Advanced Multiple-Input and Multiple-Output Technology in Wireless Communication Networks

Jiajia Guo

A dissertation submitted to the Graduate Research School of
The University of New South Wales
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

School of Electrical Engineering and Telecommunications
Faculty of Engineering

January 2018

PLEASE TYPE

THE UNIVERSITY OF NEW SOUTH WALES
Thesis/Dissertation Sheet

Surname or Family name: Guo

First name: Jiajia Other name/s:

Abbreviation for degree as given in the University calendar: Doctor of Philosophy (Research)

School: Electrical Engineering and Telecommunications Faculty: Engineering

Title: Advanced Multiple-Input and Multiple-Output Technology in Wireless Communication Networks

Abstract 350 words maximum: (PLEASE TYPE)

Nowadays, the explosive data traffic demand has been craving innovative technologies for future wireless networks. Classic theory has revealed that the capacity of a multiple-input and multiple-output (MIMO) channel can increase linearly with the number of antennas. However, besides deploying multiple antennas at base stations, there are many challenges to develop MIMO to further boost the wireless network capacity.

In this thesis, advanced MIMO technologies are studied to exploit the degree of freedom gain under a variety of proposals for future wireless networks.

First, a new linear vector physical-layer network coding scheme is proposed for a MIMO two-way relay channel where the channel state information is unavailable at transmitters. We present an explicit network coding method that minimizes the error probability at high signal-to-noise ratios (SNRs). We propose a novel typical error event analysis and show that the proposed scheme achieves the optimal error rate performance at high SNRs. Numerical results show that the proposed scheme significantly outperforms existing schemes.

Second, a new caching scheme is proposed for a random wireless device-to-device (D2D) network, where each node is equipped with a local cache and intends to download files from a prefixed library via D2D links. The distributed MIMO technology is employed between source nodes and neighbours of the destination node for cache deliveries. The induced multiplexing gain and diversity gain increase the number of simultaneous transmissions, improving the network throughput. The average aggregate throughput scales almost linearly with the number of nodes, with a vanishing outage probability, and outperforms existing ones when the cache size is limited.

Third, a hybrid D2D-cellular scheme is proposed to make use of the standby users who possess D2D communication capabilities in close proximity to each other, and to improve the rate performance for cellular users. Through D2D links, a virtual antenna array is formed by sharing antennas across different terminals to realize the diversity gain of MIMO channels. We then design an orthogonal D2D multiple access protocol and formulate the optimization problem of joint cellular and D2D resource allocation. Extensive system-level simulations demonstrate that the cellular rate performance is significantly improved.

FOR OFFICE USE ONLY Date of completion of requirements for Award:

ORIGINALITY STATEMENT

'I hereby declare that this submission is my own work and to the best of my knowledge it contains no materials previously published or written by another person, or substantial proportions of material which have been accepted for the award of any other degree or diploma at UNSW or any other educational institution, except where due acknowledgement is made in the thesis. Any contribution made to the research by others, with whom I have worked at UNSW or elsewhere, is explicitly acknowledged in the thesis. I also declare that the intellectual content of this thesis is the product of my own work, except to the extent that assistance from others in the project's design and conception or in style, presentation and linguistic expression is acknowledged.'

Signed ……..

Date ……..

ABSTRACT

Nowadays, the explosive data traffic demand has been craving innovative technologies for future wireless networks. Classic theory has revealed that the capacity of a multiple-input and multiple-output (MIMO) channel can increase linearly with the number of antennas. However, besides deploying multiple antennas at base stations, there are many challenges to develop MIMO to further boost the wireless network capacity.

In this thesis, advanced MIMO technologies are studied to exploit the degree of freedom gain under a variety of proposals for future wireless networks.

First, a new linear vector physical-layer network coding scheme is proposed for a MIMO two-way relay channel where the channel state information is unavailable at transmitters. We present an explicit network coding method that minimizes the error probability at high signal-to-noise ratios (SNRs). We propose a novel typical error event analysis and show that the proposed scheme achieves the optimal error rate performance at high SNRs. Numerical results show that the proposed scheme significantly outperforms existing schemes.

Second, a new caching scheme is proposed for a random wireless device-to-device (D2D) network, where each node is equipped with a local cache and intends to download files from a prefixed library via D2D links. The distributed MIMO technology is employed between source nodes and neighbours of the destination node for cache deliveries. The induced multiplexing gain and diversity gain increase the number of simultaneous transmissions, improv-

ing the network throughput. The average aggregate throughput scales almost linearly with the number of nodes, with a vanishing outage probability, and outperforms existing ones when the cache size is limited.

Third, a hybrid D2D-cellular scheme is proposed to make use of the standby users who possess D2D communication capabilities in close proximity to each other, and to improve the rate performance for cellular users. Through D2D links, a virtual antenna array is formed by sharing antennas across different terminals to realize the diversity gain of MIMO channels. We then design an orthogonal D2D multiple access protocol and formulate the optimization problem of joint cellular and D2D resource allocation. Extensive system-level simulations demonstrate that the cellular rate performance is significantly improved.

ACKNOWLEDGEMENTS

During my four year Ph.D study, I always feel lucky to get such great help and support, and there are so many to thank. First of the first, I would like to thank my supervisor Prof. Jinhong Yuan, who gave me the opportunity to pursue this degree in the first place, and has been always supporting me on my study all these years. He guided me into good research topics, helped me formulate research problems and was generous to share his time on my doubts and confusions. His constant hunger for innovative technologies and solid research results, and his gentle while firm way of handling people and staff have deeply impressed and inspired me.

I would like to thank my collaborator A/Prof. Andrew J. Zhang, who has been closely working with me and has given me many good suggestions on my research. He is so kind and thoughtful that each time he proposed to help me before I even asked. His hard working and dedication also impressed me a lot.

I am also grateful to work with Prof. Wei Yu in University of Toronto, Canada. He is so gifted and so hard working at the same time, and his remarkable perception on research and his efficient working style have been invaluable.

I also would like to thank Dr. Tao Yang, who had supported me in my first year of Ph.D study when I was the most confused in the very beginning.

I would also like to thank all my friends and colleagues at the Wireless Communication Lab at UNSW, for many discussions we had leading to a better

understanding of wireless communication theories, and the great pleasure we have working together.

I cannot thank enough to my family. None of this would happen without them.

Contents

Abbreviations

5G	The fifth generation.
AF	Amplify-and-forward.
AOAs	Angles of arrive.
AWGN	Additive white Gaussian noise.
BS	Base station.
CDF	Cumulative distribution function.
CF	Compute-and-forward.
CPO	Carrier-phase offset.
CSI	Channel state information.
CSIT	Transmitter-side channel state information.
dB	Decibel.
DF	Decode-and-forward.
D2D	Device-to-device.
EE	Energy efficiency.
eMBB	Enhanced mobile broadband.
FDMA	Frequency division multiple access.
i.i.d	Independent and identically distributed.
KKT	Karush-Kuhn-Tucker.
LOS	Line-of-sight.
MBS	Macro-cell base station.
MEC	Mobile edge computation.
MIMO	Multiple input multiple output.

MISO	Multiple-input single-output.
MMSE	Minimum mean square error.
mMTC	Massive machine type communications.
mmWave	Millimeter wave.
MSE	Mean square error.
MWRC	Multi-way relay channel.
NC	Network coding.
NOMA	Non-orthogonal multiple access.
PAM	Pulse-amplitude modulated.
PDF	Probability density function.
PNC	Physical layer network coding.
QAM	Quadrature amplitude modulation.
QF	Quantize-and-forward.
QoS	Quality of service.
s.t.	Subject to.
SBS	Small-cell base station.
SDMA	Space-division multiple access.
SE	Spectral efficiency.
SIC	Successive interference cancellation.
SINR	Signal-to-interference-plus-noise ratio.
SISO	Single-input single-output.
SNR	Signal-to-noise ratio.
SVD	Singular value decomposition.
TDMA	Time division multiple access.
TWRC	Two-way Relay Channel.
URLLC	Ultra-Reliable and Low Latency Communications.
V2V	Vehicle-to-vehicle.
WMMSE	Weighted minimum mean square error.

List of Notations

Boldface upper-case letters denote matrices, boldface lower-case letters denote vectors, and italics denote scalars.

$\mathbf{x}$	A column vector.
$\mathbf{X}$	A matrix.
$\mathbf{X}^T$	Transpose of $\mathbf{X}$.
$\mathbf{X}^*$	Conjugate transpose of $\mathbf{X}$.
$\mathbf{X}^\dagger$	Pseudo inverse of $\mathbf{X}$.
$\det(\mathbf{X})$	Determinant of $\mathbf{X}$.
$\mathrm{Tr}(\mathbf{X})$	Trace of $\mathbf{X}$.
$\lvert x \rvert$	Absolute value (modulus) of the scalar x.
$\lVert \cdot \rVert$	Frobenius norm of a vector or a matrix.
$\mathbf{0}$	Zero matrix. A subscript can be used to indicate the dimension.
$\mathbf{I}$	Identity matrix. A subscript can be used to indicate the dimension.
$\mathbb{E}[\cdot]$	Statistical expectation.
$\mathbb{C}$	Complex number set.
$\mathcal{CN}$	Complex Gaussian distribution.
$\mathcal{CN}$	Complex Gaussian distribution.
$\ln(\cdot)$	Natural logarithm.
$\Pr(\cdot)$	Probability.
$\exp(\cdot)$	Exponential function.
$\lim$	Limit.
$\max\{\cdot\}$	Maximization.
$\min\{\cdot\}$	Minimization.

List of Figures

List of Publications

Journal Articles:

1. **Jiajia Guo**, W. Yu and J. Yuan, "Enhancing Cellular Performance through Device-to-Device Distributed MIMO, " submitted to *IEEE Transactions on Wireless Communications*, 2017.

2. **Jiajia Guo**, Tao Yang, Jinhong Yuan and Jian A. Zhang, "A Novel Linear Physical-layer Network Coding Scheme for Y-channel without Transmitter CSI," submitted to *IEEE Transactions on Vehicle Technology*, 2017.

3. **Jiajia Guo**, J. Yuan and Jian A. Zhang, "The Throughput Scaling Law of Wireless Device-to-device Caching Networks with Distributed MIMO and Hierarchical Cooperations," *IEEE Transactions on Wireless Communications*, vol. 17, no. 1, pp. 492-505, Jan. 2018.

4. **Jiajia Guo**, Tao Yang, Jinhong Yuan and Jian A. Zhang, "Linear Vector Physical-layer Network Coding for MIMO Two-Way Relay Channels: Design and Performance Analysis," *IEEE Transactions on Communications*, vol. 63, no. 7, pp. 2591-2604, July 2015.

Conference Articles:

1. Zhiqiang Wei, Lou Zhao, **Jiajia Guo**, Wing Kwan Ng and Jinhong Yuan, "A Multi-Beam NOMA Framework for Hybrid mmWave Systems," *IEEE ICC* 2018.

2. **Jiajia Guo**, J. Yuan and Jian A. Zhang, "Wireless Device-to-device Caching Networks with Distributed MIMO and Hierarchical Cooperations," *IEEE GlobeCom* 2017.

3. Zhiqiang Wei, **Jiajia Guo**, Derrick Wing Kwan Ng, Jinhong Yuan, "Fairness Comparison of Uplink NOMA and OMA," *IEEE VTC* 2017 Spring.

4. **Jiajia Guo**, Tao Yang, Jinhong Yuan and Jian A. Zhang, "Linear Physical-layer Network Coding for the Fading Y-channel without Transmitter Channel State Information," *IEEE VTC* 2016 Fall.

5. **Jiajia Guo**, Tao Yang, Jinhong Yuan and Jian A. Zhang, "Design of Linear Physical-layer Network Coding for MIMO Two-way Relay Channels without Transmitter CSI," *IEEE WCNC* 2015.

6. **Jiajia Guo**, Tao Yang and Jinhong Yuan, "Multivariate Linear Physical-layer Network Coding for MIMO Two-way," Poster in *AusCTW*, 2014.

Chapter 1

Overview

1.1 Introduction

Driven by an explosion of personal computers, tablets and smart phones, the demand of data traffic has increased by a hundred times in the past decade and an accelerating trend is expected by 2020 [1]. In addition to the data volume, the data rates and the number of devices will continue to grow exponentially. New applications such as virtual reality (VR), augmented reality (AR), and immersive gaming are expected to consume even more bandwidth and sustained up to Gbps connections. The number of devices also could reach tens or even hundreds of billions, due to many new applications on automotive, transport, security and healthcare [2,3]. Industrial and academia societies have reached a consensus that the fifth generation (5G) communications will target at three difference classes of user cases: eMBB (enhanced Mobile Broadband), mMTC (massive Machine Type Communications) and URLLC (Ultra-Reliable and Low Latency Communications) [4–6].

Under this context, incremental improvements for current wireless networks are genuinely insufficient and it is crucial for researchers to develop innovative new technologies to meet these intense demands.

Classic theory has revealed great benefits of equipping multiple antennas in wire-

less communications [7]. Especially, the capacity of a multiple-input and multiple-output (MIMO) channel can increase linearly with the number of antennas in a rich scattering environment. This stupendous improvement can benefit wireless networks by providing high data rate, relieving the spectrum deficiency (SE), and enhancing the ability of encountering interference.

For the time being, evolution of mobile broadband (i.e., developing eMBB) will remain most important because ongoing growth in demand for it is proven to be strong and commercially profitable [4]. Among the many proposals to 5G, increasing SE/EE through advances in MIMO, to support more bits/s/Hz per node has been identified as one of the key technologies to get a 1000× data rate for 5G communications [8].

However, other than the traditional method of deploying multiple antennas at base stations, there are many challenges to further apply the MIMO technology in future wireless networks.

One challenge is how to apply MIMO in the relay channels, where using relays has been proven to be able to provide robustness against channel variations, extend coverage and improve energy efficiency (EE) [9]. In relay channels, signals from multi-users are superimposed at the relay, which brings interferences between different users on top of interferences between different data streams induced by MIMO multiplexing. This imposes as a major challenge in relay channels in terms of how to efficiently manage interferences and improve SE/EE of the network.

Another difficulty is that due to the microwave carrier frequency for current wireless networks, the interspace requirement of deploying multiple antennas according to the electromagnetic prorogation theory cannot be satisfied at the user side, given the physical size of common end-user equipments. Without multiple antennas at the user-side, a full MIMO channel capacity can not be realized in many cases.

Therefore, this thesis is dedicated to ingeniously employing MIMO in promising proposals for next generation communications.

First, physical-layer network coding (PNC) has been proven to be able to largely increase the network throughput of multi-user wireless communication networks [10–

13]. For example, PNC can double the network throughput for a two-way relay channel (TWRC), in which two users exchange information via an intermediate relay simultaneously [14–16]. Instead of treating one user's signal as interference or completely decoding the two users's messages, the relay generates network coding (NC) messages and broadcasts them to users. Each user can resolve the intended message using the knowledge of its own messages. However, when considering MIMO PNC for TWRCs, new challenges for a MIMO PNC scheme appear such as how to generate NC messages at the relay so that the MIMO capacity of the single user-to-relay transmission can be retained or not degraded from the other user's concurrent transmission, and also the multiple data streams can be recovered at the destination. This thesis tackles this problem for the first time.

Second, wireless caching has been proposed as a cost-effective way to handle the high traffic requirement caused by content delivery applications, especially on-demand video streaming [17–21]. Especially, in a cached device-to-device (D2D) network, each device is equipped with a local cache and would like to download its requested file from a pre-fixed library through D2D links. This setup is also applicable for mobile edge computing and the content distribution network. For such a network, most works focused on exploiting spatial reuse of concurrent multiple short range D2D transmissions. However, the spatial degrees of freedom in a MIMO transmission are not exploited in these works. The question of how the caching network capacity would scale by exploiting spatial degrees of freedom is not answered so far, and is thoroughly studied in this thesis. Moreover, compared with the vast library that users may request, the local cache size is rather limited in realistic networks. This conscious also motivates us to design a caching scheme for the small cache case.

Last but not least, the integration of D2D communications into a traditional cellular network has been proposed as one of the promising technologies for 5G networks [22, 23]. Studies have shown that this hybrid network can significantly increase network SE and EE and alleviate the core network congestion [24]. It has been recognized that in future cellular networks, there will be numerous standby users

possessing D2D communication capabilities in close proximity to each other, while these standby users do not necessarily request D2D communications all the time. This D2D capability balance can be used to improve cellular performance. While most of existing works proposing using the D2D link as a multi-hop relay, this thesis proposes to use D2D communications to realize the diversity gain of MIMO channels, therefore enhancing the capability of encountering the inter-cell interference.

To summarize, this thesis aims to study advanced MIMO technology to exploit the degrees of freedom gain under a variety of proposals for future wireless networks.

1.2 Background and Motivations

1.2.1 MIMO Theory

The concept of MIMO can be traced back to researches in 1970s reducing the interference (crosstalk) between multiplexed pulse-amplitude modulated (PAM) signals in multi-channel digital transmission systems, such as multi-pair cables and multi-terminal systems [25, 26]. Although these works did not exploit multi-path propagation in a wireless channel, some of the mathematical techniques in dealing with mutual interference are proved to be useful thereafter. In the mid-1980s, Bell Laboratories took this research a step further by introducing the MIMO concept to multi-user systems with time-division multiplexing.

The idea of improving the performance of cellular wireless networks by MIMO was first consolidated in the early 1990s through space-division multiple access (SDMA) [27, 28]. Using directional or smart antennas to communicate with users in different locations within the range of the same base station, the multiple antennas allow spatial separation of the signals from the different users. In the mid 1990s, it has been further observed that a similar effect can occur for a point-to-point channel with multiple transmit and receive antennas, even when the transmit antennas are not geographically far apart [29]. The mathematical expression of this result is presented

as follows.

Let matrix $\mathbf{H} \in \mathcal{C}^{N \times M}$ represent a narrow band wireless channel with M transmit and N receive antennas, whose element $h_{i,j}$ is the channel gain from transmit antenna j to receive antenna i. Then, the corresponding channel is described by

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{n}, \tag{1.1}$$

where $\mathbf{x} \in \mathcal{C}^{M \times 1}$, $\mathbf{y} \in \mathcal{C}^{N \times 1}$ and $\mathbf{n} \sim \mathcal{CN}(0, N_0\mathbf{I_N})$ denote the transmitted signal, received signal and white Gaussian noise respectively. Note that the transmitted signals must satisfy a total power constraint P.

For a deterministic MIMO channel model, applying a method of singular value decomposition (SVD), the capacity of a time-invariant MIMO channel with known channel state information (CSI) at both the transmitter and the receiver is [30]

$$\mathbf{C} = \sum_{i=1}^{n_r} \log\left(1 + \frac{P_i \lambda_i^2}{N_0}\right), \tag{1.2}$$

where n_r is the rank of matrix $\mathbf{H}$, λ_i is the i-th ordered singular value of $\mathbf{H}$ and P_i is the allocated power on the i-th eigenchannel. Each non-zero eigenchannel can support a data stream and thus the MIMO channel can support n_r spatial multiplexing streams. When the scattering environment is rich enough (i.e., $\mathbf{H}$ is not ill-ranked), having multiple antennas provides a degree-of-freedom gain of $\min\{M, N\}$ [7].

For a statistical MIMO channel model, the capacity of an $M \times N$ i.i.d. Rayleigh fading MIMO channel requiring only receiver CSI is [31]

$$\mathbf{C} = \log\det\left(1 + \frac{PHH^{\dagger}}{MN_0}\right), \tag{1.3}$$

At a high signal-to-noise ratio (SNR) regime, the capacity can be approximated as $\min\{M, N\}\log\left(\frac{P}{N_0}\right)$ bits/s/Hz.

Under both situations, an additional degree of freedom $\min\{M, N\}$ is provided and can be exploited by spatially multiplexing $\min\{M, N\}$ data streams. In other words, the capacity of a MIMO channel with N transmit and receive antennas is proportional to N and can linearly increase with N. Thus, MIMO techniques have

been widely used as a primary tool to significantly increase capacity in the high SNR regime.

1.2.2 Physical-layer Network Coding

The PNC technology is able to improve the throughput and reliability of some multi-user wireless communication networks [10–13].

The simplest setup for PNC is a two-way relay channel (TWRC), where two users exchange information via an intermediate relay simultaneously. Compared with conventional schemes such as amplify-forward and decode-forward, PNC allows the relay to efficiently reconstruct and forward appropriate functions of the users messages. In this way, considerable coding gain and degrees of freedom gain may be achieved.

The relaying strategy

In a wireless network, signals transmitted from a single node can be heard not only by the intended receiver, but also by other nearby nodes. In traditional communications, signals from unintended users in the same signal domain (in time, frequency or space) are usually regarded as interference. Techniques on interference avoidance have to be used to ensure the intended signal stand out against interferences. However, these techniques usually come with diminishing rates. Recent work on cooperative communications has shown that this penalty can be overcome by adopting new strategies at the physical layer. The key idea is that users should help to relay each other's messages by exploiting the broadcast and multiple-access properties of the wireless medium [32]. Most proposed schemes can be categorized into the following three relaying strategies:

- Amplify-and-Forward (AF): The relay sends an amplified version of the received signal to the next receiver, maintaining a fixed average transmit power [33, 34].
- Quantize-and-Forward (QF): The relay transmits a scaled version (or vector

quantized, which is also referred to as compress-and-forward) of its observation [35, 36].

- Decode-and-Forward (DF): The relay decodes the source message and transmits the re-encoded message to the next receiver [37].

- Compute-and-Forward (CF): The relay computes some network-coded message based on the received messages. With specific code design, the destination will be able to recover the messages [38].

The AF strategy is relatively easy to implement and requires much less computing power as no decoding or quantizing operation is performed at the relay [33,34]. However, the noise at the relay is also amplified which is a main reason of the restricted performance. Moreover, in a TWRC, the power of transmitting the other user's signal is wasted when performing the substraction operation at the destination user. Fig. 1.1 shows a outage rate comparison of DF, QF and CF strategies in a MIMO i.i.d. Rayleigh slow fading channel [38]. From this figure, we observe that CF performs the best in the moderate SNR regime and QF is a good strategy at the low SNR regime. This is because when the SNR is low, the effective noise introduced by vector quantization at the relays is not significant as in high SNRs. DF is not as efficient as CF at the high SNR regime because the relay must either treat one of the messages as noise or decode both. To summarize, compared with existing relay strategies, CF is the best strategy in the moderate SNR regime.

Note that the above relay strategies are not only important in the classic relay scenarios, but also apply to many advanced network architectures such as the cloud radio access network (Cloud-RAN). For example, in Cloud-RANs, the BSs are connected to a cloud-computing-based central processor (CP) with backhaul links with finite capacities [6]. Here, each BS in Cloud-RANs can be regarded as a relay from the user to the CP, and many researches have been done adopting QF and CF respectively [39, 40]. Our work is a fundamental research on the CF relay strategy

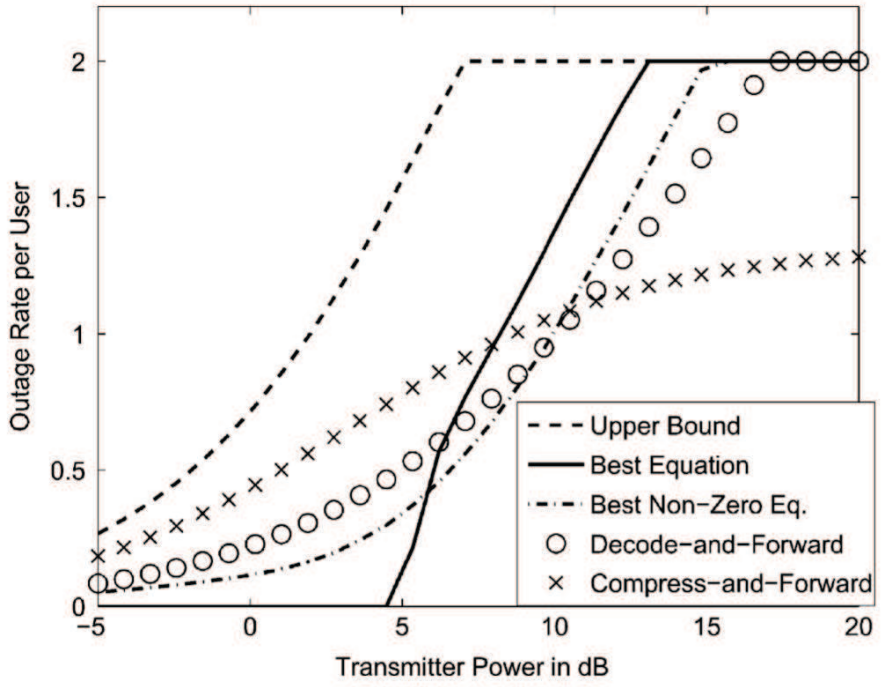


Figure 1.1. Comparisons between Quantize-and-Forward (Compress-and-Forward), Decode-and-Forward and Compute-and-forward (referred to as "Best Non-Zero Eq." in the figure) in [38]. Symmetric outage rates for the 2×2 MIMO multiple access channel with i.i.d. Rayleigh fading only known at the receivers with per channel use rate of 2 and outage probability of $\frac{1}{4}$.

and provides theoretical references and insights for the BS operation design in future Cloud-RAN. In addition, there have been studies trying to solve the collision problem of massive connectivity by using the concept of PNC and the method in our work [41]. This shows that PNC is an important research topic and has a wide range of application under a variety of scenarios.

Physical-layer network coding basics

The pioneering work of PNC is first presented by [10] in 2006 and it has been shown to be able to dramatically improve the throughput and reliability in a multi-way relay channel (MWRC) [11, 12].

The basic idea of PNC can be summarized as follows. Consider a TWRC where user A and B want to exchange messages via a relay node as shown in Fig. 1.2. Each

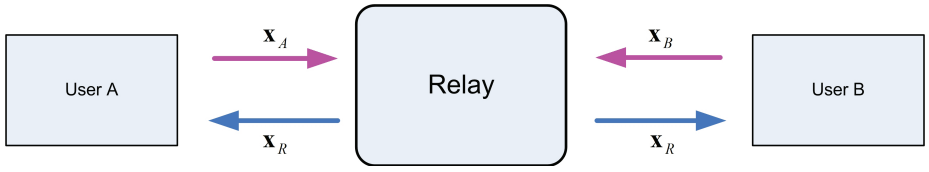


Figure 1.2. System model for a two way relay channel.

round of information exchange runs in two equal-duration time-slots, referred to as the uplink phase and the downlink phase.

In the uplink phase, both users transmit signals x_A and x_B respectively, simultaneously to the relay. In the downlink phase, the relay broadcasts a signal vector x_R (or a signal scaler under the single-antenna setup) which contains some function of the two users' signals, i.e.

$$x_R = f\left(x_A, x_B\right). \tag{1.4}$$

Upon receiving x_R, each user extracts the other user's information by exploiting the perfect knowledge of its own message. This finishes one round of message exchange. The original work on PNC suggests using a XOR function of the two users's messages, which is a well-known function for PNC operations. In fact, $f(.)$ could be linear function [42, 43] or non-linear function [12] [44], according to different system setups and designs.

From the information theoretic perspective, PNC agrees with the spirits of CF and has been shown that PNC can achieve within 1/2 bit of the capacity of a Gaussian TWRC and is asymptotically optimal at high SNRs [14, 15].

Existing works and challenges: PNC for TWRC without transmitter-side CSI

Despite its potential advantages in theory, the design of a high performance PNC scheme in a practical fading wireless channel environment remains as a challenging task. The main difficulty arises from the fact that the signals transmitted by the two users will arrive at the receiver with a relative carrier-phase offset (CPO). For

an ideal scenario where the CSI is globally known, the CPO may be compensated at the transmitter sides, as assumed in most existing works [45–51]. However, such a compensation requires very high-precision power amplifier and carrier-phase shifter, which will be very demanding to implement [52].

In many practical scenarios where the CSI is not globally known at the transmitters, it is desirable to design reliable PNC for TWRC without transmitter-side CSI (CSIT). However, it is known to be a challenging task, even for the single-input single-output (SISO) scenario [52, 53]. This is primarily due to the existence of CPO between the two users. In such a scenario, the error probability performance of the conventional PNC can be significantly degraded.

In [52], the authors proposed an improved belief propagation algorithm for the case where the CPO cannot be compensated at the transmitters. The performance penalty caused by the CPO can be alleviated by utilizing the feature of a belief propagation algorithm. However, this result only applies to a regime with a low-level modulation and a low-rate channel code. In that regime, the gain of PNC is known to be quite limited [14, 15]. In [12], the authors proposed an adaptive PNC modulation design method to optimize the signal constellation at the relay. However, due to the non-linear PNC mapping they employed, their result is largely based on numerical exhaustive search, which has a high computational complexity and lacks scalability. The work in [54] addressed the design of irregular repeat accumulate (IRA) codes for PNC under the binary-input Gaussian TWRC, and the constructed IRA codes have considerable performance improvement over the existing codes. However, their results cannot be extended to the fading TWRC. The CF theoretical framework introduces lattice code as a powerful tool and provides a potential solution to this problem when transmitters lack CSI [38]. Following this spirit, the authors in [55] proposed a lattice coded PNC for the TWRC, where two users employ the same lattice codebook and the relay attempts to decode the lattice coded network codes of the two users messages by using an iterative belief propagation decoder. Yet, results on the practical coding and modulation design for CF remain limited.

Recently, a linear PNC scheme was proposed for a SISO TWRC [43, 56, 57] without CSIT. It focused on the optimization of the PNC constellation by exploiting the structure of linear PNC mapping. This scheme can be regarded as a practical embodiment of the information theoretic concept of CF. Compared to the non-linear PNC scheme in [12], the linear PNC scheme offers low computational complexity, scalability, and clear insight for the optimized design solution. By carefully designing the linear PNC coefficients, the error probability of the scheme is minimized, leading to significantly improved error performance over conventional PNC schemes.

In contrast to the former work [16, 44, 46, 50] assuming perfect global CSI, this thesis will focus on the setup where CSI is not available at the transmitters, which is a more practical assumption. This applies to practical scenarios where there is no feedback from the receiver to the transmitter or the channel reciprocity does not hold.

Existing works and challenges: MIMO PNC for TWRC

Nowadays, wireless communication systems and devices are featuring multi-antenna and MIMO techniques. From the literature, PNC has been extended to MIMO TWRCs and is shown to dramatically outperform non-PNC schemes in terms of achievable information rate [16, 45–51, 58]. However, perfect global CSI at the transmitters was assumed in these works. Specifically, joint precoding was employed therein to strictly align the signal directions of the two users to support PNC. Unfortunately, it is known that the assumption of global CSI at the transmitters is a strong one in practice and delivering the full CSI requires considerable overhead [44, 59].

So far, the design of a reliable PNC scheme for a MIMO TWRC without CSIT remains as a challenging task. This is primarily due to 1) each user transmits multiple streams from its multiple antennas and 2) the transmitted signals will arrive at the relay with different signal directions. Results in this direction are quite limited.

In [44], the authors considered the non-linear PNC scheme in [12] for the MIMO scenario. Owing to the non-linear PNC mapping, the scheme in [44] also has a high computational complexity and lacks scalability and design insight. In [60] and [61],

the authors consider TWRCs with two antennas at the relay and one antenna at each user, and proposed a PNC scheme based on the QR decomposition of the channel matrix. However, no spatial multiplexing can be carried out in such setting and only a diversity order of one is achieved. Also, they both use quadrature phase shift keying (QPSK) modulation while further work using high-level modulations should be considered. The work on CF in [38] computes the NC codewords independently at each antenna, while jointly generating multiple NC codewords based on all received signals could be considered. The integer-forcing linear receiver studied in [62] inherits the notion of CF in designing an efficient detection algorithm for a point-to-point MIMO system. However, the focus of that work is on reducing the complexity of a linear MIMO detector, rather than dealing with efficient communication over a MIMO TWRC.

The linear PNC scheme for the SISO TWRCs in [43, 56, 57] has shown advantages over the non-linear scheme. Yet, the generalization to the MIMO case is not straightforward. The main challenges are:

1) the SISO linear PNC scheme re-constructs a single NC message from a one dimension received signal space. However for spatial multiplexing MIMO scenario, in order that each user can recover the desired multiple information streams from the other user, one NC message determined by one NC coefficient vector is not enough. Instead, it is essential for the relay to re-construct a vector of multiple NC messages by designing multiple NC coefficient vectors. Therefore, it is challenging to find the multiple vectors of NC coefficients that guarantee that the information streams can be recovered by each user at the destination and the decoding error probability is minimized;

2) For MIMO TWRCs with PNC, in the uplink phase, each information stream does not only experience interference from the other user as in SISO TWRCs. It is also interfered by other streams of the same user. Thus, new characterization of the superposition of the two user's multiple signal streams in a multi-dimensional received signal space is required to analyze the performance of the proposed scheme.

To summarize, it is desirable and challenging to design a reliable PNC scheme for MIMO TWRCs without CSIT. Section 2.2 - 2.6 will present my work on MIMO PNC for TWRCs.

Existing works and challenges: PNC for MWRC

Beyond TWRCs, there have also been many investigations on MWRCs in which a relay (or a system of relays) interconnect more than two end nodes [63–65]. Multi-way communications were first studied by Shannon in [66] where a two-way channel was considered. The concept of MWRCs is proposed by combining relaying and multi-way communications. This TWRC setup was studied in [63,67] leading to an approximate characterization of the capacity region of the Gaussian case. The MWRC with more nodes was also studied in [9] in a multi-cast scenario. A similar setup, where all users belong to the same cluster and all channel gains are equal was studied in [68], and the sum-capacity of this Gaussian setup with more than two users was obtained.

A broadcast variant of this multi-way relaying setup, the so called Y-channel, was considered in [69]. In the Y-channel, three users attempt to communicate with each other via a relay. Each user sends two independent messages, one to each other user. Each node in the Y-channel is thus a source of two messages and a destination of two messages. Studies on this area are intensely focused on achievable degrees of freedom using signal alignment [70–72]. For example, in [70], the setup where the nodes have multiple antennas was considered. A transmission scheme exploiting interference alignment was proposed, and its corresponding achievable degrees of freedom were calculated. Note that the capacity of the Y-channel is not known in general. Most existing works, especially those on the degrees of freedom for Y-channel, require global channel state information (CSI) at each node [73–75]. Unfortunately, global CSI requires considerable overhead cost in real networks, and studies without CSIT are important for real networks.

In particular, the no-CSIT Y-channel setup can be very useful in vehicle-to-vehicle (V2V) communications, where information exchange between vehicles can be achieved

with reduced latency and increased coverage through either relaying base-stations or vehicles [76]. In such communication scenarios, the network topology is usually dynamic and changes rapidly, wherein a no-CSIT scheme can significantly reduce the overhead cost.

When considering a PNC scheme for Y-channel, with three users transmitting to the relay simultaneously, the CPO between all the received signals at the relay is more difficult to deal with than that in TWRCs. Thus, designing a Y-channel PNC scheme not requiring CSIT is quite challenging. Section 2.7 will present my work on the Y-channel PNC by further extending the idea of linear PNC.

1.2.3 Wireless Caching Networks

With the wireless data traffic predicted to increase dramatically in the next few years, tremendous efforts of academia and industry have been made to satisfy this explosive demand while minimizing the energy expenditure. Among the many proposals trying to solve this issue, wireless caching has been proposed as a cost-effective way to handle the high traffic requirement [17–19, 77, 78].

Although the caching technique has been widely used in computer networks, it is until recently that wireless communication society realized its usefulness after exploiting two basic facts: One, the content delivered in the networks is "cacheable" which means that the same content is requested by different users, though the requests usually occur at different times; Two, caching shortens the distance between the content and the requester, which naturally alleviates the network burden and reduces the energy consumption. These properties endow the wireless caching with a promising future for the next-generation cellular networks [20, 21, 79].

In a caching scheme, users request the intended file from a pre-fixed library without traversing the base station or the core network. The caching scheme usually consists of two phases. The first phase is the *placement phase*, where the cached messages are stored solely based on the statistics of the user demands. In this phase, the network is

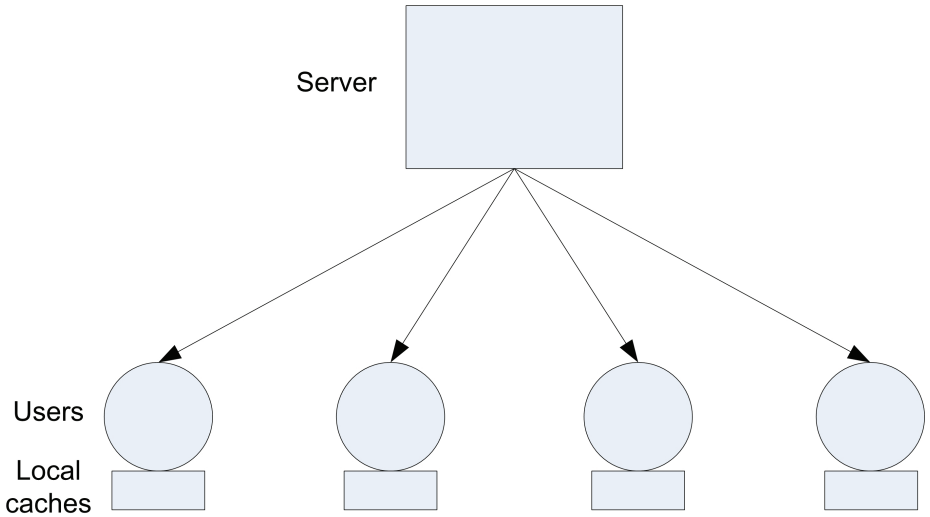


Figure 1.3. A centralized network model.

not congested, and the main limitation is the size of the cache memories. The second phase is the *delivery phase*, which is performed once the actual demands of the users have been revealed. In this phase, the network is congested, and the main limitation is the rate required to serve the requested content.

So far, there have been several works dedicating to improve the network performance by exploiting the caching techniques. Here, existing works are categorized and summarized by the considered networks as follows.

Caching in centralized networks

Consider a centralized network with one server (or macro-cell base station (MBS)) and several users (or small-cell base stations (SBS)), as depicted in Fig. 1.3. Each user is equipped with a local cache of size M_c, which is filled with content files from the core networks through a backhaul link [80–82]. The backhaul link from the server to users is limited of either low-capacity or expensive.

In [21], the authors considered the transmission between BSs and end users in a cellular network. They tried to improve the network performance by optimizing the content distribution on SBS caches. They formulated the problem as finding a solution to maximize the sum probability that any end user requests a file that is accessible

through its local helpers (BSs), weighted with its corresponding delay downloading from its helper. Their simulation results demonstrated performance improvements on the order of 400% - 500% more users at reasonable QoS levels. It provides an effective tool for content placement optimization.

K. Poularakis et al. [17] proposed an optimization framework combining caching and multicasting techniques in heterogeneous networks. They assume that an end user can be served either by the MBS or by a covering SBS provided that the latter has cached the requested content file. Their main idea is to optimize the content distribution on both MBS and SBS caches so that the chance of multicasting is utilized and the energy consumption of the entire cell is minimized.

For this model, an important information theoretical work by M.A. Maddah-Ali and U. Niesen [83] focused on minimizing the backhaul requirement for any random requests (from a set of known files) of users . They formulated the problem by the concept of an achievable memory-rate pair (M_c, R). Denote the smallest rate that (M_c, R) is achievable by $R^*(M_c)$, which describes the *memory-rate trade-off* of the caching problem.

$$R^*(M_c) \triangleq \inf\{R : (M_c, R) \text{ is achievable}\}. \tag{1.5}$$

In their proposed scheme, the files in the library are divided into packets and the users cache coded subsets of such packets. In the content delivery phase, multicasting opportunities are created through coding across the required packets. Each user will be able to resolve its wanted file by combining this multicasted message and its locally cached coded packets.

Then, assuming that the library size is m and the number of users is K, the authors provided an achievable rate $R^c(M_c)$ as an upper bound of $R^*(M_c)$ as follows.

$$R^c(M_c) = K\left(1 - \frac{M_c}{m}\right) \cdot \frac{1}{1 + \frac{KM_c}{m}}. \tag{1.6}$$

Their result suggests that by doing so, the number of users K and the size of local caches M_c can almost linearly contribute to the achievable rate. This scheme is of

zero outage and all the users can be served by one common message from the server. However, its required code length is combinatorial with the number of users, which is prohibitive for networks with many users. Also changing a single file in the library requires a significant reconfiguration of the user caches, which makes the cache update very difficult.

Note that given the caching policy and the user demands, the problem of minimizing the load over the shared link is essentially an index coding problem, which is not covered in this thesis.

Caching in decentralized networks

Consider a network with n wireless nodes, which are uniformly and independently distributed in a unit square as depicted in Fig. 1.4. Each node has a transmit power constraint and a local cache size of M_c files. The size of caches in this thesis is represented by the number of standard files. Without the presence of a centralized server or BS, communications between nodes are considered in this network. Each node intends to download its requested file from a library with m files through this cached random network.

Conventional ad-hoc networks

The capacity scaling law is recognized as a way of studying the fundamental trade-offs in large wireless networks [84] and has been intensely studied since the initial work [85] by Gupta and Kumar in 2000. The introduced network model in [85], which has been widely used, considers n nodes randomly distributed in a unit area and arbitrarily grouped into source-destination pairs. The scaling law characterizes the behavior of the capacity of such networks when the number of nodes n becomes large. The works in [85, 86] show that, for such networks, a multihop strategy is capable of achieving an aggregate throughput scaling of $\Theta(\sqrt{n})$. This important scaling result indicates that the capacity can increase rather than saturate, with the number of users n. Note that this result is achieved based on a *protocol model*, where nodes within a certain distance are able to communicate and only one node is allowed to

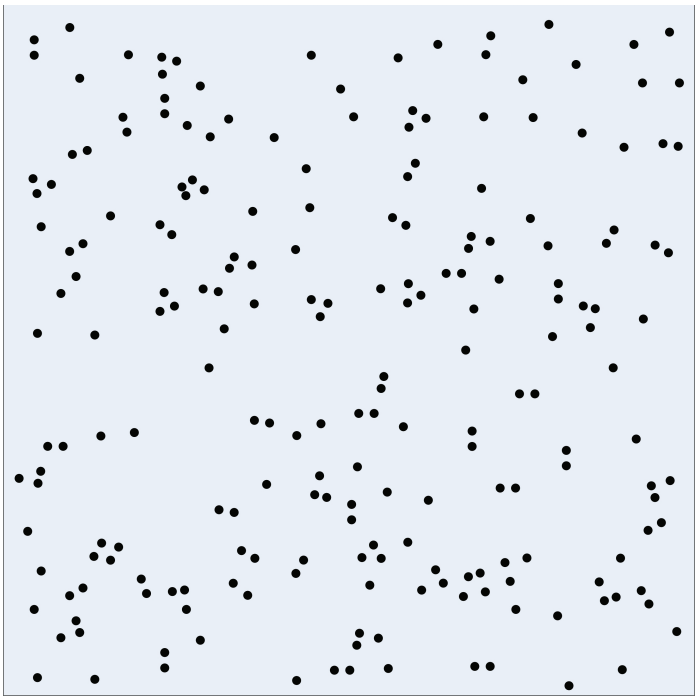


Figure 1.4. A decentralized network model.

transmit at a time within a radius [85]. The protocol model assumes that the signals received from nodes other than the source are interference and regarded as noise, which degrades the link performance. Under this assumption, the optimal strategy is to maximize the number of simultaneous transmissions (spatial reuse) [88].

On the other hand, it has been widely accepted that broadcast and superposition are two main distinguishing features of wireless medium [87]. Beyond the protocol model, a *physical model* was introduced in [88] to take into consideration wireless channel properties such as pathloss, fading, additive white Gaussian noise (AWGN) noise and interference. Based on this model, when the network is dense (fixed area with node density growing), a hierarchical cooperation scheme proposed in [88–90] achieves a near-linear throughput scaling of $\Theta\left(n^{1-\in}\right)$ where $\in$ can be arbitrarily small. Furthermore, applying the treating-interference-as-noise (TIN) condition in [91], Hong and Caire [92] characterized the achievable average per-link rate (not just the scaling law) for both hierarchical cooperation and multihop schemes. They showed that the signal-to-noise ratio and the number of hierarchical stages in [88–90] can be optimized,

and the optimized hierarchical cooperation schemes yield significant rate gains over multihop strategy, even under realistic network conditions.

Caching networks

A decentralized D2D caching network was first studied by Ji, Caire and Molisch in [93–95]. The considered caching network consists of n nodes randomly distributed in a unit area, which is similar to the conventional ad-hoc model. The difference is that each node in a caching network is equipped with a local cache and requests files from a library according to a priori popularity distribution.

Note that this setup is also applicable to many network such as mobile edge computing and content delivery networks (CDN) for future communications. For example, mobile edge computing provides IT and cloud-computing capabilities in close proximity to mobile subscriber. By offloading the computation via wireless access to the resource-rich cloud infrastructure, mobile cloud computing can augment the capabilities of mobile devices for resource-hungry applications [96]. Since the computation can only be offloaded after the information is distributed among mobile devices, this D2D caching network model can be regarded as a preparation step for mobile edge computing. On the other hand, since CDN is a geographically distributed network of proxy servers and their data centers, the problem of distributed caching with limited capacity in a content distribution network is essential in order to distribute service spatially relative to end-users. The results on D2D caching network can also provide theoretical support for this important problem in CDN [97].

In [93–95], Ji, Caire and Molisch formulated the problem by using a throughput-outage trade-off as follows. For a given network and request probability mass function, an outage-throughput pair (p, t) is achievable if there exists a cache placement scheme and an admission control and transmission scheduling policy with outage probability $p_0 \leq p$ and minimum per-user average throughput $T_{\min} \geq t$. The largest per-user average throughput $T^*(p)$ that (p, t) is achievable describes the *throughput-outage*

trade-off of the caching problem.

$$T^{*}(p) \triangleq \sup\{t : (p,t) \text{ is achievable}\}. \tag{1.7}$$

Then the authors in [93] try to solve following optimization problem:

$$\max_{\text{placement/delivery policy}} T_{\min}, \text{ subject to } p_0 \leq p. \tag{1.8}$$

The goal of this caching network is to get each node's request served through D2D links. Following the protocol model in [85], the network is divided into clusters of equal size, which is independent of the users' request and the cache placement. A user can look for the requested file only inside its own cluster.

They first maximized the probability that any user finds its requested file inside its own cluster by optimizing the caching distribution for given request probability mass function. The given solution suggested an interesting water-filling-like caching placement strategy. Then, with this optimized caching distribution, the corresponding $T^{*}(p)$ is calculated, which describes the throughput-outage trade-off of D2D caching schemes.

They showed that for a single-hop cache network, with an optimized caching placement strategy and cluster size, the network throughput scales as $\Theta\left(\frac{nM_c}{m}\right)$, growing linearly with the size of local caches M. This demonstrates that equipping caches is a cost-effective solution since the cost of providing caches is much lower than providing bandwidth. This nice scaling is achieved by exploiting spatial reuse of concurrent multiple short range D2D transmissions allowed by local caches at each node [93]. With the size of local caches growing, the requested content can be served by a shorter range D2D transmission and therefore the number of simultaneous active links increases, improving the network throughput. They later improved this scaling to $\Theta\left(n\sqrt{\frac{M_c}{m}}\right)$ in [98,99] by adopting the multihop strategy, where nodes get served by multi-relaying through the network. This even better scaling is achieved by reducing the number of transmissions within each hopping cell when increasing the size of local cache. With a larger local cache, the requested content can be served by a shorter range D2D transmission, which reduces the number of data paths within each hopping cell. Assuming

a constant aggregate achievable rate for one hopping cell, the achievable rate of each data path is increased, leading to an improved network throughput.

[78] improved the scaling law of the ad hoc networks from $\Theta\left(\frac{1}{N}\right)$ (N is the number of nodes in the network) to $\Theta(1)$ by optimizing the content distribution so that homogeneous opportunistic coordinated multi-point (CoMP) transmissions can be supported. However, this opportunistic CoMP is available by caching the same content on different helpers, which actually reduces the cache size for placement. The trade-off between the opportunistic CoMP gain and the cache size reduction is not mentioned by the authors.

Motivations

Based on the reviews of existing works, this thesis mainly focus on the decentralized approach, since it largely counts on the cooperation of end users where an potential improvement from the distributed MIMO technology and a proper relaying strategy can be expected. Especially, the key of designing a decentralized scheme is to optimize the content distribution and transmission scheduling so that 1) the probability that any user find its requested file in its accessible neighbours is maximized and 2) the interference of different transmitters is minimized.

Current studies in [93–95, 98, 99] are based on the protocol model in [85]. With the success of employing the physical model for conventional ad-hoc networks, we are curious about how the caching network capacity would scale if the physical model is used and more wireless channel properties are exploited. We notice that the work in [21] uses the physical model and enlarges the network throughput by employing MDS codes for the cached content to create cache-induced opportunistic CoMP. However, we are further interested in whether hierarchical cooperations in caching networks can achieve a better throughput scaling than the multihop scheme as in conventional ad-hoc networks. To address these questions, we investigate the caching network based on the physical model as well as considering hierarchical cooperations.

Another motivation of our work is the concern about the limited cache size in

practice. We notice that compared with the vast library that users may request, the local cache size at each node is rather limited in realistic networks. Especially, in a D2D network, the cache size of a single device cannot be sufficiently large in many cases. For example, consider the on-demand video streaming case, the file library could be up to 1000 TB while the available cache size of a user device is usually less than 1 TB. In addition, users usually are willing to contribute only a small fraction of their local caches. This conscious motivates us to design a caching scheme for the small cache case (small $\frac{M_c}{m}$), where the network throughput will be poor according to existing works. Thus, in this paper, our primary focus is on the small cache case when designing the caching scheme.

A recent work in [100] also investigated the caching network based on hierarchical cooperations. The authors proposed a hierarchical cache placement method and employed hierarchical cooperation as a basic transmission unit. Their results showed that for the "heavy tail" Zipf request distribution, the network performance is similar with that in multihop schemes if the pathloss factor is no less than three. However, whether a better caching scaling can be achieved, especially when the cache size is small, is still unknown. In our work, our focus is not on whether a better throughput scaling with the size of caches can be achieved. Instead, we mainly investigate if a good scaling with the number of users can be achieved, even when the cache size is small compared with the size of library.

1.2.4 D2D Communications in Cellular Networks

D2D communications

The 5G wireless networks are expected to meet the booming data traffic demand spurred by the popularity of smart phones and electronic tablets, as well as to provide a better quality of service and the user experience. To this aim, the integration of D2D communications into a traditional cellular network has been proposed as one of the promising technologies for 5G networks [22–24, 101]. D2D communications in

cellular networks are mainly considered for local traffic handling, which enable direct communications between two mobile users without traversing the BS or core network. Studies have shown that this hybrid network can significantly increase network SE and EE, provide low transmission delay and alleviate the core network congestion [22,23].

In a traditional cellular network, all communications must go through the BS even if both communicating parties are in range for D2D communication. This architecture suits the conventional low data rate mobile services such as voice call and text message in which users are not usually close enough to have direct communication. However, mobile users in todays cellular networks use high data rate services (e.g., video sharing, gaming, proximity-aware social networking) in which they could potentially be in range for direct communications (i.e., D2D) [101]. In addition, with the popularity of smart phones, tablets, etc., the D2D communication capability within the network has increased dramatically compared with a decade ago. Hence, D2D communications in such scenarios can highly increase the SE of the network. Nevertheless, the advantages of D2D communications are not only limited to enhanced SE, and this thesis provides a new approach to further make use of D2D communications by exploiting the diversity gain of a MIMO channel.

D2D communications in cellular networks can take place on the cellular spectrum (i.e., inband) or unlicensed spectrum (i.e., outband). The majority of current literature proposes to use the cellular spectrum for both D2D and cellular communications (i.e., underlay inband D2D). The arising co-channel interference between cellular users and D2D users is a major issue. Sophisticated resource allocation methods can be used to alleviate interference and to improve SE and EE, but most of them have high complexity [102–104]. The orthogonalization of D2D and cellular communications over the cellular spectrum (i.e., overlay inband D2D) can completely avoid such interference [105] [106], but at the expense of a lower SE improvement as compared to underlay proposals.

Outband D2D genuinely enjoys interference-free transmissions for cellular and D2D users, which largely simplifies the user management at the BS [107]. However,

it has been widely accepted that the traditional D2D technologies are inadequate. First, the widely known D2D technologies, such as Bluetooth and WiFi, work at the 2.4 GHz unlicensed band. This band is rather crowded and thus the interference is uncontrollable. In addition, both Bluetooth and WiFi require manual pairing between two devices, which causes inconvenience in customer experience [108].

A promising technology of the future 5G networks is millimeter wave (mmWave) communication, providing multi-gigabits-per-second to the end users [8], making them potentially applicable to D2D communications. More importantly, the interference problem would be largely alleviated due to the highly directional antennas and large propagation loss in mmWave communications. In several proposed mmWave D2D communications, the pairing of D2D devices can be handled by BSs and thus provides better user experience. For these reasons, this thesis proposes to use mmWave D2D communications to complement current mircoWave cellular networks.

Cooperative communications

Cooperative communications have been intensively studied in literatures and have been regarded as a focal technology in the cellular networks today [109] [39]. In general, D2D communications can be utilized for cooperative communications, e.g., packet forwarding and relaying, and their impact is expected to be remarkable [110]. A major application of D2D relaying in cellular networks is multi-hop, which has been recognized to be capable of reducing transmission power and increasing network capacity. For example, when the D2D pairs are far away from each other, the direct link between the users is not good enough for communication and other devices can then relay signals between the D2D pairs [111]. In [112] [113], multihop has also been proposed to enhance the cellular transmissions, where the user with a strong channel can forward the received message from the BS to a weak user via D2D links. This kind of relaying introduces so-called "hop gain", which can be seen as a power gain.

Another way of relaying is to exploit the diversity gain through the distributed multiple-input and multiple-output (MIMO) technology. In [114], the concept of a

virtual antenna array was proposed, where mobile users are clustered to form a virtual antenna array which emulates a MIMO device via D2D communications. A similar concept called "distributed antenna systems" has also gained attentions and has been shown to be able to cover the dead spots in wireless networks, extend service coverage, improve spectral efficiency and mitigate interference [115] [116]. Specifically, such a system architecture can realize the potential diversity gain by sharing antennas across the different terminals to form a virtual MIMO system [117]. The authors of [107] proposed a shared user equipment-side distributed antenna system, which utilizes mmWave D2D links to enable a spatial multiplexing gain for single-antenna end users to improve the energy efficiency of outdoor-to-indoor communications. Yet, studies on this direction remains limited.

Most current studies tend to treat the D2D links as ideal high-rate connections and focus on the allocation of cellular resources [107] [114]. We note that D2D connections are over wireless medium and D2D channel conditions may widely vary from one link to another due to fading and path loss, resulting in different D2D link capacities. In addition, different standby users may make different contributions to the rate of the destination users due to independent cellular channel realizations. Based on the above considerations, the following questions arises i) which standby users should be enabled, and ii) how many D2D resources should be allocated to different standby users. These questions have been dealt with in [106]. However, their proposed algorithm is based on a multi-hop D2D-cellular scheme.

For a MIMO relay scheme, the joint optimization of the MIMO transmission and relaying strategies is a difficult task, because the information theoretical capacity of the relay channel is still an open problem [118]. The authors of [119] consider the scenario of a single cellular user with a single multi-antenna relay and then address the problem of jointly optimizing the transmit covariance and quantization noise covariance matrix. Their results confirm that optimized relaying can lead to significant throughput improvement. In order to simplify cooperative transmissions and reduce complexity at each relay (standby user), it is desirable to perform inde-

pendent individual quantization at multiple single-antenna relays, which is different from the joint quantization at a single multi-antenna relay in [119]. Besides, individual relay-destination capacity constraints need to be considered instead of a sum relay-destination capacity constraint. In addition, using the sum rate of multi-cell cellular users as the performance merit instead of a single user rate is of more practical meaning. These considerations impose new challenges in the joint optimizing MIMO transmission and relaying.

Motivations

Most current works are dedicated to increasing the SE and EE from the resource utility point of view, instead of improving the cellular user performance. A typical scenario under the current D2D-cellular framework is that of two user terminals who communicate with each other directly through D2D links. But this is not the only scenario where D2D communication is useful. In this thesis, we make an observation that with a large number of end-users in the network, there are numerous devices possessing D2D communication capabilities, but do not have a need of performing communications at any given time. We refer to the users who do not request cellular connections or D2D communications as *standby users*. Our idea here is to make use of the D2D capability of these standby users to improve the active cellular user performance. Since we only use standby users without D2D communication demand, our proposal can be applied on top of current D2D framework in a dense cellular network.

Classic communication theory has revealed great benefits of equipping multiple antennas on users in wireless channels [120]. Especially, the multiple antennas are able to enhance the capability of encountering inter-cell interferences, which accounts for the main cause of rate degradation in dense cellular networks. However, with the size limit of user equipment, it is difficult to equip the end users with many antennas in the current micro-wave cellular system and for this reason, comparative studies on the benefit of user side MIMO are limited. For example, in a multi-cell interfering

broadcast channel, with multiple antennas at end-users, how many users should be scheduled as a function of transmit antennas at BS? How much performance gain can be obtained in practical systems? These questions have not yet been answered to the best of our knowledge. With the virtual antenna array formed via D2D links, multiple antennas at end-users in micro-wave cellular system becomes a real possibility. This thesis examines these questions and quantifies the benefits of the proposed scheme.

1.3 Contributions and Organization of the Thesis

Chapters 1 provides background information of my thesis work. In the following, Chapters 2 - 4 will present my novel research results on the advanced MIMO technology for future wireless networks.

1.3.1 Contributions of Chapter 2

The work on MIMO PNC for TWRCs and MWRCs is presented in Chapter 2. As discussed in Section 1.2.2, when considering MIMO PNC for TWRCs or MWRCs, since each user is transmitting in a spatial multiplexing way to achieve the full MIMO capacity, at the relay, each information stream will experience interference from the other user, on top of interference from the other streams of the same user. This characteristic brings new challenges for a MIMO PNC scheme such as how to reconstruct multiple NC messages at the relay so that the MIMO capacity of the single user-to-relay transmission can be retained or not degraded from the other user's concurrent transmission, and also the multiple data streams can be recovered at the destination. This thesis tackles this problem and presents an analytical error rate performance expression for the proposed PNC method.

The contributions of this work are summarized as follows.

- A new linear vector PNC scheme for spatial multiplexing MIMO TWRC without CSIT is proposed. With M antennas at each user and N antennas at the

relay, each user transmits M independent QAM signal streams, and the two users transmit simultaneously. Based on the receiver-side CSI, the relay selects a generator matrix that contains M coefficient vectors for linear network coding. Given the received signal vector,the relay then jointly reconstructs the associated M linear combinations of all messages (from all antennas of both users), w.r.t. the selected NC generator matrix. In the downlink phase, the resultant linear message-combinations are forwarded to the users. Each user then recovers the other users M messages. Our proposed scheme inherits the notion of linear PNC in [43, 56, 57], and is different from [44] which considered non-linear PNC and phase shift keying (PSK) modulations.

- An explicit solution for the NC generator matrix that minimizes the error probability of the scheme at a high SNR is presented. Insights on the NC generator matrix design are obtained from our derived optimized solution. In particular, our result shows that the optimized generator matrix spans a linear subspace that is orthogonal to a set of difference vectors. These difference vectors yield the so-called "minimum distance shortening" events specified in [12] (or "singular fade subspace" in [44]). We also present an efficient method for finding such an optimized solution based on the list sphere decoding algorithm.

- A closed-form expression on the average error probability of the proposed scheme over a Rayleigh fading MIMO TWRC is derived. Our analytical result shows that, as SNR tends to infinity, the average error probability of our proposed scheme approaches that of a single-user lower bound. This demonstrates the asymptotically optimal error probability performance of the proposed scheme. In particular, our proof is based on a new typical error event analysis that exploits a novel characterization of artificial deep-fade events that lead to the minimum distance shortening effect.

- Numerical results show that, for Rayleigh fading channel, the proposed scheme significantly outperforms conventional PNC schemes for a wide range of an-

tenna configurations and various QAM modulations. For example, for a MIMO TWRC where each node has two antennas, the proposed scheme offers a 4.5 dB improvement over the benchmark scheme. It is demonstrated that our derived closed-form analytical result matches well with the numerical result.

- A PNC scheme for a Y-channel model without CSIT is proposed. Under a multi-antenna relay setup, we present an explicit solution for NC generator matrices that minimize the NC error probability at a high SNR. We also provide and prove an approximation of the average NC error probability for the proposed scheme at a high SNR. In addition, the case of a single-antenna relay with multiple time-slots is discussed and different transmission strategies are compared through numerical results.

These results have been published on one journal paper [121] and two conference papers [122, 123], and have been submitted to one journal paper.

- **Jiajia Guo**, Tao Yang, Jinhong Yuan and Jian A. Zhang, "Linear Vector Physical-layer Network Coding for MIMO Two-Way Relay Channels: Design and Performance Analysis," *IEEE Transactions on Communications*, vol. 63, no. 7, pp. 2591-2604, July 2015.

- **Jiajia Guo**, Tao Yang, Jinhong Yuan and Jian A. Zhang, "Linear Physical-layer Network Coding for the Fading Y-channel without Transmitter Channel State Information," *IEEE VTC* 2016 Fall.

- **Jiajia Guo**, Tao Yang, Jinhong Yuan and Jian A. Zhang, "Design of Linear Physical-layer Network Coding for MIMO Two-way Relay Channels without Transmitter CSI," *IEEE WCNC* 2015.

- **Jiajia Guo**, Tao Yang, Jinhong Yuan and Jian A. Zhang, "A Novel Linear Physical-layer Network Coding Scheme for Y-channel without Transmitter CSI," submitted to *IEEE Transactions on Vehicle Technology*, 2017.

1.3.2 Contributions of Chapter 3

The work on the scaling law of the wireless D2D caching network is presented in Chapter 3, as discussed in Section 1.2.3. For a cached D2D network, most works focused on exploiting spatial reuse of concurrent multiple short range D2D transmissions. With the size of local caches growing, the requested content can be served by a shorter range D2D transmission and therefore the number of simultaneous transmissions increases. However, the spatial degrees of freedom in a MIMO transmission are not exploited in these works. The question of how the caching network capacity would scale by exploiting spatial degrees of freedom is not answered so far, and is thoroughly studied in this thesis. Moreover, compared with the vast library that users may request, the local cache size is rather limited in realistic networks. This conscious also motivates us to design a caching scheme for the small cache case.

The contributions of this work are summarized as follows.

- A new caching scheme employing distributed MIMO and hierarchical cooperations is proposed for a D2D caching network. Different from the protocol model widely used in D2D literatures [80, 81, 93–95, 98], our proposed scheme is based on *the physical model* [88–90] considering wireless channel properties such as pathloss and interference. The distributed MIMO technology is used in the cache delivery phase utilizing the "overheard" signals introduced by the physical model. This is different from most D2D works [80, 81, 93–95, 98] considering only intra-cluster or short-range communications. Hierarchical cooperations are used to facilitate the transmissions between neighbours and the destination node. This design exploits spatial degrees of freedom in addition to spatial reuse.

- A random independent caching placement strategy is proposed to serve the proposed caching delivery, and an asymptotic expression of the cache-hit probability is derived. In the proposed design, the files in the library are divided into packets and each node randomly caches uncoded packets from different

files, which lays a foundation for employing the distributed MIMO technique in the caching delivery phase. The asymptotic cache-hit probability is derived by transforming the cache placement into a special type of the Coupon Collectors problem in probability theory. This novel transformation provides a new perspective of examining the cache placement for caching networks.

- The throughput scaling law for the proposed scheme is derived. Our analysis shows that the average aggregate throughput of the network scales almost linearly with the number of nodes n, which outperforms the current scaling when the local cache size is small. We also prove that the outage probability approaches zero as n goes sufficiently large.

- An explicit expression of the optimal throughput as a function of system parameters such as pathloss factor under a target outage probability is derived. Our results show that there are two fundamental trade-offs in the network: the multiplexing-backhaul trade-off and the throughput outage trade-off, which determine the network throughput. We also show that one of the trade-offs is dominant, depending on the values of system parameters. Numerical results show that the throughput of the network is mainly limited by the multiplexing-backhaul trade-off under realistic parameters and our proposed scheme outperforms typical existing schemes when the local cache size is limited.

These results have been published on one journal paper [124] and one conference paper [125].

- **Jiajia Guo**, J. Yuan and Jian A. Zhang, "The Throughput Scaling Law of Wireless Device-to-device Caching Networks with Distributed MIMO and Hierarchical Cooperations," *IEEE Transactions on Wireless Communications*, vol. 17, no. 1, pp. 492-505, Jan. 2018.

- **Jiajia Guo**, J. Yuan and Jian A. Zhang, "Wireless Device-to-device Caching

Networks with Distributed MIMO and Hierarchical Cooperations," *IEEE GlobeCom* 2017.

1.3.3 Contributions of Chapter 4

The work on the D2D communications in cellular networks is presented in Chapter 4. As discussed in Section 1.2.4, when integrating D2D communications into cellular networks, most of existing works proposed using the D2D link as a multi-hop relay. This thesis proposes to use D2D communications to realize the diversity gain of MIMO channels, therefore enhancing the capability of encountering the inter-cell interference. The corresponding D2D multiple access and resource allocation problems are also addressed in this thesis.

The contributions of this work are summarized as follows.

- A hybrid D2D-cellular scheme applying the distributed MIMO technology is proposed to assist the cellular communications in 5G networks. With a user-clustering strategy, the active cellular user and its nearby standby users form a cluster. A MIMO channel is then emulated through D2D links from the BS to the cluster, providing diversity gain for single-antenna end users. This enhances the capability of inter-cell interference mitigation at end users, and thus improves the rate performance, especially for cell-edge users.

- Millimetre wave (mmWave) communications for D2D links are employed to avoid interference from cellular transmissions. We then propose an orthogonal D2D multiple access protocol to manage D2D interferences. The proposed protocol consists of time-division multiple access (TDMA) within one cluster and frequency division multiple access (FDMA) among different clusters, making use of the ample bandwidth provided by the mmWave communications.

- A joint optimization problem of maximizing the sum-rate over cellular transmit beamformers, cluster receive beamformers and D2D resource allocations is

formulated for downlink transmissions of our proposed scheme. We obtain a closed-form solution for the D2D resource allocation problem, thereby quantitatively revealing the impacts of cellular signal strength and D2D link quality at each standby user for practical system design.

- Extensive system-level simulations are performed to demonstrate the effectiveness of the proposed scheme. Up to 2.5x improvement in terms of the cell edge user rate is observed when each cellular user is helped by 9 standby users. The comparison with equal resource allocation and our proposed algorithm confirms the necessity and advantages of optimizing the D2D resource allocation in our proposed scheme. In addition, we observe that fully loading the BSs is always the optimal strategy when the number of standby users exceeds a certain threshold, which is a useful insight for system design.

These results have been submitted to an IEEE Transaction journal.

- **Jiajia Guo**, W. Yu and J. Yuan, "Enhancing Cellular Performance through Device-to-Device Distributed MIMO, " submitted to *IEEE Transactions on Wireless Communications*, 2017.

Each chapter is affiliated with an appendix containing detailed proofs and derivations if applicable. In the end, Chapter 5 summarizes the works presented in this thesis and provides concluding remarks.

Chapter 2

Physical-layer Network Coding for MIMO Multi-way Relay Channel

2.1 Introduction

In this chapter, we first propose a new linear vector PNC scheme for spatial MIMO TWRCs where the CSI is not available at the transmitters.

In this scheme, each user transmits M independent quadrature amplitude modulation signal streams respectively from its M antennas to the relay. Based on the receiver-side CSI, the relay determines a NC generator matrix for linear vector network coding, and reconstructs the associated M linear combinations of all messages. We present an explicit solution for the generator matrix that minimizes the error probability at a high SNR, as well as an efficient algorithm to find the optimized solution.

We propose a novel typical error event analysis that exploits a new characterization of the deep fade events for the TWRC. We derive a new closed-form expression for the average error probability of the proposed scheme over a Rayleigh fading MIMO TWRC. Our analysis shows that the proposed scheme achieves the optimal error rate performance at a high SNR. Numerical results show that the proposed scheme

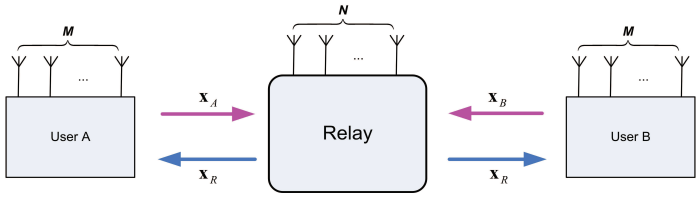


Figure 2.1. System model for MIMO TWRCs.

signicantly outperforms existing schemes, and match well with our analytical results.

Last but not least, for a typical MWRC model, the Y-channel, the case of a multiple-antenna relay with one time-slot and the case of a single-antenna relay with multiple time-slots are designed and discussed respectively.

This chapter is organized as follows. First, Section 2.2 - 2.6 present the proposed MIMO PNC work for TWRCs. Then, Section 2.7 extends the idea of the linear PNC to MWRCs, especially the Y-channel. Last, Section 2.8 concludes this chapter and detailed derivations are provided in Section 2.9.

2.2 The MIMO TWRC System Model

Consider a MIMO TWRC where user A and user B exchange messages via a relay node as shown in Fig. 2.8. Each user has M antennas and the relay has N antennas. Each round of information exchange runs in two equal-duration time-slots as in the standard protocol of PNC. We refer to these two time-slots as the uplink phase and the downlink phase.

2.2.1 Uplink Phase

In the uplink phase, both users transmit signals simultaneously to the relay with spatial multiplexing mode.

For the ease of presentation, we first present a real-valued system model, i.e., the transmitted signals and channel coefficients are real-valued. Later we will show that a complex-valued model can be easily transformed into a real-valued model.

Consider an un-channel-coded real-valued MIMO TWRC model. Denote the signals from user k, $k \in \{A, B\}$, by a length-M signal vector $\mathbf{x}_k$ with average energy normalized to one. The fading channel coefficients between user k and the relay are denoted by a matrix $\mathbf{H}_k$, whose (n, m)th element is the channel coefficient from the mth antenna of user k to the nth antenna of the relay, $n \in \{1, \cdots, N\}$, $m \in \{1, \cdots, M\}$. The received signal vector at the relay, denoted by $\mathbf{y}$, is

$$\mathbf{y} = \sqrt{E_s}\mathbf{H}_A\mathbf{x}_A + \sqrt{E_s}\mathbf{H}_B\mathbf{x}_B + \mathbf{z}. \tag{2.1}$$

where E_s is the average transmitted power per symbol vector, and $\mathbf{z}$ is the AWGN vector at the relay whose entries are i.i.d. with zero mean and variance σ_z^2. The SNR is defined as $\rho = E_s/\sigma_z^2$.

Remark 1 (Degree of CSI)**.** *In this section, we assume that CSI is not available at the transmitters, and is perfectly known at the intended receivers. This applies to practical scenarios where there is no feedback from the receiver to the transmitter or the channel reciprocity does not hold.*

In practical systems, the transmitted signals and channel coefficients are both complex-valued. Here we consider a complex-valued MIMO TWRC model. For the uplink phase, let $\tilde{\mathbf{x}}_k$, $\tilde{\mathbf{H}}_k$, $\tilde{\mathbf{y}}$ and $\tilde{\mathbf{z}}$ denote the complex-valued transmitted signal vector, channel coefficient matrix, received signal vector and noise vector, respectively. The complex-valued model is represented as

$$\tilde{\mathbf{y}} = \sqrt{E_s}\tilde{\mathbf{H}}_A\tilde{\mathbf{x}}_A + \sqrt{E_s}\tilde{\mathbf{H}}_\mathbf{B}\tilde{\mathbf{x}}_\mathbf{B} + \tilde{\mathbf{z}}. \tag{2.2}$$

It is well known that the above $N \times M$ complex-valued model is equivalent to the following $2N \times 2M$ real-valued model

$$\begin{aligned}\begin{bmatrix} \mathrm{Re}(\tilde{\mathbf{y}}) \\ \mathrm{Im}(\tilde{\mathbf{y}}) \end{bmatrix} =& \sqrt{E_s} \begin{bmatrix} \mathrm{Re}(\tilde{\mathbf{H}}_A) & -\mathrm{Im}(\tilde{\mathbf{H}}_A) \\ \mathrm{Im}(\tilde{\mathbf{H}}_A) & \mathrm{Re}(\tilde{\mathbf{H}}_A) \end{bmatrix} \begin{bmatrix} \mathrm{Re}(\tilde{\mathbf{x}}_A) \\ \mathrm{Im}(\tilde{\mathbf{x}}_A) \end{bmatrix} \\ &+ \sqrt{E_s} \begin{bmatrix} \mathrm{Re}(\tilde{\mathbf{H}}_B) & -\mathrm{Im}(\tilde{\mathbf{H}}_B) \\ \mathrm{Im}(\tilde{\mathbf{H}}_B) & \mathrm{Re}(\tilde{\mathbf{H}}_B) \end{bmatrix} \begin{bmatrix} \mathrm{Re}(\tilde{\mathbf{x}}_B) \\ \mathrm{Im}(\tilde{\mathbf{x}}_B) \end{bmatrix} + \begin{bmatrix} \mathrm{Re}(\tilde{\mathbf{z}}) \\ \mathrm{Im}(\tilde{\mathbf{z}}) \end{bmatrix}. \end{aligned} \tag{2.3}$$

Throughout the section, we will use the real-valued model in (2.1) for the design and analysis of a complex-valued MIMO TWRC.

2.2.2 Downlink Phase

Upon receiving $\mathbf{y}$, the relay generates a signal vector $\mathbf{x}_R$ which contains some function of the two users' signals, denoted by

$$\mathbf{x}_R = f\left(\mathbf{x}_A, \mathbf{x}_B\right). \tag{2.4}$$

This signal vector is then broadcasted to the two users. Here, note that $f(\cdot)$ could be some linear functions [56] or non-linear functions [12] [44] of the two users' messages. In this section, we focus on linear functions due to its low computational complexity and scalability as we will see later.

In the downlink phase, upon receiving $\mathbf{x}_R$, each user extracts the other user's message with the help of the perfect knowledge of its own message. This finishes one round of message exchange.

Let $\hat{\mathbf{x}}_k$, $k \in \{A, B\}$, be the recovered messages. An error is declared if $\hat{\mathbf{x}}_k \neq \mathbf{x}_k$. This section is primarily concerned with the design of the relay function $f(\cdot)$ such that the error probability $\Pr(\hat{\mathbf{x}}_k \neq \mathbf{x}_k)$, $k \in \{A, B\}$, is minimized. A more rigorous problem formulation will be given in (2.65) in Section 2.3.

2.3 Proposed Linear Vector PNC for MIMO TWRC

In this section, we propose a new linear vector PNC scheme for the MIMO TWRC. We focus on the setup of $M \leq N$. Our proposed scheme and results can be extended to the case of $M > N$ by incorporating with space-time coding[1].

[1]When $M > N$, each user will transmit N independent message streams via M antennas. In this case, space-time coding will be used to code the N messages over M transmit antennas.

2.3.1 Uplink Phase

Denote user k's messages by a length-M vector $\mathbf{w}_k$, $k \in \{A, B\}$. The entries of $\mathbf{w}_k$ are i.i.d. and are uniformly drawn from an finite integer set $\{0, 1, \cdots, q-1\}$.

Using pulse amplitude modulation, user k-th transmitted message vector $\mathbf{w}_k$ is one-to-one mapped to a modulated symbol vector $\mathbf{x}_k$ by

$$\mathbf{x}_k = \frac{1}{\gamma}\left[\left(\mathbf{w}_k - \frac{q-1}{2}\right)\right] \tag{2.5}$$

where γ is a normalization factor which ensures that $E\left(\|\mathbf{x}_k^2\|\right) = 1$. In the uplink phase, user A and B transmit $\mathbf{x}_A$ and $\mathbf{x}_B$ simultaneously. The relay receives a signal vector $\mathbf{y}$ as given in (2.1).

2.3.2 Linear Network Coding

Let

$$\mathbf{w} \triangleq [(\mathbf{w}_A)^T, (\mathbf{w}_B)^T]^T, \tag{2.6}$$

which collects the transmitted messages of all antennas of both users. A linear combination of the transmitted messages in a size-q finite integer set is represented by a scalar u as below:

$$u = \mathbf{g}^T \otimes \mathbf{w} \tag{2.7}$$

where $\mathbf{g}$ specifies the coefficients of the linear combination whose elements belongs to $\{0, 1, \cdots, q-1\}$ and $\otimes$ represents the matrix product in $\{0, 1, \cdots, q-1\}$, i.e.,

$$\mathbf{g}^T \otimes \mathbf{w} \triangleq \mathrm{mod}(\mathbf{g}^T\mathbf{w}, q). \tag{2.8}$$

In the following, we refer to $\mathbf{g}$ as a NC coefficient vector. In this section, we focus on the case where q is a prime number. Note that the finite integer set becomes a finite field under the modulo-q operation when q is a prime number.

In the scheme for a SISO TWRC [43], the relay only need to construct one linear combination of the two user's messages. In the MIMO two-way relay system considered in this section, each user transmits M independent messages. Thus, a scalar of

single linear combination of all users' message is not sufficient for the users to recover their desired M messages. Therefore, the relay needs to construct and forward a *vector* of multiple message-combinations, subject to a certain rank constraint. The details are described below.

Let $\mathbf{g}_1, \mathbf{g}_2, \cdots, \mathbf{g}_M$ denote M NC coefficient vectors. The associated vector of M linear message-combinations is denoted by

$$\mathbf{u} = [\mathbf{g}_1^T \otimes \mathbf{w}, \mathbf{g}_2^T \otimes \mathbf{w}, \cdots, \mathbf{g}_M^T \otimes \mathbf{w}]^T = \mathbf{G}^T \otimes \mathbf{w}. \tag{2.9}$$

We refer to the vector $\mathbf{u}$ as a linear *NC codeword* and $\mathbf{G} = [\mathbf{g}_1, \mathbf{g}_2, \cdots, \mathbf{g}_M]$ as an *NC generator matrix*. We emphasize that $\mathbf{G}$ must satisfy the following condition.

Condition 2.1. *Let* $\mathbf{G} = \begin{bmatrix} \mathbf{G}_A \\ \mathbf{G}_B \end{bmatrix}$ *where each of* $\mathbf{G}_A$ *and* $\mathbf{G}_B$ *is an* $M \times M$ *matrix. The matrices* $\mathbf{G}_A$ *and* $\mathbf{G}_B$ *are both of full rank* M*.*

For the ease of presentation, we say that a size $2M$-by-M matrix $\mathbf{G}$ is *block-wise full rank* if $\mathbf{G}$ satisfies Condition 2.1.

Explanation: Condition 2.1 is necessary to guarantee that with the NC generator matrix $\mathbf{G}$, each user is able to recover the other users' messages from all M antennas. It is clear that user A is not able to recover user B's messages of all M antennas if $\mathbf{G}_B$ is not of rank M.

2.3.3 Relay's Operation

The operation at the relay for a given channel realization is described as below and is depicted in Fig. 2.2.

Step 1: Based on the receiver-side CSI on $\mathbf{H}_A, \mathbf{H}_B$, the relay selects an NC generator matrix $\mathbf{G}$ subject to Condition 2.1;

Step 2: Given the N-dimension received signal $\mathbf{y}$ in (2.1), the relay attempts to re-construct the linear NC codeword $\mathbf{u}$ in (2.9) w.r.t. to the selected NC generator matrix $\mathbf{G}$. Denote the decision by $\hat{\mathbf{u}}$.

Step 3: The relay modulates $\hat{\mathbf{u}}$ and broadcasts it to both users.

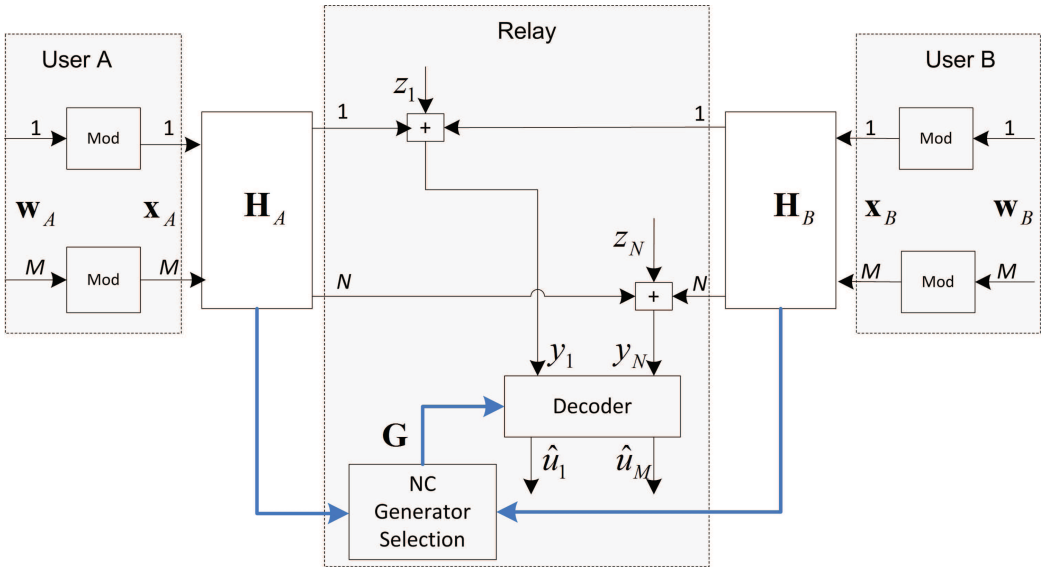


Figure 2.2. Block diagram of the proposed linear vector PNC scheme.

Remark 2.1. *There are various linear NC generator matrices satisfying Condition 2.1. It is important to note that the selection of the NC generator matrix* $\mathbf{G}$ *in Step 1 is critical for the error probability performance. Let* $\mathbf{u} \neq \hat{\mathbf{u}}$ *be referred to as an NC error. We will see later that the NC error probability is determined by the choice of the NC generator matrix. The optimal* $\mathbf{G}$ *matrix that minimizes the NC error probability will be studied in Section 2.4.*

Remark 2.2. *In practice, since* $\mathbf{G}$ *only needs to be selected once for each channel realization, the resultant overhead is negligible for a slow fading channel of a reasonably large channel coherence time. We assume that the selected* $\mathbf{G}$ *is delivered to the two users via two reliable links. This can be done by including the index of the NC generator matrix* $\mathbf{G}$ *in the overhead of the downlink transmission. The resultant overhead is very minor for a slow fading channel.*

Note that our proposed scheme is different from those in [16,46,50], where the CSI is assumed to be known to all transmitters and joint precoding is employed to align the signal directions of the two users. In these schemes, the strict signal direction alignment requires very accurate CSI feedback from the relay and it is very difficult

to realize in practice due to the CPO. In addition, the delivering of the continuous-valued full CSI from the relay to the users requires a much larger overhead than that in our proposed scheme, which only sends the index of the NC generator matrix.

2.3.4 Downlink Phase

Let the NC codeword in (2.9) be re-written as

$$\mathbf{u} = \mathbf{G}_A \otimes \mathbf{w}_A \oplus \mathbf{G}_B \otimes \mathbf{w}_B. \tag{2.10}$$

Suppose that $\mathbf{u}$ is correctly delivered to the two users. User A first removes its own messages $\mathbf{w}_A$ from $\mathbf{u}$, and obtains the resultant message vector

$$\mathbf{u} \ominus (\mathbf{G}_A \otimes \mathbf{w}_A) = \mathbf{G}_B \otimes \mathbf{w}_B,$$

which forms M linear equations for $\mathbf{w}_B$. Recall that $\mathbf{G}_B$ is of full rank from Condition 2.1. By solving the linear equations, user A can recover the desired messages from user B. Meanwhile, user B recovers its desired message $\mathbf{w}_A$ in a similar way.

In this section we mainly consider the uplink transmission phase and the reasons are as follows. First, as we will see later, the error performance in the uplink phase is primarily subject to the design of the generator matrix, and is different from the traditional point-to-point communication, while the error performance in the downlink phase is a *standard* point-to-point communication. From the PNC design point of view, we are most interested in the error performance at the relay in the uplink phase. Second, the error performance of the downlink phase remains the same for any generator matrix selected at the relay. This is because in our proposed scheme, the generated NC codewords have entries that are uniformly distributed in $\{0, 1, \cdots, q-1\}$, regardless of the selection of the generator matrix.

2.4 Asymptotically Optimal Design of the Proposed Linear Vector PNC Scheme

In this section, we investigate the optimal design of the proposed linear vector PNC scheme for a MIMO TWRC. Specifically, for any given channel realization, we will present an explicit expression for the NC generator matrix $\mathbf{G}$ that minimizes the NC error probability at a high SNR.

We note that the method provided for the SISO case in [43] cannot be directly used in the MIMO case due to multi-stream and multi-user interference. The main challenge of finding the optimized NC generator matrix $\mathbf{G}$ for the MIMO scenario lies in characterizing the superposition of the $2M$ signals (from all antennas of the two users) in an N-dimension signal space. Therefore, we first present some preliminaries.

2.4.1 Preliminaries

Re-construction of the NC codeword u

Consider a given channel realization of $\mathbf{H}_A$ and $\mathbf{H}_B$. Define

$$\mathbf{w}_s \triangleq \mathbf{H}_A\mathbf{w}_A + \mathbf{H}_B\mathbf{w}_B, \tag{2.11}$$

which is referred to as a *superimposed* (SI) *symbol vector*.[2]

Let the set

$$\mathcal{W}_s = \{\mathbf{w}_s : \mathbf{w}_s = \mathbf{H}_A\mathbf{w}_A + \mathbf{H}_B\mathbf{w}_B\} \tag{2.12}$$

collects all possible SI symbol vectors. The cardinality of the set is q^{2M} with probability one.

For a given NC generator $\mathbf{G}$, we define the following set

$$\mathcal{W}_s^{(\mathbf{G})}(\mathbf{u}) \triangleq \left\{ \mathbf{w}_s : \begin{array}{l} \mathbf{w}_s = \mathbf{H}_A\mathbf{w}_A + \mathbf{H}_B\mathbf{w}_B \\ \mathbf{G}^T \otimes \mathbf{w} = \mathbf{u} \end{array} \right\}, \tag{2.13}$$

[2]Recall that $\mathbf{w}_m$ and $\mathbf{x}_m$, $m \in \{A, B\}$ are one-to-one mapped. Therefore, we can use either $\mathbf{w}_m$ or $\mathbf{x}_m$. For simplicity, we just use $\mathbf{w}_m$ here.

where $\mathbf{w} = \left[\mathbf{w}_A^T, \mathbf{w}_B^T\right]^T$. Here, $\mathcal{W}_s^{(\mathbf{G})}(\mathbf{u})$ collects all the SI symbol vectors whose underlying NC codewords are identical and equal to $\mathbf{u}$. Clearly we have

$$\mathcal{W}_s = \left\{ \mathcal{W}_s^{(\mathbf{G})}(\mathbf{u}),\ \forall\ \mathbf{u} \in \{0, 1, \cdots, q-1\}^M \right\}. \tag{2.14}$$

Here, (2.14) can be viewed as a partition of all q^{2M} SI symbol vectors into q^M sets, where each set corresponds to a specific NC codeword.

For a given generator matrix $\mathbf{G}$, the maximum likelihood decoding rule for deciding $\hat{\mathbf{u}}$ is:

$$\hat{\mathbf{u}} = \arg\max_{\mathbf{u}} \Pr\left(\mathbf{y} | \mathcal{W}_s^{(\mathbf{G})}(\mathbf{u})\right) = \arg\max_{\mathbf{u}} \sum_{\mathbf{w}_s \in \mathcal{W}_s^{(\mathbf{G})}(\mathbf{u})} \Pr\left(\mathbf{y} | \mathbf{w}_s\right)$$

Conventional Distance Spectrum

Denote the squared Euclidean distance between any two SI symbol vectors $\mathbf{w}_s$ and $\mathbf{w}_s'$ by

$$d(\mathbf{w}_s, \mathbf{w}_s') = \left\| \mathbf{w}_s - \mathbf{w}_s' \right\|^2. \tag{2.15}$$

The minimum distance between any two SI symbol vectors is given by

$$d_1 = \min_{\mathbf{w}_s, \mathbf{w}_s' \in \mathcal{W}_s, \mathbf{w}_s \neq \mathbf{w}_s'} d(\mathbf{w}_s, \mathbf{w}_s'), \tag{2.16}$$

Similarly, denote the ith smallest distance by d_i. Also, let $A(d_i)$ denote the number of neighbours with distance d_i, i.e., the multiplicity w.r.t. the d_i events. Define

$$\mathcal{A} \triangleq \{A(d_1), \cdots, A(d_i), \cdots\}. \tag{2.17}$$

We refer to $\mathcal{A}$ as the *conventional distance spectrum* of SI symbol vectors. Note that this is the distance spectrum if the relay attempts to distinguish all individual messages from users A and B as in complete decoding scheme.

NC Set-distance Spectrum

In light of the spirit of PNC, the relay does not need to distinguish all individual messages of both users. Instead, the relay only needs to distinguish the SI symbol

vectors whose underlying NC codewords are different. Recall the notation of d_i in conventional distance spectrum. Only the d_i events in which the SI symbol vectors correspond to different NC codewords, will contribute to the NC error probability. In this case, using the conventional distance spectrum is not appropriate any more. Here, we introduce new definitions as follows.

Definition 2.1 (NC set-distance spectrum)**.** *For a given* $\mathbf{G}$*, define*

$$A^{(\mathbf{G})}(d_i) \triangleq \left|\left\{\mathbf{w}_s, \mathbf{w}'_s : d\left(\mathbf{w}_s, \mathbf{w}'_s\right) = d_i, \mathbf{w}_s \in \mathcal{W}_s^{(\mathbf{G})}(\mathbf{u}), \mathbf{w}'_s \in \mathcal{W}_s^{(\mathbf{G})}(\mathbf{u}'), \mathbf{u} \neq \mathbf{u}'\right\}\right|. \tag{2.18}$$

Here, $A^{(\mathbf{G})}(d_i)$ *is the number of events that the squared Euclidean distance of two SI symbol vectors, with different underlying NC codewords, is equal to* d_i*. Let*

$$\mathcal{A}^{(\mathbf{G})} \triangleq \left\{A^{(\mathbf{G})}(d_1), A^{(\mathbf{G})}(d_2), \cdots\right\}, \tag{2.19}$$

which is referred to as the NC set-distance spectrum*.*

Property 2.1. *With a certain NC generator matrix* $\mathbf{G}$*, some SI symbol vectors are clustered into one NC set* $\mathcal{W}_s^{(\mathbf{G})}(\mathbf{u})$*. Therefore, some elements in conventional distance spectrum will become zero in the NC set-distance spectrum.*

Remark 2.3. *It is clear that the error probability of the proposed linear vector PNC scheme is determined by the NC set-distance spectrum in (2.19) rather than the conventional distance spectrum in (2.17).*

In particular, the distance w.r.t. the first non-zero term in the NC set-distance spectrum is the minimum distance between SI symbol vectors w.r.t. different NC codewords. We refer to this distance as *minimum set-distance*, defined as follows.

Definition 2.2 (Minimum set-distance)**.** *For a given* $\mathbf{G}$*, the minimum set-distance (MSD) is defined as*

$$d_{MSD} \triangleq d_{i^*}, \text{ where } i^* = \arg \min_{A^{(\mathbf{G})}(d_i) \neq 0} i. \tag{2.20}$$

This is an important concept that we will use in the following of the section.

Error Probability

The pair-wise error probability of confusing $\mathbf{u}$ with an erroneous one $\mathbf{u}'$ is

$$P_e\left(\mathbf{u}\to\mathbf{u}'\right)=\sum_{\substack{\mathbf{w}_s\in\mathcal{W}_s^{\mathbf{G}}(\mathbf{u})\\ \mathbf{w}'_s\in\mathcal{W}_s^{\mathbf{G}}(\mathbf{u}')\\ \mathbf{u}\neq\mathbf{u}'}}P_e\left(\mathbf{w}_s\to\mathbf{w}'_s\right)=\sum_{\substack{\mathbf{w}_s\in\mathcal{W}_s^{\mathbf{G}}(\mathbf{u})\\ \mathbf{w}'_s\in\mathcal{W}_s^{\mathbf{G}}(\mathbf{u}')\\ \mathbf{u}\neq\mathbf{u}'}}Q\left(\sqrt{\frac{\rho}{2\gamma^2}d(\mathbf{w}_s,\mathbf{w}'_s)}\right).$$

Thus, the NC error probability of the proposed scheme, averaged over all NC codewords $\mathbf{u}$, can be upper bounded by

$$\begin{aligned}P_e\leq&\sum_{\mathbf{u}}p(u)P_e(u)\\&=\sum_{\mathbf{u}}\frac{1}{q^M}\sum_{\mathbf{u}'}P_e(u\to u')\\&=\frac{1}{q^M}\sum_{\mathbf{u}}\sum_{\mathbf{u}'}\sum_{\substack{\mathbf{w}_s\in\mathcal{W}_s^{\mathbf{G}}(\mathbf{u})\\ \mathbf{w}'_s\in\mathcal{W}_s^{\mathbf{G}}(\mathbf{u}')\\ \mathbf{u}\neq\mathbf{u}'}}Q\left(\sqrt{\frac{\rho}{2\gamma^2}d(\mathbf{w}_s,\mathbf{w}'_s)}\right)\end{aligned}\tag{2.21}$$

where we have utilized the union bound [7] and the fact that $p(\mathbf{u})=1/q^M$. Using the NC set-distance spectrum described in Definition 2.1, (2.21) can be further written as

$$P_e\leq\frac{1}{q^M}\left(A^{(\mathbf{G})}(d_1)Q\left(\sqrt{\frac{\rho}{2\gamma^2}d_1}\right)+A^{(\mathbf{G})}(d_2)Q\left(\sqrt{\frac{\rho}{2\gamma^2}d_2}\right)+\cdots\right).\tag{2.22}$$

As $\rho\to\infty$ and $q<\infty$, due to the approximate exponential feature of Q-function, P_e will be dominated by the first non-zero term in the NC set-distance spectrum, i.e., the MSD and its multiplicity. Then, as $\rho\to\infty$, the error probability P_e becomes

$$P_e\lessapprox\frac{A^{(\mathbf{G})}(d_{MSD})}{q^M}Q\left(\sqrt{\frac{\rho}{2\gamma^2}d_{MSD}}\right).\tag{2.23}$$

2.4.2 Design Problem Formulation

For given channel realizations of $\mathbf{H}_A$ and $\mathbf{H}_B$, the conventional distance spectrum of SI symbol vectors is determined. From Property 2.1, we know that in NC set-distance spectrum, some elements $A^{(\mathbf{G})}(d_i)$ are zero. Moreover, different NC generator

matrices will result in different zero elements in the NC set-distance spectrum, which lead to different NC error probability. The optimal matrix $\mathbf{G}$ that minimizes the error probability in (2.22) can be formulated as

$$\mathbf{G}^{opt} = \arg\min_{\mathbf{G}} P_e \text{ , s.t. Condition 2.1} \tag{2.24}$$

From (2.23), an optimized linear vector PNC scheme should make the MSD as large as possible. The problem (2.24) then becomes

$$\mathbf{G}^{opt} = \arg\max_{\mathbf{G}} d_{MSD}, \tag{2.25}$$

subject to Condition 2.1. The remainder of this section is devoted to finding the solution to this problem.

2.4.3 Solution to the Problem in (2.25)

Let $\boldsymbol{\delta}_A = \mathbf{w}_A - \mathbf{w}'_A$, $\boldsymbol{\delta}_B = \mathbf{w}_B - \mathbf{w}'_B$, and

$$\boldsymbol{\delta} \triangleq \left[\boldsymbol{\delta}_A^T, \boldsymbol{\delta}_B^T\right]^T. \tag{2.26}$$

We refer to $\boldsymbol{\delta}$ as the *difference vector* (DV) of $\mathbf{w}_s$ and $\mathbf{w}'_s$. Then, the squared Euclidean distance between two SI symbol vectors $\mathbf{w}_s$ and $\mathbf{w}'_s$, $\mathbf{w}_s \neq \mathbf{w}'_s$, is

$$\begin{aligned} d(\mathbf{w}_s, \mathbf{w}'_s) &= \|\mathbf{H}_A(\mathbf{w}_A - \mathbf{w}'_A) + \mathbf{H}_B(\mathbf{w}_B - \mathbf{w}'_B)\|^2 \\ &= \left\|[\mathbf{H}_A, \mathbf{H}_B]\left[\boldsymbol{\delta}_A^T, \boldsymbol{\delta}_B^T\right]^T\right\|^2 \\ &= \|[\mathbf{H}_A, \mathbf{H}_B]\boldsymbol{\delta}\|^2. \end{aligned} \tag{2.27}$$

Note that since $\mathbf{w}_s \neq \mathbf{w}'_s$, the DV vector belongs to $\{1-q, \cdots, 0, \cdots, q-1\}^{2M}$ and $\|\boldsymbol{\delta}\| \neq 0$.

Recall the conventional distance spectrum, and denote the DV corresponding to d_i by

$$\boldsymbol{\delta}_i \triangleq \arg_{\boldsymbol{\delta}} \left(\|[\mathbf{H}_A, \mathbf{H}_B]\boldsymbol{\delta}\|^2 = d_i\right). \tag{2.28}$$

We are now ready to present explicit solution to (2.25), which solves (2.24) at a high SNR.

Theorem 2.1. *As $\rho \to \infty$, the NC generator matrix $\mathbf{G}$ that minimizes the NC error probability satisfies*

$$\mathbf{G}^T \otimes \mathbf{P} = \mathbf{0} \tag{2.29}$$

where

$$\mathbf{P} = \mathrm{mod}([\boldsymbol{\delta}_{i_1}, \cdots, \boldsymbol{\delta}_{i_M}], q) \tag{2.30}$$

and i_1, $\cdots$, i_M are the first M indices such that $\mathbf{P}$ is of column rank M and is block-wise full rank in the finite field.

Remark 2.4. *We can view the $\mathbf{P}$ matrix composed by the M DVs (after modulo-q operation) as a parity-check matrix w.r.t. the NC generator matrix $\mathbf{G}$. We note that, with (2.29), it can be shown that $\mathbf{G}$ is block-wise full rank if and only if $\mathbf{P}$ is block-wise full rank. Thus, Condition 2.1 is automatically met.*

Proof. In general, there exists only one vector $\boldsymbol{\delta}_1$ that corresponds to the distance d_1. (Here, $\boldsymbol{\delta}_1$ and $-\boldsymbol{\delta}_1$ are viewed as the same vector.). Using the distance clustering method in [43], a matrix $\mathbf{G}$ satisfying

$$\mathrm{mod}(\mathbf{G}^T \otimes (\mathbf{w} \pm \boldsymbol{\delta}_1), q) = \mathrm{mod}(\mathbf{G}^T \otimes \mathbf{w}, q), \tag{2.31}$$

or equivalently

$$\mathbf{G}^T \otimes \boldsymbol{\delta}_1 = 0 \tag{2.32}$$

leads to $A^{(\mathbf{G})}(d_1) = 0$. In other words, if (2.32) is met, any two SI symbol vectors with distance d_1 are *clustered* into the same NC set $\mathcal{W}_s^{(\mathbf{G})}(\mathbf{u})$. Therefore, the element $A^{(\mathbf{G})}(d_1)$ in the NC set-distance spectrum $\mathcal{A}^{(\mathbf{G})}$ becomes zero. Then, the minimum set-distance satisfies $d_{MSD} > d_1$. On the other hand, it can be shown that any choice of $\mathbf{G}$ matrix which does not satisfy (2.29) will lead to $A^{(\mathbf{G})}(d_1) \neq 0$, and its d_{MSD} strictly equals to d_1.

Next, consider that $[\mathbf{p}_1, \cdots, \mathbf{p}_M] = \mathrm{mod}([\boldsymbol{\delta}_1, \cdots, \boldsymbol{\delta}_M], q)$ happens to be full rank and satisfy the block-wise full rank condition. Then, by letting

$$\mathbf{G}^T \otimes \mathbf{p}_i = 0, i = 1, \cdots, M, \tag{2.33}$$

it is guaranteed that $A^{(\mathbf{G})}(d_i) = 0$ for $i = 1, \cdots, M$ and thus the minimum set-distance satisfies $d_{MSD} > d_M$. Also, the resultant $\mathbf{G}$ from (2.33) is block-wise full rank as $\mathbf{P}$ is block-wise full rank, and Condition 2.1 is met.

Let us now consider the case that $[\mathbf{p}_1, \cdots, \mathbf{p}_M] = \mathrm{mod}([\boldsymbol{\delta}_1, \cdots, \boldsymbol{\delta}_M], q)$ is full rank but is not block-wise full rank. It can be shown that, in this case, there does not exist a $\mathbf{G}$ matrix that satisfies (2.33) while being block-wise full rank. Thus, it is required to choose the M DVs such that they are collectively full rank and block-wise full rank after modulo-q operation, while the corresponding distances are as small as possible. Let such DVs be denoted by a matrix $\mathbf{P}$ after the modulo-q operation. Such a matrix $\mathbf{P}$ is specified in the line below (2.30). Denote the smallest distance whose corresponding DV results in the block-wise full rank condition doesn't hold by d_f, where $f = \underset{i_m \neq i}{\arg\min} m$. Then the MSD satisfies $d_{MSD} > d_f$. In addition, it can be shown that the minimum set-distance with other choice of $\mathbf{P}$ is not greater than d_f. This finishes the proof. □

2.4.4 Low-complexity Implementation of (2.30)

In the preceding subsection, we have derived a solution to the optimized linear vector PNC scheme at a high SNR. The majority of the computational complexity lies in finding the $\mathbf{P}$ matrix. To arrive at the solution in Theorem 2.1, we need to find the DVs $\boldsymbol{\delta}^{i_1}, \cdots, \boldsymbol{\delta}^{i_M}$, that lead to the first several smallest distance events $\|[\mathbf{H}_A, \mathbf{H}_B]\boldsymbol{\delta}\|^2$. We call these DVs *the candidate list*. The computational complexity for this will increase with the number of antennas. In this part, we investigate how to efficiently find the solution in (2.30) based on a list sphere decoding (LSD) algorithm.

The LSD algorithm has been applied for MIMO detection in many literatures [126, 127]. The sphere decoder avoids an exhaustive search and examines only the points that lie inside a sphere, whose radius r is chosen to be large enough to contain the solution. We next present the LSD algorithm can be used to find this candidate list with significantly reduced searches.

Here, we attempt to find

$$\|[\mathbf{H}_A, \mathbf{H}_B]\boldsymbol{\delta}\|^2 \leq r^2. \tag{2.34}$$

With a proper choice of r, we can obtain a list of multiple candidates through a single search instead of exhaustive search of each individual candidate for multiple times.

Let $[\mathbf{H}_A, \mathbf{H}_B] = \mathbf{QR}$ where $\mathbf{Q}$ is a unitary matrix and R is an $N \times 2M$ upper triangular matrix. We have

$$\|[\mathbf{H}_A, \mathbf{H}_B]\boldsymbol{\delta}\|^2 = \|\mathbf{R}\boldsymbol{\delta}\|^2 = \boldsymbol{\delta}^T\mathbf{R}^T\mathbf{R}\boldsymbol{\delta}. \tag{2.35}$$

Recall the setup of $M \leq N$, the LSD algorithm is conducted for the following two cases.

Case 1: $M \leq N \leq 2M$

The upper triangular matrix $\mathbf{R}$ will consist of an $N \times N$ left upper triangular matrix and an $N \times (2M - N)$ right non-zero matrix. Then, (2.34) can be rewritten as

$$\boldsymbol{\delta}^T\mathbf{R}^T\mathbf{R}\boldsymbol{\delta} = \sum_{i=1}^{N}\left(r_{ii}\delta_i + \sum_{j=i+1}^{2M} r_{ij}\delta_j\right)^2 \leq r^2 \tag{2.36}$$

where r_{ij} is the (i,j)th entry of matrix $\mathbf{R}$. Each term in the summation over i in (2.36) is nonnegative.

We start with finding candidates for δ_N, ..., δ_{2M}. s.t.

$$\left(\sum_{j=N}^{2M} r_{Nj}\delta_j\right)^2 \leq r^2 \tag{2.37}$$

Then, starting with $i = N$ and $i = N - 1$, and disregarding the terms in $i = 1, ..., N - 2$, we get

$$\left(\sum_{j=N}^{2M} r_{N,j}\delta_j\right)^2 + \left(r_{N-1,N-1}w_{N-1} + \sum_{j=N}^{2M} r_{Nj}\delta_j\right)^2 \leq r^2 \tag{2.38}$$

or

$$\frac{-\sqrt{r^2 - \left(\sum_{j=N}^{2M} r_{N,j}\delta_j\right)^2} - \sum_{j=N}^{2M} r_{Nj}\delta_j}{r_{N-1,N-1}} \leq w_{N-1} \leq \frac{\sqrt{r^2 - \left(\sum_{j=N}^{2M} r_{N,j}\delta_j\right)^2} - \sum_{j=N}^{2M} r_{Nj}\delta_j}{r_{N-1,N-1}} \tag{2.39}$$

We now obtain a boundary for δ_{N-1} instead of examining all possible points in the constellation. Next, the boundary for $\delta_{N-2}, ..., \delta_1$ is determined in a similar way.

Case 2: $N > 2M$

The upper triangular matrix $\mathbf{R}$ will be made up of a $2M \times 2M$ top upper triangular matrix and an $(N - 2M) \times 2M$ bottom zero matrix. Since the zero matrix does not contribute to the value of $\|\mathbf{R}\boldsymbol{\delta}\|^2$, the sphere decoding algorithm is also applicable in this case and the implementation is similiar.

The proposed LSD algorithm has a complexity in the order of $O((2M)^3)$ at a high SNR for proper choices of r [127]. This significantly reduces the computation complexity compared with exhaustive search, whose complexity is exponential in $2M$.

2.5 Asymptotic Error Probability Performance of the Proposed Scheme

In this section, we analyze the average error probability performance of our proposed scheme, where a complex-valued Rayleigh fading MIMO TWRC is considered. The performance analysis of the proposed scheme in the MIMO scenario is very challenging. We introduce new analytical tools and a novel concept on typical deep-fade events for the TWRC to tackle the challenges.

In the analysis of SISO linear PNC scheme, one only needs to characterize the typical error events w.r.t. a one-dimension superposition of the two users' signals, by considering the interference structure of each user's single signal stream. While for the MIMO case, the key challenge lies in that the interference structure of each signal stream becomes much more complicated, as each user's multiple streams are already interfering with each other and on top of that they also experience the interference from the other user. Therefore, we develop a new characterization of deep fade events in a multi-dimension space to conduct the typical error event analysis for the proposed scheme.

In this section, we start with presenting the main analytical result in Theorem 2.2. The proof of the theorem will be given at the end of this section.

Theorem 2.2. *As $\rho \to \infty$, the average error probability of the proposed scheme is given by*

$$P_e \lessapprox \frac{(\eta_1+\eta_2)(1+2^{-\frac{N}{2}})}{q^{2M}} \cdot \frac{4^N+3^{N+1}}{6} \left(\frac{\gamma^2}{\rho}\right)^N. \tag{2.40}$$

where

$$\eta_1 = \sum_{i=0}^{M} \sum_{\substack{j=0 \\ i+j\le M}}^{M} \binom{M}{i} \binom{M-i}{j} 4^{i+j} (q-2)^{2M-2i-j} (4M-2i-j), \tag{2.41}$$

and

$$\begin{aligned}\eta_2 =& \sum_{i=0}^{M} \sum_{\substack{j=0 \\ i+j\le M}}^{M} \binom{M}{i} \binom{M-i}{j} 4^{i+j} (q-2)^{2M-2i-j} (4M-3i-2j) \\ &+ 8M(M-1)q^2(q-1)^2. \end{aligned} \tag{2.42}$$

Clearly, our proposed scheme achieves a full diversity of N. However, the coding gain obtained by optimizing the NC matrix $\mathbf{G}$ is not clear yet. Therefore, we next establish a lower bound on the average error probability of any scheme for the Rayleigh fading MIMO TWRC.

Next, let us consider a case whose error performance will serve as a lower bound in our analysis. Consider two decoupled one-way relay channels, i.e. two users transmit signals to their relay respectively. In this case, each user's transmission is free of interference from the other user. If any of the two users makes an error at the relay, an error will be declared. We refer to this event as an interference-free error event as in [43]. Since this senario is the ideal case for TWRCs, we refer to this interference-free error probability as *interference-free lower bound*. A closed-form result on this interference-free lower bound is given as follows. The proof is given in Appendix 2.9.1.

Lemma 2.1. *The average error probability of interference-free lower bound is*

$$P_e^{LB} = \frac{(\eta_1+\eta_2)(1+2^{-\frac{N}{2}})}{q^{2M}} \cdot \frac{4^N+3^{N+1}}{6} \left(\frac{\gamma^2}{\rho}\right)^N. \tag{2.43}$$

where η_1 and η_2 are given in (2.41) and (2.42) respectively.

Corollary 2.1. *As $\rho \to \infty$, the average NC error probability of the linear vector PNC scheme with a prime q approaches that of interference-free lower bound.*

Remark 2.5. *Corollary 1 means that the proposed scheme achieves the optimal error probability at a high SNR for a Rayleigh fading MIMO TWRC.*

The remainder of this section is devoted to the proof of Theorem 2.2. We first present some preliminaries that will be used in the proof.

Definition 2.3 (Channel deep-fade event)**.** *Consider a fading channel coefficient matrix $\mathbf{H} \in \mathbb{R}^{N\times M}$. Define the following set*

$$\mathcal{H} \triangleq \left\{ \begin{array}{l} \mathbf{H} : \exists \boldsymbol{\delta},\ \boldsymbol{\delta} \in \{1-q, ..., q-1\}^M,\ \text{such that:} \\ \qquad \|\mathbf{H}\boldsymbol{\delta}\|^2 = K\rho^{-\in},\ \text{where} \in \geq 1 \text{ and } K \text{ is a constant.} \end{array} \right\}, \tag{2.44}$$

and $\mathcal{H}$ is referred to as the set of channel deep-fade events.

Remark 2.6. *Intuitively, for a single-user point-to-point MIMO system, when the channel deep-fade event specified in Definition 2 happens, there exists a DV $\boldsymbol{\delta}$ such that the distance $\|\mathbf{H}\boldsymbol{\delta}\|^2$ is on the order of $o(\rho^{-1})$, and the received signal is highly likely in error at a high SNR.*

Definition 2.4 (Artificial deep-fade event)**.** *In a MIMO TWRC, define*

$$\mathcal{F} \triangleq \left\{ \begin{array}{l} [\mathbf{H}_A, \mathbf{H}_B] : \mathbf{H}_A \notin \mathcal{H} \ \cap\ \mathbf{H}_B \notin \mathcal{H}, \\ \qquad \exists \boldsymbol{\delta},\ \boldsymbol{\delta} \in \{1-q, ..., q-1\}^{2M},\ \text{such that:} \\ \qquad \|[\mathbf{H}_A, \mathbf{H}_B]\,\boldsymbol{\delta}\|^2 = K'\rho^{-\in'}, \text{where} \in' \geq 1 \text{ and } K' \text{ is a constant.} \end{array} \right\}. \tag{2.45}$$

We call $\mathcal{F}$ the set of artificial deep-fade events.

Remark 2.7. *The artificial deep-fade event defined above characterizes the channel realizations of $\mathbf{H}_A$ and $\mathbf{H}_B$ that, although each individual $\mathbf{H}_A$ and $\mathbf{H}_B$ is not a deep-fade event, the superposition of the two users' signals results in that the conventional*

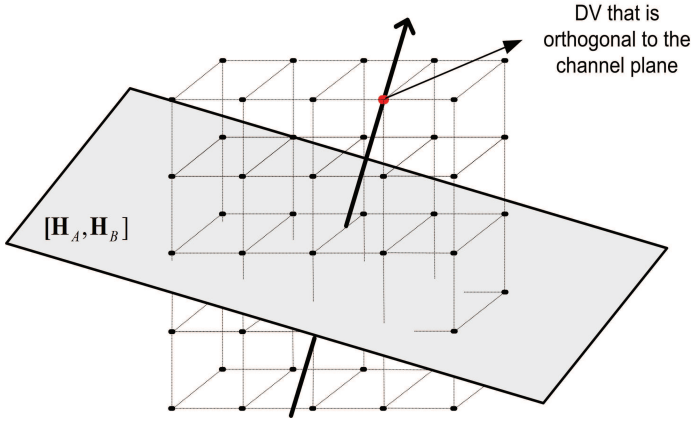


Figure 2.3. Geometrical illustration of the artificial deep-fade event.

minimum distance is very small, i.e., of order $o(\rho^{-1})$*. As the very small minimum distance is caused by the channel superposition instead of by the nature of each user's channel, we refer to such events as "artificial deep-fade events"Ẇe note that the artificial deep-fade event characterizes the so-called "removable singular fade subspace" in [44] or "minimum distance shortening events" in [12].*

A geometrical illustration for the artificial deep-fade event is depicted in Fig. 2.3. Assume that $\mathbf{H}_A$ and $\mathbf{H}_B$ do not belong to $\mathcal{H}$, i.e., each of them is not a channel deep-fade event. For illustration purpose, the N dimenional subspace spanned by $[\mathbf{H}_A, \mathbf{H}_B]$ in the $2M$-dimension whole space is depicted as a plane. In addition, all possible DVs form a bounded grid of integers in the $2M$-dimension space. As the SNR tends to infinity, for a certain realization of $[\mathbf{H}_A, \mathbf{H}_B]$, an artificial deep-fade event happens if there exists a DV which is very close to being perpendicular to the plane of $[\mathbf{H}_A, \mathbf{H}_B]$. An example of such DV is depicted in Fig. 2.3.

The key challenge in analyzing the performance of the proposed scheme is how to characterize the artificial deep-fade events. An approach is provided by following propositions. Their proofs are given in Appendix 2.9.2 - 2.9.4.

Proposition 2.1. *For any channel realization, there are at most* $2M - N$ *linearly independent DVs that are perpendicular to* $[\mathbf{H}_A, \mathbf{H}_B]$*.*

Suppose that there are L, $L \leq (2M - N)$, linearly independent DVs whose as-

sociated distances are equal to zero. Stack these L DVs into a matrix $\mathbf{\Delta}$, and we have

$$\|[\mathbf{H}_A, \mathbf{H}_B]\,\mathbf{\Delta}\| = 0. \tag{2.46}$$

Rewrite $\mathbf{\Delta}$ as $\mathbf{\Delta} = \left[\mathbf{\Delta}_A^T, \mathbf{\Delta}_B^T\right]^T$, we have the following proposition.

Proposition 2.2. *If $\mathbf{H}_A$ and $\mathbf{H}_B$ are both full rank, $\mathbf{\Delta}_A, \mathbf{\Delta}_B$ are both full-rank.*

Proposition 2.3. *Consider that the optimized $\mathbf{G}$ matrix specified in Theorem 2.1 is used. As $\rho \to \infty$, when $(\mathbf{H}_A \notin \mathcal{H}) \cap (\mathbf{H}_B \notin \mathcal{H})$, for all the values of DVs, $\boldsymbol{\delta} \in \{1-q, ..., q-1\}^{2M}$, the MSD satisfies $d_{MSD} = K\rho^{-\in}$ with $\in < 1$ and K is a constant.*

Remark 2. *If the artificial deep-fade events happens, the minimum distance between SI symbols will be on the order of $o(\rho^{-\in})$ where $\in \geq 1$. In this circumstance, for conventional PNC scheme, the MSD will remain as this vanishing minimum distance and will contribute to the NC error probability. However, Proposition 2.3 shows that using the optimized NC generator $\mathbf{G}$ in Theorem 2.1, the MSD will be larger, i.e., on the order of $o(\rho^{-\in})$ where $\in < 1$. Therefore, all the artificial deep-fade events will not contribute to the NC error probability. As a result, our proposed method will improve the decoding error performance of the MIMO TWRCs, in terms of coding gain, compared with complete decode-and-forward or PNC scheme with fixed NC generator.*

With the three propositions, we are in the position of proving Theorem 2.2.

Proof. Given channel realization $\mathbf{H}_A, \mathbf{H}_B$, the error probability of the proposed MIMO linear vector PNC scheme as $\rho \to \infty$ is upper bounded by

$$P_e/\mathbf{H}_A, \mathbf{H}_B \leq \frac{A^{(\mathbf{G})}(d_{MSD})}{q^{2M}} Q\left(\sqrt{\frac{\rho}{2\gamma^2} d_{MSD}}\right) \tag{2.47}$$

Case 1. $(\mathbf{H}_A \in \mathcal{H}) \cap (\mathbf{H}_B \notin \mathcal{H}.)$ In this case, only user A' channel is in channel deep-fade event. And $(\mathbf{H}_A \in \mathcal{H}) \cap (\mathbf{H}_B \notin \mathcal{H})$ can be interpreted as that $\|\mathbf{H}_A\boldsymbol{\delta}_A\|^2 = K_A\rho^{-\in_A}$with $\in_A \geq 1$ and $\|\mathbf{H}_B\boldsymbol{\delta}_B\|^2 = K_B\rho^{-\in_B}$with $\in_B < 1$.

The conventional minimum distance between SI symbol vectors can be written as

$$\begin{aligned} d_1 &= \min_{\boldsymbol{\delta}_A,\boldsymbol{\delta}_B} \|\mathbf{H}_A\boldsymbol{\delta}_A + \mathbf{H}_B\boldsymbol{\delta}_B\|^2 \\ &\leq \min_{\boldsymbol{\delta}_A,\boldsymbol{\delta}_B} \left(\|\mathbf{H}_A\boldsymbol{\delta}_A\| + \|\mathbf{H}_B\boldsymbol{\delta}_B\|\right)^2 \\ &= \min_{\boldsymbol{\delta}_A,\boldsymbol{\delta}_B} \left(\|\mathbf{H}_A\boldsymbol{\delta}_A\|^2 + \|\mathbf{H}_B\boldsymbol{\delta}_B\|^2 + 2\|\mathbf{H}_A\boldsymbol{\delta}_A\|\,\|\mathbf{H}_B\boldsymbol{\delta}_B\|\right) \end{aligned} \tag{2.48}$$

It is clear that as $\rho \to \infty$, $\|\mathbf{H}_A\boldsymbol{\delta}_A\|^2 \ll \|\mathbf{H}_B\boldsymbol{\delta}_B\|^2$. Note that the DV can be written as $\boldsymbol{\delta}_1 = \left[\boldsymbol{\delta}_{1,A}^T, \boldsymbol{\delta}_{1,B}^T\right]^T$. Thus, the DV w.r.t. the minimum distance satisfies $\boldsymbol{\delta}_{1,B} = \mathbf{0}$. This means as $\rho \to \infty$,

$$d_1 = \|\mathbf{H}_A\boldsymbol{\delta}_{1,A}\|^2. \tag{2.49}$$

In addition, any matrix containing a column $\left[\boldsymbol{\delta}_{1,A}^T, \mathbf{0}^T\right]^T$ is not block-wise full rank. Therefore, even for the optimized $\mathbf{G}$, $A^{(\mathbf{G})}(d_1)$ is non-zero and $d_{MSD} = d_1$.

Thus, the error probability is upper bounded by

$$P_e^{Case1}/\mathbf{H}_A, \mathbf{H}_B \leq \frac{A^{(\mathbf{G})}(d_1)}{q^{2M}} Q\left(\sqrt{\frac{\rho}{2\gamma^2}\|\mathbf{H}_A\boldsymbol{\delta}_{1,A}\|^2}\right) \tag{2.50}$$

It can be shown that the averaged multiplicity $A^{(\mathbf{G})}(d_1)/\ q^{2M}$ is the same as that where only user A exists. Let such multiplicity for the single-user case be $A'(d_1)$, we have

$$\begin{aligned} P_e^{Case1} &\leq \int_{\mathbf{H}_A\in\mathcal{H}_A\cap\mathbf{H}_B\notin\mathcal{H}_B} \frac{A'(d_1)}{q^M} Q\left(\sqrt{\frac{\rho}{2\gamma^2}\|\mathbf{H}_A\boldsymbol{\delta}_{1,A}\|^2}\right) p(\mathbf{H}_A)p(\mathbf{H}_B)d\mathbf{H}_A d\mathbf{H}_B \\ &= \int_{\mathbf{H}_A\in\mathcal{H}_A} \frac{A'(d_1)}{q^M} Q\left(\sqrt{\frac{\rho}{2\gamma^2}\|\mathbf{H}_A\boldsymbol{\delta}_{1,A}\|^2}\right) p(\mathbf{H}_A)d\mathbf{H}_A \int_{\mathbf{H}_B\notin\mathcal{H}_B} p(\mathbf{H}_B)d\mathbf{H}_B \\ &\leq \int_{\mathbf{H}_A} \frac{A'(d_1)}{q^M} Q\left(\sqrt{\frac{\rho}{2\gamma^2}\|\mathbf{H}_A\boldsymbol{\delta}_{1,A}\|^2}\right) p(\mathbf{H}_A)d\mathbf{H}_A \\ &\quad - \int_{\mathbf{H}_A\notin\mathcal{H}_A} \frac{A'(d_1)}{q^M} Q\left(\sqrt{\frac{\rho}{2\gamma^2}\|\mathbf{H}_A\boldsymbol{\delta}_{1,A}\|^2}\right) p(\mathbf{H}_A)d\mathbf{H}_A \end{aligned} \tag{2.51}$$

Here, the first term is equivalent to the error probability that only user A exists, which can be shown to be proportional to ρ^{-N}. We know that $\mathbf{H}_A \notin \mathcal{H}_A$ means

$\|\mathbf{H}_A\boldsymbol{\delta}_{1,A}\|^2$ be represented by $K\rho^{-\in}$ where $\in< 1$. So the second term

$$\begin{aligned}
&\int\limits_{\mathbf{H}_A\notin\mathcal{H}_A} \frac{A'(d_1)}{q^M}Q\left(\sqrt{\frac{\rho}{2\gamma^2}\|\mathbf{H}_A\boldsymbol{\delta}_{1,A}\|^2}\right)p(\mathbf{H}_A)d\mathbf{H}_A \\
&= \int\limits_{\mathbf{H}_A\mathcal{H}_A} \frac{A'(d_1)}{q^M}Q\left(\sqrt{\frac{K\rho^{1-\in}}{2\gamma^2}}\right)p(\mathbf{H}_A)d\mathbf{H}_A \\
&= \frac{A'(d_1)}{q^M}Q\left(\sqrt{\frac{K\rho^{1-\in}}{2\gamma^2}}\right)\int\limits_{\mathbf{H}_A\mathcal{H}_A} p(\mathbf{H}_A)d\mathbf{H}_A \propto e^{\rho^{\in-1}}
\end{aligned}$$

which decays much faster than the first term as $\rho\to\infty$. Then we get

$$P_e^{Case1}\lessapprox\int\limits_{\mathbf{H}_A}\frac{A'(d_1)}{q^M}Q\left(\sqrt{\frac{\rho}{2\gamma^2}\left\|\mathbf{H}_A\boldsymbol{\delta}_A^1\right\|^2}\right)p(\mathbf{H}_A)d\mathbf{H}_A. \tag{2.52}$$

Case 2. $(\mathbf{H}_A\notin\mathcal{H})\cap(\mathbf{H}_B\in\mathcal{H})$. In this case, only user B' channel is in channel deep-fade event. Applying similar analysis to that for Case 1, as $\rho\to\infty$, we have

$$d_{MSD}=d_1=\|\mathbf{H}_B\boldsymbol{\delta}_{1,B}\|^2. \tag{2.53}$$

and

$$P_e^{Case2}\leq\int\limits_{\mathbf{H}_B}\frac{A'(d_1)}{q^M}Q\left(\sqrt{\frac{\rho}{2\gamma^2}\|\mathbf{H}_B\boldsymbol{\delta}_{1,B}\|^2}\right)p(\mathbf{H}_B)d\mathbf{H}_B. \tag{2.54}$$

Case 3. $(\mathbf{H}_A\in\mathcal{H})\cap(\mathbf{H}_B\in\mathcal{H})$. In this case, both users' channels are in channel deep-fade event. Then,

$$\begin{aligned}
P_e^{Case3} &= \int\limits_{\mathbf{H}_A\in\mathcal{H}_A\cap\mathbf{H}_B\in\mathcal{H}_B} P_e^{Case3}/\mathbf{H}_A,\mathbf{H}_B p(\mathbf{H}_A)p(\mathbf{H}_B)d\mathbf{H}_A d\mathbf{H}_B \\
&\leq \int\limits_{\mathbf{H}_A\in\mathcal{H}_A\cap\mathbf{H}_B\in\mathcal{H}_B} p(\mathbf{H}_A)p(\mathbf{H}_B)d\mathbf{H}_A d\mathbf{H}_B \\
&= \int\limits_{\mathbf{H}_A\in\mathcal{H}_A} p(\mathbf{H}_A)d\mathbf{H}_A \int\limits_{\mathbf{H}_B\in\mathcal{H}_B} p(\mathbf{H}_B)d\mathbf{H}_B.
\end{aligned} \tag{2.55}$$

It can be shown that P_e^{Case3} is proportional to ρ^{-2N}.

Case 4. $(\mathbf{H}_A\notin\mathcal{H})\cap(\mathbf{H}_B\notin\mathcal{H})\cap([\mathbf{H}_A,\mathbf{H}_B]\in\mathcal{F})$. In this case, neither of the two users' channel is in channel deep-fade event but their superposition makes them in an artificial deep-fade event.

Using Proposition 2.3, with Theorem 2.1, d_{MSD} can always be represented by $K\rho^{-\in}$ where $\in< 1$.

$$\begin{aligned}
P_e^{Case4} &= \int\limits_{[\mathbf{H}_A,\mathbf{H}_B]\in\mathcal{F}} P_e^{Case4}/\mathbf{H}_A,\mathbf{H}_B p(\mathbf{H}_A)p(\mathbf{H}_B)d\mathbf{H}_A d\mathbf{H}_B \\
&\leq \int\limits_{[\mathbf{H}_A,\mathbf{H}_B]\in\mathcal{F}} \frac{A^{(\mathbf{G})}(d_{MSD})}{q^{2M}} Q\left(\sqrt{\frac{\rho}{2\gamma^2}d_{MSD}}\right) p(\mathbf{H}_A)p(\mathbf{H}_B)d\mathbf{H}_A d\mathbf{H}_B \\
&= \frac{A^{(\mathbf{G})}(d_{MSD})}{q^{2M}} Q\left(\sqrt{\frac{K\rho^{1-\in}}{2\gamma^2}}\right) \int\limits_{[\mathbf{H}_A,\mathbf{H}_B]\in\mathcal{F}} p(\mathbf{H}_A)p(\mathbf{H}_B)d\mathbf{H}_A d\mathbf{H}_B \\
&< \frac{A^{(\mathbf{G})}(d_{MSD})}{q^{2M}} Q\left(\sqrt{\frac{K\rho^{1-\in}}{2\gamma^2}}\right) \propto e^{\rho^{\in-1}}. \qquad (2.56)
\end{aligned}$$

Case 5. $(\mathbf{H}_A \notin \mathcal{H}) \cap (\mathbf{H}_B \notin \mathcal{H}) \cap ([\mathbf{H}_A, \mathbf{H}_B] \notin \mathcal{F})$. In this case, neither of the two users' channel is in channel deep-fade event, and their superposition does not result in artificial deep-fade events either.

From Definition 2, $[\mathbf{H}_A, \mathbf{H}_B] \notin \mathcal{F}$ means that $\forall \boldsymbol{\delta} \in \{1-q, ..., q-1\}^{2M}$, $\|[\mathbf{H}_A, \mathbf{H}_B]\,\boldsymbol{\delta}\|^2 = K\rho^{-\in}$ where $\in< 1$. This means in this case the distance between any different SI symbol vectors is always non-vanishing. The error probability for Case 5 is proportional to $e^{\rho^{\in-1}}$.

With above categorization, as $\rho \to \infty$, we have

$$\begin{aligned}
P_e &= P_e^{Case1} + P_e^{Case2} + P_e^{Case3} + P_e^{Case4} + P_e^{Case5} \\
&\leq \int\limits_{\mathbf{H}_A} \frac{A'(d_1)}{q^M} Q\left(\sqrt{\frac{\rho}{2\gamma^2}\left\|\mathbf{H}_A\boldsymbol{\delta}_{1,A}\right\|^2}\right) p(\mathbf{H}_A)d\mathbf{H}_A \\
&\quad + \int\limits_{\mathbf{H}_B} \frac{A'(d_1)}{q^M} Q\left(\sqrt{\frac{\rho}{2\gamma^2}\left\|\mathbf{H}_B\boldsymbol{\delta}_{1,B}\right\|^2}\right) p(\mathbf{H}_B)d\mathbf{H}_B \\
&\quad + o(\rho^{-2N}) + o\left(e^{\rho^{\in-1}}\right) + o\left(e^{\rho^{\in-1}}\right) \\
&\lessapprox \int\limits_{\mathbf{H}_A} \frac{A'(d_1)}{q^M} Q\left(\sqrt{\frac{\rho}{2\gamma^2}\left\|\mathbf{H}_A\boldsymbol{\delta}_{1,A}\right\|^2}\right) p(\mathbf{H}_A)d\mathbf{H}_A \\
&\quad + \int\limits_{\mathbf{H}_B} \frac{A'(d_1)}{q^M} Q\left(\sqrt{\frac{\rho}{2\gamma^2}\left\|\mathbf{H}_B\boldsymbol{\delta}_{1,B}\right\|^2}\right) p(\mathbf{H}_B)d\mathbf{H}_B, \qquad (2.57)
\end{aligned}$$

as the first two terms are proportional to ρ^{-N} while the third term is $o(\rho^{-2N})$ and the last two terms are $o\left(e^{\rho^{\epsilon-1}}\right)$. In addition, we can derive a closed-form expression on the first two terms, given as

$$P_e \lessapprox P_e^{Case1} + P_e^{Case2} = \frac{(\eta_1+\eta_2)(1+2^{-\frac{N}{2}})}{q^{2M}} \cdot \frac{4^N+3^{N+1}}{6} \left(\frac{\gamma^2}{\rho}\right)^N. \tag{2.58}$$

where $\eta_1$and η_2 are given in (2.41) and (2.42) respectively. The proof of the expression in (2.58) is given in the Appendix 2.9.

This completes the proof. □

2.6 Numerical Results for MIMO TWRC

In this section, we present numerical results for the error-rate performance of the proposed linear vector PNC scheme over MIMO TWRCs. The results are obtained by averaging over more than 10,000,000 channel realizations, where the fading channel coefficients follow i.i.d. Rayleigh distribution. For comparison purpose, we also show the performance of a complete decoding scheme, where the relay completely decodes all users' messages, and a non-optimized linear vector PNC scheme where a fixed generator matrix is used for all channel realizations. We also include the numerical results of an interference free lower bound, obtained by assuming that there is no interference between users' signals at the relay. Moreover, the derived asymptotic error probability of the proposed scheme in (2.40) is also shown in the figures.

Fig. 2.4 presents the error-rate performance of an $M = N = 2$ system where $q = 3$ (9-QAM for each antenna of each user). We observe that our proposed scheme exhibits a 4.5 dB improvement over complete decoding scheme and the non-optimized PNC scheme at the error rate of 10^{-4}. Note that the non-optimized PNC scheme has the same performance as the complete decoding scheme. This is because a fixed generator matrix cannot optimize the minimum set-distance for all channel realizations and thus for most cases, the minimum set-distance remains as the minimum SI symbol distance. In addition, we observe that at a sufficiently high SNR, the proposed

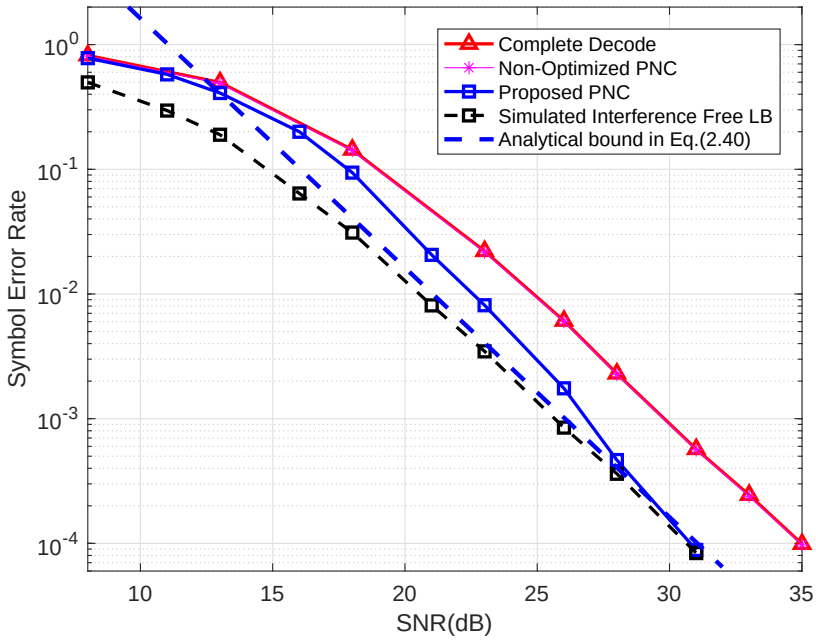


Figure 2.4. Error-rate performance of MIMO PNC scheme in a Rayleigh fading TWRC (9-QAM, $M = 2, N = 2$).

linear vector PNC scheme achieves the interference free lower bound. Also, it is demonstrated that our analytical error probability performance in (2.40) matches with the numerical result. This agrees with Theorem 2.2 and Corollary 2.1, which demonstrates the asymptotic optimality of error-rate performance of the proposed scheme. Fig. 2.5 presents the error-rate performance of an $M = N = 2$ system with $q = 7$. Similar performance trend as in Fig. 2.4 is observed. Moreover, we observe that the performance improvement of the proposed scheme increases as q increases.

Fig. 2.6 presents the error-rate performance of an $M = N = 3$ system where $q = 3$. We observe that the proposed scheme outperforms complete decoding scheme by up to about 3.8 dB in the high SNR region. Fig. 2.7 presents the error-rate performance of an $M = 3$ and $N = 4$ system where $q = 3$. We see that the proposed scheme outperforms complete decoding scheme by about 1.2 dB in the high SNR region. In both figures, the proposed linear vector PNC scheme achieves the interference free lower bound at a sufficiently high SNR and our analytical error probability

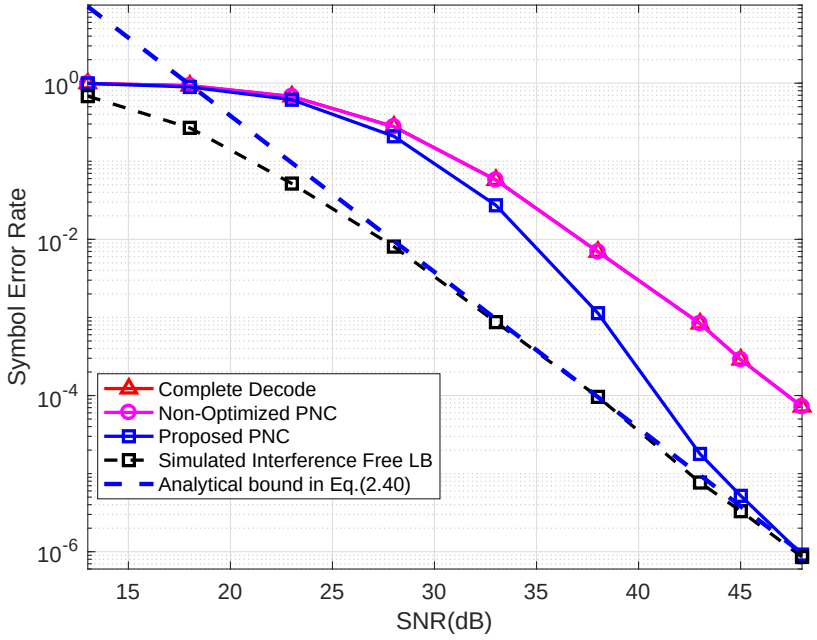


Figure 2.5. Error-rate performance of MIMO PNC scheme in a Rayleigh fading TWRC (49-QAM, $M = N = 2$).

performance matches with the numerical result. This demonstrates the asymptotic optimality of error-rate performance of the the proposed scheme. From Fig. 2.6 and Fig. 2.7, we observe that the proposed scheme yields a greater improvement over the benchmark scheme as M/N increases.

2.7 Physical-layer Network Coding for Y-channel

In this section, we propose a PNC scheme for a Y-channel model where the transmitters have no CSI. Here, we consider the full data exchange case where each user wants all messages from all other users. Under a multi-antenna relay setup, we present an explicit solution for NC generator matrices that minimize the NC error probability at a high SNR. We also provide and prove an approximation of the average NC error probability for the proposed scheme at a high SNR. In addition, the case of a single-antenna relay with multiple time-slots is discussed and different transmission

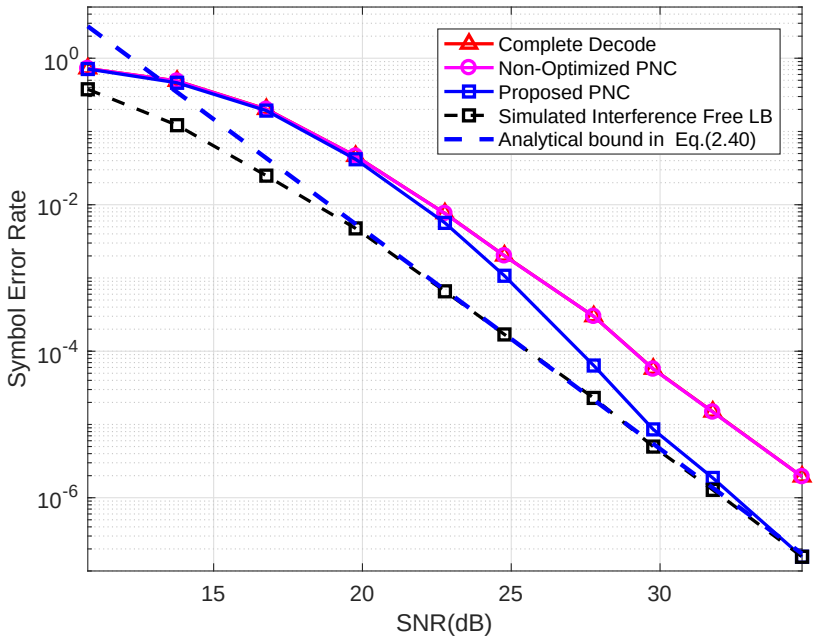


Figure 2.6. Error-rate performance of MIMO PNC scheme in a Rayleigh fading TWRC (9-QAM, $M = N = 3$).

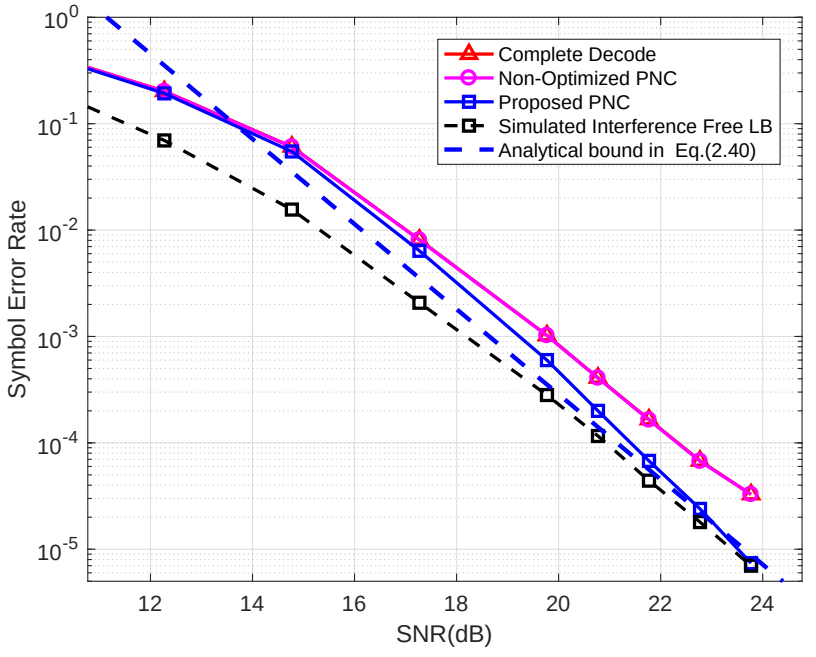


Figure 2.7. Error-rate performance of MIMO PNC scheme in a Rayleigh fading TWRC (9-QAM, $M = 3, N = 4$).

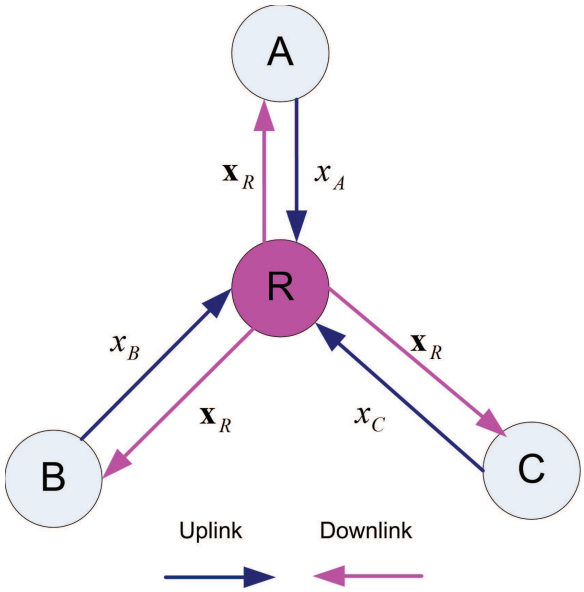


Figure 2.8. System model for the Y-channel.

strategies are compared through numerical results.

2.7.1 System Model and Proposed Scheme for Y-channel

In this section, we consider a Y-channel model with three single-antenna users A, B, C and one two-antenna relay[3]. Each user has an independent message to the other two users. Each round of information exchange consists of two stages, the uplink phase and the downlink phase as shown in Fig. 2.8. We assume a flat block fading for both uplink and downlink.

Uplink

We assume that one time-slot is allocated to the uplink phase. Each user transmits two independent messages via the in-phase and quadrature-phase, respectively. Denote the two messages of user l by w_l^{Re} and w_l^{Im}, $l \in \{A, B, C\}$, which are inde-

[3]In a MWRC model, we will not consider the case that the number of antennas at the relay is larger than the total number of antennas of all the users. This is because the relay will be able to fully decode all users' messages under this circumstance

pendently drawn from a finite field $GF(q)$.

Using q^2-order QAM, user l's transmitted messages are one-to-one mapped to a modulated symbol x_l by

$$x_l = \frac{1}{\gamma}\left[\left(w_l^{\mathrm{Re}} - \frac{q-1}{2}\right) + j\left(w_l^{\mathrm{Im}} - \frac{q-1}{2}\right)\right], \tag{2.59}$$

where γ is a normalization factor ensuring $E(|x_l^2|) = 1$.

In the uplink phase, three users transmit their signals to the relay simultaneously. Let $\mathbf{h}_l \in \mathbb{C}^{2\times 1}$ denote the fading channel coefficient between user l and the relay. The received signal at the relay is

$$\mathbf{y} = \sqrt{E_s}\mathbf{h}_A x_A + \sqrt{E_s}\mathbf{h}_B x_B + \sqrt{E_s}\mathbf{h}_C x_C + \mathbf{z}, \tag{2.60}$$

where E_s is the average transmitted power per symbol, and $\mathbf{z}$ is the complex-valued AWGN at the relay whose entries are i.i.d. with zero mean and variance σ_z^2. The SNR per user is defined as $\rho = E_s/|\sigma_z^2|$. We assume that the transmitter-side CSI is not available while the receiver-side CSI is perfectly known as in the TWRC case.

Relay

We propose a linear PNC strategy at the relay as follows. Based on the receiver CSI, the relay first selects NC coefficients from the same finite field $GF(q)$ for each user, and then generates the NC codewords u_n, $n \in \{1,2,3,4\}$ from the received signals. More specifically,

$$\begin{aligned} u_n \;=\; & \left(v_{A_n}^{\mathrm{Re}} \otimes w_{A_n}^{\mathrm{Re}}\right) \oplus \left(v_{A_n}^{\mathrm{Im}} \otimes w_{A_n}^{\mathrm{Im}}\right) \oplus \left(v_{B_n}^{\mathrm{Re}} \otimes w_{B_n}^{\mathrm{Re}}\right) \\ & \oplus \left(v_{B_n}^{\mathrm{Im}} \otimes w_{B_n}^{\mathrm{Im}}\right) \oplus \left(v_{C_n}^{\mathrm{Re}} \otimes w_{C_n}^{\mathrm{Re}}\right) \oplus \left(v_{C_n}^{\mathrm{Im}} \otimes w_{C_n}^{\mathrm{Im}}\right), \end{aligned} \tag{2.61}$$

where the NC coefficients $v_{l_n}^{\mathrm{Re}}$,$v_{l_n}^{\mathrm{Im}} \in GF(q)$. Denote a genuine message vector by

$$\mathbf{w} = \left[\; w_A^{\mathrm{Re}} \quad w_A^{\mathrm{Im}} \quad w_B^{\mathrm{Re}} \quad w_B^{\mathrm{Im}} \quad w_C^{\mathrm{Re}} \quad w_C^{\mathrm{Im}} \;\right]^T. \tag{2.62}$$

Then, we have the *linear NC codeword*

$$\mathbf{u} = \left[\; u_1 \quad u_2 \quad u_3 \quad u_4 \;\right]^T = \left[\; \mathbf{V}_A \quad \mathbf{V}_B \quad \mathbf{V}_C \;\right] \otimes \mathbf{w}, \tag{2.63}$$

where

$$\mathbf{V}_l = \begin{bmatrix} v_{l_1}^{\mathrm{Re}} & v_{l_2}^{\mathrm{Re}} & v_{l_3}^{\mathrm{Re}} & v_{l_4}^{\mathrm{Re}} \\ v_{l_1}^{\mathrm{Im}} & v_{l_2}^{\mathrm{Im}} & v_{l_3}^{\mathrm{Im}} & v_{l_4}^{\mathrm{Im}} \end{bmatrix}^T, \; l \in \{A, B, C\}, \tag{2.64}$$

and it is referred to as the*NC generator matrix*. Note that the additions and multiplications here are operated on the finite field $GF(q)$.

Note that within one coherent time, the NC generator matrices are selected and delivered only once. We assume that all the three users are aware of the selected generator matrices via reliable links.

In order to enable each user to recover all the other two users' messages, the following condition must be satisfied:

Full-Rank Condition: Matrices $[\ \mathbf{V}_A \quad \mathbf{V}_B\ ]$, $[\ \mathbf{V}_A \quad \mathbf{V}_C\ ]$ and $[\ \mathbf{V}_B \quad \mathbf{V}_C\ ]$ must be of full rank.

In our proposed PNC scheme, instead of completely decoding all users' messages $\mathbf{w}$, the relay directly decodes the NC codeword $\mathbf{u}$ in (2.63), based on the superimposed signal $\mathbf{y}$. Assume that the decoded NC codeword by the relay is $\hat{\mathbf{u}}$. The NC error probability $P_e^{NC} \triangleq \Pr\{\mathbf{u} \neq \hat{\mathbf{u}}\}$ can be minimized through the design of NC generator matrix.

Downlink

In the downlink phase, the relay modulates $\hat{\mathbf{u}}$ and broadcasts the modulated signal $\mathbf{x}_R$ to all users. We see that two time-slots are required for the downlink phase since $\hat{\mathbf{u}}$ consists of four messages. Suppose that a correct NC codeword $\mathbf{u}$ is decoded and delivered to each user. User A first removes its own messages w_A^{Re}, w_A^{Im} by computing

$$\begin{aligned} &\mathbf{u} \ominus \left(\mathbf{V}_A \otimes [\ w_A^{\mathrm{Re}} \quad w_A^{\mathrm{Im}}\ ]^T \right) \\ &= [\ \mathbf{V}_B \quad \mathbf{V}_C\ ] \otimes [\ w_B^{\mathrm{Re}} \quad w_B^{\mathrm{Im}} \quad w_C^{\mathrm{Re}} \quad w_C^{\mathrm{Im}}\ ]^T. \end{aligned}$$

The resultant message vector forms a set of linear equations with undetermined variables w_B^{Re}, w_B^{Im}, w_C^{Re} and w_C^{Im}. Since matrix $[\ \mathbf{V}_B \quad \mathbf{V}_C\ ]$ is of full rank, user A can recover its desired messages from user B and user C by solving this linear equation

set. It can be seen that one linear NC codeword $\mathbf{u}$ must consist of four elements so that the intended four messages for each user can be resolved through the linear equation set. User B and user C resolve their desired messages in a similar way. This finishes one round of message exchange.

2.7.2 Design of Linear Physical-layer Network Coding for Y-channel

2.7.3 Problem Formulation

Our goal here is to design the NC generator matrices that minimize the NC error probability. The problem is formulated as

$$\begin{aligned} \mathbf{V}_A^{Opt}, \mathbf{V}_B^{Opt}, \mathbf{V}_C^{Opt} &= \arg \min_{\mathbf{V}_A, \mathbf{V}_B, \mathbf{V}_C} P_e^{NC} , \\ &\quad \text{s.t. Full-Rank Condition} \end{aligned} \tag{2.65}$$

Denote that $w_l = w_l^{\mathrm{Re}} + j w_l^{\mathrm{Im}}$. For a given channel realization of $\mathbf{h}_A$, $\mathbf{h}_B$ and $\mathbf{h}_C$, define

$$\mathbf{w}_s \triangleq \mathbf{h}_A w_A + \mathbf{h}_B w_B + \mathbf{h}_C w_C. \tag{2.66}$$

We refer to $\mathbf{w}_s$ as a *superimposed* (SI) *symbol*. Note that a SI symbol $\mathbf{w}_s$ is not on finite field as w_l is. Then, the distance between two different SI symbols $\mathbf{w}_s$ and $\mathbf{w}_s'$ is $d = \|\mathbf{w}_s - \mathbf{w}_s'\|^2$.

For any two SI symbols whose corresponding NC codewords $\mathbf{u}$ and $\mathbf{u}'$ are different, the minimum distance between them is defined as the minimum set-distance (MSD), i.e.,

$$d_{MSD} \triangleq \min_{w_s, w_s', \mathbf{u} \neq \mathbf{u}'} d. \tag{2.67}$$

We see that at a high SNR, the NC error probability P_e^{NC} is dominated by the minimum set-distance d_{MSD} [43]. The problem in (2.65) is now transformed into

$$\begin{aligned} \mathbf{V}_A^{Opt}, \mathbf{V}_B^{Opt}, \mathbf{V}_C^{Opt} &= \arg \max_{\mathbf{V}_A, \mathbf{V}_B, \mathbf{V}_C} d_{MSD} . \\ &\quad \text{s.t. Full-Rank Condition} \end{aligned} \tag{2.68}$$

2.7.4 Solution to (2.68)

For two different SI symbols $\mathbf{w}_s$ and $\mathbf{w}'_s$, let $\delta_A = w_A - w'_A$, $\delta_B = w_B - w'_B$, $\delta_C = w_C - w'_C$, and we call $\boldsymbol{\delta} = [\ \delta_A^{\mathrm{Re}}\ \ \delta_A^{\mathrm{Im}}\ \ \delta_B^{\mathrm{Re}}\ \ \delta_B^{\mathrm{Im}}\ \ \delta_C^{\mathrm{Re}}\ \ \delta_C^{\mathrm{Im}}\ ]^T$ the *difference vector* (DV) of w_s and w'_s. Note that $\boldsymbol{\delta} \in \{1-q, \cdots, 0, \cdots, q-1\}^6$ and $\|\boldsymbol{\delta}\| \neq 0$. We see that the distance between w_s and w'_s is

$$d = \left\| \begin{bmatrix} \mathbf{h}_A^{\mathrm{Re}} & -\mathbf{h}_A^{\mathrm{Im}} & \mathbf{h}_B^{\mathrm{Re}} & -\mathbf{h}_B^{\mathrm{Im}} & \mathbf{h}_C^{\mathrm{Re}} & -\mathbf{h}_C^{\mathrm{Im}} \\ \mathbf{h}_A^{\mathrm{Im}} & \mathbf{h}_A^{\mathrm{Re}} & \mathbf{h}_B^{\mathrm{Im}} & \mathbf{h}_B^{\mathrm{Re}} & \mathbf{h}_C^{\mathrm{Im}} & \mathbf{h}_C^{\mathrm{Re}} \end{bmatrix} \cdot \boldsymbol{\delta} \right\|^2. \tag{2.69}$$

Let $\boldsymbol{\delta}^*$ denote the DV corresponding to the minimum distance of SI symbols subject to that $|\delta_A| \neq 0$, $|\delta_B| \neq 0$ and $|\delta_C| \neq 0$, i.e.,

$$\boldsymbol{\delta}^* = \underset{\substack{|\delta_A|\neq 0, |\delta_B|\neq 0, \\ |\delta_C|\neq 0}}{\arg\min}\ d. \tag{2.70}$$

Note that for the same channel matrix in (2.69), the following DV

$$\widetilde{\boldsymbol{\delta}}^* = \begin{bmatrix} -\delta_A^{*\mathrm{Im}} & \delta_A^{*\mathrm{Re}} & -\delta_B^{*\mathrm{Im}} & \delta_B^{*\mathrm{Re}} & -\delta_C^{*\mathrm{Im}} & \delta_C^{*\mathrm{Re}} \end{bmatrix}^T \tag{2.71}$$

corresponds to the same distance d as $\boldsymbol{\delta}^*$.

The solution to (2.68) at high SNRs is as follows.

Theorem 2.3. *As $\rho \to \infty$, the NC generator matrices that minimize the NC error probability satisfy*

$$[\ \mathbf{V}_A\ \ \mathbf{V}_B\ \ \mathbf{V}_C\ ] \otimes \bmod \left(\begin{bmatrix} \boldsymbol{\delta}^* & \widetilde{\boldsymbol{\delta}}^* \end{bmatrix}, q \right) = 0. \tag{2.72}$$

It can be shown that, with (4.41), matrices $[\ \mathbf{V}_A\ \ \mathbf{V}_B\ ]$, $[\ \mathbf{V}_A\ \ \mathbf{V}_C\ ]$ and $[\ \mathbf{V}_B\ \ \mathbf{V}_C\ ]$ are of full rank if and only if $\boldsymbol{\delta}^*$ satisfies that $|\delta_A| \neq 0$, $|\delta_B| \neq 0$ and $|\delta_C| \neq 0$. For example, if $|\delta_A| = 0$, $[\ \mathbf{V}_B\ \ \mathbf{V}_C\ ]$ is not of full rank; and if $|\delta_A| = 0$ and $|\delta_B| = 0$, $\mathbf{V}_C$ itself is not of full rank. And vice versa. Thus, the constraint for the DV $\boldsymbol{\delta}^*$ in (2.70) ensures that the Full-Rank Condition for the NC generator matrices is met.

Proof. Denote the minimum distance between any two SI symbols by d_1. We see that d_1 is independent of any constraints on δ_A, δ_B and δ_C.

Let us first consider the case that the DV corresponding to d_1 happens to satisfy that $|\delta_A| \neq 0$, $|\delta_B| \neq 0$ and $|\delta_C| \neq 0$. Suppose that the DVs w.r.t. the minimum distance d_1 are $\boldsymbol{\delta}^*$ and $\widetilde{\boldsymbol{\delta}}^*$ ($\boldsymbol{\delta}^*$ and $-\boldsymbol{\delta}^*$ are regarded as the same vector.). We then let the NC generator matrices satisfy

$$\begin{aligned}
&[\ \mathbf{V}_A \quad \mathbf{V}_B \quad \mathbf{V}_C\ ] \otimes \mathrm{mod}(\mathbf{w} \pm \boldsymbol{\delta}^*), q) \\
&= [\ \mathbf{V}_A \quad \mathbf{V}_B \quad \mathbf{V}_C\ ] \otimes \mathrm{mod}([\mathbf{w} \pm \widetilde{\boldsymbol{\delta}}^*), q) \\
&= [\ \mathbf{V}_A \quad \mathbf{V}_B \quad \mathbf{V}_C\ ] \otimes \mathrm{mod}(\mathbf{w}, q),
\end{aligned} \tag{2.73}$$

which is equivalent to (4.41). With (4.41), any two SI symbols with distance d_1 correspond to the same linear NC codeword. Thus, the distance between any two SI symbols w.r.t. different NC codewords will be larger than d_1, i.e., $d_{MSD} > d_1$. In this manner, the minimum set-distance is increased, which reduces the NC error probability at high SNRs.

We then consider the case that the DV corresponding to d_1 has $|\delta_A| = 0$, $|\delta_B| = 0$ or $|\delta_C| = 0$. In this case, there does not exist matrices $\mathbf{V}_A$, $\mathbf{V}_B$ and $\mathbf{V}_C$ that satisfy (4.41) and the Full-Rank Condition at the same time. In other words, there is no solution guaranteeing $d_{MSD} > d_1$ under the Full-Rank Condition. In fact, this case corresponds to the deep-fade events that cannot be eliminated through (4.41), which account for the main cause of decoding errors in our scheme.

So far we have shown that (4.41) is sufficient, next we prove that it is also necessary. We assume choosing NC generator matrices $\mathbf{V}_A^{/}$, $\mathbf{V}_B^{/}$ and $\mathbf{V}_C^{/}$ which do not satisfy (4.41), i.e.,

$$[\ \mathbf{V}_A^{/} \quad \mathbf{V}_B^{/} \quad \mathbf{V}_C^{/}\ ] \otimes \mathrm{mod}\left(\left[\ \boldsymbol{\delta}^* \quad \widetilde{\boldsymbol{\delta}}^*\ \right], q\right) \neq 0. \tag{2.74}$$

In this case, any two SI symbols with distance d_1 will correspond to two different linear NC codewords. Thus, d_{MSD} strictly equals to d_1. Therefore, the NC generator matrices not satisfying (4.41) will result in a smaller NC set-distance d_{MSD}, compared to those satisfying (4.41). In other words, $[\ \mathbf{V}_A^{/} \quad \mathbf{V}_B^{/} \quad \mathbf{V}_C^{/}\ ]$ cannot minimize the NC error probability at high SNRs. This completes the proof. □

2.7.5 Average Error Probability Performance Analysis of the Proposed Scheme

In this section, we analyse the average error probability of the proposed scheme over all channel realizations.

Preliminaries

We first present some preliminaries used in the analysis by introducing two basic schemes and related deep-fade events.

Consider a basic single user scheme where only user l transmits to the relay, i.e., $y = h_l x_l + n$. The average error probability is denoted by

$$P_e^l \triangleq \Pr\left\{\widehat{w}_l \neq w_l\right\}. \tag{2.75}$$

Consider a basic two-user scheme where two users l and k transmit to the relay simultaneously, i.e., $y = h_l x_l + h_k x_k + n$. Denote the average error probability by

$$P_e^{lk} \triangleq \Pr\left\{[\widehat{w}_l, \widehat{w}_k] \neq [w_l, w_k]\right\}. \tag{2.76}$$

Definition 2.5 (Multi-user deep-fade event)**.** *In a multi-way relay system where all users can be represented by a set $\mathcal{L}$, define*

$$\mathcal{F} \triangleq \left\{ \begin{array}{l} [\ \mathbf{h}_A \quad \mathbf{h}_B \quad \cdots\ ] : \mathbf{h}_l \notin \mathcal{H}, l \in \mathcal{L} \\ \qquad\qquad \exists \boldsymbol{\delta},\ \boldsymbol{\delta} \in \{1-q, ..., q-1\}^{|\mathcal{L}|},\ \textit{such that:} \\ \qquad\qquad \left\| [\ \mathbf{h}_A \quad \mathbf{h}_B \quad \cdots\ ]\boldsymbol{\delta} \right\|^2 = K' \rho^{-\theta'}, \\ \qquad\qquad \textit{where } \theta' \geq 1 \textit{ and } K' \textit{ is a constant.} \end{array} \right\}.$$

We call $\mathcal{F}$ the set of $|\mathcal{L}|$-user artificial deep-fade events.

Intuitively, for a multi-way relay system, if the $|\mathcal{L}|$-user deep-fade event happens, the minimum distance between SI symbols will be on the order of $o(\rho^{-\theta})$ where $\theta \geq 1$, and the decoding of the received signal at the relay is highly likely in error at a high SNR. Particularly, when $|\mathcal{L}| = 1$, the set $\mathcal{F}$ corresponds to the channel deep fade event for a standard point-to-point transmission.

For the Y-channel setup where three users transmits to the relay simultaneously, the single-user deep-fade events, the two-user deep-fade events and the three-user deep-fade events all contribute to the error probability at high SNRs. However, we next show that the contribution of the three-user deep-fade events to the NC error probability is eliminated, with properly chosen generator matrices using Theorem 2.3.

Proposition 2.4. *As $\rho \to \infty$, when $[\ \mathbf{h}_A \quad \mathbf{h}_B \quad \mathbf{h}_C\ ] \in \mathcal{F}$, using the optimized generator matrices in Theorem 2.3, the MSD satisfies $d_{MSD} = K\rho^{-\theta}$ where $\theta < 1$ and K is a constant.*

Proof. According to Definition 2, when $\left[\begin{array}{ccc} \mathbf{h}_A & \mathbf{h}_B & \mathbf{h}_C \end{array}\right] \in \mathcal{F}$, there exists a DV $\boldsymbol{\delta}$ such that $\left\| \left[\begin{array}{ccc} \mathbf{h}_A & \mathbf{h}_B & \mathbf{h}_C \end{array}\right] \boldsymbol{\delta} \right\|^2$ is of order $K\rho^{-\theta}$, $\theta \geq 1$. Note that for any channel realization, there are at most one linearly independent DVs that are perpendicular to $\left[\begin{array}{ccc} \mathbf{h}_A & \mathbf{h}_B & \mathbf{h}_C \end{array}\right]$. In other words, there exists at most one such DV whose distance are of order $K\rho^{-\theta}$, $\theta \geq 1$. With the equation (4.41) in Theorem 2.3, any SI symbols whose distance is of order $K\rho^{-\theta}$, $\theta \geq 1$, are clustered together. In other words, the multiplicity of all the distances of order $K\rho^{-\theta}$, $\theta \geq 1$, is now zero. Therefore, the contribution of the three-user deep-fade events to the NC error probability is negligible. This completes the proof. □

NC error-probability approximations

With above preliminaries, we obtain an approximation of the average NC error probability for the proposed scheme at high SNRs as follows.

Theorem 2.4. *For a PNC method based on Theorem 2.3, as $\rho \to \infty$, the average NC error-probability is approximated as*

$$P_e^{NC} \to P_e^* \triangleq P_e^A + P_e^B + P_e^C + P_e^{AB} + P_e^{AC} + P_e^{BC}. \tag{2.77}$$

Proof. The proof can be completed by performing a pairwise error analysis, and using the concept of "Multi-user deep-fade event" and Proposition 2.4. □

2.7.6 Numerical Results

This section presents numerical results for the error-rate performance of the proposed scheme. We assume that the fading channel coefficients follow i.i.d. Rayleigh distribution and the results are obtained by averaging over 100,000 channel realizations. We compare our proposed scheme with the performance of a complete decoding scheme, where the relay completely and jointly decodes all users' messages. Numerical results of the error-probability approximation in (2.77) are also shown in the figures.

Fig. 2.9 shows the error-rate performance for 9-QAM signals. We observe that using the optimized generator matrices in Theorem 2.3, our proposed scheme outperforms the complete decoding scheme by 1 dB at the error-rate of 10^{-3}. In addition, it is shown that the proposed linear PNC scheme approaches the error-probability approximation in (2.77) at high SNRs.

Similar performance behavior is observed for higher level modulations. For example, Fig. 2.10 shows that when $q = 7$, a 2.5 dB improvement over the complete decoding scheme is achieved at the error-rate of 10^{-3}, and the error-probability approximation in (2.77) is approached at high SNRs.

2.7.7 Discussions for a Single Antenna Scenario

We then present some discussions for a single-antenna relay setup to provide further understanding for the Y-channel model. Suppose that T time-slots are used for the uplink phase. Note that we will not consider the case where $T > 2$. This is because for $T > 2$, the relay will be able to well distinguish the three user's messages based on its observations in three or more time-slots and this is not the appealing zone of network coding. Also, it is of more interest when we use as less time-slots as possible.

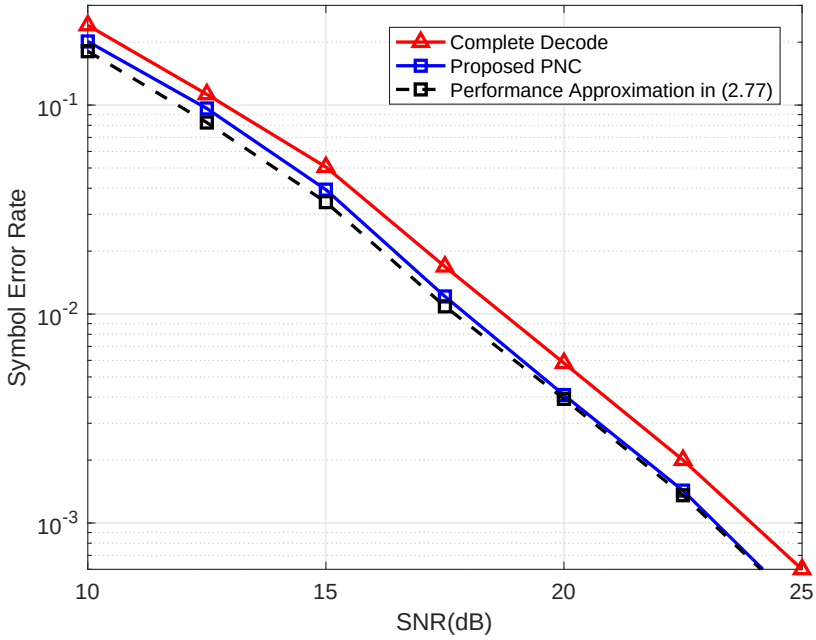


Figure 2.9. Error-rate performance of the proposed PNC scheme in a Rayleigh fading Y-channel, 9-QAM.

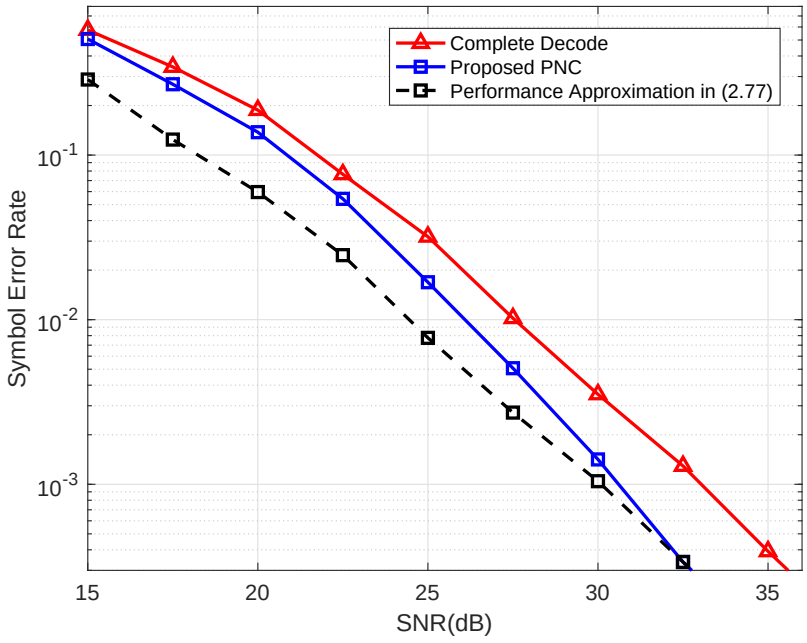


Figure 2.10. Error-rate performance of the proposed PNC scheme in a Rayleigh fading Y-channel where the uplink phase has one time-slot, 49-QAM.

$T = 1$

The three users transmit their signals simultaneously to the relay. The received signal at the relay is

$$y = \sqrt{E_s}h_A x_A + \sqrt{E_s}h_B x_B + \sqrt{E_s}h_C x_C + n \tag{2.78}$$

where $n \sim \mathcal{N}(0, \sigma_z^2)$. We see that the solution to maximize the MSD is similar to that in Theorem 2.3.

$T = 2$

Here, we propose two options for the case where two time-slots are available in the uplink phase.

Option 1: Separate Transmission Since we consider a flat block fading channel with a single antenna at each node, there is no diversity we can explore by transmitting in different time-slots. By introducing more time-slots, a basic thought is to minimize the interference between all users' transmitted signals.

Then, we have an intuitive uplink strategy as follows. Only one user transmits in the first time-slot and the rest two users transmit simultaneously in the second time-slot. For example, the received signals at the relay in each time-slot are

$$y_1 = h_A x_A + n_1,$$

and

$$y_2 = h_B x_B + h_C x_C + n_2$$

respectively. In this case, the d_{MSD} can be maximized by properly clustering the superimposed symbols of user B's and user C's messages.

First, the classic maximum-likelihood single-user detection can be performed based on the received signal in the first time-slot y_1. Then, the linear NC method for TWRCs [43] can be performed based on the received signal in the second time-slot y_2. This method has been shown to maximize the d_{MSD} between the superimposed

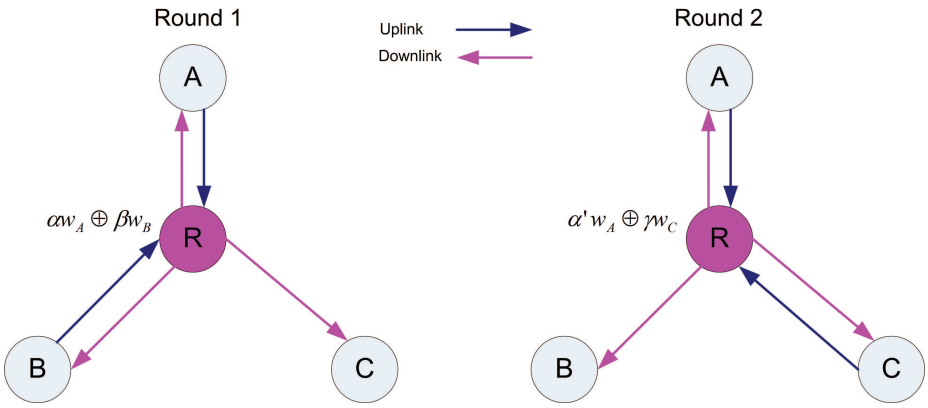


Figure 2.11. The proposed TDD-TWRC scheme for the Y-channel.

symbols of two users' messages. Denote the selected NC coefficient for user B and user C by β and γ respectively, $\beta, \gamma \in GF(q)$, $\beta \neq 0$, $\gamma \neq 0$. Then, we have the NC generator vectors as follows.

$$\mathbf{v}_A = \begin{bmatrix} 1 & 0 \end{bmatrix}^T, \mathbf{v}_C = \begin{bmatrix} 0 & \beta \end{bmatrix}^T, \mathbf{v}_C = \begin{bmatrix} 0 & \gamma \end{bmatrix}^T. \tag{2.79}$$

It can be easily verified that they satisfy the Full-Rank Condition.

Remark 2.8. *Option 1 simplifies the PNC problem for three users to that for two users, and utilizes existing PNC work for TWRCs. Its computation complexity is in the order of* q^4.

Option 2: TDD-TWRC We next consider a time-division duplex (TDD) TWRC scheme, where only two users transmit signals at each time-slot in the uplink phase as shown in Fig. 2.11.

Considering the power constrain for each user ($E(\|\mathbf{x}_l^2\|) = 1$), the received signals at the relay are

$$y_1 = \frac{\sqrt{2}}{2} h_A x_A + h_B x_B + n_1, \tag{2.80}$$

$$y_2 = \frac{\sqrt{2}}{2} h_A x_A + h_C x_C + n_2. \tag{2.81}$$

We can apply two types of relay operations for Option 2 as follows.

- Separate Processing

Since the received signal at each time-slot is a superposition of two user's signals, we can apply the networking method (for example, [43]) for TWRCs respectively for each time-slot.

Denote the optimized NC coefficients by α and β, α' and γ respectively, α, $\beta, \alpha', \gamma \in GF(q)$, $\alpha \neq 0$, $\beta \neq 0$, $\alpha' \neq 0$, $\gamma \neq 0$. The NC generator vectors are

$$\mathbf{v}_A = \begin{bmatrix} \alpha & \alpha' \end{bmatrix}^T, \mathbf{v}_B = \begin{bmatrix} \beta & 0 \end{bmatrix}^T, \mathbf{v}_C = \begin{bmatrix} 0 & \gamma \end{bmatrix}^T, \tag{2.82}$$

and they also satisfy the Full-Rank Condition.

Remark 2.9. *Similar to Option 1, the separate processing for Option 2 also simplifies the PNC problem for three users to that for two users. Its computation complexity is also in the order of q^4. However, the computation expense is at least twice that of Option 1, as the linear NC method for TWRCs have to be performed twice for Option 2.*

- Improvement using Joint Processing

We can also jointly process the received signal in two time-slots. Similar to Theorem 2.3, we first find the DV vector that corresponds to the minimum distance of SI symbols, i.e.,

$$\mathbf{\Delta} = \underset{|\delta_A|\neq 0, |\delta_B|\neq 0, |\delta_C|\neq 0}{\arg\min} \left\| \begin{bmatrix} \frac{\sqrt{2}}{2}h_A & h_B & 0 \\ \frac{\sqrt{2}}{2}h_A & 0 & h_C \end{bmatrix} \begin{bmatrix} \delta_A \\ \delta_B \\ \delta_C \end{bmatrix} \right\|. \tag{2.83}$$

Then, we determine the NC coefficients by solving (4.41).

Remark 2.10. *The joint processing for Option 2 utilizes the benefits of joint processing and Theorem 2.3. However, its computation complexity is in the order of q^6.*

Numerical Results

Here, we present numerical results for the error-rate performance of the considered two situations. The results are obtained by averaging over more than 1,000,000 channel realizations, where the fading channel coefficients follow i.i.d. Rayleigh distribution.

Fig. 2.12 shows the error-rate performance when one time-slot is allocated to the uplink phase and $q = 3$ (9-QAM for each user). For comparison purpose, we also show the performance of a complete decoding scheme, where the relay completely decodes all users' messages. We also include the numerical results of the lower bound, as discussed in Section 2.7.5. Moreover, the derived asymptotic error probability of the proposed scheme in (2.77) is also shown in the figures. We observe that our proposed scheme exhibits a 3 dB improvement over complete decoding scheme and a non-optimized at the error rate of 10^{-3}. In addition, we observe that at a sufficiently high SNR, the proposed linear PNC scheme achieves the lower bound.

Fig. 2.13 shows the error-rate performance of Option 1 and Option 2, where two time-slots are allocated to the uplink phase and $q = 3$ (9-QAM for each user). We can see from the numerical results that for Option 2, joint processing shows 1.5 dB improvement over separate processing. This agrees with our expectation that joint processing should behave better. On the other hand, we also see that the performance of Option 1 is almost as good as the joint processing for Option 2. Considering that the computation complexity for Option 1 is in the order of q^4 while that for Option 2 is in the order of q^6, the separate transmission in Option 1 is a more practical choice. This reminds us that TDD-TWRC is not always the better choice for Y-channel scheme without CSIT if more time-slots in the uplink are considered.

2.8 Conclusions

To summarize, this chapter first proposed a new linear vector physical-layer network coding scheme for spatial MIMO TWRC where CSI was not available at the transmit-

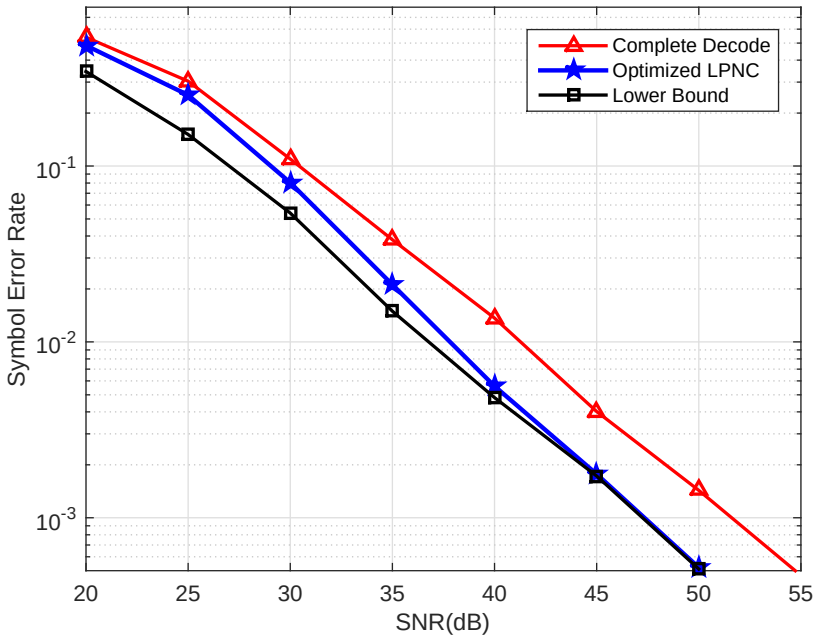


Figure 2.12. Error-rate performance of the proposed SISO PNC scheme in a Rayleigh fading Y-channel where the uplink phase has one time-slot, 9-QAM.

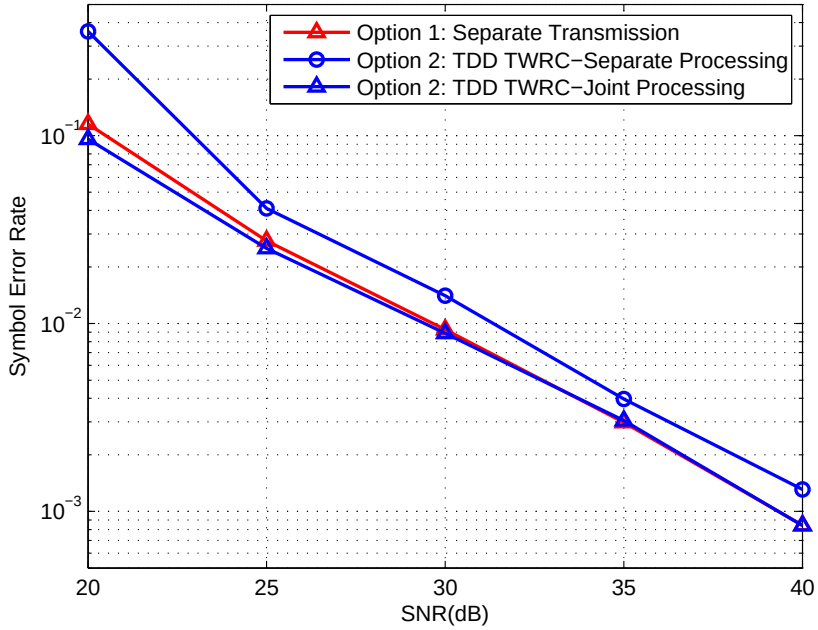


Figure 2.13. Error-rate performance of the proposed SISO PNC scheme in a Rayleigh fading Y-channel where the uplink phase has two time-slots, 9-QAM.

ters. We presented an explicit solution for the NC generator matrix that minimized the error probability at a high SNR, as well as an efficient algorithm to find the optimized solution. We derived a new closed-form expression on the average error probability of the proposed scheme over a Rayleigh fading MIMO TWRC. The derived result shows that the proposed scheme achieves the optimal error rate performance at a sufficiently high SNR. Numerical results match well with the analytical result and shows that the proposed scheme significantly outperforms existing schemes.

Then, we proposed a linear PNC scheme for a Y-channel without CSIT. We presented an explicit solution for the NC generator matrices that minimize the NC error probability at a high SNR. We also analyzed the NC error performance and provided an error-probability approximation at high SNRs. Numerical results matched well with our analytical approximation and showed that our proposed method outperformed existing schemes at high SNRs. We also presented discussions for a single-antenna relay to provide further understanding for the Y-channel model.

2.9 Appendix

2.9.1 Proof of (2.58)

Proof. We first derive an expression of the average pair-wise error probability of an $M \times N$ MIMO spatial multiplexing scheme. Rewrite the spatial multiplexing MIMO model $\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{z}$ into a MISO system with block-length N. Let $\mathbf{X}$ and $\mathbf{h}$ be

$$\mathbf{X} = \underbrace{\begin{bmatrix} \mathbf{x} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \mathbf{x} \end{bmatrix}}_{N} \tag{2.84}$$

and

$$\mathbf{h} = \left[\mathbf{H}\left(1,:\right), \cdots, \mathbf{H}\left(N,:\right)\right]. \tag{2.85}$$

Then, the received signal can be rewritten as

$$\mathbf{y} = \mathbf{h}\mathbf{X} + \mathbf{z}. \tag{2.86}$$

The average pair-wise error probability of confusing $\mathbf{X}$ with $\mathbf{X}'$ is

$$\begin{aligned} Pe_p(\mathbf{x} \to \mathbf{x}') &= E\left[Q\left(\frac{\|\mathbf{H}(\mathbf{x}-\mathbf{x}')\|}{\sqrt{2N_0}}\right)\right] \\ &= E\left[Q\left(\frac{\|\mathbf{h}(\mathbf{X}-\mathbf{X}')\|}{\sqrt{2N_0}}\right)\right] \\ &= E\left[Q\left(\sqrt{\frac{\rho}{2}\cdot\mathbf{h}(\mathbf{X}-\mathbf{X}')(\mathbf{X}-\mathbf{X}')^*\mathbf{h}^*}\right)\right], \end{aligned} \tag{2.87}$$

where $\mathbf{x}$ is the transmitted codeword, $\mathbf{x}'$ is the erroneous one. The matrix $(\mathbf{X}-\mathbf{X}')(\mathbf{X}-\mathbf{X}')^*$ is Hermitian. So, it can be written as

$$(\mathbf{X}-\mathbf{X}')(\mathbf{X}-\mathbf{X}')^* = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^*, \tag{2.88}$$

where $\mathbf{U}$ is a unitary matrix and $\mathbf{\Lambda}$ is a diagonal matrix. Since $(\mathbf{X}-\mathbf{X}')(\mathbf{X}-\mathbf{X}')^*$ is of rank N, the number of non-zero eigenvalues for $\mathbf{\Lambda}$ should be N, i.e. $\mathbf{\Lambda} = diag\left\{\lambda_1^2, \cdots, \lambda_N^2, 0, \cdots, 0\right\}$. By defining $\tilde{\mathbf{h}} = \mathbf{h}\mathbf{U}$, the pair-wise error probability can be represented as

$$\begin{aligned} Pe_p(\mathbf{x} \to \mathbf{x}') &= E\left[Q\left(\sqrt{\frac{\rho}{2}\cdot\tilde{\mathbf{h}}\mathbf{\Lambda}\tilde{\mathbf{h}}^*}\right)\right] \\ &= E\left[Q\left(\sqrt{\frac{\rho}{2}\cdot\sum_{l=1}^{N}\left\|\tilde{h}_l\right\|^2\lambda_l^2}\right)\right], \end{aligned} \tag{2.89}$$

where $\tilde{\mathbf{h}}$ has the same distribution as $\mathbf{h}$ since $\mathbf{U}$ is a unitary matrix. From literature [128], we know a tight approximation of Eric-function: $\operatorname{erf} c(x) \approx \frac{1}{6}e^{-x^2} + \frac{1}{2}e^{-4x^2/3}$. So (2.89) is upper bounded by

$$\begin{aligned} Pe_p(\mathbf{x} \to \mathbf{x}') &\le E\left[\frac{1}{12}e^{-\frac{\rho}{4}\cdot\sum_{l=1}^{N}\left\|\tilde{h}_l\right\|^2\lambda_l^2} + \frac{1}{4}e^{-\frac{\rho}{3}\cdot\sum_{l=1}^{N}\left\|\tilde{h}_l\right\|^2\lambda_l^2}\right] \\ &= E\left[\frac{1}{12}\prod_{l=1}^{N}e^{-\frac{\rho}{4}\cdot\left\|\tilde{h}_l\right\|^2\lambda_l^2} + \frac{1}{4}\prod_{l=1}^{N}e^{-\frac{\rho}{3}\cdot\left\|\tilde{h}_l\right\|^2\lambda_l^2}\right]. \end{aligned} \tag{2.90}$$

Using the fact that for a unit mean exponential random variable X, $E\left[e^{sX}\right] = 1/(1-s)$ for $s < 1$, we derive that for Rayleigh fading channels,

$$\begin{aligned} Pe_p(\mathbf{x} \to \mathbf{x}') &= \frac{1}{12}\prod_{l=1}^{N}\frac{1}{1+\frac{\rho}{4}\lambda_l^2} + \frac{1}{4}\prod_{l=1}^{N}\frac{1}{1+\frac{\rho}{3}\lambda_l^2} \\ &\leq \frac{1}{12}\frac{4^N}{\rho^N\prod_{l=1}^{N}\lambda_l^2} + \frac{1}{4}\frac{3^N}{\rho^N\prod_{l=1}^{N}\lambda_l^2} \\ &= \frac{4^N + 3^{N+1}}{12}\cdot\frac{1}{\rho^N}\cdot\frac{1}{(d)^N} \end{aligned} \tag{2.91}$$

where $d = \|(\mathbf{x} - \mathbf{x}')\|^2$. Note that $Pe_p(\mathbf{x} \to \mathbf{x}')$ can be written as $Pe_p(d)$.

For the codeword $\mathbf{x}$, the average error probability is

$$Pe(\mathbf{x}) = \sum_{\mathbf{x}', \mathbf{x}' \neq \mathbf{x}} Pe_p(\mathbf{x} \to \mathbf{x}'). \tag{2.92}$$

Then, the union bound of the average error probability for a MIMO spatial multiplexing scheme is

$$Pe^{\mathrm{Si}ngleUser} \leq \sum_{\mathbf{x}} \Pr(\mathbf{x}) Pe(\mathbf{x}) = \frac{1}{q^{2M}} \sum_{\mathbf{x}} \sum_{\mathbf{x}', \mathbf{x}' \neq \mathbf{x}} Pe_p(\mathbf{x} \to \mathbf{x}'). \tag{2.93}$$

This means the average error probability at the receiver can be categorized using the transmitted message's constellation. Denote the minimum distance in the transmitted message's constellation by d_1, the second minimum distance by d_2, $\cdots$. And denote the number of neighbors with distance d_1 by η_1, the number of neighbors with distance $d_2$by η_2, $\cdots$. Then, (2.93) can be rewritten as

$$Pe^{\mathrm{Si}ngleUser} \leq \frac{1}{q^{2M}}(\eta_1 Pe_p(d_1) + \eta_2 Pe_p(d_2) + \cdots). \tag{2.94}$$

For high SNR situation, the average error probability is dominated by the first serval terms, corresponding to the first serval minimum distance. We notice that preserving the first two terms is sufficiently accurate for most MIMO cases where $M > 1$ and $N > 1$.

Then, we obtain that for an M by N MIMO spatial multiplexing scheme using q-order QAM, the average error probability

$$Pe^{\mathrm{Si}ngleUser} \lessapprox \frac{(\eta_1+\eta_2)(1+2^{-\frac{N}{2}})}{q^{2M}} \cdot \frac{4^N+3^{N+1}}{12}\left(\frac{\gamma^2}{\rho}\right)^N \tag{2.95}$$

where $\eta_1$and η_2 are multiplicities w.r.t. the minimum and second minimum distance events respectively, and are given in (2.41) and (2.42).

Finally, we achieve the interference-free lower bound for MIMO TWRCs. For this interference-free lower bound, the two users are assumed to transmit their signals to the relay without each other's interference. Therefore, the interference-free lower bound is given by

$$Pe^{LB} = 2Pe^{SingleUser}. \tag{2.96}$$

This completes the proof. □

2.9.2 Proof of Proposition 2.1

Proof. The probability of randomly generated $[\mathbf{H}_A, \mathbf{H}_B]$ being linearly independent is 1. So, $[\mathbf{H}_A, \mathbf{H}_B]$ can be regarded as spanning an N-dimensional subspace in a $2M$-dimensional vector space. Therefore, if $d(\mathbf{w}_s, \mathbf{w}'_s) = \|[\mathbf{H}_A, \mathbf{H}_B]\,\boldsymbol{\delta}\|^2 = 0$ happens, the associated DV $\boldsymbol{\delta}$ must lie in the $(2M-N)$-dimensional subspace that is orthogonal to the subspace spanned by $[\mathbf{H}_A, \mathbf{H}_B]$. Recalling that the DVs belong to an integer set $\{1-q, \cdots, 0, \cdots, q-1\}^{2M}$ and can not reach any real-valued vectors, the number of linearly independent DVs can only be equal to or be smaller than $2M-N$. □

2.9.3 Proof of Proposition 2.2

Proof. We first relax the integer constraint that the columns of $\boldsymbol{\Delta}$ belong to $\{1-q, \cdots, 0, \cdots, q-1\}^{2M}$. Then, it is clear that $\begin{bmatrix} \boldsymbol{\Delta}_A \\ \boldsymbol{\Delta}_B \end{bmatrix}$ are in a subspace spanned by

$\begin{bmatrix} \mathbf{H}_A^{-1} \\ -\mathbf{H}_B^{-1} \end{bmatrix}$, i.e.,

$$\begin{bmatrix} \boldsymbol{\Delta}_A \\ \boldsymbol{\Delta}_B \end{bmatrix} = \begin{bmatrix} \mathbf{H}_A^{-1} \\ -\mathbf{H}_B^{-1} \end{bmatrix} \Theta. \tag{2.97}$$

where Θ is an $N \times L$ full-rank matrix. Since $\mathbf{H}_A^{-1}$ and $-\mathbf{H}_B^{-1}$ are both full rank, $\boldsymbol{\Delta}_A, \boldsymbol{\Delta}_B$ are both full-rank.

We next consider the integer constraint. Suppose $\boldsymbol{\Delta}_A$ or $\boldsymbol{\Delta}_B$ is rank-deficient. Then, it can be shown that $\mathbf{H}_A$ or $\mathbf{H}_B$ must be rank-deficient. However, this leads to absurdity as $\mathbf{H}_A$ and $\mathbf{H}_B$ are full rank a priori. Therefore, $\boldsymbol{\Delta}_A$ and $\boldsymbol{\Delta}_B$ must be both full rank. This finishes the proof. □

2.9.4 Proof of Proposition 2.3

Proof. In Theorem 2.1, the NC generator matrix $\mathbf{G}$ is chosen by $\mathbf{G}^T \otimes \mathbf{P} = \mathbf{0}$ where $\mathbf{P}$ consists of the first M message DVs such that the obtained $\mathbf{P}$ is block-wise full rank. According to Proposition 2.9.2, $\mathbf{P}$ contains up to $L \leq (2M - N)$ DVs whose distances are of order $K\rho^{-\epsilon}$, $\epsilon \geq 1$. According to Proposition 2.9.3, the matrix that consists of these DVs are block-wise full rank. Note that the resultant $\mathbf{P}$ is block-wise full rank as well. As such, it is for sure that there exists a block-wise full rank $\mathbf{G}$ such that $\mathbf{G}^T \otimes \mathbf{P} = \mathbf{0}$, and thus the multiplicities of all the distances of order $K\rho^{-\epsilon}$, $\epsilon \geq 1$, will be zero. In other words, any distance that is of order $K\rho^{-\epsilon}$, $\epsilon \geq 1$, will be removed from the set-distance spectrum. This finishes the proof. □

Chapter 3

An Achievable Throughput Scaling Law of Wireless Device-to-device Caching Networks with Distributed MIMO and Hierarchical Cooperations

3.1 Introduction

In this chapter, we propose a new caching scheme for a random wireless device-to-device (D2D) network of n nodes with local caches, where each node intends to download les from a prexed library via D2D links. Our proposed caching delivery includes two stages, employing distributed MIMO and hierarchical cooperations respectively. The distributed MIMO is applied to the rst stage between source nodes and neighbours of the destination node. The induced multiplexing gain and diversity gain increase the number of simultaneous transmissions, improving the throughput of the network. The hierarchical cooperations are applied to the second stage to fa-

cilitate the transmissions between the destination node and its neighbours. The two stages together exploit spatial degrees of freedom as well as spatial reuse. We develop an uncoded random caching placement strategy to serve this cooperative caching delivery. Analytical results show that the average aggregate throughput of the network scales almost linearly with n, with a vanishing outage probability. Furthermore, we derive an explicit expression of the optimal throughput as a function of system parameters such as pathloss factor under a target outage probability. Analytical and numerical results demonstrate that our proposed scheme outperforms existing ones when the local cache size is limited

We highlight that the main difference between our proposed scheme and existing ones is the use of the spatial network resources. In fact, no matter one-hop [95], multi-hop [99] or hierarchical cooperations [100], the nice scaling with the size of caches is mainly achieved by exploiting the spatial reuse in the network. The growth of the cache size increases the chance of spatial reuse, leading to a growing network throughput. However, since our work aims at the small cache case, we do not focus on the spatial reuse in designing our caching scheme. Instead, we try to exploit the spatial degree of freedom and increase the cache-hit probability by allowing one node to access more other nodes, to better facilitate the design for the small cache situation.

This chapter is organized as follows. In Section 3.2, we introduce the system model and formulate the problem. We state the main results of this chapter in Section 3.3. Section 3.4 presents an achievable scheme consisting of both the caching placement strategy and the delivery policy. For the proposed scheme, we then derive the throughput and the outage probability, obtain the scaling law and optimize the throughput in Section 3.5. Section 3.3 provides some numerical results.

3.2 System Model and Problem Formulation

Consider a network with n wireless nodes, which are uniformly and independently distributed in a unit square, as depicted in Fig. 2.8. Each node u, $u \in \mathcal{U} = \{1, \cdots n\}$,

has an average transmit power of P watts and a local cache size of M_c files. The size of caches in this thesis is represented by the number of standard files (F bits per file). Without the presence of a server or a base station, D2D communications between nodes are considered in this network. A library with m files is denoted by $\mathcal{F}$, $\mathcal{F} = \{f_1, \cdots, f_m\}$. Each node intends to download its requested file from the library $\mathcal{F}$ through this cached D2D network.

A caching scheme is performed in two phases: caching placement phase and caching delivery phase. The library $\mathcal{F}$ is generated once and kept fixed during the two phases.The cached messages are stored at each node in the caching placement phase and remain fixed during the delivery phase. Each node requests its own intended file from the library and gets served through wireless transmissions in the delivery phase. Multiple requests made by one node can be handled consecutively. In this chapter, we will work on maximizing the network throughput by jointly designing both phases. To this end, we introduce some useful definitions as follows.

Definition 3.1 (**Caching placement strategy**). *A cache placement strategy Π_c is a rule to assign files from the library $\mathcal{F}$ to a node's local cache in the caching placement phase. Let $\mathcal{M}_u$ represent the local storage of node u. Then, a particular cache placement G_u at node u can be viewed as a mapping from the library $\mathcal{F}$ to the local memory$\mathcal{M}_u$. One realization of caching placement at all nodes is denoted by $G \triangleq \{G_u, u \in \mathcal{U}\}$. Note that the caching placement is done without a priori knowledge of the nodes' actual requests.*

Definition 3.2 (**Users' request**). *At each request time, each node makes a request to a file f_i, $f_i \in \mathcal{F}$, randomly and independently. Denote the index of node u's requested file by i_u. Then, one realization of all nodes' requests, denoted by R, can be represented by*

$$R = \{i_1, \cdots, i_n\}. \tag{3.1}$$

In this thesis, we consider that for any node in the network, its requested file index i_u is uniformly distributed over $\{1, \cdots, m\}$. This stands as the worst case scenario

where users' requests are quite scattered rather than gathered. This assumption is also adopted in [83] and fits the spirit in [93–95] considering a "heavy tail" Zipf request distribution.

Definition 3.3 (**Transmission policy**). *A transmission policy Π_t is a rule of designing and scheduling the D2D transmissions in the delivery phase. Generally, it includes two parts: one is the transmission protocol $\mathcal{P}$ describing the feasible D2D links in the network, such that these D2D links can provide reliable communications, considering the physical constraints such as power and interference; and the other is the transmission scheduling $\mathcal{S}$ describing the activated D2D links at one time. Denote by y a set of simultaneous transmission links.*

Definition 3.4 (**Average aggregate throughput**). *For a given caching placement G, node requests R and a set of transmission links y, define $t_u(G,R,y)$ as the number of useful received information bits per time-slot by node u during one delivery phase. Adding $t_u(G,R,y)$ from all nodes together, we have the aggregate throughput of the network as*

$$\widetilde{T}_n(G,R,y) = \sum_{u\in\mathcal{U}} t_u(G,R,y). \tag{3.2}$$

Let T_n denote the average aggregate throughput*, and*

$$\begin{aligned} T_n &= E_R\left[E_S\left[\widetilde{T}_u(G,R,y)\right]\right] \\ &= \sum_R\sum_y \widetilde{T}_u(G,R,y)\Pr\{y|R,G\}\Pr\{R\}, \end{aligned} \tag{3.3}$$

where E_R means averaging over the randomness of nodes' actual requests and E_S means averaging over the randomness caused by scheduling during all delivery phases. We see that T_n is a function of the caching placement strategy Π_c and the transmission policy Π_t, i.e.,

$$T_n \sim f(\Pi_c, \Pi_t). \tag{3.4}$$

Definition 3.5 (**Outage probability**). *For a given caching placement G, given node requests R and a given set of transmission links y, if a node's request cannot be*

satisfied through the delivery phase, we say that this node is in outage. Denote the number of nodes in outage by $N_o(G,R,y)$ *in one delivery phase, i.e.,*

$$N_o(G,R,y) = \sum_u \mathbf{1}\left(t_u(G,R,y) = 0\right). \tag{3.5}$$

Define the *outage probability* of the network as

$$p_{out} = \frac{E_R\left[E_S\left[N_o(G,R,y)\right]\right]}{n}, \tag{3.6}$$

where E_R and E_S are described under (3.3). Similarly, p_{out} is a function of the caching placement strategy Π_c and the transmission policy Π_t.

In this thesis, we focus on the scaling of the average aggregate throughput T_n with an increasing number of n, while the power constraint of each node, the allocated bandwidth and the occupied spacial area of all nodes remain constant. In addition, we will assume that the size of the library m and the size of local cache M_c increase with n. Otherwise, if m and M_c remain constant when $n \to \infty$, a node can find its intended file within its neighbours with probability one regardless of any Π_c and Π_t. Thus, throughout the thesis, we assume that

$$m = n^{\alpha} \text{ and } M_c = n^{\beta}, \tag{3.7}$$

where $\alpha > 0$ and $\beta > 0$. We further assume that $0 < \alpha - \beta < 1$. This is because if $\alpha - \beta < 0$, each node is able to store the entire library in its local memory, which is not the focus of this thesis. On the other hand, if $\alpha - \beta > 1$, the total memory size of the network nM_c is smaller than the library size m. This will cause that the outage probability goes essentially large [93], making the system not useful for practical deployment.

The main goal of this chapter is to study the following two optimization problems. Let $T_n = \Theta\left(n^{\varphi}\right), \varphi \geq 0$. The first goal is, in the study of scaling laws, to jointly optimize the caching strategy Π_c and the transmission policy Π_t so that the exponent

φ is maximized while the outage probability is vanishing, i.e.,

$$\mathcal{P}_0 : \max_{\Pi_c, \Pi_t} \varphi$$
$$s.t. \lim_{n\to\infty} p_{out} = 0. \tag{3.8}$$

Beyond the scaling law, it is desirable to jointly optimize Π_c and Π_t so that the throughput T_n is maximized with a target outage probability ϵ when designing a wireless caching network. This can be represented by problem $\mathcal{P}_1$ as follows.

$$\mathcal{P}_1 : \max_{\Pi_c, \Pi_t} T_n$$
$$s.t.\ p_{out} \leq \epsilon \tag{3.9}$$

With the open and general definitions mentioned above, we see that optimizing Π_c and Π_t over all possible candidates has a prohibitive complexity. To address above problems, in this chapter, we focus on designing a transmission policy Π_t exhibiting a good scaling law of the average aggregate throughput T_n. A caching placement strategy Π_c is designed to serve the transmissions.

Note that our work does not pursue a better throughput scaling with the size of caches as in conventional caching studies. Instead, our main objective is to achieve a good throughput scaling with the number of users, especially when the cache size is small compared with the size of library.

3.3 Main Results

This section states the main results of this chapter. We first introduce an achievable scaling law when the number of nodes n is sufficiently large as follows.

Theorem 3.1. *For the considered caching wireless network with a library of m files and local caches of M_c files, if $nM_c > m$, as $n \to \infty$, the following scaling law is achievable with high probability:*

$$T_n = \Theta\left(n^{\frac{t}{t+1}}\right), \tag{3.10}$$

where t, $t \geq 1$, is an integer constant independent of n.

Proof. An achievable scheme is presented in Section 3.4 and the proof is given in Section 3.5. The choice of the parameter t will be discussed in Section 3.5. □

Remark 3.1. *From (3.10), with t increases, $\frac{t}{t+1} \to 1$. This indicates that with the help of caching, the average throughput of the network can increase almost linearly with the number of nodes n. Note that the scaling $\Theta\left(n^{\frac{t}{t+1}}\right)$ in this thesis is different from the scaling $\Theta\left(\frac{nM_c}{m}\right)$ in [93] [83] or $\Theta\left(n\sqrt{\frac{M_c}{m}}\right)$ in [98] [99]. Fig. 3.1 compares the throughput scaling laws of the caching network in [93] [83], [98] [99] and Theorem 3.1. Recall that the most interesting region is that $nM_c > m$, i.e., $0 < \alpha - \beta < 1$. We see that $1 \leq \frac{nM_c}{m} < n$, $\sqrt{n} \leq n\sqrt{\frac{M_c}{m}} < n$, and in our result $\sqrt{n} \leq n^{\frac{t}{t+1}} < n$ (since $t \geq 1$). This is shown in the figure where these three scalings exhibit different throughput ranges in the region of $0 < \alpha - \beta < 1$. It is also immediate to notice that the scalings in [93] [83] and [98] [99] are both functions of $\alpha - \beta$, the relative exponent of n between the library size m and the cache size M_c, however the scaling in Theorem 3.1 is irrelevant with $\alpha - \beta$. We will discuss the reason in Section 3.5.*

Furthermore, it can be observed that our scaling performs better under a cache-size limited situation (with $\alpha - \beta$ large). For example, consider a caching network with $\alpha = 1$ and $\beta = 0.2$. This setup corresponds to a disadvantageous caching case where there are many files in the library however the local cache at each node is relatively small. For example, when $n = 10^8$, there are 10^8 files in the library while each node only has a cache size of 40 files. The aggregate average throughput in [93] [83] and [98] [99] will be $\Theta\left(n^{0.2}\right)$ and $\Theta\left(n^{0.6}\right)$ respectively, while our scheme achieves $\Theta\left(n^{0.75}\right)$ when a typical value $t = 3$ is applied according to (3.10). It is also worth mentioning that since our scheme is specifically designed for the small cache case, it does not perform as good as current schemes when $\alpha - \beta < \frac{1}{t+1}$ or $\frac{2}{t+1}$.

In addition to the scaling law, for our proposed achievable scheme, we also derive an explicit expression of the throughput T_n and then solve the optimization problem

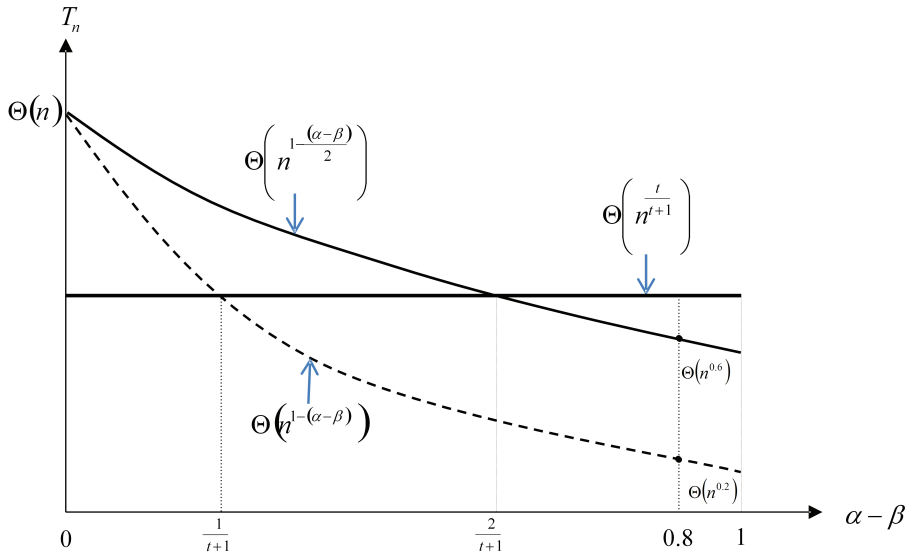


Figure 3.1. Achievable throughput scaling laws of $\Theta\left(\frac{nM_c}{m}\right)$ in [93] [83], $\Theta\left(n\sqrt{\frac{M_c}{m}}\right)$ in [98] [99] and $\Theta\left(n^{\frac{t}{t+1}}\right)$ in (3.10) respectively. Our proposed scheme outperforms current schemes when $\alpha-\beta > \frac{1}{t+1}$ or $\frac{2}{t+1}$, while is inferior when $\alpha-\beta < \frac{1}{t+1}$ or $\frac{2}{t+1}$.

of maximizing T_n when achieving a given outage probability ϵ. The result is shown in the following theorem.

Theorem 3.2. *For the considered caching wireless network with n nodes, a library of m files and local caches of M_c files where $nM_c > m$, to achieve a target outage probability ϵ, the following average aggregate throughput is achievable when n is sufficiently large:*

$$T_n = \begin{cases} \frac{1}{t+1}\left(\frac{n}{C}\right)^{\frac{t}{t+1}}, & \text{if } B_U \geq \left(\frac{n}{C}\right)^{\frac{t}{t+1}} \\ \frac{nB_U}{n+tCB_U^{\frac{t+1}{t}}}, & \text{if } B_U < \left(\frac{n}{C}\right)^{\frac{t}{t+1}}, \end{cases} \tag{3.11}$$

where C is a system-dependent constant and B_U is an upper bound for the number of packets of a file, depending on the target outage probability ϵ, the library size m and the cache size M_c.

Proof. An achievable scheme is presented in Section 3.4 and the proof is given in Section 3.5. Details of parameters C and B_U are defined later in (3.49) and (3.34) respectively. □

Remark 3.2. *From Theorem 3.2, we see that the optimal throughput can be in two regimes, depending on system parameters such as the target outage probability ϵ, the library size m and the cache size M_c. For example, if M_c and ϵ are large, the optimal throughput T_n will probably fall in the first line of (3.11). This is different from the pure scaling in Theorem 3.1, which is irrelevant with these parameters.*

In the rest of this chapter, we will present an achievable scheme with the scaling law in (3.10) in Section 3.4, and then prove these two theorems In Section 3.5.

3.4 An Achievable Scheme

We note that the throughput of a dense network is fundamentally limited by the interference between simultaneous transmissions. The following scheme is proposed to deal with the interference by integrating distributed MIMO, hierarchical cooperations and caching.

3.4.1 Caching Placement Phase

In this chapter, we propose an uncoded, distributed and randomized placement strategy Π_c.

Before the cache placement, each file in the library is partitioned into B packets of equal size. Here, we assume that the size of each file is sufficiently large to support the required number of packet B as determined later.[1] More specifically, we partition a file $f_i \in \mathcal{F}$ into packets of equal size $b_{i,j}$, $j \in \{1, \cdots, B\}$, i.e.,

$$f_i = \bigcup_{j \in \{1, \cdots, B\}} b_{i,j}. \tag{3.12}$$

Thus, there are mB packets in the library. Note that B is an important parameter to be determined.

[1] Although the file size F, when it is small, can limit the performance of caching networks, it will not become a bounding factor in our work. This is because for our proposed scheme, B is always much smaller than the file size in practical scenarios.

During the cache placement, each node fills in its local cache in an i.i.d. manner according to a same placement strategy Π_c. Each node randomly chooses $M_C B$ packets from the mB packets of the library and stores them in its own cache as depicted in Fig. 3.2. Denote the lth packet in node u's cache $\mathcal{M}_u$ by c_u^l and we have

$$\mathcal{M}_u = \underset{l\in[1,M_cB]}{\cup} c_u^l. \tag{3.13}$$

In order to enable the multiplexing gain in a MIMO transmission in the delivery phase, it is required that different packets of a file are simultaneously transmitted from different nodes. To achieve this, our caching placement must guarantee that each node caches packets from different files. We describe this process in details as follows.

First, each node u decides which files to choose from, by randomly generating M_cB file indexes $\left\{i_u^l\right\}$, $l \in \{1, \cdots, M_cB\}$, $i_u^l \in \{1, \cdots, m\}$. We use the following rule to determine the chosen file indexes for node u.

1. If the number of packets stored in each node M_cB is less than the number of files m (i.e., $M_cB \leq m$), the packets for each node cache should be chosen from different files. That is, $\forall l_1, l_2 \in [1, M_cB]$, $i_u^{l_1} \neq i_u^{l_2}$. This can be recognized as a similar process of blindly drawing M_cB balls from a box of m different balls at a time.

2. If $M_cB > m$, the file indexes are generated as follows. First, all file indexes $1, \cdots, m$ are chosen for $W = \left\lfloor \frac{M_cB}{m} \right\rfloor$ times. That is, for $w \in [1, W]$, $\forall l \in [(w-1)\,m + 1, wm]$, $i_u^l = l - (w-1)\,m$. Then, for the rest of $(M_cB - Wm)$ packets, their file indexes must be different. That is, $\forall l_1, l_2 \in [Wm + 1, M_cB]$, $i_u^{l_1} \neq i_u^{l_2}$.

After the file indexes are determined, node u randomly picks up one of the packets from file $f_{i_u^l}$ with probability $\frac{1}{B}$ and fills it in the local cache $\mathcal{M}_u$.

Denote by j_u^l the lth packet chosen by node u in the i_u^l file, we have that

$$c_u^l = b_{i_u^l, j_u^l}, \text{ and } \mathcal{M}_u = \underset{l\in[1,M_cB]}{\cup} b_{i_u^l, j_u^l}. \tag{3.14}$$

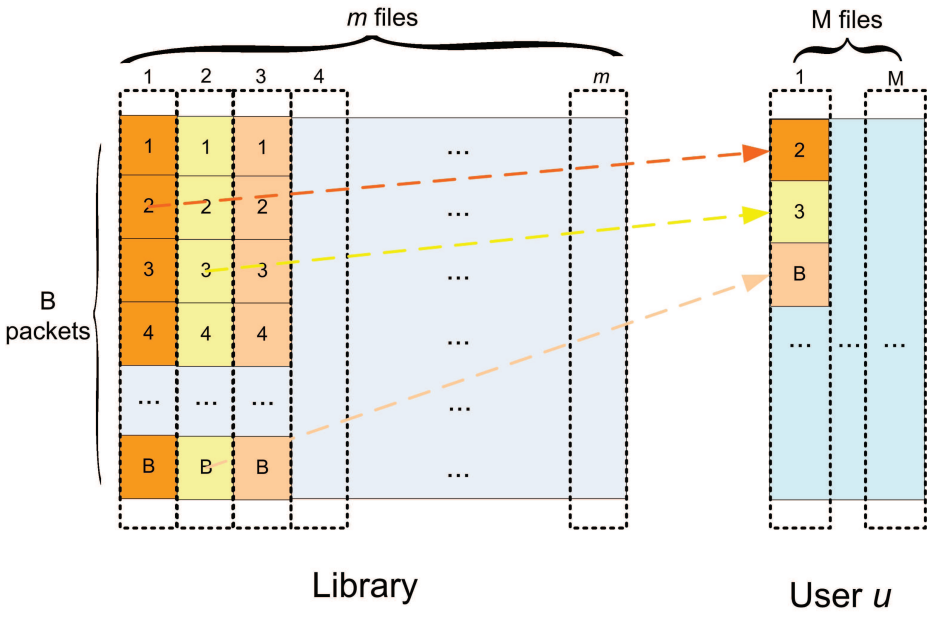


Figure 3.2. The proposed caching placement strategy.

These two steps ensure that the packets stored at one node come from different files to the most extent, which facilitates the delivery phase to be discussed next.

3.4.2 Caching Delivery Phase

Preliminaries

Channel model: We use the line-of-sight *physical model* in [88–90]. Denote the channel coefficient from node k to node u by h_{uk}, then

$$h_{uk} = \sqrt{G_A}\,(r_{uk})^{-\frac{\gamma}{2}} \exp\left(j\theta_{uk}\right), \tag{3.15}$$

where r_{uk} is the distance between node k to node u, θ_{uk} is the random phase, uniformly distributed in $[0, 2\pi)$, G_A is the antenna gain and γ is the pathloss exponent of the environment. We have $\gamma \geq 2$ for the far-field assumption.

User clustering: Divide the entire network into square cells of area A_c as shown in Fig. 3.3a. We call each square cell a *cluster* in this chapter. Let A_c be a function

of n and recall that the entire network is a unit square, then we have

$$A_c = n^{-\eta},\ \eta > 0. \tag{3.16}$$

We see that η is an important parameter determining the cluster size, which will be chosen later.

The number of nodes in one cluster is denoted by N_c. From literatures [85] [98], we know that

$$N_c = \Theta\left(n^{1-\eta}\right). \tag{3.17}$$

We denote by D the cluster that the destination node belongs to and by $\mathcal{N}$ all the neighboring clusters of D. We then design the caching delivery phase in two stages as follows.

Stage I: Distributed MIMO

In this stage, we design the transmission policy Π_t utilizing the fact that each node caches different packets of different files from previous caching placement (We will show that this is achieved with high probability in Section 3.5.). We first consider the transmission policy for serving a single node, and then extend it to serving all nodes in the network.

Recall that node u requires one file f_{i_u} with B packets. In order to collect all B packets of file f_{i_u}, we need to determine B source nodes, that store the B packets in their caches respectively. Then, to serve node u, a virtual multiplexing MIMO transmission can be formed from these B source nodes to the destination cluster D. For any requested file f_{i_u} of node u, B source nodes, $u_1^s, \cdots, u_B^s \in \mathcal{U}$, are selected according to the following criteria[2].

[2]When the total number of packets M_cB at one node is larger than the total number of files m, there must be more than one packet that is already stored in the local cache. In this case, less than B packets are required to transmit through the network. However, this does not change the scaling of the network throughput as the number of nodes n goes to infinity. Therefore, in this thesis, we consider that each source node needs B packets and derive the throughput scaling.

1. Each of the source nodes has one different packet of the requested file f_{i_u} in its local cache, i.e.,

$$\exists l_1, \cdots, l_B \in [1, M_c B],\ c_{u_1^s}^{l_1} \cup c_{u_2^s}^{l_2} \cup \cdots \cup c_{u_B^s}^{l_B} = f_{i_u}. \tag{3.18}$$

2. All the source nodes cannot lie in the neighboring clusters of node u, i.e.,

$$u_1^s, \cdots, u_B^s \notin \mathcal{N}, \tag{3.19}$$

which will be justified in Appendix 3.8.1.

Note that if there are multiple candidates of source nodes, we can randomly pick up any B candidates that meet the criteria above. The criteria allow a destination node to choose source nodes from the entire network except for the neighbouring clusters $\mathcal{N}$.

In the following, we first focus on the case where all source nodes are not in the destination cluster D. We will then show that it is equivalent to the case where some of the source nodes are in cluster D in terms of achieving the same aggregate throughput.

After selecting the source nodes $u_1^s, \cdots, u_B^s$ as transmitters, we choose all the N_c nodes in the destination cluster D as receivers. Each packet of file f_{i_u} is simultaneously sent from the B transmitters to the N_c receivers in a multiplexing way as shown in Fig. 3.3b. An MIMO transmission will be formed as long as the observations at all receivers can be jointly processed. We will discuss how to realize this joint processing in Stage II. Note that to achieve multiplexing transmission, it is required that all the source nodes are synchronized for their transmissions.

So far, we have described the transmission policy Π_t for serving a single node. For serving all nodes in the network, the same Π_t is applied in a time-division manner.

Stage II: Hierarchical Cooperations

The goal of this stage is to collect and jointly process the MIMO observations from Stage I at the destination node u. To this end, we apply hierarchical operations

similar to that in [84] [92] as follows.

Within a destination cluster, each node quantizes the received signal and sends it to the destination node. The destination node then jointly processes the N_c copies of superimposed signals received from previous multiplexing transmissions nodes. Thus an MIMO transmission from the source nodes to the destination node is formed through the two-stage cooperations. Note that the quantization does not change the linear scaling (of the number of independent transmitting streams) of MIMO capacity, which is proven in [88].

We notice that each node in the cluster wants to send independent messages to all other nodes in the same cluster. This communication problem is referred to as a *network multiple access problem* in [84] [92]. In this problem, each of the n nodes in the network has $n-1$ independent information messages, one for other $n-1$ node (all messages have the same size in bits). For this problem, a hierarchical three-phase cooperative transmission was proposed employing a similar idea of clustering, distributed MIMO and quantize-and-forward [84]. A hierarchical scheme with an improved scheduling, referred to as Method 4 in [92], achieves the best scaling and transmission rate by far to our best knowledge. We adopt this design for the network multiple access problem in Stage II to achieve the scaling.

3.5 Performance of the Proposed Scheme

In this section, we first derive an explicit expression of the aggregate throughput for the proposed scheme. Then, we analyse the outage probability and derive an asymptotic expression of its upper bound. Using these results, we then prove Theorem 3.1 and Theorem 3.2.

3.5.1 Aggregate Throughput of the Proposed Scheme

We calculate the aggregate throughput for the proposed scheme as follows.

Required time-slots for Stage I

The distributed MIMO transmissions in Stage I can be explicitly described in the following. Let $\mathbf{x}$ be a $B \times 1$ vector denoting the signals transmitted from the source nodes, and let $\mathbf{y}$ be an $N_c \times 1$ vector denoting the signals received at the destination cluster D. Then, we have that

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{z}, \tag{3.20}$$

where $\mathbf{H}$ is the $N_c \times B$ channel matrix where the $\{u, k\}$-th element h_{uk} is defined in (3.15), and $\mathbf{z}$ is an i.i.d. Gaussian noise vector. In order to guarantee reliable transmissions from the source nodes to the destination cluster, it is required that the number of independent transmitting streams B should be no more than the number of receiving antennas N_c, i.e.,

$$B \leq N_c. \tag{3.21}$$

Let the geographical distance between transmitter k and destination cluster D be r_{Dk} as shown in Fig. 3.3b. We use the following transmit power control mechanism such that

$$\mathbf{E}\left[|x_k|^2\right] = \frac{P}{B}\left(r_{Dk}\right)^{\gamma}. \tag{3.22}$$

We see that the transmit power control mechanism is similar to that in [84] [88]. We then have the following lemma.

Lemma 3.1. *Under the channel model described in (3.15), applying the transmit power control in (3.22), a long-distance distributed MIMO transmission from the source nodes to the destination cluster D achieves an aggregate rate scaling linearly with the number of transmitters B.*

The proof is given in Appendix 3.8.1.

We see that the transmit power constraint at each node is satisfied in our design. This is because when serving one node, each node only needs to transmit $\frac{P}{B}\left(r_{Dk}\right)^{\gamma}$ watts and only a fraction of $\frac{B}{n}$ nodes are transmitting. So in order to serve all n nodes

in the network, on average, each node needs a total transmit power

$$\frac{P}{B}\left(r_{Dk}\right)^{\gamma}\cdot\frac{B}{n}\cdot n=P\left(r_{Dk}\right)^{\gamma}\leq P, \tag{3.23}$$

which satisfies the transmit power constraint.

Based on our designs of Stage I, we calculate the required time-slots for serving all nodes in Stage I as follows. We know that each packet has $\frac{F}{B}$ bits and all the B packets of a file are simultaneously transmitted in a multiplexing way. Using Lemma 3.1, for any requested file of one node, a reliable transmission from source nodes to the destination cluster will take $\frac{F}{B}$ time slots. Therefore, in total, $\frac{F}{B}n$ time slots are needed for serving all n nodes in Stage I.

Required time-slots for Stage II

In this stage, we use hierarchical cooperations within one cluster. Recalling the results in [92], the time to complete the network multi-access transmission for a network of n nodes with t hierarchical stages and Q-bit quantization is given by

$$t_n=tL^2Q^{\frac{t-1}{2}}n^{\frac{t+1}{t}}, \tag{3.24}$$

where

$$L=\left\lceil 2^{\frac{3+\frac{\gamma}{2}}{\gamma}}+1\right\rceil \tag{3.25}$$

is the reuse factor, meaning that each cluster has one transmission opportunity every L^2 time-slots. By substituting the number of nodes n by N_c in our problem, we obtain the time t_{N_c} for completing transmissions for a cluster of n nodes as

$$t_{N_c}=\frac{F}{B}tL^2Q^{\frac{t-1}{2}}N_c^{\frac{t+1}{t}}. \tag{3.26}$$

Here, t is an integer representing the number of hierarchical stages, and an optimal t for maximizing the throughput is determined by number of nodes N_c according to [92].

Here, we discuss the case where some of the source nodes are in the destination cluster D. If some packets of the requested file f_{i_u}, say N_D packets, can be found

within cluster D, in Stage I, the number of transmitted packets will be $(N_c - N_D)$. Thus, we can reduce the number of receivers to $(N_c - N_D)$ and the consumed time slots for MIMO transmissions in Stage I remain unchanged. In Stage II, the number of transmitted packets to node u will remain as $(N_c - 1)$. This is because node u will only need to receive MIMO observations from $(N_c - N_D - 1)$ nodes, in addition to receiving the N_D packets from its neighbours. It is convenient to arrange that the neighbours with N_D packets are not scheduled as receivers in MIMO transmissions. Thus, Stage II can still be seen as a network multiple access problem with N_c nodes and it will also occupy the same time-slots. As a result, we see that in terms of the aggregate throughput, the case where some packets can be found within cluster D is equivalent to the case where no source nodes are in cluster D.

Aggregate throughput

From previous designs for both stages, we know that nF bits in total are delivered to their destinations in $\frac{F}{B}n + t_{N_c}$ time slots. Thus, the aggregate throughput is

$$\begin{aligned} T_n &= \frac{nF}{\frac{F}{B}n + t_{N_c}} \\ &= \frac{nB}{n + tL^2 Q^{\frac{t-1}{2}} N_c^{\frac{t+1}{t}}}. \end{aligned} \tag{3.27}$$

It is clear that in order to maximize T_n, the number of nodes in one cell N_c should be as small as possible. Since $B \leq N_c$, the optimal N_c^* should be

$$N_c^* = B. \tag{3.28}$$

Remark 3.3. *From (3.28), we see that the optimal number of nodes in one cluster N_c and the number of packets of a file B are coupled. On one hand, if choosing a large cluster size (a large N_c), the number of packets of a file B should also be large. This is essentially attributed to our design in Stage I: with large clusters (a large N_c), the number of independent transmitting streams B should also be large in order to utilize the multiplexing gain to the most extent; On the other hand, with a large B, N_c must be large enough to support reliable multiplexing transmissions in Stage I.*

Then, with (3.28), we get that

$$T_n = \frac{nB}{n + tL^2 Q^{\frac{t-1}{2}} B^{\frac{t+1}{t}}}. \tag{3.29}$$

We see that the number of packets of a file B is a critical design parameter in both the caching placement and delivery phase. By optimizing over B, the caching strategy Π_c and the transmission policy Π_t are jointly optimized. Therefore, in our proposed scheme, the throughput T_n no longer needs to be optimized over Π_c and Π_t as in problem $\mathcal{P}_0$ and $\mathcal{P}_1$. Instead, we only need to optimize T_n over the number of packets B.

3.5.2 Outage Probability of the Proposed Scheme

In this section, we examine the outage probability of our proposed scheme and derive an asymptotic expression of the upper bound of the outage probability.

Theorem 3.3. *As $n \to \infty$, the outage probability in our scheme is upper bounded by*

$$p_{out} \leq 1 - e^{-mB \cdot e^{-\frac{\left(n-\left(L^2-1\right)N_c\right)M_c}{m}}}. \tag{3.30}$$

The proof is given in Appendix 3.8.2.

According to (3.30), We see that p_{out} increases with the number of nodes in one cluster N_c. Recall that $N_c \geq B$. Therefore, to achieve a small outage probability p_{out}, it is required that

$$N_c^* = B. \tag{3.31}$$

This agrees with the requirement for maximizing T_n in (3.28). Thus, we have

$$p_{out} \leq p_{out}^U \triangleq 1 - e^{-mB \cdot e^{-\frac{\left(n-\left(L^2-1\right)B\right)M}{m}}}. \tag{3.32}$$

where p_{out}^U is referred to as the outage probability upper bound.

Remark 3.4. *In order to achieve a target outage probability ϵ, it is suffice to let*

$$p_{out}^U \leq \epsilon. \tag{3.33}$$

Let $B = B_U$ denote the solution to the equation

$$p_{out}^{U} = \epsilon. \tag{3.34}$$

Here, B_U represents the maximum number of packets of a file, allowed by a given library size m, local cache size M_c, and the target outage probability ϵ. We see that B_U decreases with m, and increases with M_c and ϵ. Especially, increasing the local cache size M_c would allow an exponentially larger number of packets B. To achieve the outage probability ϵ, it is suffice to let

$$B \leq B_U. \tag{3.35}$$

3.5.3 Throughput Scaling Law of the Proposed Scheme

Based on (3.29), the scaling law of the throughput can be easily obtained as

$$T_n = \frac{nB}{n + tL^2 Q^{\frac{t-1}{2}} B^{\frac{t+1}{t}}} = \frac{nB}{n + \Theta\left(B^{\frac{t+1}{t}}\right)}. \tag{3.36}$$

By maximizing T_n over the number of packets B, we obtain

$$T_n^* = \Theta\left(n^{\frac{t}{t+1}}\right) \tag{3.37}$$

when

$$B^* = \Theta\left(n^{\frac{t}{t+1}}\right) \tag{3.38}$$

It is interesting to see that the optimal throughput scaling T_n^* and the corresponding B^* are on the same order, i.e.,

$$T_n^* = \Theta\left(B^*\right). \tag{3.39}$$

This is because the improvement of the throughput from $\Theta(1)$ to $\Theta\left(n^{\frac{t}{t+1}}\right)$ is mainly attributed to the multiplexing gain achieved by distributed MIMO. For instance, a smaller B means that one file is divided into fewer packets, and the number of simultaneous transmissions is reduced, which decreases the throughput T_n.

However, it is also important to note that T_n^* does not always increase as B increases. In fact, with B increasing, on one hand, the number of simultaneous transmissions in Stage I increases, leading to an increasing number of total transmitted useful bits nB. On the other hand, the size of the receiving cluster also increases, consuming more time-slots for completing the traffic within one cluster in Stage II. Therefore, the optimal scaling T_n is not achieved when B is maximized, but when the trade-off between the total traffic and the consumed time-slots in Stage II is optimized. Although from Section 3.5 we can learn that B_U, the upper bound of B, increases exponentially with the size of local caches M_c, the throughput scaling T_n can not increases exponentially with M_c. Thus, for our proposed scheme, the optimal scaling T_n is independent of M_c and m, as long as $nM_c \geq m$ is satisfied.

Remark 3.5. *We further explain this relationship in the context of the distributed MIMO channel. Our two-stage caching delivery can be regarded as an equivalent distributed MIMO channel with finite backhaul capacity. The transmitting and receiving antennas respectively correspond to the source nodes and the receiving nodes in our scheme. In order to achieve the capacity of this channel, the backhaul capacity should increase with the number of antennas (nodes). Otherwise, when the number of antennas (nodes) goes large, the channel capacity will be limited by the backhaul capacity and it will not linearly increase with the number of antennas (nodes). We call this the* multiplexing-backhaul trade-off *in our scheme. Therefore, as the size of cache increases, the number of transmit antennas (source nodes) grows and then the network throughput in the first stage increases due to a larger spatial degree of freedom gain. However, to guarantee the linear scaling of MIMO transmissions, the required number of receive antennas (the number of nodes within the receiving cluster) also increases. This increases the traffic within the receiving cluster and decreases the network throughput in the second stage. This explains why our achievable throughput scaling is not a function of the size of the cache and the library, but is fundamentally determined by the design of the caching network.*

Furthermore, using the above results, we can also determine the cluster size in our design as follows. We see that to achieve the optimal throughput, $N_c^* = B$ and $B_n^* = \Theta\left(n^{\frac{t}{t+1}}\right)$ must be satisfied at the same time. Recall (3.17), η must satisfy that $\Theta\left(n^{\frac{t}{t+1}}\right) = \Theta\left(n^{1-\eta}\right)$, i.e.,

$$\eta = \frac{1}{t+1}. \tag{3.40}$$

We now prove that the throughput $\Theta\left(n^{\frac{t}{t+1}}\right)$ can be achieved with a vanishing outage probability as $n \to \infty$. To see more clearly about the scaling of p_{out} with n, we further extend (3.32) as follows. It is easy to verify that as $n \to \infty$, $\frac{n-\left(L^2-1\right)B}{n} \to 1$. Therefore, $n - \left(L^2 - 1\right)B$ can be approximated by n when n goes sufficiently large. Recall that $B^* = \Theta\left(n^{\frac{t}{t+1}}\right)$, then we have

$$p_{out} \leq 1 - e^{-n^{\varepsilon}e^{-n^{\tau}}}, \tag{3.41}$$

where

$$\varepsilon = \alpha + \frac{t}{t+1}, \text{ and } \tau = 1 - (\alpha - \beta). \tag{3.42}$$

Applying L'Hospital's Rule, we have that

$$\lim_{n\to\infty} n^{\varepsilon}e^{-n^{\tau}} = \lim_{n\to\infty} \frac{\varepsilon n^{\varepsilon-\tau+1}}{\tau e^{n^{\tau}}} \overset{(a)}{=} 0, \tag{3.43}$$

where the equality (a) is obtained by keeping differentiating the numerator and the denominator until the power exponent of the numerator is negative. Thus, we have

$$\lim_{n\to\infty} p_{out} = 0. \tag{3.44}$$

meaning that the scaling $T_n = \Theta\left(n^{\frac{t}{t+1}}\right)$ is achieved with high probability.

This finishes the proof of Theorem 3.1.

3.5.4 Optimized Throughput under a Target Outage Probability

Based on the throughput and outage probability analysis in Section 3.4 and Section 3.5, we see that problem $\mathcal{P}_1$ in (3.9) now becomes

$$\mathcal{P}_2 : \max_{B} \frac{nB}{n + tL^2 Q^{\frac{t-1}{2}} B^{\frac{t+1}{t}}}$$
$$s.t.\ B \leq B_U \tag{3.45}$$

Since the constraint is linear and the objective function is concave, which can be verified by taking the second derivative in B, $\mathcal{P}_2$ is a convex optimization problem. Now introducing a Lagrange multiplier λ, $\lambda \geq 0$, we form the Lagrangian

$$L(B, \lambda) = -\frac{nB}{n + tL^2 Q^{\frac{t-1}{2}} B^{\frac{t+1}{t}}} + \lambda (B - B_U). \tag{3.46}$$

The solution to $\mathcal{P}_2$ can be found by applying the Karush-Kuhn-Tucker (KKT) condition to (4.33) and we obtain

$$\begin{cases} \left(-\frac{nB}{n+tL^2 Q^{\frac{t-1}{2}} B^{\frac{t+1}{t}}} \right)' + \lambda = 0 \\ \lambda (B - B_U) = 0 \end{cases}. \tag{3.47}$$

Then we get

$$B^* = \min \left\{ B^{opt}, B_U \right\}. \tag{3.48}$$

where $B^{opt} \triangleq \left(\frac{n}{L^2 Q^{\frac{t-1}{2}}} \right)^{\frac{t}{t+1}}$. Define a system-dependent constant

$$C \triangleq L^2 Q^{\frac{t-1}{2}}. \tag{3.49}$$

Thus, the corresponding optimal throughput can be written as

$$T_n = \begin{cases} \frac{1}{t+1} \left(\frac{n}{C} \right)^{\frac{t}{t+1}}, & \text{if } B_U \geq \left(\frac{n}{C} \right)^{\frac{t}{t+1}} \\ \frac{nB_U}{n + tCB_U^{\frac{t+1}{t}}}, & \text{if } B_U < \left(\frac{n}{C} \right)^{\frac{t}{t+1}}, \end{cases} \tag{3.50}$$

This establishes Theorem 3.2.

Remark 3.6. *According to (3.30), p_{out} increases with the number of packets B. In order to achieve a low outage probability, the corresponding B should be small. However, a smaller B also decreases the throughput T_n according to the second line in (3.50). And vice versa. We refer to this as the* throughput-outage tradeoff*, similar to that in [93].*

The throughput in (3.50) is illustrated in Fig. 3.4. Recall that B_U is the maximum number of packets of a file determined by a given library size m, local cache size M_c, and the target outage probability ϵ. For given values of m, M_c, and ϵ, if $B_U \geq B^{opt}$, as shown in the position of long dash line in Fig. 3.4, the optimal T_n^* can be achieved when $B^* = B^{opt}$. However, if $B_U < B^{opt}$, as shown in the position of solid line, the throughput will be limited by U_B and cannot achieve the optimal T_n. This reveals that in the caching network with distributed MIMO and hierarchical cooperations, the limits of the throughput are twofold. One is the multiplexing-backhaul trade-off between the total traffic and the consumed time-slots in Stage II, which corresponds to the first line in (3.50). The other one is the throughput-outage tradeoff as shown in the second line in (3.50). Especially, a low target outage probability ϵ would result in a rather small U_B, decreasing the network throughput. In other words, in this situation the network throughput will be limited by the throughput-outage tradeoff. Recall that U_B increases with the local cache size M_c. Therefore, in this regime, the local cache size M_c is able to contribute to the throughput, since a large M_c would result in a large U_B and an increased throughput.

3.6 Numerical Results

We use examples to demonstrate the performance of our proposed scheme and compare our scheme with the multi-hop caching scheme in [98] [99]. For a network of area $A = 1$ km^2, we consider two networks operating at microWave and mmWave respectively. One network operates at the carrier frequency of 3 GHz ($\lambda = 0.1$m) with the number of users $n \leq 10^4$. The other one operates at a carrier frequency of

28 GHz ($\lambda \approx 0.01$m) with the number of users $n \leq 10^5$. This meets the spatial degree of freedom condition in [89], which allows distributed MIMO and hierarchical cooperations. Also, this is a realistic setup which can be found in many scenarios. The work in [92] shows that for such a network, the optimal number of hierarchical stages $t = 3$ and the optimal quantization bit $Q = 1$. We set the target outage probability $\epsilon = 10^{-4}$ and let $\gamma = 3$ for the microWave network, and let $\epsilon = 10^{-5}$, $\gamma = 7$ for the mmWave network respectively. We fix $\alpha = 1$ for simplicity, which corresponds to the case where the number of files in the library equals to the number of nodes. Applying (3.50), we get the optimized throughput of the network as

$$T_n = \begin{cases} 0.02n^{\frac{3}{4}}, & \text{if } B_U \geq 0.09n^{\frac{3}{4}} \\ \frac{nB_U}{n+75(B_U)^{\frac{4}{3}}}, & \text{if } B_U < 0.09n^{\frac{3}{4}} \end{cases} \tag{3.51}$$

We can further derive that for the multihop caching scheme [92] [99]

$$T_n \geq \frac{n}{2L^2}\sqrt{\frac{M_c}{m}}, \tag{3.52}$$

where L is determined through (3.25).

Fig. 3.5a and Fig. 3.5b compare the throughput in [98] [99] and in (3.50) for these two networks respectively. Both figures show that our proposed scheme exhibits a near-linear behavior as n grows. This means that the network is mainly limited by the *multiplexing-backhaul trade-off* (shown in the first line in (3.50)) under realistic parameters. We observe that in both figures our proposed scheme outperforms the multihop caching scheme when the local cache size is limited (small β).

For example, in Fig. 3.5a, the throughput of our proposed scheme with $\beta = 0.4$ is better than the multihop caching scheme with $\beta = 0.56$. This corresponds to a magnificent difference of the local size caches considering that the number of nodes n is large. For example, when $n = 10^4$, $\beta = 0.4$ means that the local cache has 40 files while $\beta = 0.56$ represents a local cache with 174 files. In Fig. 3.5b, the throughput of a mmWave network demonstrates a similar behaviour. This is because the distributed MIMO technology increases the number of simultaneous transmissions

and improves the throughput of the network. If the local cache size is abundant ($\beta > 0.6$), the multihop caching scheme outperforms our proposed scheme as can be seen from both figures. This is because in this case our proposed scheme is limited by the multiplexing-backhaul trade-off, while the throughput of the multihop caching scheme grows unboundedly with the size of local caches. However, $\beta = 0.6$ is a fairly high requirement for the local cache size. For example, when $n = 10^5$, $\beta = 0.6$ requires a local cache of 1000 files. This means our proposed scheme is more "cache-economic" than the multihop caching scheme.

Moreover, we infer from Theorem 3.1 that our scaling of $\Theta(n)^{\frac{t}{t+1}}$ outperforms multihop scaling of $\Theta(n)^{1-\frac{\alpha-\beta}{2}}$ when $\alpha - \beta > \frac{2}{t+1}$. With $\alpha = 1$ and $t = 3$ in this example, our scheme will outperform multihop schemes when $\beta < 0.5$. This conjecture is confirmed by the observation in both figures that our scheme outperforms multihop schemes roughly when $\beta < 0.5$.

3.7 Conclusions

We have investigated the throughput scaling problem in a wireless D2D network where each node is equipped with a local cache and would like to download files from a pre-fixed library. Our proposed two-stage caching delivery can be regarded as an equivalent distributed MIMO channel with finite backhaul capacity. We applied distributed MIMO between source nodes and the neighbours of the destination node to increase the number of simultaneous transmissions in the network. We used hierarchical cooperations within one cluster based on existing works in [84] [92] to provide a high backhaul capacity. We established an uncoded random caching placement strategy to serve this cooperative caching delivery. Our analytical results showed that the average aggregate throughput of the proposed scheme scales almost linearly with n, with a vanishing outage probability. Furthermore, for the proposed scheme, we derived an explicit expression of the optimal throughput T_n as a function of system parameters under a target outage probability ϵ. Numerical results showed that our

proposed scheme outperforms typical existing schemes when the local cache size is limited. Future work includes characterizing the performance gap between our proposed scheme and the multihop scheme and further exploring the physical model in the cache network.

3.8 Appendix

3.8.1 Proof of Lemma 3.1

Recall that the distance between transmitter k and receive cluster D is r_{Dk}. Observe that for any node $i \in D$,

$$r_{Dk} \leq r_{ik} \leq r_{Dk} + 2\sqrt{2A_c}. \tag{3.53}$$

When node i is not in the neighboring cluster of D, we also have $r_{Dk} \geq \sqrt{A_c}$. Then,

$$1 \leq \frac{r_{ik}}{r_{Dk}} \leq 1 + 2\sqrt{2}. \tag{3.54}$$

This means that

$$(r_{ik})^{-\gamma} = \rho_{_{ik}} (r_{Dk})^{-\gamma}, \tag{3.55}$$

where $\rho_{_{kj}} \in [b, 1]$ and b is independent of B or n. Note that a boundary from the cluster D to the source nodes such as $r_{kD} \geq \sqrt{A_c}$ must be satisfied to reach (3.55). This implies that the source nodes can not be chosen from the neighbouring clusters.

Then, (3.20) can be rewritten as

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{z} = \sqrt{\frac{GP}{B}}\mathbf{F}\widetilde{\mathbf{x}} + \mathbf{z}. \tag{3.56}$$

where $f_{ik} = \rho_{_{ik}} \exp(j\theta_{ik})$ is a scaled version of h_{ik}, $\widetilde{\mathbf{x}} = [\widetilde{x}_1, \cdots, \widetilde{x}_B]$and $\mathbf{E}\left(|\widetilde{x}_k|^2\right) = 1$.

The mutual information is

$$I(\mathbf{x}; \mathbf{y}, \mathbf{F}) = E\left[\log\det\left(I + \frac{G_A P}{N_0 W B}\mathbf{F}\mathbf{F}^*\right)\right]. \tag{3.57}$$

Applying a similar proof with Appendix I in [88], we can get that for any $v < b$,

$$I(\mathbf{x}; \mathbf{y}, \mathbf{F}) \geq B \log(1 + SNRv) \frac{(b^2 - v)^2}{2}. \tag{3.58}$$

As v is independent of n, we see that $I(\mathbf{x}; \mathbf{y}, \mathbf{F})$ grows at least linearly with B. This implies the linear scaling for the distributed MIMO transmission. This establishes Lemma 3.1.

3.8.2 Proof of Theorem 3.3

We first decompose the outage probability p_{out} and obtain that $p_{out} \leq 1 - p_{hit}$ where p_{hit} is defined in Definition 6. We then derive an asymptotic expression of p_{hit} and prove Theorem 3.3.

Decomposition of the outage probability p_{out}

Rewrite the outage probability p_{out} defined in Section 3.2 as

$$\begin{aligned} p_{out} &= \frac{\sum_R \sum_y \sum_u \mathbf{1}\{t_u(G, R, y) = 0\} \Pr\{y|R, G\} \Pr\{R\}}{n} \\ &\overset{(a)}{\leq} \sum_R \sum_y \mathbf{1}\{t_u(G, R, y) = 0\} \Pr\{y|R, G\} \Pr\{R\}, \end{aligned} \tag{3.59}$$

where the inequality (a) comes from the following inequality

$$\sum_u \mathbf{1}\{t_u(G, R, y) = 0\} \leq n \cdot \mathbf{1}\{t_u(G, R, y) = 0\},$$

which is derived from the fact that for different destination nodes, their corresponding $\mathbf{1}\{t_u(G, R, y) = 0\}$ are not independent. This is because the potential source nodes for different destination nodes may overlap, and also different destination nodes may request the same file.

We then decompose p_{out} into two parts by considering the causes of outage. Generally, a node's request cannot be satisfied due to two reasons: one is that this node

cannot find its intended file within its communication range (referred to as *no cache-hit*) and the other is that the communication fails in actual delivery phases. We have that

$$\begin{aligned}&\sum_{y}\mathbf{1}\left\{t_u\left(G,R,y\right)=0\right\}\Pr\left\{y|R,G\right\}\\&=p_{out}^{delivery}+\Pr\left\{\text{No cache-hit for a given request } R \text{ of one user}\right\}.\end{aligned}\tag{3.60}$$

As such, (3.59) can be further derived as

$$\begin{aligned}p_{out}&\leq\sum_{R}\left(p_{out}^{delivery}+\Pr\left\{\text{No cache-hit for a given request } R \text{ of one user}\right\}\right)\Pr\left\{R\right\}\\&=\sum_{R}p_{out}^{delivery}\Pr\left\{R\right\}+\Pr\left\{\text{No cache-hit for any request of one user}\right\}.\end{aligned}$$

Recall that in our transmission policy Π_t, we consider reliable transmissions in each delivery stage, and hence we have $p_{out}^{delivery}=0$. Thus, in our scheme,

$$\begin{aligned}p_{out}&\leq\Pr\left\{\text{No cache-hit for any request of one user}\right\}\\&=1-\Pr\left\{\text{Cache-hit for any request of one user}\right\}.\end{aligned}\tag{3.61}$$

We will focus on the cache-hit probability in the following analysis.

Definition 3.6 (Cache-hit probability)**.** *Define $\mathcal{A}_u$ as the cache-hit event for user u, where for any requested file of node u, its B packets can be found in B different nodes among the entire networks* except for its neighboring clusters*, i.e.,*

$$\mathcal{A}_u\triangleq\left\{\begin{array}{c}\forall f_u\in\{1,\cdots m\},\ \exists u_1^s,\cdots,u_B^s\in\{1,\cdots n\}\setminus\mathcal{N},\\ \exists l_1,\cdots,l_B\in\{1,\cdots M_cB\},\ c_{u_1^s}^{l_1}\cap c_{u_2^s}^{l_2}\cap\cdots\cap c_{u_B^s}^{l_B}=f_u.\end{array}\right\}\tag{3.62}$$

The probability of $\mathcal{A}_u$, denoted by p_{hit}, is referred to as the cache-hit probability *in our scheme, i.e.,*

$$p_{hit}\triangleq\Pr\left\{\mathcal{A}_u\right\}.\tag{3.63}$$

Then,

$$p_{out}\leq1-p_{hit}.\tag{3.64}$$

Analysis of the cache-hit probability p_{hit}

It is difficult to directly calculate this cache-hit probability due to the complexity of the event $\mathcal{A}_u$ and our proposed caching strategy Π_c. To tackle this problem, we derive p_{hit} in the following way.

For a destination node u, the set of total source nodes is $\mathcal{U}\backslash\mathcal{N}$. The total caches of all potential source nodes are denoted as

$$\mathcal{M}_u^s = \underset{u_s \in \mathcal{U}\backslash\mathcal{N}}{\cup} \mathcal{M}_{u_s}. \tag{3.65}$$

To proceed, we need the following lemma.

Lemma 3.2. *Based on our proposed caching strategy Π_c, for node u, the cache-hit event $\mathcal{A}_u$ happens as long as every packet of its requested file is placed in the total source cache $\mathcal{M}_u^s$ i.e.,*

$$\Pr\{\mathcal{A}_u\} = \Pr\{\forall f_{i_u} \in \mathcal{F},\, \forall j \in \{1, \cdots, B\},\, b_{i_u,j} \in \mathcal{M}_u^s.\} \tag{3.66}$$

Lemma 3.2 can be proved by contradiction based on our proposed Π_c in Section 3.3, and the proof is omitted due to page limit. Lemma 3.2 indicates that the B packets of a file can be found at different nodes with probability one, as long as all the B packets are placed in the total source cache $\mathcal{M}_u^s$ (no matter at which node). To ensure $\mathcal{A}_u$, it is required that all packets of all files in the library have been placed in the total source cache $\mathcal{M}_u^s$. Since the goal here is to calculate the probability of collecting all mB packets in total caches, we do not need to distinguish which file a collected packet belongs to, when calculating this probability. Therefore, we transform the problem of calculating $\Pr\{\mathcal{A}_u\}$ into the following question Ψ:

$$\begin{aligned} \Psi \;:\; & \text{According to the placement strategy } \Pi_c, \text{ given } mB \text{ packets in total,} \\ & \text{how many source nodes (Denote this number by } V) \text{ are required} \\ & \text{before having placed all } mB \text{ packets in the total source cache } \mathcal{M}_u^s? \end{aligned} \tag{3.67}$$

For node u, the cache-hit event $\mathcal{A}_u$ happens when the required number of source nodes V is smaller than the actual number of source nodes. Recall that for node u,

there are a total of $n-(L^2-1)N_c$ potential source nodes. Thus,

$$p_{hit} = \Pr\{\mathcal{A}_u\} = \Pr\left(V \leq n - \left(L^2-1\right)N_c\right). \tag{3.68}$$

The problem Ψ is a special type of the Coupon Collector's problem in probability theory: coupons in groups of constant size. In the literature, there is an expression of the expected number of groups to complete the collection. However, this expression does not present an explicit analytical result. Hence, we need to develop a different analytical expression for our problem. Let p_0 be the probability of any packet in the library being chosen at each node, we have the following lemma.

Lemma 3.3. *According to our proposed caching strategy Π_c, the probability of any packet in the library being picked at each node is $\frac{M}{m}$, i.e.,*

$$p_0 = \frac{M}{m}. \tag{3.69}$$

Proof. We omit the details of the proof due to the page limit. The approach is to first consider that each node randomly chooses M_cB files from m files with probability $\frac{M_cB}{m}$, and then picks up one of the packets of the chosen files with probability $\frac{1}{B}$. □

The caching placement with our proposed Π_c can be regarded as picking one packet at a time, with a probability p_0 determined in Lemma 3.3. Denote a packet in the library by k, $k \in \{1, 2, \cdots, mB\}$. Similar to the Coupon Collector's problem, we can then obtain the probability of getting a new packet k as

$$p_k = \frac{mB-(k-1)}{mB} \cdot M_cB. \tag{3.70}$$

Asymptotic expression of the cache-hit probability

Based on the analysis above, we now treat the whole caching placement process as selecting one packet at a time with probability p_0.

Lemma 3.4. *When the number of nodes n is sufficiently large, the probability of the required number of source nodes V less than the number of source nodes $n-(L^2-1)N_c$*

is

$$\Pr\left(V < n - \left(L^2 - 1\right) N_c\right) = \Pr\left(V < \frac{mB}{M_c B} \log mB + a \frac{mB}{M_c B}\right) \to e^{-e^{-a}}. \quad (3.71)$$

where $a = \left(n - \left(L^2 - 1\right) N_c\right) \frac{M_c B}{mB} - \log mB$.

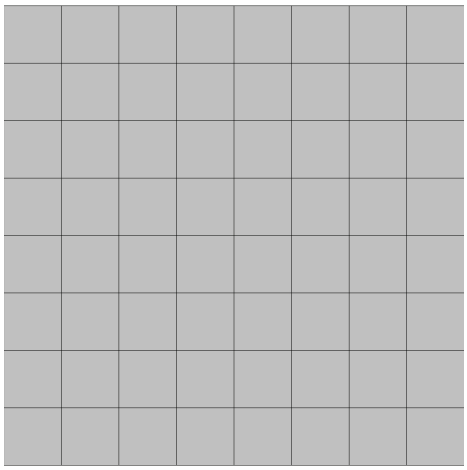
Proof. The proof is similar to that in [129], only with a different drawing probability $p_0 = \frac{M}{m}$. We omit the details of the proof due to the space limit. □

Then,

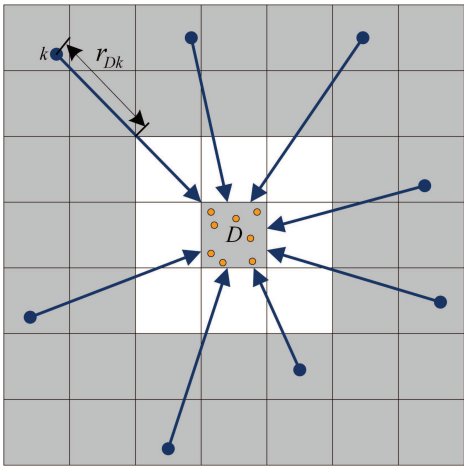
$$p_{hit} = \Pr\left(V < n - KN_c\right) \to e^{-e^{-a}} = e^{-mB \cdot e^{-\frac{\left(n - \left(L^2-1\right)N_c\right)M}{m}}}. \quad (3.72)$$

Since $0 < \alpha - \beta < 1$, as $n \to \infty$, we have $n^{1-(\alpha-\beta)} \to \infty$, $e^{-\frac{nM}{m}} \to 0$ and $p_{hit} \to 1$. This indicates that all mB packets can be collected from $n - \left(L^2 - 1\right) N_c$ source node candidates with probability one as n becomes very large. The larger the size of local memory M_c is, the more quickly that $p_{hit} \to 1$ as $n \to \infty$.

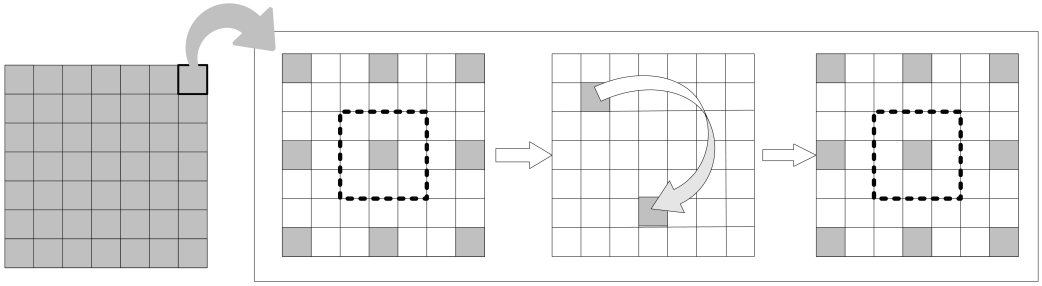
This proves the outage probability results in Theorem 3.3.



(a) Clustering: Divide the entire network into square cells of area A_c where $A_c = n^{-\eta}$, $\eta > 0$. η is a parameter determining the cluster size. The bigger η is, the smaller each cluster is.



(b) Stage I: Distributed MIMO. The middle grey cluster represents the current destination cluster D. The white area denotes its neighbouring clusters $\mathcal{N}$. The surrounding grey area contains all candidates of source nodes. The distance from transmitter k to the destination cluster D is denoted by r_{Dk}.



(c) Stage II: Hierarchical Cooperations. All clusters work simultaneously. Within one cluster, a three-phase hierarchical cooperation is employed including clustering, distributed MIMO and quantize-and-forward. A time-division protocol is used for the first and third phases. For example, in this figure, only one cluster is allowed to transmit among 9 clusters ($L = 3$).

Figure 3.3. Caching delivery phase of the proposed scheme: (a) Clustering; (b) Stage I: Distributed MIMO; (c) Stage II: Hierarchical Cooperations.

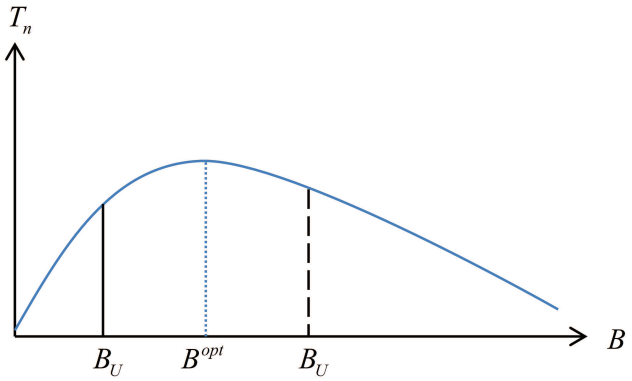
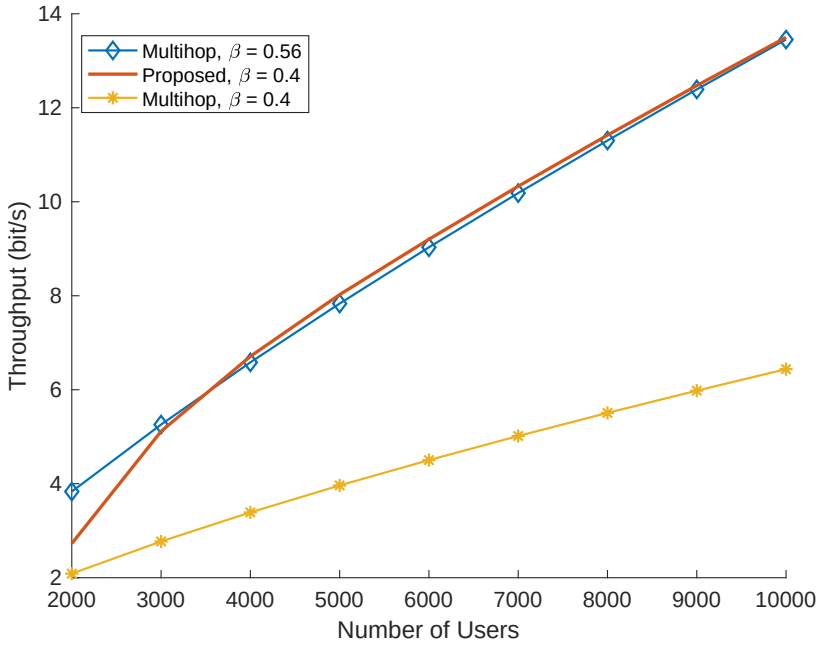
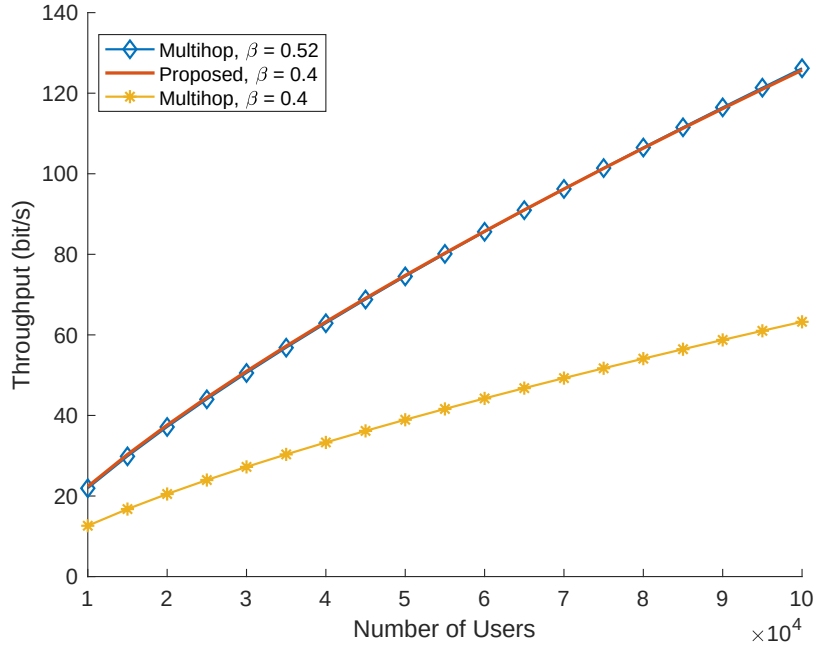


Figure 3.4. Achievable throughput as the function of B.



(a) $n \leq 10^4$ users in 1 km^2, operating at 3 GHz, $\alpha = 1$, $\gamma = 3$, $\epsilon = 10^{-4}$.



(b) $n \leq 10^5$ users in 1 km^2, operating at 28 GHz, $\alpha = 1$, $\gamma = 7$, $\epsilon = 10^{-5}$.

Figure 3.5. Achievable throughput of a D2D caching network as a function of the number of users n.

Chapter 4

Enhancing Cellular Performance through Device-to-Device Distributed MIMO

4.1 Introduction

The integration of local D2D communications and cellular connections has been intensively studied to satisfy co-existing D2D and cellular communication demand. In future cellular networks, there will be numerous standby users possessing D2D communication capabilities in close proximity to each other. Considering that these standby users do not necessarily request D2D communications all the time, this chapter proposes a hybrid D2D-cellular scheme to make use of these standby users and to improve the rate performance for cellular users. More specically, through D2D links, a virtual antenna array can be formed by sharing antennas across different terminals to realize the diversity gain of MIMO channels. This thesis considers the use of millimeter wave (mmWave) links to enable high data rate D2D communications. We observe that although the mmWave communication employs directional antenna beams, it can still suffer interference. We then design an orthogonal D2D multiple access protocol

and formulate the optimization problem of joint cellular and D2D resource allocation for downlink transmissions of our proposed scheme. We obtain a closed-form solution for D2D resource allocation, which reveals useful insights for practical system design. Numerical results from extensive system-level simulations demonstrate that the rate performance of cellular users is signicantly improved through our proposed scheme and the resource allocation algorithm.

The chapter is organized as follows. In Section 4.2, we introduce the system model, present our proposed hybrid D2D-cellular scheme and formulate the utility maximization problem. For the formulated problem, we propose a two-step optimization algorithm in Section 4.3. Section 4.4 provides numerical results of a system-level simulation and discussions of insight.

4.2 System Model and Proposed Hybird D2D-Cellular Scheme

Consider a downlink multi-cell interfering broadcast channel with N base stations (BSs) serving K active users in each cell. Each BS b, $b \in \{1, \cdots, N\}$, is equipped with L transmit antennas and each user, no matter active or non-active, has a single receive antenna. For the considered system model, it is widely recognized that inter-cell interference is the main cause of the rate degradation, especially in a dense cellular networks or for cell-edge users. In this thesis, we propose a novel outband D2D cellular scheme, to enhance the capability of encountering the inter-cell interference for end-users.

As illustrated in Fig. 4.1, our proposed scheme consists of a two-phase cooperative transmission procedure. In the first phase, the BSs transmits data to active users and standby users receive signals by "overhearing". In the second phase, standby users help relay the received data to the active users so that multiple observations of the transmitted signals are jointly processed at the active users. The application of the

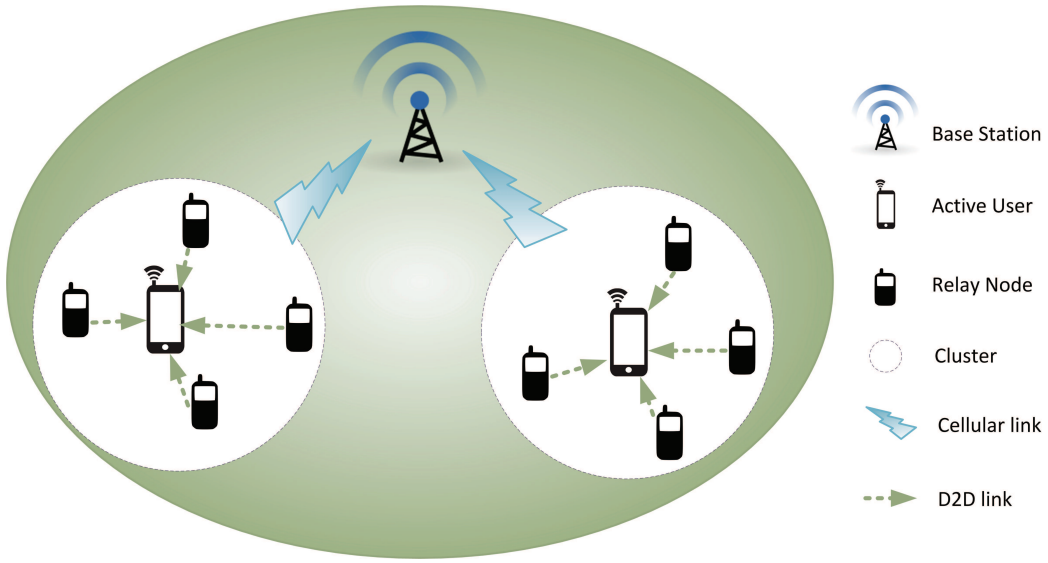


Figure 4.1. Proposed cellular system with outband D2D communications.

distributed-MIMO technique here enhances the capability of inter-cell interference mitigation of end users, and thus improves the rate performance, especially for the cell-edge users.

In general, when designing a D2D communication scheme, peer discovery, physical layer procedures and radio resource management algorithms are the three main issues to be considered [130]. We note that the peer discovery required in our scheme is in line with conventional D2D communication schemes. We do not focus on this problem in the thesis since existing peer discovery methods in literatures can be adopted. Instead, we focus on physical layer procedures in this section, and then radio resource management algorithms in Section 4.3.

4.2.1 Pre-Transmission: Clustering

In each BS, there are three type of users: active cellular users, users requesting D2D connections and standby users who do not request cellular nor D2D connections. Note that the number of these standby users is usually much larger than the number of active users. We propose a clustering strategy to make use of these standby users to enhance the rate performance of active users as follows.

We first define a maximum D2D communication range $d_{\max}$, which is the maximum geographical distance between two users allowed by the user power budget and the employed D2D communication technique. We then present some definitions for our proposed scheme as follows.

- *Relay node*: For each active user k in each cell, all its nearby standby users within the distance $d_{\max}$ are its potential relay nodes. Among these nodes, those who are willing to help the active users[1] are referred to as relay nodes. We assume that one relay node only helps one active user at the same time. We also assume that each BS is aware of the relay decisions of all relay nodes within its cell.

- *Cluster*: After relaying decisions at potential relay nodes, an active user and its relay nodes thereby form a cluster. We assume that each active user has the same number of relay nodes for simplicity. We see that there are K clusters of identical size within each cell. We assume that there are $M-1$ relay nodes for each active user, so that together with the direct link from the BS to each user, an $L \times M$ MIMO channel can be formed for each active user.

Note that this clustering strategy limits the required transmission range of D2D links, which avoids the disadvantage of large propagation loss for long distance transmissions when operating at the mmWave carrier frequency. It should be noted that only devices with two wireless interfaces (LTE and mmWave) can use outband D2D, as these users have both D2D and cellular communication capabilities.

4.2.2 Phase I: BS-to-user Transmission

In our proposed scheme, for each user, there is one direct link from the BS to the active user, and $M-1$ one-hop relay links. We denote the k-th active user in the cell

[1]Users incentive for helping others has been studied in many literatures and is not the focus of this thesis.

of BS b, $k \in \{1, \cdots, K\}$ by user b_k in this thesis. We also assume that each active user has only a single data stream for simplicity. Then, let $\mathbf{v}_{b_k} \in \mathbb{C}^{L\times 1}$ denote the beamformer that BS b uses to transmit signals to user b_k, and the transmitted signal from BS b can be written as

$$\mathbf{x}_b = \sum_{j=1}^{K} \mathbf{v}_{b_k} s_{b_k}, \tag{4.1}$$

where s_{b_k} is the message for user b_k and $\mathbf{E}\left[s_{b_k}^2\right] = 1$. We assume that s_{b_k} is chosen from a Gaussian codebook and messages for different users are independent from each other and from receiver noises. Collecting all beamformers $\mathbf{v}_{b_k}$ used by BS b, we obtain the beamformer matrix $\mathbf{V}_b = [\mathbf{v}_{b_1}, \cdots, \mathbf{v}_{b_K}] \in \mathbb{C}^{L\times K}$. Then, the power budget of each BS can be presented as

$$\mathbf{Tr}\left(\mathbf{V}_b \mathbf{V}_b^*\right) \leq P_B. \tag{4.2}$$

The directly received signal at user b_k can be written as

$$y_{b_k} = \mathbf{h}_{b,b_k}^T \mathbf{v}_{b_k} s_{b_k} + \sum_{\substack{j=1 \\ j\neq k}}^{K} \mathbf{h}_{b,b_k}^T \mathbf{v}_{b_j} s_{b_j} + \sum_{\substack{i=1 \\ i\neq b}}^{N} \sum_{\substack{j=1 \\ j\neq k}}^{K} \mathbf{h}_{i,b_k}^T \mathbf{v}_{i_j} s_{i_j} + n_{b_k}, \tag{4.3}$$

where $\mathbf{h}_{i,b_k} \in \mathbb{C}^{L\times 1}$, $i \in \{1, \cdots, N\}$, represents the channel state information (CSI) vector from BS i to user b_k, and n_{b_k} is the received noise at user b_k and follows $\mathbb{C}\mathcal{N}(0, \sigma^2)$.

Due to the broadcasting property of the wireless channel, the relay nodes also receive the transmitted signals from BS by "overhearing". The received signals at the m-th relay node of user b_k, can be described as

$$y_{b_k}^m = \sum_{i=1}^{N} \sum_{j=1}^{K} \left(\mathbf{h}_{i,b_k}^m\right)^T \mathbf{v}_{i_j} s_{i_j} + n_{b_k}^m, m \in \{1, \cdots, M-1\}, \tag{4.4}$$

where $\mathbf{h}_{i,b_k}^m \in \mathbb{C}^{L\times 1}$ is the channel coefficient from BS i to the m-th relay node of user b_k and $n_{b_k}^m \sim \mathcal{CN}(0, \sigma^2)$ is the additive white Gaussian noise at the m-th relay node of user b_k. We consider a same level of thermal noise σ^2 at different users for simplicity.

4.2.3 Phase II: Intra-cluster User Cooperation

D2D communications between users within a cluster are enabled in Phase II to realize user cooperations and ultimately facilitate BS-to-user transmissions. We assume that D2D communications take place at the mmWave frequency in this thesis.

Operations at each relay node

In this phase, relay nodes relay the received signals from Phase I to its helped active users. The relaying strategy at each relay node is chosen as follows. We note that our proposed scheme can be regarded as a general Gaussian relay channel, in which the channel of first hop (BS-to-relay) is weak (low signal-to-noise ratio) and the channel of second hop (relay-to-user) is relatively strong (high signal-to-noise ratio). Information theoretic considerations reveal that the compress-and-forward strategy is appropriate for such a relay channel [131]. Therefore, each observation at the m-th relay node of user b_k, $y_{b_k}^m$, is compressed and then forwarded to their active users.

We assume that the compression at each relay node is performed independently to simplify the intra-cluster cooperation. The compression procedure is modeled as the following forward test channel

$$\widetilde{y}_{b_k}^m = y_{b_k}^m + e_{b_k}^m, \tag{4.5}$$

where $e_{b_k}^m \in \mathbb{C}$ is the quantization noise independent of $y_{b_k}^m$ and it is assumed to be Gaussian distributed with zero mean and variance $q_{b_k}^m$. After performing this compression at relay node m, the corresponding information rate of $\widehat{y}_{b_k}^m$ is

$$C_{b_k}^m = \log\left(1 + \frac{\beta_{b_k}^m}{q_{b_k}^m}\right), \tag{4.6}$$

where

$$\beta_{b_k}^m = \sum_{i=1}^{N}\sum_{j=1}^{K} \left(\mathbf{h}_{i,b_k}^m\right)^T \mathbf{v}_{i_j}\mathbf{v}_{i_j}^* \left(\left(\mathbf{h}_{i,b_k}^m\right)^T\right)^* + \sigma^2. \tag{4.7}$$

We can regard $\beta_{b_k}^m$ as a parameter related with the rate contribution of relay node m to the user rate R_{b_k}. Smaller $\beta_{b_k}^m$ leads to a lower information rate of the compressed signal $\widehat{y}_{b_k}^m$, while smaller quantization noise $q_{b_k}^m$ leads to a higher compression rate.

Now let us denote by t_c the frame duration of cellular transmissions and by B_c the cellular bandwidth. Then, during one frame duration, the required number of information bits to be transmitted from the m-th relay node to user b_k is $t_c B_c C_{b_k}^m$.

Multiple access protocol

We now consider the multiple access aspect of multiple relay nodes transmitting to their targeted cellular user. We note that for mmWave communications in a dense network, the angles of arrive (AOAs) of signals from different relay nodes to one destination user might be rather close, and these signals would interfere with each other if they are transmitted simultaneously. Such interference can happen within one cluster as well as between different clusters. For example, relay nodes may interfere a un-targeted cellular user when their AOAs are close to the AOAs of the intended signals at this cellular user.

Thus, the interference between simultaneous D2D transmissions must be managed. One possible solution here is to consider non-orthogonal multiple access (NOMA) from relay nodes to the destination user. However, the corresponding power allocation algorithm would be very complicated due to the large number of relay nodes. In addition, it has been widely recognized that successive interference cancellation (SIC) is required at the receiver for realizing NOMA, which would increase the implementation complexity at end-users. Furthermore, the error propagation issue of SIC is usually difficult to deal with [132]. For this reason, this thesis adopts orthogonal multiple access to avoid interference and to simplify the operations at end-users.

We propose orthogonal D2D transmissions as follows. As depicted in Fig. 4.2, we adopt FDMA among different clusters and TDMA within one cluster.

Among different clusters, we assume that a same bandwidth of B_d is allocated to each cluster. Thus, a bandwidth of KB_d is required for serving all K active users within one cell. To deal with D2D interference from adjacent cells, a multi-cell frequency reuse model is used as depicted in Fig. 4.3.[2] With this frequency reuse

[2]Denote the minimum distance between the sectors using the same frequency by $d_{\min}$. Using a

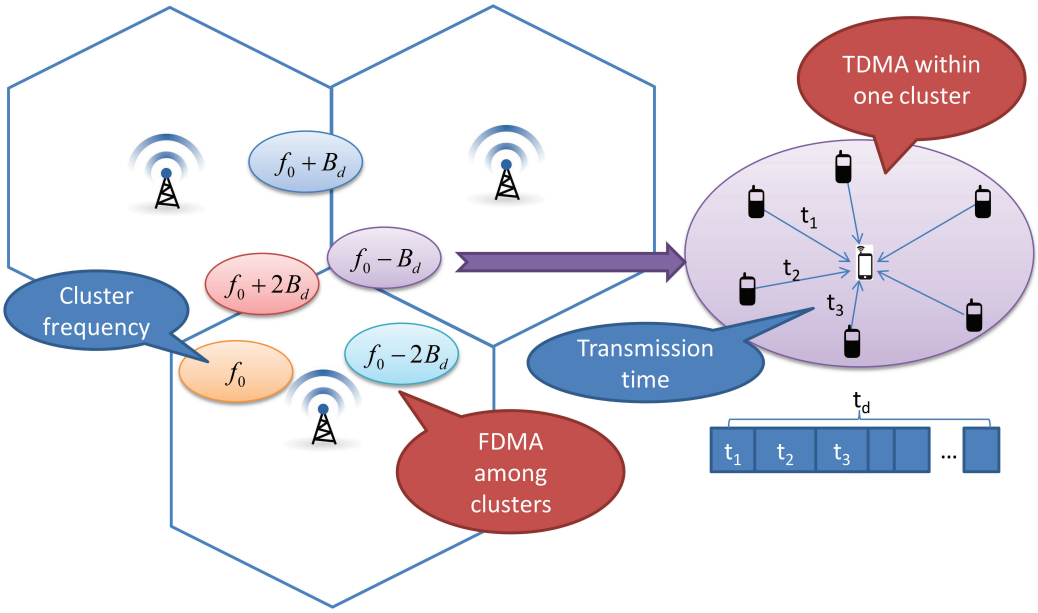


Figure 4.2. Proposed multiple access of D2D links. Here, f_0 is the carrier frequency of D2D links, and B_d is the bandwidth shared by one cluster. We propose FDMA among different clusters and TMDA within one cluster.

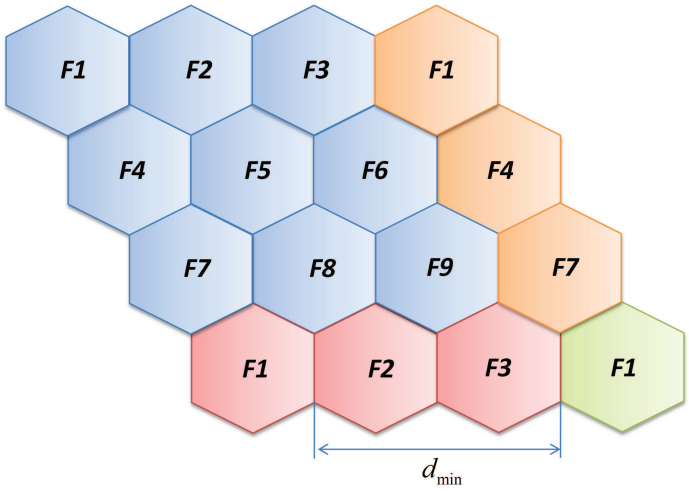


Figure 4.3. Frequency reuse pattern of the proposed FDMA protocol.

pattern, the total required bandwidth is $9KB_d$. This design makes use of the ample bandwidth exhibited by the mmWave communications. For example, at the E-band carrier frequency, a total of 10 GHz is available. We may allocate a typical 200 MHz bandwidth for each user allowing around 50 FDMA clusters, which is capable of accommodating $5-6$ active cellular users within each cell.

Since FDMA and frequency reuse are employed, all clusters are allowed to transmit simultaneously. Let us assume that a same time duration t_d is available for each cluster in Phase II. Within each cluster, each relay node is allocated with an exclusive transmission duration, which is the D2D resource allocation parameter to be optimized. In this manner, interferences between D2D links are thoroughly avoided.

Denote by $t_{b_k}^m$ the allocated time duration to the D2D link from relay node m to user b_k. The D2D link constraint can then be described as

$$\sum_{m=1}^{M-1} t_{b_k}^m \leq t_d, \ \ \forall b, k. \tag{4.8}$$

Assuming a power budget P_D for D2D transmissions at each relay node, we further obtain that the actual channel capacity $r_{b_k}^m$ allowed by the D2D channel is

$$r_{b_k}^m = \log\left(1 + \frac{P_D \left|l_{b_k}^m\right|^2}{N_0 B_d}\right), \tag{4.9}$$

where $l_{b_k}^m$ represents the channel coefficient of the D2D link from relay node m to user b_k, and N_0 is noise power spectral density. Thus, we see that a total number of $t_{b_k}^m B_d r_{b_k}^m$ information bits can be transmitted via the D2D link from relay node m to user b_k.

Therefore, to guarantee no information loss through the D2D transmissions, for the m-th relay node of user b_k, the number of transmitted information bits via the D2D

common path loss model in [133,134] and a power budget 20 dBm, we calculate that as long as $d_{\min}$ is larger than 940 m, the D2D inter-cell interference is less than 5% of noise power (noise power spectral density $N_0 = 169$ dBm/Hz). From Fig. 4.3, we see that $d_{\min}$ larger than 940 m corresponds to a cell radius no less than 235 m. This fits the current network setups. Therefore, with this frequency reuse pattern, the inter-cell interference is negligible in determining the rate performance.

link $t_{b_k}^m B_d r_{b_k}^m$ must be larger than the required number of compression bits $t_c B_c C_{b_k}^m$. That is to say, the condition $t_{b_k}^m B_d r_{b_k}^m \geq t_c B_c C_{b_k}^m$ must be satisfied for any b, k, m. In other words, we see that the allocated time duration for the m-th relay node of user b_k must satisfy

$$t_{b_k}^m \geq \frac{t_c B_c C_{b_k}^m}{B_d r_{b_k}^m}, \quad \forall b, k, m. \tag{4.10}$$

The ratio $\frac{t_c B_c}{B_d}$ can be regarded as a constant determined by system parameters. We see that the larger the required compression information rate $C_{b_k}^m$ and the smaller the actual channel capacity $r_{b_k}^m$, the more D2D time resource is required for relay node m.

Remark 4.1. *Phase I and Phase II must happen consecutively without overlapping. This is because in our design, one relay node is required to deliver information to the active user in Phase II, after collecting all the received signals from Phase I.*

4.2.4 Equivalent Transmission Model

As illustrated in Fig. 4.4, for active user b_k, upon collecting all the compressed observations $\widehat{y}_{b_k}^m$ from its relay nodes in Phase II and its direct observation y_{b_k}, the received signals at user b_k can be arranged as

$$\mathbf{y}_{b_k} = \left[\begin{array}{cccc} \widetilde{y}_{b_k}^1 & \widetilde{y}_{b_k}^2 & \cdots & \widetilde{y}_{b_k}^{M-1} & y_{b_k} \end{array} \right]^T. \tag{4.11}$$

The equivalent transmission model from BS b to user b_k can be further written as

$$\mathbf{y}_{b_k} = \sum_{i=1}^{N} \sum_{j=1}^{K} \mathbf{H}_{i,b_k} \mathbf{v}_{i_j} s_{i_j} + \mathbf{n}_{b_k} + \mathbf{e}_{b_k}, \tag{4.12}$$

where

$$\mathbf{H}_{i,b_k} = \left[\begin{array}{cccc} \mathbf{h}_{i,b_k}^1 & \mathbf{h}_{i,b_k}^2 & \cdots & \mathbf{h}_{i,b_k}^{M-1} & \mathbf{h}_{i,b_k} \end{array} \right]^T \in \mathbb{C}^{M \times L}, \tag{4.13}$$

$$\mathbf{n}_{b_k} = \left[\begin{array}{cccc} n_{b_k}^1 & n_{b_k}^2 & \cdots & n_{b_k}^{M-1} & n_{b_k} \end{array} \right]^T \in \mathbb{C}^{M \times 1}, \tag{4.14}$$

and

$$\mathbf{e}_{b_k} = \left[\begin{array}{cccc} e_{b_k}^1 & e_{b_k}^2 & \cdots & e_{b_k}^{M-1} & 0 \end{array} \right]^T \in \mathbb{C}^{M \times 1}. \tag{4.15}$$

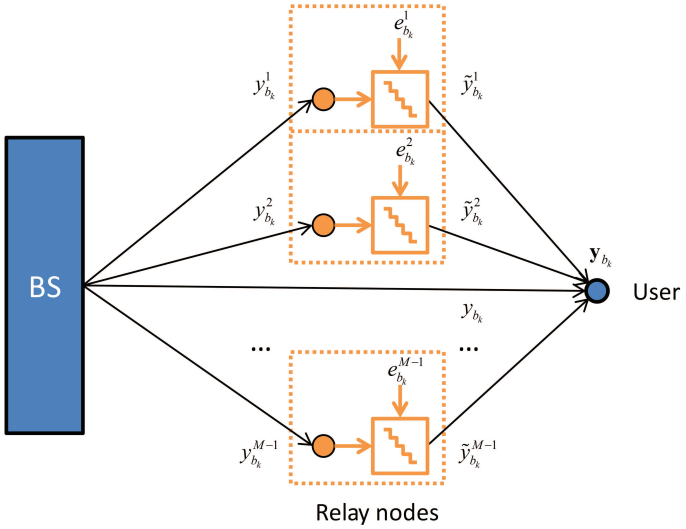


Figure 4.4. Equivalent model of the proposed scheme.

We see that $\mathbf{n}_{b_k}$ is a collection of M i.i.d. Gaussian random variables and $\mathbf{n}_{b_k} \sim \mathcal{CN}(0, \sigma^2\mathbf{I}_M)$. The vector $\mathbf{e}_{b_k}$ collects independent quantization noises from relay nodes, whose last element is zero since there is no quantization noise in the direct BS-to-user transmission.

Thus, by jointly processing the compressed observations $\widetilde{y}_{b_k}^1, \cdots, \widetilde{y}_{b_k}^{M-1}$ and the direct observation y_{b_k}, user b_k and its relay nodes realize a distributed-MIMO channel. Then, the achievable rate for user b_k is written as

$$\widetilde{R}_{b_k} = \log\left(1 + \mathbf{v}_{b_k}^* \mathbf{H}_{b,b_k}^* \mathbf{J}_{b_k}^{-1} \mathbf{H}_{b,b_k} \mathbf{v}_{b_k}\right), \tag{4.16}$$

where

$$\mathbf{J}_{b_k} = \sum_{\substack{j=1 \\ j\neq k}}^{K} \mathbf{H}_{b,b_k} \mathbf{v}_{b_j} \mathbf{v}_{b_j}^* \mathbf{H}_{b,b_k}^* + \sum_{\substack{i=1 \\ i\neq b}}^{N} \sum_{\substack{j=1 \\ j\neq k}}^{K} \mathbf{H}_{i,b_k} \mathbf{v}_{i_j} \mathbf{v}_{i_j}^* \mathbf{H}_{i,b_k}^* + \sigma^2 \mathbf{I}_M + \begin{bmatrix} \mathbf{Q}_{b_k} & 0 \\ 0 & 0 \end{bmatrix}, \tag{4.17}$$

and

$$\mathbf{Q}_{b_k} = \begin{bmatrix} q_{b_k}^1 & 0 & \cdots & \cdots \\ 0 & q_{b_k}^2 & 0 & \cdots \\ \cdots & 0 & \cdots & 0 \\ \cdots & \cdots & 0 & q_{b_k}^{M-1} \end{bmatrix}. \tag{4.18}$$

Considering the fact that Phase I and Phase II happens consecutively, the actual information rate of user b_k is

$$R_{b_k} = \frac{t_c}{t_d + t_c} \widetilde{R}_{b_k}. \tag{4.19}$$

We see that by permitting adjacent users to cooperate with each other, users located in different locations can form a virtual antenna array. This allows the deployment of MIMO techniques to single-antenna users and enhances the channel capacity by diversity gain.

We further notice that with the multiple observations at each user, receive beamforming is also required for our proposed scheme to better realize the MIMO capacity. In this thesis, we treat interference as noise and consider a linear receive beamforming strategy. So the estimated message is given by

$$\widehat{s}_{b_k} = \mathbf{u}_{b_k}^* \mathbf{y}_{b_k}, \tag{4.20}$$

where $\mathbf{u}_{b_k} = \left[u_{b_k}^1, u_{b_k}^2, \cdots, u_{b_k}^M\right] \in \mathbb{C}^{M \times 1}$ is the receive beamformer at the user b_k. We see that the receive beamformer is also a parameter to be optimized.

4.2.5 Problem Formulation

Under the context of the multi-cell interfering broadcast channel, a popular utility maximization problem is to find the optimal transmit and receive beamformers, so that the sum-rate $\sum_{b=1}^{N}\sum_{k=1}^{K} R_{b_k}$ is maximized. On top of this, we note that the resource allocation for D2D links is also important for our proposed scheme. This is because from (4.6) and (4.10) we see that the quantization noise variance matrix $\mathbf{Q}_{b_k}$ is determined by the allocated time duration $t_{b_k}^m$ of D2D links, which ultimately affects the user rate R_{b_k}. This section formulates a problem of multicell sum-rate maximization by jointly optimizing transmit beamformers, receive beamformers and quantization noise variance matrices. We also note that the CSI knowledge in this thesis includes the CSI from BSs to active users as well as that from BSs to relay nodes.

Intuitively, there are two primary factors affecting the allocated time duration $t_{b_k}^m$ for each relay node: one is its rate contribution to cellular user rate R_{b_k}, brought by

the cellular observation $y_{b_k}^m$ at relay node m, and the other is the D2D link quality from relay node m to user b_k. These two factors may vary at different relay nodes due to different cellular and D2D channel realizations. It is immediate to see that if a relay node has very weak cellular observations and poor D2D connections, we should avoid allocating much transmission time to it. And it is also obvious that a relay node with strong cellular observations and good D2D connections should be allocated with a fairly amount of transmission time. However, for the cases between these extreme situations, how do we quantitatively allocate a reasonable transmission time for each relay node? This D2D resource allocation problem is important yet challenging since there is not a straightforward relationship between the individual cellular observation at each relay node, the D2D link quality and the rate of the destination user.

In this thesis, the network utility maximization problem can now be formulated as follows.

$$\mathcal{P}_0: \quad \max_{\mathbf{v}_{b_k},\mathbf{u}_{b_k},t_{b_k}^m} \sum_{b=1}^{N}\sum_{k=1}^{K} R_{b_k} \quad \text{s.t.} \quad \begin{cases} \text{C1:} & t_{b_k}^m \geq \frac{t_c B_c C_{b_k}^m}{B_d r_{b_k}^m}, \forall m,b,k \\ \text{C2:} & \sum_{m=1}^{M-1} t_{b_k}^m \leq t_d, \forall b,k \\ \text{C3:} & \mathbf{Tr}\left(\mathbf{V}_b \mathbf{V}_b^*\right) \leq P_B, \forall b \end{cases}, \tag{4.21}$$

where R_{b_k} is defined as in (4.19), C1 is the information flow constraint that accounts for lossless information passing through D2D links, C2 is the transmission time constraint for D2D links and C3 is the BS transmission power constraint.

We see that C1 must be satisfied with equality, i.e.,

$$t_{b_k}^m = \frac{t_c B_c C_{b_k}^m}{B_d r_{b_k}^m}, \tag{4.22}$$

This is because the objective function monotonically decreases with $q_{b_k}^m$, which also decreases with the capacity of the D2D link $C_{b_k}^m$. To maximize R_{b_k}, $C_{b_k}^m$ should be as large as possible.

To better understand the property of problem $\mathcal{P}_0$, we transform it into optimizing over the quantization noise covariance $q_{b_k}^m$ instead of the allocated time $t_{b_k}^m$ as follows.

Denote

$$w_{b_k}^m = \frac{t_c B_c}{t_d B_d r_{b_k}^m}, \tag{4.23}$$

which can be interpreted as a parameter measuring the D2D link quality. The smaller the $w_{b_k}^m$ is, the better the D2D link is. Thus, using this notation and together with (4.6), we combine C1, C2 and obtain the new D2D link constraints as

$$\begin{cases} \text{C4:} \ \sum\limits_{m=1}^{M-1} w_{b_k}^m \log\left(1 + \frac{\beta_{b_k}^m}{q_{b_k}^m}\right) \leq 1. \\ \text{C5:} \ q_{b_k}^m \geq 0. \end{cases} \tag{4.24}$$

Then, we obtain the following problem

$$\begin{aligned} \mathcal{P}_1 : \quad & \max_{\mathbf{v}_{b_k}, \mathbf{u}_{b_k}, q_{b_k}^m} \sum_{b=1}^{N} \sum_{k=1}^{K} R_{b_k} \\ & \text{s.t.} \quad \text{C3, C4, C5} \end{aligned} \tag{4.25}$$

We see that $\mathcal{P}_1$ is not a convex problem. The major difficulty in solving $\mathcal{P}_1$ arises from the fact that the objective function and the constraints are both concave in transmit beamformer $\mathbf{v}_{b_k}$ and convex in quantization noise covariance $q_{b_k}^m$.

4.3 Enhanced Rate-performance of Cellular Users

In this section, we propose a two-step iterative optimization method for the formulated problem $\mathcal{P}_1$. In the following, we show that for given beamformers $\mathbf{v}_{b,k}$, $\mathbf{u}_{b,k}$, the optimization of the quantization noise covariance $q_{b_k}^m$ can be transformed into a convex problem and solved in closed-form by applying the Karush-Kuhn-Tucker (KKT) conditions. Then, the optimization of beamformers $\mathbf{v}_{b_k}$, $\mathbf{u}_{b_k}$ under fixed quantization noise covariance $q_{b_k}^m$ can be regarded as a general multi-cell sum-rate maximization problem. An iterative optimization algorithm is proposed between the two steps, leading to an appropriate use of D2D links.

4.3.1 Optimized D2D Resource Allocation

We first assume that all transmit and receive beamformers are fixed. Then, for given beamformers $\mathbf{v}_{b_k}$ and $\mathbf{u}_{b_k}$, the problem of quantization noise covariance optimization at the relay node for each single-antenna user can be described as follows.

$$\mathcal{P}_2 : \quad \begin{array}{l} \max\limits_{q_{b_k}^m} R_{b_k} \\ \text{s.t.} \quad \text{C4, C5} \end{array} \tag{4.26}$$

Problem $\mathcal{P}_2$ here is a non-convex problem and we transform it to a convex problem as follows. Note that $\mathcal{P}_2$ is equivalent to minimizing the mean square error (MSE) since a single user rate is considered. Thus, instead of solving $\mathcal{P}_2$, we study the following problem.

$$\mathcal{P}_3 : \quad \begin{array}{l} \min\limits_{q_{b_k}^m} MSE_{b_k} \\ \text{s.t.} \quad \text{C4, C5} \end{array} \tag{4.27}$$

where

$$\begin{aligned} MSE_{b_k} &= E\left[\left(\widehat{s}_{b_k} - s_{b_k}\right)\left(\widehat{s}_{b_k} - s_{b_k}\right)^*\right] && (4.28) \\ &= \mathbf{u}_{b_k}^* \left(\sum_{i=1}^{N} \sum_{j=1}^{K} \mathbf{H}_{i,b_k} \mathbf{v}_{i_j} \mathbf{v}_{i_j}^* \mathbf{H}_{i,b_k}^* + \sigma^2 \mathbf{I}_M \right) \mathbf{u}_{b_k} && (4.29) \\ &\quad + \mathbf{u}_{b_k}^* \begin{bmatrix} \mathbf{Q}_{b_k} & 0 \\ 0 & 0 \end{bmatrix} \mathbf{u}_{b_k} - 2\,\mathrm{Re}\left\{\mathbf{u}_{b_k}^* \mathbf{H}_{b,b_k} \mathbf{v}_{b_k}\right\} \\ &= \sum_{m=1}^{M-1} q_{b_k}^m u_{b_k}^{m*} u_{b_k}^m + Const. && (4.30) \end{aligned}$$

Then, by substituting (4.6) into (4.30), we have

$$\mathcal{P}_4 : \quad \begin{array}{l} \min\limits_{C_{b_k}^m} \sum\limits_{m=1}^{M-1} \frac{\beta_{b_k}^m u_{b_k}^{m*} u_{b_k}^m}{2^{C_{b_k}^m} - 1} \\ \text{s.t.} \sum\limits_{m=1}^{M-1} w_{b_k}^m C_{b_k}^m \leq 1 \;\&\; C_{b_k}^m \geq 0. \end{array} \tag{4.31}$$

We see that the above problem is a convex optimization problem, since the constraint is linear and the objective function is convex, which can be verified by taking the

second derivative in $C_{b_k}^m$. Now introducing Lagrange multiplier μ and $\boldsymbol{\lambda} \in \mathbb{R}^{M-1}$, μ, $\boldsymbol{\lambda} \geq 0$, we form the Lagrangian

$$L\left(C_{b_k}^m, \mu, \boldsymbol{\lambda}\right) = \sum_{m=1}^{M-1} \frac{\beta_{b_k}^m u_{b_k}^{m*} u_{b_k}^m}{2^{C_{b_k}^m} - 1} + \mu \left(\sum_{m=1}^{M-1} w_{b_k}^m C_{b_k}^m - 1 \right) - \sum_{m=1}^{M-1} \lambda_k C_{b_k}^m. \tag{4.32}$$

Taking the derivative of the above with respect to $C_{b_k}^m$, we apply the KKT condition as follows.

$$\frac{\partial L}{C_{b_k}^m} = -\frac{\beta_{b_k}^m u_{b_k}^{m*} u_{b_k}^m}{\left(2^{C_{b_k}^m} - 1\right)^2} 2^{C_{b_k}^m} \ln 2 + \mu w_{b_k}^m - \lambda_k = 0. \tag{4.33}$$

Note that $\lambda_k = 0$ whenever $C_{b_k}^m > 0$. Now, the optimal $C_{b_k}^m$ must satisfy the D2D constraint C4 with equality, i.e.,

$$\sum_{m=1}^{M-1} w_{b_k}^m C_{b_k}^m = 1. \tag{4.34}$$

This is because the objective function in $\mathcal{P}_4$ monotonically decreases with $C_{b_k}^m$. Solving the condition (4.33), we obtain the following optimal $C_{b_k}^m$ as

$$\left(C_{b_k}^m\right)^* = \log\left(\frac{a + 2 + \sqrt{a^2 + 4a}}{2} \right), \tag{4.35}$$

where

$$a = \frac{u_{b_k}^{m*} u_{b_k}^m \ln 2}{\mu} \cdot \frac{\beta_{b_k}^m}{w_{b_k}^m} \tag{4.36}$$

and μ is chosen such that (4.34) is satisfied.

We see the parameter a here represents a quantitative and comprehensive description of the cellular observation strength and the D2D link quality of relay node m. Especially, the ratio $\frac{\beta_{b_k}^m}{w_{b_k}^m}$ is the ultimate parameter when allocating D2D resources. Recall that $\beta_{b_k}^m$ represents the cellular rate contribution while $w_{b_k}^m$ indicates the D2D link quality. This water-filling like solution means that we should always allocate more time to relay nodes with strong cellular observations and good D2D link qualities.

Remark 4.2. *With (4.22) and (4.23), we have*

$$t_{b_k}^m = t_d w_{b_k}^m C_{b_k}^m. \tag{4.37}$$

With the solution for optimal $C_{b_k}^m$ in (4.35), we then notice that the cellular rate contribution parameter $\beta_{b_k}^m$ affects the allocated time duration $t_{b_k}^m$ through a log function. The saturation of log function at a large-value variable indicates that if the received signals at all relay nodes are strong ($\beta_{b_k}^m$ large for all m), the D2D link quality $w_{b_k}^m$ is the main issue to be considered when allocating the D2D time resources within one cluster. On the other hand, when the D2D link qualities of the relay nodes are roughly at the same level within one cluster, a small $t_{b_k}^m$ for relay node m means that the rate contribution from relay node m to the destination b_k is rather negligible. In this case, we can remove relay node m to save its power. Thus, this solution can be used as a criterion of selecting relay nodes when determining user clusters.

4.3.2 Joint Optimization of Tx. Beamformer and Rx. Beamformer

With the optimized D2D allocation and corresponding quantization noise variances at relay nodes, we then consider the beamformer optimization for our proposed scheme as follows.

$$\mathcal{P}_5: \quad \max_{\mathbf{v}_{b_k}, \mathbf{u}_{b_k}} \sum_{b=1}^{N} \sum_{k=1}^{K} R_{b_k} \quad \text{s.t.} \quad \text{C3.} \tag{4.38}$$

We note that such a sum-rate maximization problem has been intensely addressed in many literatures. Especially, if considering coordinated beamforming at different BSs, a weighted minimum mean square error (WMMSE) approach has been proposed to solve a similar problem [135] and shown great effectiveness though numerical experiments. However, this method has a high computation complexity when considering a large multi-cell model with multiple antennas at each user. Since our focus here is to study the benefits of user cooperations in the cellular network, we do not consider high-complexity algorithms for BS cooperations in this thesis.

Instead, we assume a simple situation that each BS performs random scheduling and zero-forcing beamforming for the active users within its own cell. Under this

assumption, the benefits of user cooperations through D2D links alone are examined and advanced BS cooperations can be further carried out on top of our proposed scheme.

Given the considered non-cooperative BSs, we adopt a random scheduling strategy at each BS as follows. In each cell, S out of the K active users are randomly scheduled and all K users must have been scheduled once after $\lceil \frac{K}{S} \rceil$ frame durations[3]. Denote by $\mathcal{S}_{b,t}$ the set of scheduled users in the cell of BS b at time t, $t \in \{1, , \cdots, \lceil \frac{K}{S} \rceil\}$. Then, problem $\mathcal{P}_5$ becomes

$$\mathcal{P}_6 : \quad \begin{array}{l} \forall t, \ \max\limits_{\mathbf{v}_{b_k}, \mathbf{u}_{b_k}} \sum\limits_{b=1}^{N} \sum\limits_{k=1}^{K} R_{b_k} \\ \text{s.t.} \quad \begin{cases} \text{C3}, \\ \forall b_k \notin \mathcal{S}_{b,t}, \mathbf{v}_{b_k} = \mathbf{0}. \end{cases} \end{array}$$

Then, the actual rate of user b_k is

$$\overline{R}_{b_k} = \frac{S}{K} R_{b_k}, \tag{4.39}$$

since each user is served every $\frac{K}{S}$ frame durations.

For each scheduled user b_s, fixing all the transmit beamformers and minimizing the MSE, we obtain the well-known MMSE receiver:

$$\mathbf{u}_{b_k} = \mathbf{J}_{b_k}^{-1} \mathbf{H}_{b,b_k} \mathbf{v}_{b_k}. \tag{4.40}$$

Then, the equivalent channel from BS b to user b_k can be seen as $\widetilde{\mathbf{h}}_{b_k} = \left(\mathbf{u}_{b_k}^* \mathbf{H}_{b,b_k}\right)^T$. Collecting all $\widetilde{\mathbf{h}}_{b_k}$ for scheduled users of BS b at time-slot t, we have $\widetilde{\mathbf{H}}_b = \left[\cdots, \widetilde{\mathbf{h}}_{b_k}, \cdots\right]^T$. Zero-forcing transmit beamforming can be easily applied using the pseudo inverse of $\widetilde{\mathbf{H}}_b$ [136], i.e.,

$$\mathbf{V}_b = \widetilde{\mathbf{H}}_b^{\dagger} = \widetilde{\mathbf{H}}_b^* \left(\widetilde{\mathbf{H}}_b \widetilde{\mathbf{H}}_b^*\right)^{-1}. \tag{4.41}$$

We then propose an iterative approach for solving $\mathcal{P}_1$. Specifically, we first find the optimal quantization noise covariance q_{b_k} for given beamformers using the solution in Section 4.3, and then find the optimal $\mathbf{u}_{b_k}$ for given transmit beamformers and

[3]Note that $\frac{K}{S}$ is usually set to be an integer number for simplicity.

Algorithm 4.1 Resource Allocation Algorithm

1: Initialization:

2: Generate the random user scheduling sets: $\forall b$, randomly generate $\frac{K}{S}$ integer sets $\mathcal{S}_{b,t}$, $t \in \left\{1, \cdots, \frac{K}{S}\right\}$ such that $\underset{t}{\cup}\mathcal{S}_{b,t} = \{1, \cdots, K\}$ and $\mathcal{S}_{b,t_1} \cap \mathcal{S}_{b,t_2}$, for any $t_1 \neq t_2$.

3: Set $\mathbf{v}_{b_k}$ such that $\mathbf{Tr}\left(\mathbf{v}_{b_k}\mathbf{v}_{b_k}^*\right) = \frac{P_B}{S}$ and $\mathbf{v}_{b_k}^T\mathbf{v}_{b_j} = 0$, $\forall b, k, j \neq k$.

4: Let $\mathbf{Q}_{b_k} = \mathbf{0}_M$, compute the MMSE receiver $\mathbf{u}_{b_k}$ according to (4.40), $\forall b, k$.

5: **for** $t \leftarrow 1$ **to** $t \leftarrow \frac{K}{S}$ **do**

6: **repeat**

7: $\forall b, k \in \mathcal{S}_{b,t}$

8: 1. Compute $\mathbf{Q}_{b_k}$ according to (4.35) and (4.36).

9: 2. Update $\mathbf{u}_{b_k}$ according to (4.40).

10: 3. Compute the achievable rate R_{b_k} according to (4.19).

11: 4. Find the ZF transmit beamformer $\mathbf{v}_{b_k}$ under above $\mathbf{u}_{b_k}$, according to (4.41).

12: **until** convergence

13: **end for**

14: Compute the actual rate for each user: $\overline{R}_{b_k} = \frac{S}{K}R_{b_k}$, $\forall b, k$.

quantization noise covariances using (4.40). Then the transmit beamformer $\mathbf{v}_{b_k}$ is updated based on (4.41). By doing so, we integrate two optimization problems $\mathcal{P}_4$ and $\mathcal{P}_5$, and thus jointly optimize the cellular and D2D resource allocation. We use Algorithm 1 to further illustrate our proposed resource allocation algorithm.

4.4 Numerical Results

In this section, numerical simulations are conducted to show the effectiveness of the proposed algorithms.

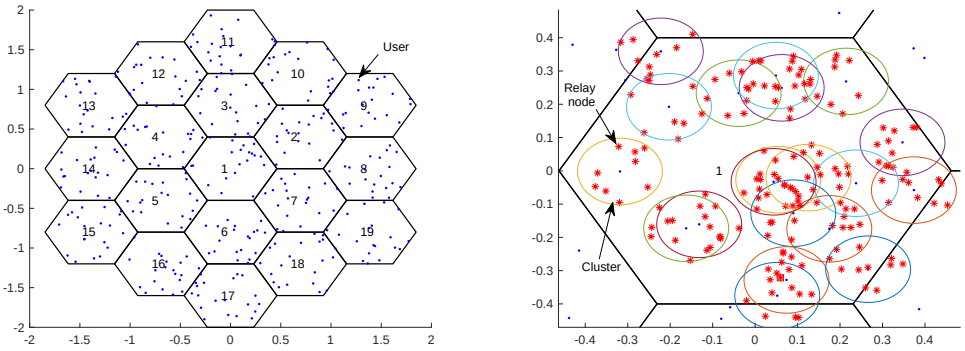
4.4.1 Simulation Setups

To fully demonstrate the inter-cell interference mitigation, we consider a 19-cell wrapped-around network with the simulation parameters listed in Table 4.1. Each cell is a regular hexagon with a single BS located at the center, within which cellular users are randomly distributed as shown in Fig. 4.5a. A circle centered at the each cellular user with a radius of d_{max} illustrate the user clustering as shown in Fig. 4.5b.

The cellular wireless channel is centered at a frequency of 2 GHz and has a bandwidth $B_c = 20$ MHz, following the 3GPP LTE-A standard. The mmWave wireless channel is centered at a frequency of 73 GHz, and has several orthogonal sub-bands of 200 MHz so that an exclusive bandwidth can be occupied by each cluster, which is shared within one cluster via TDMA. The channel from the relay nodes to each cellular user is LoS, with the path loss given by $69.7 + 24log10(d_m)$ dB [133]. In addition, we consider a Nakagami fading with the Nakagami parameter $\alpha = 4$ as assumed in many mmWave D2D works. Each relay node is assumed to transmit at a fixed power of 20 dBm with an antenna gain of 27 dBi [134].

We compare the performance of the following benchmark schemes with our proposed scheme by extensive system-level simulations.

Benchmark scheme 1: No user cooperation. We assume that each single-



(a) 19-cell wrapped around model used in simulations.

(b) User clustering illustrated in one cell.

Figure 4.5. 19-cell wrapped around network. Dots represent the active cellular users and stars stand for the relay nodes. Each active cellular user and its relay nodes are circled out by an oval, representing one cluster.

Table 4.1. SIMULATION PARAMETERS.

Cellular Layout	Hexagonal 19-cell wrapped-around
Cellular bandwidth	20 MHz
Cellular frame duration	1.25 ms
D2D bandwidth for one cluster	200 MHz
Distance between cells	0.8 km
Max. D2D transmission range	100 m
Num. of users	20
Max. Tx power for BSs	43 dBm
Max. Tx power for Relay nodes	20 dBm [134]
Cellular Antenna gain	15 dBi
D2D Antenna gain(mmWave)	27 dBi [134]
Background noise	169dBm/Hz
Path loss from BS to user	$128.1 + 37.6log10(d_{km})$ dB
Path loss of D2D link	$69.7 + 24log10(d_m)$ dB [133]
Log-normal shadowing	8 dB
Rayleigh small scale fading	0 dB
Nakagami parameter of D2D links	4

antenna user has no relay node. This can be regarded as the performance baseline. If the rate performance of our proposed scheme is even worse than this benchmark, there is no need to perform user cooperation as proposed.

Benchmark scheme 2: Ideal user cooperation. Consider an ideal case of infinite D2D link capacity, which is obtained by assuming no quantization noise at the relay nodes, i.e., fixing $\mathbf{Q}_{b,k} = \mathbf{0}_M$. This can be seen as a performance upper bound of our proposed scheme. In this case, since an infinite D2D link capacity is assumed, having $M-1$ relay nodes for each cellular user is equivalent to that each user is equipped with M geographically dispersed receive antennas.

Benchmark scheme 3: Equal D2D resource allocation. Consider an equal D2D resource allocation, where each relay node occupies a same length of transmission time, i.e.,

$$t_{b_k}^m = \frac{t_d}{M-1}. \tag{4.42}$$

Benchmark scheme 4: Multi-hop D2D cooperation. Consider a multi-hop D2D cooperation scheme as in many literatures [112] [113]. Within one cluster, the relay node with the strongest channel will decode the message from BS and forward it to the user through the D2D connection. We assume that the cellular and D2D transmissions happen consecutively with a negligible time delay caused by establishing the relay link. Also, the D2D link capacity is much larger than the cellular transmission rate. Thus, the achievable rate of user b_k is the maximum rate within its cluster, i.e.,

$$R_{b_k}^{\mathbf{Multi-hop}} = \max\left\{c_{b_k}, \max_m c_{b_k}^m\right\}, \tag{4.43}$$

where

$$c_{b_k} = \log\left(1 + \frac{\mathbf{h}_{b,b_k}^T \mathbf{v}_{b_k} \mathbf{v}_{b_k}^* \left(\mathbf{h}_{b,b_k}^T\right)^*}{\sum\limits_{\substack{j=1 \\ j\neq k}}^{K} \mathbf{h}_{b,b_k}^T \mathbf{v}_{b_j} \mathbf{v}_{b_j}^* \left(\mathbf{h}_{b,b_k}^T\right)^* + \sum\limits_{\substack{i=1 \\ i\neq b}}^{N} \sum\limits_{\substack{j=1 \\ j\neq k}}^{K} \mathbf{h}_{i,b_k}^T \mathbf{v}_{i_j} \mathbf{v}_{i_j}^* \left(\mathbf{h}_{i,b_k}^T\right)^* + \sigma^2}\right), \tag{4.44}$$

and

$$c_{b_k}^m = \log\left(1 + \frac{\left(\mathbf{h}_{b,b_k}^m\right)^T \mathbf{v}_{b_k}\mathbf{v}_{b_k}^* \left(\left(\mathbf{h}_{b,b_k}^m\right)^T\right)^*}{\sum\limits_{\substack{j=1\\ j\neq k}}^{K}\left(\mathbf{h}_{b,b_k}^m\right)^T \mathbf{v}_{b_j}\mathbf{v}_{b_j}^* \left(\left(\mathbf{h}_{b,b_k}^m\right)^T\right)^* + \sum\limits_{\substack{i=1\\ i\neq b}}^{N}\sum\limits_{\substack{j=1\\ j\neq k}}^{K}\left(\mathbf{h}_{i,b_k}^m\right)^T \mathbf{v}_{i_j}\mathbf{v}_{i_j}^* \left(\left(\mathbf{h}_{i,b_k}^m\right)^T\right)^* + \sigma^2}\right).$$

For above benchmark schemes and our proposed scheme, Algorithm 1[4] is executed over T different channel realizations to measure the cellular rate performance by the averaged long-term rate. We have $T = 100$ for the shown results in this thesis.

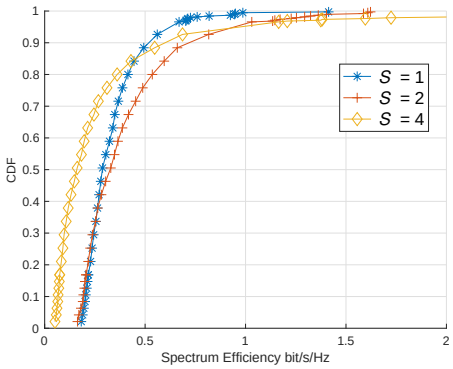
4.4.2 Simulation Results

How much performance gain can be achieved considering multiple antennas at end-users?

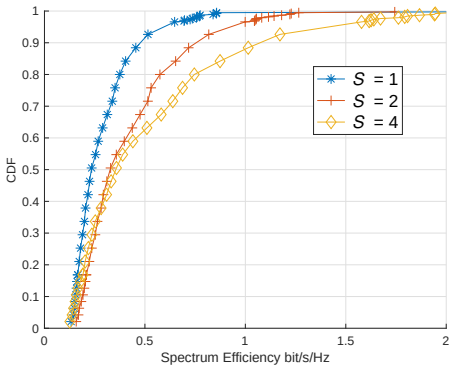
In order to examine whether our proposed framework has stupendous potential under practical system settings, we compare Benchmark scheme 1 and Benchmark scheme 2 to check how much rate improvement can be achieved under infinite D2D link capacity.

We first examine the effects of the number of scheduled users S on the rate performance. Fig. 4.6 shows the cumulative distributions of the long-term average user rates, with various number of receive antennas M and number of scheduled users S when the number of transmit antennas $L = 4$. Fig. 4.7 shows the cumulative distributions with the same parameters except for $L = 10$. We notice that the case $M = 1$ stands for the Benchmark scheme 1 without user cooperations. Comparing Fig. 4.6a - Fig. 4.6d, we observe that when the number of receive antennas M is limited, the relationship of rate performance and the number of scheduled users S is not straightforward. For example, as shown in Fig. 4.6a, when $L = 4$ and there is no relay node, scheduling 2 users at a time ($S = 2$) achieves the best performance. While

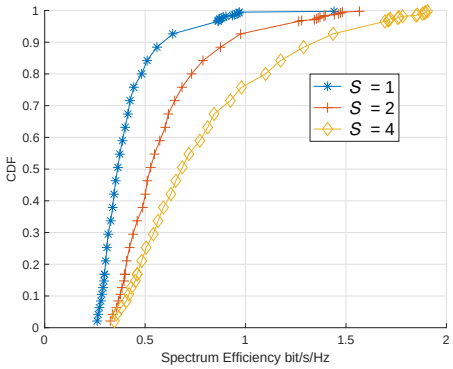
[4]If Benchmark scheme 1 and 4 is considered, omit the step 1 within each iteration.



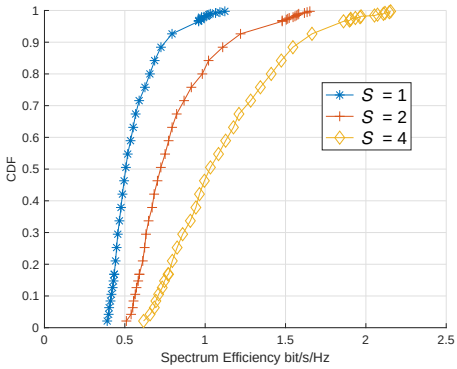
(a) $M = 1$, Benchmark 1 (No user cooperation).



(b) $M = 2$, Benchmark 2 (Ideal cooperation).

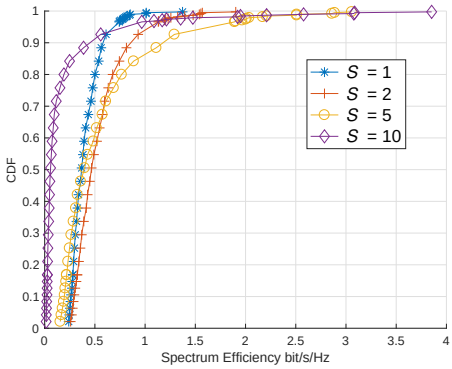


(c) $M = 5$, Benchmark 2 (Ideal cooperation).

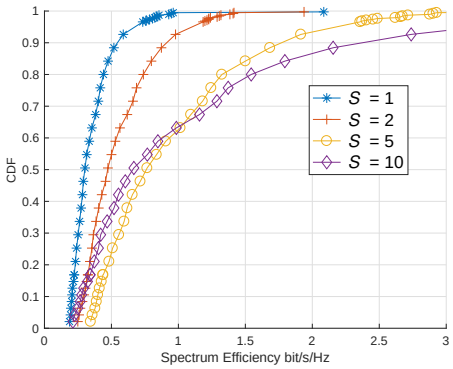


(d) $M = 10$, Benchmark 2 (Ideal cooperation).

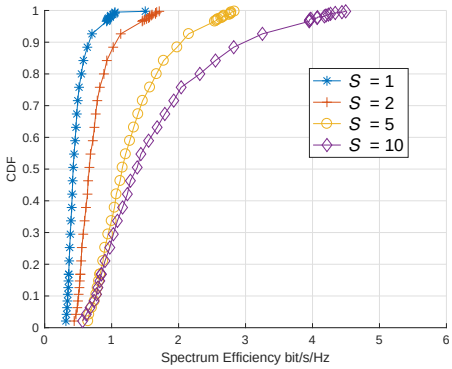
Figure 4.6. Cumulative distribution function of user data rate comparison with different number of scheduled users S, the number of transmit antennas at each BS $L = 4$, 19-cell wrapped-around.



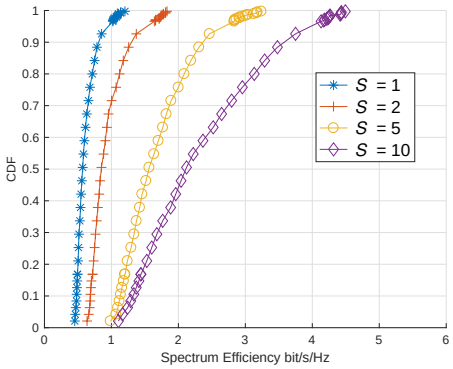
(a) $M = 1$, Benchmark 1 (No user cooperation).



(b) $M = 2$, Benchmark 2 (Ideal cooperation).

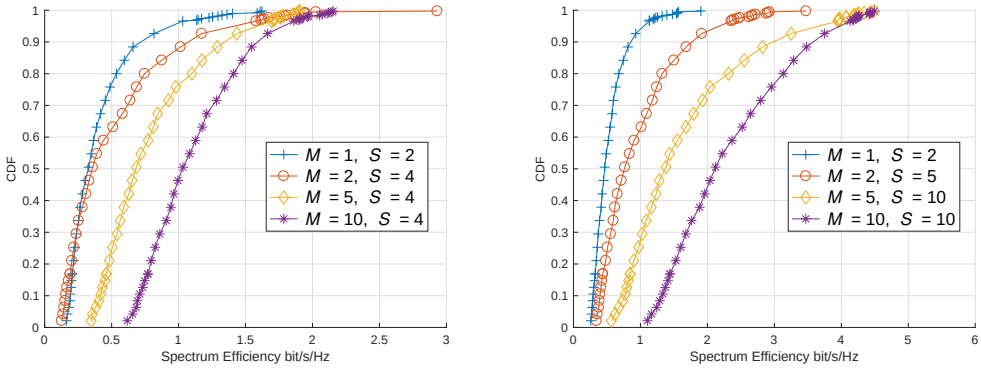


(c) $M = 5$, Benchmark 2 (Ideal cooperation).



(d) $M = 10$, Benchmark 2 (Ideal cooperation).

Figure 4.7. Cumulative distribution function of user data rate comparison with different number of scheduled users S, the number of transmit antennas at each BS $L = 10$, 19-cell wrapped-around.



(a) $L = 4$, Benchmark 2 (Ideal cooperation).

(b) $L = 10$, Benchmark 2 (Ideal cooperation).

Figure 4.8. Cumulative distribution function of user data rate comparison with different number of received antennas M.

with one relay node at each user, Fig. 4.6b shows that the case of $S = 4$ achieves the best performance.

However, it is interesting to note that when the number of receive antennas M exceeds a certain threshold, full loading the BS, i.e., setting the number of scheduled users S to be equal to the number of BS transmit antennas L, is always the optimal strategy. For instance, we see that full loading achieves the best SE when $M \geq 2$ in Fig. 4.6 and when $M \geq 5$ in Fig. 4.7. This is a practical insight for designing user scheduling of future multiple antenna systems.

Next, we eliminate the influence of the user scheduling and focus on examining the benefits of user cooperations. That is, for given numbers of transmit and receive antennas L, M, we pick up the cases with the best rate performance and then compare their SE in Fig. 4.8a and Fig. 4.8b, respectively. We see that both figures show great improvement in terms of rate performance of the cellular users as the number of receive antennas M increases. For example, in Fig. 4.8b, when $L = 10$, an improvement up to 4.5x is observed with $M = 10$ for the 50th percentile users compared with the case of $M = 1$.

We list the rate improvement for the 10th and 50th percentile users when $L = 4$ in Table 4.2. A huge rate improvement is observed for both percentile users and

the rate improvement increases as the number of receive antennas M increases. For instance, when $L = 4$, compared with the non-cooperative Benchmark scheme 1, for the 10th percentile user, we obtain a 117.7% rate improvement when $M = 5$, while 279.2% more rate is achieved when $M = 10$. Due to different user path loss, the rate improvement is not linear with the number of receive antennas M as indicated by MIMO theory. However, the improvement brought by this D2D cellular framework is still significant and rather useful for encountering inter-cell interferences. In addition, we also note that this improvement for low-rate users is especially meaningful since it largely enhances the performance of cell-edge users and thus improves user experience.

Table 4.2. User data rate comparisons, $L = 4$.

Scheme	10th Percentile Rate (Mbps)	Gain	50th Percentile Rate (Mbps)	Gain
Benchmark 1 (No-user-cooperation)	3.85	N/A	6.58	N/A
Benchmark 3 (Ideal - 1 relay node)	3.17	-17.7%	7.10	7.9%
Benchmark 3 (Ideal - 4 relay nodes)	8.38	117.7%	13.70	108.2%
Proposed (4 relay nodes)	6.25	62.3%	10.78	63.9%
Benchmark 3 (Ideal - 9 relay nodes)	14.60	279.2%	20.58	212.8%
Proposed (9 relay nodes)	9.69	151.7%	14.54	121.0%

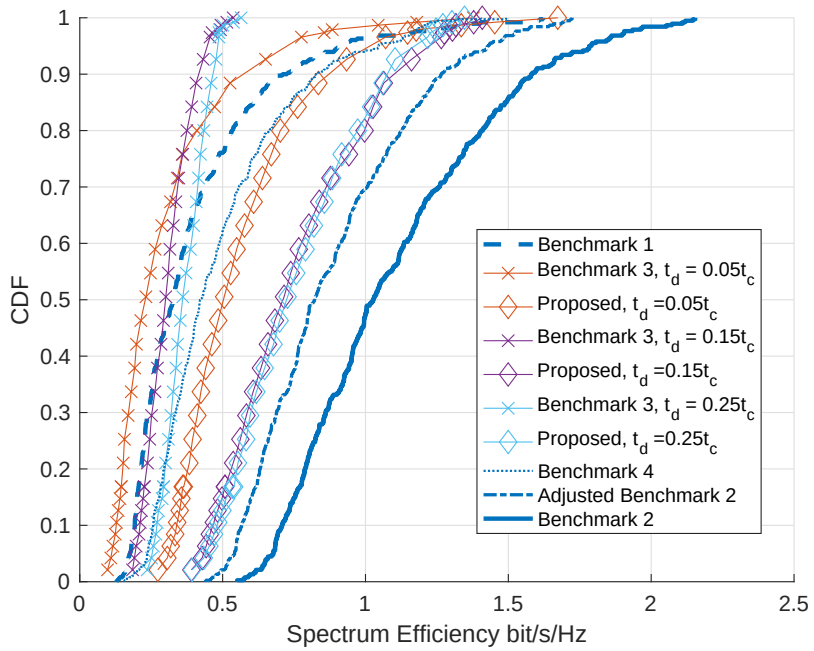


Figure 4.9. Cumulative distribution function of user data rate comparison with Benchmark 3, $L = 4$, $M = 10$.

The effectiveness of our proposed resource allocation

Next, we consider the physical limits of D2D links and examine the effectiveness of our proposed resource allocation algorithm. Fig. 4.9 and Fig. 4.10 compare the cumulative distributions of the long-term average user rates for the Benchmark scheme 3 (equal allocation), the Benchmark scheme 4 (multi-hop) and our proposed algorithm. We also show the curve of Benchmark 1 as a baseline and that of Benchmark 2 as an upper bound in the figures.

Both figures demonstrate an impressive enhancement brought by our proposed allocation algorithm. For example, if using the equal allocation strategy, the rate performance shows a rather limited improvement. What's worse, as shown in Fig. 4.9, when the D2D transmission time $t_d = 0.05t_c$ (t_c is the cellular frame duration), the rate might be even degraded since the small rate increase brought by the D2D links cannot compensate the consumed time duration for D2D links. On the other hand, we can see that our optimized allocation guarantees an improvement of user

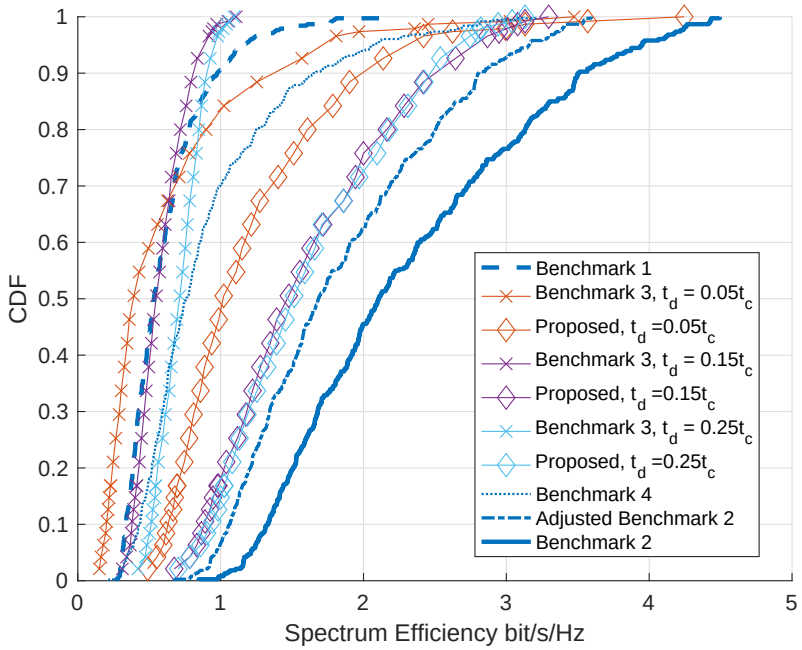


Figure 4.10. Cumulative distribution function of user data rate comparison with Benchmark 3, $L = 10$, $M = 10$.

rate under different D2D transmission time t_d. We also observe that our proposed algorithm, which utilizes a diversity gain, largely outperforms the Benchmark scheme 4 (multi-hop D2D cooperation) which achieves a power gain under various system setups.

We also list the improvement for the 10th and 50th percentile users when $L = 4$ in Table 4.2. Similarly, we see that a huge rate improvement is observed for both percentile users and the rate improvement increases as the number of relay nodes $M - 1$ increases. Especially, a 2.5x improvement in terms of the 10th percentile user rate is observed when each cellular user is helped by 9 standby users. It is interesting to observe that the optimized D2D allocation for $t_d = 0.15t_c$ and $t_d = 0.25t_c$ achieves almost the same rate performance. This can be interpreted as that the rate increase brought by better D2D connections from $t_d = 0.15t_c$ and $t_d = 0.25t_c$ is cancelled by the increasing consumed time duration on D2D links.

We further notice that although there seems a big gap from the rate performance

of our optimized allocation to that of Benchmark 2, our curve is already close to the ideal case if only assuming that the D2D link capacity is infinite. This is because the gap here mainly comes from the average over the total transmission time $(t_d + t_c)$ when calculating the rate, while in Benchmark 2 we have $t_d = 0$. If we also multiply the ratio $\frac{t_c}{t_d+t_c}$ to the rate of Benchmark 2 (shown as "Adjusted Benchmark 2" in Fig. 4.9 and Fig. 4.10), we can see that the rate of our proposed method $t_d = 0.25t_c$ approaches that of Benchmark 2.

We would like to point out that the CSI from the BS to relay nodes is essential in our scheme. This is because without the relay CSI, the transmit beamformer is still designed based on a single receive antenna model, and thus the interference cannot be effectively managed. Numerical results have also confirmed this and we do not show here due to the length limit.

4.5 Conclusion

In this chapter, we exploited the D2D communication capability of standby users and proposed a hybrid D2D-cellular scheme applying the distributed MIMO technology to improve the data rate and user experience for cellular users. More specifically, a virtual antenna array was formed by sharing antennas across different terminals to realize the diversity gain of MIMO systems. To achieve this, mmWave D2D links were considered to enable high data rate D2D communications. We then designed a D2D multiple access protocol and formulated the optimization problem of joint cellular and D2D resource allocation for downlink transmissions of our proposed scheme. We obtained a closed-form solution for the D2D resource allocation, which revealed useful insights for D2D resource allocation. Extensive system-level simulations were performed to demonstrate the effectiveness of the proposed scheme. Extensive system-level simulations demonstrated that the rate performance of cellular users is significantly improved through our proposed scheme and the resource allocation algorithm.

Chapter 5

Conclusion

This thesis studied advanced MIMO technology to exploit the degrees of freedom gain under a variety of proposals for future wireless networks.

First, a new linear vector PNC scheme for a spatial multiplexing MIMO TWRC where CSIT was not available are proposed in Chapter 2. Then, a new caching scheme for a random wireless device-to-device (D2D) network was proposed in Chapter 3, where each node is equipped with a local cache and intends to download files from a prefixed library via D2D links. Last but not least, a hybrid D2D-cellular scheme is proposed to make use of the standby users who possess D2D communication capabilities in close proximity to each other, and to improve the rate performance for cellular users in Chapter 5. For the proposed three schemes, an optimized design and corresponding performance analysis were presented respectively. In addition, a great performance (symbol error rate, network throughput and average rate) improvement of the proposed scheme and design were demonstrated through both analytical and numerical results.

The author would like to conclude that although the topic of MIMO had been the research focus during the last decade and seemed "old-fashioned" nowadays, the tremendous "capacity boosting" capability of MIMO technology still makes it appealing in many proposals and will continue to play an important role in future wireless

networks.

Bibliography

[1] Cisco, Visual Networking Index, Feb. 2014, white paper at Cisco.com.

[2] A. Maeder, P. Rost, and D. Staehle, "The challenge of M2M communications for the cellular radio access network," in *Proc. Wrzburg Workshop IP, Joint ITG Euro-NF Workshop Vis. Future Gener. Netw. EuroView*, pp. 12, Aug. 2011.

[3] S. Hilton, "Machine-to-Machine Device Connections: Worldwide Forecast 20102020, Analysys Mason Report," 2010.

[4] "The path to 5G: as much evolution as revolution, " 3GPP Report, 2016.

[5] "Evolving LTE to fit the 5G future, " Ericsson Technology Review, 2017.

[6] "5G Network Architecture: A High-Level Perspective, " Whitepaper, Huawei Technologies co., ltd., 2016.

[7] D. Tse and P. Visanath, *Fundamentals of wireless communicationts*, Cambridge University Press, 2006.

[8] J. G. Andrews et al., "What Will 5G Be?," *IEEE Journal on Selected Areas in Communications*, vol. 32, no. 6, pp. 1065-1082, June 2014.

[9] D. Gunduz, A. Yener, A. Goldsmith, and H. V. Poor, "The multi-way relay channel," in *Proc. of IEEE ISIT*, Seoul, South Korea, Jun. 2009.

[10] S. Zhang, S. Liew, and P. Lam, "Physical-layer network coding," in ACM Mobicom, 2006.

[11] P. Popovski and H. Yomo, "Physical Network Coding in Two-Way Wireless Relay Channels," IEEE International Conference on Communications, Glasgow, pp. 707-712, 2007.

[12] T. Koike-Akino, P. Popovski, and V. Tarokh, "Optimized constellations for two-way wireless relaying with physical network coding," *IEEE J. Sel. Area. Comm.*, vol. 27, no. 5, pp. 773-787, June 2009.

[13] Xiaojun Yuan, "MIMO multiway relaying with clustered full data exchange: Signal space alignment and degrees of freedom," *IEEE Trans. on Wireless Commun.*, vol. 13, no. 12, pp. 6795-6808, Dec. 2014.

[14] W. Nam, S. Chung, Y. H. Lee, "Capacity of the Gaussian two-way relay channel to within 1/2 bit," *IEEE Trans. Inf. Theory*, vol. 56, no. 11, pp. 5488-5494, Nov. 2010.

[15] M. P. Wilson, K. Narayanan, H. D. Pfister and A. Sprintson, "Joint physical layer coding and network coding for bidirectional relaying," *IEEE Trans. Inf. Theory*, vol. 56, no. 11, pp. 5641-5654, Nov. 2010.

[16] H. J. Yang, J. Chun, and A. Paulraj, "Asymptotic capacity of the separated MIMO two-way relay channel," *IEEE Trans. Inf. Theory*, vol.57, no. 11, pp. 7542-7554, Nov. 2011.

[17] K. Poularakis, G. Iosifidis, V. Sourlas and L. Tassiulas, "Exploiting Caching and Multicast for 5G Wireless Networks," *Wireless Communications, IEEE Transactions on*, vol.PP, no.99, pp.1-1, 2016.

[18] M. Tao, E. Chen, H. Zhou and W. Yu, "Content-Centric Sparse Multicast Beamforming for Cache-Enabled Cloud RAN," *Wireless Communications, IEEE Transactions on*, vol.PP, no.99, pp.1-1, 2016.

[19] H. Zhou, M. Tao, E. Chen and W. Yu, "Content-Centric Multicast Beamforming in Cache-Enabled Cloud Radio Access Networks," *IEEE Global Communications Conference (GLOBECOM)*, San Diego, CA, pp. 1-6, 2015.

[20] Binbin Dai and Wei Yu, "Joint user association and content placement for Cache-enabled wireless access networks," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3521-3525, Shanghai, 2016.

[21] A. Liu and V. K. N. Lau, "Asymptotic Scaling Laws of Wireless Ad Hoc Network With Physical Layer Caching," *Wireless Communications, IEEE Transactions on*, vol. 15, no. 3, pp. 1657-1664, March 2016.

[22] K. Doppler, M. Rinne, C. Wijting, C. B. Ribeiro and K. Hugl, "Device-to-device communication as an underlay to LTE-advanced networks," *IEEE Communications Magazine*, vol. 47, no. 12, pp. 42-49, Dec. 2009.

[23] D. Feng, L. Lu, Y. Yuan-Wu, G. Y. Li, S. Li and G. Feng, "Device-to-device communications in cellular networks," *IEEE Communications Magazine*, vol. 52, no. 4, pp. 49-55, April 2014.

[24] X. Lin, J. G. Andrews, A. Ghosh and R. Ratasuk, "An overview of 3GPP device-to-device proximity services," *IEEE Communications Magazine*, vol. 52, no. 4, pp. 40-48, April 2014.

[25] A. Kaye and D. George, "Transmission of multiplexed PAM signals over multiple channel and diversity systems," *IEEE Transactions on Communication Technology*, vol. 18, no. 5, pp. 520-526, Oct. 1970.

[26] L. H. Brandenburg and A. D. Wyner, "Capacity of the Gaussian channel with memory: The multivariate case," *The Bell System Technical Journal*, vol. 53, no. 5, pp. 745-778, May-June 1974.

[27] R. Roy, A. Paulraj, and T. Kailath, "Method for estimating signal source locations and signal parameters using an array of signal sensor pairs," *U.S. Patent, 4 750 147*, June 7, 1988.

[28] B. Ottersten, "Asymptotic Analysis of SVD-Based Array Processing Algorithms," *SIAM Conference on Linear Algebra in Signals, Systems and Control*, San Francisco, CA, Nov. 1990.

[29] G. G. Raleigh and J. M. Cioffi, "Spatio-temporal coding for wireless communication," *IEEE Transactions on Communications*, vol. 46, no. 3, pp. 357-366, Mar 1998.

[30] G. J. Foschini, "Layered space-time architecture for wireless communication in a fading environment when using multi-element antennas," *Bell Labs Tech. J.*, vol. 1, no. 2, pp. 41-59, 1996.

[31] A. M. Sayeed, "Deconstructing multiantenna fading channels," *IEEE Transactions on Signal Processing*, vol. 50, no. 10, pp. 2563-2579, Oct 2002.

[32] R. U. Nabar, H. Bolcskei and F. W. Kneubuhler, "Fading relay channels: performance limits and space-time signal design," *IEEE Journal on Selected Areas in Communications*, vol. 22, no. 6, pp. 1099-1109, Aug. 2004.

[33] P. Popovski and H. Yomo, "Wireless network coding by amplify-and-forward for bi-directional traffic flows," *IEEE Communications Letters*, vol. 11, no. 1, pp. 16-18, Jan. 2007.

[34] J. Joung and A. H. Sayed, "Multiuser Two-Way Amplify-and-Forward Relay Processing and Power Control Methods for Beamforming Systems," *IEEE Transactions on Signal Processing*, vol. 58, no. 3, pp. 1833-1846, March 2010.

[35] M. Katz and S. Shamai, "Relaying protocols for two colocated users," *IEEE Transactions on Information Theory*, vol. 52, no. 6, pp. 2329-2344, June 2006.

[36] S. Yao, T. T. Kim, M. Skoglund and H. V. Poor, "Half-Duplex Relaying Over Slow Fading Channels Based on Quantize-and-Forward," *IEEE Transactions on Information Theory*, vol. 59, no. 2, pp. 860-872, Feb. 2013.

[37] T. Wang, A. Cano, G. B. Giannakis and J. N. Laneman, "High-Performance Cooperative Demodulation With Decode-and-Forward Relays," *IEEE Transactions on Communications*, vol. 55, no. 7, pp. 1427-1438, July 2007.

[38] B. Nazer and M. Gastpar, "Compute-and-forward: harnessing interference with structured codes," *IEEE Trans. Inform. Theory*, vol. 57, no. 10, pp. 6463-6486, Oct. 2011.

[39] B. Dai and W. Yu, "Sparse Beamforming and User-Centric Clustering for Downlink Cloud Radio Access Network," *IEEE Access*, vol. 2, pp. 1326-1339, 2014.

[40] T. Yang, Q. T. Sun, J. A. Zhang and J. Yuan, "A Linear Network Coding Approach for Uplink Distributed MIMO Systems: Protocol and Outage Behavior," *IEEE Journal on Selected Areas in Communications*, vol. 33, no. 2, pp. 250-263, Feb. 2015.

[41] Zhuo Sun, Lei Yang, Jinhong Yuan and M. Chiani, "A novel detection algorithm for random multiple access based on physical-layer network coding," *IEEE International Conference on Communications Workshops (ICC)*, p. 608-613, 2016.

[42] F. Dong, A. Burr, J. Yuan, "Linear Physical-Layer Network Coding Over Hybrid Finite Ring for Rayleigh Fading Two-Way Relay Channels," *IEEE Transactions on Communications*, vol. 62, no. 9, pp. 3249-3261, Sept. 2014.

[43] T. Yang and I. B. Collings, "On the optimal design and performance of linear physical-layer network coding for fading two-way relay channels," *IEEE Trans. Wireless Comm.*, vol. 13, no. 2, pp. 956-967, Feb. 2014.

[44] V.T. Muralidharan, and B.S. Rajan, "Wireless Network Coding for MIMO Two-Way Relaying," *IEEE Trans. Wireless Comm.*, vol. 7, no.11, pp. 3566-3577, Jul. 2013.

[45] G. Hui, L. Tiejun, Z. Shengli, Y. Chau, and Y. Shaoshi, "Zero-Forcing Based MIMO Two-Way Relay with Relay Antenna Selection: Transmission Scheme and Diversity Analysis," *IEEE Trans. Wireless Comm.*, vol. 11, pp. 4426-4437, Dec. 2012.

[46] X. Yuan, T. Yang and I. B. Collings, "MIMO two-way relaying: a space division approach," *IEEE Trans. Inf. Theory*, vol. 59, no. 10, pp. 6421-6440, Oct. 2013.

[47] K. Young-Tae, L. Kwangwon, P. Moonseo, L. Kyoung-Jae, and L. Inkyu, "Precoding Designs Based on Minimum Distance for Two-Way Relaying MIMO Systems with Physical Network Coding," *IEEE Trans. Comm.*, vol. 61, pp. 4151-4160, Oct. 2013.

[48] Z. Ding and H. V. Poor, "A General Framework of Precoding Design for Multiple Two-Way Relaying Communications," *IEEE Trans. Signal Process.*, vol. 61, pp. 1531-1535, Mar. 2013.

[49] K. Lee, N. Lee and I. Lee, "Achievable Degrees of Freedom on MIMO Two-way Relay Interference Channels," *IEEE Trans. Wireless Comm.*, vol. 12, pp. 1472-1480, Apr. 2013.

[50] T. Yang, X. Yuan, P. Li, I. B. Collings and J. Yuan, "A new physical-layer network coding scheme with eigen-direction alignment precoding for MIMO two-way relaying," *IEEE Trans. Comm.*, vol. 61, no. 3, pp. 973-986, Mar. 2013.

[51] C. Li, L. Yang, and W. Zhu, "Two-Way MIMO Relay Precoder Design with Channel State Information," *IEEE Trans. Comm.*, vol. 58, pp. 3358-3363, Dec. 2010.

[52] L. Lu and S. C. Liew, "Asynchronous physical-layer network coding," *IEEE Trans. Wireless Comm.*, vol. 11, no. 2, pp. 819-831, Feb. 2012.

[53] C. Feng, D. Silva and F. R .Kschischang, "An algebraic approach to physical-layer network coding," in *Proc. IEEE ISIT*, 2010.

[54] T. Huang, T. Yang, J. Yuan, and I. Land, "Design of irregular repeat-accumulate coded physical-layer network coding for Gaussian two-way relay channels," *IEEE Trans. Comm.*, vol. 61, no. 3, pp.897-909, Feb. 2013.

[55] M. Qiu, L. Yang and J. Yuan, "Irregular Repeat-Accumulate Lattice Network Codes for Two-Way Relay Channels," *IEEE Global Communications Conference (GLOBECOM)*, 2016.

[56] T. Yang and I. B. Collings, "Asymptotically optimal error-rate performance of linear physical-layer network coding in Rayleigh fading two-way relay channels," *IEEE Comm. Lett.*, vol. 16, no. 7, pp. 1068-1071, Jul. 2012.

[57] T. Yang and I. B. Collings, "Design criterion of linear physical-layer network coding for fading two-way relay channels," *Proc. IEEE Intern. Conf. Comm.*, pp. 3302-3306, 2013.

[58] D. Hwang, S. Hong, and T. Lee, "Multiuser Two Way Relaying Schemes in the Future Cellular Network," *IEEE Trans. Wireless Comm.*, vol. 12, pp. 5200-5207, Otc. 2013.

[59] T. Koike-Akino, "Adaptive Network Coding in Two-Way Relaying MIMO Systems," *IEEE GLOBECOM*, 2010.

[60] S. Zhang, C. Nie, L. Lu, S. Zhang, and G. Qian, "MIMO physical layer network coding based on VBLAST detection," *Wireless Communications and Signal Processing (WCSP), International Conference on*, 2012.

[61] S. Zhang and S. Liew, "Physical Layer Network Coding with Multiple Antennas," *Wireless Communications and Networking Conference (WCNC), IEEE*, 2010.

[62] J. Zhan, B. Nazer and U. Erez and M. Gastpar, "Integer-forcing linear receivers," in *Proc. IEEE ISIT*, 2010.

[63] A. S. Avestimehr, A. Sezgin, and D. Tse, "Capacity of the two-way relay channel within a constant gap," *Eur. Trans. Telecommun.*, vol. 21, no. 4, pp. 363-374, Jun. 2010.

[64] Xiaochen Lin, Meixia Tao, Youyun Xu and Rui Wang, "Outage probability and finite-SNR diversity-multiplexing tradeoff for two-way relay fading channels," *IEEE Trans. on Vehicular Technology*, vol. 62, no. 7, pp. 3123-3136, Sept. 2013.

[65] Q. Li, H. Li, G. Wu and S. Li, "Retrospective Network Coding Alignment Over K-User MIMO Y Channel," *IEEE Communications Letters*, vol. 20, no. 3, pp. 502-505, March 2016.

[66] C. E. Shannon, "Two-way communication channels," in *Proc. Berkeley Symp. Math. Statistics Probability*, vol. 1, pp. 611-644, 1961.

[67] B. Rankov and A. Wittneben, "Spectral efficient signaling for halfduplex relay channels," in *Proc. Conf. Rec. 39th Asilomar Conf. Signals, Syst., Comput.*, Pacific Grove, CA, USA, pp. 1066-1071, Oct./Nov. 2005.

[68] L. Ong, C. Kellett, and S. Johnson, "Capacity theorems for the AWGN multiway relay channel," in *Proc. of IEEE ISIT*, 2010.

[69] N. Lee and J.-B. Lim, "A novel signaling for communication on MIMO Y channel: Signal space alignment for network coding," in *Proc. of IEEE ISIT*, vol. 1, Seoul, pp. 2892-2896, Jun. 2009.

[70] N. Lee, J.-B. Lim, and J. Chun, "Degrees of freedom of the MIMO Y channel: Signal space alignment for network coding," *IEEE Trans. Inf. Theory*, vol. 56, no. 7, pp. 3332-3342, Jul. 2010.

[71] Chaaban, A.; Sezgin, A., "Signal space alignment for the Gaussian Y-channel," *Proc. of IEEE ISIT*, pp.2087,2091, 1-6 July 2012.

[72] Hua Mu; Tugnait, J.K., "Achievable Degrees of Freedom for K -User MIMO Y Channels Using Signal Group Based Alignment," *Wireless Communications, IEEE Transactions on*, vol.13, no.8, pp.4520-4533, Aug. 2014.

[73] N. Wang, Z. Ding, X. Dai and A. V. Vasilakos, "On Generalized MIMO Y Channels: Precoding Design, Mapping, and Diversity Gain," *IEEE Trans. on Vehicular Technology*, vol. 60, no. 7, pp. 3525-3532, Sept. 2011.

[74] K. Liu, X. Yuan and M. Tao, "Optimal Degrees of Freedom Region for the Asymmetric MIMO Y Channel," *IEEE Communications Letters*, vol. 20, no. 12, pp. 2454-2457, Dec. 2016.

[75] T. Ding, X. Yuan and S. C. Liew, "On the Degrees of Freedom of the Symmetric Multi-Relay MIMO Y Channel," *IEEE Trans. on Wireless Comm.*, vol. PP, no. 99, pp. 1-1, 2017.

[76] K. Zheng, F. Liu, Q. Zheng, W. Xiang and W. Wang, "A Graph-Based Cooperative Scheduling Scheme for Vehicular Networks," *IEEE Trans. on Vehicular Technology*, vol. 62, no. 4, pp. 1450-1458, May 2013.

[77] Y. Cui, D. Jiang and Y. Wu, "Analysis and Optimization of Caching and Multicasting in Large-Scale Cache-Enabled Wireless Networks," *Wireless Communications, IEEE Transactions on*, vol. 15, no. 7, pp. 5101-5112, July 2016.

[78] N. Golrezaei; K. Shanmugam; A.G. Dimakis, A.F. Molisch, and G. Caire, "FemtoCaching: Wireless video content delivery through distributed caching helpers," *INFOCOM, 2012 Proceedings IEEE*, vol., no., pp.1107-1115, March 2012.

[79] L. Zhang, M. Xiao, G. Wu and S. Li, "Efficient Scheduling and Power Allocation for D2D-Assisted Wireless Caching Networks," *Communications, IEEE Transactions on*, vol. 64, no. 6, pp. 2438-2452, June 2016.

[80] J. Rao, H. Feng, C. Yang, Z. Chen and B. Xia, "Optimal caching placement for D2D assisted wireless caching networks," *IEEE International Conference on Communications (ICC)*, pp.1-6, 2016.

[81] H. Wu, L. Wang, T. Svensson and Z. Han, "Resource allocation for wireless caching in socially-enabled D2D communications," *IEEE International Conference on Communications (ICC)*, pp. 1-6, 2016.

[82] M. Gregori, J. Gómez-Vilardebó, J. Matamoros and D. Gündüz, "Wireless Content Caching for Small Cell and D2D Networks," *Selected Areas in Communications, IEEE Journal on*, vol. 34, no. 5, pp. 1222-1234, May 2016.

[83] M. A. Maddah-Ali and U. Niesen, "Fundamental Limits of Caching," *Information Theory, IEEE Transactions on*, vol.60, no.5, pp.2856-2867, May 2014.

[84] A. Ozgur and O. Leveque, "Throughput-delay trade-off for hierarchical cooperation in ad hoc wireless networks," *2008 International Conference on Telecommunications, St. Petersburg*, 2008, pp. 1-5.

[85] P. Gupta and P. R. Kumar, "The capacity of wireless networks," *IEEE Transactions on Information Theory*, vol. 46, no. 2, pp.388-404, Mar 2000.

[86] M. Franceschetti, O. Dousse, D. N. C. Tse and P. Thiran, "Closing the Gap in the Capacity of Wireless Networks Via Percolation Theory," *IEEE Transactions on Information Theory*, vol. 53, no. 3, pp. 1009-1018, March 2007.

[87] A. S. Avestimehr, S. N. Diggavi, and D. N. C. Tse, "Wireless Network Information Flow: A Deterministic Approach," *IEEE Transactions on Information Theory*, vol. 57, no. 4, pp.1872-1905, 2011.

[88] A. Ozgur, O. Leveque and D. Tse, "Hierarchical Cooperation Achieves Optimal Capacity Scaling in Ad Hoc Networks," *Information Theory, IEEE Transactions on*, vol. 53, no. 10, pp. 3549-3572, Oct. 2007.

[89] A. Ozgur, O. Leveque and D. Tse, "Spatial Degrees of Freedom of Large Distributed MIMO Systems and Wireless Ad Hoc Networks," *Selected Areas in Communications, IEEE Journal on*, vol. 31, no. 2, pp. 202-214, February 2013.

[90] A. Ozgur, R. Johari, and D. Tse, "Operating Regimes of Large Wireless Networks," *Foundations and Trends in Networking*, Vol. 5, No. 1, DOI: 10.1561/1300000016, 2010.

[91] C. Geng, N. Naderializadeh, A. S. Avestimehr, S. A. Jafar, "On the optimality of treating interference as noise," *Information Theory, IEEE Transactions on*, vol. 61, no. 4, pp. 1753-1767, Apr. 2015.

[92] S. N. Hong and G. Caire, "Beyond Scaling Laws: On the Rate Performance of Dense Device-to-Device Wireless Networks" *IEEE Transactions on Information Theory*, vol. 61, no. 9, pp. 4735-4750, Sept. 2015.

[93] M. Ji, G. Caire and A. F. Molisch, "Optimal throughput-outage trade-off in wireless one-hop caching networks," *Information Theory Proceedings (ISIT), 2013 IEEE International Symposium on*, pp.1461-1465, 7-12 July 2013.

[94] M. Ji, G. Caire and A. F. Molisch, "Wireless Device-to-Device Caching Networks: Basic Principles and System Performance," *Selected Areas in Communications, IEEE Journal on*, vol.34, no.1, pp.176-189, Jan. 2016.

[95] M. Ji, G. Caire and A. F. Molisch, "Fundamental Limits of Caching in Wireless D2D Networks," *Information Theory, IEEE Transactions on*, vol.62, no.2, pp.849-869, Feb. 2016.

[96] X. Chen, L. Jiao, W. Li and X. Fu, "Efficient Multi-User Computation Offloading for Mobile-Edge Cloud Computing," *IEEE/ACM Transactions on Networking*, vol. 24, no. 5, pp. 2795-2808, October 2016.

[97] W. Hoiles, O. N. Gharehshiran, V. Krishnamurthy, N. D. o and H. Zhang, "Adaptive Caching in the YouTube Content Distribution Network: A Revealed Preference Game-Theoretic Learning Approach," *IEEE Transactions on Cognitive Communications and Networking*, vol. 1, no. 1, pp. 71-85, March 2015.

[98] S. W. Jeon, S. N. Hong, M. Ji and G. Caire, "Caching in wireless multihop device-to-device networks," *IEEE International Conference on Communications (ICC)*, pp. 6732-6737, 2015.

[99] S. W. Jeon, S. N. Hong, M. Ji, G. Caire and A. F. Molisch, "Wireless Multihop Device-to-Device Caching Networks," *Information Theory, IEEE Transactions on*, vol. 63, no. 3, pp. 1662-1676, March 2017.

[100] A. Liu, V. Lau, and G. Caire. "Cache-induced Hierarchical Cooperation in Wireless Device-to-Device Caching Networks," arXiv preprint arXiv:1612.07417, 2016.

[101] A. Asadi, Q. Wang and V. Mancuso, "A Survey on Device-to-Device Communication in Cellular Networks," *IEEE Communications Surveys & Tutorials*, vol. 16, no. 4, pp. 1801-1819, 2014.

[102] C. Xu et al., "Efficiency Resource Allocation for Device-to-Device Underlay Communication Systems: A Reverse Iterative Combinatorial Auction Based Approach," *IEEE Journal on Selected Areas in Communications*, vol. 31, no. 9, pp. 348-358, September 2013.

[103] D. Feng, L. Lu, Y. Yuan-Wu, G. Y. Li, G. Feng and S. Li, "Device-to-Device Communications Underlaying Cellular Networks," *IEEE Transactions on Communications*, vol. 61, no. 8, pp. 3541-3551, August 2013.

[104] X. Lin, J. G. Andrews and A. Ghosh, "Spectrum Sharing for Device-to-Device Communication in Cellular Networks,"*IEEE Transactions on Wireless Communications*, vol. 13, no. 12, pp. 6727-6740, Dec. 2014.

[105] Y. Pei and Y. C. Liang, "Resource Allocation for Device-to-Device Communications Overlaying Two-Way Cellular Networks," *IEEE Transactions on Wireless Communications*, vol. 12, no. 7, pp. 3611-3621, July 2013.

[106] B. Zhou, H. Hu, S. Q. Huang and H. H. Chen, "Intracluster Device-to-Device Relay Algorithm With Optimal Resource Utilization," *IEEE Transactions on Vehicular Technology*, vol. 62, no. 5, pp. 2315-2326, Jun 2013.

[107] D. W. K. Ng, et al., "Energy-Efficient 5G Outdoor-to-Indoor Communication: SUDAS Over Licensed and Unlicensed Spectrum," *IEEE Transactions on Wireless Communications*, vol. 15, no. 5, pp. 3170-3186, May 2016.

[108] H. Min, J. Lee, S. Park and D. Hong, "Capacity Enhancement Using an Interference Limited Area for Device-to-Device Uplink Underlaying Cellular Networks," *IEEE Transactions on Wireless Communications*, vol. 10, no. 12, pp. 3995-4000, December 2011.

[109] C. Lin and M. Gerla, "Adaptive clustering for mobile wireless networks," *IEEE Journal on Selected Areas in Communications*, vol. 15, no. 7, pp.1265-1275, 1997.

[110] F. Fitzek, M. Katz and Q. Zhang, "Cellular Controlled Short-Range Communication for Cooperative P2P Networking," *Wireless Personal Communications*, vol. 48, issue 1, pp 141-155, January 2009.

[111] H. Nishiyama, M. Ito and N. Kato, "Relay-by-smartphone: realizing multihop device-to-device communications," *IEEE Communications Magazine*, vol. 52, no. 4, pp. 56-65, April 2014.

[112] Y. Cao, T. Jiang and C. Wang, "Cooperative device-to-device communications in cellular networks," *IEEE Wireless Communications*, vol. 22, no. 3, pp. 124-129, June 2015.

[113] F. Hou, L. X. Cai, P. H. Ho, X. Shen and J. Zhang, "A cooperative multicast scheduling scheme for multimedia services in IEEE 802.16 networks," *IEEE Transactions on Wireless Communications*, vol. 8, no. 3, pp. 1508-1519, March 2009.

[114] M. Dohler, "Virtual antenna arrays," PhD thesis, University of London, 2004.

[115] R. Heath, S. Peters, Y. Wang, and J. Zhang, "A current perspective on distributed antenna systems for the downlink of cellular systems," *IEEE Communications Magazine*, vol. 51, no. 4, pp. 161-167, Apr. 2013.

[116] H. Zhu, "On frequency reuse in cooperative distributed antenna systems," *IEEE Communications Magazine*, vol. 50, no. 4, pp. 85-89, Apr. 2012.

[117] J. Jiang, J. S. Thompson and H. Sun, "A Singular-Value-Based Adaptive Modulation and Cooperation Scheme for Virtual-MIMO Systems," *IEEE Transactions on Vehicular Technology*, vol. 60, no. 6, pp. 2495-2504, July 2011.

[118] G. Kramer, M. Gastpar and P. Gupta, "Cooperative Strategies and Capacity Theorems for Relay Networks," *IEEE Transactions on Information Theory*, vol. 51, no. 9, pp. 3037-3063, Sept. 2005.

[119] S. A. Ayoughi and W. Yu, "Optimized MIMO transmission and compression for interference mitigation with cooperative relay," *IEEE International Conference on Communications (ICC),* pp. 4321-4326, 2015.

[120] Q. H. Spencer, A. L. Swindlehurst and M. Haardt, "Zero-forcing methods for downlink spatial multiplexing in multiuser MIMO channels," *IEEE Transactions on Signal Processing*, vol. 52, no. 2, pp. 461-471, Feb. 2004.

[121] Jiajia Guo, Tao Yang, Jinhong Yuan and Jian A. Zhang, "Linear Vector Physical-layer Network Coding for MIMO Two-Way Relay Channels: Design and Performance Analysis," *IEEE Transactions on Communications*, vol. 63, no. 7, pp. 2591-2604, July 2015.

[122] Jiajia Guo, Tao Yang, Jinhong Yuan and Jian A. Zhang, "Linear Physical-layer Network Coding for the Fading Y-channel without Transmitter Channel State Information," *IEEE VTC* 2016 Fall.

[123] Jiajia Guo, Tao Yang, Jinhong Yuan and Jian A. Zhang, "Design of Linear Physical-layer Network Coding for MIMO Two-way Relay Channels without Transmitter CSI," *IEEE WCNC* 2015.

[124] Jiajia Guo, J. Yuan and Jian A. Zhang, "The Throughput Scaling Law of Wireless Device-to-device Caching Networks with Distributed MIMO and Hierarchical Cooperations," *IEEE Transactions on Wireless Communications*, vol. 17, no. 1, pp. 492-505, Jan. 2018.

[125] Jiajia Guo, J. Yuan and Jian A. Zhang, "Wireless Device-to-device Caching Networks with Distributed MIMO and Hierarchical Cooperations," *IEEE GlobeCom* 2017.

[126] B.M. Hochwald and S. ten Brink, "Achieving Near-Capacity on a Multiple-Antenna Channel," *IEEE Trans. Comm.* vol. 51, no. 3, pp. 389-399, Mar. 2003.

[127] M. O.Damen, K.Abed-Meraim, and M. S.Lemdani, "Further results on the sphere decoder," *Proc. IEEE ISIT*, 2001.

[128] M. Chiani, D. Dardari, and Marvin K. Simon, "New exponential bounds and approximations for the computation of error probability in fading channels," *IEEE Trans. Wireless Comm.*, vol. 2, no. 4, pp. 840-845, Jul. 2003.

[129] Erdös, Paul; Rényi, Alfréd , "On a classical problem of probability theory," Publ. Math. Inst. Hung. Acad. Sci., Ser. A 6, 215-220 (1961), 60C05.

[130] G. Fodor et al., "Design aspects of network assisted device-to-device communications," *IEEE Communications Magazine*, vol. 50, no. 3, pp. 170-177, March 2012.

[131] Abbas El Gamal and Young-Han Kim, "Network Information Theory," Cambridge University Press, 2011.

[132] Z. Wei, D. W. K. Ng, J. Yuan and H. M. Wang, "Optimal Resource Allocation for Power-Efficient MC-NOMA With Imperfect Channel State Information," *IEEE Transactions on Communications*, vol. 65, no. 9, pp. 3944-3961, Sept. 2017.

[133] R. G. Stephen and R. Zhang, "Joint Millimeter-Wave Fronthaul and OFDMA Resource Allocation in Ultra-Dense CRAN," *IEEE Transactions on Communications*, vol. 65, no. 3, pp. 1411-1423, March 2017.

[134] T. S. Rappaport, G. R. MacCartney, M. K. Samimi, S. Sun, "Wideband millimeter-wave propagation measurements and channel models for future wireless communication system design," *IEEE Transactions on Communications*, vol. 63, no. 9, pp. 3029-3056, Sep. 2015.

[135] Q. Shi, et al., "An Iteratively Weighted MMSE Approach to Distributed Sum-Utility Maximization for a MIMO Interfering Broadcast Channel," *IEEE Transactions on Signal Processing*, vol. 59, no. 9, pp. 4331-4340, Sept. 2011.

[136] Taesang Yoo and A. Goldsmith, "On the optimality of multiantenna broadcast scheduling using zero-forcing beamforming," *IEEE Journal on Selected Areas in Communications*, vol. 24, no. 3, pp. 528-541, March 2006.