ_ν^j as well as the relation between $\alpha_{\vartheta_{\nu-1},j}$ and $\alpha_{\vartheta_\nu,j}$. Consequently, we consider the derivatives of $\alpha_{\vartheta_{\nu-1},j}$ and $\alpha_{\vartheta_\nu,j}$ in ϕ_ν^j as follows

$$\frac{\partial \phi_\nu^j}{\partial \alpha_{\vartheta_{\nu-1},j}} = \left(g_{\nu-1,j}-1+\left(1-g_{\nu,j}\right)\left(1-\alpha_{\vartheta_\nu,j}\right)\right)\prod_{t=1}^{\nu-2}\alpha_{\vartheta_t,j}s_j, \tag{F.4}$$

$$\frac{\partial \phi_\nu^j}{\partial \alpha_{\vartheta_\nu,j}} = -\left(1-g_{\nu,j}\right)\prod_{t=1}^{\nu-1}\alpha_{\vartheta_t,j}s_j. \tag{F.5}$$

Let $\Delta_\nu^j = \frac{\partial \phi_\nu^j}{\partial \alpha_{\vartheta_\nu,j}} - \frac{\partial \phi_\nu^j}{\partial \alpha_{\vartheta_{\nu-1},j}}$, and we obtain that

$$\Delta_\nu^j = \sum_{j=1}^{N}\left(1-g_{\nu,j}\right)\left(\alpha_{\vartheta_\nu,j}-\alpha_{\vartheta_{\nu-1},j}\right)\prod_{t=1}^{\nu-2}\alpha_{\vartheta_t,j}s_j + \left(g_{\nu,j}-g_{\nu-1,j}\right)\prod_{t=1}^{\nu-2}\alpha_{\vartheta_t,j}s_j. \tag{F.6}$$

Because $\Delta_\nu^j = 0$ indicates that $\alpha_{\vartheta_{\nu-1},j}$ and $\alpha_{\vartheta_\nu,j}$ are interchangeable, here we focus on the case when $\Delta_\nu^j \neq 0$. If $\Delta_\nu^j > 0$, it follows that the derivative of $\alpha_{\vartheta_\nu,j}$ in ϕ_ν^j is higher than that of $\alpha_{\vartheta_{\nu-1},j}$, which is to say, the weight for $\alpha_{\vartheta_\nu,j}$ in terms of the weighted summation ϕ_ν^j is higher. Hence, we should let $\alpha_{\vartheta_\nu,j} \leq \alpha_{\vartheta_{\nu-1},j}$ in order to minimize the objective function $R_{\text{multicast}}^{\text{MDS}}$. On the contrary, if $\Delta_\nu^j < 0$, then it holds true that $\alpha_{\vartheta_\nu,j} \geq \alpha_{\vartheta_{\nu-1},j}$. Consequently, assuming that $\Delta_\nu^j < 0$, we obtain $\alpha_{\vartheta_\nu,j} \geq \alpha_{\vartheta_{\nu-1},j}$ and hence the right side of (F.6) is always non-negative since $g_{\nu,j} \leq 1$ and $g_{\nu-1,j} \leq g_{\nu,j}$, which conflicts with the assumption. Hence, it holds true that $\Delta_\nu^j \geq 0$ and $\alpha_{\vartheta_\nu,j} \leq \alpha_{\vartheta_{\nu-1},j}$ and the lemma is then proved.

Based on the definition of β^j and the conclusion drawn above, we derive that $\beta^j = [\alpha_{\theta_1,j}, \alpha_{\theta_2,j}, \ldots, \alpha_{\theta_K,j}]$. As a consequence, the optimal $\varphi_{t,j}^*$ can be written as $\varphi_{t,j}^* = \left(1-\beta_{t,j}\right)\prod_{\nu=1}^{t-1}\beta_{\nu,j}$. The corresponding values of $\{y_{k,j}^t\}$ can easily be calculated as given in (6.18).

Appendix G

Proof of *Lemma 6.5*

Given some cooperative caching policy $(\{x^t_{k,j}\},\{m_{k,j}\})$, the costs for fetching content from neighboring cells are the same in the coded and uncoded caching scenarios. Therefore, the difference in the backhaul cost shows up most clearly in the UA costs. When uncoded fragments are stored, all the fragments except the ones that are either stored in local cache or fetched from the neighboring cells are needed from the MBS via backhaul to each cell requesting the particular file. Considering the possible content overlap amongst those fragments, the number of unique fragments for file j available at cell $k\in\mathscr{K}_{\pi_j}$ would always be less than or equal to $\sum_t x^t_{k,j}$ for a certain user request profile π_j which leads to a higher backhaul rate than that in the MDS coded case. If the fragments are assumed to be randomly selected to be stored in the cells and then sent to the neighboring cells *equiprobably*, the probability of each fragment of file j needing to be sent to cell k via backhaul, i.e., not being stored locally or sent to the particular cell k from other SBSs, would be given by

$$\widehat{\rho}_{k,j}=\prod_{t=1}^{K}\left(\frac{\binom{n_j-1}{m_{t,j}}}{\binom{n_j}{m_{t,j}}}+\frac{\binom{n_j-1}{m_{t,j}-1}}{\binom{n_j}{m_{t,j}}}\frac{\binom{m_{t,j}-1}{x^t_{k,j}}}{\binom{m_{t,j}}{x^t_{k,j}}}\right)=\prod_{t=1}^{K}\left(1-\frac{x^t_{k,j}}{n_j}\right). \tag{G.1}$$

In this case, the average UA cost can be written as

$$C^{\text{uncoded}}_{\text{coop}}=\sum_{j=1}^{N}\sum_{k=1}^{K}\left[\widehat{\rho}_{k,j}f^M_k+\sum_{t=1}^{K}\frac{x^t_{k,j}}{n_j}f^t_k\right]s_j(1-\alpha_{k,j}). \tag{G.2}$$

Compared with the UA cost in (6.27), if we can prove that

$$\prod_{t=1}^{K}\left(1-\frac{x^t_{k,j}}{n_j}\right)\geq 1-\min\left(1,\sum_{t=1}^{K}\frac{x^t_{k,j}}{n_j}\right),\forall k,j, \tag{G.3}$$

then it holds true that $C_{\text{coop}}^{\text{MDS}} \leq C_{\text{coop}}^{\text{uncoded}}$. Hence, here we focus on the proof of the result (G.3). As can be observed, when $\sum_{t=1}^{K} \frac{x_{k,j}^t}{n_j} \geq 1$, (G.3) is always true. When $\sum_{t=1}^{K} \frac{x_{k,j}^t}{n_j} < 1$, the right hand side of (G.3) equals to $\left(1 - \sum_{t=1}^{K} \frac{x_{k,j}^t}{n_j}\right)$. In this case, we prove (G.3) using mathematical induction.

To be brief, we mathematically reformulate the problem into a general problem, which reads

$$\prod_{t=1}^{K} (1 - \chi_t) \geq 1 - \sum_{t=1}^{K} \chi_t, \tag{G.4}$$

where $\chi_t \in [0,1]$. Obviously, when $K = 1$ or 2, the statement is always true as expected. Now assuming that (G.4) holds for $K = \kappa$, we hence have

$$\prod_{t=1}^{\kappa} (1 - \chi_t) \geq 1 - \sum_{t=1}^{\kappa} \chi_t. \tag{G.5}$$

Then it follows that

$$\begin{aligned} \prod_{t=1}^{\kappa+1} (1 - \chi_t) &= \prod_{t=1}^{\kappa} (1 - \chi_t) - \prod_{t=1}^{\kappa} (1 - \chi_t)\, \chi_{\kappa+1} \\ &\geq \left(1 - \sum_{t=1}^{\kappa} \chi_t\right) - \chi_{\kappa+1}, \end{aligned} \tag{G.6}$$

due to the fact that $0 \leq \prod_{t=1}^{\kappa} (1 - \chi_t) \leq 1$ as well as the inequality (G.5). Now we are able to conclude that the statement is true for all available K via induction. Then going back to the original problem and letting $\chi_t = \frac{x_{k,j}^t}{n_j}$ for any given k, we have proved the statement

$$\prod_{t=1}^{K} \left(1 - \frac{x_{k,j}^t}{n_j}\right) \geq 1 - \sum_{t=1}^{K} \frac{x_{k,j}^t}{n_j}, \forall k, j. \tag{G.7}$$

Based on this analysis, $C_{\text{coop}}^{\text{MDS}} \leq C_{\text{coop}}^{\text{uncoded}}$ is then proved.

Appendix H

Proof of *Lemma 6.6*

Considering multicast-aware cooperative caching, the UA cost can be written as

$$C_{\text{coop}}^{\text{Mul}} = \sum_{j=1}^{N} \sum_{\pi_j \in \Pi_j} \left[\left(1 - \min_{k \in \mathscr{K}_{\pi_j}} \sum_{t=1}^{K} z_{k,j}^{t} \right) \max_{k \in \mathscr{K}_{\pi_j}} f_k^M + \sum_{k \in \mathscr{K}_{\pi_j}} \sum_{t=1}^{K} z_{k,j}^{t} f_k^t \right] P_r(\pi_j) s_j. \quad \text{(H.1)}$$

As we can see, the first item denotes the backhaul cost while the second item presents the cost for content sharing among the cooperative SBSs. For each given user request profile for a particular file π_j, the cost for fetching content from the cooperative SBSs at cell k appears only when file j is requested by the users in cell k which means that $\pi_j(k) = 1$ regardless of the individual user request profiles in other cells. It is easy to prove $P_r(\pi_j|_{\pi_j(k)=1}) = 1 - \alpha_{k,j}$, and so (6.28).

Appendix I

Proof of *Lemma 6.7*

If $(\{x^t_{k,j}\},\{m_{k,j}\})$ is given, then the costs for fetching content from neighboring cells will be the same in the coded and uncoded caching scenarios. As a result, the comparison is focused on the backhaul costs in the two scenarios. When uncoded fragments are stored, all the fragments except for the ones that can be fetched at all of the cells requesting the file either from local cache or from the neighboring cells are needed to be sent from the MBS via multicast transmission. Assuming that the fragments are randomly selected to be stored in the cells and then sent to the neighboring cells *equiprobably*, the probability of each fragment of file j available at all of the cells requesting the file either from local cache or from the neighboring cells would be given by

$$\widetilde{\rho}_{\pi_j} = \prod_{k\in\mathscr{K}_{\pi_j}} (1-\widehat{\rho}_{k,j}), \tag{I.1}$$

where $\widehat{\rho}_{k,j}$ is the probability of each fragment of file j not being stored locally or sent to the particular cell k from other SBSs given by (G.1) in Appendix A. Similar to the multicast-aware case, the average UA cost can be written as

$$C^{\text{uncoded}}_{\text{mult,coop}} = \sum_{j=1}^{N}\left[\sum_{\pi_j\in\Pi_j}\left(1-\widetilde{\rho}_{\pi_j}\right)\max_{k\in\mathscr{K}_{\pi_j}} f^M_k P_r(\pi_j) + \sum_{k=1}^{K}\sum_{t=1}^{K} z^t_{k,j} f^t_k (1-\alpha_{k,j})\right] s_j. \tag{I.2}$$

According to (G.1) and (G.7), we obtain

$$\widetilde{\rho}_{\pi_j} \le \prod_{k\in\mathscr{K}_{\pi_j}} (\sum_{t=1}^{K} \frac{x^t_{k,j}}{n_j}). \tag{I.3}$$

As $0 \leq \sum_{t=1}^{K} \frac{x_{k,j}^t}{n_j} \leq 1, \forall k \in \mathscr{K}_{\pi_j}$, it holds true that $\widetilde{\rho}_{\pi_j} \leq \min_{k \in \mathscr{K}_{\pi_j}} \sum_{t=1}^{K} z_{k,j}^t$. Compared with the average UA cost in (6.28), we derive that $C_{\text{mult,coop}}^{\text{MDS}} \leq C_{\text{mult,coop}}^{\text{uncoded}}$.

Appendix J

Proof of *Lemma 6.8*

To proceed, we sort $\lambda^j = \{\lambda_{k,j}, k \in \mathscr{S}_l^u\}$ in an ascending order and define the sorted vector as ψ^j with $\psi_{k,j} = \lambda_{\vartheta_k,j}$ and $\psi_{k,j} \leq \psi_{k+1,j}, \forall k \in \mathscr{S}_l^u \setminus |\mathscr{S}_l^u|$. For instance, if $\vartheta_1 = k$, it means that $\lambda_{k,j}$ equals to $\psi_{1,j}$ and is therefore the lowest. On the contrary, if $\vartheta_{|\mathscr{S}_l^u|} = k$, it means that $\lambda_{k,j}$ equals to $\psi_{|\mathscr{S}_l^u|,j}$ and is hence the highest. Consequently, the objective function in (6.36) can be rewritten as

$$\widetilde{C}_l^u = \sum_{j=1}^{N} \sum_{k \in \mathscr{S}_l^u} \Bigg[\left(1 - \psi_{k,j}\right) f_u^M (1 - \alpha_{\vartheta_k,j}) \prod_{\nu=1}^{k-1} \alpha_{\vartheta_\nu,j} + \psi_{k,j} f_l^u (1 - \alpha_{\vartheta_k,j}) \Big] s_j. \quad \text{(J.1)}$$

The reformulated problem can then be written as

$$\min_{\{\psi_{k,j}\}} \ \widetilde{C}_l^u \quad \text{(J.2a)}$$

$$\text{s.t.} \ \ \psi_{k,j} \leq \psi_{k+1,j}, \forall k \in \mathscr{S}_l^u \setminus |\mathscr{S}_l^u|, \forall j, \quad \text{(J.2b)}$$

$$0 \leq \psi_{k,j} \leq 1, \ \forall k \in \mathscr{S}_l^u, \forall j, \quad \text{(J.2c)}$$

$$q_{l,j}^u \leq \psi_{k,j} \leq |\mathscr{S}_l^u| q_{l,j}^u, \ \forall k \in \mathscr{S}_l^u, \forall j. \quad \text{(J.2d)}$$

Apparently, $\psi_{k,j}, \forall k \in \mathscr{S}_l^u$ are treated similarly in the constraints (J.2c)-(J.2d) regardless of the values of $\{\vartheta_k\}$. Given any $\{\vartheta_k\}$, we want to find the actual relation of the optimal $\psi_{k,j}, \forall k \in \mathscr{S}_l^u$ to minimize the objective function in (J.1). Furthermore, the objective function and constraints are independent towards of different files in (J.2), and hence the UA cost minimization problem for each cluster can be further decomposed into N sub-problems each minimizing the associated cost for a

particular file defined as $\widetilde{C}^u_{l,j}, \forall j$. Thus, we consider the derivatives of $\{\psi_{k,j}\}$ in $\widetilde{C}^u_{l,j}$ given by

$$\frac{\partial \widetilde{C}^u_{l,j}}{\partial \psi_{k,j}} = \left(Q^u_l - \prod_{\nu=1}^{k-1} \alpha_{\vartheta_\nu,j} \right) f^M_u (1-\alpha_{\vartheta_k,j}) s_j, \forall k \in \mathscr{S}^u_l \backslash 1, \tag{J.3}$$

$$\frac{\partial \widetilde{C}^u_{l,j}}{\partial \psi_{1,j}} = (Q^u_l - 1) f^M_u (1-\alpha_{\vartheta_k,j}) s_j, \tag{J.4}$$

where $Q^u_l = f^u_l / f^M_u$ denotes the ratio between the costs of fetching content via backhaul and from the cluster. Since $0 < Q^u_l < 1$, it holds true that $\frac{\partial \widetilde{C}^u_{l,j}}{\partial \psi_{1,j}} < 0$. For any $\psi_{k,j}, \forall k \in \mathscr{S}^u_l \backslash 1$ satisfying the constraints, $\widetilde{C}^u_{l,j}$ reaches its lowest when we let $\psi_{1,j} = \psi_{2,j}$ since a larger $\psi_{1,j}$ contributes to a lower $\widetilde{C}^u_{l,j}$. In the same way, it can be proved that the relation between $\psi_{k,j}$ and $\psi_{k+1,j}$ is subject to the value of $\left(Q^u_l - \prod_{\nu=1}^{k-1} \alpha_{\vartheta_\nu,j} \right)$. Note that $\prod_{\nu=1}^{k-1} \alpha_{\vartheta_\nu,j}$ always decreases with the increase of k which indicates that if $\prod_{\nu=1}^{k-1} \alpha_{\vartheta_\nu,j} \leq Q^u_l$, we always have $\prod_{\nu=1}^{t-1} \alpha_{\vartheta_\nu,j} < Q^u_l, \forall t > k$. Hence, we discuss about the relation among $\{\psi_{k,j}\}$ in two kinds of conditions. In the first case, we assume that $Q^u_l \leq \prod_{\nu=1}^{|\mathscr{S}^u_l|-1} \alpha_{\vartheta_\nu,j}$, and it is easy to prove that $\psi_{1,j} = \psi_{2,j} = \cdots = \psi_{|\mathscr{S}^u_l|,j}$ by iteratively utilizing the similar trick for proving $\psi_{1,j} = \psi_{2,j}$. Otherwise, when

$$Q^u_l \geq \prod_{\nu=1}^{k-1} \alpha_{\vartheta_\nu,j} = \begin{cases} < 0, & k \in [1,\ldots,t], \\ \geq 0, & k \in [t+1,\ldots,|\mathscr{S}^u_l|], \end{cases} \tag{J.5}$$

it is still possible to prove that $\psi_{1,j} = \psi_{2,j} = \cdots = \psi_{t+1,j}$ by fixing $\psi_{t+1,j}$. While for $k \in [t+1,\ldots,|\mathscr{S}^u_l|]$ when $\widetilde{C}^u_{l,j}$ decreases with the decline of $\psi_{k,j}$, we let $\psi_{t+1,j} = \psi_{t+2,j} = \cdots = \psi_{|\mathscr{S}^u_l|,j}$ to get the lowest cost $\widetilde{C}^u_{l,j}$ for any given $\psi_{t+1,j}$ using (J.2b). It is then proved that $\psi_{1,j} = \psi_{2,j} = \cdots = \psi_{|\mathscr{S}^u_l|,j}$. As a result, we derive that $\lambda_{k,j} = \lambda_{t,j}, \forall k,t \in \mathscr{S}^u_l$.

Bibliography

[1] M. Agiwal, A. Roy, and N. Saxena, "Next Generation 5G Wireless Networks: A Comprehensive Survey," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 3, pp. 1617–1655, 3rd Quart. 2016.

[2] D. Niyato, E. Hossain, M. M. Rashid, and V. K. Bhargava, "Wireless Sensor Networks with Energy Harvesting Technologies: A Game-Theoretic Approach to Optimal Energy Management," *IEEE Trans. Wireless Commun.*, vol. 14, no. 4, pp. 90–96, Aug. 2007.

[3] L. Hou, S. Tan, "A Preliminary Study of Thermal Energy Harvesting for Industrial Wireless Sensor Networks," in *Proc. 10th Int. Conf. Sens. Technol. (ICST)*, Oct. 2016, pp. 1–5.

[4] I. Krikidis, S. Timotheou, S. Nikolaou, G. Zheng, D. W. K. Ng and R. Schober, "Simultaneous wireless information and power transfer in modern communication systems," *IEEE Commun. Magazine*, vol. 52, no. 11, pp. 104–110, Nov. 2014.

[5] L. R. Varshney, "Transporting Information and Energy Simultaneously," *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Oct. 2008, pp. 1612–1616.

[6] I. J. Yoon, "Wireless Power Transfer in the Radiating Near-Field Region," *Proc. USNC URSI Radio Sci. Meeting (Joint AP S Symp.)*, Oct. 2015, pp. 334.

[7] X. Zhou, R. Zhang, and C. K. Ho, "Wireless Information and Power Transfer in Multiuser OFDM Systems," *IEEE Trans. Wireless Commun.*, vol. 13, no. 4, pp. 2282–2294, Apr. 2014.

[8] Z. Hu, N. Wei, and Z. Zhang, "Optimal Resource Allocation for Harvested Energy Maximization in Wideband Cognitive Radio Network with SWIPT," *IEEE Access*, vol. 5, pp. 23383–23394, 2017.

[9] Y. Zeng and R. Zhang, "Full-Duplex Wireless-Powered Relay with Self-Energy Recycling," *IEEE Wireless Commun. Lett.*, vol. 4, no. 2, pp. 201–204, Apr. 2015.

[10] R. Zhang and C. K. Ho, "MIMO broadcasting for simultaneous wireless information and power transfer," *IEEE Trans. Wireless Commun.*, vol. 12, no. 5, pp. 1989–2001, May 2013.

[11] R. I. Ansari, S. A. Hassan and C. Chrysostomou, "A SWIPT-based Device-to-Device Cooperative Network," in *Proc. 24th International Conference on Telecommunications (ICT)*, Limassol, 2017, pp. 1–5.

[12] S. Timotheou, G. Zheng, C. Masouros, and I. Krikidis, "Symbollevel Precoding in Miso Broadcast Channels for SWIPT Systems," in *Proc. 23rd Int. Conf. Telecommun. (ICT)*, Thessaloniki, Greece, 2016, pp. 1–5.

[13] A. Nasir, X. Zhou, S. Durrani, and R. Kennedy, "Relaying Protocols for Wireless Energy Harvesting and Information Processing," *IEEE Trans. Wireless Commun.*, vol. 12, no. 7, pp. 3622–3636, Jul. 2013.

[14] Z. Ding et al., "A Survey on Non-Orthogonal Multiple Access for 5g Networks: Research Challenges and Future Trends, " *IEEE J. Sel. Areas Commun.*, vol. 35, no. 10, pp. 2181–2195, Oct. 2017.

[15] T. A. Khan, A. Alkhateeb and R. W. Heath, "Millimeter Wave Energy Harvesting," *IEEE Trans. Wireless Commun.*," vol. 15, no. 9, pp. 6048–6062, Sept. 2016.

[16] S. Lohani, E. Hossain and V. K. Bhargava, "On Downlink Resource Allocation for SWIPT in Small Cells in a Two-Tier HetNet," *IEEE Tran. Wireless Commun.*, vol. 15, no. 11, pp. 7709–7724, Nov. 2016.

[17] K. Liang; L. Zhao; Z. Ding; H. H. Chen, "Double Side Signal Splitting SWIPT for Downlink CoMP Transmissions with Capacity Limited Backhaul," *IEEE Commun. Lett.*, vol.PP, no.99, pp.1–1, Aug. 2016.

[18] W. N. S. F. Wan Ariffin; X. Zhang; M. R. Nakhai, "Sparse Beamforming for Real-time Resource Management and Energy Trading in Green C-RAN," *IEEE Trans. Smart Grid*, vol.PP, no.99, pp.1-1, Sept. 2016.

[19] C. Wang, J. Li, F. Ye, and Y. Yang, "NETWRAP: An NDN-Based Real-Time Wireless Recharging Framework for Wireless Sensor Networks," *IEEE Trans. Mobile Comput.*, vol. 13, no. 6, pp. 1283–1297, Jun. 2014.

[20] Z. Ding et al., "Application of Smart Antenna Technologies in Simultaneous Wireless Information and Power Transfer," *IEEE Commun. Mag.*, vol. 53, no. 4, pp. 86–93, Apr. 2007.

[21] L. Liu, R. Zhang, and K.-C. Chua, "Secrecy Wireless Information and Power Transfer with MISO Beamforming," *IEEE Trans. Signal Processing*, vol. 62, no. 7, pp. 1850–63, Apr. 2014.

[22] T. D. Ponnimbaduge Perera, D. N. K. Jayakody, S. K. Sharma, S. Chatzinotas and J. Li, "Simultaneous Wireless Information and Power Transfer (SWIPT): Recent Advances and Future Challenges,"*IEEE Commun. Surveys Tuts.*, vol. 20, no. 1, pp. 264–302, First quarter 2018.

[23] Cisco, "Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2016-2021 White Paper [OL]," http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/mobile-white-paper-c11-520862.html, 2017.

[24] J. G. Andrews, S. Buzzi, W. Choi et al., "What Will 5G Be?" *IEEE Commun. Mag.*, vol. 32, no. 6, pp. 1065–1082, Jun. 2014.

[25] A. Gupta and R. K. Jha, "A Survey of 5G Network: Architecture and Emerging Technologies, " *IEEE Access*, vol. 52, 1206–1232, Aug. 2015.

[26] M.-G. D. Benedetto and B. R. Vojcic, "Ultra Wide Band Wireless Communications: A Tutorial ," *Journ. Commun. and Networks*, vol. 5, no. 4, pp. 290–302, Dec. 2003.

[27] Z. Xiang, M. Tao, and X.Wang, "Massive MIMO Multicasting in Noncooperative Cellular Networks, " *IEEE J. Selected Areas in Commun. (JSAC)*, vol. 32, no. 6, pp. 1180–1193, Jun. 2014.

[28] R. Taori and A. Sridharan, "Point-to-multipoint in-band mmwave backhaul for 5G networks," *IEEE Commun. Mag.*, vol. 53, no. 1, pp. 195–201, Jan. 2015.

[29] J. G. Andrews, "Seven Ways that HetNets Are a Cellular Paradigm Shift," *IEEE Commun. Mag.*, vol. 51, no. 3, pp. 136–144, Mar. 2013.

[30] E. Bastug, M. Bennis et al., "Big data meets telcos: A proactive caching perspective," *Journ. Commun. and Networks, Special Issue on Big Data Networking Challenges and Applications*, vol. 17, no. 6, pp. 549–557, Dec. 2015.

[31] E. Zeydan, E. Bastug, M. Bennis, M. A. Kader et al., "Big Data Caching for Networking: Moving from Cloud to Edge," *IEEE Commun. Mag.*, vol. 54, no. 9, pp. 36–42, Sep. 2016.

[32] N. Golrezaei, A. F. Molisch, A. G. Dimakis et al., "Femtocaching and Device-to-Device Collaboration: A New Architecture for Wireless Video Distribution," *IEEE Commun. Mag.*, vol. 51, no. 4, pp. 142–149, Apr. 2013.

[33] M. Mahloo, P. Monti, J. Chen and L. Wosinska, "Cost Modeling of Backhaul for Mobile Networks," in *Proc. IEEE International Conference on Communications Workshops (ICC)*, Sydney, NSW, pp. 397–402, Jun 2014.

[34] G. Paschos, E. Bastug, I. Land, G. Caire and M. Debbah,"Wireless caching: technical misconceptions and business barriers," *IEEE Commun. Mag.*, vol. 54, no. 8, pp. 16–22, August 2016.

[35] X. Wang, M. Chen, T. Taleb, A. Ksentini, and V. C. M. Leung, "Cache in the Air: Exploiting Content Caching and Delivery Techniques for 5G Systems", *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 131–139, Feb 2014.

[36] V. Sourlas, P. Georgatsos, P. Flegkas, and L. Tassiulas, "Partition-based caching in information-centric networks," in *Proc. IEEE Intl. Workshop on Network Science for Commun. Networks (NetSciCom)*, Hong Kong, Apr. 2015, pp. 396–401.

[37] T. A. Shanmugasundaram and A. Nachiappan, "Impact of Doppler shift on the performance of RS coded non-coherent MFSK under Rayleigh and Rician fading channels," 2013 Int. Conf. Human Computer Interactions (ICHCI), pp. 1–5, Chennai, India, Aug. 2013.

[38] Simon, M. G. and M. S. Alouini, "Digital communications over fading channels," New York: John Wiley & Sons, 2000.

[39] B. Sklar, "Rayleigh fading channels in mobile digital communication systems. I. Characterization," *IEEE Commun. Mag.*, vol. 35, no. 7, pp. 90–100, July 1997.

[40] Gradshteyn, I. S. and I. M. Ryzhik, "Table of Integrals, Series, and Products, Orlando, FL: Academic Press," 5th ed., 1994.

[41] A. Tulino, A. Lozano, and S. Verdu, "Impact of antenna correlation on the capacity of multiantenna channels," *IEEE Trans. Inf. Theory*, vol 51, pp. 2491–2509, 2005.

[42] M. K. Ozdemir and H. Arslan, "Channel estimation for wireless ofdm systems,"*IEEE Commun. Surveys Tuts.*, vol. 9, no. 2, pp. 18–48, Second Quarter 2007.

[43] E. Bjornson, B. Ottersten, "A Framework for Training-Based Estimation in Arbitrarily Correlated Rician MIMO Channels with Rician Disturbance," *IEEE Trans. Signal Process.*, vol 58, pp. 1807–1820, 2010.

[44] C. Shin, R. W. Heath and E. J. Powers, "Blind Channel Estimation for MIMO-OFDM Systems," *IEEE Trans. Veh. Technol.*, vol. 56, no. 2, pp. 670–685, March 2007.

[45] J. Xiong, D. Ma, K. Wong and J. Wei, "Robust Masked Beamforming for MISO Cognitive Radio Networks With Unknown Eavesdroppers," *IEEE Trans. Veh. Technol.*, vol. 65, no. 2, pp. 744-755, Feb. 2016.

[46] G. Zheng, K.-K. Wong, and B. Ottersten, "Robust cognitive beamforming with bounded channel uncertainties," *IEEE Trans. Signal Process.*, vol. 57, no. 12, pp. 4871–4881, Dec. 2009.

[47] M. Bengtsson and B. Ottersten, "Optimal and suboptimal transmit beamforming," *Handbook of Antennas in Wireless Communications*, L. C. Godara, Ed. Boca Raton, FL, USA: CRC, Aug. 2001.

[48] P.-J. Chung, H. Du, and J. Gondzio, "A probabilistic constraint approach for robust transmit beamforming with imperfect channel information," *IEEE Trans. Signal Process.*, vol. 59, no. 6, pp. 2773–2782, Jun. 2011.

[49] Shannon, C. E. and W.Weaver, "A Mathematical Theory of Communication," Urbana, IL: Univ. of Illinois Press, 1949.

[50] B. Blaszczyszyn and A. Giovanidis, "Optimal geographic caching in cellular networks," in *Proc. IEEE Intl. Conf. on Commun. (ICC)*, Jun. 2015, pp. 3358–3363.

[51] N. Golrezaei, K. Shanmugam, A. G. Dimakis, A. F. Molisch, and G. Caire, "Femtocaching: Wireless video content delivery through distributed caching helpers,"in *Proc. IEEE INFOCOM*, pp. 1107–1115, March 2012.

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

[53] P. Grover and A. Sahai, "Shannon Meets Tesla: Wireless Information and Power Transfer," in *Proc. IEEE Sym. Inf. Theory (ISIT)*, June 2010, pp. 2363–2367.

[54] J. Xu, L. Liu and R. Zhang, "Multiuser MISO Beamforming for Simultaneous Wireless Information and Power Transfer," *IEEE Trans. Signal Process.*, vol. 62, no. 18, pp. 4798–4810, Sept. 2014.

[55] A. A. Nasir, X. Zhou, S. Durrani and R. A. Kennedy, "Wireless-Powered Relays in Cooperative Communications: Time-Switching Relaying Protocols and Throughput Analysis," *IEEE Trans. Commun.*, vol. 63, no. 5, pp. 1607–1622, May 2015.

[56] Q. Shi, L. Liu, W. Xu, and R. Zhang, "Joint transmit beamforming and receive power splitting for MISO SWIPT systems," *IEEE Trans. Wireless Commun.*, vol. 13, no. 6, pp. 3269–3280, Jun. 2014.

[57] J. Tang, D. K. C. So, A. Shojaeifard, K. Wong and J. Wen, "Joint Antenna Selection and Spatial Switching for Energy Efficient MIMO SWIPT System," *IEEE Trans. Wireless Commun.*, vol. 16, no. 7, pp. 4754–4769, July 2017.

[58] S. Timotheou and I. Krikidis, "Joint Information and Energy Transfer in the Spatial Domain with Channel Estimation Error," in *Proc. IEEE Online Conf. Green Commun.*, Oct. 2013, pp. 115–20.

[59] S. Goel and R. Negi, "Guaranteeing Secrecy Using Artificial Noise," *IEEE Trans. Wireless Commun.*, vol. 7, no. 6, pp. 2180–2189, Jun. 2008.

[60] S. Timotheou, I. Krikidis, G. Zheng, and B. Ottersten, "Beamforming for MISO Interference Channels with QoS and RF Energy Transfer," *IEEE Trans. Wireless Commun.*, vol. 13, no. 5, pp. 2646–2658, May 2014.

[61] X. Chen, Z. Zhang, H. h. Chen and H. Zhang, "Enhancing wireless information and power transfer by exploiting multi-antenna techniques," *IEEE Commun. Magazine*, vol. 53, no. 4, pp. 133–141, April 2015.

[62] Q. Shi, W. Xu, J. Wu, E. Song and Y. Wang, "Secure Beamforming for MIMO Broadcasting With Wireless Information and Power Transfer," *IEEE Trans. Wireless Commun.*, vol. 14, no. 5, pp. 2841–2853, May 2015.

[63] D. Hwang, D. I. Kim and T. Lee, "Throughput Maximization for Multiuser MIMO Wireless Powered Communication Networks," *IEEE Trans. Veh. Technol*, vol. 65, no. 7, pp. 5743–5748, July 2016.

[64] X. Zhou, R. Zhang, and C. K. Ho, "Wireless information and power transfer: Architecture design and rate-energy tradeoff," *IEEE Trans. Commun.*, vol. 61, no. 11, pp. 4754–4767, Nov. 2013.

[65] L. Liu, R. Zhang, and K. C. Chua, "Wireless information transfer with opportunistic energy harvesting," *IEEE Trans. Wireless Commun.*, vol. 12, no. 1, pp. 288–300, Jan. 2013.

[66] Q. Shi, W. Xu, T. H. Chang, Y. Wang and E. Song, "Joint Beamforming and Power Splitting for MISO Interference Channel With SWIPT: An SOCP Relaxation and Decentralized Algorithm," *IEEE Trans. Signal Process.*, vol. 62, no. 23, pp. 6194–6208, Dec. 2014.

[67] M. R. A. Khandaker and K. K. Wong, "SWIPT in MISO Multicasting Systems" *IEEE Wireless Commun. Lett.*, vol. 3, no. 3, pp. 277–280, June 2014.

[68] Z. Zhu, Z. Wang, X. Gui, and X. Gao, "Robust downlink beamforming and power splitting design in multiuser MISO SWIPT system," in *Proc. IEEE/CIC Int. Conf. Commun. China*, Oct. 2014, pp. 271–275.

[69] Z.-Q. Luo, W.-K. Ma, A. M.-C. So, Y. Ye, and S. Zhang, "Semidefinite Relaxation of Quadratic Optimization Problems," *IEEE Signal Process. Mag.*, vol. 27, no. 3, pp. 20–34, May 2010.

[70] F. Alizadeh and D. Goldfarb, "Second-order cone programming," *Math. Program. B*, vol. 95, no. 1, pp. 3–51, 2003.

[71] P. Richtarik and M. Takac, "Iteration Complexity of Randomized Block-Coordinate Descent Methods for Minimizing a Composite Function," *Math. Program.*, vol. 144, no. 1, pp. 1–38, 2014.

[72] E. Boshkovska, D. W. K. Ng, N. Zlatanov, A. Koelpin and R. Schober, "Robust Resource Allocation for MIMO Wireless Powered Communication Networks Based on a Non-Linear EH Model," *IEEE Trans. Commun.*, vol. 65, no. 5, pp. 1984–1999, May 2017.

[73] M. Grant, and S. Boyd, *CVX: Matlab software for disciplined convex programming*, version 2.0 beta. http://cvxr.com/cvx, Sep. 2013.

[74] J. Liao, M. R. A. Khandaker and K. K. Wong, "Robust Power-Splitting SWIPT Beamforming for Broadcast Channels," *IEEE Commun. Lett.*, vol. 20, no. 1, pp. 181–184, Jan. 2016.

[75] J. Wang, "A Survey of Web Caching Schemes for The Internet," *ACM Comp. Commun. Review*, vol. 29, no. 5, pp. 36–46, 1999.

[76] A. Passarella, "A Survey on Content-Centric Technologies for The Current Internet: CDN and P2P Solutions," *IEEE Commun. Surveys Tuts.*, vol. 35, no. 2012, pp. 1–32, 2011.

[77] V. Pacifici and G. Dan, "Distributed Caching Algorithms for Interconnected Operator CDNs," *IEEE Journal on Selected Areas in Commun.*, vol. 35, no. 2, pp. 380–391, Feb. 2017.

[78] G. Xylomenos, C. N. Ververidis, V. A. Siris, N. Fotiou, C. Tsilopoulos et al., "A Survey of Information-Centric Networking Research," *IEEE Commun. Surveys Tuts.*, vol. 16, no. 2, pp. 1024–1049, Apr. 2014.

[79] C. Fang, F. R. Yu, T. Huang, J. Liu, and Y. Liu, "A Survey of Green Information-Centric Networking: Research Issues and Challenges," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 3, pp. 1455–1469, Oct. 2015.

[80] D. Liu, B. Chen, C. Yang, and A. F. Molisch, "Caching at the Wireless Edge: Design Aspects, Challenges, and Future Directions," *IEEE Commun. Mag.*, vol. 54, no. 9, pp. 22–28, Sep. 2016.

[81] U. Niesen, D. Shah, and G. Wornell, "Caching in Wireless Networks," in *Proc. IEEE Intern. Sym. on Inf. Theory (ISIT)*, Seoul, Korea, Jun. 2009, pp. 2111–2115.

[82] J. Zhang, X. Lin and X. Wang, "Coded caching under arbitrary popularity distributions," in *Proc. IEEE Inf. Theory Applications Workshop (ITA)*, San Diego, CA, Feb. 2015, pp. 98–107.

[83] J. Zhang, X. Lin, C.-C. Wang, and X. Wang,"Coded caching for files with distinct file sizes," in *Proc. IEEE Sym. Inf. Theory (ISIT)*, Jun. 2015, pp. 1686–1690.

[84] S. Wang, W. Li, X. Tian, and H. Liu, "Fundamental limits of heterogenous cache," arXiv preprint arXiv:1504.01123v1.

[85] M. Ji, A. M. Tulino, J. Llorca and G. Caire, "On the average performance of caching and coded multicasting with random demands," in *Proc. IEEE Int. Sym. Wireless Commun. Systems (ISWCS)*, Aug. 2014, pp. 922–926.

[86] A. Sengupta, R. Tandon and T. C. Clancy, "Fundamental Limits of Caching with Secure Delivery," *IEEE Trans. Inform. Forensics Security*, vol. 10, no. 2, pp 355–370, January 2015.

[87] M. A. Maddah-Ali and U. Niesen, "Cache-aided interference channels," in *Proc. IEEE Sym. Inf. Theory (ISIT)*, Jun. 2015, pp. 809–813.

[88] M. Ji, G. Caire and A. F. Molisch, "Wireless device-to-device caching networks: Basic principles and system performance," *IEEE J. Select. Areas Commun.*, vol. 34, no. 1, pp. 176–189, Jan. 2016.

[89] R. Tandon and O. Simeone, "Cloud-aided wireless networks with edge caching: Fundamental latency trade-offs in fog Radio Access Networks," in *Proc. IEEE Sym. Inf. Theory (ISIT)*, Barcelona, Spain, July 2016

[90] N. Golrezaei, K. Shanmugam, A. G. Dimakis, A. F. Molisch and G. Caire, "FemtoCaching: Wireless Video Content Delivery through Distributed Caching Helpers," in *Proc. IEEE INFOCOM*, Orlando, FL, 2012, pp. 1107–1115.

[91] A. Khreishah and J. Chakareski, "Collaborative Caching for Multicell-Coordinated Systems," in *Proc. IEEE Conf. on Computer Commun. Workshops (INFOCOM WKSHPS)*, Hong Kong, Apr. 2015, pp. 257–262.

[92] M. A. Maddah-Ali and U. Niesen, "Coding for Caching: Fundamental Limits and Practical Challenges," *IEEE Commun. Mag.*, vol. 54, no. 8, pp. 23–29, Aug. 2016.

[93] C. Yang, Y. Yao, Z. Chen, and B. Xia, "Analysis on Cache-Enabled Wireless Heterogeneous Networks," *IEEE Trans. Wireless Commun.*, vol. 15, no. 1, pp. 131–145, Jan. 2016.

[94] K. Shanmugam, N. Golrezaei, A. G. Dimakis, A. F. Molisch, and G. Caire, "Femtocaching: Wireless Content Delivery through Distributed Caching Helpers," *IEEE Trans. Inf. Theory*, vol. 59, no. 12, pp. 8402–8413, Dec. 2013.

[95] M. Ji, G. Caire, and A. F. Molisch, "Fundamental limits of caching in wireless D2D networks,"*IEEE Trans. Inf. Theory*, vol. 62, no. 2, pp. 849869, Feb. 2016.

[96] M. Gregori, J. Gomez-Vilardebo, J. Matamoros, and D. Gunduz, "Wireless Content Caching for Small Cell and D2D Networks," *IEEE Journal on Selected Areas in Commun.*, vol. 34, no. 5, pp. 1222–1234, May 2016.

[97] 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, Sept. 2016.

[98] R. Tandon and O. Simeone, "Cloud-aided wireless networks with edge caching: Fundamental latency trade-offs in fog radio access networks," in *Proc. IEEE Intl. Sym. on Inform. Theory (ISIT)*, Barcelona, Spain, Jul. 2016, pp. 2029–2033.

[99] X. Peng, J. C. Shen, J. Zhang and K. B. Letaief, "Joint data assignment and beamforming for backhaul limited caching networks," in *Proc. IEEE Sym. Personal, Indoor and Mobile Radio Commun. (PIMRC)*, Washington DC, USA, Sept. 2014.

[100] A. Liu and V. K. N. Lau, "Mixed-timescale precoding and cache control in cached MIMO interference network," *IEEE Trans. Signal Process.*, vol. 61, no. 24, pp. 6320–6332, Dec. 2013.

[101] E. Bastug, M. Bennis, M. Kountouris, and M. Debbah, "Cache-enabled small cell networks: Modeling and tradeoffs," *EURASIP J. Wireless Commun. Netw.*, no. 1, pp. 1–11, Feb. 2015.

[102] C. Yang, Y. Yao, Z. Chen, and B. Xia, "Analysis on cache-enabled wireless heterogeneous networks," *IEEE Trans. Wireless Commun.*, vol. 15, no. 1, pp. 131–145, Jan 2016.

[103] Z. Chen, J. Lee, T. Q. Quek, and M. Kountouris, "Cooperative caching and transmission design in cluster-centric small cell networks," arXiv preprint arXiv:1601.00321, 2016.

[104] S. T. ul Hassan, M. Bennis, P. H. J. Nardelli, and M. Latva-aho, "Caching in wireless small cell networks: A storage bandwidth tradeoff," *IEEE Commun. Lett.*, vol. PP, no. 99, pp. 1175–1178, 2016.

[105] Y. Cui, D. Jiang, "Analysis and Optimization of Caching and Multicasting in Large-Scale Cache-Enabled Heterogeneous Wireless Networks," *IEEE Trans. Wireless Commun.*, vol.PP, no.99, pp.1–1.

[106] K. Poularakis, L. Tassiulas, "On the Complexity of Optimal Content Placement in Hierarchical Caching Networks," *IEEE Trans. Commun.*, vol. 64, no. 5, pp. 2092–2103, March 2016.

[107] R. Wang, X. Peng, J. Zhang, and K. B. Letaief, "Mobility-aware caching for content-centric wireless networks: Modeling and methodology," *IEEE Commun. Mag.*, vol. 54, no. 8, pp. 77–83, Aug. 2016.

[108] K. Poularakis, L. Tassiulas, "Exploiting User Mobility for Wireless Content Delivery," in *Proc. IEEE Sym. Inf. Theory (ISIT)*, Istanbul, Turkey, July 2013.

[109] K. Yang, Y. Shi, and Z. Ding, "Low-Rank Matrix Completion for Mobile Edge Caching in Fog-RAN via Riemannian Optimization," arXiv preprint arXiv:1608.07800.

[110] P. Blasco and D. Gunduz, "Learning-based optimization of cache content in a small cell base station," in *Proc. IEEE Commun. Conf. (ICC)*, Sydney, Australia, Jun. 2014.

[111] K. Hamidouche, W. Saad, and M. Debbah, "Many-to-many matching games for proactive social-caching in wireless small cell networks," in *Proc. of Wiopt, WNC3 Workshop*, Hammanet, Tunisia, May, 2014.

[112] M. Gregori, J. Gomez-Vilardebo, J. Matamoros and D. Gunduz, "Joint transmission and caching policy design for energy minimization in the wireless

backhaul link," in *Proc. IEEE Sym. Inf. Theory (ISIT)*, Hong Kong, China, Jun. 2015.

[113] L. Li, G. Zhao and R. S. Blum, "A Survey of Caching Techniques in Cellular Networks: Research Issues and Challenges in Content Placement and Delivery Strategies," *IEEE Commun. Surveys Tuts.*, March 2018.

[114] E. Bastug, M. Bennis, and M. Debbah, "Cache-Enabled Small Cell Networks: Modeling and Tradeoffs," in *Proc. 11th International Sym. on Wireless Communications Systems (ISWCS)*, Barcelona, Spain, Aug. 2014, pp. 649–653.

[115] B. Chen, C. Yang, and Z. Xiong, "Optimal Caching and Scheduling for Cache-Enabled D2D Communications," *IEEE Commun. Lett.*, vol. pp, no. 99, pp. 1–1, Jan. 2017.

[116] S.-C. Hung, H. Hsu, S.-Y. Lien, and K.-C. Chen, "Architecture Harmonization between Cloud Radio Access Networks and Fog Networks," *IEEE Access*, vol. 3, pp. 3019–3034, Dec. 2015.

[117] S.-H. Park, O. Simeone, and S. Shamai, "Joint Optimization of Cloud and Edge Processing for Fog Radio Access Networks," *IEEE Trans.Wireless Commun.*, vol. 15, no. 11, pp. 7621–7632, Nov. 2016.

[118] A. Shokrollahi, "Raptor Codes," *IEEE Trans. Inf. Theory*, vol. 52, no. 6, pp. 2551–2567, Jun. 2006.

[119] D. J. MacKay, "Fountain Codes," *IEE Proceedings Communications*, vol. 152, no. 6, pp. 1062–1068, Dec. 2005.

[120] M. A. Maddah-Ali and U. Niesen, "Decentralized coded caching attains order-optimal memory-rate tradeoff," *IEEE/ACM Trans. Netw.*, no. 99, 2014.

[121] R. Pedarsani, M. A. Maddah-Ali, and U. Niesen, "Online Coded Caching," *IEEE/ACM Trans. Networking*, vol. 24, no. 2, pp. 836–1040, Apr. 2016.

[122] M. Arlitt, L. Cherkasova, J. Dilley, R. Friedrich, and T. Jin, "Evaluating content management techniques for web proxy caches," *ACM SIGMETRICS Performance Evaluation Review*, vol. 27, no. 4, pp. 3–11, Mar. 2000.

[123] K. Poularakis, G. Iosifidis, V. Sourlas, and L. Tassiulas, "Exploiting caching and multicast for 5G wireless networks," *IEEE Trans. Wireless Commun.*, vol. 15, no. 4, pp. 2995–3007, Jan. 2016.

[124] Z. Chen and M. Kountouris, "Cache-Enabled Small Cell Networks with Local User Interest Correlation," in *Proc. IEEE 16th Intl. Workshop on Signal Process. Advances in Wireless Commun. (SPAWC)*, Stockholm, Sweden, Jun. 2015, pp. 680–684.

[125] K. Hamidouche, W. Saad and M. Debbah, "Many-to-Many Matching Games for Proactive Social-Caching in Wireless Small Cell Networks," in *Proc. 12th International Sym. on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt)*, Hammamet, 2014, pp. 569–574.

[126] F. Shen, K. Hamidouche, E. Bastug and M. Debbah, "A Stackelberg Game for Incentive Proactive Caching Mechanisms in Wireless Networks," in *Proc. Global Communications Conference (GLOBECOM)*, Washington, DC, Dec. 2016, pp. 1–6.

[127] A. Sengupta, S. Amuru, R. Tandon, R. M. Buehrer and T. C. Clancy, "Learning Distributed Caching Strategies in Small Cell Networks," in *Proc. 11th International Sym. on Wireless Commun. Systems (ISWCS)*, Barcelona, 2014, pp. 917–921.

[128] H. Tuy, *Convex Analysis and Global Optimization*, Boston, MA: Kluwer Academic, 2000.

[129] A. Bel-Tal and A. Nemirovski, *Lectures on Modern Convex Optimization. SIAM Series on Optimization*, Philadelphia, PA: SIAM, 2001.

[130] A. H. Phan, H. D Tuan, H. H. Kha and D. T. Ngo, "Nonsmooth Optimization for Efficient Beamforming in Cognitive Radio Multicast Transmission," *IEEE Trans. Signal Process.*, vol. 60, no. 6, pp. 2941–2951, June. 2012.

[131] K. Huang and E. Larsson, "Simultaneous information and power transfer for broadband wireless systems," *IEEE Trans. Signal Process.*, vol. 61, no. 23, pp. 5972–5986, Dec. 2013.

[132] X. Tang and Y. Hua, "Optimal design of non-regenerative MIMO wireless relays," *IEEE Trans. Wireless Commun.*, vol. 6, no. 4, pp. 1536–1276, Apr. 2007.

[133] Z. Fang, Y. Hua, and J. C. Koshy, "Joint source and relay optimization for a non-regenerative MIMO relay," in *Proc. IEEE Workshop Sensor Array Multi. Process.*, Waltham, USA, 12-14 Jul. 2006, pp. 239–243.

[134] Y. Rong and Y. Hua, "Optimality of diagonalization of multi-hop MIMO relays," *IEEE Trans. Wireless Commun.*, vol. 8, no. 12, pp. 6068–6077, Dec. 2009.

[135] Z. He, W. Jiang, and Y. Rong, "Robust design for amplify-and-forward MIMO relay systems with direct link and imperfect channel information," *IEEE Trans. Wireless Commun.*, vol. 14, no. 1, pp. 353–363, Jan. 2015.

[136] Y. Chen, Z. Wen, S. Wang, J. Sun, and M. Li, "Joint relay beamforming and source receiving in MIMO two-way AF relay network with energy harvesting," in *Proc. IEEE Veh. Technol. Conf. Spring*, Glasgow, Scotland, May 2015, pp. 1–5.

[137] K. Xiong, P. Fan, C. Zhang, and K. B. Letaief, "Wireless information and energy transfer for two-hop non-regenerative MIMO-OFDM relay networks," *IEEE J. Select. Areas Commun.*, vol. 33, no. 8, pp. 1595–1611, Aug. 2015.

[138] Y. Zeng and R. Zhang, "Full-duplex wireless-powered relay with self-energy recycling," *IEEE Wireless Commun. Lett.*, vol. 4, pp. 201–204, Apr. 2015.

[139] E. Telatar, "Capacity of multi-antenna gaussian channels," Technical Report, Bell Labs, 1995.

[140] Y. Cui, F. Lai, S. Hanly and P. Whiting, "Optimal caching and user association in cache-enabled heterogeneous wireless networks," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Washington DC, USA, Dec. 2016, pp. 1–6.

[141] E. Altman, K. Avrachenkov, and J. Goseling, "Coding for caches in the plane," arXiv preprint arXiv:1309.0604, 2013.

[142] D. Jiang, and Y. Cui, "Partition-based caching in large-scale SIC-enabled wireless Networks," arXiv preprint arXiv:1610.09526, 2016.

[143] V. Bioglio, F. Gabry, and I. Land, "Optimizing MDS codes for caching at the edge," in *Proc. IEEE GLOBECOM*, San Diego, USA, Dec. 2015.

[144] N. Abedini and S. Shakkottai, "Content caching and scheduling in wireless networks with elastic and inelastic traffic," *IEEE/ACM Trans. Network.*, vol. 22, no. 3, pp. 864–874, June 2014.

[145] J. Liao, K. K. Wong, M. R. A. Khandaker, and Z. Zheng, "Optimizing cache placement for heterogeneous small cell networks," *IEEE Commun. Lett.*, no.99, pp.1–1, Sept. 2016.

[146] 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, July 2016.

[147] A. Gharaibeh, A. Khreishah, B. Ji, M. Ayyash, "A Provably Efficient Online Collaborative Caching Algorithm for Multicell-Coordinated Systems," *IEEE Trans. Mobile Computing*, vol. 15, no. 10, pp. 1863–1876, aug. 2016.

[148] A. Khreishah, J. Chakareski and A. Gharaibeh, "Joint Caching, Routing, and Channel Assignment for Collaborative Small-Cell Cellular Networks," *IEEE J. Select. Areas in Commun. (JSAC)*, vol. 34, no. 8, pp. 2275–2284, Aug. 2016.

[149] F. Pantisano, M. Bennis, W. Saad and M. Debbah, "In-network caching and content placement in cooperative small cell networks," in *Proc. 1st Int. Conf. 5G for Ubiquitous Connectivity (5GU)*, pp.128–133, Nov. 2014.

[150] X. Peng, J. Zhang, S. H. Song and K. B. Letaief, "Cache size allocation in backhaul limited wireless networks," in *Proc. IEEE Int. Conf. Commun.(ICC)*, pp. 1–6, Kuala Lumpur, Malaysia, May 2016.

[151] Y. Wang, Z. Li, G. Tyson, S. Uhlig and G. Xie, "Design and Evaluation of the Optimal Cache Allocation for Content-Centric Networking," *IEEE Trans. Computers*, vol. 65, no. 1, pp. 95–107, Jan. 2016.

[152] K. Poularakis, G. Iosifidis and L. Tassiulas, "Approximation Algorithms for Mobile Data Caching in Small Cell Networks," *IEEE Trans. Commun.*, vol. 62, no. 10, pp. 3665–3677, Oct. 2014.

[153] D. Lopez-Perez, I. Guvenc, G. Roche, M. Kountouris, T. Quek and J. Zhang, "Enhanced intercell interference coordination challenges in heterogeneous networks," *IEEE Wireless Commun.*, vol. 18, no. 3, pp. 22–30, June 2011.

[154] L. Breslau, P. Cao, L. Fan, G. Phillips, and S. Shenker, "Web caching and Zipf-like distributions: Evidence and implications," in *Proc. of IEEE INFOCOM 99*, New York, U.S.A, Mar. 1999.

[155] GUROBI 6.5, GUROBI Optimization Inc., 2016.

[156] D. W. Matolak and R. Sun, "Unmanned aircraft systems: Air-ground channel characterization for future applications," *IEEE Vehic. Tech. Mag.*, vol. 10, no. 2, pp. 79–85, Jun. 2015.

[157] Y. Zeng, R. Zhang, and T. J. Lim, "Wireless communications with unmanned aerial vehicles: Opportunities and challenges," *IEEE Commun. Mag.*, vol. 54, no. 5, pp. 36–42, May 2016.

[158] P. Kamalinejad et al., "Wireless energy harvesting for the Internet of Things," *IEEE Commun. Mag.*, vol. 53, no. 6, pp. 102–108, Jun. 2015.

[159] S. Vassilaras and G. C. Alexandropoulos, "Cooperative beamforming techniques for energy efficient IoT wireless communication," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Paris, France, Jun 2017, pp. 1–6.

[160] H. Matsumoto and K. Hashimoto, "Report of the URSI inter-commission working group on SPS and appendices," Solar Power Satellite Syst. Gen. Assembly Sci. Symp. Int. Union Radio Sci., Ghent, Belgium, White Paper, 2006.