UNIVERSITY OF TECHNOLOGY SYDNEY
Faculty of Engineering and Information Technology

Research on Key Technologies of Low-Latency Uplink Non-Orthogonal Multiple Access

by

Jie Zeng

A Thesis Submitted
in Partial Fulfillment of the
Requirements for the Degree

Doctor of Philosophy

Sydney, Australia

2020

Certificate of Authorship/Originality

I, Jie Zeng, declare that this thesis, is submitted in fulfilment of the requirements for the award of doctor of philosophy, in the Faculty of Engineering and Information Technology at the University of Technology Sydney. This thesis is wholly my own work unless otherwise reference or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis. I certify that the work in this thesis has not previously been submitted for a degree nor has it been submitted as part of the requirements for a degree except as fully acknowledged within the text. This thesis is the result of a research candidature jointly delivered with Beijing University of Posts and Telecommunications as part of a Collaborative Doctoral Research Degree. This research is supported by the Australian Government Research Training Program.

Signature: Production Note:
Signature removed prior to publication.

Date: 19 / 2 / 2021

ABSTRACT

Research on Key Technologies of Low-Latency Uplink Non-Orthogonal Multiple Access

by

Jie Zeng

5G is expected to significantly reduce latency in numerous emerging services in the Internet of Things (IoT). Non-orthogonal multiple access (NOMA) has been widely regarded as one promising technology to enable 5G. NOMA can allow multiple users to send signals in the same radio resource simultaneously, and can distinguish the signals of users with multi-user detection (MUD) at the receiver. NOMA can be combined with multiple-input and multiple-output (MIMO), advanced modulation and coding, and full-duplex (FD) to ensure low latency communications with high reliability. Generally speaking, the non-orthogonal superposition of signals from access users and grant-free scheduling can reduce the access latency; MIMO and FD can shorten the transmission latency by increasing spectrum efficiency. This thesis studies the design and enhancement of NOMA to guarantee low latency in the uplink (UL), under the assumption of massive accessed users, a few transmit antennas, shadow fading, and imperfect channel state information, which reflect the characteristics of IoT services. Meanwhile, the effectiveness of the proposed schemes is evaluated via the novel finite blocklength information theory, thereby complying with the small packet size in IoT. The main contributions of this thesis are summarized as follows.

1. The rate splitting algorithms and successive interference cancellation detection in multi-user MIMO (MU-MIMO) NOMA are proposed to minimize the maximum transmission latency of users. The achievable data rates are derived, and two corresponding rate splitting algorithms are proposed. Numerical results validate that the

rate splitting MU-MIMO NOMA can efficiently shorten the transmission latency and processing latency.

2. The sparse code multiple access enhanced FD (FD-SCMA) scheme is designed to operate UL and downlink (DL) simultaneously. This thesis derives the error probability with imperfect self-interference suppression in FD. FD-SCMA is proved to achieve lower transmission latency than existing SCMA and FD schemes in time-invariant flat-fading channels and time-invariant frequency-selective fading channels by theoretical calculation and simulation.

3. Low latency transmission of the emerging MU-MIMO NOMA is studied. Under log-normal shadow fading, this thesis derives the probability density function of effective SNRs, and calculates the error probability given transmission latency. Further, the error probability can be minimized by adjusting the length of pilots. Simulation results verify that the MU-MIMO NOMA enables low latency transmissions under moderate shadow fading for massive accessed users.

Overall, NOMA can remarkably lower access latency, transmission latency, and processing latency in UL.

Acknowledgements

First, I would like to thank my principal supervisor, Prof. Ren Ping Liu, for his guidance and support with patience through the three-year scientific research. He has a thorough knowledge of the information and communication technologies, and has a strong influence in the academic and industrial fields. I would like to thank Prof. Liu for providing me with an excellent working team and cutting-edge research topics during my study. He has given me profound advice on key decisions, which benefit me a lot.

I also would like to express my deep appreciation to my co-supervisors, Prof. Xiaojing Huang and Dr. Wei Ni, for their valuable advice and guidance on my research. From them, I learned a lot of useful experience and the style of well-known scholars.

Special thanks to the University of Technology Sydney (UTS) and Beijing University of Posts and Telecommunications (BUPT) for providing the scholarship and financial support for my study.

To my family, thank you all for your love to encourage me to complete my oversea study.

Jie Zeng

Sydney, Australia, 2020.

List of Publications

J-1 **J. Zeng, T. Lv, R. P. Liu, X. Su, Y. J. Guo, N. C. Beaulieu.** *"Enabling Ultra-reliable and Low Latency Communications Under Shadow Fading by Massive MU-MIMO" in* ***IEEE Internet of Things Journal***, vol. 7, no. 1, pp. 234–246, Jan. 2020.

J-2 **J. Zeng, T. Lv, Z. Lin, *et al.*** *"Achieving Ultrareliable and Low-Latency Communications in IoT by FD-SCMA" in* ***IEEE Internet of Things Journal***, vol. 7, no. 1, pp. 363–378, Jan. 2020.

J-3 **J. Zeng, T. Lv, W. Ni, R. P. Liu, N. C. Beaulieu, Y. J. Guo.** *"Ensuring Max-Min Fairness of UL SIMO-NOMA: A Rate Splitting Approach" in* ***IEEE Transactions on Vehicular Technology***, vol. 68, no. 11, pp. 11080–11093, Nov. 2019.

J-4 **J. Zeng, T. Lv, R. P. Liu, X. Su, N. C. Beaulieu, Y. J. Guo** *"Linear Minimum Error Probability Detection for Massive MU-MIMO With Imperfect CSI in URLLC" in* ***IEEE Transactions on Vehicular Technology***, vol. 68, no. 11, pp. 11384–11388, Nov. 2019.

J-5 **G. J. Sutton, J. Zeng, R. P. Liu, *et al.*** *"Enabling Technologies for Ultra-Reliable and Low Latency Communications: From PHY and MAC Layer Perspectives" in* ***IEEE Communications Surveys & Tutorials***, vol. 21, no. 3, pp. 2488–2524, Third Quart. 2019.

J-6 **C. Xiao, J. Zeng, W. Ni, R. P. Liu, X. Su, J. Wang.** *"Delay Guarantee and Effective Capacity of Downlink NOMA Fading Channels" in* ***IEEE Journal of Selected Topics in Signal Processing***, vol. 13, no. 3, pp. 508–523, Jun. 2019.

J-7 **C. Xiao, J. Zeng, W. Ni, *et al.*** *"Downlink MIMO-NOMA for Ultra-*

Reliable Low-Latency Communications" in ***IEEE Journal on Selected Areas in Communications***, vol. 37, no. 4, pp. 780–794, Apr. 2019.

J-8 **J. Zeng, T. Lv, R. P. Liu, *et al.*** *"Investigation on Evolving Single-Carrier NOMA Into Multi-Carrier NOMA in 5G" in* ***IEEE Access***, vol. 6, pp. 48268–48288, 2018.

J-9 **T. Lv, Y. Ma, J. Zeng, P. T. Mathiopoulos.** *"Millimeter-Wave NOMA Transmission in Cellular M2M Communications for Internet of Things" in* ***IEEE Internet of Things Journal***, vol. 5, no. 3, pp. 1989–2000, Jun. 2018.

J-10 **G. J. Sutton, J. Zeng, R. P. Liu, *et al.*** *"Enabling Ultra-Reliable and Low-Latency Communications through Unlicensed Spectrum" in* ***IEEE Network***, vol. 32, no. 2, pp. 70–77, Mar. 2018.

J-11 **J. Zeng, D. Kong, B. Liu, X. Su, T. Lv.** *"RIePDMA and BP-IDD-IC Detection" in* ***EURASIP Journal on Wireless Communications and Networking***, vol. 2017, no. 1, pp. 1-12, Dec. 2017.

Contents

List of Figures

Abbreviation

3rd generation partnership project - 3GPP

Additive white Gaussian noise - AWGN

Base station - BS

Belief propagation - BP

Bits per channel use - bpcu

Block error rate - BLER

Channel use - CU

Channel state information - CSI

China academy of telecommunications technology - CATT

Co-channel interference - CCI

Downlink - DL

Enhanced mobile broadband - eMBB

The fifth generation - 5G

Finite blocklength - FBL

Full-duplex - FD

Golden section search method - GSSM

Infinite blocklength - IBL

Interference cancellation - IC

International telecommunication union - ITU

Internet of Things - IoT

Large-scale fading - LSF

Length of pilots - LoP

Lower bound - LB

Massive machine-type communications - mMTC

Maximal-ratio combining - MRC

Maximal-ratio combining SIC - MRC-SIC

Maximum a posteriori - MAP

Maximum likelihood - ML

Message passing algorithm - MPA

Minimum mean square error - MMSE

Minimum mean square error SIC - MMSE-SIC

Multi-carrier NOMA - MC-NOMA

Multiple access channel - MAC

Multiple-input and multiple-output - MIMO

Multiple-input multiple-output NOMA - MIMO-NOMA

Multi-user detection - MUD

Multi-user interference - MUI

Multi-user multiple-input multiple-output - MU-MIMO

Multi-user shared access - MUSA

New radio - NR

Non-orthogonal multiple access - NOMA

Next generation node B - gNB

Ordered SIC - OSIC

Orthogonal frequency division multiple access - OFDMA

Orthogonal frequency division multiplexing - OFDM

Orthogonal multiple access - OMA

Pattern division multiple access - PDMA

Perfect interference cancelation - PIC

Pilot-assistant channel estimation - PACE

Power-domain NOMA - PD-NOMA

Probability density function - pdf

Receiving antennas - Rx antennas

Resource block - RB

Right-hand side - RHS

Resource element - RE

SCMA enhanced FD - FD-SCMA

Self-interference suppression - SIS

Signal to interference plus noise ratio - SINR

Signal to noise ratio - SNR

Small-scale fading - SSF

Sparse code multiple access - SCMA

Spectral efficiency - SE

Successive interference cancellation - SIC

System-level simulation - SLS

Transmit antennas - Tx antennas

Transmit power - Tx power

Ultra-reliable and low latency communications - URLLC

Uplink - UL

Upper bound - UB

User equipment - UE

Zero forcing – ZF

Chapter 1

Introduction

1.1 Background

To support the rapid development of mobile Internet and the Internet of Things (IoT) to a higher standard, the fifth generation (5G) mobile networks have been significantly improved in spectrum efficiency, energy efficiency, and the number of simultaneously connected devices. 5G can be deployed in three typical application scenarios:

1) enhanced mobile broadband (eMBB) can significantly increase user data rates;

2) massive machine-type communications (mMTC) will connect a considerable number of low data rate IoT devices;

3) ultra-reliable and low latency communications (URLLC) need to support ultra-high reliability with low latency in short packet communications.

Specifically, the user plane latency for transmitting a 32-byte packet cannot exceed 1 ms in URLLC [1]. This feature will undoubtedly lead to several promising applications, and some typical low-latency user cases are shown in Fig. 1.1.

Simultaneously, the requirements for ultra-high data rates, low latency, high reliability, and massive connectivity bring new challenges to multiple access technologies. Traditional orthogonal multiple access (OMA) has been challenging to satisfy these requirements at the same time since it is limited by the single-user capacity bound and the division of radio resources. Therefore, non-orthogonal multiple access (NOMA), while achieving the higher requirements of 5G, is considered to be the potential breakthrough direction of wireless communications. NOMA can superimpose and transmit signals from different users on the same radio resources. Then, advanced multi-user detection (MUD) is applied at the receiver to distinguish

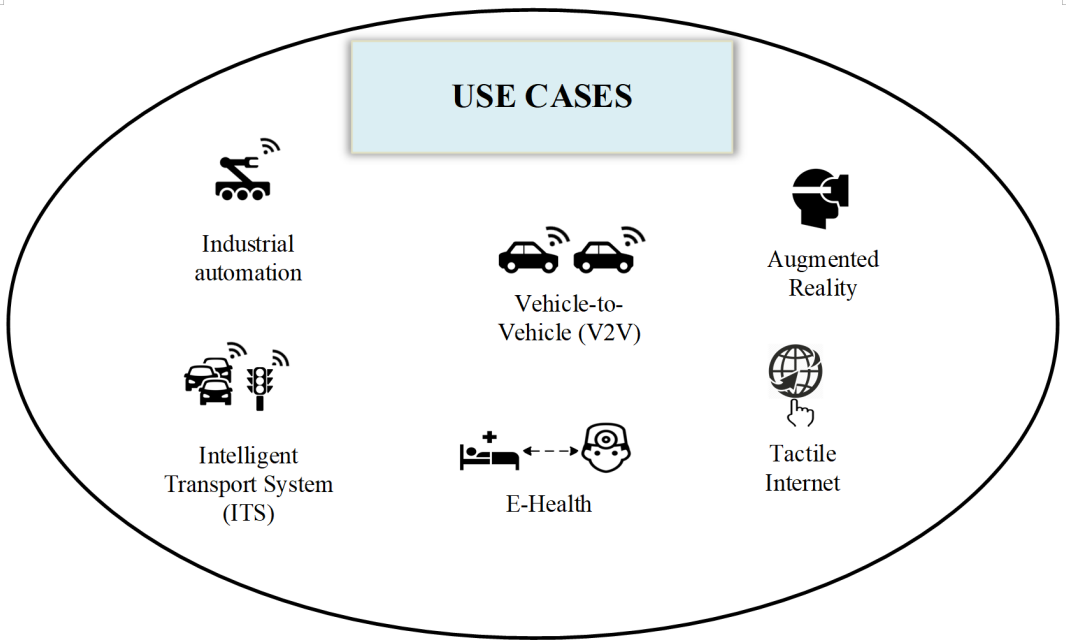

Figure 1.1 : Typical use cases of 5G low latency communications.

received signals from multiple users. Consequently, NOMA significantly improves frequency efficiency and connection density in various application scenarios. Furthermore, NOMA can reduce access latency by simplifying signaling with grant-free transmission. Compared with OMA, NOMA can achieve significant gains in improvement of multi-user capacity, increase of simultaneous accessing users amount, reduction of scheduling overhead and latency. Thus, it can attain 5G core performance indicators and support a variety of critical IoT applications. The research on NOMA, such as key technologies, design, and evaluation, has been actively promoted by the international telecommunication union (ITU) and the 3rd generation partnership project (3GPP).

1.2 Motivation

Currently, 5G mobile networks are widely expected to support the rapidly developing IoT applications, especially those with ultra-low latency that could not be ensured by existing mobile networks. The superimposing transmission and advanced MUD of NOMA can reduce transmission latency by the reuse of radio resources.

Meanwhile, NOMA can shorten the access latency by employing grant-free transmission for multiple users. It is worth to have further exploration on NOMA key technologies to realize low latency in the uplink (UL) IoT applications.

Considering the time and frequency selective fading, diversity plays a vital role in reliability improvement in UL short packet transmission. However, the time domain diversity brought by retransmission is significantly restricted due to the low latency constraint. It is more appropriate to incorporate the diversity gains from the frequency and space domains into NOMA. Accordingly, this thesis will focus on the research of multi-carrier NOMA (MC-NOMA) and multiple-input multiple-output NOMA (MIMO-NOMA). Practically, a large number of receiving antennas (Rx antennas) and a small number of transmit antennas (Tx antennas) are equipped in the UL of 5G IoT deployments. Thus, the combination of multi-carrier and multi-receiver diversity, the design of low-latency UL transmission, and stable detection with low complexity are thoroughly studied. These studies are significant and rarely mentioned in the previous literature, which enhances the necessity of our work.

The integration of NOMA and full-duplex (FD) will encounter self-interference and inter-user interference, which could become more complicated when the number of Rx antennas increases. Previous studies usually assume a single Rx antenna is equipped and refrain from complicated interference. With multiple Rx antennas, this thesis designs a NOMA system that supports simultaneous UL and downlink (DL) transmission to further reduce transmission latency. Considering the effects of self-interference suppression (SIS), this thesis explores how to guarantee reliability given ultra-low latency in UL and DL. For instance, the block error rate (BLER) can be lowered to less than 10^{-5} given 0.5 ms transmission latency.

Most of the existing research on NOMA is based on Shannon's information theory and assumes that a coding block is infinitely long. This theory has limitations in IoT applications that mainly transmit short packets. The recently proposed finite block length (FBL) information theory is utilized in NOMA to analyze the performance of reliability and latency. Based on the FBL information theory, the evaluation and comparison of the proposed and existing schemes are more generally applicable in

IoT.

Finally, the performance of MUD with multiple Rx antennas is highly dependent on accurate channel state information (CSI). However, within limited latency, it is difficult to allocate a large number of pilots for accurate CSI. Considering the impact of imperfect CSI, the length of pilots (LoP) in pilot-assistant channel estimation (PACE) should be optimized. Then, a compromise between the accuracy of channel estimation and the data rate can be achieved. This thesis examines how to improve the reliability of UL low-latency transmission under imperfect CSI, which has great practical value and has not been investigated in previous studies.

1.3 Research Status

To date, a variety of NOMA technologies, including power domain NOMA (PD-NOMA), pattern division multiple access (PDMA), sparse code multiple access (SCMA), multi-user shared access (MUSA) and so on, have been proposed. The design principle, key characteristics, pros, and cons of NOMA technologies are introduced and compared in [2]. Also, the authors point out the opportunities, challenges, and potential research trends. Key NOMA technologies in grant-free UL transmission are discussed in [3]. Finally, a summary of the concepts, challenges, applications, and research trends of NOMA technologies is performed in [4].

NTT DOCOMO has proposed PD-NOMA, which can superimpose signals of multiple users at the same radio resources and distinguish the signals by using serial interference cancellation (SIC) at the receiver [5, 6]. System-level simulation (SLS) results indicate that PD-NOMA can achieve better performance than traditional OMA in various application environments [7].

China academy of telecommunications technology (CATT) has proposed PDMA, which can introduce unequal diversity between multiple users to raise the sum data rate. PDMA superimposes and transmits signals on multiple dimensions (the time domain, frequency domain, power domain, and space domain) according to the designed PDMA pattern matrix, resulting in a higher multiplexing gain [8]. PDMA uses advanced MUD to approach the capacity boundary of the multiple access chan-

nel (MAC). The PDMA pattern matrix design can be independently or jointly performed in multiple dimensions to improve the performance of reliability and latency in different scenarios.

Huawei has proposed SCMA, which uses the sparse expansion of the code domain and message passing algorithm (MPA), to improve the spectrum efficiency and access density [9]. SCMA combines low-density encoding and modulation techniques to generate a codebook set by constellation rotation, and assigns a specific codebook to each user for signal mapping. SCMA can effectively reduce the latency with grant-free transmission, and thus, it is suitable for low-latency UL IoT applications.

The NOMA technologies mentioned above have the advantages of high spectral efficiency (SE), high flexibility, and low latency. However, it is challenging to realize higher performance MUD due to complex multi-user interference (MUI) caused by superimposing transmission. As we all know, the maximum a posterior (MAP) is the optimal detection algorithm, and in the case of uniform priori probabilities, maximum likelihood (ML) can be equivalent to MAP [10]. However, when the number of users is large and the modulation order is high, the computational complexity and the processing latency are particularly high in MAP and ML detection. Generally, linear MUD with lower complexity can be performed using the minimum mean square error (MMSE) algorithm, the zero-forcing (ZF) algorithm, and so on. At the same time, iterative detection algorithms, such as belief propagation (BP) and MPA, can also be implemented to achieve reliability close to MAP [11].

One critical challenge in low-latency communications is to maximize reliability with affordable computational complexity at the receiver. A low-complexity MPA for UL SCMA is proposed to achieve a compromise in processing latency and reliability by partial marginalization in [12]. In addition, a weighted MPA is presented for SCMA to reduce complexity by removing the iterative process [13]. Moreover, an iterative detection and decoding mechanism for SCMA is suggested to significantly improve reliability with acceptable computational complexity [14]. The codeword level and symbol level SIC detection in PD-NOMA is analyzed, and the influence of error propagation on reliability is evaluated [15]. It can be seen that the research

on low-latency and low-complexity MUD affects the promotion and applications of NOMA critically.

NOMA can be classified into single-carrier NOMA and MC-NOMA by the mapping pattern of signals. When the same symbol is mapped onto multiple carriers at the transmitter, it is considered herein as MC-NOMA. Recent studies have also attempted to apply PD-NOMA in multi-carrier orthogonal frequency division multiple access (OFDMA) systems [16, 17]. It is generally believed that PDMA and SCMA, which were proposed in the early stage of 5G research and have been rapidly developed, are two representative MC-NOMA technologies. Meanwhile, they can combine with OFDMA maturely and can adopt the sparsity of the mapping matrix to reduce complexity. It should be noted that most of the spread spectrum-based NOMA technologies can be naturally included in the scope of MC-NOMA. Companies from the industry mainly promote these MC-NOMA technologies which need to be in-depth investigated theoretically.

The challenges and applications of URLLC have been mentioned in technical reports and specifications published by standards organizations [1]. Reliability can be defined as the probability of transmitting a short packet under low latency successfully. In URLLC, the user plane latency in UL and DL is restricted under 0.5 ms. For 32-byte short packet transmission, the reliability is generally required to be 99.999% and above. The effective bandwidth and effective capacity are used to calculate the maximum achievable data rate, and a shorter frame can be designed to reduce the latency in [18]. On the other hand, the headers and data are proposed to be combined and coded effectively to support faster and more reliable transmission [19]. Moreover, reference [20] explores the trade-off between bandwidth, coding scheme, signal-to-noise ratio (SNR), diversity order, and reliability when transmitting 100-bit short packets within 100 μs. Shorter frames can ensure low-latency UL transmission under required reliability. Besides, installing adequate Rx antennas can guarantee reliable low-latency UL transmission via spatial diversity.

1.4 Main Contributions

This thesis mainly studies the key technologies of NOMA in low-latency UL communications, including receiver design, latency reduction, short packets transmission, and impact of imperfect CSI. Innovative research has been undertaken, and the main contributions are listed as follows.

1. This thesis comprehensively overviews the key technologies of NOMA and low-latency communications, and summarize the developments and trends of NOMA. MC-NOMA combined with OFDMA is a vital candidate technology for 5G due to its higher spectrum efficiency, higher access density, and lower access latency than traditional OMA. At the same time, the emerging FBL information theory is applied to analyze the reliability of short packet services given transmission latency. The emerging grant-free NOMA, which eliminates the granting and scheduling processes in the UL random access, is adopted to significantly reduce the user plane latency.

2. Multi-user multiple-input multiple-output (MU-MIMO) NOMA is explored to guarantee the user data rate. This thesis introduces the concept of symmetric capacity and establishes the system model of UL MU-MIMO NOMA. The equivalent SNR and the minimum user data rate of multi-user and multi-stream detection are derived for SIC-based algorithms: MMSE-SIC and maximal-ratio combining (MRC)-SIC. At the same, sufficient conditions for stable SIC detection are given. In UL, a rate splitting algorithm is proposed to maximize the minimum user data rate. Further, a two-layer rate splitting algorithm is presented to achieve a lower bound (LB) of the minimum user data rate with low-complexity MRC-SIC detection. Consequently, maximum transmission latency and processing latency can be optimized in UL. Simulation results validate that the proposed MU-MIMO NOMA technology significantly reduces the maximum transmission latency with affordable computational complexity.

3. FD-SCMA is proposed to satisfy the requirements of URLLC in short-packet applications. In FD-SCMA, multiple UL users and DL users map their signals according to the SCMA codebook and transmit the signals simultaneously in the FD

mode. The effective SNRs of UL users are derived, and then, the error probability is calculated based on the FBL information theory. FD-SCMA is theoretically proved to achieve higher reliability than the existing FD, SCMA, and OMA technologies with limited latency in time-invariant flat-fading channels. Theoretical analysis and numerical results also verify that FD-SCMA outperforms the existing technologies in time-invariant frequency-selective fading channels. It shows that the proposed FD-SCMA has significant advantages in supporting low-latency short packet transmission in UL.

4. The MU-MIMO NOMA is studied with perfect CSI and imperfect CSI. This thesis reveals that equipping multiple Rx antennas improves reliability via space diversity. Then, the impacts of estimation errors, which are caused by PACE, on UL MU-MIMO NOMA are evaluated. The effective SNRs of ZF detection are derived under perfect and imperfect CSI. Assuming the users are homogeneously and randomly deployed, the probability density functions (pdfs) of effective SNRs under perfect and imperfect CSI are derived under lognormal shadow fading. The error probability of ZF detection given transmission latency is calculated based on the FBL information theory. Finally, the golden section search method (GSSM) is proposed to reduce the error probability by optimizing the LoP. Simulation results verify that the proposed MU-MIMO NOMA can ensure URLLC for a large number of accessing users with extremely low-complexity ZF detection.

Overall, the research in this thesis responds to the demand for 5G ultra-low latency with the innovative NOMA concept, FD, and MU-MIMO technologies. Our work contributes to reducing the access, transmission, and processing latency of UL accessing users in IoT applications.

1.5 Thesis Organization

This thesis is divided into six chapters and is structured as follows.

Chapter 1 introduces the research background, motivation, and research status of NOMA-based low latency communications, and gives the main contributions and structure of this thesis.

Chapter 2 compares the key technologies of NOMA, including single-carrier NOMA and MC-NOMA. Meanwhile, it highlights the method of maintaining reliability in low-latency communications, and emphasizes that grant-free NOMA can further reduce the latency in UL.

Chapter 3 studies MU-MIMO NOMA to optimize multi-user UL superposition transmission via spatial diversity. With the proposed rate splitting and stable SIC detection, the minimum user data rate can be improved compared with OMA technologies. To this end, the maximum user transmission latency can be directly reduced in short packet transmission.

Chapter 4 explores the proposed FD-SCMA to demonstrate its superiority over other existing technologies in two typical time-invariant fading channels. Considering the characteristics of short packets in IoT applications, the FBL information theory can be applied to model achievable data rates.

In Chapter 5, the NOMA combined with MU-MIMO is investigated to adapt to the features of IoT, such as short packets, imperfect CSI, and severe shadow fading. Based on PACE, an algorithm is proposed to optimize LoP and effectively resist shadow fading under latency constraints.

Chapter 6 summarizes the thesis and points out the limitations and future research directions.

Chapter 2

Key Technologies of NOMA and Low-latency Transmission

Low-latency communications are characterized by several unique features, such as small data packet size, small transmission latency, low detection complexity, and few retransmissions. To this end, the research on the key technologies of low-latency UL NOMA can be implemented in two aspects. The first one is latency reduction through redesigning the transmitter and receiver of NOMA, and the second one is incorporation of the mature technologies by supporting low-latency communications for NOMA. This chapter firstly introduces the research progress of existing NOMA technologies, especially in low-complexity detection. Then, a comprehensive investigation has been performed on technologies that can effectively guarantee the reliability of low-latency communications, such as diversity, advanced modulation and coding, and FD. In particular, the superimposed users' signals can be distinguished by advanced MUD, so grant-free transmission can be implemented in UL NOMA to further reduce the access latency.

2.1 NOMA Techniques

NOMA techniques can be classified into the single-carrier NOMA technique and the MC-NOMA technique, according to the way of mapping users' signals on occupied carriers. The schematic diagrams of single-carrier NOMA and MC-NOMA are shown in Fig. 2.1.

2.1.1 Single-Carrier NOMA Technique

In single-carrier NOMA, the transmitting signals of two or more users are superimposed on one carrier. Meanwhile, a representative single-carrier NOMA scheme is PD-NOMA [21], in which superimposed users' signals are distinguished by different

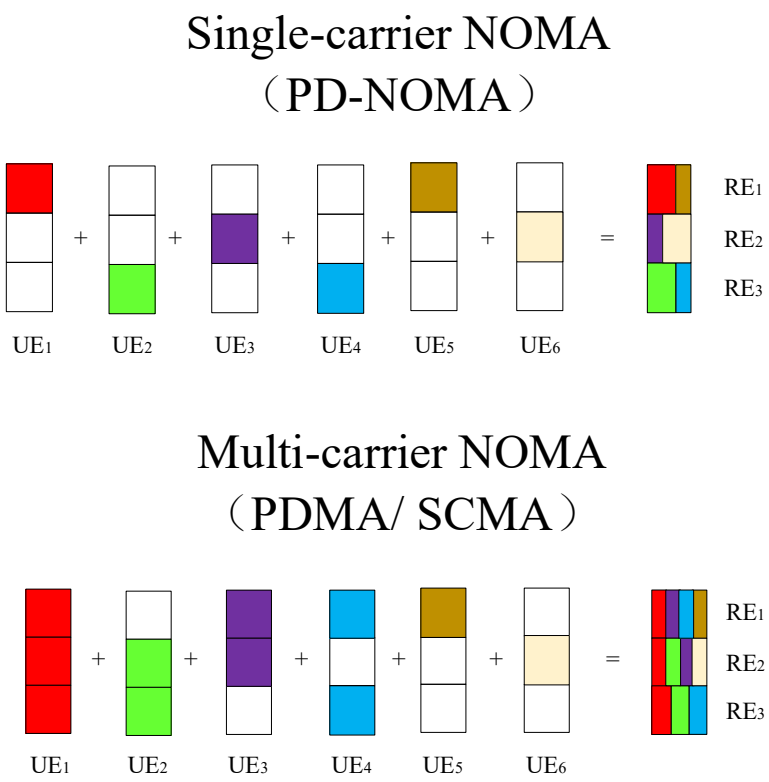

Figure 2.1 : Illustrations of single-carrier NOMA and MC-NOMA.

power levels. Below is a brief outline of PD-NOMA from the perspectives of the system model and receiver design. PD-NOMA supports reliable multiple access by superimposing signals from different users on the power domain and then distinguishing them by different power levels. PD-NOMA adopts advanced SIC-enabled MUD in DL and UL, which can fully utilize the differences between users' channel conditions. As can be verified, it is superior to OMA in terms of spectrum efficiency and user fairness [22, 23].

In DL PD-NOMA, as shown in Fig. 2.2, a base station (BS) transmits the signal x_i to user i with transmit power (Tx power) P_i^{DL} $(i = 1,\ 2)$, where $P_1^{\mathrm{DL}} + P_2^{\mathrm{DL}} \leq P$ and $\mathbb{E}[|x_i|^2] = 1$. In DL PD-NOMA, the superimposed signal transmitted by BS can be given by $x = \sqrt{P_1^{\mathrm{DL}}}x_1 + \sqrt{P_2^{\mathrm{DL}}}x_2$. Then, the signal received by user i is

$$y_i = h_i x + n_i = y_i = h_i(\sqrt{P_1^{\mathrm{DL}}}x_1 + \sqrt{P_2^{\mathrm{DL}}}x_2) + n_i \tag{2.1}$$

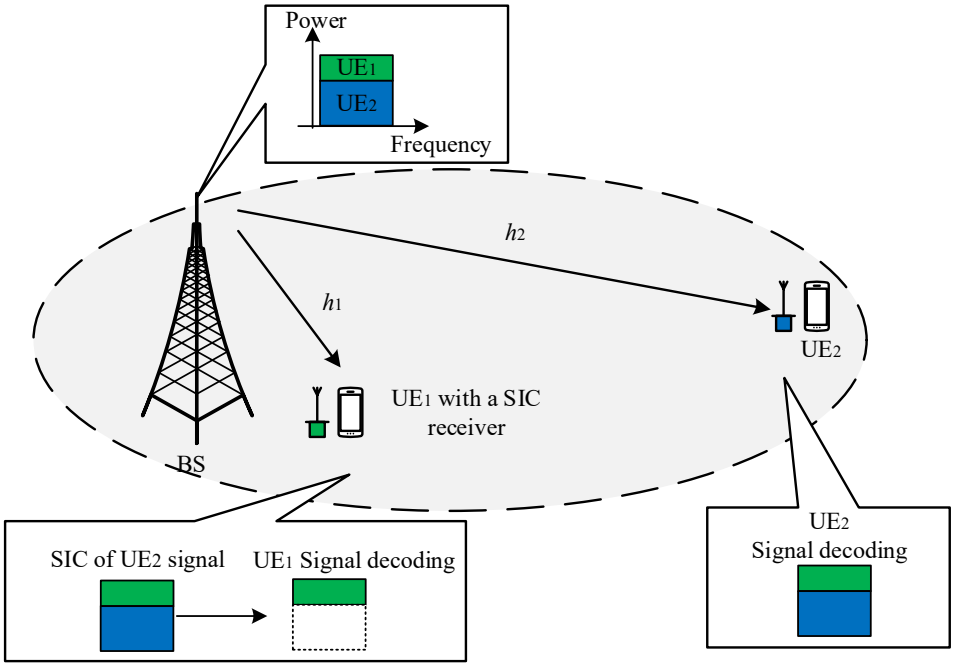

Figure 2.2 : Schematic diagram of DL PD-NOMA [24].

where h_i is the channel gain between BS and user i, and n_i indicates additive white Gaussian noise (AWGN) at the receiver of user i with power N_i. It is assumed that user 1 locates at the cell center, user 2 locates at the cell edge, and $|h_1| > |h_2|$. At this time, user 1 performs SIC detection in the ascending order of channel gains. To this end, the achievable data rates of users 1 and 2 can be respectively expressed as

$$R_1^{\mathrm{DL}} = \log_2(1 + \frac{P_1^{\mathrm{DL}}|h_1|^2}{N_1}) \tag{2.2}$$

and

$$R_2^{\mathrm{DL}} = \log_2(1 + \frac{P_2^{\mathrm{DL}}|h_2|^2}{P_1^{\mathrm{DL}}|h_2|^2 + N_2}). \tag{2.3}$$

PD-NOMA utilizing SIC detection can obtain gains of more than 20% in the total DL throughput and the cell edge user's throughput, respectively [25].

In UL PD-NOMA, as shown in Fig. 2.3, the signal sent by user i is denoted as x_i $(i = 1,\ 2)$, and the Tx power is P_i^{UL} with $\mathbb{E}[|x_i|^2] = 1$. Then, the superimposed

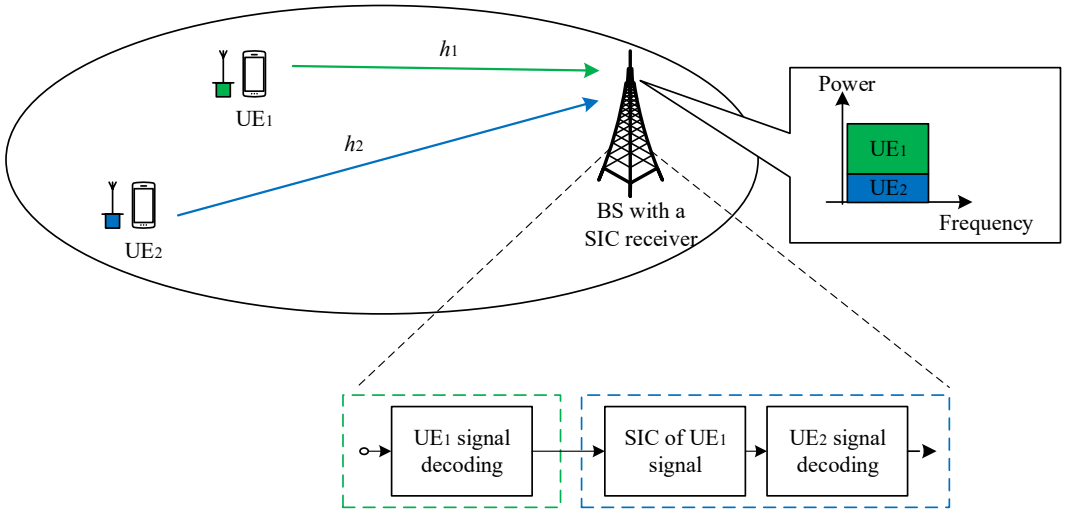

Figure 2.3 : Schematic diagram of UL PD-NOMA [24].

signal received by the BS can be expressed as

$$y = h_1\sqrt{P_1^{\mathrm{UL}}}x_1 + h_2\sqrt{P_2^{\mathrm{UL}}}x_2 + n_0 \tag{2.4}$$

where h_i is the channel gain between user i and BS, n_0 representing the AWGN with power N_0 at the BS. The BS performs SIC detection according to the descending order of channel gains. Consequently, the achievable data rates of user 1 and user 2 can be expressed as

$$R_1^{\mathrm{UL}} = \log_2(1 + \frac{P_1^{\mathrm{UL}}|h_1|^2}{P_2^{\mathrm{UL}}|h_2|^2 + N_0}) \tag{2.5}$$

and

$$R_2^{\mathrm{UL}} = \log_2(1 + \frac{P_2^{\mathrm{UL}}|h_2|^2}{N_0}). \tag{2.6}$$

It can be seen that the cell edge user's throughput can be markedly improved, since the interference from the cell center user has been removing by SIC detection.

2.1.2 MC-NOMA Technique

MC-NOMA, such as PDMA and SCMA, distinguishes users' signals through different codewords and mapping patterns. PDMA carries out non-orthogonal su-

perimposition of different users' signals according to a pattern matrix, and separates the signals by detection algorithms, such as BP and MPA. At the same time, SCMA directly maps the coded bits of users to multi-dimensional codewords in a sparse mode, and detects superimposed signals by MPA and so on.

PDMA

PDMA can simultaneously superimpose signals of multiple users on radio resources from the time, frequency, and space domains to effectively increase the number of accessing users [8, 26]. PDMA maps the users' signals sparsely to different numbers of carriers, and gains diversity with reduced system complexity. PDMA can increase reliability through advanced MUD based on MPA and BP.

It can be found that more reliable data transmission can be obtained with a broader diversity gain.To this end, the signal of the user with more diversity should be decoded with priority to reduce error propagation. The joint of diversity, the overload factor, and complexity are essential in the design of the PDMA pattern matrix. Tang *et al.* [27] have verified that grant-free PDMA effectively supports massive UL users. Advanced MUD can be utilized at the receiver, such as the rapidly developing BP algorithm, to mitigate inter-user interference. The sparsity of the PDMA pattern matrix can reduce the complexity of a BP receiver, and the different orders of transmission diversity can accelerate the convergence of BP [8]. Better performance can be achieved at the receiver through interference cancellation (IC) schemes, such as MMSE-IC [28], ML-IC, and BP-IDD-IC. BP-IDD has also been proposed to increase reliability by adding external iterations [29] .

SCMA

Nikopour *et al.* [9] have proposed SCMA, which uses sparse code expansion and multi-dimensional modulation to reduce the complexity of MUD. In SCMA, different users are assigned a unique codebook, which indicates how signals of one user are mapped to multiple carriers. With different codebooks, users can be distinguished by differences in occupied carriers. Codebook design is one vital scheme in SCMA. It consists of the design of the factor graph and multi-dimensional constellation and

determines the reliability and processing latency of transmission. Each codebook can be generated from the parent constellation with a certain rotation, so users' signals can be pseudo-orthogonal. System-level codebook design has been described and verified to be practical and effective [9, 30], and has been verified to be practical and effective.

To meet the requirements of massive connectivity, SCMA utilizes the MPA scheme based on low-density signature sequences. However, due to the relatively high computational complexity of MPA, it is necessary to develop a low-complexity solution without significant reliability decrement. An iterative MUD has been proposed to effectively reduce complexity in demodulation and decoding, fully utilizing the coding gain and the diversity gain from the structure and factor graph of the SCMA codebook [12, 31]. Besides, two low-complexity MPA algorithms for SCMA have been proposed to ensure reliability [32, 33]. A threshold-based low-complexity MPA, which performs a deterministic check through a specific threshold during the iterative process, reduces unnecessary iterations via timely removal of the successfully decoded users [34]. At present, the research on the receiver design of SCMA mainly focuses on improving reliability and throughput without increasing complexity significantly. The research on low-complexity receiver design is particularly worthy of attention since it is more practical in low-latency communications.

2.1.3 Research Prospects

Compared to traditional OMA technologies, NOMA technologies can provide higher spectral efficiency (SE) and achievable data rates. To date, the research on single-carrier NOMA has been relatively mature, but MC-NOMA has not been extensively studied. The combinations of MC-NOMA and other key technologies of 5G, such as cooperative MC-NOMA and MIMO enhanced MC-NOMA, are worth further exploration in the coming 5G era.

2.2 Key Technologies for Low-latency Communications

At present, several physical layer technologies that can assure reliability in low-latency communications. The newly proposed FBL information theory can be utilized to facilitate the transmitter and receiver design in short packet transmission. Meanwhile, it is more effective to improve reliability from frequency/space diversity than from time diversity when the retransmission is critically limited. Advanced modulation and coding schemes can support the detection of short data packets with low processing latency. Besides, FD technology can reduce access latency via simultaneous UL and DL transmission.

2.2.1 FBL Information Theory

It is generally assumed that the random coding scheme with infinite blocklength (IBL) can achieve the Shannon capacity. When packet size is sufficiently large, the Shannon capacity can effectively approximate the achievable data rate. However, the Shannon information theory cannot obtain high accuracy when the packet size is small in typical IoT applications. To this end, the relationship between the achievable data rate and error probability needs to be revealed with the assumption of FBL. Recently, the emerging FBL information theory is applied to model and analyze the achievable data rate given error probability, and adequately fits the needs of low-latency IoT applications.

Given blocklength and error probability in short packet transmission, Polyanskiy *et al.* [35] firstly proposed the FBL information theory to derive the approximation and bounds of the achievable data rate in the AWGN channel and revealed the relationship among reliability, bandwidth, and SNR. Meanwhile, the FBL information theory was extended to multi-antenna scenarios in Rayleigh fading channels, and the achievable data rate given error probability was derived [36, 37]. Following that, the FBL information theory was developed to discover the upper and lower bounds of the achievable data rate given SNR and error probability in multi-carrier MIMO [38]. Furthermore, the error probability and sufficient capacity of incorporating single-carrier NOMA into relays can be analyzed with the same theoretical

tool [39].

2.2.2 Diversity-based Technology

Diversity is critical in reliability improvement, which is always an essential goal of wireless communications. Generally, diversity can be obtained from the space, frequency, time, and code domains. Space diversity obtained from multiple propagation paths can reduce the error probability to a deficient level. Frequency diversity achieved from the multi-carrier technology can adequately combat frequency-selective fading. Implementing retransmission in time-varying fading channels can yield time diversity. Moreover, adapting low-rate modulation and coding to the channel conditions gains diversity from the code domain.

In the frequency-selective fading channel without precise CSI, frequency hopping can enable transmissions in a predefined frequency sequence. It can achieve a high frequency diversity gain in rich scattering environments. Furthermore, fast frequency hopping can be combined with the forward error correction design to allow duplicated bits to be sent together in time, while also having low-correlated noise and fading.

The FBL model was utilized to study the effects of space and frequency diversity on required bandwidth and reliability [40]. Recently, massive MIMO, with the capability of mainly increasing space diversity, has become more critical in improving reliability. In the UL case, Panigrahi *et al.* pointed out that the vast antenna array equipped at the BS could guarantee ultra-high reliability, even though only one or two antennas are equipped at UL users [41]. Considering the possible deep-fading channel in IoT, appropriate diversity-based techniques play a significant role in maintaining robust low-latency transmission.

2.2.3 Modulation and Coding Techniques for Short Packets

Modulation and coding techniques are essential methods to ensure reliability by mitigating the effects of channel fading. Currently, to satisfy the critical processing latency, low-latency and reliable encoding/decoding and modulation/demodulation

schemes for short packets are in urgency.

A resource allocation scheme in hierarchical multicast communications was proposed to sharply contract the decoding operations in random linear network coding [42]. A low-complexity rate adaptation technique was studied to minimize the concurrent transmission latency and processing latency of sensor nodes [43]. Punctured trellis-coded modulation was proposed to achieve higher SE through a dynamic coding rate and low complexity decoding [44].

2.2.4 FD Technology

In FD communications, signals are transmitted and received simultaneously on the same antenna array with the help of SIS [45]. FD devices can suppress or eliminate their transmitting signals at the receiver, through effective signal processing and antenna design.

The resource allocation optimization in FD under the constraints of power and load was presented, and then a method to minimize the data queuing latency of users was proposed [46]. NOMA, combined with FD, was studied in ultra-dense networks, and could achieve much higher data rates than OMA and NOMA combined with Half-Duplex (HD) [47]. FD can significantly increase SE through simultaneous transmission and reception. However, self-interference cannot be completely removed at the receiver with practical SIS. When applying FD in low-latency communications, it is necessary to model the impact of imperfect SIS on robust detection.

2.2.5 Combination of the Existing Techniques

Generally, it is not entirely straightforward to associate the existing techniques, since they might operate under conflicting assumptions. For example, retransmission within the coherence time of the fading channels are unable to gain time diversity adequately in the limited latency budget. In FD networks, employing multiple Tx antennas directly leads to the complex and severe inter-user interference and self-interference. Furthermore, the emerging techniques are facing realistic propagation channels without perfect real-time CSI, where the interference could not be mitigated

absolutely.

2.3 Low-latency Grant-free NOMA

Grant-free NOMA mitigates the requesting and scheduling processes at the access stage. Users transmit on pre-configured radio resources without scheduling, thereby significantly reducing signaling overheads and access latency. The time and frequency resources, as well as the dedicated pilots, for UL grant-free transmission, are semi-statically pre-configured to users. Grant-free NOMA can be an effective way to enable URLLC if collision probability could be further reduced.

Chen *et al.* [48] studied grant-free NOMA to mitigate random access collisions with a heavy load, thus ensuring reliability with reduced latency. It is verified to be practical to reduce the latency significantly through a combination of NOMA and analog fountain codes. A dynamic compressed sensing-based MUD is proposed to utilize the time correlation of active users [49]. The scheme could estimate the set of active users in the current slot and regard the set as a priori information of the next time slot. Besides, an adaptive subspace tracking algorithm assisted by a priori information was proposed to guarantee reliable MUD [50].

However, the computational complexity of grant-free NOMA can be extraordinarily high with iterated MUD. The reduction of processing latency is the top priority in order to deploy grant-free NOMA widely. In addition, when multiple pairs of grant-free transmission are simultaneously supported in one network, in-band interference and out-of-band interference becomes more complicated and need to be further alleviated. When the multi-user beamforming and space-division multiplexing schemes are adopted, large numbers of Tx and Rx antennas are always equipped. Consequently, the CSI estimation and feedback processes are high in computational complexity, resulting in a significant latency increment. To this end, with limited latency, users with no more than two antennas are more suitable to be deployed in grant-free NOMA.

2.4 Summary of This Chapter

This chapter has summarized the key technologies of NOMA and low-latency communications. NOMA, single-carrier NOMA, MC-NOMA, and their technical principles have been introduced, respectively. The transmitter and receiver design for two representative MC-NOMA technologies, PDMA and SCMA, have been reviewed. To realize low-latency communications, diversity, advanced modulation and coding technology, and FD technology can also be applied. Meanwhile, it is necessary to analyze the error probability of short packets through the FBL information theory. Particularly, UL grant-free NOMA, which simplifies UL access processes, has become a potential enabler of low-latency transmission for massive short packets.

Chapter 3

Layered MU-MIMO NOMA Transmission and SIC Detection

In this chapter, multi-layer superposition transmission is proposed and studied in the MU-MIMO NOMA. The achievable data rate regions of two detection methods (MMSE-SIC and MRC-SIC) are derived. At the same time, a multi-layer rate splitting scheme is proposed to approach symmetric capacity. Furthermore, a two-layer rate splitting scheme is proposed to ensure an affordable minimum user data rate with low detection complexity and processing latency. The schemes proposed in this chapter can improve the minimum data rate of multiple UL users when equipping multiple Rx antennas, thereby reducing the maximum transmission latency of UL users. Incorporating SIC into linear MMSE and MRC detection can reduce the complexity of each SIC operation, and is superior to other SIC algorithms in processing latency reduction.

3.1 System Model of MU-MIMO NOMA

3.1.1 Symmetric Capacity

Symmetric capacity is a crucial indicator of fairness in the MAC. It is defined as the maximum data rate at which all users in the system can reliably communicate simultaneously [51]. In typical application scenarios, such as autonomous driving, public safety, and drone communications, UL users are randomly located and experiencing different fading channels. In order to reduce the transmission latency, each user also needs to maximize the achievable data rate. In these scenarios, it is challenging to increase the symmetric capacity that is equivalent to maximizing the minimum achievable data rate of users. Symmetric capacity can be approached through implementation of ML detection [51] and iterative BP detection [52], but

the computational complexity can be extremely high. In most cases, limited by demodulation and decoding latency, it is necessary to consider low-complexity SIC-based detection to approach symmetric capacity [52], thereby ensuring user fairness and minimizing the maximum transmission latency of users.

Ding *et al.* proposed a general framework of MIMO-NOMA [53–55], which can effectively increase the system data rate and reduce the outage probability. Although UL transmission plays an essential role in IoT applications, UL MIMO-NOMA's current research progress is significantly slower than DL MIMO-NOMA's. On the one hand, the signal alignment MIMO-NOMA [53] ideally assumes that the number of Tx antennas is greater than or equal to half of the number of Rx antennas. However, there is only one or a small amount of Tx antennas at each user in typical UL IoT applications. Therefore, signal alignment MIMO-NOMA cannot be directly applied in most UL IoT applications where the number of Tx antennas is small. On the other hand, although the implementation of iterative detection to achieve the capacity region of MIMO-NOMA has been revealed [56, 57], approaching symmetric capacity by low-complex and low-latency algorithms without iterations has not been studied in the existing literature. Therefore, it is necessary to discover an effective method that maximizes the minimum user data rate in MU-MIMO NOMA.

The energy-efficient precoding, joint power control of pilot and data, and UL training were studied to improve performance further [58–60]. In particular, a general framework based on signal alignment has been proposed to support UL applications where the number of Tx antennas is greater than half of the number of Rx antennas [53]. In addition, it has been proposed to approach the boundary of the capacity of MIMO-NOMA by iterative MMSE detection [56, 57]. The challenges in designing MUD and determining the detection order of users to ensure stable SIC detection were raised in [61].

Rate splitting can be managed to reach the entire capacity region [62], realizing by the well-known Han-Kobayashi coding scheme [63]. Studies have discovered that the applications of rate splitting in MIMO can provide higher reliability, SE, and energy efficiency [64]. Meanwhile, rate splitting and ZF beamforming can be

combined to increase the data rate [65]. In DL MIMO, rate splitting and power control were combined for a higher data rate, and the closed-form expression of optimal power allocation is given by Dai *et al.* [66]. Besides, precoding and rate splitting can optimize the sum data rate of DL MU-MISO while maintaining fairness among users [67, 68]. In addition, NOMA could apply a heuristic rate splitting scheme to achieve almost the same data rate as BP detection with much lower decoding complexity [69].

Group decoding with interference management has been studied to suppress inter-user interference in UL MUD [70, 71]. Data are encoded into multiple layers at each user, and stable group SIC is performed at the BS to eliminate part of the interference first and achieve a higher sum data rate [72]. Multi-layer superposition transmission with group decoding achieves significant throughput gain [70]. Receiver-centric group decoding schemes can be used to design multicast beamforming in multi-cell networks to increase data rates or reduce power consumption with lower complexity [71]. The combination of rate splitting and group decoding is adopted in two-way relay networks based on NOMA. It can markedly increase the ergodic data rate and reduce the outage probability [73].

Inspired by recent research on rate splitting and group decoding, this chapter proposes a method for layered superposition transmission and SIC detection, including MMSE-SIC and MRC-SIC, to approach symmetric capacity in MU-MIMO NOMA, as shown in Fig. 3.1.

First, we derive the achievable data rate region through MMSE-SIC detection and MRC-SIC detection in MU-MIMO NOMA. Second, this chapter proves that the multi-layer superposition transmission combined with MMSE-SIC detection can approach the symmetric capacity. Then, a new UL multi-user rate splitting algorithm is proposed to ensure the minimum user data rate is close to symmetric capacity with convergent detection. This chapter also proves that stable MRC-SIC detection can be obtained under two-layer rate splitting while raising the minimum user data rate.

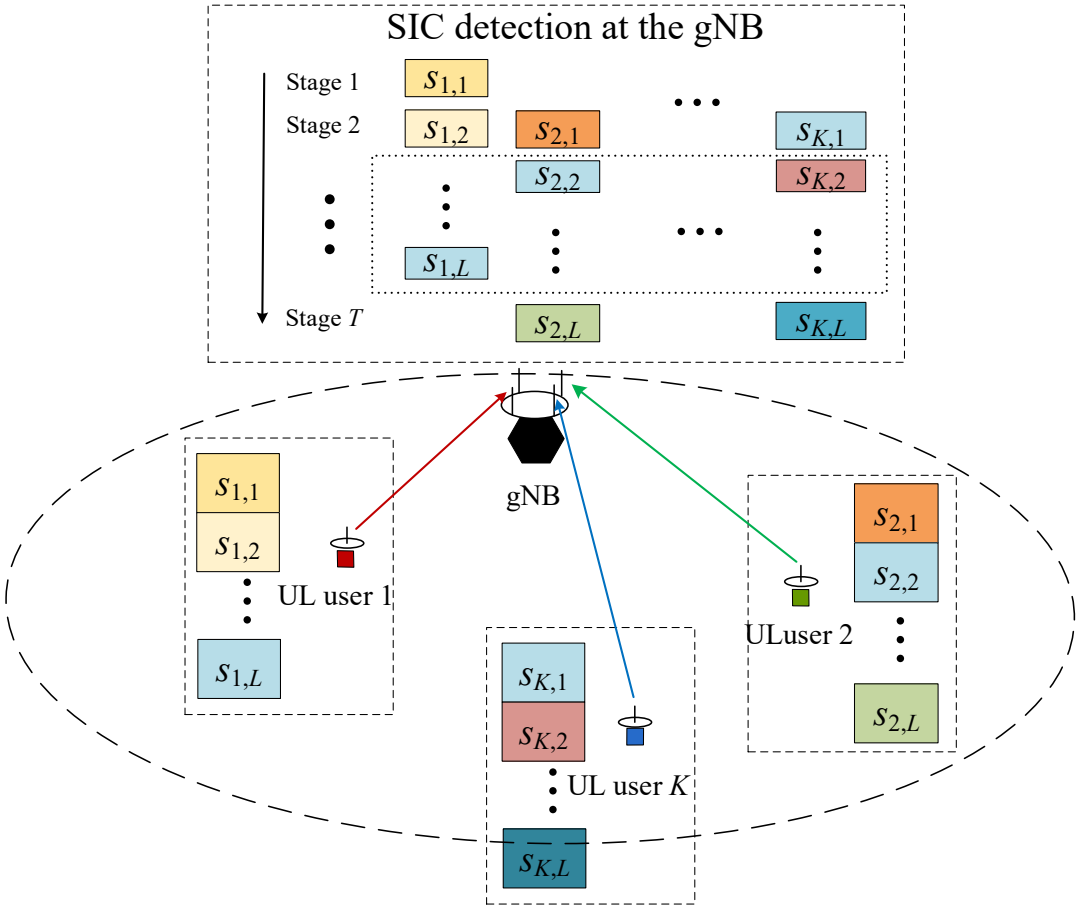

Figure 3.1 : An illustration of the MU-MIMO NOMA system.

3.1.2 System Model

In this chapter, $\mathbf{h}_k = \varpi_k \mathbf{g}_k$ are used to represent the channel coefficients from user k to the gNB, where $\mathbf{g}_k = [g_{k,1}, ..., g_{k,M}]^T \sim \mathcal{CN}(0, \mathbf{I}_M)$ denotes the Rayleigh fading coefficients of user k, $\varpi_k = d_k^{-\alpha^{\mathrm{PL}}}$ represents the LSF, where d_k is the distance from user k to the gNB, α^{PL} denotes the path loss exponent.

In multi-layer MU-MIMO NOMA, the received signal at the gNB can be given

by

$$\begin{aligned}\mathbf{y} &= \sum_{i=1}^{K} \mathbf{h}_i \sum_{j=1}^{L} \sqrt{\alpha_{i,j}p}\, s_{i,j} + \mathbf{z} \\ &= \underbrace{\mathbf{h}_k \sqrt{\alpha_{k,l}p}\, s_{k,l}}_{\text{signal of user } k \text{ layer } j} + \underbrace{\mathbf{h}_k \sum_{j=1, j\neq l}^{L} \sqrt{\alpha_{k,j}p}\, s_{k,j}}_{\text{interference of intra-user signals}} \\ &\quad + \underbrace{\sum_{i=1, i\neq k}^{K} \sum_{j=1}^{L} \mathbf{h}_i \sqrt{\alpha_{i,j}p}\, s_{i,j}}_{\text{interference of inter-user signals}} + \mathbf{z},\end{aligned} \tag{3.1}$$

where $\sum_{j=1}^{L} \sqrt{\alpha_{i,j}p}\, s_{i,j}$ is the superposed transmit signal of the L data streams from user i; p is the maximum Tx power of UL users; $s_{k,l}$ represents the data stream on the layer l of user k, $\alpha_{k,l}$ is the corresponding power allocation coefficient on the layer l of user k ($1 \leq k \leq K, 1 \leq l \leq L$), with $\mathbf{E}(s_{k,l}) = 0$, $\mathbf{E}(|s_{k,l}|) = 1$, and $\mathbf{E}(s_{k,l}s_{i,j}) = 0$, $\forall (k,l) \neq (i,j)$; $\mathbf{z} \sim \mathcal{CN}(0, \sigma^2 \mathbf{I}_M)$ is the complex additive white Gaussian noise at the gNB.

$\mathbf{r}_K = [r_1, r_2, ..., r_K]^T$ represents the data rates of different users, where r_k denotes the data rate of user k. $\mathbf{R}_{K,L} = [r_{k,l}]_{K\times L}$ represents the data rates of different layers of different users, where $r_{k,l}$ denotes the data rate of the layer l of user k ($r_k = \sum_{l=1}^{L} r_{k,l}$). The power allocated to different users and different layers can be expressed as $\mathbf{p}_K = [p_1, p_2, ..., p_K]^T$ and $\mathbf{P}_{K,L} = [p_{k,l}]_{K\times L}$, where $p_k = \alpha_k p$ and $p_{k,l} = \alpha_{k,l} p$ are Tx power of user k and the layer l of user k. α_k and $\alpha_{k,l}$ denote the power control and power splitting factors ($\sum_{l=1}^{L} \alpha_{k,l} = \alpha_k \leq 1$), which should be carefully designed to ensure the stable SIC.

In general, the symmetric capacity in MU-MIMO NOMA can be defined as follows,

$$\begin{aligned}&\mathcal{C}_{\text{sym}}(\mathbf{p}_K, \mathbf{H}_K, \sigma^2) \\ &= \max\left\{\bar{r} \mid |\mathcal{S}|\bar{r} \leq \mathcal{C}_{\mathcal{S}}(\mathbf{p}_K, \mathbf{H}_K, \sigma^2), \forall \mathcal{S} \subseteq \{1, ..., K\}\right\},\end{aligned} \tag{3.2}$$

where $\mathbf{H}_K = [\mathbf{h}_1, \mathbf{h}_2, ..., \mathbf{h}_K]^T$, and $\mathcal{C}_{\mathcal{S}}(\mathbf{p}_K, \mathbf{H}_K, \sigma^2) = \log_2 \det\left(\mathbf{I}_M + \sum_{k\in\mathcal{S}} \frac{p_k}{\sigma^2} \mathbf{h}_k \mathbf{h}_k^H\right)$ is the supremum of the sum data rate of the users in the set $\mathcal{S}$ [51]. In this sense, symmetric capacity is the maximum data rate that all UL users can achieve si-

multaneously. The process of approaching symmetric capacity can be equivalent to increasing the minimum user data rate and improving fairness among UL users. Symmetric capacity is also related to the number of UL users. With more UL users, more capacity constraints must be satisfied, and thus the symmetric capacity decreases.

3.2 SIC Detection with Multiple Rx Antennas

3.2.1 MMSE-SIC for Maximizing Sum Data Rate

Data streams can be successfully detected by MMSE and then be reconstructed and subtracted from the received signal at each SIC stage in MMSE-SIC. Without loss of generality, the set $\mathcal{U}$ can be used to represent the remaining data streams; the other data streams have been successfully detected. When simultaneously detecting all data streams $s_{k_1,l_1}, s_{k_2,l_2}, ..., s_{k_{|\mathcal{D}|},l_{|\mathcal{D}|}}$ in the set $\mathcal{D}$ with MMSE, where $\mathcal{D} \subseteq \mathcal{U} \subseteq \mathcal{K}_{K,L} = \{(k,l) \mid 1 \leqslant k \leqslant K, 1 \leqslant l \leqslant L\}$, the received signal that has successfully removed all detected data streams can be expressed as

$$\begin{aligned} \mathbf{y}_{\mathcal{U}} &= \sum_{(i,j)\in\mathcal{U}} \mathbf{h}_i \sqrt{\alpha_{i,j} p} s_{i,j} + \mathbf{z} \\ &= \underbrace{\sum_{(k,l)\in\mathcal{D}} \mathbf{h}_k \sqrt{\alpha_{k,l} p} s_{k,l}}_{\text{desired data streams}} + \underbrace{\sum_{(i,j)\in\mathcal{U}\setminus\mathcal{D}} \mathbf{h}_i \sqrt{\alpha_{i,j} p} s_{i,j}}_{\text{interference of other data streams}} + \mathbf{z}. \end{aligned} \tag{3.3}$$

We assume that $\mathbf{H}_{\mathcal{D}} = [\sqrt{\alpha_{i,j} p}\mathbf{h}_i]_{(i,j)\in\mathcal{D}}$, and the elements in $\mathbf{H}_{\mathcal{D}}$ is sorted in the ascending order of the values of $k \times L + l$, for example,

$$\mathbf{H}_{\mathcal{D}} = [\sqrt{\alpha_{1,1} p}\mathbf{h}_1, ..., \sqrt{\alpha_{1,L} p}\mathbf{h}_1, \sqrt{\alpha_{2,1} p}\mathbf{h}_1, ..., \sqrt{\alpha_{K,L} p}\mathbf{h}_K.$$

Thus, the MMSE detection can be expressed by the following formula.

$$\begin{aligned}
&\mathbf{V}_{\mathcal{D},\mathcal{U}}^{\mathrm{MMSE}}\mathbf{y}_{\mathcal{U}} = \mathbf{H}_{\mathcal{D}}^{H}\mathbf{F}_{\mathcal{U}}^{-1}\mathbf{y}_{\mathcal{U}} \\
=&\mathbf{H}_{\mathcal{D}}^{H}\mathbf{F}_{\mathcal{U}}^{-1}\mathbf{H}_{\mathcal{D}}\left[\sqrt{\alpha_{k_1,l_1}p}s_{k_1,l_1},...,\sqrt{\alpha_{k_{|\mathcal{D}|},l_{|\mathcal{D}|}}p}s_{k_{|\mathcal{D}|},l_{|\mathcal{D}|}}\right]^T \\
&+\mathbf{H}_{\mathcal{D}}^{H}\mathbf{F}_{\mathcal{U}}^{-1}\left(\sum_{(i,j)\in\mathcal{U}\setminus\mathcal{D}}\mathbf{h}_i\sqrt{\alpha_{i,j}p}s_{i,j}+\mathbf{z}\right),
\end{aligned} \tag{3.4}$$

where $\mathbf{v}_{\mathcal{D},\mathcal{U}}^{\mathrm{MMSE}} = \mathbf{H}_{\mathcal{D}}^{H}\mathbf{F}_{\mathcal{U}}^{-1}$ and $\mathbf{F}_{\mathcal{U}} = \sigma^2\mathbf{I}_M + \mathbf{H}_{\mathcal{U}}\mathbf{H}_{\mathcal{U}}^{H} = \sigma^2\mathbf{I}_M + \sum_{(i,j)\in\mathcal{U}}\alpha_{i,j}p\mathbf{h}_i\mathbf{h}_i^{H}$. And, the μ-th row of $\mathbf{V}_{\mathcal{D},\mathcal{U}}^{\mathrm{MMSE}}\mathbf{y}_{\mathcal{U}}$ is an estimation of data stream s_{k_μ,l_μ}. Consquently, the detector used to detect data stream $s_{k,l}$ can be expressed as $\mathbf{v}_{(k,l),\mathcal{U}}^{\mathrm{MMSE}} = \sqrt{\alpha_{k,l}p}\mathbf{h}_k^{H}\mathbf{F}_{\mathcal{U}}^{-1} = \nu_{(k,l),\mathcal{U}}\mathbf{h}_k^{H}\mathbf{F}_{\mathcal{U}\setminus(k,l)}^{-1}$, where $\nu_{(k,l),\mathcal{U}} = \sqrt{\alpha_{k,l}p}(1+\alpha_{k,l}p\mathbf{h}_k^{H}\mathbf{F}_{\mathcal{U}\setminus(k,l)}^{-1}\mathbf{h}_k)^{-1}$ is a scalar, and the second equality of the above equation can be derived from the Sherman-Morrison-Woodbury formula [74, eq. 1].

When the detector $\mathbf{v}_{(k,l),\mathcal{U}}^{\mathrm{MMSE}}$ is adopted, an estimation of $s_{k,l}$ can be given by

$$\begin{aligned}
&\mathbf{v}_{(k,l),\mathcal{U}}^{\mathrm{MMSE}}\mathbf{y}_{\mathcal{U}} \\
=&\nu_{(k,l),\mathcal{U}}\Bigg(\underbrace{\mathbf{h}_k^{H}\mathbf{F}_{\mathcal{U}\setminus(k,l)}^{-1}\mathbf{h}_k\sqrt{\alpha_{k,l}p}s_{k,l}}_{\text{desired data stream}} \\
&\underbrace{+\mathbf{h}_k^{H}\mathbf{F}_{\mathcal{U}\setminus(k,l)}^{-1}\Big(\sum_{(i,j)\in\mathcal{U}\setminus(k,l)}\mathbf{h}_i\sqrt{\alpha_{i,j}p}s_{i,j}+\mathbf{z}\Big)}_{\text{interference from other data streams and noise}}\Bigg).
\end{aligned} \tag{3.5}$$

Hence, the equivalent SNR of the detecting data stream $s_{k,l}$ is

$$\begin{aligned}
&\mathrm{SINR}_{(k,l),\mathcal{U}}^{\mathrm{MMSE}} \\
=&\frac{\parallel \mathbf{h}_k^{H}\mathbf{F}_{\mathcal{U}\setminus(k,l)}^{-1}\left(\mathbf{h}_k\sqrt{\alpha_{k,l}p}\right)\parallel^2}{\mathbf{E}\left\{\parallel \mathbf{h}_k^{H}\mathbf{F}_{\mathcal{U}\setminus(k,l)}^{-1}\left(\sum_{(i,j)\in\mathcal{U}\setminus(k,l)}\mathbf{h}_i\sqrt{\alpha_{i,j}p}s_{i,j}+\mathbf{z}\right)\parallel^2\right\}} \\
=&\alpha_{k,l}p\parallel \mathbf{h}_k^{H}\mathbf{F}_{\mathcal{U}\setminus(k,l)}^{-1}\mathbf{h}_k\parallel .
\end{aligned} \tag{3.6}$$

When the following constraints are satisfied, MMSE can successfully detect the data stream $s_{k,l}$ at each SIC stage.

$$r_{k,l} \le \mathcal{R}_{(k,l),\mathcal{U}}^{\mathrm{MMSE}} \triangleq \log_2\left(1+\alpha_{k,l}p\parallel \mathbf{h}_k^{H}\mathbf{F}_{\mathcal{U}\setminus(k,l)}^{-1}\mathbf{h}_k\parallel\right). \tag{3.7}$$

If (3.7) is true, all data streams in set $\mathcal{D}$ can be correctly detected at the same time. When multiple data streams are detected simultaneously, each data stream suffers from interference from other data streams. Maximizing the minimum user data rate can be effectively achieved by detecting a data stream at one SIC stage, i.e., a single layer SIC detection. It can increase the equivalent SNR of one user by subtracting the detected signals timely.

3.2.2 Low-latency and Low-complexity MRC-SIC

MMSE-SIC and MRC-SIC are two classic MUD schemes. MMSE-SIC can maximize the data rate, while MRC-SIC can significantly reduce computational complexity to support low-latency access for multiple users [51]. The MRC detector is $\mathbf{V}_{\mathcal{D}}^{\mathrm{MRC}} = \mathbf{H}_{\mathcal{D}}^{H}$, and the processed signal at the time of detection is as follows.

$$\begin{aligned}\mathbf{V}_{\mathcal{D}}^{\mathrm{MRC}}\mathbf{y}_{\mathcal{U}} =&\mathbf{H}_{\mathcal{D}}^{H}\mathbf{H}_{\mathcal{D}}\left[s_{k_1,l_1},...,s_{k_{|\mathcal{D}|},l_{|\mathcal{D}|}}\right]^{T}\\ &+\mathbf{H}_{\mathcal{D}}^{H}\sum_{(i,j)\in\mathcal{U}\setminus\mathcal{D}}\mathbf{h}_i\sqrt{\alpha_{i,j}p}s_{i,j}+\mathbf{H}_{\mathcal{D}}^{H}\mathbf{z},\end{aligned} \tag{3.8}$$

where the μ-th row of $\mathbf{V}_{\mathcal{D}}^{\mathrm{MRC}}\mathbf{y}_{\mathcal{U}}$ can be used to estimate the data stream s_{k_μ,l_μ}, and it can be further expanded to

$$\begin{aligned}&\sqrt{\alpha_{i,j}p}\mathbf{h}_k^{H}\mathbf{y}_{\mathcal{U}}\\ =&\sqrt{\alpha_{i,j}p}\parallel\mathbf{h}_k\parallel^2\Bigg(\underbrace{\frac{\mathbf{h}_k^{H}}{\parallel\mathbf{h}_k\parallel^2}\mathbf{h}_k\sqrt{\alpha_{k,l}p}s_{k,l}}_{\text{desired data stream}}\\ &+\underbrace{\frac{\mathbf{h}_k^{H}}{\parallel\mathbf{h}_k\parallel^2}\sum_{(i,j)\in\mathcal{U}\setminus(k,l)}\mathbf{h}_i\sqrt{\alpha_{i,j}p}s_{i,j}}_{\text{interference of other data streams}}+\underbrace{\frac{\mathbf{h}_k^{H}}{\parallel\mathbf{h}_k\parallel^2}\mathbf{z}}_{\text{noise}}\Bigg).\end{aligned} \tag{3.9}$$

When using MRC detection, the equivalent received power of the desired data stream $s_{k,l}$ is $\alpha_{k,l}p\parallel\mathbf{h}_k\parallel^2$, the noise power is $\mathbf{E}\left\{\parallel\mathbf{h}_k^{H}\mathbf{z}\mathbf{z}^{H}\mathbf{h}_k\parallel/\parallel\mathbf{h}_k\parallel^2\right\}=\sigma^2$, and the power of equivalent interference can be expressed as

$$\mathbf{E}\left\{\parallel\mathbf{h}_k^{H}\Big(\sum_{(i,j)\in\mathcal{U}\setminus(k,l)}\mathbf{h}_i\sqrt{\alpha_{i,j}p}s_{i,j}\Big)\parallel^2/\parallel\mathbf{h}_k\parallel^2\right\}=\sum_{(i,j)\in\mathcal{U}\setminus(k,l)}\alpha_{i,j}p\parallel\mathbf{h}_k^{H}\mathbf{h}_i\parallel^2/\parallel\mathbf{h}_k\parallel^2 .$$

Therefore, the equivalent SNR of $s_{k,l}$ could be expressed as follows.

$$\begin{aligned}&\mathrm{SINR}_{(k,l),\mathcal{U}}^{\mathrm{MRC}}\\=&\frac{\alpha_{k,l}p \parallel \mathbf{h}_k \parallel^2}{\sum_{(i,j)\in\mathcal{U}\setminus(k,l)} \alpha_{i,j}p \parallel \mathbf{h}_k^H\mathbf{h}_i \parallel^2 / \parallel \mathbf{h}_k \parallel^2 + \sigma^2}.\end{aligned} \tag{3.10}$$

When $r_{k,l}$ satisfies the following constraints, MRC can successfully detect the data stream $s_{k,l}$.

$$\begin{aligned}&r_{k,l} \leq \mathcal{R}_{(k,l),\mathcal{U}}^{\mathrm{MRC}}\\\triangleq&\log_2\left(\frac{\sum_{(i,j)\in\mathcal{U}} \alpha_{i,j}p \parallel \mathbf{h}_k^H\mathbf{h}_i \parallel^2 / \parallel \mathbf{h}_k \parallel^2 + \sigma^2}{\sum_{(i,j)\in\mathcal{U}\setminus(k,l)} \alpha_{i,j}p \parallel \mathbf{h}_k^H\mathbf{h}_i \parallel^2 / \parallel \mathbf{h}_k \parallel^2 + \sigma^2}\right).\end{aligned} \tag{3.11}$$

In addition, when the above formulas hold for $\forall(k,l) \in \mathcal{D}$, all data streams in the set $\mathcal{D}$ can be detected by MRC simultaneously.

3.3 Achievable Data Rate Based on Stable SIC Detection

This section describes the conditions for SIC stability and the theoretical achievable data rate region by MMSE-SIC and MRC-SIC detection, which plays an essential role in achieving multi-layer superposition transmission to approach symmetric capacity.

3.3.1 Conditions for Stable SIC Detection

In the case of single-layer SIC detection, the receiver detects one specified data stream at a time in a particular order and continuously subtracts the successfully detected data streams. Here, a mapping function $\pi : \mathcal{K}_{K,L} \longmapsto \mathcal{L}_{KL} = \{1, 2, ..., KL\}$ is defined to map the index of one data stream to the detection order of this data stream at the gNB. $\pi((k,l)) < \pi((i,j))$ specifies that the layer l of user k is detected before the layer j of user i. Meanwhile, the UL multi-layer MU-MIMO NOMA system can be represented by a K-user L-layer configuration $\mathcal{F}(\mathbf{P}_{K,L}, \mathbf{H}_K, \mathbf{R}_{K,L}, \sigma^2)$, including the allocated power $\mathbf{P}_{K,L}$, user data rates $\mathbf{R}_{K,L}$, channel coefficients $\mathbf{H}_K$, and noise power σ^2.

Here, the conditions for stable SIC detection are given. UL user layers can be stably demodulated by MMSE-SIC when there is a mapping function $\pi : \mathcal{K}_{K,L} \longmapsto \mathcal{L}_{KL}$ satisfying the following formula.

$$r_{k,l} \leq \mathcal{R}^{\mathrm{MMSE}}_{(k,l),\mathcal{W}_{(k,l)}}, \quad \forall (k,l) \in \mathcal{K}_{K,L}, \tag{3.12}$$

where $\mathcal{W}_{(k,l)} = \{(i,j) | \pi((i,j)) \geq \pi((k,l))\}$ represents the set of remaining data streams when $s_{k,l}$ is being detected. At the same time, if the following formula holds, the UL system can be stably detected by MRC-SIC.

$$r_{k,l} \leq \mathcal{R}^{\mathrm{MRC}}_{(k,l),\mathcal{W}_{(k,l)}}, \quad \forall (k,l) \in \mathcal{K}_{K,L}. \tag{3.13}$$

3.3.2 Minimum Achievable User Data Rate by MMSE-SIC

Let $\mathcal{U}$ denote the set of remaining data streams, and the supremum of the sum data rate of all data streams in $\mathcal{D}$ is given by:

$$\mathcal{R}^{\mathrm{SUP}}_{\mathcal{D},\mathcal{U}} = \log_2 \det \left(\mathbf{I}_{|\mathcal{D}|} + \mathbf{H}^H_{\mathcal{D}} \mathbf{F}^{-1}_{\mathcal{U}\backslash\mathcal{D}} \mathbf{H}_{\mathcal{D}} \right), \tag{3.14}$$

where $\mathcal{R}^{\mathrm{SUP}}_{\mathcal{D},\mathcal{D}} = \log_2 \det \left(\mathbf{I}_M + \frac{1}{\sigma^2} \mathbf{H}_{\mathcal{D}} \mathbf{H}^H_{\mathcal{D}} \right)$. In fact, $\mathcal{R}^{\mathrm{SUP}}_{\mathcal{D},\mathcal{U}}$ can be achieved by a single-layer MMSE-SIC of any detection order, as described in the following theorem.

Theorem 1. *Applying an arbitrary mapping function $\pi : \mathcal{D} \longmapsto \mathcal{L}_{|\mathcal{D}|} = \{1, 2, ..., |\mathcal{D}|\}$, the supremum of the sum data rate of data streams in $\mathcal{D}$ can be reached by the single-layer MMSE-SIC detection. It can be expressed as*

$$\sum\nolimits_{(k,l)\in\mathcal{D}} \mathcal{R}^{MMSE}_{(k,l),\mathcal{W}_{(k,l)}} = \mathcal{R}^{SUP}_{\mathcal{D},\mathcal{U}}, \tag{3.15}$$

where $\mathcal{W}_{(k,l)} = \mathcal{U}\backslash\{(i,j) | \pi((i,j)) < \pi((k,l))\}$ represents the remaining data streams at the time of detecting $s_{k,l}$.

Proof. Without loss of generality, an arbitrary mapping $\pi : \mathcal{D} \longmapsto \mathcal{L}_{|\mathcal{D}|}$ can be implemented to decide the detection order $(k_1, l_1) \rightarrow (k_2, l_2) ... \rightarrow (k_{|\mathcal{D}|}, l_{|\mathcal{D}|})$. When

adopting the successive detection process, the right-hand side (RHS) of (3.15) can be rewritten as follows

$$\begin{aligned}\mathcal{R}_{\mathcal{D},\mathcal{U}}^{\mathrm{SUP}} &= \log_2 \det\left(\mathbf{I}_{|\mathcal{D}|} + \mathbf{H}_{\mathcal{D}}^H \mathbf{F}_{\mathcal{U}\setminus\mathcal{D}}^{-1} \mathbf{H}_{\mathcal{D}}\right) \\ &= \log_2 \det\begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{pmatrix} \\ &= \log_2 \left(\det(\mathbf{A}) \det(\mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B})\right),\end{aligned} \tag{3.16}$$

where $\mathbf{A} = \mathbf{I}_{|\mathcal{D}|-1} + \mathbf{H}_{\mathcal{D}\setminus(k_1,l_1)}^H \mathbf{F}_{\mathcal{U}\setminus\mathcal{D}}^{-1} \mathbf{H}_{\mathcal{D}\setminus(k_1,l_1)}$, $\mathbf{B} = \vartheta \mathbf{H}_{\mathcal{D}\setminus(k_1,l_1)}^H \mathbf{F}_{\mathcal{U}\setminus\mathcal{D}}^{-1} \mathbf{h}_{k_1}$, $\mathbf{C} = \vartheta \mathbf{h}_{k_1}^H \mathbf{F}_{\mathcal{U}\setminus\mathcal{D}}^{-1} \mathbf{H}_{\mathcal{D}\setminus(k_1,l_1)}$, and $\mathbf{D} = 1 + \vartheta^2 \mathbf{h}_{k_1}^H \mathbf{F}_{\mathcal{U}\setminus\mathcal{D}}^{-1} \mathbf{h}_{k_1}$, and $\vartheta = \sqrt{\alpha_{k_1,l_1} p}$.

Then, the Sherman-Morrison-Woodbury formula [74] can be performed. Hence, we have,

$$\begin{aligned}&\det(\mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B}) \\ =&\mathbf{D} - \mathbf{C}\left(\mathbf{I}_{|\mathcal{D}|-1} - \mathbf{H}_{\mathcal{D}\setminus(k_1,l_1)}^H \mathbf{X} \mathbf{F}_{\mathcal{U}\setminus\mathcal{D}}^{-1} \mathbf{H}_{\mathcal{D}\setminus(k_1,l_1)}\right)\mathbf{B} \\ =&\mathbf{D} - \mathbf{C}\mathbf{H}_{\mathcal{D}\setminus(k_1,l_1)}^H \mathbf{X} \mathbf{F}_{\mathcal{U}\setminus\mathcal{D}}^{-1} \mathbf{h}_{k_1} \\ =&1 + \vartheta^2 \mathbf{h}_{k_1}^H \left(\mathbf{F}_{\mathcal{U}\setminus\mathcal{D}} + \mathbf{H}_{\mathcal{D}\setminus(k_1,l_1)} \mathbf{H}_{\mathcal{D}\setminus(k_1,l_1)}^H\right)^{-1} \mathbf{h}_{k_1} \\ =&1 + \vartheta^2 \mathbf{h}_{k_1}^H \mathbf{F}_{\mathcal{U}\setminus(k_1,l_1)}^{-1} \mathbf{h}_{k_1},\end{aligned} \tag{3.17}$$

where $\mathbf{X} = \left(\mathbf{I}_M + \mathbf{F}_{\mathcal{U}\setminus\mathcal{D}}^{-1} \mathbf{H}_{\mathcal{D}\setminus(k_1,l_1)} \mathbf{H}_{\mathcal{D}\setminus(k_1,l_1)}^H\right)^{-1}$.

Through the above derivation, (3.16) can be further expanded as

$$\begin{aligned}\mathcal{R}_{\mathcal{D},\mathcal{U}}^{\mathrm{SUP}} &= \log_2 \left(\det(\mathbf{A}) \det(\mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B})\right) \\ =&\log_2 \prod\nolimits_{1 \le i \le |\mathcal{D}|} \left(1 + \alpha_{k_i,l_i} p \mathbf{h}_{k_i}^H \mathbf{F}_{\mathcal{U}\setminus\{(k_1,l_1),\ldots,(k_i,l_i)\}}^{-1} \mathbf{h}_{k_i}\right) \\ =&\sum\nolimits_{(k,l)\in\mathcal{D}} \log_2 \left(1 + \alpha_{k,l} p \mathbf{h}_k^H \mathbf{F}_{\mathcal{W}_{(k,l)}\setminus(k,l)}^{-1} \mathbf{h}_k\right) \\ =&\sum\nolimits_{(k,l)\in\mathcal{D}} \mathcal{R}_{(k,l),\mathcal{W}_{(k,l)}}^{\mathrm{MMSE}}.\end{aligned} \tag{3.18}$$

This proves the theorem. □

In particular, for a subset of the admitted users $\mathcal{S}$, MMSE-SIC detection can achieve the supremum of the sum data rate of the users from $\mathcal{S}$ ($\forall k \in \mathcal{S}$, $(k,l) \in \mathcal{U}, \mathcal{D}$

always holds), which is given by

$$\sum\nolimits_{k\in\mathcal{S}}\sum\nolimits_{l=1}^{L}\mathcal{R}^{\mathrm{MMSE}}_{(k,l),\mathcal{W}_{(k,l)}}=\mathcal{C}_{\mathcal{S}}(\mathbf{p}_K,\mathbf{H}_K,\sigma^2). \tag{3.19}$$

As can be seen, the sum of the data rates remains unchanged when the detection order of two adjacent data streams is changed. In fact, MMSE-SIC with enough layers can be implemented to approach the symmetric capacity, as shown in the following theorem.

Theorem 2. *For arbitrary $\epsilon > 0$ and a K-user configuration $\mathcal{F}(\hat{\mathbf{p}}_K,\hat{\mathbf{H}}_K,\hat{\mathbf{r}}_K,\sigma^2)$ that satisfies*

$$\hat{r}_k\le\mathcal{C}_{sym}(\hat{\mathbf{p}}_K,\hat{\mathbf{H}}_K,\sigma^2)-\epsilon,\quad\forall k\in\{1,2,...,K\}, \tag{3.20}$$

there always exits a λ_K. When $L\ge\lambda_K$, at least one MMSE-SIC stable K-user L-layer configuration $\mathcal{F}(\hat{\mathbf{P}}_{K,L},\hat{\mathbf{H}}_K,\hat{\mathbf{R}}_{K,L},\sigma^2)$ meets the following constraints,

$$\hat{p}_k=\sum\nolimits_{l=1}^{L}\hat{p}_{k,l}, \tag{3.21a}$$

$$\hat{r}_k=\sum\nolimits_{l=1}^{L}\hat{r}_{k,l}. \tag{3.21b}$$

Proof. The following lemma is claimed for a concise proof. To derive the original Theorem 2 from Lemma 1 is straightforward.

Lemma 1. *For arbitrary $\epsilon>0$, there exists a $\lambda_K<\infty$. $\forall L\ge\lambda_K$, at least one MMSE-SIC stable one-user L-layer configuration $\mathcal{F}(\hat{\mathbf{P}}_{K,L},\hat{\mathbf{H}}_K,\hat{\mathbf{R}}_{K,L},\sigma^2)$ satisfies* (3.21a), (3.21b), *and $C\left(\hat{p}_{k,l}\parallel\mathbf{h}_k\parallel^2,\sigma^2\right)\triangleq\log_2(1+\hat{p}_{k,l}\parallel\mathbf{h}_k\parallel^2/\sigma^2)\le\epsilon,\forall k,l$.*

When $K=1$, Theorem 2 is true with $L\ge1$. Further, for arbitrary $\epsilon>0$, there exists a $\lambda_1<\infty$, Lemma 1 is true.

When $K=2$, it is easy to verify Theorem 2 is true when $L\ge2$. For $\hat{r}_k\le\mathcal{C}_{\mathrm{sym}}(\hat{\mathbf{p}}_2,\hat{\mathbf{H}}_2,\sigma^2),k=1,2$, we can construct a new two-user two-layer configuration $\mathcal{F}(\hat{\mathbf{P}}_{2,2},\hat{\mathbf{H}}_2,\hat{\mathbf{R}}_{2,2},\sigma^2)$. The power allocated to layers is set as $\hat{p}_{1,1}=\hat{p}_1$, $\hat{p}_{1,2}=0$, $\hat{p}_{2,1}=\delta_2\hat{p}_2$, $\hat{p}_{2,2}=(1-\delta_2)\hat{p}_2$, and δ_2 is the unique number that satisfies $\hat{r}_1=\log_2\left(1+\hat{p}_1\parallel\mathbf{h}_1^H(\sigma^2\mathbf{I}_M+\delta_2\hat{p}_2\mathbf{h}_2\mathbf{h}_2^H)^{-1}\mathbf{h}_1\parallel\right)$. The data rates are set as $\hat{r}_{1,1}=\hat{r}_1$,

$\hat{r}_{1,2} = 0$, and $\hat{r}_{2,1} = \log_2\left(1 + \hat{p}_{2,1} \parallel \mathbf{h}_2 \parallel^2 / \sigma^2\right)$, $\hat{r}_{2,2} = \hat{r}_2 - \hat{r}_{2,1}$. From (3.20) and (3.2), we have $0 \leq \delta_2 \leq 1$. It is straightforward to prove that $\hat{p}_1 = \hat{p}_{1,1} + \hat{p}_{1,2}$, $\hat{p}_2 = \hat{p}_{2,1} + \hat{p}_{2,2}$, $\hat{r}_1 = \hat{r}_{1,1} + \hat{r}_{1,2}$, $\hat{r}_2 = \hat{r}_{2,1} + \hat{r}_{2,2}$ hold, and thus the constraints in (3.21a) and (3.21b) are satisfied. In addition, with the detection order $(2,2) \rightarrow (1,1) \rightarrow (1,2) \rightarrow (2,1)$, this new configuration is MMSE-SIC stable. Hence, for arbitrary $\epsilon > 0$, there exists a $\lambda_2 < \infty$, we can further split the layers of each user into no more than λ_2 layers, without changing the detection order and sum data rate, to satisfy Lemma 1.

When $K = K_0 + 1$, we assume that Lemma 1 is true for $1 \leq K \leq K_0$. Due to space limitation, we briefly prove that Lemma 1 is true for $K = K_0 + 1$ as follows.

First, without loss of generality, we assume that the user with the largest receive power $\hat{p}_k \| \mathbf{h}_k \|^2$ is user $K_0 + 1$, $\mathcal{S}_{K_0} = \{1, 2, ..., K_0\}$, $\mathcal{S}_{K_0+1} = \mathcal{S}_{K_0} \cup \{K_0 + 1\}$, the symmetric capacity of $\mathcal{S}_{K_0}$ is $\bar{r}_{K_0} = \mathcal{C}_{\text{sym}}(\hat{\mathbf{p}}_{K_0}, \hat{\mathbf{H}}_{K_0}, \sigma^2)$, and the symmetric capacity of $\mathcal{S}_{K_0+1}$ is $\bar{r}_{K_0+1} = \mathcal{C}_{\text{sym}}(\hat{\mathbf{p}}_{K_0+1}, \hat{\mathbf{H}}_{K_0+1}, \sigma^2)$.

Considering the symmetric capacity is lower with more limitations, $\bar{r}_{K_0+1} \leq \bar{r}_{K_0}$. From Lemma 1, for arbitrary $\epsilon > 0$, there exists an integer $\lambda_{K_0} < \infty$, we have an MMSE-SIC stable K_0-user λ_{K_0}-layer configuration $\mathcal{F}(\hat{\mathbf{P}}_{K_0,\lambda_{K_0}}, \hat{\mathbf{H}}_{K_0}, \hat{\mathbf{R}}_{K_0,\lambda_{K_0}}, \sigma^2)$ satisfies $\hat{p}_k = \sum_{l=1}^{\lambda_{K_0}} \hat{p}_{k,l}$, $\sum_{l=1}^{\lambda_{K_0}} \hat{r}_{k,l} = \bar{r}_{K_0} - \epsilon, \forall k \in \mathcal{S}_{K_0}$, and $C\left(\hat{p}_{k,l} \parallel \mathbf{h}_k \parallel^2, \sigma^2\right) \leq \epsilon, \forall k \in \mathcal{S}_{K_0}, \forall l \in \{1, ..., \lambda_{K_0}\}$.

If the achievable data rate of user $K_0 + 1$ is higher than $\bar{r}_{K_0+1}$, we have

$$\log_2\left(1 + \hat{p}_{K_0+1} \parallel \mathbf{h}_{K_0+1}^H \mathbf{F}_{\mathcal{K}_{K_0,\lambda_{K_0}}}^{-1} \mathbf{h}_{K_0+1} \parallel\right) > \bar{r}_{K_0+1}.$$

Then, Theorem 2 is definitely true when user $K_0 + 1$ is detected first, and we can split user $K_0 + 1$ into no more than λ_{K_0+1} layers to ensure Lemma 1 is true.

Otherwise, we gradually increase the detection order of the un-layered user K_0+1 from 1 to $K\lambda_{K_0} + 1$. The period between adjacent increments is called a phase, and $\hat{r}_{K_0+1} \geq \bar{r}_{K_0+1}$ always holds at the last phase.

At each phase, we select user $K^{'}$ with the largest achievable data rate, and select the latest detected layer $L^{'}$ of user $K^{'}$. Then, the layer $(K^{'}, L^{'})$ is moved to be

detected first without changing the relative order of the other layers of all the users. Next, the current latest detected layer $L^{''}$ of user $K^{'}$ is chosen and moved to be detected first. From (3.19), each movement of the layer of user $K^{'}$ increases the achievable data rates of all the other users. The movement of the layers of user $K^{'}$ is undertook until $\bar{r}_{K_0+1} - \epsilon \leq \sum_{l=1}^{\lambda_{K_0}} \hat{r}_{K^{'},l} < \bar{r}_{K_0+1}$ holds. After that, we select user $K^{''}$ with the largest achievable data rate, and repeat the above operations. When the detection order of user $K_0 + 1$ increases, the next phase starts.

Then, there always exists a phase, including the final phase in which user K_0+1 is last detected, and the detection order of user $K_0 + 1$ cannot increase anymore. In this phase, the users with at least one layer detected after user $K_0 + 1$ cannot be selected as the user acquiring the largest achievable data rate. Suppose they belong to the set $\mathcal{S}^{'} \subseteq \mathcal{S}_{K_0}$. Considering the movement can always be performed, the users in $\mathcal{S}^{''} \subseteq \mathcal{S}_{K_0} \setminus \mathcal{S}^{'}$ are successfully detected with achievable data rate larger than $\bar{r}_{K_0+1} - \epsilon$ and the other users in $\mathcal{S}_{K_0+1} \setminus \mathcal{S}^{''}$. Therefore, the users in $\mathcal{S}^{''}$ can be detected while the constraints are satisfied and the detection order of the layers of the other users does not change. When we split user $K_0 + 1$ to ensure the users in $\mathcal{S}_{K_0+1} \setminus \mathcal{S}^{''}$ satisfying the constraints, the case is equivalent to $K = |\mathcal{S}_{K_0+1} \setminus \mathcal{S}^{''}|$. Known from Lemma 1, we can spit the user $K_0 + 1$ into at most $\lambda_{|\mathcal{S}_{K_0+1} \setminus \mathcal{S}^{''}|}$ layers, to maintain $\hat{r}_{K_0+1} = \bar{r}_{|\mathcal{S}_{K_0+1} \setminus \mathcal{S}^{''}|} - \epsilon \geq \bar{r}_{K_0+1} - \epsilon$. Therefore, Theorem 2 is true when $K = K_0 + 1$.

Finally, we split the user $K_0 + 1$ into at most $\lambda_{K_0+1} < \infty$ layers to satisfy $C\left(\hat{p}_{K_0+1,l} \parallel \mathbf{h}_{K_0+1} \parallel^2, \sigma^2\right) \leq \epsilon, \forall l$. Thus, Lemma 1 is true when $K = K_0 + 1$.

In summary, Lemma 1 can be proved, and Theorem 2 is also true. □

When the number of layers is large enough, the symmetric capacity can be approached by rate splitting and stable MMSE-SIC detection.

3.3.3 Minimum User Data Rate Achieved by MRC-SIC

From (3.10), the equivalent interference power is lower than the total power of the other remaining data streams. To this end, a LB for the achievable sum data

rate of the data streams in $\mathcal{D}$ is

$$\mathcal{R}^{\mathrm{LB}}_{\mathcal{D},\mathcal{U}} = C\Big(\sum_{(k,l)\in\mathcal{D}} p_{k,l} \parallel \mathbf{h}_k \parallel^2, \sum_{(i,j)\in\mathcal{U}\setminus\mathcal{D}} p_{i,j} \parallel \mathbf{h}_i \parallel^2 + \sigma^2 \Big). \tag{3.22}$$

$\mathcal{R}^{\mathrm{LB}}_{(k,l),\mathcal{W}_{(k,l)}} \leq \mathcal{R}^{\mathrm{MRC}}_{(k,l),\mathcal{W}_{(k,l)}}$ can always be verified as true. Considering the condition of SIC stability, when there is a mapping function π and the K-user L-layer UL configuration $\mathcal{F}(\mathbf{P}_{K,L}, \mathbf{H}_K, \mathbf{R}_{K,L}, \sigma^2)$ satisfies $r_{k,l} \leq \mathcal{R}^{\mathrm{LB}}_{(k,l),\mathcal{W}_{(k,l)}}, \forall (k,l) \in \mathcal{K}_{K,L}$, the UL configuration can be stably detected by MRC-SIC at the receiver.

3.4 Maximizing the Minimum User Data Rate by Rate Splitting

In this section, a receiver-centric algorithm is designed to update the split power and data rate of each user to increase the minimum data rate for UL users. The proposed solution has two advantages. Firstly, UL users' minimum data rate is increased by the scheduling of the gNB, thus ensuring fairness among UL users. Secondly, when the target data rate of all potential users is given, more users can be admitted to access the gNB simultaneously. To ensure stable MMSE-SIC detection, the gNB can calculate the optimal rate split configuration for each user after receiving the requested data rate and obtaining CSI from potential users. The gNB then announces the power allocation and data rate allocation factors to all users during transmission. The gNB scheduling process of rate splitting in MMSE-SIC is presented here. The data rate of each layer $r_{k,l} = \beta_l \bar{r}, (1 \leq l \leq L)$ is determined by the target data rate of user $\bar{r}$, where β_l is a predefined rate splitting factor, and the rate splitting factor can be set to $\beta_l = \dfrac{2l}{L(1+L)}$. At each stage, the gNB selects a user and split its power into two parts, one for supporting the target data rate of the current layer, and the other one is the unallocated power. First, the gNB sorts the users in descending order of the equivalent SNRs. Second, the gNB estimates the data rate gap for each user, which is equal to the difference between the achievable data rate of unallocated power and the sum of the required data rates of unallocated layers. Third, the user with the minimum data rate gap is selected, and one layer is

Algorithm 1 The Multi-Layer Rate Splitting scheme to Support Stable MMSE-SIC

Initialize $\mathcal{U} = \mathcal{K}_{K,L}$, $t = 1$, $\mathcal{V} = \varnothing$, $\hat{\mathbf{p}}_K$, $\hat{\mathbf{H}}_K$, and $\mathbf{R}_{K,L}$
repeat
  Achievable data rate$\hat{r}_k = \log_2\left(1 + \hat{p}_k \parallel \hat{\mathbf{h}}_k^H \mathbf{F}_{\mathcal{V}}^{-1} \hat{\mathbf{h}}_k \parallel\right)$
  $\delta_{k,l} = \min_{(i,j)\in\mathcal{U}} (\hat{r}_k - r_{i,j})$
  $(k,l) = \arg\min_{(i,j)\in\mathcal{U}} (\hat{r}_k - r_{i,j})$
  Allocate data rate $\hat{r}_{k,l} = \min\{\hat{r}_k, r_{k,l}\}$, $\mathcal{V}_t = (k,l)$
  Allocate power $\hat{p}_{k,l} = \dfrac{2^{\hat{r}_{k,l}} - 1}{\parallel \hat{\mathbf{h}}_k^H \mathbf{F}_{\mathcal{V}}^{-1} \hat{\mathbf{h}}_k \parallel}$
  Update $\mathcal{U} \leftarrow \mathcal{U} \setminus \mathcal{V}_t$ and $\mathcal{V} \leftarrow \mathcal{V} \cup \mathcal{V}_t$ for next rate splitting
  $t = t + 1$, $\hat{p}_k = \hat{p}_k - \hat{p}_{k,l}$
until $\mathcal{U} = \varnothing$
$t = 1$
repeat
  Determine the SIC detection order $\pi(\mathcal{V}_t) = KL - t$, $t = t + 1$
until $t = KL$
Output the allocated power $\hat{\mathbf{P}}_{K,L}$, allocated data rates $\hat{\mathbf{R}}_{K,L}$ and the SIC detection order π

split for this user. If the achievable data rate of unallocated power is greater than or equal to the target data rate of the layer, the unique power value that achieves the desired data rate is assigned to this layer. The allocated data rate for that layer is equal to the target data rate. Otherwise, all unallocated power will be allocated to the layer, and this layer's allocated data rate is equal to the unallocated power's achievable data rate. Finally, the process terminates when all layers of all users have been determined with allocated power and data rates. After the detection order and the data rate splitting factors at the gNB are obtained, the feedback of rate splitting will be transferred to UL users to improve UL transmission. As shown in Algorithm 1, the optimal detection order can be determined. The power and data rates of UL users can approach the symmetric capacity with a stable MMSE-SIC detection.

The similar receiver-centric algorithm can be designed to increase the minimum data rate for UL users when MRC-SIC detection is adopted.

Table 3.1 : Simulation Parameters

Parameter	Value
Bandwidth (MHz)	1
Cell range (m)	30∼150
Data packet size (Byte)	64
Noise power (dB)	-110
Tx power (dBm)	-5∼15
Path loss exponent	3
Antenna configuration	1Tx, 4Rxs

3.5 Simulation Results and Analysis

Theoretical calculations and simulation results are demonstrated here to verify the above findings. Comparing the minimum user data rates of different schemes, it is verified that the proposed schemes can approach symmetric capacity. Then, the number of UL users and the complementary cumulative distribution function (CCDF) of the transmission latency are presented. It is assumed that UL users are evenly dispersed in the cell with their distances to the gNB ranged from 30 m to 150 m, and the channel coefficients are randomly generated for 1000 times. Table 3.1 lists the detailed parameters of the simulation.

3.5.1 Maximizing the Minimum User Data Rate

Fig. 3.2 illustrates the minimum data rate of UL users for different schemes in MU-MIMO NOMA. MMSE-SIC NOMA can achieve nearly 90% of the symmetric capacity, while MRC-SIC NOMA can only achieve 70% of the symmetric capacity. At the same time, when Tx power is high, the proposed schemes are significantly superior to the existing ordered SIC (OSIC) NOMA and OMA schemes. Notably, the minimum user data rates for OSIC NOMA and OMA decrease significantly at this time. The minimum user data rate for different numbers of admitted users are also compared. Generally, MMSE-SIC NOMA can approach the symmetric capacity with increasing admitted users.

Fig. 3.3 demonstrates the maximum number of admitted users when there are 20 potential users. The admitted users can be successfully detected with the required

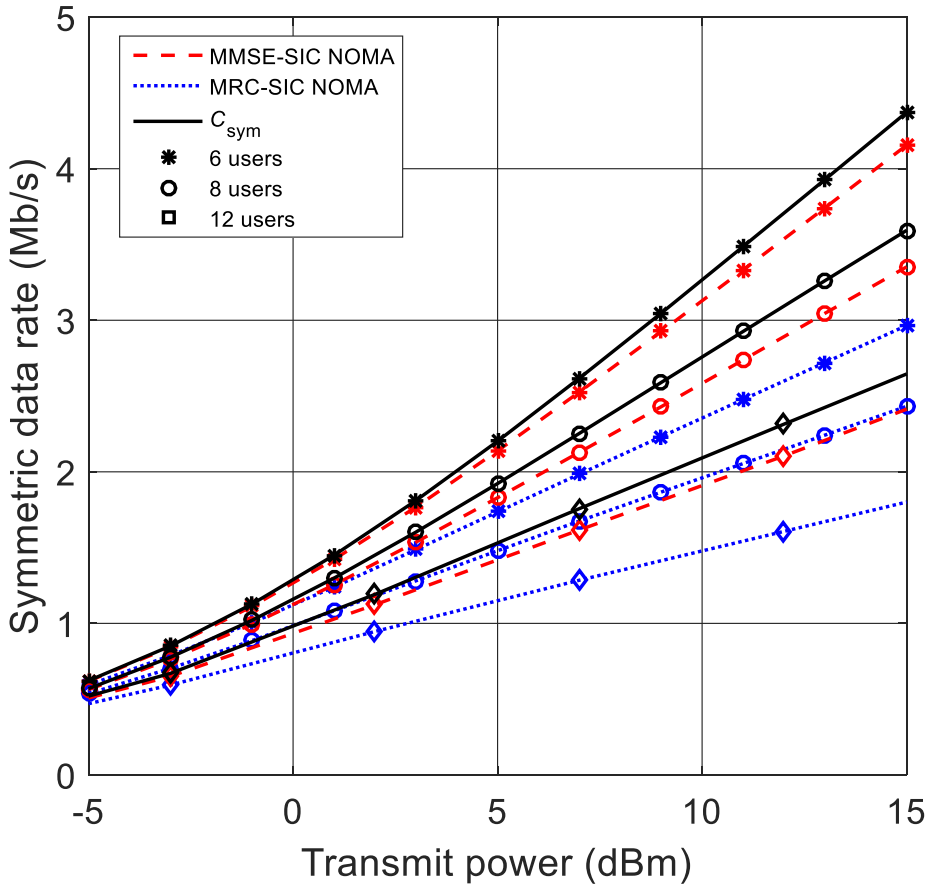

Figure 3.2 : Minimum user data rates achieved by different schemes when each user is split to 6 layers.

data rate $\bar{r}$ =2 bit/s/Hz. MMSE-SIC NOMA and MRC-SIC NOMA indicate great advantages to the existing schemes. When Tx power is 10 dBm, MMSE-SIC NOMA can support three times more admitted users than OSIC NOMA. Meanwhile, as Tx power increases, the maximum number of admitted users in the proposed schemes increases.

3.5.2 Reduce Detection Complexity and Latency

The reduction in computational complexity plays an essential role in the realization of SIC detection, especially when the number of UL users is large. Therefore,

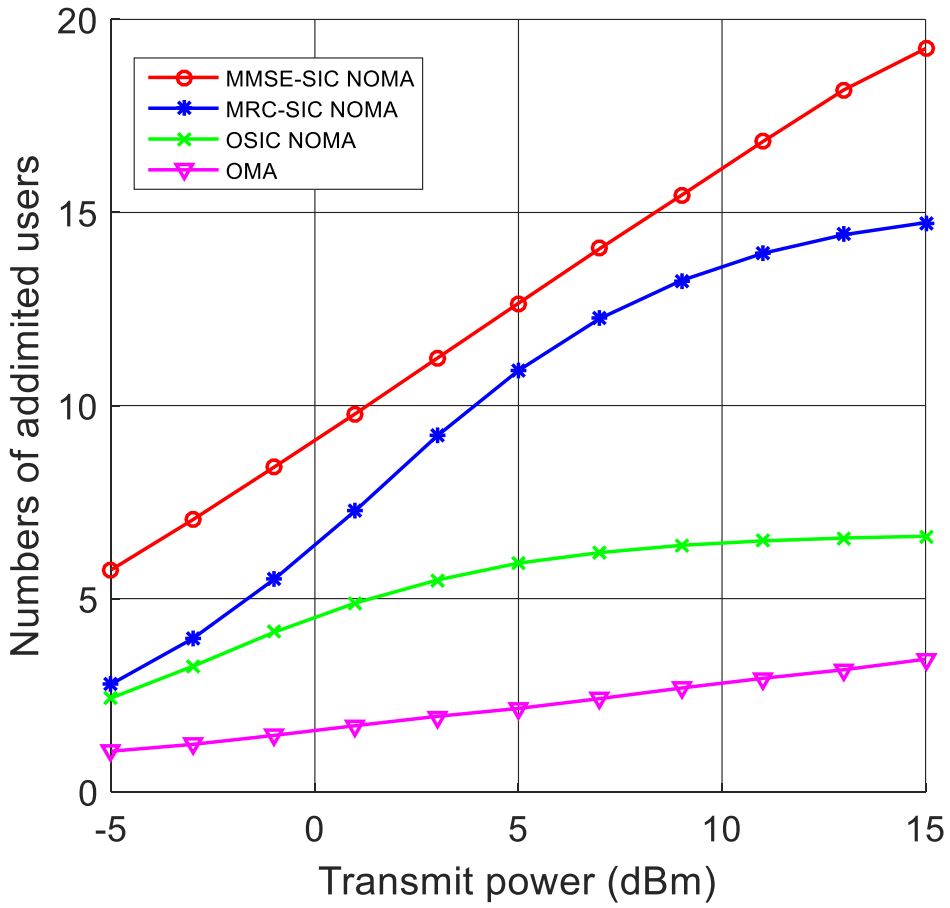

Figure 3.3 : Maximum number of admitted users when there are 20 potential users.

it leads to the limitation of the number of matrix inversions and SIC operations to a tolerable level, which is usually determined by the number of admitted users and the number of split layers. In addition, the processing latency decreases with the reduction of the computational complexity and the number of SICs. On the one hand, when a moderate number of admitted users exist, the data of each user can be split into enough layers and MMSE-SIC detection tends to approach the symmetric capacity closely. Then, the matrix inversion should be performed for $\mathcal{O}(KL)$ times, and the number of single-layer SIC operations can be $\mathcal{O}(KL)$. If group SIC detection is implemented at the receiver, the matrix inversion operations can still be $\mathcal{O}(KL)$, and the number of group SIC times is drastically reduced to approxi-

mate $\mathcal{O}(K^{\frac{1}{2}}L^{\frac{1}{2}})$. Moreover, when massive users are admitted to accessing the gNB, the data can be divided into only two layers and detected by MRC-SIC at the receiver. Therefore, the number of matrix inversion operations and single-layer SICs decreases sharply to $\mathcal{O}(1)$ and $\mathcal{O}(2K)$ and respectively. Meanwhile, the number of SICs for OSIC NOMA is $\mathcal{O}(K)$. Although the symmetric capacity of MU-MIMO NOMA can be achieved by nonlinear and iterative linear detection, they have very high computational complexity, usually from $\mathcal{O}(K^2)$ to $\mathcal{O}(K^3)$.

Considering the detection order, the data stream of one UL user may not be detected until a partial data stream has been successfully detected and subtracted.

Fig. 3.4 illustrates the reduction in the number of SICs when group SIC detection is implemented. When there are K users and L layers, the single-layer SIC should be executed for $KL-1$ times. The ratio of grouped SIC times to single-layer SIC times reflects the reduction in processing latency. Although the number of single-layer SICs increases as the number of admitted users increases, the ratio of grouped SIC times to single-layer SIC times reduces. Hence, group SIC detection has a sharp decrease in the number of SICs with a lower processing latency than single-layer SIC detection.

3.5.3 Reducing Transmission Latency

Fig. 3.5 presents the CCDF of transmission latency when the packet size is 64 bytes. The proposed scheme achieves the lowest transmission latency with negligible fluctuation among users. It is always superior to OMA. Meanwhile, although 50% of OSIC NOMA users can achieve almost the same transmission latency as MMSE-SIC NOMA users, 20% of OSIC NOMA users have much higher transmission latency than MMSE-SIC NOMA users. Therefore, rate splitting effectively increases fairness among UL users.

Fig. 3.6 illustrates the CCDF of transmission latency with different numbers of admitted users. When the number of admitted users rises, the transmission latency of all UL transmission schemes increases. In particular, MMSE-SIC NOMA and MRC-SIC NOMA have a much lower increment in transmission latency when the

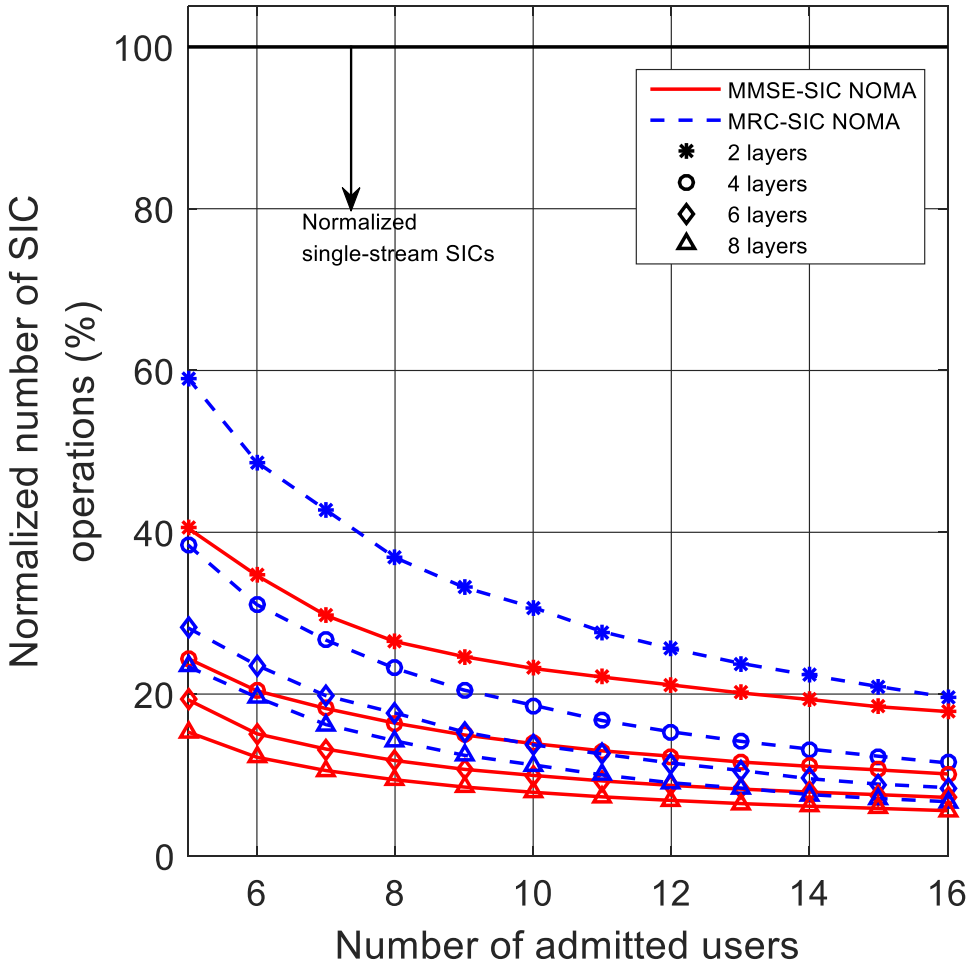

Figure 3.4 : An example of the decrease in the number of SICs when using group SIC detection.

number of admitted users becomes greater than it is in the existing schemes. When 12 admitted users exist, the probability of transmission latency exceeding 0.4 ms is less than 1% in the proposed schemes. Therefore, in IoT applications with massive potential users, MMSE-SIC NOMA and MRC-SIC NOMA are more competitive than existing schemes to satisfy the constraints on latency.

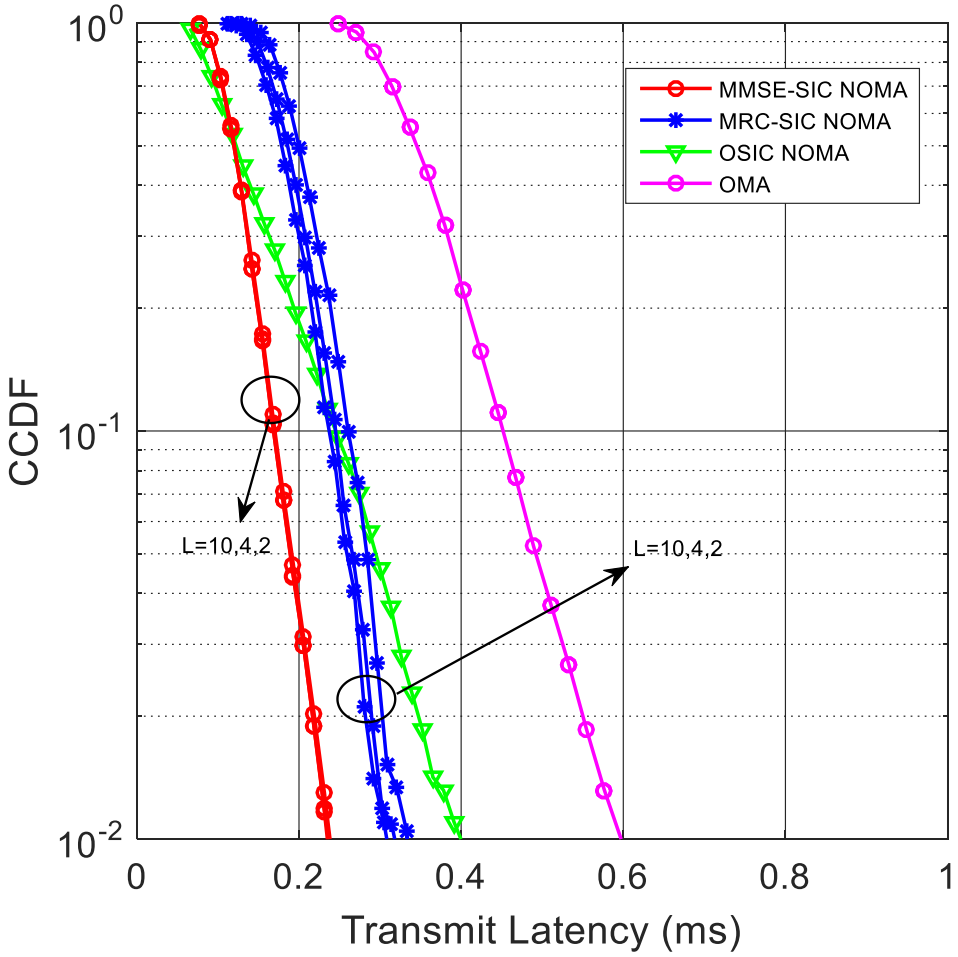

Figure 3.5 : CCDF of transmission latency with 8 users.

3.6 Summary of This Chapter

This chapter has presented the multi-layer superposition transmission enabled by rate splitting and MMSE-SIC/MRC-SIC detection. Besides, their advantages have been demonstrated by approaching the symmetric capacity in MU-MIMO NOMA. For high performance, high precision or even perfect CSI is assumed to be obtained at the BS. The detection algorithm with imperfect CSI will be analyzed and discussed in Chapter 5. This chapter has proposed a new rate splitting scheme to approach the symmetric capacity by MMSE-SIC detection. As presented in the

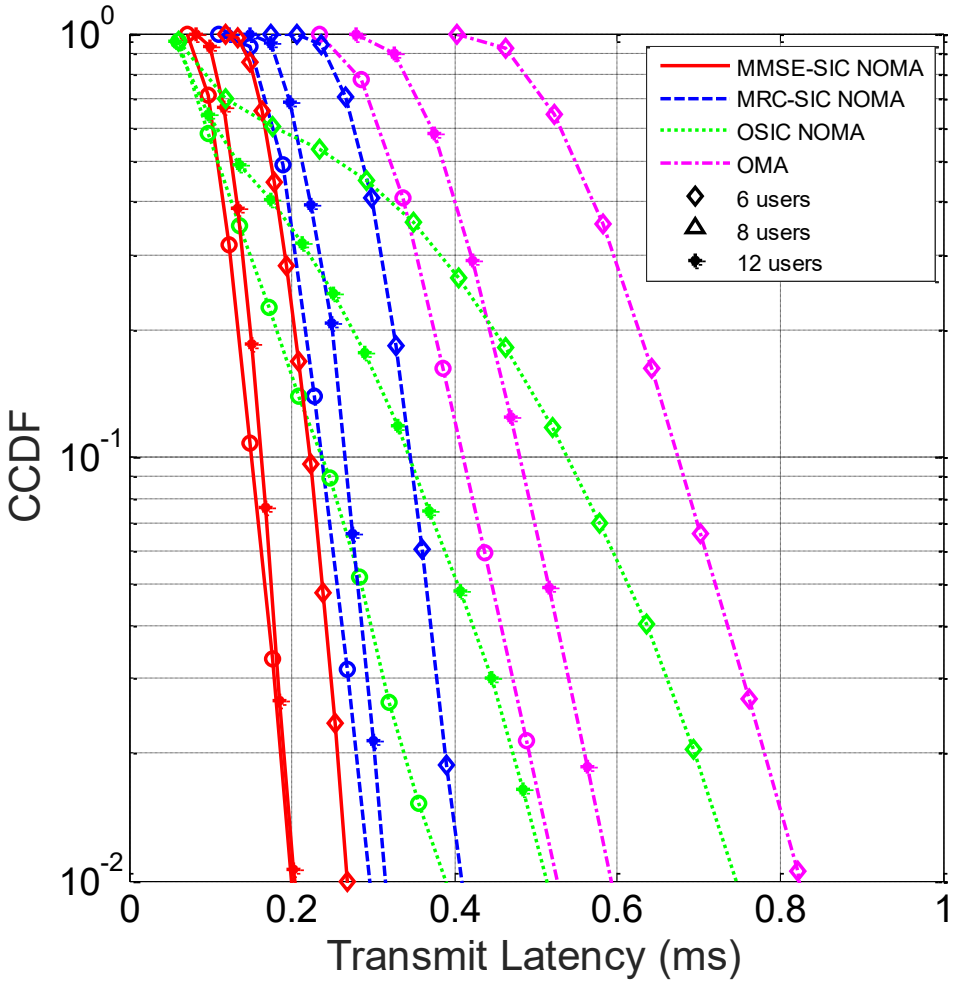

Figure 3.6 : CCDF of transmission latency with 6-layer rate splitting.

simulation results, MMSE-SIC detection can achieve 90% of the symmetry capacity without a significant increase in computational complexity and latency. Therefore, the proposed multi-layer superposition transmission based on SIC detection is superior to the OMA and OSIC NOMA schemes in maximizing the minimum user data rate. It has outstanding application prospects in the future IoT applications with massive UL users.

Chapter 4

Supporting Short-packet Transmission with FD-SCMA

This chapter combines the FD and SCMA techniques to reduce the latency and utilizes the diversity gain of multiple Rx antennas to improve reliability. Firstly, FD-SCMA is proposed to support URLLC for transmission in UL and DL simultaneously, and spatial diversity is applied to further enhance reliability by installing multiple Rx antennas. This chapter derives the effective SNR under imperfect SIS to analyze reliability, and then calculates the error probability in UL and DL, respectively. In the time-invariant flat fading channel, this chapter demonstrates that FD-SCMA increases reliability under transmission latency constraints. In the time-invariant frequency selective fading channel, this chapter derives the error probability of FD-SCMA under given transmission latency and verifies the effectiveness of FD-SCMA by theoretical calculation and Monte Carlo simulation. The proposed FD-SCMA scheme can support transmission in UL and DL simultaneously to significantly minimize the access latency. Also, the proposed scheme can achieve a good compromise in transmission latency and reliability.

4.1 Motivation

To support URLLC by 5G New Radio (NR), reliability and latency can be guaranteed from the following aspects.

First, the recently studied diversity-based approach significantly improves the reliability of UL transmission. The key factor in reliability improvement is the diversity in the space domain, frequency domain, time domain, and coding domain [75, 76]. Considering the deep fading that could be encountered in time-variant and frequency-selective channels, diversity is essential in highly reliable transmission. In

addition, MUD and accurate channel estimation also help to improve reliability in UL transmission.

Second, redefining the frame structure can significantly reduce latency. A mini-slot is defined as a short slot in 5G NR containing 1 to 13 OFDM symbols [1]. When the system operates at above 6 GHz, the subcarrier spacing of OFDM symbols can be extended from 15 kHz to 60 kHz, 120 kHz, or even 240 kHz [1]. Through adjusting the frame structure, the duration of a slot can be reduced from 500 microseconds (μs) to about 5 μs. Meanwhile, FD can simultaneously support UL and DL transmission [16].

Third, re-using radio resources can enhance reliability with lower latency. NOMA can serve multiple users in the same radio resource, with higher SE than OMA [4]. In addition, NOMA can reduce access latency by providing services to multiple users simultaneously with given radio resources, so it can reduce the time consumed on waiting for some users. SCMA spreads a user's signal across multiple carriers and then improves reliability through frequency diversity. At present, OFDMA is the basic waveform of 5G and can be adequately combined with SCMA. Further, the combination of OFDMA and SCMA can benefit from the frequency diversity remarkably in the extensively existed frequency-selective channel. With the improvement in standardization and applications, the prospect of SCMA deployment in 5G is bright.

One of the key techniques to maintain reliability in SCMA is the codebook design. Each codebook can be generated from the mother constellation with a specific phase rotation, and users are distinguished by different carriers occupied [9, 30]. Grant-free SCMA can be implemented to reduce the latency that is consumed in sending scheduling requests [77]. Codebook and MUD, which have been extensively studied, can be applied as solid bases in this chapter.

Recently, some scholars have proposed and studied the impact of the combination of FD and NOMA. The utilization of an optimal power allocation scheme can effectively reduce the outage probability of weak users in DL cooperative FD NOMA relay systems [78]. In the meantime, the outage probability of cooperative

FD NOMA relay systems may be increased by incomplete SIS [79]. In practical applications, FD can generally suppress the self-interference level to about -130 dB through SIS [80]. Under practical and imperfect SIS, it is necessary to evaluate the decline in reliability. Most of the previous studies have assumed that FD is only applied with one Rx antenna and mainly in single-carrier transmission. Subsequently, FD can only reduce the BLER to the level of nearly 10^{-2}, which still cannot achieve the required 10^{-5} BLER. Besides, previous studies have not considered the compromise between low latency and high reliability. In [81], the authors comprehensively investigate the basic concept, suppression techniques, MAC protocols, and performance of in-band full-duplex systems, and then they indicate the research trends and potential applications. Paper [82] studies the performance of spectrum sensing schemes that use either a single-channel or multi-channel energy detector for in-band full-duplex operation. Zhou *et al.* [83] take advantage of pilot-data superposition in a FD-MIMO system to reduce the pilot overhead of the uplink and improve the network performance. Biswas et al. [84] study an FD MIMO cellular system collocated with a MIMO radar system, and propose spectrum sharing algorithms to improve the QoS for cellular users.

There are still some challenges during the application of FD and NOMA schemes mentioned above. First, there is a trade-off between reliability and latency due to power and bandwidth limitations in IoT deployment [85]. Then, the direct application of multiple Tx antennas in FD may cause complicated inter-user interference and self-interference that are difficult to suppress effectively, especially in multi-cell scenarios. Finally, accurate CSI is critical for IC and MUD algorithms. However, it is challenging to achieve accurate instantaneous CSI in an ultra-low latency in current wireless networks. Through the extensive study, SCMA and FD schemes have been proven to be effective in reducing latency. Nevertheless, when there is only one Rx antenna being applied, the required 10^{-5} BLER for reliable communications cannot be directly guaranteed by SCMA or FD independently. It is possible to considerably reduce the BLER by combining the advantages of SCMA and FD and exploiting the diversity gains fully, thereby supporting reliable delivery without

retransmission in low-latency communications.

This chapter proposes FD-SCMA that the FD gNB can support multiple SCMA users in both UL and DL. Further, multiple Rx antennas are installed to exploit spatial diversity. It is assumed that accurate CSI is available at the receiver, and the FD gNB simultaneously serves UL and DL users with dedicated assigned SCMA codewords. In addition, due to the improved SE brought by FD, UL users can leverage a lower rate coding scheme that can gain diversity from the coding domain. Since the SCMA codewords, which utilized used by the UL users and the gNB, are mutually pseudo-orthogonal, the original MUD can distinguish the signals at DL users. Therefore, the intrinsic UL-to-DL interference in FD can be mitigated. The FD-SCMA scheme that gains diversity from multiple domains can significantly improve reliability in low-latency communications.

There are some studies involving the combination of FD and NOMA schemes in BSs, but they do not fully exploit the diversity gains of multiple carriers and Rx antennas. An essential feature of FD-SCMA is that it can effectively utilize the diversity gain to significantly improve reliability with low latency in UL communications. This chapter also focuses on ensuring reliability and latency constraints on the combination of FD and NOMA, which is not mentioned in the existing research. The existing research combining FD and NOMA is mainly based on the IBL hypothesis, without appropriate direct application in short-packet transmission. It is more appropriate to apply the newly proposed FBL information theory in short-packet services to analyze reliability.

4.2 FD-SCMA System Model

The FD-SCMA system, as illustrated in Fig. 4.1, consists of one FD gNB, K active DL users, and L active UL users. FD-SCMA performs short-packet transmission through N available carriers. Each carrier consists of 12 subcarriers with f_{SC} subcarrier spacing. D information bits are assumed to be included in a packet. Considering the ultra-high reliability, this chapter assumes N_{RX} Rx antennas are equipped at DL users as well as the gNB to exploit space diversity. Only one Tx

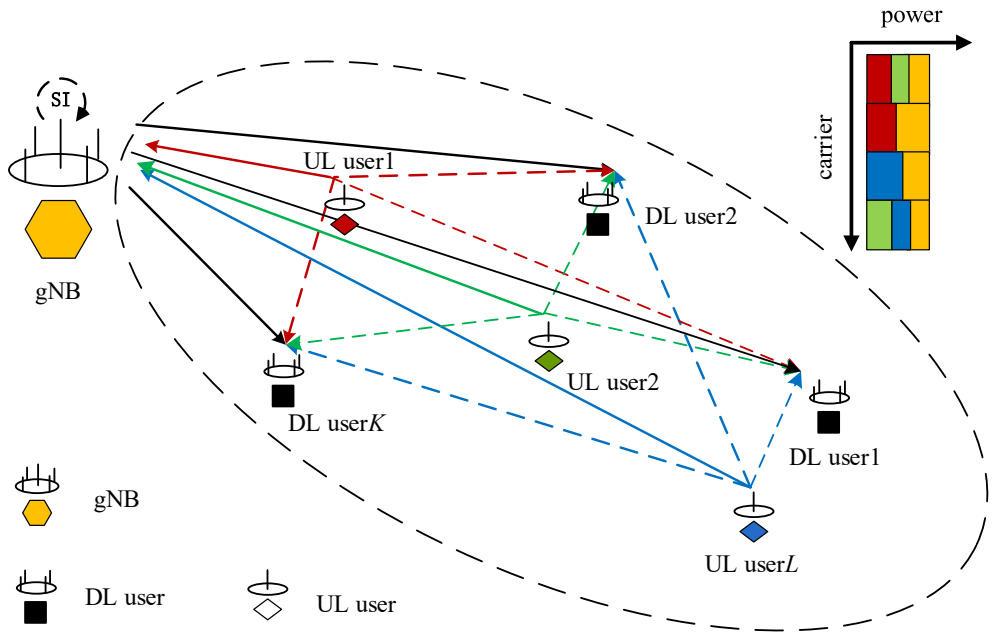

Figure 4.1 : An illustration of the FD-SCMA system with one FD gNB, K DL users, and L UL users.

antenna is installed at UL users and the gNB, respectively. The latency typically includes the access, transmission, demodulation, decoding, and retransmission latency [86]. This chapter studies the transmission latency t_{DE} that accounts for the most substantial proportion of the total 0.5 ms latency.

By appropriately assigning an SCMA codebook to each UL/DL user, the transmission signals can be generated by mapping the information data according to the codebook. The signals of different users can be transmitted simultaneously and be distinguished properly at the receiver through MUD, for example, joint ML detection.

The transmission of the FD-SCMA system follows the slot mode. A resource block (RB) lasts one time slot in the time domain and occupies one carrier in the

frequency domain. One slot includes N RBs spreading in the frequency domain, and each RB lasts for T_{SL}, the time duration of one slot. To transmit one short packet in t_{DE}, the number of occupied slots is $N_{\mathrm{SL}} = t_{\mathrm{DE}}/T_{\mathrm{SL}}$. The basic element of bearer signals in NR is RE, and M_{RB} REs are included in one RB. The coding blocklength M can be given by $M = M_{\mathrm{RB}}N_{\mathrm{SL}} = M_{\mathrm{RB}}t_{\mathrm{DE}}/T_{\mathrm{SL}}$.

One codeword is mapped onto N_{d} different carriers. The following two matrices may be applied to indicate the carrier allocation for the signals in UL and DL,

$$\mathbf{G}^{\mathrm{DL}} = [\mathbf{g}_1^{\mathrm{DL}}, \mathbf{g}_2^{\mathrm{DL}}, \cdots, \mathbf{g}_K^{\mathrm{DL}}], \tag{4.1}$$

and

$$\mathbf{G}^{\mathrm{UL}} = [\mathbf{g}_1^{\mathrm{UL}}, \mathbf{g}_2^{\mathrm{UL}}, \cdots, \mathbf{g}_L^{\mathrm{UL}}], \tag{4.2}$$

where $\mathbf{g}_k^{\mathrm{DL}} = [g_{1,k}^{\mathrm{DL}}, g_{2,k}^{\mathrm{DL}}, \cdots, g_{N,k}^{\mathrm{DL}}]^T$ indicates the SCMA mapping pattern of the signal transmitted from the gNB to DL user k, and $\mathbf{g}_l^{\mathrm{UL}} = [g_{1,l}^{\mathrm{UL}}, g_{2,l}^{\mathrm{UL}}, \cdots, g_{N,l}^{\mathrm{UL}}]^T$ indicates the SCMA mapping pattern of the signal transmitted from UL user l. $g_{n,k}^{\mathrm{DL}} \in \{0, -1\}$ and $g_{n,l}^{\mathrm{UL}} \in \{0, 1\}$ $(1 \leq k \leq K,\ 1 \leq l \leq L,\ 1 \leq n \leq N)$ denote whether carrier n will be assigned to DL user k or UL user l, respectively. $g_{n,k}^{\mathrm{DL}} = -1$, if and only if the signal of DL user k is mapped onto carrier n. $g_{n,l}^{\mathrm{UL}} = 1$, if and only if the signal of UL user l is mapped onto carrier n. The number of non-zero elements in $\mathbf{g}_k^{\mathrm{DL}}$ or $\mathbf{g}_l^{\mathrm{UL}}$ is equal to N_{d}. According to the principle of SCMA design, each column of $\mathbf{G}_{\mathrm{SCMA}}^{\mathrm{FD}} = [-\mathbf{G}^{\mathrm{DL}}, \mathbf{G}^{\mathrm{UL}}]$ has a different mapping pattern so that operating two users on the exactly same set of carriers can be avoided.

In order to simply describe the small-scale fading (SSF) and large-scale fading (LSF) coefficients of UL and DL channels, the FD gNB can be represented as DL user 0 and UL user 0. Meanwhile, a uniform variable $h_{r,n,l,k} \in \mathbb{C}$ $(1 \leq r \leq N_{\mathrm{RX}}, 1 \leq n \leq N)$ can represent the SSF coefficient on carrier n of the Rx antenna r of the channel from the gNB to DL user k $(l = 0, 1 \leq k \leq K)$, or the channel from UL user l to the gNB $(1 \leq l \leq L, k = 0)$, or the channel from UL user l to DL user k $(1 \leq l \leq L, 1 \leq k \leq K)$. In this way, a uniform matrix $\mathbf{H}_{l,k} = [\mathbf{h}_{1,l,k}, \mathbf{h}_{2,l,k}, \cdots, \mathbf{h}_{N,l,k}]$ can represent the SSF coefficients of the channel from UL user l to DL user k, where

$\mathbf{h}_{n,l,k} = [h_{1,n,l,k}, h_{2,n,l,k}, \cdots, h_{N_{\mathrm{RX}},n,l,k}]^T$ $(1 \leq n \leq N)$ is the SSF coefficients of the channels from the Tx antenna of UL user l to the Rx antennas of DL user k on carrier n. Particularly, $\mathbf{H}_{0,k}$ denotes the SSF coefficients of DL user k; and $\mathbf{H}_{l,0}$ denotes the SSF coefficients of UL user l.

The variable $\varpi_{l,k} \in \mathbb{R}^+$ can represent the LSF coefficient of the channel from the gNB to DL user k $(l = 0, 1 \leq k \leq K)$, or the channel from UL user l to DL user k $(1 \leq l \leq L, 1 \leq k \leq K)$, or the channel from UL user l to the gNB $(1 \leq l \leq L, k = 0)$. Furthermore, the LSF coefficients could be expanded and expressed as $\varpi_{l,k} = d_{l,k}^{-\alpha}$ $(0 < d_{l,k} \leq R_{\mathrm{D}})$, where α denotes the path loss exponent, R_{D} is the cell radius, and $d_{l,k}$ represents the distance between the gNB and DL user k $(l = 0, 1 \leq k \leq K)$, or the distance between UL user l and DL user k $(1 \leq l \leq L, 1 \leq k \leq K)$, or the distance between UL user l and the gNB $(1 \leq l \leq L, k = 0)$.

In one slot, the FD gNB and UL users will simultaneously transmit signals in the occupied carriers, so that DL users undergo two types of interference: MUI caused by superposed DL transmission, and UL-to-DL co-channel interference (CCI) in FD. Hence, the received signal of DL user k on carrier n of Rx antenna r can be expressed as

$$\begin{aligned} y_{r,n,k}^{\mathrm{DL}} = & \underbrace{\sqrt{p_{n,k}^{\mathrm{DL}} \varpi_{0,k}} h_{r,n,0,k} |g_{n,k}^{\mathrm{DL}}| x_{n,k}^{\mathrm{DL}}}_{\text{desired signal of the DL user } k} + \underbrace{\sum_{m=1, m \neq k}^{K} \sqrt{p_{n,m}^{\mathrm{DL}} \varpi_{0,k}} h_{r,n,0,k} |g_{n,m}^{\mathrm{DL}}| x_{n,m}^{\mathrm{DL}}}_{\text{MUI in DL}} \\ & + \underbrace{\sum_{l=1}^{L} \sqrt{q_{n,l}^{\mathrm{UL}} \varpi_{l,k}} h_{r,n,l,k} |g_{n,l}^{\mathrm{UL}}| x_{n,l}^{\mathrm{UL}}}_{\text{UL-to-DL CCI}} + \underbrace{z_{r,n,k}^{\mathrm{DL}}}_{\text{noise}}, \end{aligned} \tag{4.3}$$

where $x_{n,k}^{\mathrm{DL}} \in \mathbb{C}$ and $x_{n,l}^{\mathrm{UL}} \in \mathbb{C}$ $(1 \leq k \leq K,\ 1 \leq l \leq L,\ 1 \leq n \leq N)$ indicate the signals transmitted on carrier n from the gNB to DL user k and from UL user l to the gNB, respectively; $\mathbf{E}\left(|x_{n,k}^{\mathrm{DL}}|^2\right) = \mathbf{E}\left(|x_{n,l}^{\mathrm{UL}}|^2\right) = 1$; $z_{r,n,k}^{\mathrm{DL}} \sim \mathcal{CN}(0, \sigma_k^2)$ is the complex AWGN at DL user k on carrier n of Rx antenna r; $p_{n,k}^{\mathrm{DL}}$ indicates the Tx power of the gNB for DL user k on carrier n, and $q_{n,l}^{\mathrm{UL}}$ indicates the Tx power of UL user l on carrier n.

At the same time, the transmitted signals of UL users will suffer from two types of interference: the MUI due to UL superposed transmission, and the residual self-interference caused by the FD gNB. Thus, the received signal at the gNB on carrier n of Rx antenna r can be expressed as

$$\begin{aligned} y_{r,n}^{\mathrm{UL}} = & \underbrace{\sqrt{q_{n,l}^{\mathrm{UL}} \varpi_{l,0}} h_{r,n,l,0} |g_{n,l}^{\mathrm{UL}}| x_{n,l}^{\mathrm{UL}}}_{\text{desired signal of the UL user } l} + \underbrace{\sum_{j=1, j \neq l}^{L} \sqrt{q_{n,j}^{\mathrm{UL}} \varpi_{j,0}} h_{r,n,j,0} |g_{n,j}^{\mathrm{UL}}| x_{n,j}^{\mathrm{UL}}}_{\text{MUI in UL}} \\ & + \underbrace{\sqrt{\kappa} h_{0,0} \sum_{k=1}^{K} \sqrt{p_{n,k}^{\mathrm{DL}}} |g_{n,k}^{\mathrm{DL}}| x_{n,k}^{\mathrm{DL}}}_{\text{residual self-interference}} + \underbrace{z_{r,n}^{\mathrm{UL}}}_{\text{noise}}, \end{aligned} \tag{4.4}$$

where $\kappa \in [0, 1)$ is the SIS factor of the gNB, and can represent the ability of the gNB to eliminate SIS caused by FD; $h_{0,0} \in \mathbb{C}$ represents the SSF coefficient of the channel from the Tx antenna of the FD gNB to the Rx antennas of the FD gNB; $z_{r,n}^{\mathrm{UL}} \sim \mathcal{CN}(0, \sigma_0^2)$ is the complex AWGN of the Rx antenna r of the gNB on carrier n.

Here, it is assumed that each user maps the signals to the occupied carriers with equal power. Therefore, the Tx power for DL user k on carrier n is given by

$$p_{n,k}^{\mathrm{DL}} = |g_{n,k}^{\mathrm{DL}}| p_k^{\mathrm{DL}} / N_{\mathrm{d}}, \tag{4.5}$$

where p_k^{DL} indicates the Tx power allocated by the gNB for DL user k; $p_k^{\mathrm{DL}} = \sum_{n=1}^{N} p_{n,k}^{\mathrm{DL}}$ and $\sum_{k=1}^{K} p_k^{\mathrm{DL}} \leq P_{\mathrm{MAX}}^{\mathrm{DL}}$, where $P_{\mathrm{MAX}}^{\mathrm{DL}}$ indicates the maximum Tx power of the gNB. So as to uniformly describe the self-interference of the gNB on different carriers, it can be assumed that the maximum Tx power of the gNB on each carrier is $P_{\mathrm{MAX}}^{\mathrm{DL}}/N$, and $\sum_{k=1}^{K} p_{n,k}^{\mathrm{DL}} \leq P_{\mathrm{MAX}}^{\mathrm{DL}}/N$ $(1 \leq n \leq N)$.

The Tx power of UL user l on carrier n is represented by the following relationship

$$q_{n,l}^{\mathrm{UL}} = |g_{n,l}^{\mathrm{UL}}| q_l^{\mathrm{UL}} / N_{\mathrm{d}}, \tag{4.6}$$

where q_l^{UL} denotes the Tx power of UL user l; $q_l^{\mathrm{UL}} = \sum_{n=1}^{N} q_{n,l}^{\mathrm{UL}} \leq P_{\mathrm{MAX}}^{\mathrm{UL}}$, where

$P_{\mathrm{MAX}}^{\mathrm{UL}}$ indicates the maximum Tx power of UL users.

It is worth mentioning that the codewords causing CCI and MUI are homogeneously generated for the DL user's receiver. Consequently, the original MUD module can be applied directly to process all UL and DL interference. After MUD is adopted, it can detect and suppress strong UL-to-DL interference through the SIC at the receiver of a DL user when the interference power is stronger than the noise power. Moreover, when the interference power and the noise power are comparable, the weaker UL-to-DL interference that is hard to be detected can be allocated to non-overlapped carriers. Therefore, when the number of carriers is large enough, DL users can significantly mitigate the residual UL-to-DL CCI in FD-SCMA compared to other typical FD schemes.

4.3 Effective SNR and Error Probability in Low-Latency Communications

This section derives the effective SNRs of the gNB and DL users. In addition, an expression for the error probability is derived in the FBL.

4.3.1 Effective SNR

Since each codeword of an SCMA user is distributed on N_{d} of total N carriers, the receiver can detect the codeword by combining the signals on different carriers. Further, the maximal achievable rate can be achieved by $\tilde{R} = C(\tilde{\gamma}) = \sum_{n=1}^{N} C(\tilde{\gamma}_n)$ [87], where $\tilde{\gamma}$ is the effective SNR of the codeword. Besides, $\tilde{\gamma}_n$ indicates the effective SNR on carrier n under the condition where carrier n is not occupied by the codeword, $\tilde{\gamma}_n = 0$, and $C(\tilde{\gamma}_n) = 0$. In low-latency short-packet transmission, the duration is generally less than the coherence time of the channel. It can be reasonable to assume that the channel coefficient of each carrier remains unchanged. Thus, we

can have [87, 88]

$$\begin{aligned}\tilde{\gamma} &= C^{-1}\left(\sum_{n=1}^{N}\left(\frac{1}{M}\sum_{m=1}^{M}\log_2(1+\beta\breve{\gamma}_{m,n})\right)\right)\\ &= C^{-1}\left(\sum_{n=1}^{N}\log_2(1+\beta\breve{\gamma}_n)\right)\\ &= \prod_{n=1}^{N}(1+\beta\breve{\gamma}_n)-1,\end{aligned} \tag{4.7}$$

where $\breve{\gamma}_n$ is the effective SNR on carrier n when the interference is completely eliminated, known as perfect interference cancelation (PIC). β indicates the ability of different schemes to eliminate MUI, and it is affected by the number of active users. In the physical layer abstraction of SCMA, the recommended value β is from 0.82 to 0.97 [88]. Greater β can be achieved when implementing more efficient MUDs, such as joint ML detection, and serving fewer users. $\breve{\gamma}_{m,n}$ is the effective SNR on carrier n of the mth codeword when interference is completely cancelled, and $\breve{\gamma}_{m,n} = \breve{\gamma}_n$ $(1 \le m \le M)$ when the channel coefficient of carrier n remains static in the very short duration of short-packet transmission.

With MRC detection, the effective SNR of DL user k can be given by [10]

$$\begin{aligned}\tilde{\gamma}_k^{\mathrm{DL}} &= \prod_{n=1}^{N}(1+\beta\breve{\gamma}_{n,k}^{\mathrm{DL}})-1\\ &= \prod_{n=1}^{N}\left(1+\beta|g_{n,k}^{\mathrm{DL}}|\parallel\mathbf{h}_{n,0,k}\parallel^2\zeta_k^{\mathrm{DL}}\right)-1.\end{aligned} \tag{4.8}$$

In the same way, when the interference is completely eliminated in UL FD-SCMA, the effective SNR of UL user l on carrier n under PIC can be given by:

$$\breve{\gamma}_{n,l}^{\mathrm{UL}} = \sum_{r=1}^{N_{\mathrm{RX}}}\frac{|g_{n,l}^{\mathrm{UL}}||h_{r,n,l,0}|^2\zeta_l^{\mathrm{UL}}}{|h_{0,0}|^2\zeta_n^{\mathrm{SI}}+1} = \frac{|g_{n,l}^{\mathrm{UL}}|\parallel\mathbf{h}_{n,l,0}\parallel^2\zeta_l^{\mathrm{UL}}}{|h_{0,0}|^2\zeta_n^{\mathrm{SI}}+1}, \tag{4.9}$$

where $\zeta_l^{\mathrm{UL}} = q_l^{\mathrm{UL}}d_{l,0}^{-\alpha}/N_{\mathrm{d}}\sigma_0^2$ and $\zeta_n^{\mathrm{SI}} = \kappa\sum_{k=1}^{K}p_{n,k}^{\mathrm{DL}}/\sigma_0^2 \approx \kappa P_{\mathrm{MAX}}^{\mathrm{DL}}/N\sigma_0^2$.

For simplicity, the self-interference on the same carrier $\sum_{k=1}^{K}p_{n,k}^{\mathrm{DL}}$ can be approx-

imated by the maximum Tx power of the gNB on carrier n, which is $P_{\mathrm{MAX}}^{\mathrm{DL}}/N$. In fact, if the actual DL Tx power P^{DL} is lower than $P_{\mathrm{MAX}}^{\mathrm{DL}}$, the reliability of UL users can be improved. It can be verified in the following simulation. Therefore, through the previous analysis, the effective SNR of UL user l can be obtained as follows,

$$\tilde{\gamma}_l^{\mathrm{UL}} = \prod_{n=1}^{N} \left(1 + \frac{\beta |g_{n,l}^{\mathrm{UL}}| \parallel \mathbf{h}_{n,l,0} \parallel^2 \zeta_l^{\mathrm{UL}}}{|h_{0,0}|^2 \zeta_n^{\mathrm{SI}} + 1} \right) - 1. \tag{4.10}$$

4.3.2 Error Probability in FBL

The maximal achievable rate $R = D/M$ for a given blocklength M in a deterministic channel can be closely approximated by [35]

$$R = C(\gamma) - \sqrt{\frac{V(\gamma)}{M}} Q^{-1}(\varepsilon), \tag{4.11}$$

where $\varepsilon < 0.5$ is the error probability, γ is the effective SNR, $Q(w) = \int_w^\infty \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt$, and $V(\gamma) = \frac{\gamma(\gamma+2)}{(1+\gamma)^2} \log_2^2 e$ defines the probability that a Gaussian random variable of zero mean unit variance is greater than w.

The error probability in FBL is determined by the effective SNR and the blocklength. If D information bits are transmitted in the AWGN channel with blocklength M, an upper bound (UB) for the error probability can be expressed as [89, eq. 14]

$$\varepsilon = \begin{cases} 1 & , \quad \text{if } C(\gamma) < D/M; \\ Q\left(\sqrt{\frac{M}{V(\gamma)}} \left(C(\gamma) - D/M \right) \right) & , \quad \text{otherwise.} \end{cases} \tag{4.12}$$

In DL, when a channel realization $\mathbf{H}_{0,k}$ is determined, the error probability of DL user k can be given by

$$\Psi_k^{\mathrm{DL}} = \begin{cases} 1 & , \quad \text{if } \Lambda_k^{\mathrm{DL}} < 0; \\ Q\left(\sqrt{\frac{M_{\mathrm{RB}} t_{\mathrm{DE}} / T_{\mathrm{SL}}}{V(\tilde{\gamma}_k^{\mathrm{DL}})}} \Lambda_k^{\mathrm{DL}} \right) & , \quad \text{otherwise,} \end{cases} \tag{4.13}$$

where $\Lambda_k^{\mathrm{DL}} = C(\tilde{\gamma}_k^{\mathrm{DL}}) - DT_{\mathrm{SL}}/M_{\mathrm{RB}} t_{\mathrm{DE}} = \sum_{n=1}^{N} \log_2 \left(1 + \frac{\beta |g_{n,k}^{\mathrm{DL}}| \|\mathbf{h}_{n,0,k}\|^2 p_k^{\mathrm{DL}} d_{0,k}^{-\alpha}}{N_{\mathrm{d}} \sigma_k^2} \right) - \frac{DT_{\mathrm{SL}}}{M_{\mathrm{RB}} t_{\mathrm{DE}}}$.

Also in UL, given the realization of $\mathbf{H}_{l,0}$ and $h_{0,0}$, the error probability of UL user l can be given by

$$\Psi_l^{\mathrm{UL}} = \begin{cases} 1 & , \quad \text{if } \Lambda_l^{\mathrm{UL}} < 0; \\ Q\left(\sqrt{\frac{M_{\mathrm{RB}} t_{\mathrm{DE}}/T_{\mathrm{SL}}}{V(\tilde{\gamma}_l^{\mathrm{UL}})}} \Lambda_l^{\mathrm{UL}}\right) & , \quad \text{otherwise,} \end{cases} \tag{4.14}$$

where $\Lambda_l^{\mathrm{UL}} = C(\tilde{\gamma}_l^{\mathrm{UL}}) - DT_{\mathrm{SL}}/M_{\mathrm{RB}} t_{\mathrm{DE}} = \sum_{n=1}^{N} \log_2 \left(1 + \frac{\beta |g_{n,l}^{\mathrm{UL}}| \| \mathbf{h}_{n,l,0} \|^2 q_l^{\mathrm{UL}} d_{l,0}^{-\alpha}}{|h_{0,0}|^2 \kappa P_{\mathrm{MAX}}^{\mathrm{DL}} N_{\mathrm{d}}/N + N_{\mathrm{d}} \sigma_0^2}\right) - \frac{DT_{\mathrm{SL}}}{M_{\mathrm{RB}} t_{\mathrm{DE}}}$.

4.4 Performance in Time-invariant Flat-Fading Channels

With the utilization of the error probability derived in FBL, this section studies the reliability performance of FD-SCMA in time-invariant flat-fading channels with latency constraints.

In the time-invariant flat-fading channel, the fading coefficients of the channels between the users and the gNB stay constant through all occupied carriers during the whole short-packet transmission (i.e., $h_{r,n,l,k} = h_{r,0,l,k}, \forall n \in \{1, 2, ..., N\}$).

Without loss of generality, it can be assumed that DL user K has the weakest channel gain in DL, and UL user L has the weakest channel gain in UL. It is apparent that the bottleneck of reliability improvement is the user with the weakest channel gain in DL or UL. Therefore, this session mainly analyzes the performance of DL user K and UL user L. From (4.8), the effective SNR of DL user K can be derived as

$$\tilde{\gamma}_K^{\mathrm{DL}} = \prod_{n=1}^{N} \left(1 + \beta |g_{n,K}^{\mathrm{DL}}| \parallel \mathbf{h}_{n,0,K} \parallel^2 \zeta_K^{\mathrm{DL}}\right) - 1. \tag{4.15}$$

According to (4.10), the effective SNR of UL user L can be expressed as

$$\tilde{\gamma}_L^{\mathrm{UL}} = \prod_{n=1}^{N} \left(1 + \frac{\beta |g_{n,L}^{\mathrm{UL}}| \parallel \mathbf{h}_{n,L,0} \parallel^2 \zeta_L^{\mathrm{UL}}}{|h_{0,0}|^2 \zeta_n^{\mathrm{SI}} + 1}\right) - 1. \tag{4.16}$$

In a time-invariant flat-fading channel, $\mathbf{h}_{0,0,k} = [h_{1,0,0,k}, ..., h_{N_{\mathrm{RX}},0,0,k}]^T$ and $\mathbf{h}_{0,l,0} = [h_{1,0,l,0}, ..., h_{N_{\mathrm{RX}},0,l,0}]^T$ can represent channel fading coefficients on different carriers

in DL and UL, respectively. $\mathbf{h}_{n,0,k} = \mathbf{h}_{0,0,k}$ and $\mathbf{h}_{n,l,0} = \mathbf{h}_{0,l,0}$ $(1 \leq n \leq N)$ can always hold. At this time, the effective SNR of DL user K in the time-invariant flat-fading channel can be rewritten as

$$\gamma_K^{\text{DL\_STA}} = \left(1 + \frac{\beta \parallel \mathbf{h}_{0,0,K} \parallel^2 p_K^{\text{DL}} d_{0,K}^{-\alpha}}{N_{\text{d}} \sigma_K^2}\right)^{N_{\text{d}}} - 1. \tag{4.17}$$

In order to ensure reliability under a given transmission latency constraint t_{DE}, it should be guaranteed that $C(\gamma_K^{\text{DL\_STA}}) \geq \hat{R}$, where $\hat{R} = DT_{\text{SL}}/M_{\text{RB}}t_{\text{DE}}$ represents the achievable rate. The way to maintain sufficient effective SNR is to deploy DL users within a given range $d_0^{\text{DL}^*}$, which depends on the transmission latency constraint t_{DE} and Tx power p_K^{DL}. Thus, it can be inferred that the deployment range depends on

$$d_0^{\text{DL}^*} = \left(\frac{\beta \parallel \mathbf{h}_{0,0,K} \parallel^2 p_K^{\text{DL}}}{(2^{DT_{\text{SL}}/M_{\text{RB}}t_{\text{DE}}N_{\text{d}}} - 1)N_{\text{d}}\sigma_K^2}\right)^{1/\alpha}. \tag{4.18}$$

It is known from (4.13) that the error probability of DL user K can be given by

$$\Psi_K^{\text{DL\_STA}} = Q\left(\sqrt{\frac{M_{\text{RB}}t_{\text{DE}}/T_{\text{SL}}}{V(\gamma_K^{\text{DL\_STA}})}}\left(C\big(\gamma_K^{\text{DL\_STA}}\big) - \hat{R}\right)\right). \tag{4.19}$$

Similarly, the effective SNR of UL user L in the time-invariant flat-fading channel can be further extended to

$$\gamma_L^{\text{UL\_STA}} = \left(1 + \frac{\beta \parallel \mathbf{h}_{0,L,0} \parallel^2 q_L^{\text{UL}} d_{L,0}^{-\alpha}}{\kappa|h_{0,0}|^2 P_{\text{MAX}}^{\text{DL}} N_{\text{d}}/N + N_{\text{d}}\sigma_0^2}\right)^{N_{\text{d}}} - 1. \tag{4.20}$$

In order to maintain an effective SNR to satisfy the requirement of high reliability, the deployment range of UL users can be extended depending on

$$d_0^{\text{UL}^*} = \left(\frac{\beta \parallel \mathbf{h}_{0,L,0} \parallel^2 q_L^{\text{UL}}}{(2^{DT_{\text{SL}}/M_{\text{RB}}t_{\text{DE}}N_{\text{d}}} - 1)(\kappa\eta^{\text{SI}} + 1)N_{\text{d}}\sigma_0^2}\right)^{1/\alpha}, \tag{4.21}$$

where $\eta^{\text{SI}} = |h_{0,0}|^2 P_{\text{MAX}}^{\text{DL}}/N\sigma_0^2$. When $d_{L,0} \leq d_0^{\text{UL}^*}$, we can always have $C(\gamma_L^{\text{UL\_STA}}) \geq \hat{R}$.

From (4.14), the error probability of UL user L can be expressed as

$$\Psi_L^{\text{UL\_STA}} = Q\left(\sqrt{\frac{M_{\text{RB}} t_{\text{DE}}/T_{\text{SL}}}{V(\gamma_L^{\text{UL\_STA}})}}\left(C\big(\gamma_L^{\text{UL\_STA}}\big) - \hat{R}\right)\right). \tag{4.22}$$

Similarly, the error probabilities of HD-SCMA, FD-OMA, and HD-OMA schemes under FBL can be derived, as demonstrated later in this chapter. By comparing the error probabilities of FD-SCMA with other schemes, the following theorem is obtained.

Theorem 3. *In the time-invariant flat-fading channel, DL URLLC has a transmission latency constraint t_{DE}. When DL user K is within the operating range of the system, $d_{0,K} \leq d_0^{DL^*}$, FD-SCMA has a lower error probability than HD-SCMA, FD-OMA, and HD-OMA.*

Proof. In DL, when $d_{0,K} \leq d_0^{\text{DL}^*}$, we can have $\Lambda_K^{\text{DL}} \geq 0$, and then the error probability of FD-SCMA only depends on $\Theta(\gamma) = Q\left(\sqrt{\frac{M}{V(\gamma)}}(C(\gamma) - D/M)\right)$. Compared with FD schemes, the HD schemes utilize half of the time slots to support UL and DL, respectively. It can be obtained that $M^{\text{HD}} = M_{\text{RB}} t_{\text{DE}}/2T_{\text{SL}} = M/2$, where M^{HD} is the blocklength of reliable HD transmission with limited latency, and $\hat{R}^{\text{HD}} = D/M^{\text{HD}} = 2\hat{R}$, where $\hat{R}^{\text{HD}}$ is the achievable rate in HD schemes.

In order to prove Theorem 3 in a concise and clear manner, the proof is divided into the following three steps.

1) Similar to the minimum effective SNR of DL users in FD-SCMA, the minimum effective SNR of DL users in HD-SCMA can be given by $\gamma_K^{\text{DL\_HD\_SCMA}} = \left(1 + \beta \parallel \mathbf{h}_{0,0,K} \parallel^2 p_K^{\text{DL}} d_{0,K}^{-\alpha}/N_{\text{d}}\sigma_K^2\right)^{N_{\text{d}}} - 1$. If $C(\gamma_K^{\text{DL\_HD\_SCMA}}) - \hat{R}^{\text{HD}} < 0$, the error probability of HD-SCMA is equal to 1, and thus the error probability of FD-SCMA can not be higher than the error probability of HD-SCMA. When $C(\gamma_K^{\text{DL\_HD\_SCMA}}) - \hat{R}^{\text{HD}} \geq 0$, the error probability of HD-SCMA depends on $\Theta(\gamma)$ and it is easily verified

that the following condition is true.

$$\begin{aligned}&\sqrt{\frac{M}{V(\gamma_K^{\text{DL\_STA}})}}\left(\log_2(1+\gamma_K^{\text{DL\_STA}})-\hat{R}\right)\\ \geq&\sqrt{\frac{M/2}{V(\gamma_K^{\text{DL\_HD\_SCMA}})}}\left(\log_2(1+\gamma_K^{\text{DL\_HD\_SCMA}})-2\hat{R}\right).\end{aligned}\tag{4.23}$$

Because $Q(w)$ is a monotonically decreasing function of w, we can have

$$\begin{aligned}&Q\left(\sqrt{\frac{M}{V(\gamma_K^{\text{DL\_STA}})}}\left(\log_2(1+\gamma_K^{\text{DL\_STA}})-\hat{R}\right)\right)\\ \leq&Q\left(\sqrt{\frac{M^{\text{HD}}}{V(\gamma_K^{\text{DL\_HD\_SCMA}})}}\left(\log_2(1+\gamma_K^{\text{DL\_HD\_SCMA}})-\hat{R}^{\text{HD}}\right)\right).\end{aligned}\tag{4.24}$$

Thus, the error probability of FD-SCMA is lower than that of HD-SCMA in DL when $d_{0,K} \leq d_0^{\text{DL}^*}$.

2) When $d_{0,K} \leq d_0^{\text{DL}^*}$, the error probability of FD-SCMA in DL is lower than that of FD-OMA. When $x > 0$ and $\beta \geq N/K$, we can obtain $\frac{d(1+\beta x/N_{\text{d}})^{N_{\text{d}}}}{dx} > \frac{d(1+x)^{N/K}}{dx}$, and

$$\begin{aligned}&\left(1+\beta \parallel \mathbf{h}_{0,0,K} \parallel^2 p_K^{\text{DL}} d_{0,K}^{-\alpha}/N_{\text{d}}\sigma_K^2\right)^{N_{\text{d}}}\\ \geq&\left(1+ \parallel \mathbf{h}_{0,0,K} \parallel^2 p_K^{\text{DL}} d_{0,K}^{-\alpha}/\sigma_K^2\right)^{N/K}.\end{aligned}\tag{4.25}$$

Moreover, in order to support K DL users on N carriers, the minimum effective SNR of DL users in FD-OMA $\gamma_K^{\text{DL\_FD\_OMA}}$ satisfies the following condition.

$$\gamma_K^{\text{DL\_FD\_OMA}} \leq \left(1+ \parallel \mathbf{h}_{0,0,K} \parallel^2 p_K^{\text{DL}} d_{0,K}^{-\alpha}/\sigma_K^2\right)^{N/K} - 1.\tag{4.26}$$

Therefore, $\gamma_K^{\text{DL\_STA}} \geq \gamma_K^{\text{DL\_FD\_OMA}}$ is always true.

Considering $\Theta(\gamma)$ is a monotonically decreasing function [39], the following for-

mula holds,

$$\begin{aligned} &Q\left(\sqrt{\frac{M}{V(\gamma_K^{\text{DL\_STA}})}}\left(\log_2(1+\gamma_K^{\text{DL\_STA}})-\hat{R}\right)\right) \\ \leq &Q\left(\sqrt{\frac{M}{V(\gamma_K^{\text{DL\_FD\_OMA}})}}\left(\log_2(1+\gamma_K^{\text{DL\_FD\_OMA}})-\hat{R}\right)\right). \end{aligned} \tag{4.27}$$

Therefore, when $d_{0,K} \leq d_0^{\text{DL}^*}$, the error probability of FD-SCMA is lower than that of FD-OMA in DL.

3) The minimum effective SNR of DL users in HD-OMA can be given by

$$\gamma_K^{\text{DL\_HD\_OMA}} = \left(1+ \parallel \mathbf{h}_{0,0,K} \parallel^2 p_K^{\text{DL}} d_{0,K}^{-\alpha}/\sigma_K^2\right)^{N/K} - 1. \tag{4.28}$$

Substituting (4.25) into (4.28), we can obtain $\gamma_K^{\text{DL\_STA}} \geq \gamma_K^{\text{DL\_HD\_OMA}}$.

Similar to (4.24), for $\forall d_{0,K} \leq d_0^{\text{DL}^*}$, we can acquire the following relationship,

$$\begin{aligned} &Q\left(\sqrt{\frac{M}{V(\gamma_K^{\text{DL\_STA}})}}\left(\log_2(1+\gamma_K^{\text{DL\_STA}})-\hat{R}\right)\right) \\ \leq &Q\left(\sqrt{\frac{M^{\text{HD}}}{V(\gamma_K^{\text{DL\_HD\_OMA}})}}\left(\log_2(1+\gamma_K^{\text{DL\_HD\_OMA}})-\hat{R}^{\text{HD}}\right)\right). \end{aligned} \tag{4.29}$$

Therefore, when $d_{0,K} \leq d_0^{\text{DL}^*}$, the error probability of FD-SCMA is lower than that of HD-OMA in DL.

In summary, Theorem 3 has been proved. □

Theorem 3 illustrates that the proposed FD-SCMA scheme is superior to other existing FD and NOMA schemes, including FD-OMA, HD-SCMA, and HD-OMA, in DL when the transmission latency constraint needs to be guaranteed.

Theorem 4. *In the time-invariant flat-fading channel, it is assumed that UL URLLC has a transmission latency constraint t_{DE}. When UL user L is within the deployment range of the system, $d_{L,0} \leq d_0^{UL^*}$, the error probability of FD-SCMA is lower than that of FD-OMA.*

Proof. Supporting L users on N carriers in FD-OMA is equivalent to occupying only N/L carries per UL user, so the minimum effective SNR of UL users can be expressed as

$$\gamma_L^{\text{UL\_FD\_OMA}} = \left(1 + \frac{\parallel \mathbf{h}_{0,L,0} \parallel^2 q_L^{\text{UL}} d_{L,0}^{-\alpha}}{\kappa |h_{0,0}|^2 P_{\text{MAX}}^{\text{DL}}/N + \sigma_0^2}\right)^{N/L} - 1. \tag{4.30}$$

When $d_{L,0} \leq d_0^{\text{UL}^*}$ and $\beta \geq N/L$, similar to (4.25), we can obtain

$$\begin{aligned} &\left(1 + \frac{\beta \parallel \mathbf{h}_{0,L,0} \parallel^2 q_L^{\text{UL}} d_{L,0}^{-\alpha}}{\kappa |h_{0,0}|^2 P_{\text{MAX}}^{\text{DL}} N_{\text{d}}/N + N_{\text{d}} \sigma_0^2}\right)^{N_{\text{d}}} \\ \geq &\left(1 + \frac{\parallel \mathbf{h}_{0,L,0} \parallel^2 q_L^{\text{UL}} d_{L,0}^{-\alpha}}{\kappa |h_{0,0}|^2 P_{\text{MAX}}^{\text{DL}}/N + \sigma_0^2}\right)^{N/L}. \end{aligned} \tag{4.31}$$

Thus, $\gamma_L^{\text{UL\_STA}} \geq \gamma_L^{\text{UL\_FD\_OMA}}$ is always true.

Considering the monotonically decreasing characteristic of $\Theta(\gamma)$, the error probability of FD-SCMA is lower than that of FD-OMA in UL for $\forall d_{L,0} \leq d_0^{\text{UL}^*}$.

Therefore, Theorem 4 has been proved. □

Theorem 4 states that the proposed FD-SCMA scheme is always superior to FD-OMA in terms of reliability in UL when the two schemes face the same self-interference level caused by imperfect SIS.

Theorem 5. *In the UL time-invariant flat-fading channel of URLLC with latency constraint t_{DE}, FD-SCMA has a lower error probability than HD-SCMA and HD-OMA under the following conditions: if $\kappa \leq (\sqrt{2}\varsigma - 1)/\eta^{SI}$ (the SIS of the gNB is sufficiently effective, so the FD scheme only faces small residual self-interference), where $\varsigma = 2^{(\sqrt{2}-1)\hat{R}/N_d}$, then UL user L is within the deployment range of the system $d_{L,0} \leq d_0^{UL^*}$; on the contrary, if $\kappa > (\sqrt{2}\varsigma - 1)/\eta^{SI}$, then UL user L needs to be within a smaller deployment range given by $d_1^{UL^*} \leq d_{L,0} \leq d_0^{UL^*}$, where*

$$d_1^{\text{UL}^*} = \left(\frac{\beta \parallel \mathbf{h}_{0,L,0} \parallel^2 q_L^{\text{UL}} \left(\kappa \eta^{\text{SI}} - (\sqrt{2}\varsigma - 1)\right)}{\sqrt{2} N_{\text{d}} \sigma_0^2 (\kappa \eta^{\text{SI}} + 1)(\varsigma - 1)}\right)^{1/\alpha}. \tag{4.32}$$

Proof. In UL, when $d_{L,0} \leq d_0^{\mathrm{UL}^*}$, the error probability of FD-SCMA only depends on $\Theta(\gamma)$. For the concise and clear presentation, the proof of Theorem 5 can be divided into the following two steps.

1) If $\kappa \leq (\sqrt{2}\varsigma - 1)/\eta^{\mathrm{SI}}$, since $\frac{1}{2^{\hat{R}/N_{\mathrm{d}}}} > \frac{1}{2^{\sqrt{2}\hat{R}/N_{\mathrm{d}}}}$ and $\frac{d((1+x/(\kappa\eta^{\mathrm{SI}}+1))/2^{\hat{R}/N_{\mathrm{d}}})}{dx} > \frac{d((1+x)^{\sqrt{1/2}}/2^{\sqrt{2}\hat{R}/N_{\mathrm{d}}})}{dx}$ $(x > 0)$, we can obtain the following relationship

$$\begin{aligned} &\frac{1+\beta \parallel \mathbf{h}_{0,L,0}\parallel^2 q_L^{\mathrm{UL}} d_{L,0}^{-\alpha}/N_{\mathrm{d}}\sigma_0^2(\kappa\eta^{\mathrm{SI}}+1)}{2^{\hat{R}/N_{\mathrm{d}}}} \\ \geq &\frac{(1+\beta \parallel \mathbf{h}_{0,L,0}\parallel^2 q_L^{\mathrm{UL}} d_{L,0}^{-\alpha}/N_{\mathrm{d}}\sigma_0^2)^{\sqrt{1/2}}}{2^{\sqrt{2}\hat{R}/N_{\mathrm{d}}}}. \end{aligned} \tag{4.33}$$

Thus, we can draw the following conclusions,

$$\begin{aligned} &\sqrt{\frac{M}{V(\gamma_L^{\mathrm{UL\_STA}})}}\left(\log_2(1+\gamma_L^{\mathrm{UL\_STA}}) - \hat{R}\right) \\ \geq &\sqrt{\frac{M/2}{V(\gamma_L^{\mathrm{UL\_STA}})}}\left(\log_2(1+\gamma_L^{\mathrm{UL\_HD\_SCMA}}) - 2\hat{R}\right), \end{aligned} \tag{4.34}$$

where $\gamma_L^{\mathrm{UL\_HD\_SCMA}} = \left(1+\frac{\beta\|\mathbf{h}_{0,L,0}\|^2 q_L^{\mathrm{UL}} d_{L,0}^{-\alpha}}{N_{\mathrm{d}}\sigma_0^2}\right)^{N_{\mathrm{d}}} - 1$ represents the minimum effective SNR of UL users in HD-SCMA.

Considering $Q(w)$ is a monotonically decreasing function and $V(\gamma)$ is a monotonic increasing function, we can acquire

$$\begin{aligned} &Q\left(\sqrt{\frac{M}{V(\gamma_L^{\mathrm{UL\_STA}})}}\left(\log_2\left(1+\gamma_L^{\mathrm{UL\_STA}}\right) - \hat{R}\right)\right) \\ \leq &Q\left(\sqrt{\frac{M^{\mathrm{HD}}}{V(\gamma_L^{\mathrm{UL\_STA}})}}\left(\log_2\left(1+\gamma_L^{\mathrm{UL\_HD\_SCMA}}\right) - \hat{R}^{\mathrm{HD}}\right)\right) \\ \leq &Q\left(\sqrt{\frac{M^{\mathrm{HD}}}{V(\gamma_L^{\mathrm{UL\_HD\_SCMA}})}}\left(\log_2\left(1+\gamma_L^{\mathrm{UL\_HD\_SCMA}}\right) - \hat{R}^{\mathrm{HD}}\right)\right). \end{aligned} \tag{4.35}$$

On the other hand, if $\kappa > (\sqrt{2}\varsigma - 1)/\eta^{\mathrm{SI}}$, when $d_1^{\mathrm{UL}^*} \leq d_{L,0} \leq d_0^{\mathrm{UL}^*}$, we can still

acquire

$$\begin{aligned}&\frac{1+\beta \parallel \mathbf{h}_{0,L,0}\parallel^2 q_L^{\mathrm{UL}} d_{L,0}^{-\alpha}/N_{\mathrm{d}}\sigma_0^2(\kappa\eta^{\mathrm{SI}}+1)}{2^{\hat{R}/N_{\mathrm{d}}}}\\ \geq&\frac{1+\sqrt{1/2}\beta \parallel \mathbf{h}_{0,L,0}\parallel^2 q_L^{\mathrm{UL}} d_{L,0}^{-\alpha}/N_{\mathrm{d}}\sigma_0^2}{2^{\sqrt{2}\hat{R}/N_{\mathrm{d}}}}\\ \geq&\frac{(1+\beta \parallel \mathbf{h}_{0,L,0}\parallel^2 q_L^{\mathrm{UL}} d_{L,0}^{-\alpha}/N_{\mathrm{d}}\sigma_0^2)^{\sqrt{1/2}}}{2^{\sqrt{2}\hat{R}/N_{\mathrm{d}}}}.\end{aligned} \tag{4.36}$$

At this point, (4.35) is still true.

Therefore, when $d_1^{\mathrm{UL}^*} \leq d_{L,0} \leq d_0^{\mathrm{UL}^*}$, the error probability of FD-SCMA is lower than that of HD-SCMA in UL.

2) When $x > 0$ and $\beta \geq N/L$, $\frac{d(1+\beta x/N_{\mathrm{d}})^{N_{\mathrm{d}}}}{dx} > \frac{d(1+x)^{N/L}}{dx}$ holds, and we can still obtain the effective SNR relationship as follows

$$\gamma_L^{\mathrm{UL\_HD\_SCMA}} \geq \gamma_L^{\mathrm{UL\_HD\_OMA}}, \quad \forall d_{L,0} \leq d_0^{\mathrm{UL}^*}, \tag{4.37}$$

where $\gamma_L^{\mathrm{UL\_HD\_OMA}} = (1+ \parallel \mathbf{h}_{0,L,0} \parallel^2 q_L^{\mathrm{UL}} d_{L,0}^{-\alpha}/\sigma_0^2)^{N/L} - 1$ represents the minimum effective SNR of UL users in HD-OMA.

According to the monotonically decreasing characteristic of $\Theta(\gamma)$, the error probability performance of HD-SCMA is lower than that of HD-OMA in UL for $d_{L,0} \leq d_0^{\mathrm{UL}^*}$.

As described in step 1, since the error probability of FD-SCMA is lower than that of HD-SCMA, the error probability of FD-SCMA is lower than that of HD-OMA in UL, for $d_1^{\mathrm{UL}^*} \leq d_{L,0} \leq d_0^{\mathrm{UL}^*}$.

In summary, Theorem 5 has been proved. □

Compared with HD-SCMA and HD-OMA in UL, the reliability of FD-SCMA is affected by the capability of the SIS, as described in Theorem 5. When the SIS is sufficiently efficient, FD-SCMA has the lowest error probability. However, when the SIS is not sufficiently efficient, it can be found that FD-SCMA still guarantees better reliability than other schemes within a smaller operation range.

In general, FD-SCMA can gain reliability from the space diversity, spread spec-

trum, and MUD. In addition, it reduces access latency by simultaneous transmission in UL and DL, which is beneficial to support highly reliable communications in 5G.

4.5 Performance in Time-Invariant Frequency-Selective Fading Channels

In general, since the coherence time of the channel with limited Doppler shift is typically greater than transmission duration,it can be assumed that the fading channel is time-invariant in low-latency short-packet transmission [90]. On the other hand, in most URLLC scenarios, the carriers spacing is greater than the coherence frequency, so the channel fading coefficient will vary among carriers. This section analyzes the performance of the time-invariant frequency-selective fading channel, assuming the SSF coefficients of carriers are i.i.d. and follow the Rayleigh distribution ($h_{r,n,l,k} \sim \mathcal{CN}(0,\mu^2)$, $h_{0,0} \sim \mathcal{CN}(0,\mu_0^2)$). An essential reason for the failure of short-packet transmission is that deep fading on the occupied carrier reduces the effective SNR. In this case, the frequency diversity can be obtained from the independent channel gains of different carriers, especially in FD-SCMA. Thus, it is preferred to improve reliability when encountering the frequency-selective deep fading.

To estimate a UB for the error probability, an LB for the user's effective SNR is derived. In the time-invariant frequency-selective fading channel, an LB for the effective SNR in DL and UL can be respectively written as

$$\underline{\gamma}^{\mathrm{DL}} = \big(\sum_{n=1}^{N} |g_{n,K}^{\mathrm{DL}}| \parallel \mathbf{h}_{n,0,K} \parallel^2 \big)\beta P_{\mathrm{MAX}}^{\mathrm{DL}} d_{0,K}^{-\alpha}/KN_{\mathrm{d}}\sigma_K^2, \tag{4.38}$$

and

$$\underline{\gamma}^{\mathrm{UL}} = \frac{\big(\sum_{n=1}^{N} |g_{n,L}^{\mathrm{UL}}| \parallel \mathbf{h}_{n,L,0} \parallel^2 \big)\beta P_{\mathrm{MAX}}^{\mathrm{UL}} d_{L,0}^{-\alpha}}{\kappa |h_{0,0}|^2 N_{\mathrm{d}} P_{\mathrm{MAX}}^{\mathrm{DL}}/N + N_{\mathrm{d}}\sigma_0^2}. \tag{4.39}$$

Here, we can verify that $\underline{\gamma}^{\mathrm{DL}} \le \tilde{\gamma}_K^{\mathrm{DL}} \le \tilde{\gamma}_k^{\mathrm{DL}}$ $(1 \le k \le K)$ and $\underline{\gamma}^{\mathrm{UL}} \le \tilde{\gamma}_L^{\mathrm{UL}} \le \tilde{\gamma}_l^{\mathrm{UL}}$ $(1 \le l \le L)$.

At the same time, since the SSF coefficients obey i.i.d. Rayleigh distribution,

$\hbar^{\mathrm{SI}} \triangleq \frac{1}{\mu_0^2}|h_{0,0}|^2$, $\hbar^{\mathrm{DL}} \triangleq \frac{1}{\mu^2}\sum_{n=1}^{N}|g_{n,K}^{\mathrm{DL}}| \parallel \mathbf{h}_{n,0,K} \parallel^2$, and $\hbar^{\mathrm{UL}} \triangleq \frac{1}{\mu^2}\sum_{n=1}^{N}|g_{n,L}^{\mathrm{UL}}| \parallel \mathbf{h}_{n,L,0} \parallel^2$ obey the Chi-square distribution. We can obtain $\hbar^{\mathrm{SI}} \sim \chi^2(2)$, $\hbar^{\mathrm{DL}} \sim \chi^2(2N_{\mathrm{RX}}N_{\mathrm{d}})$, and $\hbar^{\mathrm{UL}} \sim \chi^2(2N_{\mathrm{RX}}N_{\mathrm{d}})$. (4.38) and (4.39) become

$$\underline{\gamma}^{\mathrm{DL}} = \mu^2\hbar^{\mathrm{DL}}\beta P_{\mathrm{MAX}}^{\mathrm{DL}} d_{0,K}^{-\alpha}/KN_{\mathrm{d}}\sigma_K^2, \tag{4.40}$$

and

$$\underline{\gamma}^{\mathrm{UL}} = \frac{\mu^2\hbar^{\mathrm{UL}}\beta P_{\mathrm{MAX}}^{\mathrm{UL}} d_{L,0}^{-\alpha}}{\mu_0^2\hbar^{\mathrm{SI}}\kappa N_{\mathrm{d}} P_{\mathrm{MAX}}^{\mathrm{DL}}/N + N_{\mathrm{d}}\sigma_0^2}. \tag{4.41}$$

Given a realization of $\hbar^{\mathrm{DL}}$, a UB for the DL users' error probability can be derived as follows

$$\overline{\Psi}^{\mathrm{DL}}\left(\hbar^{\mathrm{DL}}\right) = \begin{cases} 1 & , \quad \text{if } \underline{\Lambda}^{\mathrm{DL}} < 0; \\ Q\left(\sqrt{\frac{M_{\mathrm{RB}}t_{\mathrm{DE}}/T_{\mathrm{SL}}}{V(\underline{\gamma}^{\mathrm{DL}})}}\underline{\Lambda}^{\mathrm{DL}}\right) & , \quad \text{otherwise,} \end{cases} \tag{4.42}$$

where

$$\begin{aligned} \underline{\Lambda}^{\mathrm{DL}} &= C(\underline{\gamma}^{\mathrm{DL}}) - \hat{R} \\ &= \log_2\left(1 + \frac{\mu^2\hbar^{\mathrm{DL}}\beta P_{\mathrm{MAX}}^{\mathrm{DL}} d_{0,K}^{-\alpha}}{KN_{\mathrm{d}}\sigma_K^2}\right) - \frac{DT_{\mathrm{SL}}}{M_{\mathrm{RB}}t_{\mathrm{DE}}}. \end{aligned}$$

Given $\hbar^{\mathrm{UL}}$ and $\hbar^{\mathrm{SI}}$, a UB for the error probability of UL users is

$$\overline{\Psi}^{\mathrm{UL}}\left(\hbar^{\mathrm{UL}}, \hbar^{\mathrm{SI}}\right) = \begin{cases} 1 & , \quad \text{if } \underline{\Lambda}^{\mathrm{UL}} < 0; \\ Q\left(\sqrt{\frac{M_{\mathrm{RB}}t_{\mathrm{DE}}/T_{\mathrm{SL}}}{V(\underline{\gamma}^{\mathrm{UL}})}}\underline{\Lambda}^{\mathrm{UL}}\right) & , \quad \text{otherwise,} \end{cases} \tag{4.43}$$

where

$$\begin{aligned} \underline{\Lambda}^{\mathrm{UL}} &= C(\underline{\gamma}^{\mathrm{UL}}) - \hat{R} \\ &= \log_2\left(1 + \frac{\mu^2\hbar^{\mathrm{UL}}\beta P_{\mathrm{MAX}}^{\mathrm{UL}} d_{L,0}^{-\alpha}}{\mu_0^2\hbar^{\mathrm{SI}}\kappa N_{\mathrm{d}} P_{\mathrm{MAX}}^{\mathrm{DL}}/N + N_{\mathrm{d}}\sigma_0^2}\right) - \frac{DT_{\mathrm{SL}}}{M_{\mathrm{RB}}t_{\mathrm{DE}}}. \end{aligned}$$

Thus, a UB for the DL's error probability can be expressed as

$$\begin{aligned}\overline{\varepsilon}^{\mathrm{DL}} =& \mathbf{E}\left\{\overline{\Psi}^{\mathrm{DL}}\left(\hbar^{\mathrm{DL}}\right)\right\} \\ =& \int_{\Phi^{\mathrm{DL}}}^{\infty} Q\left(\sqrt{\frac{M_{\mathrm{RB}}t_{\mathrm{DE}}}{T_{\mathrm{SL}}V(\underline{\gamma}^{\mathrm{DL}})}}\underline{\Delta}^{\mathrm{DL}}\right) f_{\hbar^{\mathrm{DL}}}(t)dt \\ &+ \int_{0}^{\Phi^{\mathrm{DL}}} 1 \times f_{\hbar^{\mathrm{DL}}}(t)dt \\ =& \int_{\Phi^{\mathrm{DL}}}^{\infty} Q\left(\sqrt{\frac{M_{\mathrm{RB}}t_{\mathrm{DE}}}{T_{\mathrm{SL}}V(\underline{\gamma}^{\mathrm{DL}})}}\underline{\Delta}^{\mathrm{DL}}\right) f_{\hbar^{\mathrm{DL}}}(t)dt + \overline{\varepsilon}_{\mathrm{IBL}}^{\mathrm{DL}},\end{aligned} \tag{4.44}$$

where $\Phi^{\mathrm{DL}} = \frac{(2^{\hat{R}}-1)KN_{\mathrm{d}}\sigma_K^2}{\mu^2\beta P_{\mathrm{MAX}}^{\mathrm{DL}}d_{0,K}^{-\alpha}}$, and $\overline{\varepsilon}_{\mathrm{IBL}}^{\mathrm{DL}}$ is a UB for the error probability of DL in IBL.

The error probability in IBL is part of the error probability in FBL. When the blocklength increases, $\overline{\varepsilon}^{\mathrm{DL}}$ and $\overline{\varepsilon}_{\mathrm{IBL}}^{\mathrm{DL}}$ are close to each other. When the blocklength tends to become infinity, the two values are equivalent.

At this time, a UB for the DL's error probability in IBL can be expressed as

$$\begin{aligned}\overline{\varepsilon}_{\mathrm{IBL}}^{\mathrm{DL}} &= \Pr\left(C(\underline{\gamma}^{\mathrm{DL}}) < D/M\right) = \Pr(\hbar^{\mathrm{DL}} < \Phi^{\mathrm{DL}}) \\ &= \frac{\gamma(N_{\mathrm{RX}}N_{\mathrm{d}}, \Phi^{\mathrm{DL}}/2)}{\Gamma(N_{\mathrm{RX}}N_{\mathrm{d}})},\end{aligned} \tag{4.45}$$

where the Gamma function is expressed as $\Gamma(n) = (n-1)!$ $(n \in \mathbb{Z}^+)$, and the lower incomplete Gamma function can be expressed as $\gamma(s,x) = \int_0^x t^{s-1}e^{-t}dt$. It can be seen that when the signal is spread on N_{d} independent carriers, the effective SNR will increase, and the probability of the transmission failure decreases sharply.

In high-reliability transmission scenarios, the cumulative distribution of $\hbar^{\mathrm{UL}} \sim \chi^2(2N_{\mathrm{RX}}N_{\mathrm{d}})$ can be approximated by $\Pr(\hbar^{\mathrm{UL}} < x) \approx \frac{x^{N_{\mathrm{RX}}N_{\mathrm{d}}}}{(N_{\mathrm{RX}}N_{\mathrm{d}})!}$. Thus, a UB for the

Table 4.1 : Simulation Parameters

Parameter	Definition	Value
D	Number of information bits (bit)	256
R_{D}	Cell range (m)	200
f_{SC}	Subcarrier spacing (kHz)	120
T_{SL}	Slot duration (μs)	18
N	Number of total carriers	10
K	Number of DL users	15
L	Number of UL users	15
σ_0^2, σ_k^2	Variance of noise (dBm)	-115
N_{d}	Number of occupied carriers by one user	4
μ	Variance of Rayleigh variables $h_{r,n,l,k}$	1
μ_0	Variance of Rayleigh variables $h_{0,0}$	0.1
α	Path loss exponent	3.6
κ	SIS factor (dB)	$-100 \sim -130$
β	Capability of the MUI cancelation	0.88

UL's error probability in IBL can be calculated by

$$\begin{aligned}
\overline{\varepsilon}_{\mathrm{IBL}}^{\mathrm{UL}} =& \Pr\left(C(\underline{\gamma}^{\mathrm{UL}}) < D/M\right) \\
=& \mathbf{E}\left\{\Pr(\hbar^{\mathrm{UL}} < \Phi^{\mathrm{UL}})\right\} \approx \mathbf{E}\left\{\frac{(\Phi^{\mathrm{UL}})^{N_{\mathrm{RX}}N_{\mathrm{d}}}}{(N_{\mathrm{RX}}N_{\mathrm{d}})!}\right\} \\
=& \frac{1}{(N_{\mathrm{RX}}N_{\mathrm{d}})!}\int_0^{\infty}\left\{e^{-\hbar^{\mathrm{SI}}}\left(\frac{1}{\mu^2\beta P_{\mathrm{MAX}}^{\mathrm{UL}} d_{L,0}^{-\alpha}}\right)^{N_{\mathrm{RX}}N_{\mathrm{d}}}\right. \\
& \left.\left((2^{\hat{R}}-1)N_{\mathrm{d}}({\mu_0}^2\hbar^{\mathrm{SI}}\kappa P_{\mathrm{MAX}}^{\mathrm{DL}}/N + \sigma_0^2)\right)^{N_{\mathrm{RX}}N_{\mathrm{d}}}\right\} d\hbar^{\mathrm{SI}} \\
=& \frac{e^{\vartheta}\Gamma(1+N_{\mathrm{RX}}N_{\mathrm{d}},\vartheta)}{(N_{\mathrm{RX}}N_{\mathrm{d}})!}\left(\frac{(2^{\hat{R}}-1)N_{\mathrm{d}}\sigma_0^2}{\mu^2\beta P_{\mathrm{MAX}}^{\mathrm{UL}} d_{L,0}^{-\alpha}\vartheta}\right)^{N_{\mathrm{RX}}N_{\mathrm{d}}},
\end{aligned} \tag{4.46}$$

where $\vartheta = \frac{N\sigma_0^2}{\mu_0\kappa P_{\mathrm{MAX}}^{\mathrm{DL}}}$.

4.6 Numerical Results and Analysis

The numerical and simulation results are provided to verify the theoretical analyses in time-invariant frequency-selective Rayleigh fading channels. In order to verify the UB for the error probability, DL user K and UL user L are assumed to be placed at the cell boundary of 200 meters from the gNB. The other simulation parameters

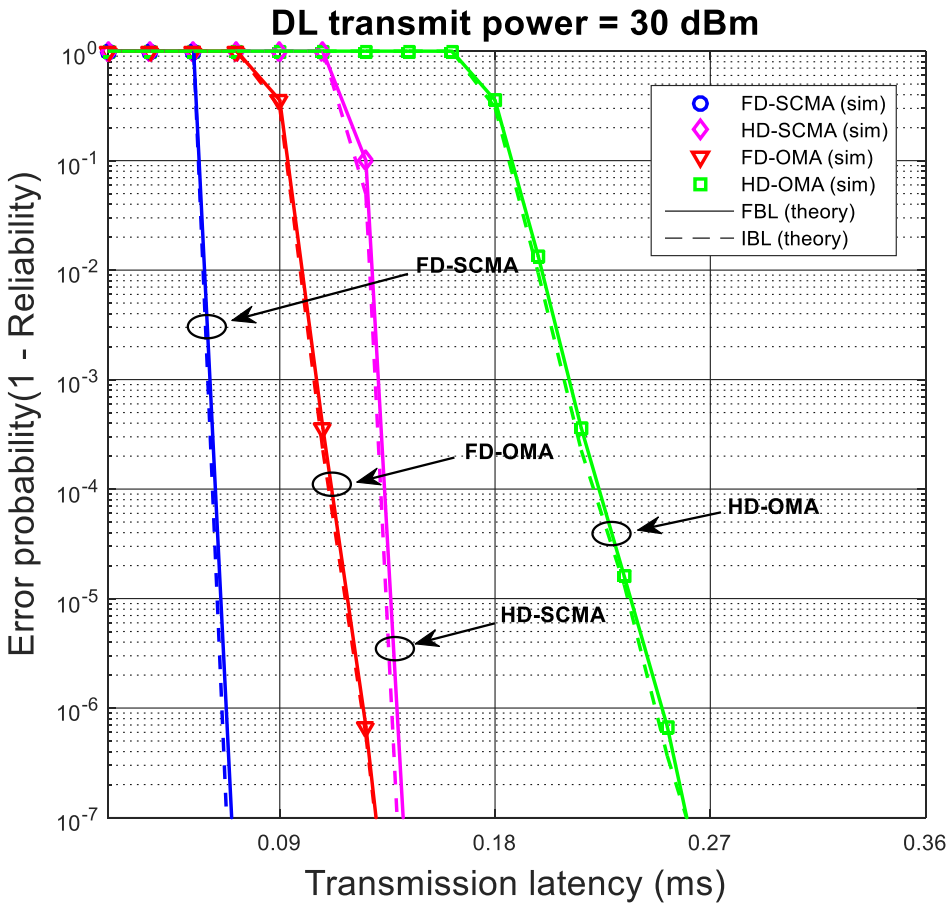

Figure 4.2 : The trade-off between reliability and transmission latency in DL, when $P^{\mathrm{DL}} = 30$ dBm.

are listed in Table 4.1.

4.6.1 DL Performance

Reliability in DL is presented in Figs. 4.2 and 4.3. The solid lines denote the theoretical results in FBL; the dotted lines denote the theoretical results in IBL; and the various markers denote the simulation results. Fig. 4.2 illustrates the trade-off between reliability and latency when DL Tx power is 30 dBm. FD-SCMA can reduce the latency by more than 0.06 ms compared to other schemes given 10^{-5} error probability. Fig. 4.3 suggests that FD-SCMA can significantly enhance

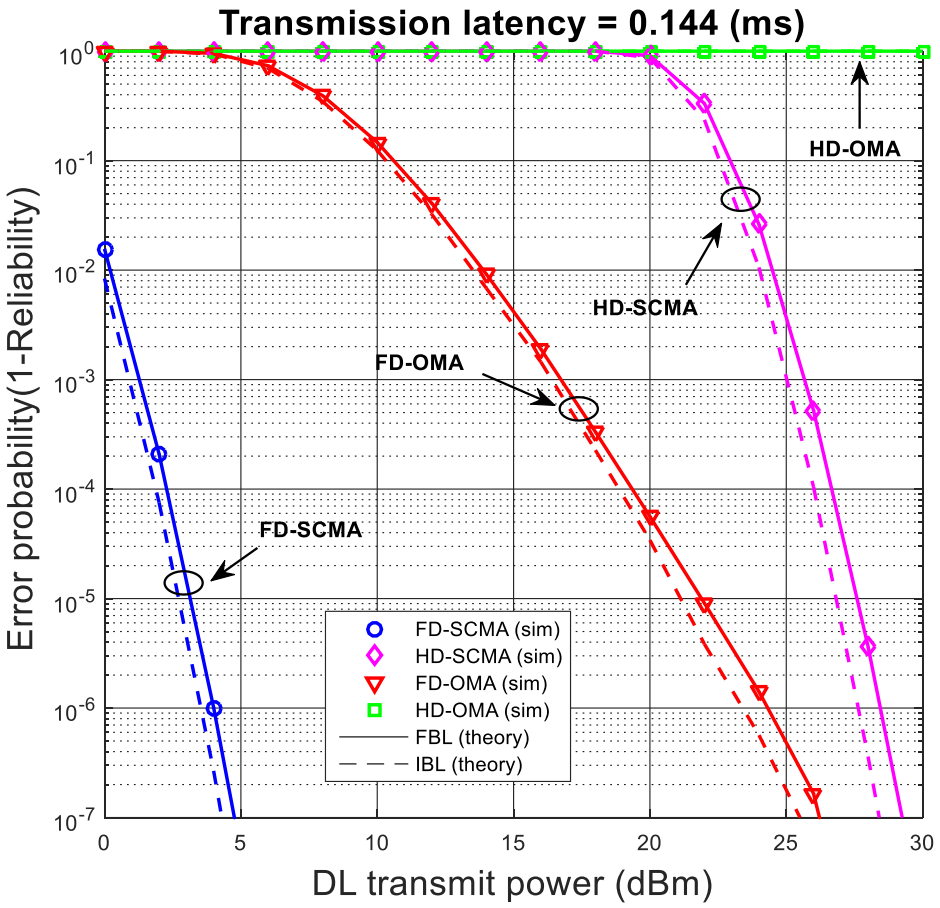

Figure 4.3 : The trade-off between reliability and transmission latency in DL, when $t_{\mathrm{DE}} = 0.144$ ms.

reliability in DL when DL Tx power increases. When DL Tx power is about 5 dBm, FD-SCMA is sufficient to support DL URLLC high-reliability communications. FD-SCMA outperforms the other schemes in terms of reliability and latency, and the Monte Carlo simulation results can match the analysis in FBL well.

Figs. 4.4 and 4.5 illustrate the relationship between reliability and transmission latency when the gNB applies different Tx power (20 dBm and 10 dBm). The figure suggests that the reliability degradation of FD-SCMA is almost negligible when the Tx power of gNB reduces to 10 dBm, compared to other existing schemes. Therefore, self-interference can be effectively minimized by reducing DL Tx power.

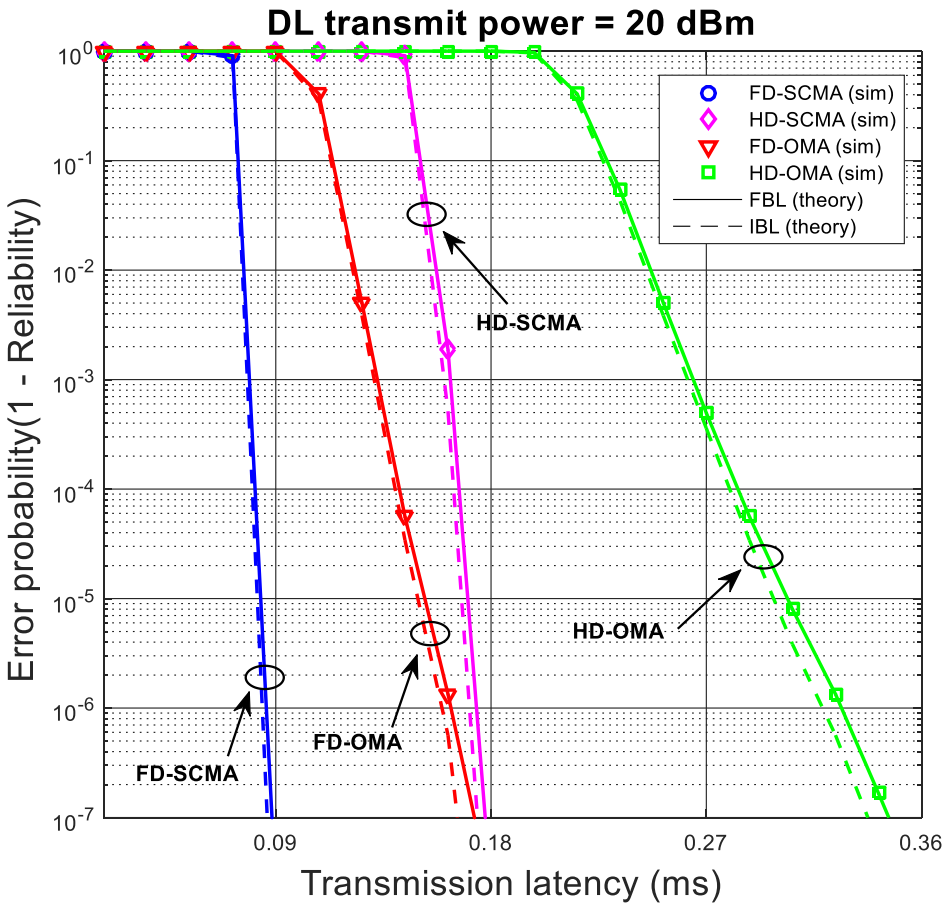

Figure 4.4 : The trade-off between reliability and transmission latency in DL, when $P^{\mathrm{DL}} = 20$ dBm.

Furthermore, when the error probability is lower than 10^{-5}, FD-SCMA can shorten the latency by approximate 0.07 ms under 20 dBm DL Tx power. At the same time, with 10 dBm DL Tx power, FD-SCMA can obtain a latency that is 0.11 ms lower than other schemes.

4.6.2 UL Performance

Reliability in UL is presented in Fig. 4.6, where DL Tx power is set to 30 dBm and the SIS factor is set to $\kappa = -130$ dB. Similarly, FD-SCMA outperforms the existing schemes, and Monte Carlo simulation results match FBL theoretical results

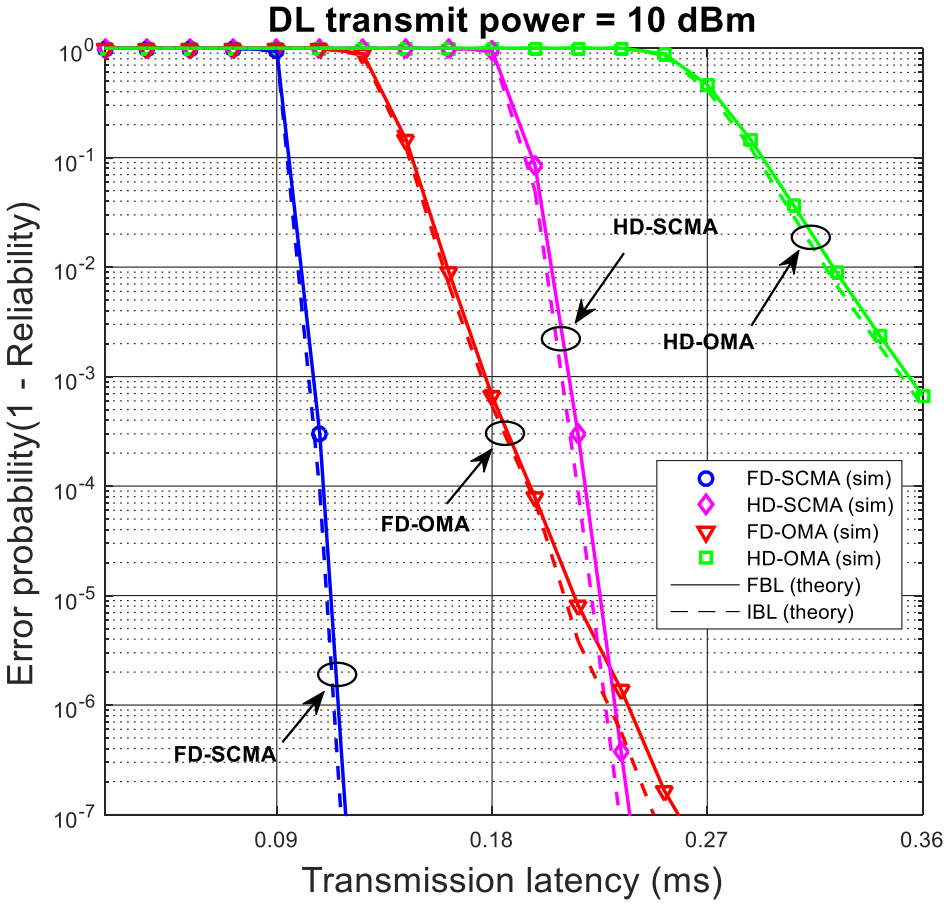

Figure 4.5 : The trade-off between reliability and transmission latency in DL, when $P^{\mathrm{DL}} = 10$ dBm.

well in UL. Given 10^{-5} error probability, the latency can be reduced by more than 0.06 ms in FD-SCMA. FD-SCMA can achieve significant reliability improvements in both UL and DL under extremely low transmission latency. This indicates the proposed FD-SCMA can outstandingly satisfy the requirements of both low latency and high reliability.

Fig. 4.7 and Fig. 4.8 illustrate the trade-off between reliability and latency for different SIS factors. Fig. 4.7 suggests reliability in UL affected by the capability of SIS. When κ decreases from -100 to -110 dB, the latency of UL users can be reduced by 0.02 ms in order to achieve 10^{-5} error probability. However, when κ de-

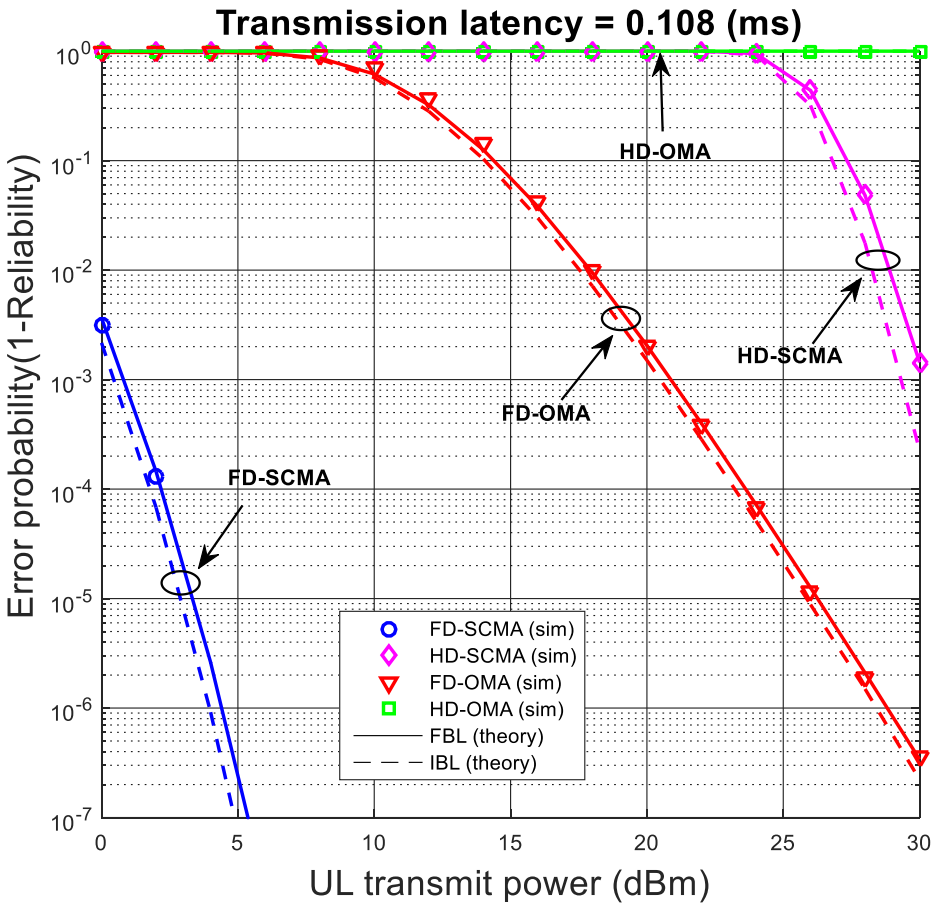

Figure 4.6 : The relationship between the reliability and Tx power in UL, under imperfect SIS ($\kappa = -130$ dB, $P^{\mathrm{DL}} = 30$ dBm).

creases from -120 to -130 dB, the latency remains almost unchanged in FD-SCMA. This indicates that FD-SCMA can achieve the desired result under incomplete SIS, which is consistent with the previous analysis. In addition, Fig. 4.8 verifies that the reliability and latency performance of FD-SCMA in UL can be improved by reducing DL Tx power under imperfect SIS. When DL Tx power is decreased from 30 dBm to 20 dBm, the latency can be reduced by more than 0.02 ms. In general, reduction on the self-interference of the gNB can increase reliability and shorten the latency in UL moderately. However, when residual self-interference is further reduced, the influence of noise at the receiver will become the dominant factor. It implies that

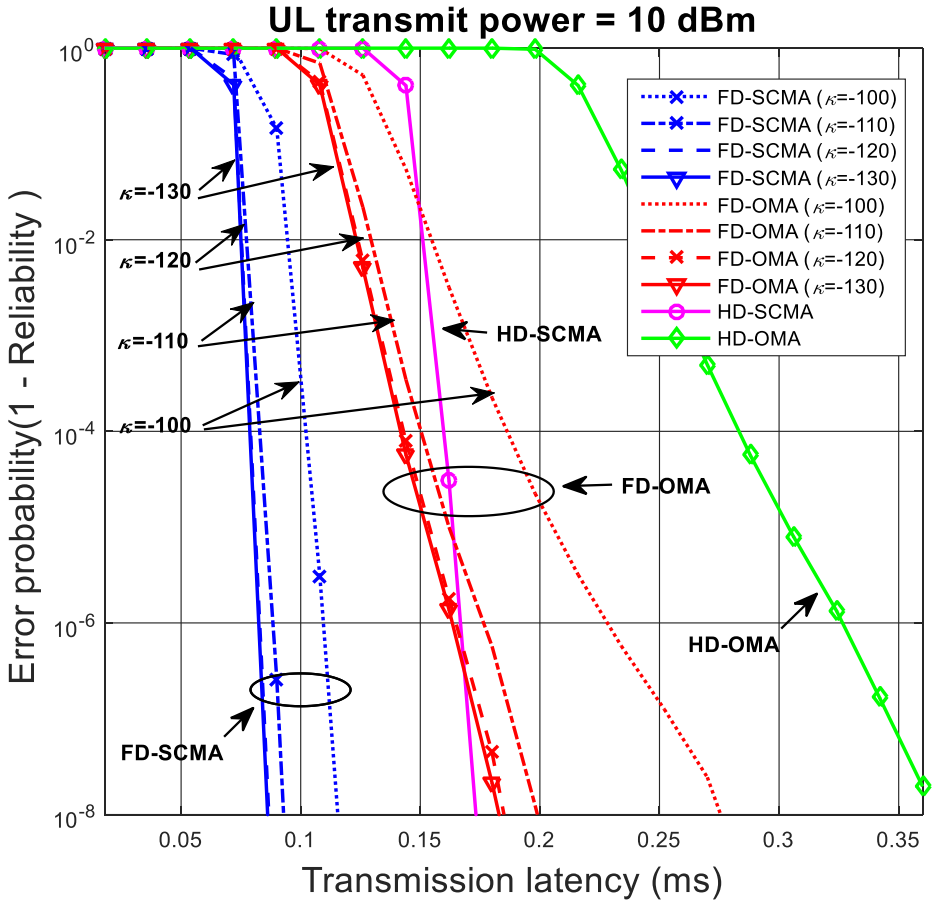

Figure 4.7 : The relationship between the reliability and transmission latency in UL, under imperfect SIS ($P^{\mathrm{DL}} = 30$ dBm).

FD-SCMA can achieve performance near perfect SIS in the case of imperfect SIS.

4.6.3 Balance Between UL and DL

The error probability of FD-SCMA in UL is affected by the capability of SIS, and the self-interference can be minimized by reducing DL Tx power. Considering that the error probability in DL can be much lower than that in UL, the UL error probability can be lowered by reducing DL Tx power, while ensuring that DL error probability is still within an acceptable range.

As demonstrated in Fig. 4.9, the self-interference of the gNB is mitigated by

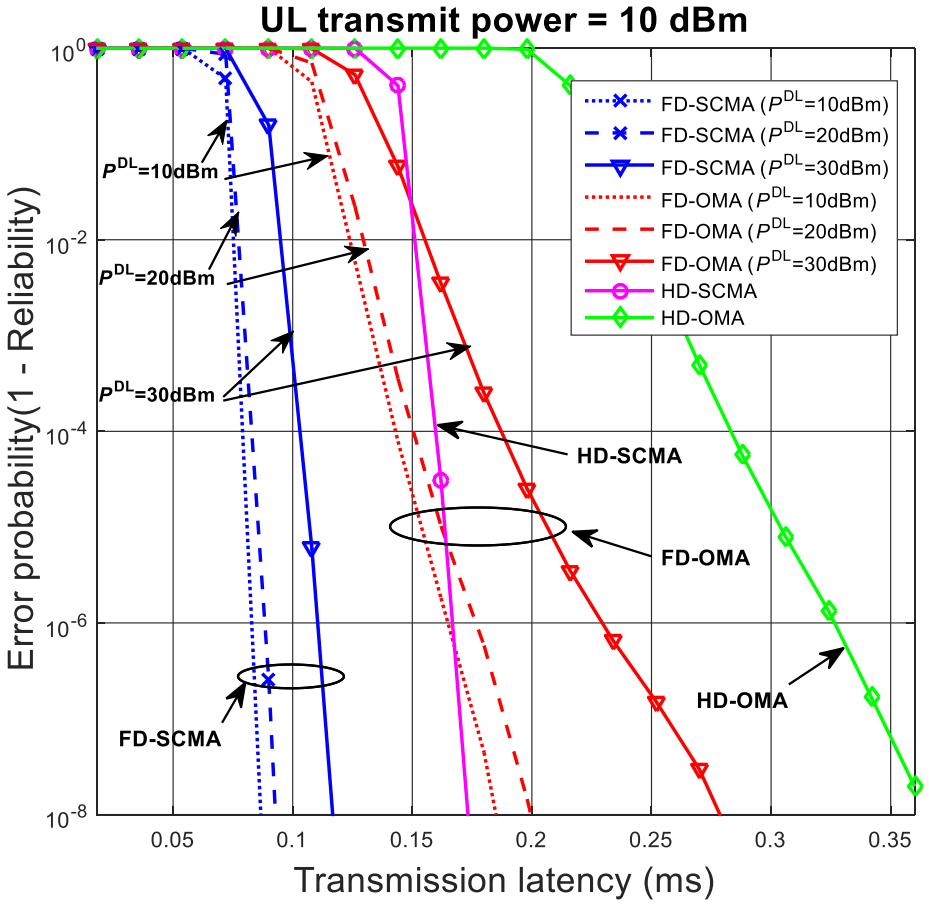

Figure 4.8 : The relationship between the reliability and transmission latency in UL, under imperfect SIS ($\kappa = -130$ dB).

lowering DL Tx power, and so a compromise between the UL and DL reliability can be achieved. When DL Tx power is reduced from 30 dBm to 10 dBm, the UL reliability will significantly increase. In addition, when DL Tx power is decreased to 10 dBm, the maximum error probability of all UL and DL users can be minimized. Therefore, in an FD system with imperfect SIS, power control in DL is effective in guaranteeing the reliability both in DL and UL transmission.

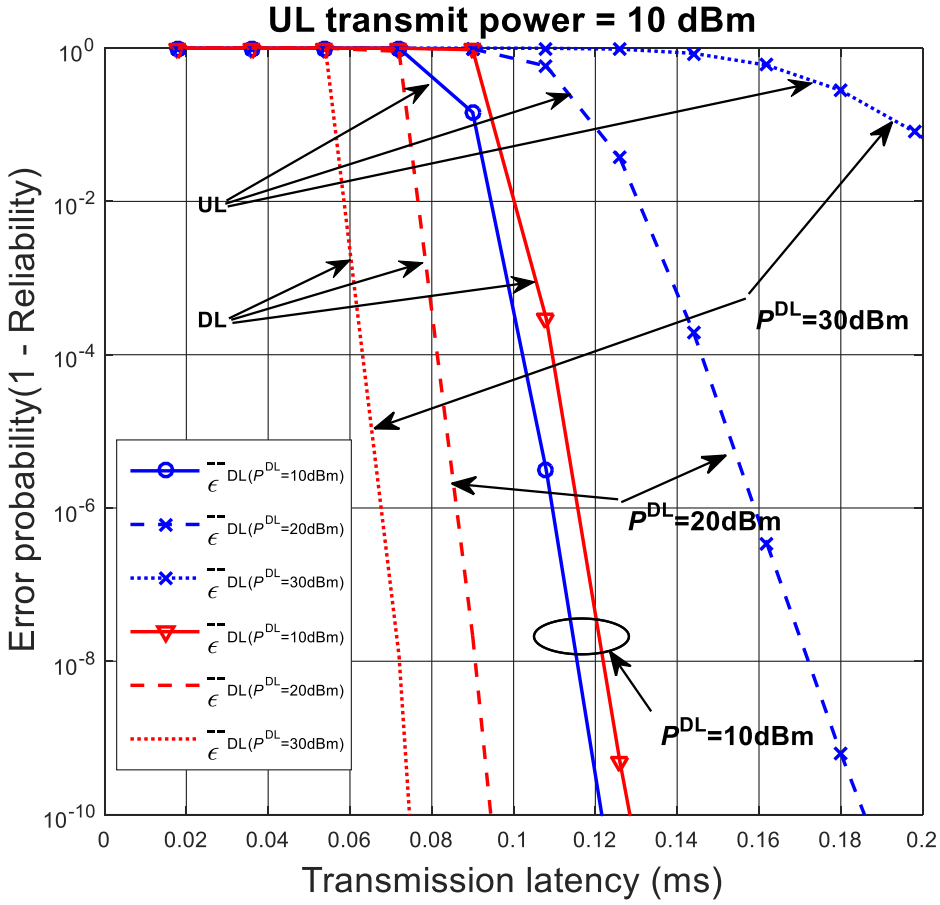

Figure 4.9 : The balance of UL and DL reliability by adjusting DL Tx power in FD-SCMA.

4.7 Summary of This Chapter

This chapter proposes FD-SCMA and analyzes its reliability with constrained latency. Then, FD-SCMA's advantages are theoretically proved in time-invariant flat-fading channels, and are verified in time-invariant frequency-selective fading channels. It can be seen that FD-SCMA can sufficiently resist the fading channel by implementing frequency diversity of multiple occupied carriers under independent channel fading. In addition, it obtains the resource-reuse through the support of FD transmission, and more radio resources can bring the diversity gain of the code domain. As demonstrated in simulation results, when the packet size is 32 byte, the

proposed scheme can achieve higher than 99.999% reliability within 0.2 ms transmission latency. Thus, the FD-SCMA scheme exhibits outstanding performance in satisfying the critical requirements of URLLC in various IoT scenarios. Like most NOMA transmission and detection schemes, the proposed FD-SCMA scheme has relatively high requirements for CSI accuracy acquired by the receiver. Its advantages under perfect CSI are verified by theoretical analysis and simulation. Under perfect CSI, FD-SCMA generally outperforms existing FD-OMA, HD-SCMA, and HD-OMA schemes in terms of error probability and transmission latency. Moreover, the performance of the superposed transmission and detection techniques under imperfect CSI will be analyzed and verified in Chapter 5.

Chapter 5

MU-MIMO NOMA Under Perfect and Imperfect CSI

Enabling 5G URLLC under severe channel fading in IoT is particularly challenging. This chapter will combine MU-MIMO with NOMA to achieve URLLC with PACE and ZF detection. Assuming UL users are uniformly and randomly distributed in the cell and face log-normal shadow fading, this chapter derives the pdf of post-processing SNRs in UL MU-MIMO NOMA with perfect and imperfect CSI obtained by PACE. Further, the FBL information theory is applied to derive the error probability of accessed users under a given latency, and thus evaluate the reliability of MU-MIMO NOMA. In addition, this chapter determines the LoP by applying GSSM to minimize the error probability. This method can converge rapidly and lower the overhead of pilots significantly. Through numerical calculations and simulation results, this chapter verifies that MU-MIMO NOMA can also support URLLC when users face severe shadow fading. The number of accessed users can be further increased with a substantial number of Rx antennas. This chapter implements ZF detection to support a high volume of UL users with lower computational complexity and reduces access and processing latency. At the same time, the reliability and transmission latency of MU-MIMO under imperfect CSI can be further reduced by optimizing the LoP.

5.1 Research Significance of MU-MIMO NOMA

Diversity plays an essential role in the reliability improvement of URLLC, and it can be acquired from the following domains. First, space diversity can be obtained from different signals of different Rx antennas that pass through different propagation paths [41, 91]. Further, frequency diversity can be acquired by spread-

ing signals on different carriers with independent SSF coefficients. Moreover, time domain diversity can be obtained from retransmission. Last, code domain diversity can be achieved by low-rate coding and modulation schemes can exploit more channel uses (CUs). Considering deep fading that might be encountered in time-variant and frequency-selective channels, proper diversity-based techniques are needed to maintain ultra-reliable transmission. However, general IoT services are supported with limited bandwidth, and with limited retransmission when the latency is strictly constrained to be less than 0.5 ms. It is challenging for IoT services to obtain sufficient diversity gains from the frequency domain and the time domain. Therefore, it is necessary to optimize the diversity gain from the potential spatial domain and the code domain, which leads this chapter to focus on the MU-MIMO technology. In general, MU-MIMO significantly improves reliability by installing a great number of Rx antennas and simultaneously adopting all available CUs to jointly serve multiple users (rather than assigning a small number of CUs to each user with limited bandwidth and latency). It is worth mentioning that the reliability of MU-MIMO is highly dependent on precise CSI. In the freqency division duplexing (FDD) system, precise CSI should be estimated and fed back in a strictly limited latency. However, in a time division duplexing (TDD) system, UL CSI could be acquired from the DL CSI measurement with reciprocity. Increasing the LoP improves the accuracy of estimated CSI, but it might reduce reliability with fewer CUs allocated to data at the same post-processing SNR. To this end, finding a proper LoP can improve reliability with constrained latency. At the same time, typical IoT services face deep fading when encountering poor propagation channels, such as non-line-of-sight propagation and severe shadow fading, while it is difficult to obtain real-time and highly accurate CSI through long-term observation due to latency constraints. In order to overcome these challenges, this chapter explores corresponding research in MU-MIMO.

Under the perfect and imperfect CSI assumptions, the LBs for the capacity of large-scale MU-MIMO detection in MRC, MMSE, and ZF were derived in [92]. The post-processing SNR of MU-MIMO using MMSE detection was derived under

the imperfect CSI assumption [93, 94]. In millimeter-wave enabled massive MIMO systems, channel variations and probabilistic constraints on reliability and latency were modeled as the network utility maximization problems, and then they were solved via Lyapunov optimization [91]. When the UE installs only one antenna, it is verified that the gNB with an appropriate number of Rx antennas could ensure URLLC [41]. Different from previous research, this chapter mainly demonstrates the following work:

- Assuming CSI can be estimated perfectly or imperfectly, this chapter deduces the pdf of post-processing SNRs with ZF detection in MU-MIMO. Then, this chapter applies the FBL information theory to derive the error probability under shadow fading in short-packet transmission.
- Based on the error probability derived, this chapter optimizes the LoP under different numbers of accessed users and different latency constraints. Furthermore, a low-complexity GSSM is proposed to effectively determine the LoP and achieve near-optimal reliability.
- Through numerical and simulation results, this chapter verifies that MU-MIMO can also support URLLC in the appearance of severe shadow fading. When the number of accessed users increases, reliability and latency can be guaranteed by increasing the number of Rx antennas equipped at the gNB.
- Compared with fixed-ratio pilot allocation, this chapter verifies the effectiveness of the proposed GSSM, which can maintain the error probability of MU-MIMO and the pilot overhead at a reasonable level simultaneously.

The significant differences between the work in this chapter and previous studies are discussed below. Firstly, current studies on MU-MIMO mainly focus on maximizing data rates and assume that the gNB can obtain accurate UL CSI. This chapter assumes CSI is mainly estimated by PACE, which has certain channel estimation errors. This chapter jointly optimizes latency and reliability, and then derives the error probability under the constrained latency. Secondly, existing research ignores some of the key features of IoT applications, including extremely

short packets, a great number of randomly deployed users, and severe shadow fading. Here, this chapter applies the emerging FBL information theory to analyze the error probability of randomly deployed users under severe shadow fading. With PACE, the optimal LoP can be determined within the given latency. The study on the reliability of accessed users in short-packet transmission has not been mentioned in previous literature on MU-MIMO. Finally, rather than studies on how to ensure URLLC when serving multiple users simultaneously, the preliminary work on MU-MIMO focuses on MMSE detection and attempts to maximize ergodic achievable sum-rate. This chapter mainly studies ZF detection, which can effectively mitigate the interference caused by other users. Also, ZF detection can significantly improve the transmission reliability through the spatial diversity brought by massive Rx antennas. When the number of Rx antennas is large, ZF can achieve near-optimal ergodic sum-rate [92], while its detection complexity is much lower than that of MMSE's.

5.2 System Model

5.2.1 Signal Model of PACE

The MU-MIMO system includes a gNB equipped with L Rx antennas and K accessed users equipped with a Tx antenna ($L \geq K$). Each user is assumed to transmit a short packet carrying D information bits through N CUs, which occupy a total of B Hz bandwidth and t_{DE} second (s) time ($N = Bt_{\mathrm{DE}}$).

In MU-MIMO, the received signal at the gNB can be given by

$$\mathbf{Y} = \sqrt{\rho}\mathbf{G}\mathbf{X} + \mathbf{Z}, \tag{5.1}$$

where ρ is the transmit SNR, $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_K]^T \in \mathbb{C}^{K\times N}$, and $\mathbf{x}_k = [x_{k,1}, \ldots, x_{k,N}]^T$ represent transmitted signals from user k; $\mathbf{G} \in \mathbb{C}^{L\times K}$ represents the channel coefficient matrix between all K users and the gNB, and $\mathbf{Z} \in \mathbb{C}^{L\times N}$ is the AWGN at the receiver with i.i.d. $\mathcal{CN}(0, 1)$ elements.

$\mathbf{g}_k = \sqrt{\beta_k s_k}\mathbf{h}_k$ represents the channel fading coefficient vector from user k to

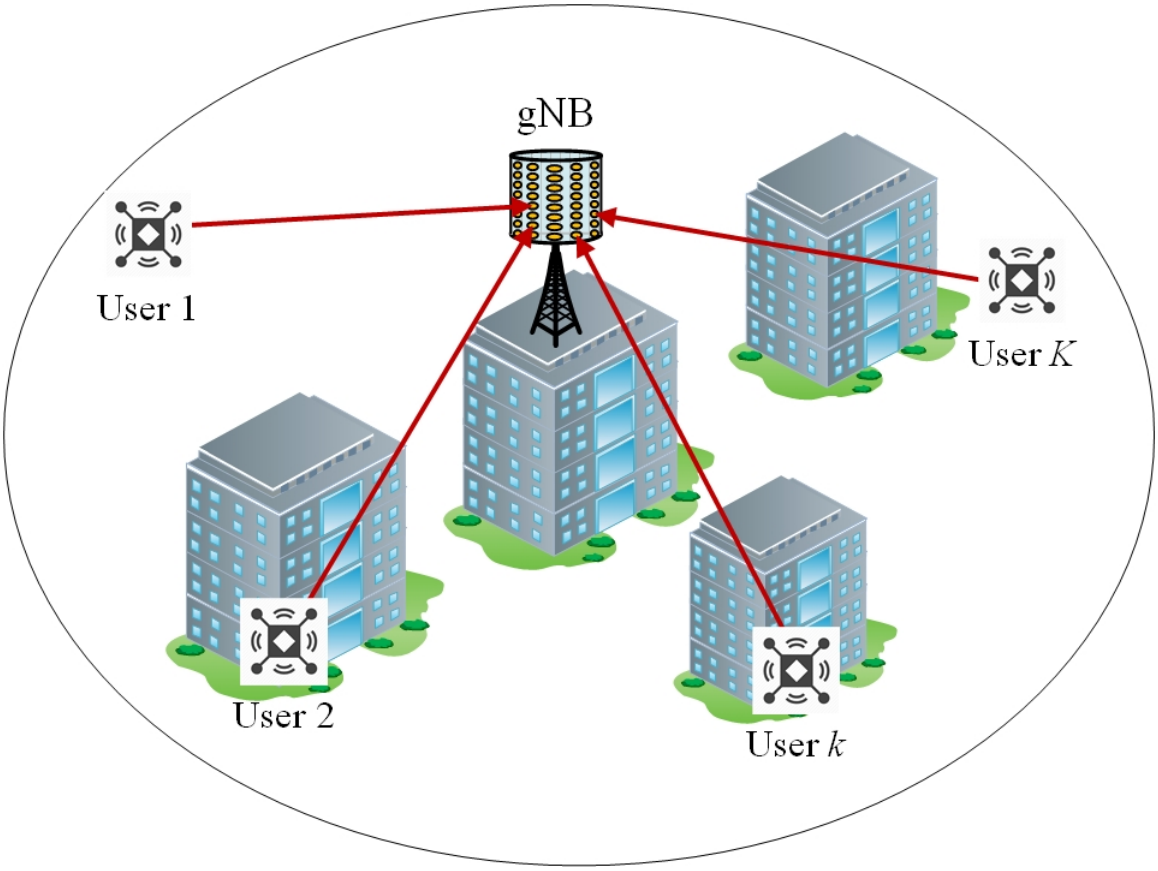

Figure 5.1 : An illustration of the UL MU-MIMO NOMA system.

the gNB, in which $\mathbf{h}_k = [h_{k,1}, ..., h_{k,L}]^T \sim \mathcal{CN}(0, \mathbf{I}_L)$ indicates the SSF channel coefficients of user k that follow i.i.d. Rayleigh distributions; β_k represents the path loss between user k and the gNB; s_k represents the shadow fading of user k that follows a log-normal distribution with variance σ_S^2.

Since path loss and shadow fading change slowly, their channel coefficients can be assumed to be perfectly estimated at the gNB [94–96].

There are two classical assumptions on the estimation of SSF channel coefficients. Most of the previous research on URLLC assumes that perfect CSI can be obtained at the gNB [39, 89], where all N CUs can be used to carry the information bits. Thus, the channel code blocklength of each short packet is N. If perfect CSI is not available, practical channel estimation should be performed at the gNB. In this chapter, PACE is implemented with m CUs bearing specific pilots, where m

represents the LoP. These known pilots are utilized at the gNB to estimate the channel coefficients of instantaneous SSF. The remaining CUs can be utilized to bear the required information bits, so the blocklength is $M = N - m$ in PACE.

For simplification, the channel matrix $\mathbf{G}$ can be rewritten as

$$\mathbf{G} = [\mathbf{g}_1, \ldots, \mathbf{g}_K] \triangleq \mathbf{H}\mathbf{D}^{\frac{1}{2}}\mathbf{S}^{\frac{1}{2}}, \tag{5.2}$$

where $\mathbf{H} = [\mathbf{h}_1, \ldots, \mathbf{h}_K] \in \mathbb{C}^{L\times K}$, $\mathbf{D} = \operatorname{diag}(\beta_1, \ldots, \beta_K) \in (\mathbb{R}^+)^{K\times K}$, and $\mathbf{S} = \operatorname{diag}(s_1, \ldots, s_K) \in (\mathbb{R}^+)^{K\times K}$ represent Rayleigh fading, path loss, and shadow fading, respectively.

Here, the pdf of the LSF channel coefficients β_k and s_k can be derived. Since the users are uniformly and randomly dispersed in the area, $\beta_k = \mu_{\mathrm{P}}(d_k/d_0)^{-\alpha_{\mathrm{P}}}$ ($d_0 \leq d_k \leq d_{\mathrm{R}}$), where d_0 is the minimum distance between users and the gNB, d_{R} is the cell radius, d_k is the distance between user k and the gNB, μ_{P} denotes path loss at the minimum distance d_0, and α_{P} denotes the path loss exponent.

Thus, the cumulative distribution function of β_k is given by

$$\begin{aligned} F_{\beta_k}(x) =& \mathbb{P}\left(\mu_{\mathrm{P}}(d_k/d_0)^{-\alpha_{\mathrm{P}}} \leq x\right) = \mathbb{P}(d_k \geq d_0\mu_{\mathrm{P}}^{\frac{1}{\alpha_{\mathrm{P}}}}x^{-\frac{1}{\alpha_{\mathrm{P}}}}) \\ =& \frac{d_{\mathrm{R}}^2 - d_0^2\mu_{\mathrm{P}}^{\frac{2}{\alpha_{\mathrm{P}}}}x^{-\frac{2}{\alpha_{\mathrm{P}}}}}{d_{\mathrm{R}}^2 - d_0^2} = \frac{\vartheta^2 - \mu_{\mathrm{P}}^{\frac{2}{\alpha_{\mathrm{P}}}}x^{-\frac{2}{\alpha_{\mathrm{P}}}}}{\vartheta^2 - 1}, \quad \forall x \in [\mu_{\mathrm{P}}\vartheta^{-\alpha_{\mathrm{P}}}, \mu_{\mathrm{P}}], \end{aligned} \tag{5.3}$$

where $\vartheta \triangleq d_{\mathrm{R}}/d_0$. At this time, the pdf of β_k is as follows,

$$f_{\beta_k}(x) = \frac{2\mu_{\mathrm{P}}^{\frac{2}{\alpha_{\mathrm{P}}}}x^{-\frac{2}{\alpha_{\mathrm{P}}}-1}}{\alpha_{\mathrm{P}}(\vartheta^2 - 1)}, \quad \forall x \in [\mu_{\mathrm{P}}\vartheta^{-\alpha_{\mathrm{P}}}, \mu_{\mathrm{P}}]. \tag{5.4}$$

When user k is stationary given path loss $\bar{\beta}_k$, the LSF $\bar{\beta}_k s_k$ is determined by the standard deviation of log-normal shadow fading σ_{S} (in dB) as follows [10, 92, 97],

$$f_{\bar{\beta}_k s_k|\bar{\beta}_k}(x) = \frac{\varphi}{\sqrt{2\pi}\sigma_{\mathrm{S}}x}\exp\left[-\frac{(\varphi\ln x - \varphi\ln\bar{\beta}_k)^2}{2\sigma_{\mathrm{S}}^2}\right], \quad \forall x > 0, \tag{5.5}$$

where $\varphi = 10/\ln 10$, and $\varphi\ln\bar{\beta}_k$ is the mean of $10\log_{10}(\bar{\beta}_k s_k)$, representing the dB

value of path loss [10, 97].

Further, the pdf of $\beta_k s_k$ can be derived using the formula of total probability as follows,

$$\begin{aligned} f_{\beta_k s_k}(x) &= \int_{\mu_{\mathrm{P}}\vartheta^{-\alpha_{\mathrm{P}}}}^{\mu_{\mathrm{P}}} f_{\beta_k}(w) f_{ws_k|w}(x)\mathrm{d}w \\ &= \int_{\mu_{\mathrm{P}}\vartheta^{-\alpha_{\mathrm{P}}}}^{\mu_{\mathrm{P}}} \frac{x\mu_{\mathrm{P}}}{w^2} \frac{2\mu_{\mathrm{P}}^{\frac{2}{\alpha_{\mathrm{P}}}-1} w^{-\frac{2}{\alpha_{\mathrm{P}}}+1}}{\alpha_{\mathrm{P}}(\vartheta^2-1)} \frac{\varphi}{\sqrt{2\pi}\sigma_{\mathrm{S}}x^2} \left[\exp\left(-\frac{(\varphi\ln x - \varphi\ln w)^2}{2\sigma_{\mathrm{S}}^2}\right)\right]\mathrm{d}w \\ &\overset{(a)}{=} \int_{x}^{x\vartheta^{\alpha_{\mathrm{P}}}} \frac{2\mu_{\mathrm{P}}^{\frac{2}{\alpha_{\mathrm{P}}}-1} (\frac{x\mu_{\mathrm{P}}}{z})^{-\frac{2}{\alpha_{\mathrm{P}}}+1}}{\alpha_{\mathrm{P}}(\vartheta^2-1)} \frac{\varphi}{\sqrt{2\pi}\sigma_{\mathrm{S}}x^2} \left[\exp\left(-\frac{\left(\varphi\ln x - \varphi\ln(\frac{x\mu_{\mathrm{P}}}{z})\right)^2}{2\sigma_{\mathrm{S}}^2}\right)\right]\mathrm{d}z \\ &= \int_{x}^{x\vartheta^{\alpha_{\mathrm{P}}}} \frac{\sqrt{2}\varphi x^{-\frac{2}{\alpha_{\mathrm{P}}}-1} z^{\frac{2}{\alpha_{\mathrm{P}}}-1}}{\sqrt{\pi}\sigma_{\mathrm{S}}\alpha_{\mathrm{P}}(\vartheta^2-1)} \exp\left(-\frac{(\varphi\ln z - \mu_{\mathrm{S}})^2}{2\sigma_{\mathrm{S}}^2}\right)\mathrm{d}z, \quad \forall x > 0, \end{aligned} \tag{5.6}$$

where $\mu_{\mathrm{S}} \overset{\Delta}{=} \varphi\ln\mu_{\mathrm{P}} = 10\log_{10}\mu_{\mathrm{P}}$ denotes path loss at the minimum distance d_0 in dB, and the equality (a) holds when letting $z \overset{\Delta}{=} \frac{x\mu_{\mathrm{P}}}{w}$ (then $\mathrm{d}z = -\frac{x\mu_{\mathrm{P}}}{w^2}\mathrm{d}w$).

5.2.2 ZF Detection Under Perfect CSI

In general, when perfect CSI is available, all elements in $\mathbf{X}$ can be applied to carry information bits, and they are assumed to be uncorrelated, as $\mathbb{E}\left(|[\mathbf{X}]_{k,l}|^2\right) = 1$ ($1 \le k \le K$ and $1 \le l \le N$) and $\mathbb{E}\left([\mathbf{X}]_{k,l}[\mathbf{X}]_{i,j}\right) = 0$ ($i \ne k$ or $j \ne l$).

In this case, ZF detection vector $\mathbf{V}^{\mathrm{ZF}} = (\mathbf{G}^H\mathbf{G})^{-1}\mathbf{G}^H$ can be utilized at the gNB [92]. At this time, the post-processing signal of ZF detection is

$$\begin{aligned} \mathbf{V}^{\mathrm{ZF}}\mathbf{Y} &= \sqrt{\rho}\left(\mathbf{G}^H\mathbf{G}\right)^{-1}\mathbf{G}^H\mathbf{G}\mathbf{X} + \left(\mathbf{G}^H\mathbf{G}\right)^{-1}\mathbf{G}^H\mathbf{Z} \\ &= \sqrt{\rho}\mathbf{X} + \left(\mathbf{G}^H\mathbf{G}\right)^{-1}\mathbf{G}^H\mathbf{Z}. \end{aligned} \tag{5.7}$$

The kth row of $\mathbf{V}^{\mathrm{ZF}}\mathbf{Y}$ can be applied to estimate the user's signal $\mathbf{x}_k$, so the post-

processing SNR of user k can be derived as follows [92]

$$\begin{aligned}\gamma_k &= \frac{\rho}{\mathbb{E}\left\{\left|\left[(\mathbf{G}^H\mathbf{G})^{-1}\mathbf{G}^H\mathbf{Z}\right]_{k,1}\right|^2\right\}} \\ &= \frac{\rho}{\left[(\mathbf{G}^H\mathbf{G})^{-1}\mathbf{G}^H\mathbf{I}_L\mathbf{G}(\mathbf{G}^H\mathbf{G})^{-H}\right]_{k,k}} \\ &= \frac{\rho}{\left[(\mathbf{G}^H\mathbf{G})^{-1}\right]_{k,k}}.\end{aligned} \tag{5.8}$$

Substituting (5.2) into (5.8) can further convert the post-processing SNR of user k to

$$\gamma_k = \frac{\rho}{\left[\left(\mathbf{S}^{\frac{1}{2}}\mathbf{D}^{\frac{1}{2}}\mathbf{H}^H\mathbf{H}\mathbf{D}^{\frac{1}{2}}\mathbf{S}^{\frac{1}{2}}\right)^{-1}\right]_{k,k}} = \frac{\rho\beta_k s_k}{\left[(\mathbf{H}^H\mathbf{H})^{-1}\right]_{k,k}}. \tag{5.9}$$

$\mathbf{H}^H\mathbf{H}$ is a $K \times K$ central complex Wishart matrix with L degrees of freedom [96, 98], so we can obtain $\frac{1}{\left[(\mathbf{H}^H\mathbf{H})^{-1}\right]_{k,k}} \sim \chi^2_{2\psi}$ ($\psi \triangleq \frac{L-K+1}{2}$), where $\chi^2_{2\psi}$ is a Chi-square distribution with 2ψ degrees of freedom [99, 100]. The distribution of the post-processing SNRs is closely related to the number of Rx antennas L and the number of accessed users K. On the one hand, when the gNB installs more Rx antennas, the degrees of freedom of the Chi-square distribution increases, so MU-MIMO can provide higher reliability. On the other hand, when the number of accessed users increases, installing more Rx antennas at the gNB can maintain the degrees of freedom of the Chi-square distribution constant, and then guarantee the reliability of UL MUD.

Further, the post-processing SNR of user k in ZF detection can be expressed as

$$\gamma_k = \beta_k s_k \tau_k, \tag{5.10}$$

where $\tau_k = \frac{\rho}{\left[(\mathbf{H}^H\mathbf{H})^{-1}\right]_{k,k}} \sim \rho\chi^2_{2\psi}$. As a result, the pdf of τ_k is expressed as

$$f_{\tau_k}(x) = \frac{1}{\rho}\frac{(\frac{x}{\rho})^{\psi-1}\exp(-\frac{x}{2\rho})}{2^\psi\Gamma(\psi)} = \frac{x^{\psi-1}\exp(-\frac{x}{2\rho})}{2^\psi\Gamma(\psi)\rho^\psi}, \quad \forall x > 0, \tag{5.11}$$

where $\Gamma(\psi)=\int_0^\infty x^{\psi-1}e^{-x}\mathrm{d}x$.

From (5.6) and (5.11), the pdf of γ_k is given by

$$
\begin{aligned}
&f_{\gamma_k}(x)=\int_0^\infty \frac{1}{y}f_{\tau_k}(\frac{x}{y})f_{\beta_k s_k}(y)\mathrm{d}y\\
=&\int_0^\infty \frac{1}{y}\left(\frac{(\frac{x}{y})^{\psi-1}\exp\left(-\frac{x}{2\rho y}\right)}{2^\psi\Gamma(\psi)\rho^\psi}\right)\int_y^{y\vartheta^{\alpha_\mathrm{P}}}\left[\frac{\sqrt{2}\varphi y^{-\frac{2}{\alpha_\mathrm{P}}-1}z^{\frac{2}{\alpha_\mathrm{P}}-1}}{\sqrt{\pi}\sigma_\mathrm{S}\alpha_\mathrm{P}(\vartheta^2-1)}\exp\left(-\frac{(\varphi\ln z-\mu_\mathrm{S})^2}{2\sigma_\mathrm{S}^2}\right)\right]\mathrm{d}z\mathrm{d}y\\
=&\frac{2e^{2(\sigma_\mathrm{S}^2\alpha_\mathrm{P}^{-2}+\mu_\mathrm{S}\alpha_\mathrm{P}^{-1})}x^{\psi-1}}{\sqrt{\pi}2^\psi\Gamma(\psi)\alpha_\mathrm{P}(\vartheta^2-1)\rho^\psi}\int_0^\infty e^{-\frac{x}{2\rho y}}y^{-\psi-\frac{2}{\alpha_\mathrm{P}}-1}\int_y^{y\vartheta^{\alpha_\mathrm{P}}}\Bigg[\\
&\frac{\varphi}{\sqrt{2}\sigma_\mathrm{S}z}e^{-\frac{(\varphi\ln z-\mu_\mathrm{S})^2}{2\sigma_\mathrm{S}^2}+\frac{2(\varphi\ln z-\mu_\mathrm{S})}{\alpha_\mathrm{P}}-\frac{2\sigma_\mathrm{S}^2}{\alpha_\mathrm{P}^2}}\Bigg]\mathrm{d}z\mathrm{d}y\\
\overset{(a)}{=}&\frac{2e^{2(\sigma_\mathrm{S}^2\alpha_\mathrm{P}^{-2}+\mu_\mathrm{S}\alpha_\mathrm{P}^{-1})}x^{\psi-1}}{\sqrt{\pi}2^\psi\Gamma(\psi)\alpha_\mathrm{P}(\vartheta^2-1)\rho^\psi}\int_0^\infty e^{-\frac{x}{2\rho y}}y^{-\psi-\frac{2}{\alpha_\mathrm{P}}-1}\left[\int_{\frac{\ln y-\mu_\mathrm{S}-2\sigma_\mathrm{S}^2/\alpha_\mathrm{P}}{\sqrt{2}\sigma_\mathrm{S}}}^{\frac{\ln y-\mu_\mathrm{S}-2\sigma_\mathrm{S}^2/\alpha_\mathrm{P}+\alpha_\mathrm{P}\ln\vartheta}{\sqrt{2}\sigma_\mathrm{S}}}e^{-w^2}\mathrm{d}w\right]\mathrm{d}y\\
\overset{(b)}{=}&\frac{2e^{2(\sigma_\mathrm{S}^2\alpha_\mathrm{P}^{-2}+\mu_\mathrm{S}\alpha_\mathrm{P}^{-1})}x^{\psi-1}}{\sqrt{\pi}2^\psi\Gamma(\psi)\alpha_\mathrm{P}(\vartheta^2-1)\rho^\psi}\int_{-\infty}^\infty e^{(\psi+\frac{2}{\alpha_\mathrm{P}})t-\frac{xe^t}{2\rho}}\left[\int_{\frac{t+\mu_\mathrm{S}+2\sigma_\mathrm{S}^2/\alpha_\mathrm{P}-\alpha_\mathrm{P}\ln\vartheta}{\sqrt{2}\sigma_\mathrm{S}}}^{\frac{t+\mu_\mathrm{S}+2\sigma_\mathrm{S}^2/\alpha_\mathrm{P}}{\sqrt{2}\sigma_\mathrm{S}}}e^{-w^2}\mathrm{d}w\right]\mathrm{d}t,
\end{aligned}
\tag{5.12}
$$

Considering that $\mathrm{erf}(x)=\frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2}\mathrm{d}t$, (5.12) can be further converted to

$$
\begin{aligned}
f_{\gamma_k}(x)=&\frac{\exp\left(2\sigma_\mathrm{S}^2\alpha_\mathrm{P}^{-2}+2\mu_\mathrm{S}\alpha_\mathrm{P}^{-1}\right)x^{\psi-1}}{2^\psi\Gamma(\psi)\alpha_\mathrm{P}(\vartheta^2-1)\rho^\psi}\int_{-\infty}^\infty\exp\left((\psi+2\alpha_\mathrm{P}^{-1})t-\frac{xe^t}{2\rho}\right)\\
&\left[\mathrm{erf}\left(\frac{t+\nu_\mathrm{S}}{\sqrt{2}\sigma_\mathrm{S}}\right)-\mathrm{erf}\left(\frac{t+\nu_\mathrm{S}-\alpha_\mathrm{P}\ln\vartheta}{\sqrt{2}\sigma_\mathrm{S}}\right)\right]\mathrm{d}t,
\end{aligned}
\tag{5.13}
$$

where $\nu_\mathrm{S}=\mu_\mathrm{S}+2\sigma_\mathrm{S}^2\alpha_\mathrm{P}^{-1}$.

5.2.3 ZF Detection Under Imperfect CSI

In actual deployment, it is challenging to obtain UL perfect CSI at the gNB within limited latency. This section will study the performance of ZF detection under imperfect CSI after PACE estimates SSF channel coefficients.

In PACE, each user's UL signal consists of pilots and data. Let $\mathbf{X}^{(\mathrm{p})}\in\mathbb{C}^{K\times m}$ and $\mathbf{X}^{(\mathrm{d})}\in\mathbb{C}^{K\times M}$ represent the pilots and data transmitted from users, respectively. In this way, UL signals can be expressed as $\mathbf{X}=\left[\mathbf{X}^{(\mathrm{p})},\ \mathbf{X}^{(\mathrm{d})}\right]$, and the received signals

at the gNB can be rewritten as

$$\left[\mathbf{Y}^{(\mathrm{p})}, \mathbf{Y}^{(\mathrm{d})}\right] = \sqrt{\rho}\mathbf{G}\left[\mathbf{X}^{(\mathrm{p})}, \mathbf{X}^{(\mathrm{d})}\right] + \left[\mathbf{Z}^{(\mathrm{p})}, \mathbf{Z}^{(\mathrm{d})}\right], \tag{5.14}$$

where $\mathbf{Y}^{(\mathrm{p})} \in \mathbb{C}^{L\times m}$ and $\mathbf{Y}^{(\mathrm{d})} \in \mathbb{C}^{L\times M}$ represent the received signals containing pilots and the received signals containing data, respectively; $\mathbf{Z}^{(\mathrm{p})} \in \mathbb{C}^{L\times m}$ and $\mathbf{Z}^{(\mathrm{d})} \in \mathbb{C}^{L\times M}$ have i.i.d. $\mathcal{CN}(0,1)$ elements. Typically, users' UL pilots are designed to be orthogonal to each other, and thus $\mathbf{X}^{(\mathrm{p})}\left(\mathbf{X}^{(\mathrm{p})}\right)^H = m\mathbf{I}_K$ could be ensured. The elements in $\mathbf{X}^{(\mathrm{d})}$ are uncorrelated, so $\mathbb{E}\left(|[\mathbf{X}^{(\mathrm{d})}]_{k,l}|^2\right) = 1$ ($1 \le k \le K$ and $1 \le l \le M$), $\mathbb{E}\left([\mathbf{X}^{(\mathrm{d})}]_{k,l}[\mathbf{X}^{(\mathrm{d})}]_{i,j}\right) = 0$ ($i \ne k$ or $j \ne l$).

The received signals including pilots at the gNB can be given by

$$\mathbf{Y}^{(\mathrm{p})} = \sqrt{\rho}\mathbf{G}\mathbf{X}^{(\mathrm{p})} + \mathbf{Z}^{(\mathrm{p})}. \tag{5.15}$$

At this time, the least squares estimation of the channel matrix $\mathbf{G}$ can be expressed as [93, 94]

$$\begin{aligned}\hat{\mathbf{G}} &= \frac{1}{\sqrt{\rho}m}\mathbf{Y}^{(\mathrm{p})}\left(\mathbf{X}^{(\mathrm{p})}\right)^H \\ &= \frac{1}{m}\mathbf{G}\mathbf{X}^{(\mathrm{p})}\left(\mathbf{X}^{(\mathrm{p})}\right)^H + \frac{1}{\sqrt{\rho}m}\mathbf{Z}^{(\mathrm{p})}\left(\mathbf{X}^{(\mathrm{p})}\right)^H \\ &= \mathbf{G} + \mathbf{W},\end{aligned} \tag{5.16}$$

where $\mathbf{W} \triangleq \frac{1}{\sqrt{\rho}m}\mathbf{Z}^{(\mathrm{p})}\left(\mathbf{X}^{(\mathrm{p})}\right)^H$ and has i.i.d. $\mathcal{CN}(0, \frac{1}{\rho m})$ elements that denote the channel estimation errors. In order to ensure accurate channel estimation for user k, $\frac{1}{\rho m} \ll \beta_k s_k$ can be assumed to be true [101, 102]. In PACE, the variance of the elements in $\mathbf{W}$ is inversely proportional to transmit SNR ρ and the LoP m. Thus, it can be available to obtain more accurate channel estimation with higher transmit SNR or more pilots. On the contrary, channel estimation errors are zero under the assumption of perfect CSI, so all the elements in $\mathbf{W}$ are equal to 0.

Similarly, the received signals containing data can be expressed as

$$\mathbf{Y}^{(\mathrm{d})} = \sqrt{\rho}\mathbf{G}\mathbf{X}^{(\mathrm{d})} + \mathbf{Z}^{(\mathrm{d})} = \sqrt{\rho}\hat{\mathbf{G}}\mathbf{X}^{(\mathrm{d})} - \sqrt{\rho}\mathbf{W}\mathbf{X}^{(\mathrm{d})} + \mathbf{Z}^{(\mathrm{d})}. \tag{5.17}$$

When ZF detection is applied at the gNB, the detection vector is given by $\hat{\mathbf{V}}^{\mathrm{ZF}} = (\hat{\mathbf{G}}^H\hat{\mathbf{G}})^{-1}\hat{\mathbf{G}}^H$ [92], so post-processing signals can be expressed as

$$\begin{aligned}
&\hat{\mathbf{V}}^{\mathrm{ZF}}\mathbf{Y}^{(\mathrm{d})} \\
=&\sqrt{\rho}\left(\hat{\mathbf{G}}^H\hat{\mathbf{G}}\right)^{-1}\hat{\mathbf{G}}^H\hat{\mathbf{G}}\mathbf{X}^{(\mathrm{d})} + \left(\hat{\mathbf{G}}^H\hat{\mathbf{G}}\right)^{-1}\hat{\mathbf{G}}^H\left(-\sqrt{\rho}\mathbf{W}\mathbf{X}^{(\mathrm{d})} + \mathbf{Z}^{(\mathrm{d})}\right) \\
=&\sqrt{\rho}\mathbf{X}^{(\mathrm{d})} + \left(\hat{\mathbf{G}}^H\hat{\mathbf{G}}\right)^{-1}\hat{\mathbf{G}}^H\left(-\sqrt{\rho}\mathbf{W}\mathbf{X}^{(\mathrm{d})} + \mathbf{Z}^{(\mathrm{d})}\right).
\end{aligned} \tag{5.18}$$

The kth row of $\hat{\mathbf{V}}^{\mathrm{ZF}}\mathbf{Y}^{(\mathrm{d})}$ can be utilized to estimate the signal of user k, so the post-processing SNR of user k when using PACE can be expressed as

$$\begin{aligned}
\hat{\gamma}_k(m) =&\frac{\rho}{\mathbb{E}\left\{\left|\left[\left(\hat{\mathbf{G}}^H\hat{\mathbf{G}}\right)^{-1}\hat{\mathbf{G}}^H\left(-\sqrt{\rho}\mathbf{W}\mathbf{X}^{(\mathrm{d})} + \mathbf{Z}^{(\mathrm{d})}\right)\right]_{k,1}\right|^2\right\}} \\
=&\frac{\rho}{\left[\left(\hat{\mathbf{G}}^H\hat{\mathbf{G}}\right)^{-1}\hat{\mathbf{G}}^H\left(\frac{K}{m}+1\right)\mathbf{I}_L\hat{\mathbf{G}}\left(\hat{\mathbf{G}}^H\hat{\mathbf{G}}\right)^{-H}\right]_{k,k}} \\
=&\frac{\rho m}{(K+m)\left[\left(\hat{\mathbf{G}}^H\hat{\mathbf{G}}\right)^{-1}\right]_{k,k}}.
\end{aligned} \tag{5.19}$$

It can be seen that $\hat{\gamma}_k(m)$ can be determined by the number of accessed users K and the LoP m. Thus, the post-processing SNR of user k can be increased by optimizing the LoP, thereby improving the reliability of UL detection.

To derive the pdf of $\hat{\gamma}_k(m)$ concisely, the estimation of the channel matrix can be decomposed as follows

$$\hat{\mathbf{G}} = \mathbf{G} + \mathbf{W} = \mathbf{H}\mathbf{D}^{\frac{1}{2}}\mathbf{S}^{\frac{1}{2}} + \mathbf{W} = \breve{\mathbf{H}}\breve{\mathbf{D}}^{\frac{1}{2}}, \tag{5.20}$$

where $\breve{\mathbf{H}} = \left(\mathbf{H}\mathbf{D}^{\frac{1}{2}}\mathbf{S}^{\frac{1}{2}} + \mathbf{W}\right)(\mathbf{D}\mathbf{S} + \frac{1}{\rho m}\mathbf{I}_K)^{-\frac{1}{2}}$ has i.i.d. $\mathcal{CN}(0,1)$ elements, and $\breve{\mathbf{D}}^{\frac{1}{2}} = (\mathbf{D}\mathbf{S} + \frac{1}{\rho m}\mathbf{I}_K)^{\frac{1}{2}} = \mathrm{diag}(\sqrt{\beta_1 s_1 + \frac{1}{\rho m}}, ..., \sqrt{\beta_K s_K + \frac{1}{\rho m}})$. In this sense, (5.19) can be transformed to

$$\hat{\gamma}_k(m) = \frac{\rho m}{(K+m)\left[\left(\breve{\mathbf{D}}^{\frac{1}{2}}\breve{\mathbf{H}}^H\breve{\mathbf{H}}\breve{\mathbf{D}}^{\frac{1}{2}}\right)^{-1}\right]_{k,k}} = \beta_k s_k \hat{\tau}_k(m), \tag{5.21}$$

where $\hat{\tau}_k(m) = \frac{\varsigma(m)}{2\left[\left(\breve{\mathbf{H}}^H\breve{\mathbf{H}}\right)^{-1}\right]_{k,k}}$, and $\varsigma(m) \triangleq \frac{2(\rho m\beta_k s_k+1)}{(K+m)\beta_k s_k}$. As mentioned above, when $\frac{1}{\rho m} \ll \beta_k s_k$ is true in PACE, $1 \ll \rho m\beta_k s_k$ can be obtained, and thus we have $\varsigma(m) \cong \frac{2\rho m}{K+m}$. Known from literature [96, 98], we can have $\frac{1}{\left[\left(\breve{\mathbf{H}}^H\breve{\mathbf{H}}\right)^{-1}\right]_{k,k}} \sim \chi^2_{2\psi}$. Note that $\chi^2_k = \Gamma(\frac{k}{2}, 2)$ and $a\Gamma(k,\theta) = \Gamma(k, a\theta)$ $(a > 0)$ hold, in which $\Gamma(k,\theta)$ denotes a gamma distribution with a shape parameter k and a scale parameter θ, we can obtain $\hat{\tau}_k(m) \sim \Gamma(\psi, \varsigma(m))$. Since $\psi = \frac{L-K+1}{2}$, when the number of Rx antennas rises, $\hat{\gamma}_k(m)$ monotonically increases. Besides, when the number of accessed users increases, $\hat{\gamma}_k(m)$ monotonically decreases. Further, the pdf of $\hat{\tau}_k(m)$ can be expressed as

$$f_{\hat{\tau}_k(m)}(x) = \frac{x^{\psi-1}\exp(-\frac{x}{\varsigma(m)})}{\Gamma(\psi)\varsigma(m)^\psi}, \quad \forall x > 0. \tag{5.22}$$

Similar to (5.12), the pdf of $\hat{\gamma}_k(m)$ can be derived as follows,

$$\begin{aligned}
&f_{\hat{\gamma}_k(m)}(x) = \int_0^\infty \frac{1}{y} f_{\hat{\tau}_k(m)}(\frac{x}{y}) f_{\beta_k s_k}(y)\mathrm{d}y \\
&= \int_0^\infty \frac{1}{y}\left(\frac{(\frac{x}{y})^{\psi-1}\exp\left(-\frac{x}{\varsigma(m)y}\right)}{\Gamma(\psi)\varsigma(m)^\psi}\right)\int_y^{y\vartheta^{\alpha_\mathrm{P}}}\left[\frac{\sqrt{2}\varphi y^{-\frac{2}{\alpha_\mathrm{P}}-1}z^{\frac{2}{\alpha_\mathrm{P}}-1}}{\sqrt{\pi}\sigma_\mathrm{S}\alpha_\mathrm{P}(\vartheta^2-1)}e^{-\frac{(\varphi\ln z-\mu_\mathrm{S})^2}{2\sigma_\mathrm{S}^2}}\right]\mathrm{d}z\mathrm{d}y \\
&\overset{(a)}{=} \frac{2e^{2(\sigma_\mathrm{S}^2\alpha_\mathrm{P}^{-2}+\mu_\mathrm{S}\alpha_\mathrm{P}^{-1})}x^{\psi-1}}{\sqrt{\pi}\Gamma(\psi)\alpha_\mathrm{P}(\vartheta^2-1)\varsigma(m)^\psi}\int_0^\infty e^{-\frac{x}{\varsigma(m)y}}y^{-\psi-\frac{2}{\alpha_\mathrm{P}}-1}\left[\int_{\frac{\ln y-\mu_\mathrm{S}-2\sigma_\mathrm{S}^2/\alpha_\mathrm{P}}{\sqrt{2}\sigma_\mathrm{S}}}^{\frac{\ln y-\mu_\mathrm{S}-2\sigma_\mathrm{S}^2/\alpha_\mathrm{P}+\alpha_\mathrm{P}\ln\vartheta}{\sqrt{2}\sigma_\mathrm{S}}}e^{-w^2}\mathrm{d}w\right]\mathrm{d}y \\
&\overset{(b)}{=} \frac{\exp\left(2\sigma_\mathrm{S}^2\alpha_\mathrm{P}^{-2}+2\mu_\mathrm{S}\alpha_\mathrm{P}^{-1}\right)x^{\psi-1}}{\Gamma(\psi)\alpha_\mathrm{P}(\vartheta^2-1)\varsigma(m)^\psi}\int_{-\infty}^\infty\left\{\exp\left((\psi+\frac{2}{\alpha_\mathrm{P}})t-\frac{xe^t}{\varsigma(m)}\right)\left[\right.\right. \\
&\left.\left.\mathrm{erf}\left(\frac{t+\mu_\mathrm{S}+2\sigma_\mathrm{S}^2/\alpha_\mathrm{P}}{\sqrt{2}\sigma_\mathrm{S}}\right)-\mathrm{erf}\left(\frac{t+\mu_\mathrm{S}+2\sigma_\mathrm{S}^2/\alpha_\mathrm{P}-\alpha_\mathrm{P}\ln\vartheta}{\sqrt{2}\sigma_\mathrm{S}}\right)\right]\right\}\mathrm{d}t \\
&\approx \frac{\exp\left(2\sigma_\mathrm{S}^2\alpha_\mathrm{P}^{-2}+2\mu_\mathrm{S}\alpha_\mathrm{P}^{-1}\right)(K+m)^\psi x^{\psi-1}}{\Gamma(\psi)\alpha_\mathrm{P}(\vartheta^2-1)(2\rho m)^\psi}\int_{-\infty}^\infty\left\{\exp\left[\,(\psi+2\alpha_\mathrm{P}^{-1})t-\frac{(K+m)xe^t}{2\rho m}\right]\right. \\
&\left.\left[\mathrm{erf}\left(\frac{t+\nu_\mathrm{S}}{\sqrt{2}\sigma_\mathrm{S}}\right)-\mathrm{erf}\left(\frac{t+\nu_\mathrm{S}-\alpha_\mathrm{P}\ln\vartheta}{\sqrt{2}\sigma_\mathrm{S}}\right)\right]\right\}\mathrm{d}t.
\end{aligned} \tag{5.23}$$

The equality (a) can be established by applying the substitution method of $w \triangleq \frac{\varphi\ln z-\mu_\mathrm{S}-2{\sigma_\mathrm{S}}^2\alpha_\mathrm{P}^{-1}}{\sqrt{2}\sigma_\mathrm{S}}$. The equality (b) can be established by letting $t \triangleq -\ln y$. Note that $\psi = \frac{L-K+1}{2}$, the pdf of the post-processing SNRs is decided by the LoP, the

number of accessed users, and the number of Rx antennas.

5.3 Analysis of Error Probability

5.3.1 Error Probability in Short-packet Transmission

From the aspect of an information theory, the data rate can be expressed in bits per channel use (bpcu). In MU-MIMO, the number of CUs occupied by each user is equal to the channel code blocklength M. To this end, the data rate can be expressed as $R = D/M$ bpcu. The maximum achievable rate R in the deterministic channel can be approximated by (4.11).

When information bits are transmitted with a given blocklength M, the error probability can be approximated by the following equation,

$$\varepsilon(M, D, \gamma) = Q\left(\sqrt{\frac{M}{V(\gamma)}}\left(C(\gamma) - D/M\right)\right). \tag{5.24}$$

5.3.2 Error Probability Under Perfect CSI

The post-processing SNR γ_k varies with the users' random deployment and the experienced shadow fading. Therefore, the error probability of user k in the fading channel under a given transmission latency t_{DE} can be given by

$$\begin{aligned}\epsilon_k(t_{\mathrm{DE}}) &= \int_0^{\infty} \varepsilon(Bt_{\mathrm{DE}}, D, x) f_{\gamma_k}(x)\mathrm{d}x \\ &= \int_0^{\infty} Q\left(\sqrt{\frac{Bt_{\mathrm{DE}}}{V(x)}}\left(C(x) - \frac{D}{Bt_{\mathrm{DE}}}\right)\right) f_{\gamma_k}(x)\mathrm{d}x.\end{aligned} \tag{5.25}$$

Since $Q(w)$ is a monotonically decreasing function of w and $f_{\gamma_k}(x)$ is independent of t_{DE}, implementing a greater number of CUs can effectively reduce the error probability. However, due to the strict constraint on latency t_{DE}, it is challenging to individually allocate a wide band to each user in IoT applications. In this way, the average number of CUs allocated per user is small. To satisfy the strict requirement of reliability, all the available CUs can be reused among users by superposed transmission, and then the channel code blocklength can be large enough.

5.3.3 Error Probability Under Imperfect CSI

When the LoP is m, the error probability of user k can be expressed by

$$\begin{aligned}\hat{\epsilon}_k(m, t_{\mathrm{DE}}) &= \int_0^{\infty} \varepsilon(Bt_{\mathrm{DE}} - m, D, x) f_{\hat{\gamma}_k(m)}(x) \mathrm{d}x \\ &= \int_0^{\infty} Q\left(\sqrt{\frac{Bt_{\mathrm{DE}} - m}{V(x)}}\left(C(x) - \frac{D}{Bt_{\mathrm{DE}} - m}\right)\right) f_{\hat{\gamma}_k(m)}(x) \mathrm{d}x,\end{aligned} \tag{5.26}$$

for $K \leq m \leq Bt_{\mathrm{DE}} - 1$.

On the one hand, if the LoP increases, $\varsigma(m) \cong \frac{2\rho m}{K+m}$ becomes larger, and users are more likely to obtain higher post-processing SNRs to improve reliability. On the other hand, when the number of CUs occupied by pilots m increases, the number of CUs carrying information bits $Bt_{\mathrm{DE}} - m$ decreases, resulting in an increment in the channel code rate and a decrement in reliability. Thus, the error probability in short-packet transmission can be reduced with the optimized LoP.

5.3.4 Optimizing the LoP

As can be seen from the previous analysis, the error probability is closely related to the LoP m, which needs to be optimized. To minimize the error probability of user k, the optimal LoP can be expressed as

$$m^*(t_{\mathrm{DE}}) = \arg \min_{K \leq m \leq Bt_{\mathrm{DE}}-1} \hat{\epsilon}_k(m, t_{\mathrm{DE}}). \tag{5.27}$$

At this time, the minimum error probability of user k can be expressed as

$$\hat{\epsilon}_k^*(t_{\mathrm{DE}}) = \hat{\epsilon}_k(m^*(t_{\mathrm{DE}}), t_{\mathrm{DE}}). \tag{5.28}$$

To determine the optimal LoP $m^*(t_{\mathrm{DE}})$, a one-dimensional numerical search algorithms can be applied. The exhaustive search method can be employed as a baseline. Meanwhile, GSSM can be utilized as an effective scheme to determine the approximate optimal LoP, and it can rapidly converge and avoid calculation of the derivatives of the $\hat{\epsilon}_k(m, t_{\mathrm{DE}})$, which is extremely intricate. GSSM can search

Algorithm 2 Implementation of GSSM

Initialize $m_{\mathrm{a}} = K$, $m_{\mathrm{b}} = Bt_{\mathrm{DE}} - 1$, $M_{\mathrm{tolerance}} = 0.5$, $m_1 = m_{\mathrm{b}} - 0.618(m_{\mathrm{b}} - m_{\mathrm{a}})$, $m_2 = m_{\mathrm{a}} + 0.618(m_{\mathrm{b}} - m_{\mathrm{a}})$
For the LoP is $m = m_1$ and $m = m_2$, calculate the error probability $\varepsilon_1 = \hat{\epsilon}_k(m_1, t_{\mathrm{DE}})$ and $\varepsilon_2 = \hat{\epsilon}_k(m_2, t_{\mathrm{DE}})$, respectively
repeat
  if $(\varepsilon_1 > \varepsilon_2)$ **then**
    Re-apply the parameters for the next iteration, $m_{\mathrm{a}} = m_1$, $m_1 = m_2$, and $\varepsilon_1 = \varepsilon_2$
    Calculate the parameters for the next iteration, $m_2 = m_{\mathrm{a}} + 0.618(m_{\mathrm{b}} - m_{\mathrm{a}})$
    For the LoP is $m = m_2$, calculate the error probability $\varepsilon_2 = \hat{\epsilon}_k(m_2, t_{\mathrm{DE}})$
  else
    Re-apply the parameters for the next iteration, $m_{\mathrm{b}} = m_2$, $m_2 = m_1$, and $\varepsilon_2 = \varepsilon_1$
    Calculate the parameters for the next iteration, $m_1 = m_{\mathrm{b}} - 0.618 * (m_{\mathrm{b}} - m_{\mathrm{a}})$
    For the LoP is $m = m_1$, calculate the error probability $\varepsilon_1 = \hat{\epsilon}_k(m_1, t_{\mathrm{DE}})$
  end if
until $|m_{\mathrm{b}} - m_{\mathrm{a}}| \leq M_{\mathrm{tolerance}}$
if $(\hat{\epsilon}_k(\lfloor \frac{m_{\mathrm{b}}+m_{\mathrm{a}}}{2} \rfloor, t_{\mathrm{DE}}) > \hat{\epsilon}_k(\lceil \frac{m_{\mathrm{b}}+m_{\mathrm{a}}}{2} \rceil, t_{\mathrm{DE}}))$ **then**
  $m^*(t_{\mathrm{DE}}) = \lceil \frac{m_{\mathrm{b}}+m_{\mathrm{a}}}{2} \rceil$
else
  $m^*(t_{\mathrm{DE}}) = \lfloor \frac{m_{\mathrm{b}}+m_{\mathrm{a}}}{2} \rfloor$
end if
Output determined optimal LoP $m^*(t_{\mathrm{DE}})$.

for the minimum point of a unimodal function by iteratively narrowing the search scope, which has a fixed narrowing ratio $\frac{\sqrt{5}+1}{2} \approx 0.618$ [103, 104]. The objective function is set to $\hat{\epsilon}_k(m, t_{\mathrm{DE}})$ and the search scope is $m \in [m_{\mathrm{a}}, m_{\mathrm{b}}]$, where $m_{\mathrm{a}} = K$, $m_{\mathrm{b}} = Bt_{\mathrm{DE}} - 1$, and the unique minimum point in the range $m^*(t_{\mathrm{DE}})$ is an integer. With the narrowing ratio $\frac{\sqrt{5}+1}{2} \approx 0.618$, the selected points within range $[m_{\mathrm{a}}, m_{\mathrm{b}}]$ are given by $m_1 = m_{\mathrm{b}} - 0.618(m_{\mathrm{b}} - m_{\mathrm{a}})$ and $m_2 = m_{\mathrm{a}} + 0.618(m_{\mathrm{b}} - m_{\mathrm{a}})$. When $\hat{\epsilon}_k(m_1, t_{\mathrm{DE}}) > \hat{\epsilon}_k(m_2, t_{\mathrm{DE}})$, $m^*(t_{\mathrm{DE}})$ locates in the range $[m_1, m_{\mathrm{b}}]$, and we update the search scope by allowing $m_{\mathrm{a}} = m_1$ and update the selected points by allowing $m_1^{'} = m_2$ and $m_2 = m_1 + 0.618(m_{\mathrm{b}} - m_1)$. On the contrary, when $\hat{\epsilon}_k(m_1, t_{\mathrm{DE}}) \leq \hat{\epsilon}_k(m_2, t_{\mathrm{DE}})$, $m^*(t_{\mathrm{DE}})$ locates in the range $[m_{\mathrm{a}}, m_2]$ and we update the search scope by allowing $m_{\mathrm{b}} = m_2$, and update the selected points by allowing $m_1^{'} = m_2 - 0.618(m_2 - m_{\mathrm{a}})$ and $m_2^{'} = m_1$. Considering $m^*(t_{\mathrm{DE}})$ is an integer, GSSM terminates after a few iterations. When the size of the search scope $m_{\mathrm{b}} - m_{\mathrm{a}}$ is

Table 5.1 : MU-MIMO Simulation Parameters

Parameter	**Value**
Bandwidth (kHz)	720
Cell range (m)	90
Minimum distance (m)	30
Packet size (bit)	30
Noise power spectral density (dBm/Hz)	-170
Tx power (dBm)	-50 $\sim$ 30
Constant path loss μ_{S} (dB)	-10
Path loss exponent	3

less than 0.5, we can directly select one neighbor integer of $\frac{m_{\mathrm{b}}+m_{\mathrm{a}}}{2}$ with minimum $\hat{\epsilon}_k(x, t_{\mathrm{DE}})$ as the near-optimal point [104, 105]. Furthermore, the implementation of GSSM is given in Algorithm 2. As verified in numerical results, it is possible to discover near-optimal $m^*(t_{\mathrm{DE}})$ by applying GSSM with low computational complexity and a limited number of iterations.

5.4 Numerical Results and Analysis

This section provides theoretical calculations and simulation results to verify the previous theoretical derivation. Users obey a uniform distribution in the cell ranging from the minimum distance to the cell radius. They experience shadow fading and Rayleigh fading, and the path loss exponent is set to 3. The channel coefficients are randomly generated for 1000 times. It is assumed that the transmission occupies 12 adjacent 60 kHz subcarriers, which form a total bandwidth of 720 kHz. To evaluate MU-MIMO conveniently, the number of Rx antennas at the gNB can be assumed to grow with the number of accessed users to maintain reliability at a consistent level. The parameters of the simulation are listed in Table 5.1.

5.4.1 Optimal LoP

Figs. 5.2-5.3 illustrate the error probability decided by different lengths of pilots, when the standard deviation of shadow fading σ_{S} is 4 dB and 6 dB, respectively. The reliability of detection is highly dependent on the data rate of users and the accuracy of estimated CSI at the receiver. Increasing the LoP improves the accuracy

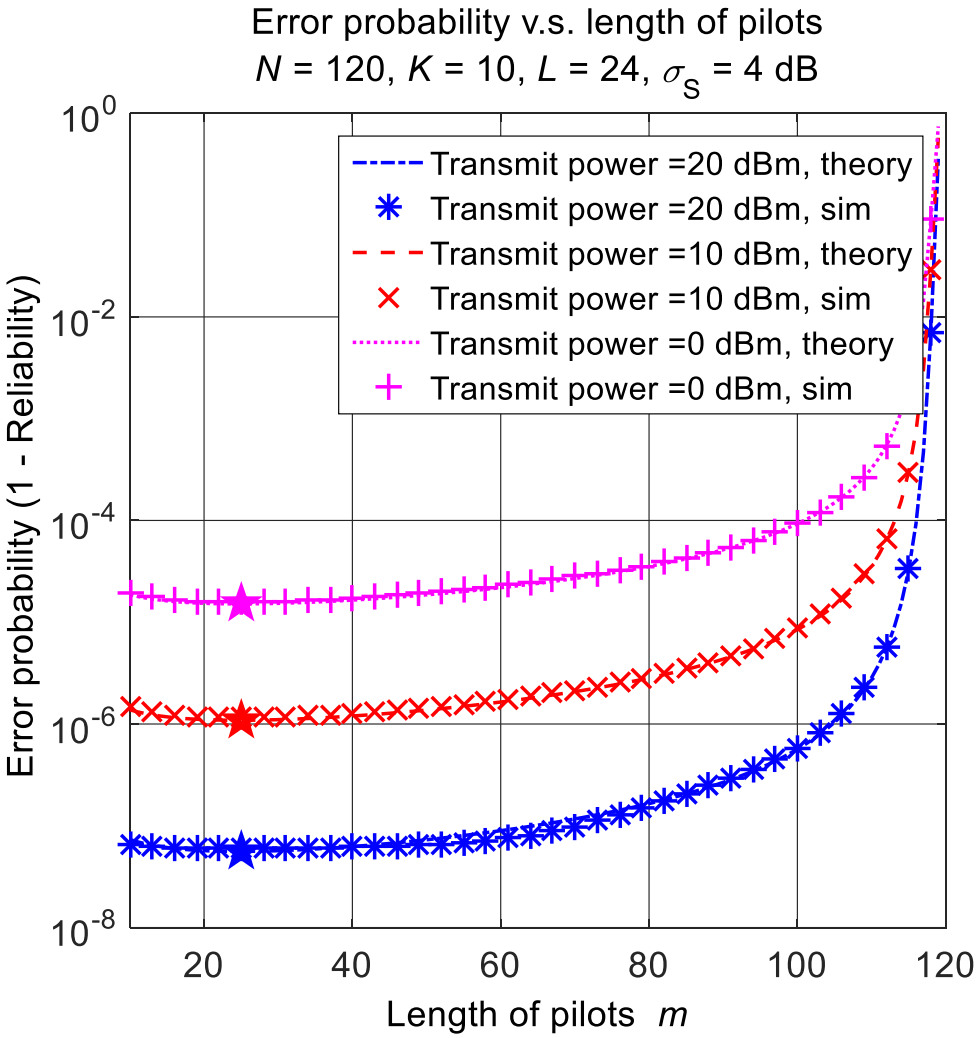

Figure 5.2 : Error probability v.s. LoP: $N = 120$, $\sigma_{\mathrm{S}} = 4$ dB.

of estimated CSI obtained by PACE, but decreases the number of CUs allocated to data and might reduce reliability at the same post-processing SNR. Thus, the error probability might increase when the LoP increases in PACE. The theoretical results comply with the simulation results, and the minimum error probability can be reduced to the level of 10^{-5} with the optimized LoP. With larger Tx power, the error probability can be significantly reduced in MU-MIMO. Furthermore, when severe shadow fading ($\sigma_{\mathrm{S}} = 6$ dB) is encountered, it is necessary to obtain a greater number of CUs (higher latency) and higher Tx power (maximum 30 dBm is sufficient) to

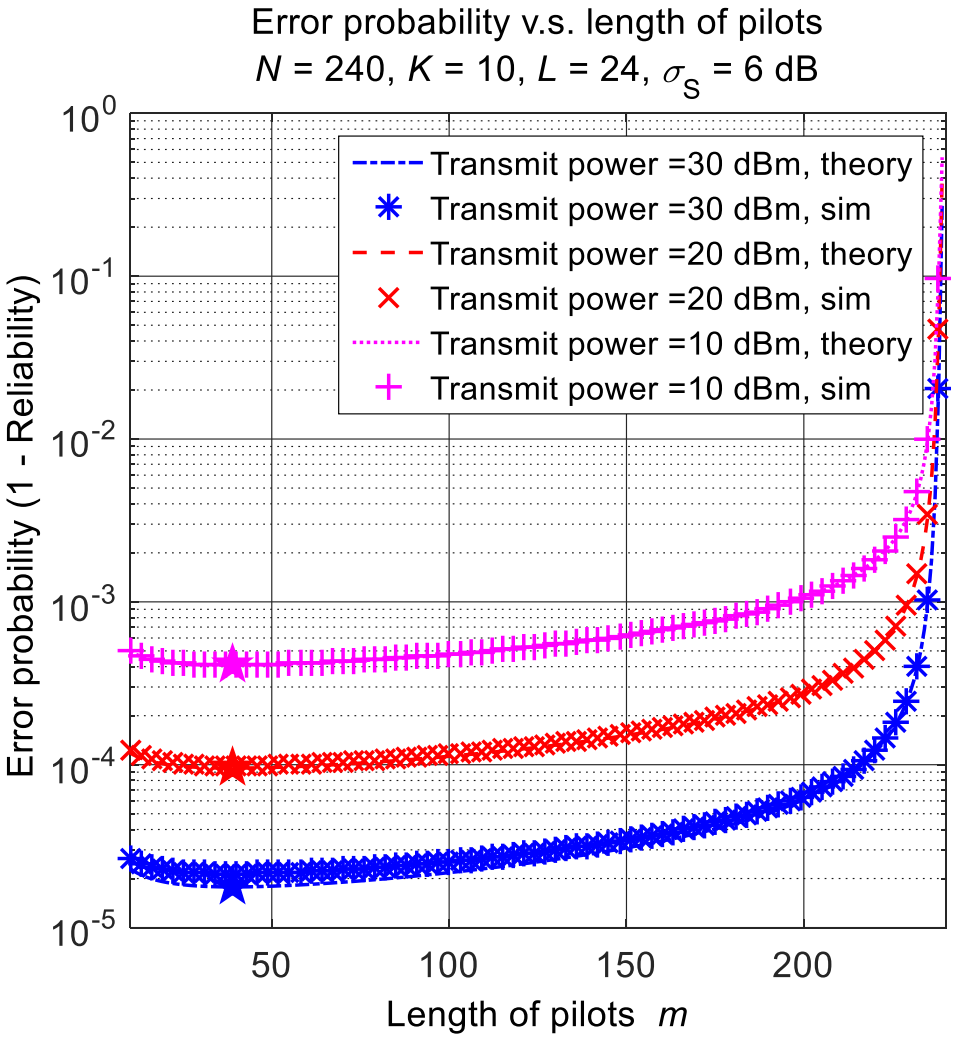

Figure 5.3 : Error probability v.s. LoP: $N = 240$, $\sigma_{\mathrm{S}} = 6$ dB.

ensure reliability.

5.4.2 Overhead of Pilots

Fig. 5.4 suggests the overhead of pilots when applying GSSM to determine the LoP. The LoP determined by GSSM is highly consistent with the optimal LoP, which verifies the convergence and validity of GSSM. Furthermore, the fixed-ratio pilot allocation can be employed as a baseline, and it allocates $\max(K, 0.2N)$ CUs to pilots. As can be seen from Fig. 5.4, GSSM can reduce the overhead of pilots to

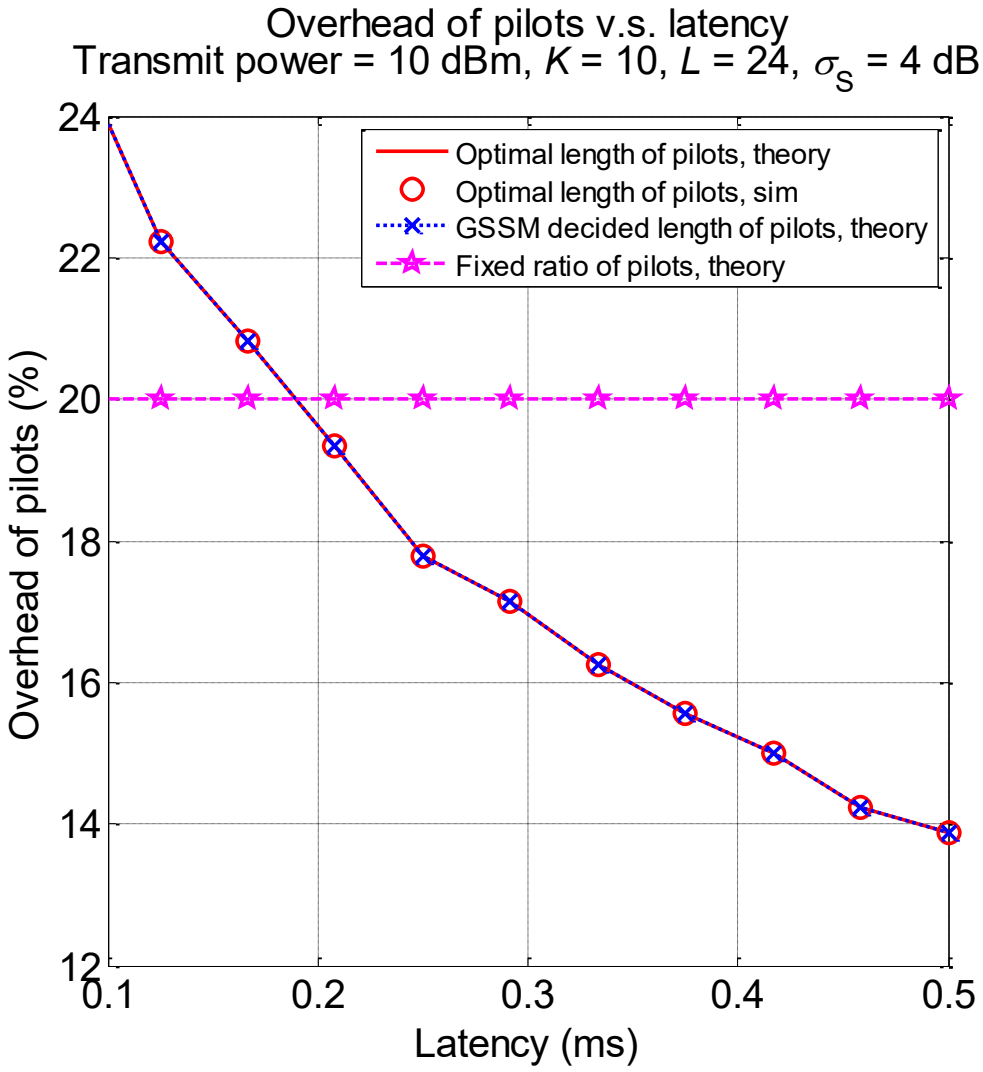

Figure 5.4 : Relationship between latency and overhead of pilots.

less than 15% while minimizing the error probability.

5.4.3 Trade-off Between Reliability and Latency

Fig. 5.5 demonstrates the trade-off between reliability and latency, when the number of accessed users $K = 16$ and the number of Rx antennas $L = 32$. Under perfect CSI, the latency should be higher than 0.15 ms to achieve 10^{-5} error probability. Under imperfect CSI, when the latency is 0.22 ms, the error probability can

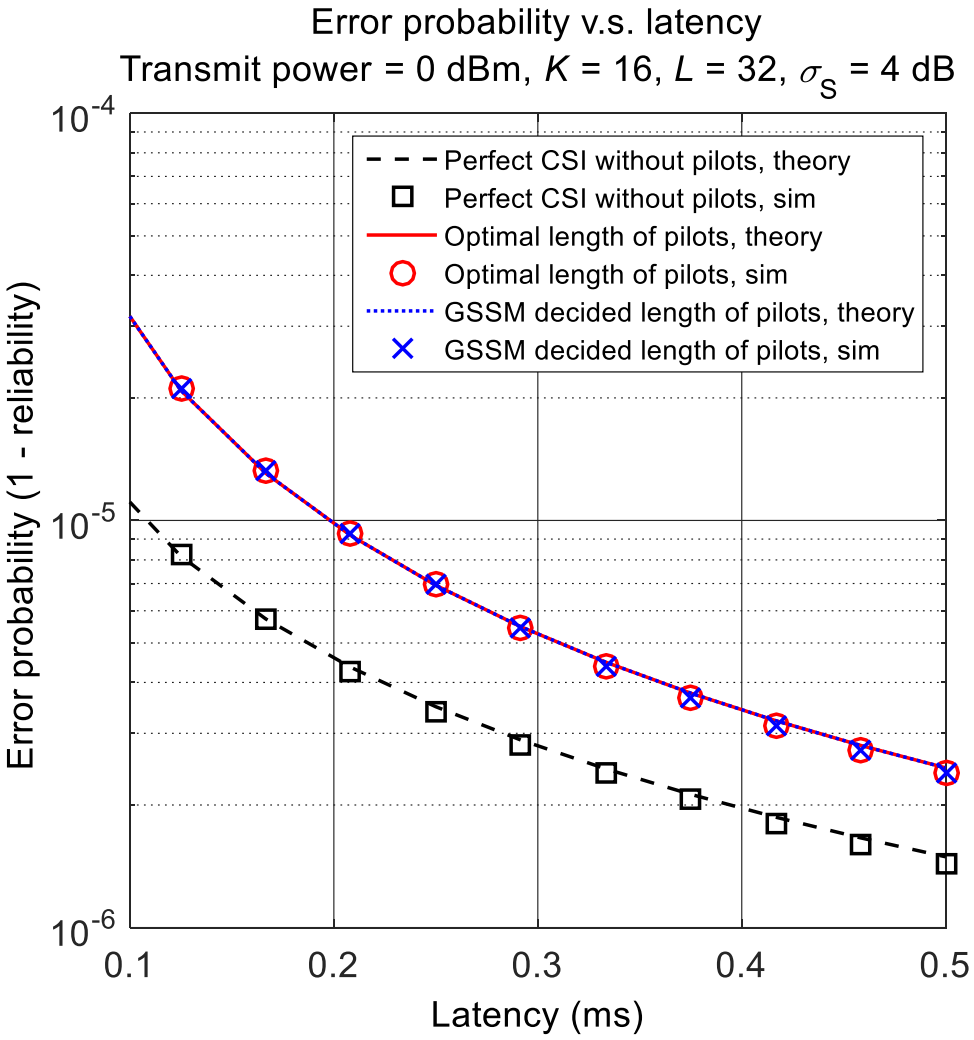

Figure 5.5 : Error probability v.s. latency, when $\sigma_{\mathrm{S}} = 4$ dB.

be lowered to 10^{-5}. Generally, as the latency decreases, the number of CUs that can be utilized to improve channel estimation and bear information bits decreases, and the reliability of MU-MIMO moderately declines. Consequently, the reliability can be compensated by an affordable increment in latency (about 0.1 ms higher).

Fig. 5.6 reveals the trade-off between the reliability and latency for different numbers of Rx antennas and different numbers of accessed users (in the following cases: $K = 10$, $L = 24$; $K = 16$, $L = 32$; $K = 24$, $L = 40$). Under perfect CSI, when the latency is larger than 0.15 ms, the error probability can be lower than 10^{-5}.

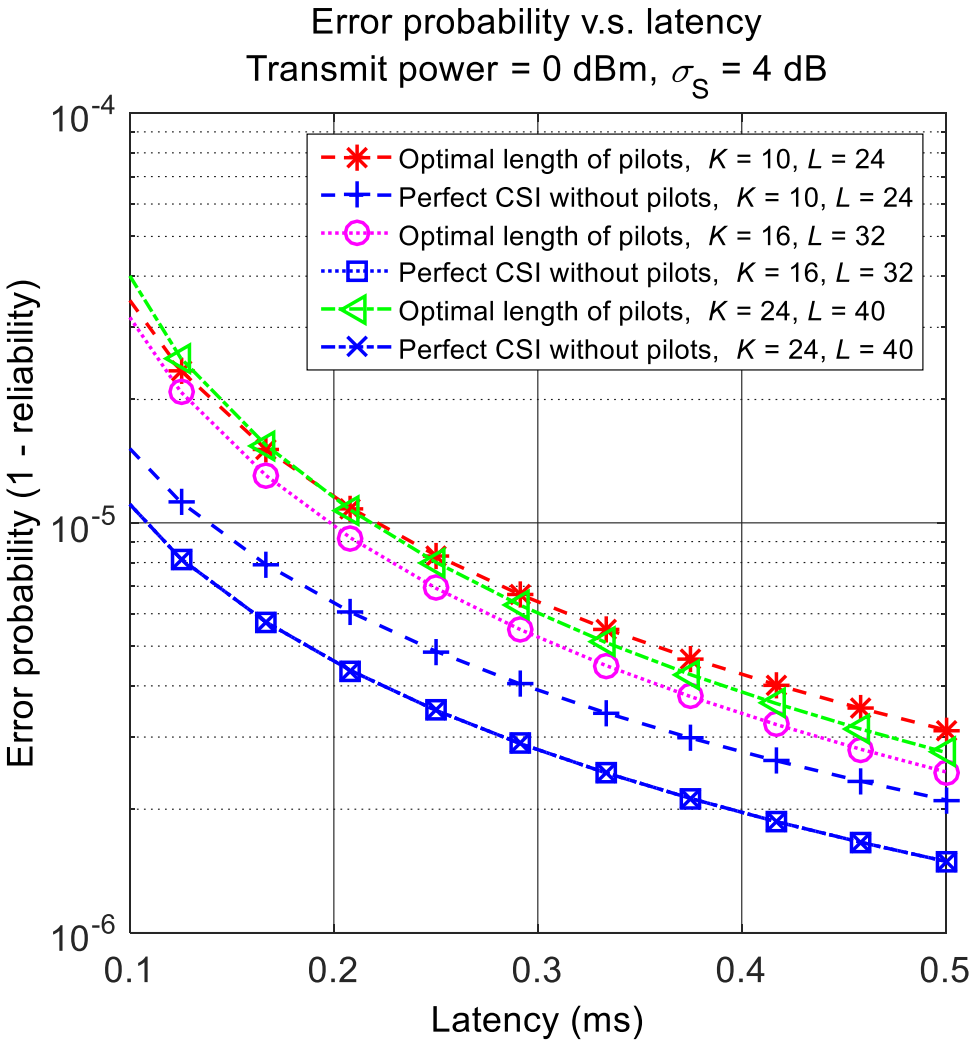

Figure 5.6 : Trade-off between the reliability and latency for different numbers of Rx antennas and different numbers of accessed users.

At the same time, under imperfect CSI, when the latency is no less than 0.22 ms, the error probability reduces to lower than 10^{-5}. In general, as the latency increases, the number of CUs that can be applied to improve channel estimation and carry information bits increases, and the reliability of MU-MIMO NOMA moderately rises.

Fig. 5.7 reveals the trade-off between reliability and latency at different levels of shadow fading. When $\sigma_S = 2$ dB, ultra-high reliability can be achieved with 0 dBm Tx power. When $\sigma_S = 4$ dB, the error probability can be reduced to 10^{-5} by

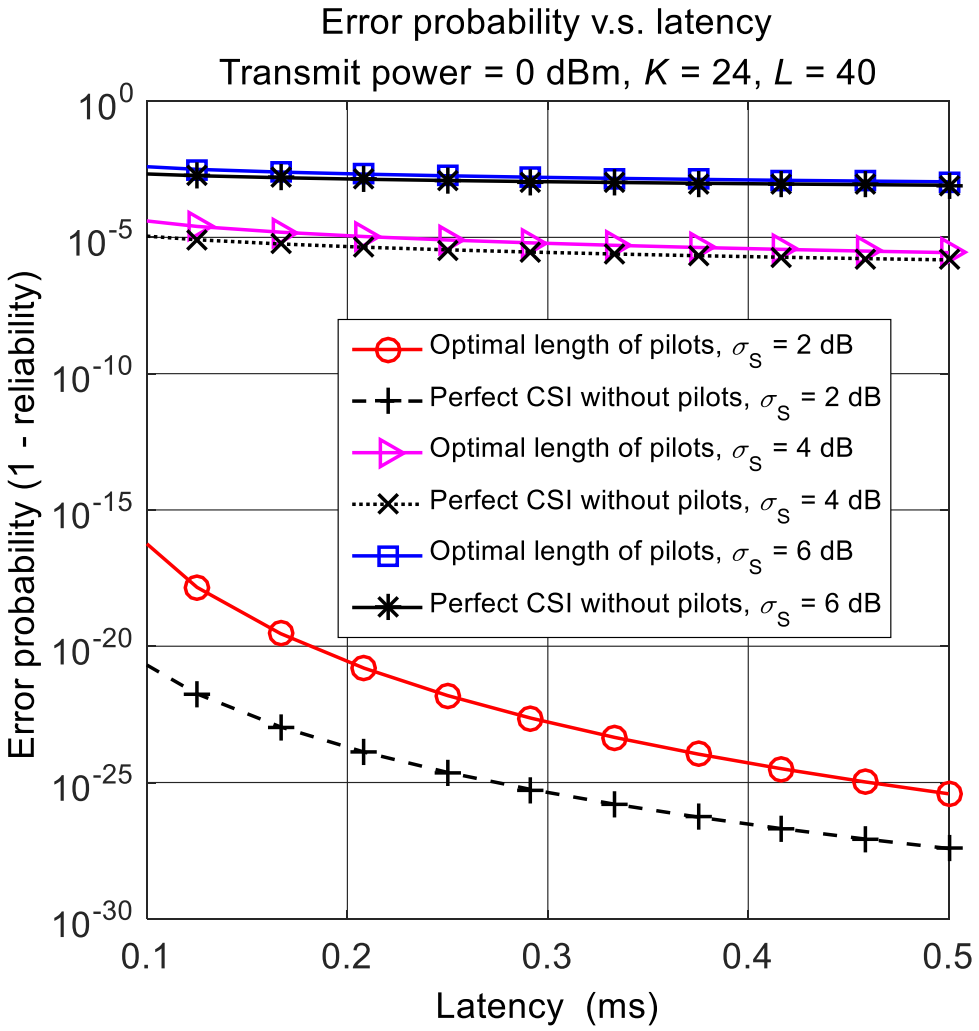

Figure 5.7 : The trade-off between reliability and latency at different levels of shadow fading.

transmitting in 0.5 ms at 0 dBm Tx power.

5.4.4 Relation Between Reliability and Tx Power

As demonstrated in Fig. 5.8, the error probability decreases significantly with increasing Tx power. When Tx power is higher than 0 dBm, the error probability can be lowered to 10^{-5}. Moreover, GSSM can achieve a near-optimal error probability and satisfy the requirement of reliability in MU-MIMO.

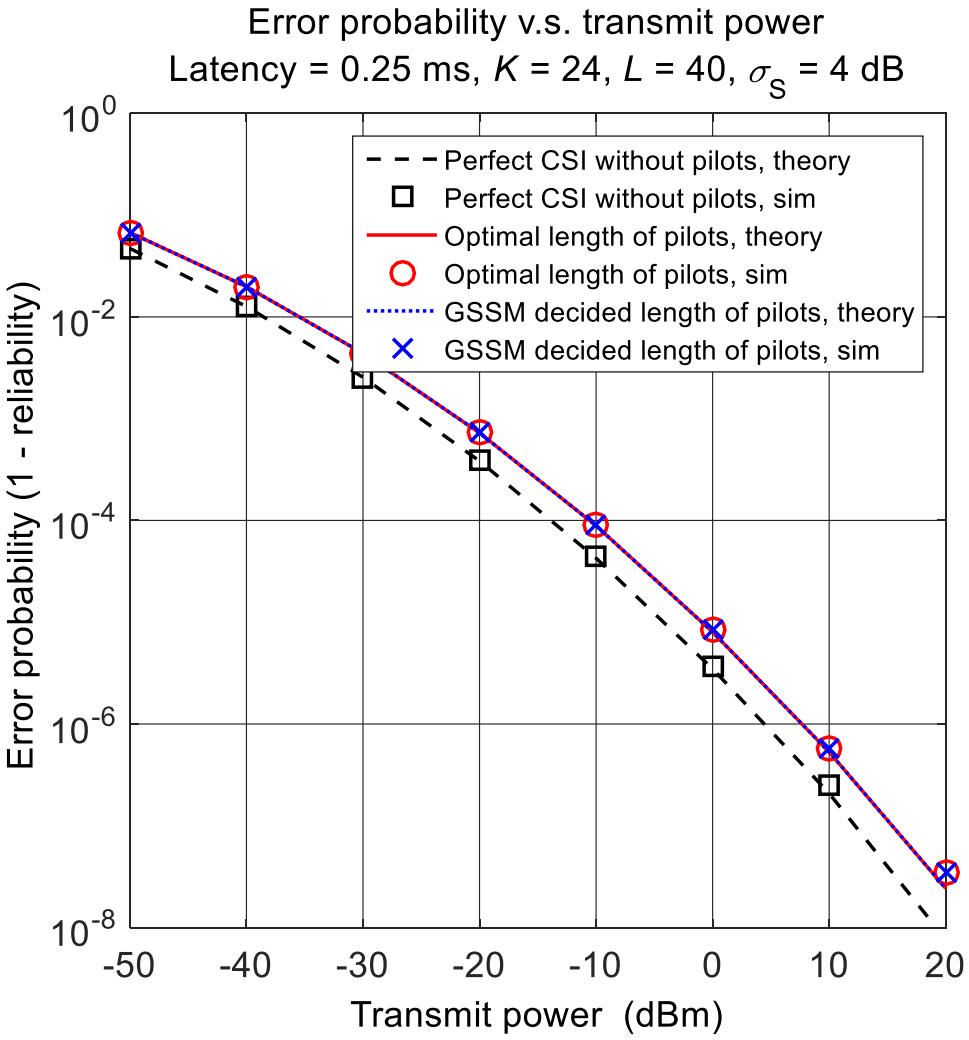

Figure 5.8 : The relationship between reliability and Tx power when using GSSM.

At different levels of shadow fading, Fig. 5.9 provides the performance of MU-MIMO with 0.25 ms latency. It has been verified that ultra-high reliability can be achieved under moderate shadow fading with 0 dBm Tx power. Under severe shadow fading ($\sigma_S = 6$ dB), higher Tx power (more than 20 dBm) or a higher latency (more than 0.25 ms) is required to guarantee 99.999% and above reliability.

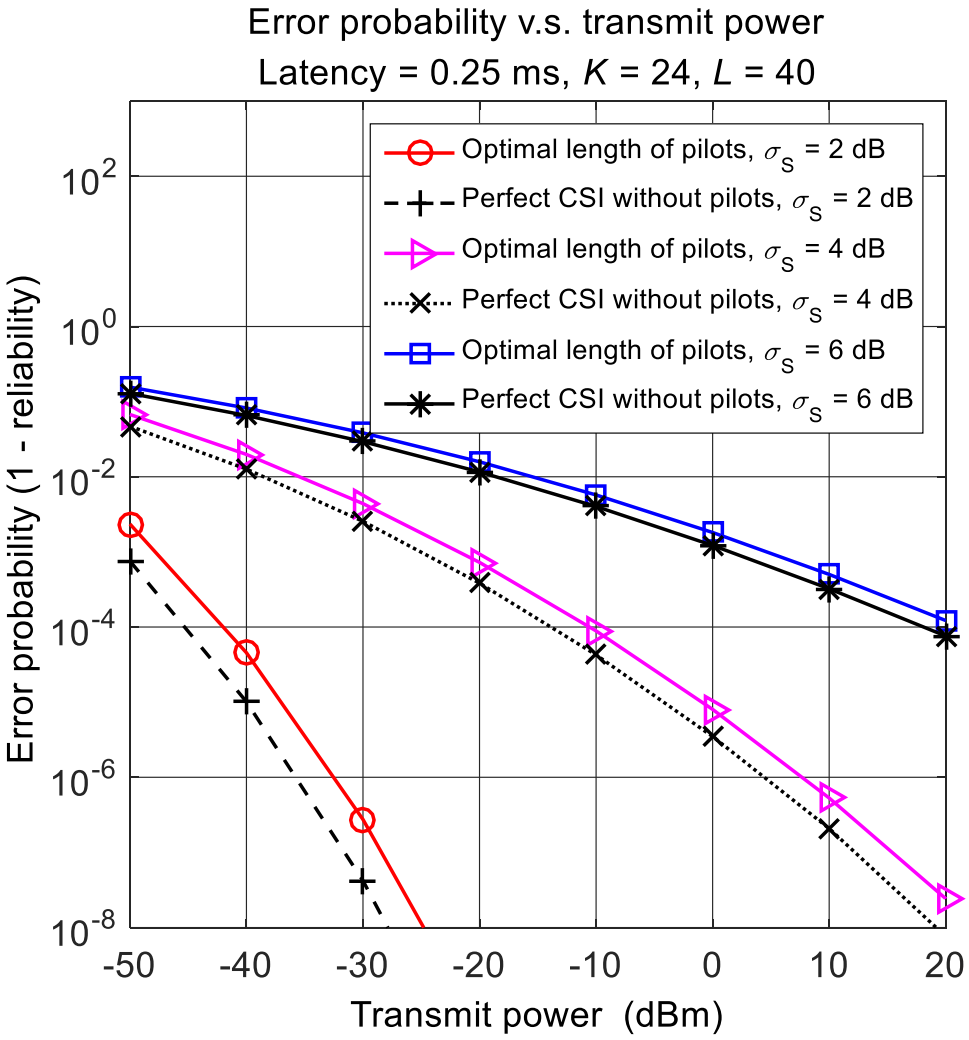

Figure 5.9 : The trade-off between reliability and Tx power at different levels of shadow fading.

5.5 Summary of This Chapter

This chapter studies MU-MIMO and analyzes the error probability under constrained latency and fading. We derive the post-processing SNRs for ZF detection. By assuming that shadow fading obeys a log-normal distribution, the pdf of post-processing SNRs has also been derived. Also, this chapter implements the FBL information theory to derive the error probability of ZF detection. Based on the theoretical results, although it is necessary to serve multiple users simultaneously,

MU-MIMO can ensure reliability in a severe fading channel by installing a great number of Rx antennas. Furthermore, GSSM is studied to determine the LoP. As has been verified in simulation results, MU-MIMO can achieve 99.999% reliability within a less than 0.5 ms transmission latency.

Chapter 6

Summary and Future Work

6.1 Summary of This Thesis

This thesis has studied NOMA and its key technologies for low-latency UL communications in 5G. Based on the characteristics of short data packets, reliability and latency are analyzed and optimized with the FBL information theory. This thesis has contributed lots of effort to the analysis of NOMA and low-latency transmission, multi-antenna enhanced NOMA, FD-SCMA, and massive MU-MIMO NOMA. We have in-depth considered the impact of the number of accessing users, the small packet size, the small number of Tx antennas at the users, the severe shadow fading, and imperfect CSI. The latency of simultaneous transmission of multiple users is effectively reduced with optimized schemes from the perspectives of access delay, transmission delay, and processing delay. The design of advanced MUD can maintain the reliability of low-latency communications at a high level.

The main work of this thesis can be summarized as follows.

1. For the transmission of IoT short packets in UL, the potential key technologies of NOMA and low-latency communications have been investigated. From the perspective of diversity, it is suggested that the diversity gained from the frequency and spatial domains can reduce UL transmission latency in NOMA without a substantial loss in reliability. The emerging grant-free NOMA can considerably shorten access latency, and processing latency can be shortened by stable and low-complexity MUD.

2. In IoT applications, the number of Tx antennas at each user is small, while the number of Rx antennas at the BS can be large. The combination of the rate splitting and SIC detection techniques in MU-MIMO NOMA can guarantee

stable demodulation and decoding in UL transmission. The minimum data rate of accessing users has been increased, thereby reducing the maximum transmission latency of accessing users. The number of SIC operations at the receiver can be substantially reduced through the two-layer rate splitting algorithm and the group SIC algorithm, thereby shortening the processing latency.

3. UL users can be further reduced with the incorporation of the FD technology into SCMA and application of simultaneous UL and DL transmission. Considering the impact of incomplete SIS at the gNB, the emerging FBL information theory is implemented to derive the error probability. Theoretical analysis and numerical results suggest that the proposed FD-SCMA outperforms the existing FD, SCMA, and OMA schemes in time-invariant fading channels. Through DL power control, the interference from DL to UL can be reduced to minimize the error probability in UL. Further, it can minimize the maximum error probability in UL and DL. Moreover, FD-SCMA can reduce access latency and transmission latency.

4. To overcome the severe channel fading and imperfect CSI in IoT, massive MU-MIMO NOMA is utilized to ensure the reliability of UL low-latency transmission. According to the distribution of users and the characteristics of channel fading factors, the pdf of effective SNRs of ZF detection under perfect and imperfect CSI is derived. Then, the FBL information theory is applied to calculate the error probability of access users given transmission latency. Optimizing the LoP through GSSM minimizes the error probability to alleviate the effects of shadow fading and imperfect CSI. The massive MU-MIMO NOMA can implement low-complexity ZF detection to support the reliable transmission of short data packets for a large number of users. Furthermore, this can shorten the access latency and processing latency.

6.2 Future Work

The following directions are listed for further study.

1. MUD leads to a trade-off between reliability and complexity. In this thesis, the computational complexity is mainly derived from theoretical estimation, and the reduction in processing latency is evaluated in simulations. In practice, accurate measurement needs to be established in actual deployment. Since the processing latency is enormously affected by the capability of devices, with the rapid deployment and commercialization of 5G, the large-scale applications of advanced devices can reduce the processing latency of MUD to some extent.

2. The numerous IoT scenarios will lead to complex channel conditions, that need to be modeled and considered in future research. Currently, some researchers are trying to unify the mathematical expressions of various classical channel models. Also, the blocking effect caused by body movements is modeled in IoT, and several innovative distributions are assumed to depict SSF coefficients. The new features of these channels may lead to modifications in the physical layer design, especially when adopting MUD with multiple Rx antennas. Many topics are worthy of further exploration and optimization in channel estimation and MUD. The evaluation and optimization of various emerging technologies and proposed algorithms are worth further study under a variety of channel models.

3. Grant-free NOMA can remove the grant-request and the scheduling process, thereby significantly reducing the signaling overheads and scheduling latency. The time-frequency resources, along with a dedicated reference signal, are pre-configured to the UE semi-statically for URLLC UL grant-free transmission. Frequencies for hopping between initial transmission and re-transmissions should also be provided to reduce repeated collisions. However, MUD in grant-free NOMA requires high computational complexity. To be deployed more widely, the process of MUD needs to be optimized. In addition, the interference becomes more pronounced under a network supporting grant-free NOMA. As such, grant-free NOMA needs to be further studied and more fully understood, thereby managing the in-band and out-of-band interference that is within the system and to other systems.

Bibliography

[1] 3GPP TR 38.913 V14.3.0, "Study on scenarios and requirements for next generation access technologies (Release 14)," Jun. 2017.

[2] L. Dai *et al.* "Non-orthogonal multiple access for 5G: Solutions, challenges, opportunities, and future research trends," *IEEE Commun. Mag.* , vol. 53, no. 9, pp. 74-81, Sep. 2015.

[3] Y. Yuan *et al.* "Non-orthogonal transmission technology in LTE evolution," *IEEE Commun. Mag.* , vol. 54, no. 7, pp. 68-74, Jul. 2016.

[4] Z. Ding, X. Lei, G. K. Karagiannidis, R. Schober, J. Yuan, and V. K. Bhargava, "A survey on non-orthogonal multiple access for 5G networks: Research challenges and future trends," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 10, pp. 2181-2195, Oct. 2017.

[5] A. Benjebbour, Y. Saito, Y. Kishiyama, A. Li, A. Harada, and T. Nakamura, "Concept and practical considerations of non-orthogonal multiple access (NOMA) for future radio access," in *Proc. International Symposium on Intelligent Signal Process. and Commun. Systems*, Naha, Japan, Nov. 2013, pp. 770-774.

[6] A. Li, A. Benjebbour, X. Chen, H. Jiang, and H. Kayama, " Uplink non-orthogonal multiple access (NOMA) with single-carrier frequency division multiple access (SC-FDMA) for 5G systems," *IEICE Trans. Commun.*, vol. E98B, no. 8, pp. 1426–1435, Aug. 2015.

[7] Y. Saito, A. Benjebbour, A. Li, K. Takeda, Y. Kishiyama, and T. Nakamura, "System-level evaluation of downlink non-orthogonal multiple access (NOMA) for non-full buffer traffic model," in *Proc. 2015 IEEE Conf. Stand. Commun. Networking, CSCN* , 2015, pp. 94–99.

[8] S. Chen, B. Ren, Q. Gao, S. Kang, *et al.*, "Pattern division multiple access (PDMA)-A novel nonorthogonal multiple access for fifth-geneation radio networks," *IEEE Trans. Veh. Tech.*, vol. 66, no. 4, pp. 3185-3196, Apr. 2016

[9] H. Nikopour, E. Yi, A. Bayesteh, and K. Au, "SCMA for downlink multiple access of 5G wireless networks," in *Proc. IEEE Global Telecommun. Conf.*, Austin, USA, Dec. 2014, pp. 3940-3945.

[10] A. Goldsmith, *Wireless Communications.* Cambridge, U.K.: Cambridge university press, 2005.

[11] S. Yang and L. Hanzo, "Fifty years of MIMO detection: The road to large-scale MIMOs," *IEEE Commun. Surv. Tutorials*, vol. 17, no. 4, pp. 1941–1988, 2015.

[12] H. Mu, Z. Ma, M. Alhaji, P. Fan, and D. Chen, "A fixed low complexity message pass algorithm detector for up-link SCMA system," *IEEE Wirel. Commun. Lett.*, vol. 4, no. 6, pp. 585–588, 2015.

[13] D. Wei, Y. Han, S. Zhang, and L. Liu, "Weighted message passing algorithm for SCMA," in *Proc. International Conf. Wireless Commun. Signal Process.*, Nanjing, China, Oct 2015, pp. 1-5.

[14] B. Xiao, K. Xiao, S. Zhang, Z. Chen, B. Xia, and H. Liu, "Wei Iterative detection and decoding for SCMA systems with LDPC codes," in *2015 International Conference on Wireless Communications & Signal Processing (WCSP)*, 2015, vol. 2, pp. 1–5.

[15] C. Yan, A. Harada, A. Benjebbour, Y. Lan, A. Li, and H. Jiang, "Receiver design for downlink non-orthogonal multiple access (NOMA)," in *Proc. IEEE 81st Veh. Technol. Conf.*, Glasgow, Glasgow, UK, May 2015, pp. 1-6.

[16] Y. Sun, D. W. K. Ng, Z. Ding and R. Schober, "Optimal joint power and subcarrier allocation for full-duplex multicarrier non-orthogonal multiple access systems," *IEEE Trans. Commun.*, vol. 65, no. 3, pp. 1077-1091, Mar. 2017.

[17] F. Fang, H. Zhang, J. Cheng, S. Roy, and V. C. M. Leung, "Joint user scheduling and power allocation optimization for energy-efficient NOMA systems with imperfect CSI," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 12, pp. 2874-2885, Dec. 2017.

[18] B. Soret, P. Mogensen, K. I. Pedersen, and M. C. Aguayo-Torres, "Fundamental tradeoffs among reliability, latency and throughput in cellular networks," in *Proc. 2014 IEEE Globecom Workshops (GC Wkshps)*, 2014, pp. 1391–1396.

[19] P. Popovski, "Ultra-reliable communication in 5G wireless systems," in *Proc. 1st International Conference on 5G for Ubiquitous Connectivity*, 2014, pp. 146–151.

[20] N. A. Johansson, Y. E. Wang, E. Eriksson, and M. Hessler, "Radio access for ultra-reliable and low-latency 5G communications," in *Proc. 2015 IEEE International Conference on Communication Workshop, ICCW*, 2015, pp. 1184–1189.

[21] Z. Ding *et al.*, "Application of non-orthogonal multiple access in LTE and 5G networks," *IEEE Commun. Mag.*, vol. 55, no. 2, pp. 185–191, Feb. 2017.

[22] K. Higuchi and A. Benjebbour, "Non-orthogonal multiple access (NOMA) with successive interference cancellation for future radio access," *IEICE Trans. Commun.*, vol. E98.B, no. 3, pp. 403-414, 2015.

[23] K. Higuchi, "NOMA for future cellular systems," in *Proc. IEEE 84th Veh. Technol. Conf.*, Montreal, QC, Canada, Sept. 2016, pp. 1-5.

[24] F. Luo and C. Zhang, "Non-orthogonal multiple access (NOMA): Concept and design," in *Proc. Signal Process. for 5G: Algorithms and Implementations, 1*, Wiley-IEEE Press, 2016, pp. 143-167.

[25] Y. Saito, Y. Kishiyama, A. Benjebbour, T. Nakamura, A. Li, and K. Higuchi, "Non-orthogonal multiple access (NOMA) for cellular future radio access," in *Proc. IEEE 77th Veh. Technol. Conf.*, Dresden, Germany, Jun. 2013, pp. 1-5. S. Chen, B. Ren, Q. Gao, S. Kang, *et al.*, "Pattern division multiple access (PDMA)-A novel nonorthogonal multiple access for fifth-geneation radio networks," *IEEE Trans. Veh. Tech.*, vol. 66, no. 4, pp. 3185-3196, Apr. 2016

[26] L. Dai, B. Wang, Z. Ding, Z. Wang, S. Chen, and L. Hanzo, "A survey of non-orthogonal multiple access for 5G," *Commun. Surveys Tuts.*, pp. 1-30, 2018.

[27] W. Tang, S. Kang, B. Ren and X. Yue, "Uplink grant-free pattern division multiple access (GF-PDMA) for 5G radio access," *China Commun.*, vol. 15, no. 4, pp. 153-163, Apr. 2018.

[28] D. Kong, J. Zeng, X. Su, L. Rong, and X. Xu, "Multiuser detection algorithm for PDMA uplink system based on SIC and MMSE," in *Proc. IEEE/CIC International Conf. Commun. in China*, Chengdu, China, July 2016.

[29] B. Ren *et al.*, "Advanced IDD receiver for PDMA uplink system," in *Proc. 2016 IEEE/CIC International Conference on Communications in China (ICCC)*, 2016, pp. 1–6.

[30] M. Taherzadeh, H. Nikopour, A. Bayesteh, and H. Baligh, "SCMA codebook design," in *Proc. IEEE 80th Veh. Technol. Conf.*, Vancouver, BC, Canada, Sept. 2014, pp. 1-5.

[31] Y. Wu, S. Zhang, and Y. Chen, "Iterative multiuser receiver in sparse code multiple access systems," in *Proc. IEEE International Conf. Commun.*, London, UK, June 2015, pp. 2918-2923.

[32] C. Zhang, Y. Luo and Y. Chen, "A low-complexity SCMA detector based on discretization," *IEEE Trans. Wireless Commun.*, vol. 17, no. 4, pp. 2333-2345, Apr. 2018.

[33] M. Jia, L. Wang, Q. Guo, X. Gu and W. Xiang, "A low complexity detection algorithm for fixed up-link SCMA system in mission critical scenario," in *IEEE Internet of Things Journal*, 2017.

[34] L. Yang, Y. Liu, and Y. Siu, "Low complexity message passing algorithm for SCMA system," *IEEE Commun. Lett.*, vol. 20, no. 12, pp. 2466-2469, Dec. 2016.

[35] Y. Polyanskiy, H. V. Poor, and S. Verdu, "Channel coding rate in the finite blocklength regime," *IEEE Trans. Inf. Theory*, vol. 56, no. 5, pp. 2307-2359, May 2010.

[36] W. Yang, G. Durisi, T. Koch, and Y. Polyanskiy, "Quasi-static multiple-antenna fading channels at finite blocklength," *IEEE Trans. Inf. Theory*, vol. 60, no. 7, pp. 4232-4265, Jul. 2014.

[37] G. Durisi, T. Koch, J. Östman, Y. Polyanskiy, and W. Yang, "Short-packet communications over multiple-antenna Rayleigh-fading channels," *IEEE Trans. Commun.*, vol. 64, no. 2, pp. 618-629, Feb. 2016.

[38] G. Durisi, T. Koch, and P. Popovski, "Toward massive, ultrareliable, and low-latency wireless communication with short packets," *Proc. IEEE*, vol. 104, no. 9, pp. 1711-1726, Sep. 2016.

[39] Y. Hu, M. C. Gursoy, and A. Schmeink, "Relaying-enabled ultra-reliable low-latency communications in 5G," *IEEE Netw.*, vol. 32, no. 2, pp. 62-68, Mar. 2018.

[40] C. She, C. Yang, and T. Q. S. Quek, "Uplink transmission design with massive machine type devices in tactile Internet," in *Proc. 2016 IEEE Globecom Workshops (GC Wkshps)*, 2016, pp. 1–6.

[41] S. R. Panigrahi, N. Bjorsell, and M. Bengtsson, "Feasibility of large antenna arrays towards low latency ultra reliable communication," in *Proc. IEEE International Conference on Industrial Technology (ICIT)*, Toronto, Canada, Mar. 22-25, 2017, pp. 1289-1294.

[42] A. Tassi, I. Chatzigeorgiou, and D. E. Lucani, "Analysis and optimization of sparse random linear network coding for reliable multicast services," *IEEE Trans. Commun.*, vol. 64, no. 1, pp. 285–299, 2016.

[43] B. Farayev, Y. Sadi, and S. C. Ergen, "Optimal power control and rate adaptation for ultra-reliable M2M control applications," in *Proc. 2015 IEEE Globecom Workshops (GC Wkshps)*, 2015, pp. 1–6.

[44] F. Schuh and J. B. Huber, "Punctured vs. multidimensional TCM - A comparison w.r.t. complexity," in *Proc. 2014 IEEE Globecom Workshops (GC Wkshps)*, 2014, pp. 1408–1413.

[45] M. S. Elbamby, M. Bennis, W. Saad, M. Debbah, and M. Latva-aho, "Resource optimization and power allocation in in-band full duplex-enabled non-orthogonal multiple access networks," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 12, pp. 2860–2873, Dec. 2017.

[46] A. Yadav, O. A. Dobre, and N. Ansari, "Energy and traffic aware full-duplex communications for 5G systems," *IEEE Access*, vol. 5, pp. 11278-11290, 2017.

[47] A. Yadav and O. A. Dobre, "All technologies work together for good: A glance at future mobile networks," *IEEE Wirel. Commun.*, vol. 25, no. 4, pp. 10–16, Aug. 2018.

[48] H. Chen *et al.*, "Ultra-reliable low latency cellular networks: Use cases, challenges and approaches," *IEEE Commun. Mag.*, vol. 56, no. 12, pp. 119–125, 2018.

[49] B. Wang, L. Dai, Y. Zhang, T. Mir, and J. Li, "Dynamic compressive sensing-based multi-user detection for uplink grant-free NOMA," *IEEE Commun. Lett.*, vol. 20, no. 11, pp. 2320–2323, 2016.

[50] Y. Du *et al.*, "Efficient multi-user detection for uplink grant-free NOMA: Prior-information aided adaptive compressive sensing perspective," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 12, pp. 2812–2828, Dec. 2017.
J. Zeng, D. Kong, B. Liu, X. Su, and T. Lv, "RIePDMA and BP-IDD-IC detection," *EURASIP Journal on Wireless Commun. and Networking*, vol. 12, no. 1, 2017.
J. Zeng, D. Kong, X. Su, L. Rong and X. Xu, "On the performance of pattern division multiple access in 5G systems," in *Proc. 8th International Conf. on Wireless Commun. & Signal Processing (WCSP)*, Yangzhou, China, 2016, pp. 1-5.

[51] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication.* Cambridge, U.K.: Cambridge Univ. Press, 2005.

[52] L. Liu, C. Yuen, Y. L. Guan, Y. Li, and C. Huang, "Gaussian message passing iterative detection for MIMO-NOMA systems with massive access," in *Proc. 2016 IEEE Global Communications Conference (GLOBECOM)*, 2016, pp. 1–6.

[53] Z. Ding, R. Schober, and H. V. Poor, "A general MIMO framework for NOMA downlink and uplink transmission based on signal alignment," *IEEE Trans. Wirel. Commun.*, vol. 15, no. 6, pp. 4438-4454, Jun. 2016.

[54] Z. Ding, L. Dai, and H. V. Poor, "MIMO-NOMA design for small packet transmission in the Internet of Things," *IEEE Access*, vol. 4, pp. 1393-1405, 2016.

[55] Z. Ding, F. Adachi, and H. V. Poor, "The application of MIMO to non-orthogonal multiple access," *IEEE Trans. Wireless Commun.*, vol. 15, no. 1, pp. 537-552, Jan. 2016.

[56] L. Liu, C. Yuen, Y. L. Guan, and Y. Li, "Capacity-achieving iterative LMMSE detection for MIMO-NOMA systems," in *Proc. 2016 IEEE International Conference on Communications (ICC)*, 2016, pp. 1–6.

[57] Y. Chi, L. Liu, G. Song, C. Yuen, Y. L. Guan, and Y. Li, "Practical MIMO-NOMA: Low complexity and capacity-approaching solution," *IEEE Trans. Wirel. Commun.*, vol. 17, no. 9, pp. 6251–6264, Sep. 2018.

[58] X. Cheng, M. Zhang, M. Wen, and L. Yang, "Index modulation for 5G: Striving to do more with less," vol. 25, no. 2, pp. 126–132, Apr. 2018.

[59] Z. Wei, D. W. K. Ng, and J. Yuan, "Joint pilot and payload power control for uplink MIMO-NOMA with MRC-SIC receivers," *IEEE Commun. Lett.*, vol. 22, no. 4, pp. 692-695, Apr. 2018.

[60] H. Wang, R. Zhang, R. Song, and S.-H. Leung, "A novel power minimization precoding scheme for MIMO-NOMA uplink systems," *IEEE Commun. Lett.*, vol. 22, no. 5, pp. 1106-1109, May 2018.

[61] Y. Huang, C. Zhang, J. Wang, Y. Jing, L. Yang, and X. You, "Signal processing for MIMO-NOMA: Present and future challenges," *IEEE Wireless Commun.*, vol. 25, no. 2, pp. 32-38, Apr. 2018.

[62] B. Rimoldi, R. Urbanke, and B. H. Rimoldi, "A rate-splitting approach to the Gaussian multiple-access channel," *IEEE Trans. Inf. Theory*, vol. 42, no. 2, pp. 364-375, Feb. 1996.

[63] K. Kobayashi, "A new achievable rate region for the interference channel," *IEEE Trans. Inf. Theory*, vol. 27, no. 1, pp. 49-60, Jan. 1981.

[64] B. Clerckx, H. Joudeh, C. Hao, M. Dai, and B. Rassouli, "Rate splitting for MIMO wireless networks: a promising PHY-layer strategy for LTE evolution," *IEEE Commun. Mag.*, vol. 54, no. 5, pp. 98-105, May 2016.

[65] C. Hao, Y. Wu, and B. Clerckx, "Rate analysis of two-receiver MISO broadcast channel with finite rate feedback: A rate-splitting approach," *IEEE Trans. Commun.*, vol. 63, no. 9, pp. 3232–3246, Sep. 2015.

[66] M. Dai, B. Clerckx, D. Gesbert, and G. Caire, "A rate splitting strategy for massive MIMO with imperfect CSIT," *IEEE Trans. Wirel. Commun.*, vol. 15, no. 7, pp. 4611-4624, Jul. 2016.

[67] H. Joudeh and B. Clerckx, "Robust transmission in downlink multiuser MISO systems: A rate-splitting approach," *IEEE Trans. Signal Process.*, vol. 64, no. 23, pp. 6227–6242, Dec. 2016.

[68] H. Joudeh and B. Clerckx, "Sum-rate maximization for linearly precoded downlink multiuser MISO systems with partial CSIT: A rate-splitting approach," *IEEE Trans. Commun.*, vol. 64, no. 11, pp. 4847–4861, Nov. 2016.

[69] Y. Zhu, Z. Zhangt, X. Wang, and X. Liang, "A low-complexity non-orthogonal multiple access system based on rate splitting," in *Proc. 2017 9th International Conference on Wireless Communications and Signal Processing (WCSP)*, 2017, pp. 1–6.

[70] C. Gong, A. Tajer, and X. Wang, "A practical coding scheme for interference channel using constrained partial group decoder," in *Proc. IEEE Glob. Telecommun. Conf. (GLOBECOM)*, 2011, pp. 1–5.

[71] M. Ashraphijuo, X. Wang, and M. Tao, "Multicast beamforming design in multicell networks with successive group decoding," *IEEE Trans. Wirel. Commun.*, vol. 16, no. 6, pp. 3492–3506, 2017.

[72] C. Gong, A. Tajer, and X. Wang, "Interference channel with constrained partial group decoding," *IEEE Trans. Commun.*, vol. 59, no. 11, pp. 3059–3071, Nov. 2011.

[73] B. Zheng, X. Wang, M. Wen, and F. Chen, "NOMA-based multi-pair two-way relay networks with rate splitting and group decoding," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 10, pp. 2328-2341, Oct. 2017.

[74] W. W. Hager, "Updating the inverse of a matrix," *SIAM review*, vol. 31, no. 2, pp. 221-239, Feb. 1989.

[75] G. J. Sutton *et al.*, "Enabling ultra-reliable and low-latency communications through unlicensed spectrum," *IEEE Netw.*, vol. 32, no. 2, pp. 70-77, Mar. 2018.

[76] P. Popovski *et al.*, "Wireless access for ultra-reliable low-latency communication: Principles and building blocks," *IEEE Netw.*, vol. 32, no. 2, pp. 16-23, Mar. 2018.

[77] J. Zhang *et al.*, "PoC of SCMA-based uplink grant-free transmission in UCNC for 5G," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 6, pp. 1353–1362, 2017.

[78] L. Zhang, J. Liu, M. Xiao, G. Wu, Y. C. Liang, and S. Li, "Performance analysis and optimization in downlink NOMA systems with cooperative full-duplex relaying," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 10, pp. 2398-2412, Oct. 2017.

[79] Z. Zhang, Z. Ma, M. Xiao, Z. Ding, and P. Fan, "Full-duplex device-to-device-aided cooperative nonorthogonal multiple access," *IEEE Trans. Veh. Technol.*, vol. 66, no. 5, pp. 4467-4471, May 2017.

[80] Y. Xin, M. Ma, Z. Zhao, and B. Jiao, "Co-channel interference suppression techniques for full duplex cellular system," *China Communications*, vol. 12, no. Supplement, pp. 18-27, Dec. 2015.

[81] D. Kim, H. Lee, and D. Hong, "A survey of in-band full-duplex transmission: From the perspective of PHY and MAC layers," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 4, pp. 2017-2046, Fourth Quart. 2015.

[82] A.-A. Boulogeorgos, H. Bany Salameh, and G. Karagiannidis, "Spectrum sensing in full-duplex cognitive radio networks under hardware imperfections," *IEEE Trans. Veh. Technol.*, vol. 66, no. 3, pp. 2072-2084, Mar. 2017.

[83] Z. Zhou, L. Liu, and J. Zhang, "FD-MIMO via pilot-data superposition: Tensor-based DOA estimation and system performance," *IEEE J. Sel. Topics Signal Process.*, vol. 12, no. Supplement, pp. 18-27, Dec. 2015.

[84] S. Biswas, K. Singh, O. Taghizadeh, and T. Ratnarajah, "Design and analysis of FD MIMO cellular systems in coexistence with MIMO radar," *IEEE Trans. Wireless Commun.*, vol. 19, no. 7, pp. 4727-4743, Jul. 2020.

[85] I. Parvez, A. Rahmati, I. Guvenc, A. I. Sarwat, and H. Dai, "A survey on low latency towards 5G: RAN, core network and caching solutions," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 4, pp. 3098-3130, Fourth Quart. 2018.

[86] H. Ji, S. Park, J. Yeo, Y. Kim, J. Lee and B. Shim, "Ultra-reliable and low-latency communications in 5G downlink: Physical layer aspects," *IEEE Wireless Commun.*, vol. 25, no. 3, pp. 124-130, Jun. 2018.

[87] S.-H. Moon, K.-J. Lee, J. Kim, and I. Lee, "Link performance estimation techniques for MIMO-OFDM systems with maximum likelihood receiver," *IEEE Trans. Wireless Commun.*, vol. 11, no. 5, pp. 1808-1816, May 2012.

[88] M. Cheng, Y. Wu, Y. Li, Y. Chen, and L. Zhang, "PHY abstraction and system evaluation for SCMA with UL grant-free transmission," in *Proc. IEEE Veh. Technol. Conf. (VTC-Spring)*, Sydney, NSW, Australia, Jun. 2017, pp. 1-5.

[89] X. Sun, S. Yan, N. Yang, Z. Ding, C. Shen and Z. Zhong, "Short-packet downlink transmission with non-orthogonal multiple access," *IEEE Trans. Wireless Commun.*, vol. 17, no. 7, pp. 4550-4564, Jul. 2018.

[90] W. Yang, G. Caire, G. Durisi, and Y. Polyanskiy, "Optimum power control at finite blocklength," *IEEE Trans. Inf. Theory*, vol. 61, no. 9, pp. 4598–4615, 2015.

[91] T. K. Vu, C.-F. Liu, M. Bennis, M. Debbah, M. Latva-aho, and C. S. Hong, "Ultra-reliable and low latency communication in mmWave-enabled massive MIMO networks," *IEEE Commun. Lett.*, vol. 21, no. 9, pp. 2041-2044, May 2017.

[92] H. Q. Ngo, E. G. Larsson, and T. L. Marzetta, "Energy and spectral efficiency of very large multiuser MIMO systems," *IEEE Trans. Commun.*, vol. 61, no. 4, pp. 1436-1449, Apr. 2013.

[93] G. Fodor, P. Di Marco, and M. Telek, "On minimizing the MSE in the presence of channel state information errors," *IEEE Commun. Lett.*, vol. 19, no. 9, pp. 1604-1607, Sep. 2015.

[94] A. Abrardo, G. Fodor, M. Moretti, and M. Telek, "MMSE receiver design and SINR calculation in MU-MIMO systems with imperfect CSI," *IEEE Wirel. Commun. Lett.*, vol. 8, no. 1, pp. 269–272, Feb. 2019.

[95] S. S. Ikki and S. Aissa, "Two-way amplify-and-forward relaying with Gaussian imperfect channel estimations," *IEEE Commun. Lett.*, vol. 16, no. 7, pp. 956-959, Jul. 2012.

Y. Chen, L. Wang, Y. Ai, B. Jiao, and L. Hanzo, "Performance analysis of noma-sm in vehicle-to-vehicle massive mimo channels," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 12, pp. 2653-2666, Dec. 2017.

[96] C. Wang *et al.*, "On the performance of the MIMO zero-forcing receiver in the presence of channel estimation error," *IEEE Trans. Wireless Commun.*, vol. 6, pp. 805-810, Mar. 2007.

[97] D. Ben Cheikh, J. M. Kelif, M. Coupechoux, and P. Godlewski, "Multicellular Alamouti scheme performance in Rayleigh and shadow fading," *Annals of Telecommunications-annales des Telecommunications*, vol. 68, no. 5-6, pp. 345-358, Jun. 2013.

[98] J. H. Winters, J. Salz, R. D. Gitlin, "The impact of antenna diversity on the capacity of wireless communication systems," *IEEE Trans. Commun.*, vol. 42, no. 2/3/4, pp. 1740-1751, Feb./Mar./Apr. 1994.

[99] W. Härdle, and L. Simar, *Applied Multivariate Statistical Analysis.* New York, NY, USA: Springer, 2012

[100] A. J. Izenman, *Modern Multivariate Statistical Techniques: Regression, Classification and Manifold Learning.* New York, NY, USA: Sringer, 2008.

[101] Q. Zhang, C. He, and L. Jiang, "Per-stream MSE based linear transceiver design for MIMO interference channels with CSI error," *IEEE Trans.Commun.*, vol. 63, no. 5, pp. 1676-1689, May 2015.

[102] H. Shen, B. Li, M. Tao, and X.Wang, "MSE-based transceiver designs for the MIMO interference channel," *IEEE Trans. Wireless Commun.*, vol. 9, no. 11, pp. 3480-3489, Nov. 2010.

[103] B. Koh, S. Choi, and J. Chun, "A SAR autofocus technique with MUSIC and golden section search for range bins with multiple point scatterers," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 8, pp. 1600-1604, Aug. 2015.

[104] D. H. Yeom, J. B. Park, and Y. H. Joo, "Selection of coefficient for equalizer in optical disc drive by golden section search," *IEEE Trans. Consum. Electron.*, vol. 56, no. 2, pp. 657-662, Jul. 2010.

[105] Y. Xiong, S. Shafer, "Depth from focusing and defocusing", in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Jun. 1993, pp. 68-73.