

Chapter 3

Multiple Access Techniques



Yijie Mao and Bruno Clerckx

Multiple access is an essential physical-layer technique in wireless communication networks that allows multiple mobile users to access the network simultaneously. Driven by the upsurge of devices expected in the fifth generation (5G) and beyond, future wireless communication networks are foreseen to operate in dynamic regimes ranging from underloaded (where the number of scheduled devices is smaller than the number of transmit antennas on each access point) to overloaded (where the number of scheduled devices is larger than the number of transmit antennas on each access point). Besides, each transmitter is required to simultaneously serve devices with heterogeneous capabilities, deployments, as well as qualities of channel state information at the transmitter (CSIT) since the devices for 5G and beyond tend to be more diverse including low-end units such as Internet of Things (IoT) and machine-type communications (MTC)-type devices and high-end equipment such as smartphones with varied user deployments and applications. The resulting requirements for massive connectivity, high throughput, as well as quality of service (QoS) heterogeneity have recently sparked interests in redesigning multiple access techniques for the downlink of communication systems.

This chapter first reviews the state-of-the-art multiple access techniques including their benefits and limitations, followed by introducing the promising multiple access candidate, *rate-splitting multiple access (RSMA)* for 5G and beyond, and a comprehensive comparison among all multiple access techniques. The challenges and future trends of using RSMA will be summarized in the end.

Y. Mao · B. Clerckx (✉)

Communication and Signal Processing Group, Department of Electrical and Electronic Engineering, Imperial College London, London, UK

e-mail: y.mao16@imperial.ac.uk; b.clerckx@imperial.ac.uk

3.1 Evolution of Multiple Access Techniques

The past decades have witnessed the development of multiple access techniques brought by the evolution of cellular networks from the first generation (1G) to 5G. From orthogonal multiple access (OMA) to non-orthogonal multiple access (NOMA) and space-division multiple access (SDMA), multiple access techniques have progressed toward serving more users non-orthogonally in each subcarrier due to the scarcity of spectrum. In this section, those existing multiple access techniques are reviewed.

3.1.1 Orthogonal Multiple Access (OMA)

The 1G wireless communication system introduced in the 1980s employs *frequency-division multiple access (FDMA)* where the frequency bandwidth is divided into nonoverlapping frequency sub-channels and each user is allocated with an independent sub-channel. It was used to support the original analog voice services. The second generation (2G) is developed in the 1990s to further enhance the voice service quality as well as to enable short messaging service. The 2G standard system, Global System for Mobile Communications (GSM), adopts *time-division multiple access (TDMA)* where the frequency domain is shared by all users, while the time domain is divided into different time slots and occupied by independent users. The third generation (3G) introduced in the 2000s opens the new dimension of code to design multiple access where *code-division multiple access (CDMA)* is commercially applied to support TV streaming, mobile video calls, and so on. Different from FDMA and TDMA, CDMA enables the simultaneous transmission for multiple users through the same sub-channels by employing the spread spectrum technology to avoid inter-user interference. In 2009, the fourth generation (4G) based on the long-term evolution (LTE) standard is developed to meet the increasing user demand for more sophisticated mobile devices. By employing *orthogonal frequency-division multiple access (OFDMA)* as the standard multiple access technique, the time and frequency resources are further divided into narrow time slots and subcarriers, respectively. The resource blocks formed by the divided time–frequency grids are allocated to the users dynamically. Compared with FDMA, TDMA, and CDMA, OFDMA is more robust and achieves a higher spectral efficiency. The robustness comes from its ability of combatting narrowband co-channel interference and multipath fading by scheduling users over orthogonal subcarriers, while the spectral efficiency comes from its ability of multiplexing users with low data rate into a wider channel with adaptive transmission rate for each user. All of the aforementioned multiple access techniques are categorized into *orthogonal multiple access (OMA)* where users are scheduled in orthogonal dimensions.

3.1.2 Space-Division Multiple Access (SDMA)

Driven by the increasing user demands, access points nowadays are commonly equipped with multiple antennas. The arisen multiple-input multiple-output (MIMO) systems have been widely used in modern wireless standards, including mobile Worldwide Interoperability for Microwave Access (WiMAX) systems, 4G LTE standard, IEEE 802.11n, and so on. The spatial dimension introduced by MIMO systems opens the door to *space-division multiple access (SDMA)*. By utilizing the spatial dimension to separate users, SDMA allows multiple users to be served simultaneously in the same time–frequency resources.

The only strategy that achieves the capacity region of the multiple-input single-output (MISO)/MIMO (Gaussian) broadcast channel (BC) with perfect CSIT is the complex dirty paper coding (DPC) [1], in which the transmitter relies on perfect CSIT to encode the user messages and perform enhanced interference cancellation such that the encoded data stream experiences no interference from previously encoded streams. However, due to the high computational burden of implementing DPC in practice, linear precoding at the transmitter is more practical and attractive since it simplifies the transmitter design [2]. SDMA is therefore commonly implemented using multi-user linear precoding (MU–LP) either in closed-form beamforming or optimized beamforming using optimization tools. Though the beamformer might be suboptimal, SDMA based on MU–LP is shown to be useful especially when users experience semi-orthogonal channels and relatively similar channel strength or long-term signal-to-noise ratio (SNR) [3]. Hence, it is well-acknowledged and becomes the fundamental multiple access of various 4G and 5G techniques such as multi-user MIMO (MU-MIMO), massive MIMO, network MIMO, millimeter-wave MIMO, and coordinated multipoint (CoMP). Figure 3.1 illustrates the system model of K -user SDMA based on MU–LP for MISO BC. The messages W_1, \dots, W_K intended for K users are independently encoded into data streams s_1, \dots, s_K and superimposed at the transmitter after linear precoding. Each

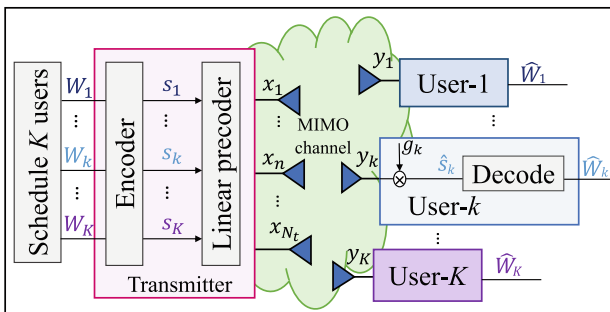


Fig. 3.1 Transmission model of K -user SDMA based on MU–LP

user directly decodes its intended stream by treating interference from streams for all other users as noise.

The main benefit of SDMA is its capability of achieving all spatial multiplexing gains of MISO BC with perfect CSIT. With the use of MU-LP, the precoder and receiver complexity remains low. However, there are three major limitations of SDMA based on MU-LP, which are summarized as follows:

- It is only suited to the underloaded regime, and its performance drops dramatically in the overloaded regime since MU-LP requires more transmit antennas than the number of users in MISO BC so as to generate orthogonal beams to manage multiuser interference efficiently. The current approach to deal with overloaded scenarios at the transmitter is to divide users into groups and schedule user groups over orthogonal resources (e.g., time/frequency). Users in the same group are served by MU-LP. However, such approach may increase latency and decrease QoS.
- The performance of SDMA based on MU-LP is sensitive to the user deployments. It is only suited when users have semi-orthogonal channels with similar channel strengths. Though there exists low-complex scheduling and user pairing algorithms to pair users with semi-orthogonal channels [2], the complexity of the scheduler increases rapidly when considering the optimal scheduling with an exhaustive search.
- Though SDMA based on MU-LP achieves the optimal degrees of freedom¹ (DoF) in MISO BC with perfect CSIT [4], there is a significant DoF and performance loss when CSIT becomes imperfect [5]. As SDMA design is motivated by perfect CSIT, the direct application of SDMA in the presence of imperfect CSIT results in additional interference coming from the imperfect linear precoder design [4].

3.1.3 *Non-orthogonal Multiple Access (NOMA)*

With the aim of further boosting the system spectral efficiency, *non-orthogonal multiple access (NOMA)* that superposes users in the same time–frequency resources via the power domain or code domain is introduced. Specifically, NOMA can be categorized into power-domain NOMA (e.g., [6]) and code-domain NOMA (e.g., sparse code multiple access (SCMA) [7]). In this chapter, we focus on power-domain NOMA² that relies on superposition coding (SC) at the transmitter and successive interference cancellation (SIC) at the receiver [6, 8–10] (which is also denoted in short as SC–SIC). The study of NOMA starts from single-input single-output (SISO) (Gaussian) BC and is further extended to multi-antenna BC. In this

¹The DoF, also known as spatial multiplexing gain, characterizes the number of interference-free streams that can be transmitted or equivalently the pre-log factor of the rate at high SNR.

²In the sequel, power-domain NOMA will be referred to simply by NOMA.

chapter, we denote NOMA in SISO BC as single-antenna NOMA while NOMA in MISO/MIMO BC as multi-antenna NOMA.

Single-Antenna NOMA

The study of single-antenna NOMA is inspired by the well-known result in the literature of information theory that SC–SIC is the capacity-achieving technique for SISO BC [11, 12]. Comparing NOMA and OMA, it is well-known that when there are certain channel strength disparities among users, the capacity region of SISO BC is achieved by NOMA, and it is larger than the rate region achieved by OMA (e.g., TDMA) [12]. However, when users experience the same channel strengths, the advantage of NOMA vanishes, and OMA is sufficient to achieve the capacity region [12].

The major benefit of single-antenna NOMA is its ability to improve the spectral efficiency in an overloaded regime by allowing multiple users (that experience different channel strengths or path losses) to be served by one transmitter with single transmit antenna on the same time–frequency resource. However, its limitation is non-negligible. For a K -user SISO BC, $K - 1$ layers of SIC are required at the user with strongest channel strength to sequentially decode the $K - 1$ streams of all other co-scheduled users before decoding its intended stream. As the number of user increases, the receiver complexity and likelihood of error propagation increase significantly. A practical system requires the number of SIC layers to be small. One method is to divide the users into small groups, apply SC–SIC in each group, and schedule groups over orthogonal resources by using OMA, which, however, would lead to some performance loss and latency issue.

Multi-antenna NOMA

Motivated by the benefits of SC–SIC in SISO BC, NOMA has been further applied to multi-antenna BC. There are two main strategies of multi-antenna NOMA, both of which rely on linearly precoded SC–SIC.

The first strategy, which is simply denoted as “SC–SIC,” is a direct application of SC–SIC to MISO/MIMO BC [13–16]. However, contrary to SISO BC, multi-antenna BC is non-degraded, i.e., users cannot be ordered according to their channel strengths in general settings. SC–SIC degrades multi-antenna BC by ordering users based on their effective scalar channels obtained at the transmitter after linear precoding. Users with stronger effective channel strengths are required to decode and remove the streams of users with weaker effective channel strengths in a successive manner. Such strategy forces the multi-antenna non-degraded channel into an effective single-antenna degraded channel since the user with the strongest channel strength is required to decode the messages of all other users. SC–SIC wastes all spatial multiplexing gains in MISO/MIMO BC and is only able to cope with the scenarios when user channels are aligned with certain channel strength

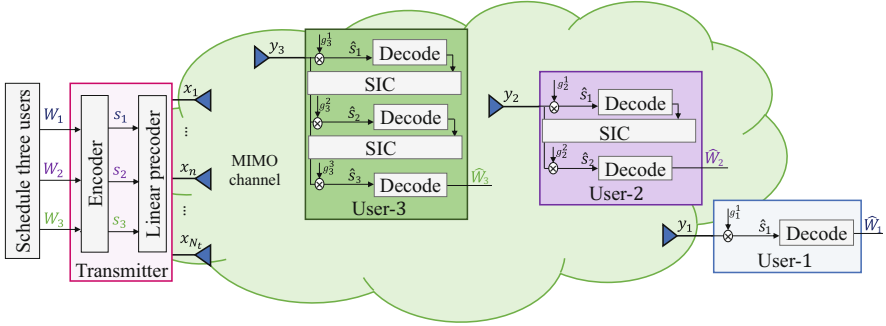


Fig. 3.2 Transmission model of three-user SC-SIC

disparities among them. From the DoF perspective, the sum-DoF achieved by SC-SIC is 1 since one receiver has to decode all streams [17]. It is equal to the DoF achieved by OMA or single-user beamforming. Compared with the sum-DoF $\min\{N_t, K\}$ (where N_t is the number of transmit antennas and K is the number of users in MISO BC) achieved by DPC and MU-LP in a MISO BC with perfect CSIT, SC-SIC in multi-antenna BC results in a significant DoF loss, and such loss comes with a dramatic burden to receivers due to the use of SIC layers. In contrast, MU-LP does not require any SIC at receiver sides, and it achieves a higher spatial multiplexing gain which drives the use of MU-MIMO in 4G [18]. To compensate the DoF loss of SC-SIC, one natural method is to consider dynamic switching between NOMA and SDMA based on the channel states [19]. Figure 3.2 illustrates the transmission model of a three-user SC-SIC with decoding order from user-1 to user-3. Hence, user-3 is required to decode all the three streams.

The second strategy, denoted as “SC-SIC per group,” divides users into disparate groups with users in the same group being served by SC-SIC and users across the groups being served by SDMA in order to coordinate inter-group interference [6, 20–24]. By combining SDMA and NOMA in SC-SIC per group, multi-antenna BC is decomposed into non-interfering single-antenna NOMA channels, and the DoF loss of SC-SIC can be recovered. However, it is only suited to an overloaded regime, and users within the same group require almost aligned channels, while users in different groups require (semi-)orthogonal channels. Figure 3.3 illustrates the transmission model of four-user SC-SIC per group with user-1 and user-2 in group 1 while user-3 and user-4 in group 2. The inner-group interference is decoded based on SC-SIC, while the inter-group interference is treated as interference based on MU-LP. By assuming the decoding order in group 1 is from the message of user-1 to that of user-2, user-2 is required to decode the messages of both user-1 and user-2 while fully treating the inter-group interference from user-3 and user-4 as noise. Similarly in group-2, the decoding order from user-3 to user-4 is assumed.

Multi-antenna NOMA also relies on perfect CSIT as SDMA. When CSIT becomes imperfect, extra multiuser interference is introduced for both SC-SIC and SC-SIC per group strategies [16]. Similarly to single-antenna NOMA, the major

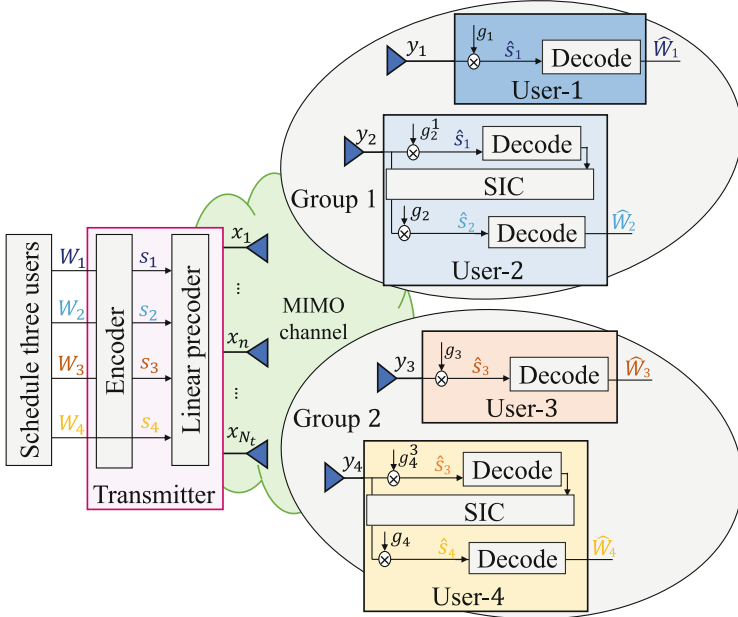


Fig. 3.3 Transmission model of four-user SC-SIC per group

benefit of multi-antenna NOMA is its ability to deal with an overloaded regime with aligned user channels and channel strength disparities. The limitations of multi-antenna NOMA are summarized as follows:

- The DoF loss of multi-antenna NOMA is severe. The fundamental reason that SC-SIC achieves the capacity region of SISO BC is due to the fact that users can be ordered based on channel strengths in such a degraded BC. However, striving to using SC-SIC in non-degraded MISO/MIMO BC degrades multi-antenna BC and results in a waste of spatial resources. Hence, there is an unavoidable DoF loss of SC-SIC in multi-antenna NOMA.
- Multi-antenna NOMA is only suited for specific user deployments when user channels are aligned with a disparity of channel strengths. It is not suited for general settings.
- There is a complexity increase at both the transmitter and the receivers in multi-antenna NOMA. At each receiver, multiple layers of SIC are required to decode and remove the interference from other users. At the transmitter equipped with multiple antennas, the optimization of user grouping, decoding orders, and precoders are coupled since the effective user channels are influenced by the precoders. For example, considering a three-user MISO BC, SC-SIC requires to jointly optimize the precoding vectors of three users and six possible decoding orders, while SC-SIC per group requires the user ordering and grouping to be jointly optimized with precoders. One commonly adopted method to reduce the

complexity at the transmitter in multi-antenna NOMA is to assign the same precoding vector to all users within the same group [6], which, however, would further deteriorate system performance as the overall searching space for optimal precoders is reduced.

- As multi-antenna NOMA is motivated in the presence of perfect CSIT, it is also sensitive to the CSIT inaccuracy as SDMA.

Based on the discussion in Sects. 3.1.2 and 3.1.3, we conclude that SDMA and NOMA are actually two extreme interference management strategies in MISO/MIMO BC where users in NOMA try to fully decode and remove interference created by other users, while users in SDMA always fully treat any residual multiuser interference as noise. Moving toward imperfect CSIT, residual inter-user interference is introduced for both SDMA and NOMA (SC-SIC per group).

3.2 Rate-Splitting Multiple Access (RSMA) for 5G and Beyond

Rate-splitting multiple access (RSMA), based on linearly precoded rate-splitting (RS) at the transmitter and SIC at the receivers, is a more general and powerful multiple access for downlink multi-antenna systems that contains SDMA, NOMA, and OMA as special cases. Apart from SDMA that fully treats interference as noise and NOMA that fully decodes interference, RSMA achieves a more dynamic interference management where the interference is partially decoded and partially treated as noise at each user [4]. At the transmitter that supports RSMA, user messages are split into common and private parts; the common messages are combined and encoded into common streams to be decoded by multiple users, while private messages are independently encoded into private streams to be decoded by the corresponding users. All streams are superimposed at the transmitter and broadcast to the users. Each user relies on layers of SIC to decode the common streams before decoding the intended private stream. By adjusting the power allocation for the common and private streams as well as the message split, RSMA automatically bridges SDMA and NOMA that solely rely on the two extreme interference management strategies or a combination thereof.

3.2.1 Literature Review

The fundamental building block of RSMA is RS technique. The previous study of RS can be categorized into communication and information theory categories. Both are summarized comprehensively in this section.

The information theoretic works on RS are summarized in Table 3.1. The idea of RS is not new. It dates back to Carleial's work and the Han and Kobayashi

Table 3.1 Summary of information theoretic literature on rate-splitting

Timeline	Ref.	Scenarios	CSIT accuracy	Network load	Metric	Main discovery
1981	[25]	SISO IC	Perfect CSIT	Underloaded	Rate region	RS based on Han and Kobayashi (HK) scheme achieves the best-known achievable rate region of the two-user interference channel
1996	[27]	SISO MAC	Perfect CSIT	Underloaded	Capacity region	RS based on successive single-user decoding and interference cancellation achieves the capacity region of the K -user Gaussian MAC
2008	[26]	SISO IC	Perfect CSIT	Underloaded	Capacity region	RS based on the HK scheme achieves rate regions within 1 bit/s/Hz of the capacity region
2013, 2016	[29, 30]	MISO BC	Imperfect CSIT	Underloaded	Sum-DoF	RS (with SIC) achieves the optimum sum-DoF of the K -user underloaded MISO BC with imperfect CSIT
2016	[37]	MISO BC	Imperfect CSIT	Underloaded	DoF region	RS achieves the entire DoF region of the K -user underloaded MISO BC with imperfect CSIT
2016	[34]	MISO BC	Imperfect CSIT	Underloaded	Symmetric DoF	RS achieves a higher symmetric (max-min) DoF than SDMA based on MU-LP for the K -user underloaded MISO BC with imperfect CSIT
2016	[37]	MISO BC	Imperfect CSIT	Overloaded	DoF region	RS achieves the entire DoF region of the K -user overloaded MISO BC with imperfect heterogeneous CSIT qualities
2017	[17]	MISO BC	Perfect CSIT	Overloaded	Symmetric DoF	RS achieves the highest symmetric DoF compared with that of SDMA based on MU-LP and NOMA based on SC-SIC for the K -user overloaded MISO BC with perfect CSIT
2017	[32]	MISO IC	Imperfect CSIT	Underloaded	DoF region	RS achieves the best-known DoF region of the K -cell MISO IC with imperfect CSIT
2017	[33]	MIMO IC, MIMO BC	Imperfect CSIT	Underloaded	DoF region	RS achieves the optimum DoF region of the two-user MIMO IC with imperfect CSIT under certain antenna configurations and CSIT qualities
2017	[36]	MISO BC	Imperfect CSIT	Underloaded	GDoF	RS-assisted interference enhancement approach achieves the entire GDoF region of the two-user underloaded MISO BC with imperfect CSIT

(HK) scheme in 1980s for the two-user SISO interference channel (IC) [25]. Such scheme is further proved in [26] to achieve rate regions within 1 bit/s/Hz of the capacity region. The terminology RSMA is first introduced in [27] for the SISO multiple access channel (MAC), where RS based on successive single-user decoding and interference cancellation has been shown to achieve the capacity region of the K -user Gaussian MAC. However, the uplink RSMA has fundamentally different motivations and structures than the downlink RSMA we considered in this chapter. The use of RS as the building block of RSMA framework is motivated by recent progresses on the fundamental limits of a multi-antenna BC and IC characterized by RS. In contrast with the conventional RS used for MAC or two-user SISO IC, the RSMA technique we introduced here is in a different setup, namely, (1) in a BC and (2) with multiple transmit (and receive) antennas. Note that the study of RS in the multi-antenna BC in both information-theoretical and communication perspectives was initiated a few years ago. In comparison, research on NOMA based on SC–SIC in a BC already appeared for several decades [11, 12]. Up to now, the capacity region of the K -user MISO BC with imperfect CSIT remains an open issue. Instead, attention has been switched to characterizing its DoF region. Surprisingly, the information theoretic upperbound on the sum-DoF of the K -user underloaded MISO BC with imperfect CSIT derived in [28] has been shown to coincides with the sum-DoF achieved by linearly precoded RS with SIC [29, 30]. It is further proved in [31] that RS achieves the entire DoF region of the underloaded MISO BC with imperfect CSIT. In comparison, the sum-DoFs achieved by SDMA based on MU–LP and multi-antenna NOMA are suboptimum. The DoF benefits of RS in imperfect CSIT have also been studied in the underloaded MISO IC [32] and underloaded MIMO IC/BC [33]. The optimum symmetric DoF (also known as max–min DoF) of RS has been studied in [34] for the underloaded MISO BC with imperfect CSIT, where RS achieves a higher symmetric DoF over that of SDMA based on MU–LP. Moving toward the overloaded scenario, the power-partitioning approach that superimposes degraded symbols for no-CSIT users on top of linearly precoded RS symbols for partial-CSIT users has been shown to achieve the entire DoF region of the K -user overloaded MISO BC with imperfect CSIT with heterogeneous CSIT qualities. When CSIT is perfect, the symmetric DoF achieved by RS has been shown to outperform that of SDMA based on MU–LP and NOMA based on SC–SIC in [17] for the K -user overloaded MISO BC with perfect CSIT. To further capture the diversity of channel strengths among users, the generalized DoF (GDoF) has been introduced [26]. The GDoF region of a two-user underloaded MISO BC with imperfect CSIT has been studied in [35, 36] where RS is considered as part of the interference enhancement scheme to achieve the entire GDoF region.

The DoF and GDoF superiority of RSMA over SDMA based on MU–LP and NOMA discovered in the information theoretic literature motivates its recent study at the finite SNR regime for rate enhancement in the practical wireless communication systems. The communication literature on RSMA is summarized in Table 3.2. The DoF improvement of RSMA over SDMA in imperfect CSIT

Table 3.2 Summary of communication literature on rate-splitting

Timeline	Ref.	Scenarios	CSIT accuracy	Network load	Metric	Precoding scheme	RSMA scheme
2016	[38]	MISO BC	Quantized feedback	Underloaded	Ergodic sum rate	Random+ZFBF	1-layer RS
2016	[30]	MISO BC	Imperfect CSIT	Underloaded	Ergodic sum rate, Ergodic rate region	Optimized linear precoding	1-layer RS
2016	[34]	MISO BC	Imperfect CSIT	Underloaded	Max-min rate	Optimized linear precoding	1-layer RS
2016	[37]	MISO BC	Imperfect CSIT	Overloaded	Sum rate	Random+ZFBF	1-layer RS with power splitting
2016	[39]	Massive MIMO	Imperfect CSIT	Underloaded	Sum rate	MRT+RZF	1-layer RS, 2-layer HRS
2017	[40]	MISO BC with hardware impairments	Imperfect CSIT	Underloaded	Sum rate	MRT+RZF	1-layer RS
2017	[17]	Multigroup multicast	Imperfect CSIT	Overloaded	Max-min rate	Optimized linear precoding	1-layer RS
2017	[41]	Millimeter-wave MISO BC	Quantized feedback, statistical CSIT	Underloaded	Sum rate	Hybrid (partially optimized) linear precoding	1-layer RS
2018	[42]	MISO BC	Perfect CSIT	Underloaded, Overloaded	Rate region, WSR	Optimized linear precoding	Generalized RS, 2-layer HRS, 1-layer RS
2018	[43]	Multicell-multigroup multicast	Perfect CSIT	Underloaded	EE	Optimized linear precoding	1-layer RS
2018	[44]	MISO BC	Imperfect CSIT	Underloaded	Sum power	Optimized linear precoding	1-layer RS
2018	[45]	MISO BC	Perfect CSIT	Underloaded	EE	Optimized linear precoding	1-layer RS
2018	[46]	MISO BC	Imperfect CSIT	Underloaded	Ergodic sum rate	ZF+THP	1-layer THPRS
2018	[47]	Multi-pair massive MIMO relaying	Imperfect CSIT	Underloaded	Sum rate	MRT+RZF	1-layer RS
2018	[48]	Multi-beam satellite networks	Imperfect CSIT	Underloaded	Sum rate	Optimized power allocation within each beam	1-layer RS
2019	[49]	Cooperative multicell MISO BC	Perfect CSIT	Underloaded	Rate region, WSR	Optimized linear precoding	Generalized RS, 2-layer HRS, 1-layer RS

(continued)

Table 3.2 (continued)

Year	Scenario	CSIT	Load	Metric	Pre-coding	RS
2019	UAV-assisted C-RAN	Perfect CSIT	Underloaded	WSR	Optimized linear precoding	Generalized RS
2019	mmWave UAV-assisted MISO BC	Perfect CSIT	Underloaded	EE	Optimized linear precoding	1-layer RS
2019	C-RAN	Perfect CSIT	Underloaded	WSR	Optimized linear precoding	1-layer RS
2019	C-RAN	Perfect CSIT	Underloaded	Max-min rate	Optimized linear precoding	Generalized RS
2019	MISO BC with SWIPT	Perfect CSIT	Underloaded	Rate region	Optimized linear precoding	1-layer RS
2019	MISO IC with SWIPT	Imperfect CSIT	Underloaded	Sum transmit power	Optimized linear precoding	1-layer RS
2019	MISO BC with cooperative transmission	Perfect CSIT	Underloaded	Rate region, WSR	Optimized linear precoding	1-layer RS
2019	Non-orthogonal unicast and multicast	Perfect CSIT, Imperfect CSIT	Underloaded, overloaded	Rate region, WSR, EE	Optimized linear precoding	Generalized RS, 2-layer HRS, 1-layer RS
2019	MISO BC	Perfect CSIT	Underloaded	WSR	Optimal common stream precoder+ZF	1-layer RS
2019	Multicarrier multigroup multicast	Perfect CSIT	Overloaded	Max-min rate	Optimized linear precoding	1-layer RS
2020	MIMO BC	Imperfect CSIT	Underloaded	Sum rate	Regularized block diagonalization	1-layer RS

has been reflected in the rate performance at the finite SNR regime according to recent studies [17, 30, 34, 37, 38]. In the presence of quantized feedback, RS reduces CSIT feedback overhead compared to MU-LP when using random beamforming for the common stream and zero-forcing beamforming (ZFBBF) for the private streams [38]. It is further shown that with optimized precoders, RS outperforms MU-LP in the underloaded MISO BC with imperfect CSIT for the ergodic sum rate maximization [30] and the worst-case rate optimization (max-min rate) [34]. When considering the overloaded scenario, RS with power-partitioning strategy has been shown to outperform its time-partitioning counterpart at finite SNR in the overloaded MISO BC with heterogeneous CSIT [37] with low-complex maximum ratio transmission (MRT) or matched filtering beamforming scheme for the common stream and regularized zero-forcing (RZF) beamforming for the private streams. The 2-layer hierarchical RS (HRS) that relies on multiple common messages decoded by different groups of users is proposed in [39] for massive MIMO. Furthermore, the generalized RS scheme of RSMA that embraces 1-layer RS and 2-layer HRS as subcases is proposed in [42] for MISO BC with perfect CSIT, where RSMA shows clear rate region and weighted sum rate (WSR) improvement over SDMA and NOMA. The comparison among SDMA, NOMA and RSMA is further analyzed in the two-user case with optimized precoders to maximize energy efficiency [45] or with low-complex precoding but optimal power allocation for common part and private parts of the user messages [58]. Besides the above studies of RSMA in MISO BC, the transceiver design of RS has been studied in other applications of multi-antenna BC, such as MISO BC with hardware impairment [40], multigroup multicast [40], millimeter-wave (mmWave) systems [41], multi-pair relaying [47], cooperative multicell MISO BC [49], cloud-radio access networks (C-RAN) [52, 53], unmanned aerial vehicle (UAV)-assisted networks [50, 51], simultaneous wireless information and power transfer (SWIPT) [54], cooperative user relaying networks [56, 61], non-orthogonal unicast and multicast [57], multi-carrier systems [59], and so on. All of the above works consider linearly precoded RS at the transmitter, and nonlinear precoder design of RS has been studied as Tomlinson–Harashima Precoded RS (THPRS) [46] and dirty paper coded RS (DPCRS) [62]. Moving toward MIMO BC, different linear combining techniques are studied in [60] with minimum mean-square error (MMSE) combiner showing the best performance.

3.2.2 RSMA Framework

RSMA is a generalized multiple access technique for exploring a larger rate region and the room of QoS enhancement. In the framework of RSMA, there are three commonly studied schemes in the literature, namely, 1-layer RS, 2-layer HRS, and the generalized RS, which are all specified in this section.

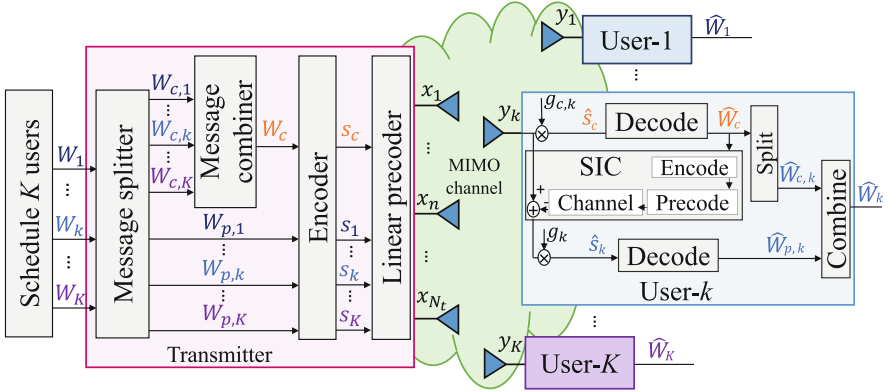


Fig. 3.4 Transmission model of K -user 1-layer RS

1-Layer RS

1-layer RS is the simplest RSMA scheme and it is the building block of the entire RSMA framework. It has been widely studied in the literature of RS in multi-antenna BC and its applications [17, 30, 34, 37, 38, 40–42, 45, 47, 50, 54, 56–59, 61] with both perfect and imperfect CSIT. Figure 3.4 illustrates the transmission model of K -user 1-layer RS with one base station (BS) equipped with N_t transmit antennas simultaneously serving K single-antenna users. The users are indexed by $\mathcal{K} = \{1, \dots, K\}$.

At the transmitter, the K messages W_1, \dots, W_K intended for the K users are passed to the message splitter. The message of each user W_k , $k \in \mathcal{K}$ is split into one common part $W_{c,k}$ and one private part $W_{p,k}$.³ The common parts $W_{c,1}, \dots, W_{c,K}$ are combined into the common message W_c and encoded into the common stream s_c to be decoded by all users. The private parts $W_{p,1}, \dots, W_{p,K}$ are independently encoded into K private streams s_1, \dots, s_K to be decoded by the corresponding users only. The encoded stream vector $\mathbf{s} = [s_c, s_1, \dots, s_K]^T \in \mathbb{C}^{(K+1) \times 1}$ is linearly precoded via precoding matrix $\mathbf{P} = [\mathbf{p}_c, \mathbf{p}_1, \dots, \mathbf{p}_K] \in \mathbb{C}^{N_t \times (K+1)}$ with $\mathbf{p}_k \in \mathbb{C}^{N_t \times 1}$, $k \in \{c\} \cup \mathcal{K}$. The resulting transmit signal is

$$\mathbf{x} = \mathbf{P}\mathbf{s} = \mathbf{p}_c s_c + \sum_{k \in \mathcal{K}} \mathbf{p}_k s_k. \quad (3.1)$$

At user sides, the signal received at each user is

³Note that it is not necessary to let all users split their messages in some cases. For example, when maximizing the sum rate without QoS rate constraint [30], one user splits its message into common and private parts which is sufficient. However, splitting the messages of all users is more general, and it becomes necessary when user fairness is considered in the design. For instance, when maximizing WSR or max–min fairness or with QoS rate constraint [17, 30, 34].

$$\begin{aligned}
y_k &= \mathbf{h}_k^H \mathbf{x} + n_k \\
&= \mathbf{h}_k^H \mathbf{p}_c s_c + \mathbf{h}_k^H \mathbf{p}_k s_k + \sum_{j \in \mathcal{K}, j \neq k} \mathbf{h}_k^H \mathbf{p}_j s_j + n_k,
\end{aligned} \tag{3.2}$$

where $\mathbf{h}_k \in \mathbb{C}^{N_t \times 1}$ is the channel between the BS and user- k . It may be perfectly known at the transmitter [42, 45, 49, 50, 53, 54, 56–59, 61] or partially known at the transmitter [17, 30, 34, 37–41, 46, 47, 57, 62] due to the quantization error or feedback delay. n_k is the additive white Gaussian noise (AWGN) at user- k that follows the distribution $\mathcal{CN}(0, \sigma_{n,k}^2)$.

Each user firstly decodes the data stream s_c by treating the interference from all private streams as noise.⁴ The signal-to-interference-pulse-noise ratio (SINR) of decoding the common stream s_c at user- k is

$$\gamma_{c,k} = \frac{|\mathbf{h}_k^H \mathbf{p}_c|^2}{\sum_{j \in \mathcal{K}} |\mathbf{h}_k^H \mathbf{p}_j|^2 + \sigma_{n,k}^2}. \tag{3.3}$$

To ensure all users can successfully decode the common stream, its achievable rate should not exceed

$$R_c = \min \{ \log_2 (1 + \gamma_{c,k}), \dots, \log_2 (1 + \gamma_{c,K}) \}. \tag{3.4}$$

Note that R_c is shared by all K users. Denote C_k as the part of rate allocated to user- k for the transmission of $W_{c,k}$, we have

$$\sum_{k \in \mathcal{K}} C_k = R_c. \tag{3.5}$$

Once s_c is successfully decoded, it is re-encoded, precoded, and subtracted from y_k . Each user then decodes its intended private stream s_k by treating the interference from the private streams of other users as noise.⁵ The SINR of decoding the private stream s_k at user- k is

⁴Please notice that the role of the common stream here is fundamentally different from a multicast stream, though both of them are decoded by all users. The common stream in RS encapsulates parts of private messages of different users. It is not entirely required by all users. In contrast, a multicast stream is encoded by a message originally intended for all users. Each user requires the full message [4].

⁵The decoding order of s_c and s_k can be further optimized. We here follow the rule that the data stream intended for more users has a higher decoding priority [63, 64] for the entire RSMA framework.

$$\gamma_k = \frac{|\mathbf{h}_k^H \mathbf{p}_k|^2}{\sum_{j \in \mathcal{K}, j \neq k} |\mathbf{h}_k^H \mathbf{p}_j|^2 + \sigma_{n,k}^2}. \quad (3.6)$$

Its corresponding private rate is $R_k = \log_2(1 + \gamma_k)$. Hence, the total achievable rate of user- k , $k \in \mathcal{K}$ is

$$R_{k,tot} = C_k + R_k. \quad (3.7)$$

Following the above-described structure of 1-layer RS, we can design the precoders $\mathbf{p}_C, \mathbf{p}_1, \dots, \mathbf{p}_K$ with different objectives, such as maximizing the WSR (or sum rate) [30, 42], maximizing the worst-case user rate [34], maximizing EE [45], minimizing transmit power [44], etc.

2-Layer HRS

2-layer HRS is originally introduced for massive MIMO networks [39] with the aim of enhancing the achievable rate of all users and reducing the CSI feedback. In a K -user 2-layer HRS network, the K users are divided into G separated groups indexed by $\mathcal{G} = \{1, \dots, G\}$ with \mathcal{K}_g , $g \in \mathcal{G}$ users in each group such that $\bigcup_{g \in \mathcal{G}} \mathcal{K}_g = \mathcal{K}$. Different from 1-layer RS where the message of each user is only split into two parts, each user in 2-layer HRS splits its message W_k , $k \in \mathcal{K}_g$ into three different parts $W_k^{\mathcal{K}}, W_k^{\mathcal{K}_g}, W_k^k$ in order to form outer-group common message and inner-group common message. The outer-group common messages $\{W_k^{\mathcal{K}} | k \in \mathcal{K}\}$ of all users are jointly combined into one common message $W_{\mathcal{K}}$ and encoded into the outer-group common stream $s_{\mathcal{K}}$ to be decoded by all users. The inner-group common messages $\{W_k^{\mathcal{K}_g} | k \in \mathcal{K}_g\}$ of users in group- g are jointly combined into the common message $W_{\mathcal{K}_g}$ and encoded into the inner-group common stream $s_{\mathcal{K}_g}$ to be decoded by all users in \mathcal{K}_g . The private messages $\{W_k^k | k \in \mathcal{K}\}$ are independently encoded into the private streams s_1, \dots, s_K for the corresponding users only. The encoded streams $\mathbf{s} = [s_{\mathcal{K}}, s_{\mathcal{K}_1}, \dots, s_{\mathcal{K}_G}, s_1, \dots, s_K]^T \in \mathbb{C}^{(K+G+1) \times 1}$ are linearly precoded via precoding matrix $\mathbf{P} = [\mathbf{p}_{\mathcal{K}}, \mathbf{p}_{\mathcal{K}_1}, \dots, \mathbf{p}_{\mathcal{K}_G}, \mathbf{p}_1, \dots, \mathbf{p}_K] \in \mathbb{C}^{N_t \times (K+G+1)}$, the shaped transmit signal is

$$\mathbf{x} = \mathbf{P}\mathbf{s} = \mathbf{p}_{\mathcal{K}}s_{\mathcal{K}} + \sum_{g \in \mathcal{G}} \mathbf{p}_{\mathcal{K}_g}s_{\mathcal{K}_g} + \sum_{k \in \mathcal{K}} \mathbf{p}_k s_k. \quad (3.8)$$

At user sides, once each user receives the signal as $y_k = \mathbf{h}_k^H \mathbf{x} + n_k$, it employs two layers of SIC to successfully decode $s_{\mathcal{K}}, s_{\mathcal{K}_g}$ and s_k , $k \in \mathcal{K}_g$. The outer-group common stream $s_{\mathcal{K}}$ is decoded first at all users by treating the interference from all other streams as noise. The corresponding SINR of decoding $s_{\mathcal{K}}$ at user- k is

$$\gamma_k^{\mathcal{K}} = \frac{|\mathbf{h}_k^H \mathbf{p}_{\mathcal{K}}|^2}{\sum_{g \in \mathcal{G}} |\mathbf{h}_k^H \mathbf{p}_{\mathcal{K}_g}|^2 + \sum_{j \in \mathcal{K}} |\mathbf{h}_k^H \mathbf{p}_j|^2 + \sigma_{n,k}^2}. \quad (3.9)$$

Once $s_{\mathcal{K}}$ is successfully decoded with its contributed part removed from the received signal, each user then decodes the inner-group common stream $s_{\mathcal{K}_g}$ by treating interference from other inner-group common streams and private streams as noise. The SINR of decoding $s_{\mathcal{K}_g}$ at user- k is

$$\gamma_k^{\mathcal{K}_g} = \frac{|\mathbf{h}_k^H \mathbf{p}_{\mathcal{K}_g}|^2}{\sum_{g' \in \mathcal{G}, g' \neq g} |\mathbf{h}_k^H \mathbf{p}_{\mathcal{K}_{g'}}|^2 + \sum_{j \in \mathcal{K}} |\mathbf{h}_k^H \mathbf{p}_j|^2 + \sigma_{n,k}^2}. \quad (3.10)$$

After removing $s_{\mathcal{K}_g}$ from the received signal, user- k decodes its private stream s_k . The SINR of decoding the private stream s_k at user- k is

$$\gamma_k = \frac{|\mathbf{h}_k^H \mathbf{p}_k|^2}{\sum_{g' \in \mathcal{G}, g' \neq g} |\mathbf{h}_k^H \mathbf{p}_{\mathcal{K}_{g'}}|^2 + \sum_{j \in \mathcal{K}, j \neq k} |\mathbf{h}_k^H \mathbf{p}_j|^2 + \sigma_{n,k}^2}. \quad (3.11)$$

Following (3.4) and (3.5), we obtain the respective achievable rate of $s_{\mathcal{K}}$, $s_{\mathcal{K}_g}$, and s_k , which are given by

$$\begin{aligned} \sum_{k \in \mathcal{K}} C_k^{\mathcal{K}} &= \min \left\{ \log_2 \left(1 + \gamma_{k'}^{\mathcal{K}} \right) \mid k' \in \mathcal{K} \right\}, \\ \sum_{k \in \mathcal{K}_g} C_k^{\mathcal{K}_g} &= \min \left\{ \log_2 \left(1 + \gamma_{k'}^{\mathcal{K}_g} \right) \mid k' \in \mathcal{K}_g \right\}, \forall g \in \mathcal{G} \\ R_k &= \log_2 (1 + \gamma_k), \forall k \in \mathcal{K}. \end{aligned} \quad (3.12)$$

where $C_k^{\mathcal{K}}$ and $C_k^{\mathcal{K}_g}$ are the parts of the rate allocated to user- k for the transmission of messages $W_k^{\mathcal{K}}$ and $W_k^{\mathcal{K}_g}$, respectively. The total achievable rate of user- k , $k \in \mathcal{K}_g$ is

$$R_{k,tot} = C_k^{\mathcal{K}} + C_k^{\mathcal{K}_g} + R_k. \quad (3.13)$$

Figure 3.5 illustrates the transmission model of four-user 2-layer HRS with one BS equipped with N_t transmit antennas simultaneously serving four single-antenna users. There are two user groups with user-1 and user-2 in group 1 and user-3 and user-4 in group 2. s_{1234} is an outer-group common stream to be decoded by all the four users, while s_{12} and s_{34} are the two inner-group common streams to be decoded by the users within the corresponding groups only. The receiver structures of user-2 and user-4 follow that of user-1 and user-2, respectively.

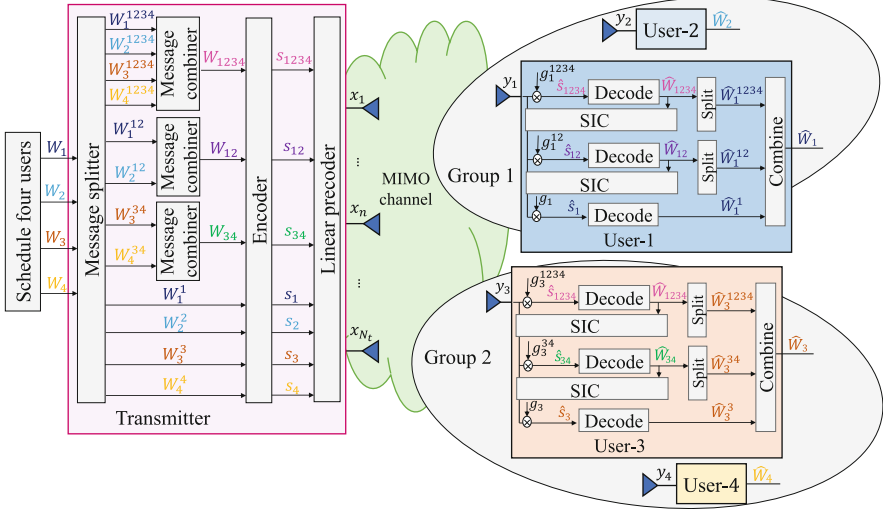


Fig. 3.5 Transmission model of four-user 2-layer HRS

Generalized RS

The generalized RS framework is proposed in [42] with the aim of identifying the largest room for rate and QoS enhancement at the expense of more layers of SIC at each user. In the K -user generalized RS framework, the number of message splits of each user increases with K so as to form common streams intended for different user subsets of \mathcal{K} . For any user subset $\mathcal{A} \subseteq \mathcal{K}$, the BS transmits a data stream $s_{\mathcal{A}}$ by loading messages of all users in the subset \mathcal{A} , and $s_{\mathcal{A}}$ needs to be decoded by all users in the subset \mathcal{A} while treated as noise by other users. The message of user- k is split into 2^{K-1} parts as $\{W_k^{\mathcal{A}'} | \mathcal{A}' \subseteq \mathcal{K}, k \in \mathcal{A}'\}$. User messages $\{W_k^{\mathcal{A}'} | k' \in \mathcal{A}\}$ with the same superscript \mathcal{A} are encoded together into the stream $s_{\mathcal{A}}$.

The concept of *stream order* is introduced here to simplify the explanation. We define the streams to be decoded by l users as l -order streams. Hence, the common stream $s_{\mathcal{K}}$ intended for all users is a K -order stream, while the private stream s_k is a 1-order stream since it is only decoded by a single user. In the K -user case, all l -order streams form the stream set $\{s_{\mathcal{A}'} | \mathcal{A}' \subseteq \mathcal{K}, |\mathcal{A}'| = l\}$, and there are in total $\binom{K}{l}$ elements within the set. Specifically, there is one K -order stream $s_{\mathcal{K}}$ and K 1-order streams s_1, \dots, s_K . We further introduce l -order data stream vector formed by all l -order streams as $\mathbf{s}_l \in \mathbb{C}^{\binom{K}{l} \times 1}$. Note that when $l = K$, there is one element within the set, and therefore, \mathbf{s}_K reduces to $s_{\mathcal{K}}$. \mathbf{s}_l is then linearly precoded via the precoding matrix \mathbf{P}_l formed by $\{\mathbf{p}_{\mathcal{A}'} | \mathcal{A}' \subseteq \mathcal{K}, |\mathcal{A}'| = l\}$, and the resulting transmit signal is

$$\mathbf{x} = \sum_{l=1}^K \mathbf{P}_l \mathbf{s}_l = \sum_{l=1}^K \sum_{\mathcal{A}' \subseteq \mathcal{K}, |\mathcal{A}'|=l} \mathbf{p}_{\mathcal{A}' s_{\mathcal{A}'}}. \quad (3.14)$$

At user sides, each user requires $2^{K-1} - 1$ layers of SIC to sequentially decode all the intended common streams. The decoding process starts from the K -order stream and then goes down to the 1-order private stream. Note that each user is involved in multiple l -order streams except the 1-order and K -order streams, and the set of l -order streams to be decoded at user- k is $\mathcal{S}_{l,k} = \{s_{\mathcal{A}'} | \mathcal{A}' \subseteq \mathcal{K}, |\mathcal{A}'| = l, k \in \mathcal{A}'\}$. We denote the decoding order of the l -order streams s_l at all users as π_l . Based on one certain decoding order π_l , we obtain the l -order stream vector to be decoded at user- k as $\mathbf{s}_{\pi_l,k} = [s_{\pi_l,k(1)}, \dots, s_{\pi_l,k(|\mathcal{S}_{l,k}|)}]^H$, where we assume $s_{\pi_l,k(i)}$ is decoded before $s_{\pi_l,k(j)}$ if $i < j$. The SINR of user- k to decode the l -order stream $s_{\pi_l,k(i)}$ is

$$\gamma_k^{\pi_l,k(i)} = \frac{|\mathbf{h}_k^H \mathbf{p}_{\pi_l,k(i)}|^2}{I_{\pi_l,k(i)} + \sigma_{n,k}^2}, \quad (3.15)$$

where

$$I_{\pi_l,k(i)} = \sum_{j>i} |\mathbf{h}_k^H \mathbf{p}_{\pi_l,k(j)}|^2 + \sum_{l'=1}^{l-1} \sum_{j=1}^{|\mathcal{S}_{l',k}|} |\mathbf{h}_k^H \mathbf{p}_{\pi_{l',k}(j)}|^2 + \sum_{\mathcal{A}' \subseteq \mathcal{K}, k \notin \mathcal{A}'} |\mathbf{h}_k^H \mathbf{p}_{\mathcal{A}'}|^2$$

is the interference received at user- k when decoding $s_{\pi_l,k(i)}$. The first term $\sum_{j>i} |\mathbf{h}_k^H \mathbf{p}_{\pi_l,k(j)}|^2$ is the interference from the remaining non-decoded l -order streams in $\mathbf{s}_{\pi_l,k}$. The second term $\sum_{l'=1}^{l-1} \sum_{j=1}^{|\mathcal{S}_{l',k}|} |\mathbf{h}_k^H \mathbf{p}_{\pi_{l',k}(j)}|^2$ is the interference from lower-order streams $\{s_{\pi_{l',k}} | l' < l\}$ to be decoded at user- k , while the third term $\sum_{\mathcal{A}' \subseteq \mathcal{K}, k \notin \mathcal{A}'} |\mathbf{h}_k^H \mathbf{p}_{\mathcal{A}'}|^2$ is the interference received from the streams that are not intended for user- k . The corresponding achievable rate of user- k for the data stream $s_{\pi_l,k(i)}$ is $R_k^{\pi_l,k(i)} = \log_2(1 + \gamma_k^{\pi_l,k(i)})$. Following (3.4), (3.5), and (3.12), the achievable rate of the $|\mathcal{A}|$ -order stream $s_{\mathcal{A}}$ ($\mathcal{A} \in \mathcal{K}$, $2 \leq |\mathcal{A}| \leq K$) shall not exceed

$$\sum_{k \in \mathcal{A}} C_k^{\mathcal{A}} = \min_{k'} \left\{ R_{k'}^{\mathcal{A}} \mid k' \in \mathcal{A} \right\}. \quad (3.16)$$

where $C_k^{\mathcal{A}}$ is the part of the common rate allocated to user- k ($k \in \mathcal{A}$) for the transmission of $W_k^{\mathcal{A}}$ via $s_{\mathcal{A}}$. Hence, the total achievable rate of user- k is

$$R_{k,tot} = \sum_{\mathcal{A}' \subseteq \mathcal{K}, k \in \mathcal{A}'} C_k^{\mathcal{A}'} + R_k. \quad (3.17)$$

Figure 3.6 illustrates a three-user example of the generalized RS. The message of each user is split into four parts, i.e., the message of user-1 is split into $\{W_1^{123}$,

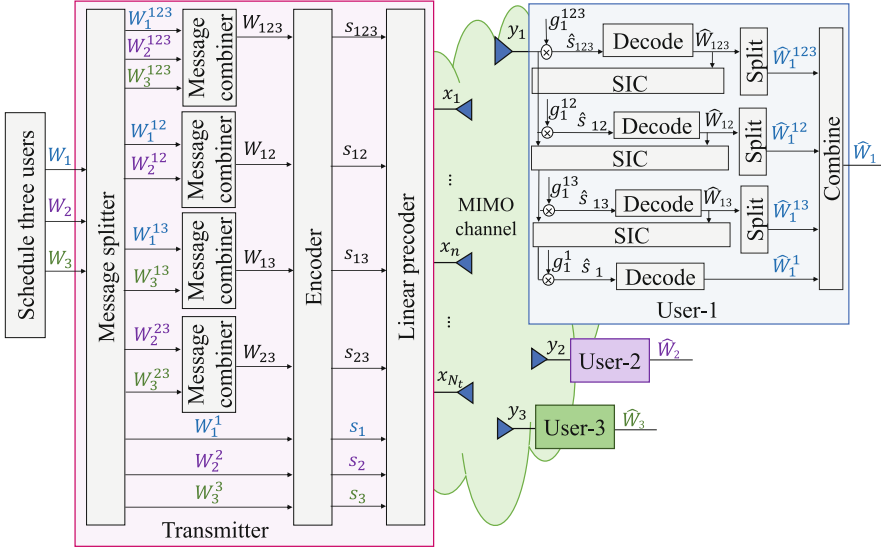


Fig. 3.6 Transmission model of three-user generalized RS

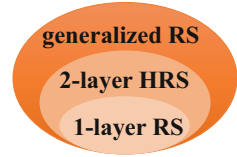
$W_1^{12}, W_1^{13}, W_1^1$). There is one 3-order stream, three 2-order streams, and three 1-order streams. The corresponding stream vectors are denoted as $\mathbf{s}_1 = [s_1, s_2, s_3]^T$, $\mathbf{s}_2 = [s_{12}, s_{13}, s_{23}]^T$, and s_{123} , respectively. Each user requires three layers of SIC to sequentially decode the intended streams. The instance of decoding order π_2 for the 2-order streams illustrated in Fig. 3.6 is $12 \rightarrow 13 \rightarrow 23$. All users follow the rule that s_{12} is decoded before s_{13} and s_{23} is decoded lastly. At user-1, only the 2-order streams s_{12} and s_{13} are decoded. The decoding order based on π_2 at user-1 is $\pi_{2,1} = 12 \rightarrow 13$. We have $s_{\pi_{2,1}(1)} = s_{12}$ and $s_{\pi_{2,1}(2)} = s_{13}$.

1-Layer RS vs. 2-Layer HRS vs. Generalized RS

The inclusive relation of the above three RSMA schemes is illustrated in Fig. 3.7. The generalized RS is the most general scheme that embraces 2-layer HRS and 1-layer RS as two sub-schemes. 2-layer HRS is a sub-scheme when only the K -order stream, $|\mathcal{K}_g|$ -order streams, and 1-order streams are active (with a nonzero power allocation) in the generalized RS, while 1-layer RS is the sub-scheme of 2-layer HRS when only the K -order stream and 1-order streams are active. All other inactive streams are allocated with zero transmit power.

In terms of the computational complexity and hardware complexity at the BS and users, 1-layer RS achieves the lowest complexity, and the generalized RS has the opposite highest complexity in the RSMA framework. In the K -user case, each user in the 1-layer RS system only requires one layer of SIC without any scheduling requirement at the transmitter, while each user requires two layers of SIC and the

Fig. 3.7 RSMA framework and the schemes included



BS requires to consider the issue of user grouping in 2-layer HRS. Both 1-layer RS and 2-layer HRS maintain relative low transceiver complexities and are practical for implementation since the number of SIC layers deployed at each user is independent from the number of users. In comparison, the generalized RS is more complex to be implemented since the number of SIC layers increases rapidly with the number of users. In the K -user case, $2^{K-1} - 1$ layers of SIC are required at each user, and the decoding order of the common streams needs to be optimized at the transmitter. However, readers are reminded that the motivation of introducing the generalized RS contrasts with the previous two low-complex schemes, that is, to identify the best possible performance of the network at the scarifies of more SIC layers at the receivers.

3.2.3 RSMA vs. NOMA/SDMA/OMA

Framework Comparison

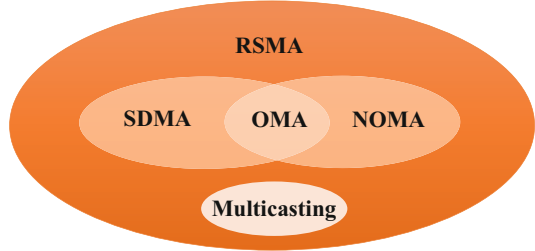
Table 3.3 summarizes the comparison of different multiple access techniques. Compared with the existing multiple access techniques, the major and unique characteristic of RSMA is its ability of partially decoding interference and partially treating interference as noise. Gaining benefits from its dynamic interference management capability, RSMA framework generalizes and encompasses multi-antenna NOMA (including SC-SIC and SC-SIC per group), SDMA (based on MU-LP), OMA (TDMA/FDMA), and multicasting as sub-schemes. Their relation is further illustrated in Fig. 3.8.

SDMA is a sub-scheme of RSMA when all common streams are turned off and the transmit power is fully allocated to the private streams. RSMA boils down to NOMA (based on SC-SIC) when each common stream is fully encoded by the entire message of a single user. OMA is a sub-scheme of SDMA, NOMA, and RSMA when the transmit power is fully allocated to a single user. Physical-layer multicasting is achieved by encoding the messages of all users into the K -order stream. Hence, multicasting is a special instance of RSMA with full transmit power being allocated to the K -order common stream (embracing partial messages of all users). Based on the above discussion, we obtain that SDMA based on MU-LP, multi-antenna NOMA, OMA, and multicasting are sub-schemes of RSMA. Most importantly, RSMA smoothly bridges all sub-schemes without hard switching among them.

Table 3.3 Comparison of different multiple access techniques

Multiple access Strategy	OMA	SDMA	NOMA	RSMA
Design principle	LP Orthogonal resource allocation to get rid of interference	MU-LP Fully treat interference as noise	SC-SIC Fully decode interference	All forms of RS Partially decode interference and partially treat interference as noise
Decoder architecture	Treat interference as noise	Treat interference as noise	SIC at receivers	SIC at receivers
User deployment scenario	Any angle between channels and any disparity in channel strengths	User channels are (semi-)orthogonal with similar channel strengths	Users experience aligned channel directions and a large disparity in channel strengths	Any angle between channels and any disparity in channel strengths
Network load	Only one active user (in each resource block)	More suited to overloaded network	More suited to overloaded network	Suited to any network load

Fig. 3.8 RSMA framework and the schemes included



	s_1	s_2	s_c
SDMA	W_1	W_2	-
NOMA	W_1	-	W_2
OMA	W_1	-	-
Multicasting	-	-	W_1, W_2
RS	$W_{p,1}$	$W_{p,2}$	$W_{c,1}, W_{c,2}$

decoded by its intended user and treated as noise by the other user
decoded by both users

Fig. 3.9 Mapping of messages to streams

Figure 3.9 illustrates the different mappings of the messages to the streams in the two-user case for all multiple access techniques. When $K = 2$, the generalized RS boils down to 1-layer RS automatically with one common stream s_c containing one part of message $W_{c,1}$ for user-1 and one part of message $W_{c,2}$ for user-2. Other parts $W_{p,1}$ and $W_{p,2}$ are independently encoded into private streams s_1 and s_2 . SDMA is obtained by allocating no power to the common stream ($\|\mathbf{p}_c\|^2 = 0$) such that W_k is encoded into s_k directly. NOMA is obtained by encoding the message of one user, i.e., W_2 entirely into s_c and W_1 into s_1 while s_2 is turned off ($\|\mathbf{p}_2\|^2 = 0$). In this example, user-1 fully decodes the interference from the message of user-2. OMA is obtained when only one user is scheduled ($\|\mathbf{p}_c\|^2 = \|\mathbf{p}_2\|^2 = 0$). Multicasting is obtained when the messages of both users W_1, W_2 are combined into s_c and the private streams are turned off ($\|\mathbf{p}_1\|^2 = \|\mathbf{p}_2\|^2 = 0$).

Complexity Comparison

The qualitative complexity of different strategies is compared in Table 3.4. SDMA based on MU-LP and OMA based on point-to-point linear precoding have the lowest receiver and encoder complexities. However, the scheduling complexity is relatively high due to the subcarrier/time-slot allocation for OMA and user selection for SDMA. As mentioned previously, SDMA based on MU-LP is only suited when the user channels are semi-orthogonal. Accurate CSIT is required to carefully design user scheduling for interference coordination.

Both SC-SIC per group and 2-layer HRS have the highest user grouping complexity. The total number of user grouping methods to be considered in both

Table 3.4 Qualitative comparison of the complexity of different multiple access techniques

Multiple access	RSMMA						
Strategy	OMA	SDMA	NOMA	1-layer RS	2-layer HRS	generalized RS	
Encoder complexity	LP Encode K streams	MU-LP Encode K streams	SC-SIC Encode K streams	SC-SIC per group Encode K streams	Encode $K + 1$ streams	Encode $K + G + 1$ streams	encode $2^K - 1$ streams
Scheduler complexity	Complex as OMA requires subcarrier/time-slot allocation to all users	Complex as MU-LP requires to pair together semi-orthogonal users with similar channel gains	Very complex as it requires to find aligned users and decide upon $K!$ user ordering	Very complex as it requires to divide users into orthogonal groups, with aligned users in each group. Decide upon $\sum_{k=1}^K S(K, k)$ grouping method and at most $K!$ decoding order for each grouping method	Simpler user scheduling as it copes with any user deployment scenario, does not rely on user grouping or user ordering	Decide upon $\sum_{k=1}^K S(K, k)$ grouping method without decoding order problem	Complex as it requires to decide upon $\prod_{k=2}^{K-1} \binom{K}{k}!$ decoding orders
Receiver complexity	Does not require any SIC	Does not require any SIC	Requires $K - 1$ layers of SIC. Subject to error propagation	Requires $ K_g - 1$ layers of SIC in each group. Subject to error propagation	Requires one layer of SIC at each user. Less subject to error propagation	Requires two layers of SIC at each user. Less subject to error propagation	Requires $2^{K-1} - 1$ layers of SIC. Subject to error propagation

schemes is $\sum_{k=1}^K S(K, k)$, where $S(K, k) = \frac{1}{k!} \sum_{i=0}^k (-1)^i \binom{k}{i} (k-i)^K$, also known as a Stirling set number [65], is the total number of methods to partition a set of K elements into k nonempty sets. As each user in 2-layer HRS sequentially decodes the outer-group common stream and the inner-group common stream followed by the intended private stream, the decoding order is determined without introducing additional scheduling complexity and only requires two layers of SIC at each user. In comparison, at most $K!$ decoding orders are required to be considered in SC–SIC per group for each grouping method, and each user is required to have $|\mathcal{K}_g| - 1$ layers of SIC. For example, for a four-user system with two groups and two users in each group, we have to consider three different user grouping methods and four different decoding orders for each grouping method. Generally, SC–SIC per group has the highest scheduling complexity compared with other schemes since the decoding order and user grouping are required to be jointly decided. Note that $K!$ is the total number of decoding orders when there is one user group. In such scenario, SC–SIC per group reduces to SC–SIC. Different from the single-antenna NOMA in SISO BC where the optimal decoding order of NOMA is determined based on the channel gain, multi-antenna NOMA based on SC–SIC requires the decoding order to be jointly decided with the precoders at the transmitter. As SC–SIC is only suited for aligned user channels with certain channel strength disparities, additional scheduler complexity is introduced for a proper user scheduling algorithm. Hence, the scheduler complexity of SC–SIC is relatively high, and each user requires $K - 1$ layers of SIC in the K -user SC–SIC system. Compared with SC–SIC per group, SC–SIC simplifies the scheduling complexity at the transmitter (since there is no requirement of user grouping) but increases the receiver complexity.

Compared with existing multiple access techniques, RSMA is able to achieve a better trade-off between performance and complexity. All RS strategies including 1-layer RS, 2-layer HRS, and generalized RS are suited for users with any channel strength disparity and channel angle in between. Specifically, 1-layer RS has the lowest scheduling complexity compared with all other schemes since it does not have any issue of user scheduling, grouping, and ordering. It also maintains very low receiver complexity since only one layer of SIC is required at each user in the K -user scenario. 1-layer RS is a sub-scheme of 2-layer HRS and the generalized RS. Compared with 1-layer RS, the complexity at the transmitter and receivers for 2-layer HRS is higher due to a higher dimensional message splits. The receiver complexity of 2-layer HRS is still low compared with other schemes since the number of SIC layers required at 1-layer RS and 2-layer HRS is independent from the number of user K . The receiver complexity is much reduced compared with SC–SIC or SC–SIC per group or the generalized RS. Though the generalized RS achieves the highest flexibility of interference management compared with all other schemes, it has a higher transmitter and receiver complexity. The generalized RS requires the decoding order of multiple streams with the same stream order to be jointly decided with the precoders, and each user requires an exponentially increasing number of SIC layers to decode the intended streams sequentially. For

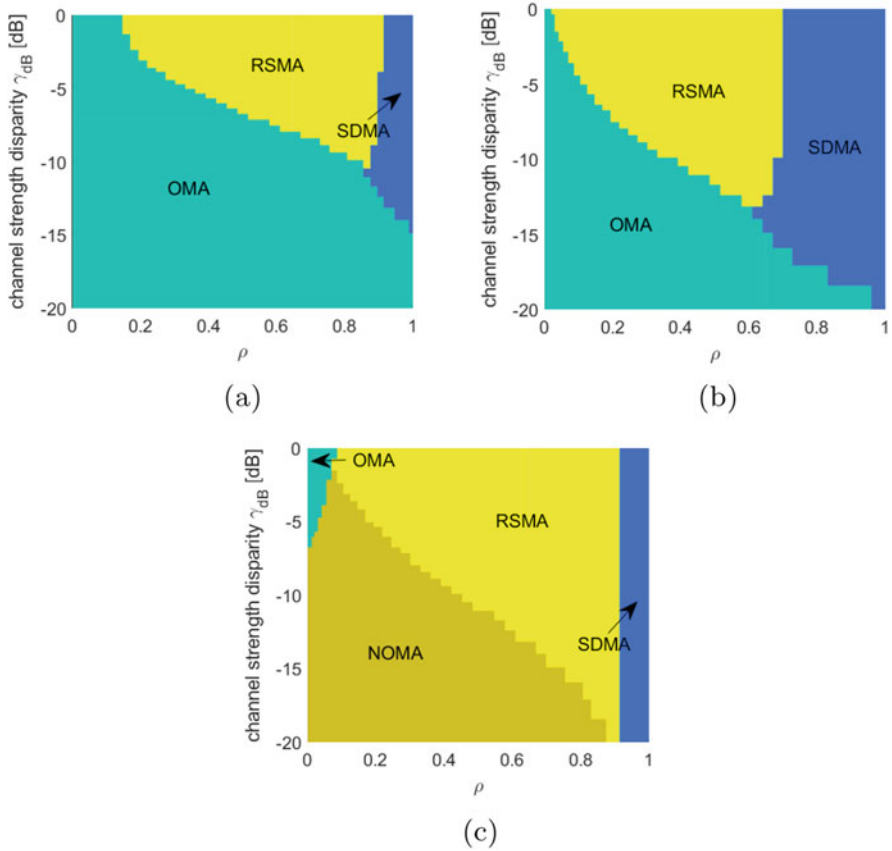


Fig. 3.10 Regions of operation for different multiple access techniques, $K = 2$, SNR = 20 dB, $\epsilon = 0.01$. (a) $u_1 = 10^{0.5}$, $u_2 = 1$. (b) $u_1 = 1$, $u_2 = 1$. (c) $u_1 = 1$, $u_2 = 10^{0.5}$

example, each user requires to decode two 2-order streams in Fig. 3.6, and in total, three layers of SIC are required at each user.

The multi-antenna NOMA and the generalized RS schemes have a number of SIC layers increasing with the number of user K , which not only lead to an increase of the scheduler and receiver complexity but also impel more error propagation in SIC. They are preferred to be applied in the scenarios when K is small so as to achieve a better trade-off between the performance improvement and transmitter/receiver complexity.

Performance Comparison

Figure 3.10 illustrates the preferred regions for the operation of OMA, SDMA, NOMA, and RSMA with perfect CSIT. Following the evaluations in [58], we

assume that the BS equipped with $N_t = 2$ transmit antennas is serving two single-antenna users ($K = 2$). The channel vectors are $\mathbf{h}_1 = 1/\sqrt{2}[1, 1]^H$ and $\mathbf{h}_2 = \gamma/\sqrt{2}[1, e^{j\theta}]^H$. As there are only two users, SC-SIC per group boils down to SC-SIC, and the generalized RS boils down to 1-layer RS. The precoders are optimized based on the weighted minimum mean square error (WMMSE) precoding optimization framework developed in [30, 42, 66] with the aim of maximizing the sum rate $\sum_{k=1,2} R_{k,tot}$. The total achievable rate of user- k for RS is given as (3.7). The sum rate formulas of SDMA and NOMA are illustrated in [42]. The colors in Fig. 3.10 illustrate the strategy that achieves the maximized WSR as a function of $\rho = 1 - \frac{|\mathbf{h}_1^H \mathbf{h}_2|^2}{\|\mathbf{h}_1\|^2 \|\mathbf{h}_2\|^2}$ (ranging from 0 to 1) and $\gamma_{dB} = 20 \log_{10}(\gamma)$ (ranging from 0 to -20 dB), i.e., user-1 and user-2 have a long-term SNR of 20dB and $0\text{dB} \leq 20\text{dB} + \gamma_{dB} \leq 20\text{dB}$, respectively. As the WSR of RSMA is always larger than or equal to that of other strategies, we follow the rules below to select the strategy:

- (i) if $|\text{WSR}_{\text{RSMA}} - \text{WSR}_{\text{OMA}}| < \epsilon$, the preferred strategy is OMA.
- (ii) if $|\text{WSR}_{\text{SDMA}} - \text{WSR}_{\text{OMA}}| > \epsilon$ and $|\text{WSR}_{\text{RSMA}} - \text{WSR}_{\text{SDMA}}| < \epsilon$, the preferred strategy is SDMA.
- (iii) if $|\text{WSR}_{\text{NOMA}} - \text{WSR}_{\text{SDMA}}| > \epsilon$ and $|\text{WSR}_{\text{RSMA}} - \text{WSR}_{\text{NOMA}}| < \epsilon$, the preferred strategy is NOMA.
- (iv) if $|\text{WSR}_{\text{RSMA}} - \text{WSR}_{\text{SDMA}}| > \epsilon$ and $|\text{WSR}_{\text{RSMA}} - \text{WSR}_{\text{NOMA}}| > \epsilon$, the preferred strategy is RSMA.

Option (iv) is selected when RSMA does not boils down to any other multiple access techniques. We observe from the figure that when equal or higher weight is allocated to the user with a stronger channel, NOMA has no benefit over SDMA at all. Only when the user fairness is taken into consideration with a higher weight allocated to the weaker user, NOMA outperforms SDMA. But NOMA is only preferred for the deployment with small ρ , i.e., users are closely aligned. SDMA is preferred whenever ρ is sufficiently large. In comparison, for all different user weights, RSMA always provides the same or better performance than SDMA, NOMA, and OMA. It unifies and outperforms existing multiple access techniques.

Figure 3.11 further illustrates the ergodic rate region of different multiple access techniques over 100 random channel realizations with imperfect CSIT. The BS is equipped with $N_t = 2$ antennas and serves two single-antenna users. The channel model specified in [30, 34] is adopted, i.e., $\mathbf{h}_k = \hat{\mathbf{h}}_k + \tilde{\mathbf{h}}_k$. The estimated channel of each user $\hat{\mathbf{h}}_k$ and channel error $\tilde{\mathbf{h}}_k$ have independent and identically distributed (i.i.d.) complex Gaussian entries that follow the distributions $\mathcal{CN}(0, \sigma_k^2)$ and $\mathcal{CN}(0, \sigma_{e,k}^2)$, respectively. The variance of error $\sigma_{e,k}^2$ scales exponentially with SNR as $\sigma_{e,k}^2 \sim \mathcal{O}(P_t^{-\alpha})$, where $\alpha \in [0, \infty)$ is interpreted as the quality of CSIT in the high SNR regime [5, 28–30, 67]. The rate region improvement of RSMA over NOMA and SDMA is significant in all subfigures. Thanks to its flexible interference management capability, RSMA is more robust to CSIT inaccuracy and channel strength disparities between the users. In contrast, NOMA is only suited when there is a certain channel strength disparity between the two users, while SDMA is suited

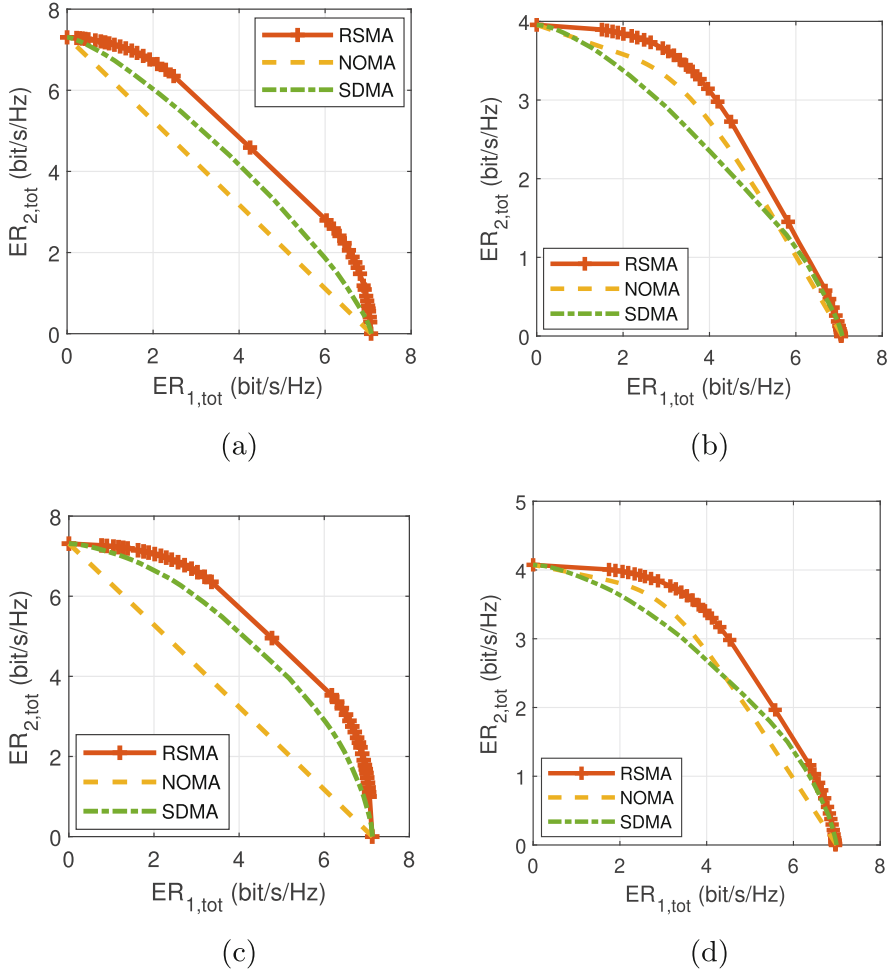


Fig. 3.11 Ergodic rate region comparison of different multiple access techniques with partial CSIT, $K = 2$, SNR = 20 dB. (a) $\alpha = 0.6, \sigma_2^2 = 1$. (b) $\alpha = 0.6, \sigma_2^2 = 0.09$. (c) $\alpha = 0.9, \sigma_2^2 = 1$. (d) $\alpha = 0.9, \sigma_2^2 = 0.09$

when users have equal channel strength. Moreover, the performance of SDMA drops as CSIT becomes inaccurate.

We further consider the three-user case. The generalized RS does not reduce to 1-layer RS, and SC-SIC per group does not reduce to SC-SIC. Figure 3.12 illustrates the ergodic sum rate versus CSIT accuracy α of different strategies over 100 random channel realizations with imperfect CSIT. Figure 3.12(a) considers an underloaded regime, while Fig. 3.12b, c shows the results of an overloaded regime but with different channel strength disparities among users. The precoders are designed to maximize the ergodic sum rate (where users have equal weights)

subject to a QoS rate constraint of each user. For $\alpha = [0.2, 0.4, 0.6, 0.8, 1]$, the corresponding rate constraint for user- k ($k \in \{1, 2, 3\}$) changes as $\mathbf{r}_k^{th} = [0.1, 0.2, 0.3, 0.4, 0.5]$ bit/s/Hz. In all subfigures, the ergodic sum rate of SC-SIC and MU-LP drops dramatically as α decreases. In contrast, the generalized RS further boosts the system performance and achieves explicit rate gain over all other strategies especially when CSIT is severely inaccurate or in the overloaded regime.

In an extremely overloaded scenario, we further show the WSR improvement of 1-layer RS with a much lower receiver complexity compared with SC-SIC in Fig. 3.13. The BS is equipped with two antennas and serves ten users. The rate of each user is averaged over the ten randomly generated channels. As SNR increases as $[0, 5, \dots, 30]$ dB, the QoS rate constraint of each user increases as

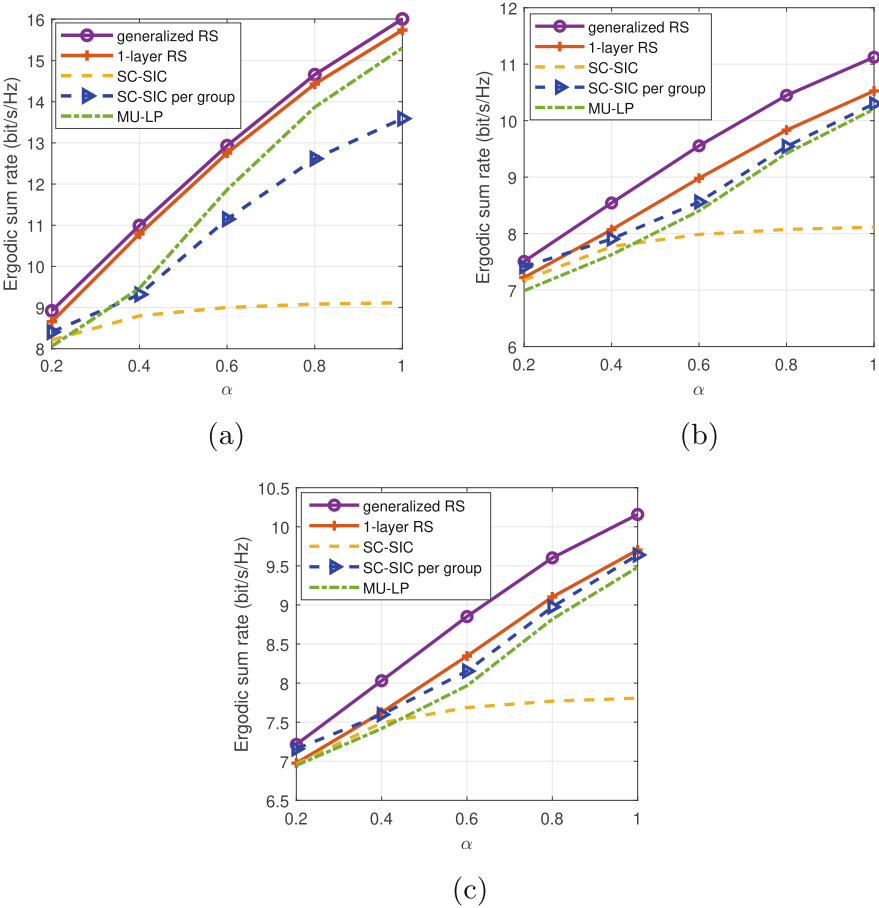


Fig. 3.12 Ergodic sum rate versus CSIT inaccuracy comparison of different multiple access techniques, averaged over 100 random channel realizations, $K = 3$, SNR = 20 dB. (a) $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = 1$, $N_t = 4$. (b) $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = 1$, $N_t = 2$. (c) $\sigma_1^2 = \sigma_2^2 = 1$, $\sigma_3^2 = 0.09$, $N_t = 2$

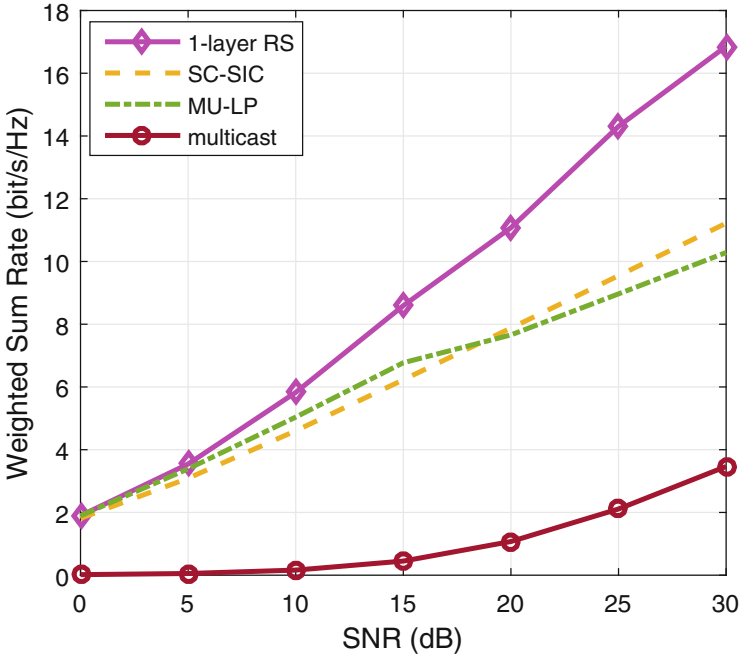


Fig. 3.13 Weighted sum rate versus SNR comparison of different multiple access techniques for overloaded ten-user deployment with perfect CSIT. $\sigma_1^2 = 1, \sigma_2^2 = 0.9, \dots, \sigma_{10}^2 = 0.1, N_t = 2$

[0, 0.001, 0.004, 0.01, 0.03, 0.06, 0.1] bit/s/Hz. We observe that 1-layer RS exhibits explicit WSR improvement over all other strategies. It achieves a sum-DoF of 2 with only a single layer of SIC deployed at each user. In contrast, the slopes of the WSRs of SC-SIC and MU-LP are the same and smaller than 1-layer RS. It implies that SC-SIC and MU-LP achieve a sum-DoF of 1. However, SC-SIC requires nine layers of SIC at each user. RS is able to exploit the largest DoF in such overloaded deployment by using the common stream to pack messages from eight users while using the two private streams to serve the remaining two users. In contrast, SC-SIC and MU-LP allocate most of power to a single user, which limits their achievable DoF.

3.2.4 Advantages of RSMA

Based on the above comparison from framework, complexity, and performance aspects, we here summarize the major advantages of RSMA:

- **Universal:** RSMA is a more general multiple access framework that outperforms and unifies OMA, SDMA based on MU-LP, and multi-antenna NOMA as sub-schemes.
- **Flexible:** RSMA is suited to all user deployments (with a diversity of channel directions, channel strengths) and network loads (underloaded and overloaded regimes). It implies that RSMA is capable of managing all different kinds of interference flexibly. RSMA automatically reduces to other multiple access techniques according to the channel conditions, i.e., it reduces to SDMA when user channels are orthogonal in the underloaded MISO BC with perfect CSIT. When the channels are aligned with certain channel strength disparities, it automatically boils down to NOMA. For other channel conditions, RS takes advance to the common streams and achieves a better interference management by partially decoding the interference and partially treating the remaining interference as noise.
- **Robust:** RSMA is robust to CSIT inaccuracy. As RSMA is primarily motivated by multi-antenna deployments with multiuser interference coming from imperfect CSIT, it compensates the DoF loss of other multiple access techniques in imperfect CSIT and is therefore less sensitive to CSIT inaccuracy.
- **Spectrally efficient:** The spectral efficiency of RSMA is always larger than or equal to that of existing multiple access techniques. Considering a MISO BC without QoS constraints, the rate region of RSMA comes much closer to the optimal DPC region than SDMA and NOMA when CSIT is perfect. When CSIT becomes imperfect CSIT, linearly precoded RSMA is able to achieve a larger rate region than complex DPC in multi-antenna BC. As RSMA achieves the optimal DoF in both perfect and imperfect CSIT, it optimally exploits the spatial dimensions and the availability of CSIT. This contrasts with SDMA and NOMA that are suboptimal.
- **Energy efficient:** As RSMA is more general than SDMA and NOMA, its energy efficiency is also larger than or equal to that of existing multiple access techniques in a wide range of user deployments.
- **Enhancing QoS and fairness:** RSMA exhibits a more explicit performance gain over other multiple access techniques when there is a QoS rate constraint for each user or when a higher weight is allocated to the user with a weaker channel condition. Therefore, the ability of a wireless network architecture to partially decode interference and partially treat interference as noise leads to enhanced QoS and user fairness.
- **Reducing complexity:** The performance gain of RSMA can come with a lower transmitter and receiver complexity than multi-antenna NOMA. In contrast to multi-antenna NOMA that requires user grouping, ordering, and switching (between NOMA and SDMA) at the transmit scheduler and multiple layers of SIC at the receivers, 1-layer RS without any user ordering, grouping, or dynamic switching at the transmit scheduler and with only one layer of SIC at each receiver is capable of achieving significant performance gain over NOMA (as illustrated in Fig. 3.13). In contrast to SDMA that requires user pairing to pair users with semi-orthogonal channels, RSMA is suited to all channel conditions,

and it does not require complex user scheduling and pairing. Moreover, RSMA is capable of further reducing CSI feedback overhead [41, 68] in the presence of quantized feedback.

3.3 Emerging Applications of RSMA

RSMA is originally proposed for MIMO BC in cellular communication networks. Recently, the applications of RSMA in other 5G technologies-enabled networks have attracted substantial interests. In massive MIMO system, 2-layer HRS proposed in [39] has been shown to achieve superior sum rate performance over conventional two-tier precoding schemes based on SDMA [69–71], and 1-layer RS has been shown to be a more robust strategy for massive MIMO in the presence of phase and amplified thermal noise since its sum rate does not saturate at high SNR [40]. The application of RSMA in the multigroup multicasting system has been shown to boost the DoF in the high SNR regime as well as to enhance the system performance in the low SNR regime [17]. By using one common stream to encapsulate parts of the multicast messages for different multicast groups, RSMA based on 1-layer RS enables the ability of partially decoding the interference and partially treating the interference as noise. Recent researches have shown that RSMA is more energy efficient in the multicell multigroup systems [43] as well as enhancing the user fairness in the multicarrier multigroup multicast systems [59]. In mmWave MIMO communication systems, the authors in [41] employ 1-layer RS and propose a one-stage feedback scheme which effectively reduces the complexity of the signaling and feedback procedure. The benefits of RSMA have been further discovered in other applications such as non-orthogonal unicast and multicast transmission (NOUM) [57], coordinated multipoint (CoMP) [49], cloud-radio access networks (C-RAN) [52, 53], simultaneous wireless information and power transfer (SWIPT) [54, 55], cooperative relaying [56, 61], wireless caching [37, 72], and unmanned aerial vehicle (UAV)-aided wireless communications [50, 51], which will not be specified here. Motivated by the benefits of RSMA discovered in cellular

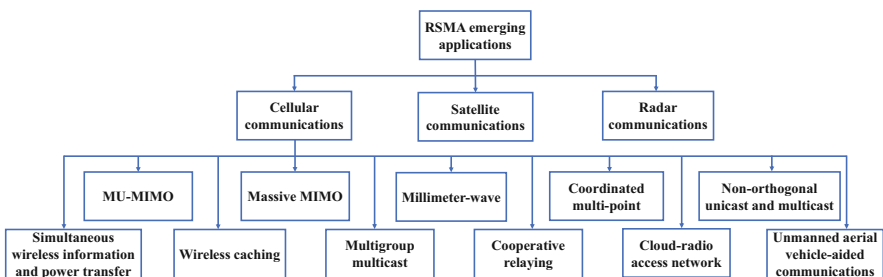


Fig. 3.14 Emerging applications of RSMA

communications, RSMA has been applied to other communication networks such as radar communications [73] and satellite communications [48, 74], which are summarized in Fig. 3.14.

3.4 Challenges and Future Trends of RSMA

The study of RSMA is still in its infancy. Even for the applications specified in Sect. 3.3, there are still many challenges and open issues that remain to be addressed. RSMA is a goldmine of research problems for academia and standard specification issues for industry. The multifarious attractive and potential research directions of RSMA are summarized in Fig. 3.15.

There are various applications of RSMA in other techniques besides those described in Sect. 3.3. Some of the techniques are complementary, and the investigation of RSMA in the combination of those techniques may collide with different sparks. For example, RSMA has shown its performance benefits respectively in cooperative relaying and NOUM. There is also a great potential of applying RSMA in NOUM with cooperative relaying. In such networks, the system performance will be further enhanced since the common stream to be forwarded from the relaying users to other users will help to enhance not only the rate of the multicast message for all users but also the unicast messages for the corresponding users.

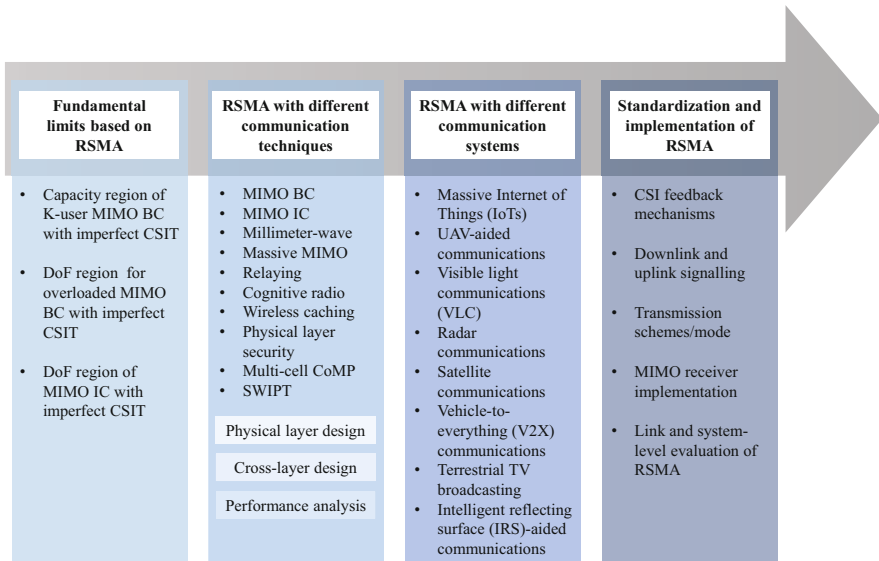


Fig. 3.15 Structure of future research directions

In addition to the promising systems summarized in Fig. 3.15, there are many other combinations of RSMA that are worth to be studied, such as RSMA in UAV-aided, radar, or satellite communication systems. In the UAV-aided communications, one major challenge is the UAV deployment and trajectory optimization. However, perfectly tracking the rapidly changed channels of the entire location map is impossible which would result in strong co-channel interference. As RSMA is superior in robust interference management and it achieves higher performance gain when CSIT is imperfect, the application of RSMA in the UAV-aided multi-antenna broadcast channel has a great potential to overcome that challenge.

The key technologies required to implement RSMA are MU-MIMO/CoMP, superposition coding at the transmitter, SIC at receivers, and non-orthogonal unicast and multicast transmission. Though the standardization of RSMA has not been taken by the 3rd Generation Partnership Project (3GPP) yet, some current work items in 3GPP can be leveraged for the implementation of RSMA. MU-MIMO and CoMP are the key technologies in LTE, which are included in 3GPP Release 8 [75] and 3GPP Release 11 [76], respectively. One major receiver technique used in RSMA is SIC, which has been incorporated in 3GPP Release 12 for network-assisted interference cancellation and suppression (NAICS) [77]. In 3GPP Release 13, superposition coding at the transmitter and successive decoding at each receiver has been further considered for LTE downlink mobile broadband (MBB) services as multiuser superposition transmission (MUST) [78]. The multicast functionality is recently included in 3GPP Release 17 for 5G with the name new radio (NR) multicast/broadcast. Besides the necessary machinery discussed or approved by 3GPP, there are some implementation issues specific to RSMA which require further study. First of all, the CSI feedback mechanisms of RSMA are unclear even though RS has been shown to reduce the CSIT feedback overhead compared to MU-LP in the presence of quantized feedback [38]. Secondly, the downlink and uplink signaling of RS remains obscure. The issue of synchronizing the knowledge of how to split/merge each stream at the transmitter and receivers needs to be tackled. Last but not least, there is still a lack of link-level and system-level evaluation of RSMA. To further evaluate the recommended configurations of RSMA in the physical layer (such as frequency band, coding scheme, modulation scheme, transceiver design, topography, etc.) or higher layers (such as scheduling, error-control scheme in the multiple access layer or QoS requirements in the application layer, etc.), the link-level and system-level performance of RSMA is of significance to be investigated thoroughly.

References

1. H. Weingarten, Y. Steinberg, S.S. Shamai, The capacity region of the Gaussian multiple-input multiple-output broadcast channel. *IEEE Trans. Inf. Theory* **52**(9), 3936–3964 (2006)
2. B. Clerckx, C. Oestges, *MIMO Wireless Networks: Channels, Techniques and Standards for Multi-antenna, Multi-user and Multi-cell Systems* (Academic Press, New York, NY, USA 2013)

3. T. Yoo, A. Goldsmith, On the optimality of multiantenna broadcast scheduling using zero-forcing beamforming. *IEEE J. Sel. Areas Commun.* **24**(3), 528–541 (2006)
4. B. Clerckx, H. Joudeh, C. Hao, M. Dai, B. Rassouli, Rate splitting for MIMO wireless networks: a promising PHY-layer strategy for LTE evolution. *IEEE Commun. Mag.* **54**(5), 98–105 (2016)
5. N. Jindal, MIMO broadcast channels with finite-rate feedback. *IEEE Trans. Inf. Theory* **52**(11), 5045–5060 (2006)
6. Y. Saito, Y. Kishiyama, A. Benjebbour, T. Nakamura, A. Li, K. Higuchi, Non-orthogonal multiple access (NOMA) for cellular future radio access, in *Proceedings of the IEEE 77th Vehicular Technology Conference (VTC Spring)*, June 2013, pp. 1–5
7. H. Nikopour, H. Baligh, Sparse code multiple access, in *Proceedings of the IEEE Annual International Symposium on Personal Indoor Mobile Radio Communications (PIMRC)*, Sept 2013, pp. 332–336
8. L. Dai, B. Wang, Y. Yuan, S. Han, I. Chih-lin, Z. Wang, Non-orthogonal multiple access for 5G: solutions, challenges, opportunities, and future research trends. *IEEE Commun. Mag.* **53**(9), 74–81 (2015)
9. Z. Ding, Y. Liu, J. Choi, Q. Sun, M. Elkashlan, I. Chih-lin, H.V. Poor, Application of non-orthogonal multiple access in LTE and 5G networks. *IEEE Commun. Mag.* **55**(2), 185–191 (2017)
10. W. Shin, M. Vaezi, B. Lee, D.J. Love, J. Lee, H.V. Poor, Non-orthogonal multiple access in multi-cell networks: theory, performance, and practical challenges. *IEEE Commun. Mag.* **55**(10), 176–183 (2017)
11. T. Cover, Broadcast channels. *IEEE Trans. Inf. Theory* **18**(1), 2–14 (1972)
12. D. Tse, P. Viswanath, *Fundamentals of Wireless Communication* (Cambridge University Press, Cambridge, U.K. 2005)
13. M.F. Hanif, Z. Ding, T. Ratnarajah, G.K. Karagiannidis, A minorization-maximization method for optimizing sum rate in the downlink of non-orthogonal multiple access systems. *IEEE Trans. Signal Process.* **64**(1), 76–88 (2016)
14. J. Choi, Minimum power multicast beamforming with superposition coding for multiresolution broadcast and application to NOMA systems. *IEEE Trans. Commun.* **63**(3), 791–800 (2015)
15. Q. Sun, S. Han, I. Chih-lin, Z. Pan, On the ergodic capacity of MIMO NOMA systems. *IEEE Wireless Commun. Lett.* **4**(4), 405–408 (2015)
16. Q. Zhang, Q. Li, J. Qin, Robust beamforming for nonorthogonal multiple-access systems in MISO channels. *IEEE Trans. Veh. Technol.* **65**(12), 10231–10236 (2016)
17. H. Joudeh, B. Clerckx, Rate-splitting for max-min fair multigroup multicast beamforming in overloaded systems. *IEEE Trans. Wireless Commun.* **16**(11), 7276–7289 (2017)
18. C. Lim, T. Yoo, B. Clerckx, B. Lee, B. Shim, Recent trend of multiuser MIMO in LTE-advanced. *IEEE Commun. Mag.* **51**(3), 127–135 (2013)
19. Z. Chen, Z. Ding, X. Dai, G.K. Karagiannidis, On the application of quasi-degradation to MISO–NOMA downlink. *IEEE Trans. Signal Process.* **64**(23), 6174–6189 (2016)
20. Z. Ding, F. Adachi, H.V. Poor, The application of MIMO to non-orthogonal multiple access. *IEEE Trans. Wireless Commun.* **15**(1), 537–552 (2016)
21. J. Choi, On generalized downlink beamforming with NOMA. *J. Commun. Netw.* **19**(4), 319–328 (2017)
22. W. Shin, M. Vaezi, B. Lee, D.J. Love, J. Lee, H.V. Poor, Coordinated beamforming for multi-cell MIMO-NOMA. *IEEE Commun. Lett.* **21**(1), 84–87 (2017)
23. V.D. Nguyen, H.D. Tuan, T.Q. Duong, H.V. Poor, O.S. Shin, Precoder design for signal superposition in MIMO–NOMA multicell networks. *IEEE J. Sel. Areas Commun.* **35**(12), 2681–2695 (2017)
24. M. Zeng, A. Yadav, O.A. Dobre, G.I. Tsiropoulos, H.V. Poor, Capacity comparison between MIMO-NOMA and MIMO-OMA with multiple users in a cluster. *IEEE J. Sel. Areas Commun.* **35**(10), 2413–2424 (2017)
25. T. Han, K. Kobayashi, A new achievable rate region for the interference channel. *IEEE Trans. Inf. Theory* **27**(1), 49–60 (1981)

26. R.H. Etkin, D.N.C. Tse, H. Wang, Gaussian interference channel capacity to within one bit. *IEEE Trans. Inf. Theory* **54**(12), 5534–5562 (2008)
27. B. Rimoldi, R. Urbanke, A rate-splitting approach to the Gaussian multiple-access channel. *IEEE Trans. Inf. Theory* **42**(2), 364–375 (1996)
28. A.G. Davoodi, S.A. Jafar, Aligned image sets under channel uncertainty: settling conjectures on the collapse of degrees of freedom under finite precision CSIT. *IEEE Trans. Inf. Theory* **62**(10), 5603–5618 (2016)
29. S. Yang, M. Kobayashi, D. Gesbert, X. Yi, Degrees of freedom of time correlated MISO broadcast channel with delayed CSIT. *IEEE Trans. Inf. Theory* **59**(1), 315–328 (2013)
30. H. Joudeh, B. Clerckx, Sum-rate maximization for linearly precoded downlink multiuser MISO systems with partial CSIT: a rate-splitting approach. *IEEE Trans. Commun.* **64**(11), 4847–4861 (2016)
31. E. Piovano, B. Clerckx, Optimal DoF region of the K-user MISO BC with partial CSIT. *IEEE Commun. Lett.* **21**(11), 2368–2371 (2017)
32. C. Hao, B. Clerckx, MISO networks with imperfect CSIT: a topological rate-splitting approach. *IEEE Trans. Commun.* **65**(5), 2164–2179 (2017)
33. C. Hao, B. Rassouli, B. Clerckx, Achievable DoF regions of MIMO networks with imperfect CSIT. *IEEE Trans. Inf. Theory* **63**(10), 6587–6606 (2017)
34. H. Joudeh, B. Clerckx, Robust transmission in downlink multiuser MISO systems: a rate-splitting approach. *IEEE Trans. Signal Process.* **64**(23), 6227–6242 (2016)
35. A.G. Davoodi, S.A. Jafar, GDoF of the MISO BC: bridging the gap between finite precision CSIT and perfect CSIT, in *Proceedings of the IEEE International Symposium on Information Theory (ISIT)*, July 2016, pp. 1297–1301
36. A.G. Davoodi, S.A. Jafar, Transmitter cooperation under finite precision CSIT: a GDoF perspective. *IEEE Trans. Inf. Theory* **63**(9), 6020–6030 (2017)
37. E. Piovano, H. Joudeh, B. Clerckx, Overloaded multiuser MISO transmission with imperfect CSIT, in *Proceedings of the 50th Asilomar Conference on Signals, Systems and Computers*, Nov 2016, pp. 34–38
38. C. Hao, Y. Wu, B. Clerckx, Rate analysis of two-receiver MISO broadcast channel with finite rate feedback: a rate-splitting approach. *IEEE Trans. Commun.* **63**(9), 3232–3246 (2015)
39. M. Dai, B. Clerckx, D. Gesbert, G. Caire, A rate splitting strategy for massive MIMO with imperfect CSIT. *IEEE Trans. Wireless Commun.* **15**(7), 4611–4624 (2016)
40. A. Papazafeiropoulos, B. Clerckx, T. Ratnarajah, Rate-splitting to mitigate residual transceiver hardware impairments in massive MIMO systems. *IEEE Trans. Veh. Technol.* **66**(9), 8196–8211 (2017)
41. M. Dai, B. Clerckx, Multiuser millimeter wave beamforming strategies with quantized and statistical CSIT. *IEEE Trans. Wireless Commun.* **16**(11), 7025–7038 (2017)
42. Y. Mao, B. Clerckx, V.O.K. Li, Rate-splitting multiple access for downlink communication systems: bridging, generalizing, and outperforming SDMA and NOMA. *EURASIP J. Wireless Commun. Netw.* **2018**(1), 133 (2018)
43. O. Tervo, L. Trant, S. Chatzinotas, B. Ottersten, M. Juntti, Multigroup multicast beamforming and antenna selection with rate-splitting in multicell systems, in *Proceedings of the IEEE International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, June 2018, pp. 1–5
44. M. Medra, T.N. Davidson, Robust downlink transmission: an offset-based single-rate-splitting approach, in *Proceedings of the IEEE International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, June 2018, pp. 1–5
45. Y. Mao, B. Clerckx, V.O.K. Li, Energy efficiency of rate-splitting multiple access, and performance benefits over SDMA and NOMA, in *Proceedings of the IEEE International Symposium on Wireless Communication Systems (ISWCS)*, Aug 2018, pp. 1–5
46. A.R. Flores, B. Clerckx, R.C. de Lamare, Tomlinson-harashima precoded rate-splitting for multiuser multiple-antenna systems, in *Proceedings of the IEEE International Symposium on Wireless Communication Systems (ISWCS)*, Aug 2018, pp. 1–6

47. A. Papazafeiropoulos, T. Ratnarajah, Rate-splitting robustness in multi-pair massive MIMO relay systems. *IEEE Trans. Wireless Commun.* **17**(8), 5623–5636 (2018)
48. M. Caus, A. Pastore, M. Navarro, T. Ramirez, C. Mosquera, N. Noels, N. Alagha, A.I. Perez-Neira, Exploratory analysis of superposition coding and rate splitting for multibeam satellite systems, in *Proceedings of the IEEE International Symposium on Wireless Communication Systems (ISWCS)*, Aug 2018, pp. 1–5
49. Y. Mao, B. Clerckx, V.O.K. Li, Rate-splitting multiple access for coordinated multi-point joint transmission, in *IEEE International Conference on Communications Workshops (ICC Workshops)*, May 2019, pp. 1–6
50. A.A. Ahmad, J. Kakar, R. Reifert, A. Sezgin, UAV-assisted C-RAN with rate splitting under base station breakdown scenarios, in *IEEE International Conference on Communications Workshops (ICC Workshops)*, May 2019, pp. 1–6
51. A. Rahmati, Y. Yapici, N. Rupasinghe, I. Guvenc, H. Dai, A. Bhuyan, Energy efficiency of RSMA and NOMA in cellular-connected mmwave UAV networks, in *IEEE International Conference on Communications Workshops (ICC Workshops)*, May 2019, pp. 1–6
52. A. Alameer Ahmad, H. Dahrouj, A. Chaaban, A. Sezgin, M. Alouini, Interference mitigation via rate-splitting and common message decoding in cloud radio access networks. *IEEE Access* **7**, 80350–80365 (2019)
53. D. Yu, J. Kim, S. Park, An efficient rate-splitting multiple access scheme for the downlink of C-RAN systems. *IEEE Wireless Commun. Lett.* **8**(6), 1555–1558 (2019)
54. Y. Mao, B. Clerckx, V.O.K. Li, Rate-splitting for multi-user multi-antenna wireless information and power transfer, in *Proceedings of the IEEE International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, July 2019, pp. 1–5
55. X. Su, L. Li, H. Yin, P. Zhang, Robust power- and rate-splitting-based transceiver design in k -user MISO SWIPT interference channel under imperfect CSIT. *IEEE Commun. Lett.* **23**(3), 514–517 (2019)
56. J. Zhang, B. Clerckx, J. Ge, Y. Mao, Cooperative rate splitting for MISO broadcast channel with user relaying, and performance benefits over cooperative NOMA. *IEEE Signal Process. Lett.* **26**(11), 1678–1682 (2019)
57. Y. Mao, B. Clerckx, V.O.K. Li, Rate-splitting for multi-antenna non-orthogonal unicast and multicast transmission: spectral and energy efficiency analysis. *IEEE Trans. Commun.* **67**(12), 8754–8770 (2019)
58. B. Clerckx, Y. Mao, R. Schober, H.V. Poor, Rate-splitting unifying SDMA, OMA, NOMA, and multicasting in MISO broadcast channel: a simple two-user rate analysis. *IEEE Wireless Commun. Lett.* **9**(3), 349–353 (2020)
59. H. Chen, D. Mi, B. Clerckx, Z. Chu, J. Shi, P. Xiao, Joint power and subcarrier allocation optimization for multigroup multicast systems with rate splitting. *IEEE Trans. Veh. Technol.* **69**(2), 2306–2310 (2020)
60. A.R. Flores, R.C. De Lamare, B. Clerckx, Linear precoding and stream combining for rate splitting in multiuser MIMO systems. *IEEE Commun. Lett.* **24**(4), 890–894 (2020)
61. Y. Mao, B. Clerckx, J. Zhang, V.O.K. Li, M. Arafah, Max-min fairness of K-user cooperative rate-splitting in MISO broadcast channel with user relaying. *IEEE Trans. on Wireless Commun.* **19**(10), 6362–6376 (2020)
62. Y. Mao, B. Clerckx, Beyond dirty paper coding for multi-antenna broadcast channel with partial CSIT: a rate-splitting approach. *IEEE Trans. on Commun.* (early access, 2020)
63. J. Zhao, D. Gündüz, O. Simeone, D. Gómez-Barquero, Non-orthogonal unicast and broadcast transmission via joint beamforming and LDM in cellular networks. *IEEE Trans. Broadcast.* **66**(2), 216–228 (2019)
64. Y.F. Liu, C. Lu, M. Tao, J. Wu, Joint multicast and unicast beamforming for the MISO downlink interference channel, in *Proceedings of the IEEE International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, July 2017, pp. 1–5
65. J. Riordan, *Introduction to combinatorial analysis* (Courier Corporation, Chelmsford, MA, USA 2012)

66. S.S. Christensen, R. Agarwal, E.D. Carvalho, J.M. Cioffi, Weighted sum-rate maximization using weighted MMSE for MIMO-BC beamforming design. *IEEE Trans. Wireless Commun.* **7**(12), 4792–4799 (2008)
67. G. Caire, N. Jindal, M. Kobayashi, N. Ravindran, Multiuser MIMO achievable rates with downlink training and channel state feedback. *IEEE Trans. Inf. Theory* **56**(6), 2845–2866 (2010)
68. C. Hao, Y. Wu, B. Clerckx, Rate analysis of two-receiver MISO broadcast channel with finite rate feedback: a rate-splitting approach. *IEEE Trans. Commun.* **63**(9), 3232–3246 (2015)
69. J. Chen, V.K.N. Lau, Two-tier precoding for FDD multi-cell massive MIMO time-varying interference networks. *IEEE J. Sel. Areas Commun.* **32**(6), 1230–1238 (2014)
70. J. Park, B. Clerckx, Multi-user linear precoding for multi-polarized massive MIMO system under imperfect CSIT. *IEEE Trans. Wireless Commun.* **14**(5), 2532–2547 (2015)
71. D. Kim, G. Lee, Y. Sung, Two-stage beamformer design for massive MIMO downlink by trace quotient formulation. *IEEE Trans. Commun.* **63**(6), 2200–2211 (2015)
72. J. Zhang, P. Elia, Fundamental limits of cache-aided wireless BC: interplay of coded-caching and CSIT feedback. *IEEE Trans. Inf. Theory* **63**(5), 3142–3160 (2017)
73. C. Xu, B. Clerckx, S. Chen, Y. Mao, J. Zhang, Rate-splitting multiple access for multi-antenna joint communication and radar transmissions. *IEEE International Conference on Communications Workshops (ICC Workshops)*, Dublin, Ireland, 2020, pp. 1–6
74. L. Yin, B. Clerckx, Rate-splitting multiple access for multibeam satellite communications. *arXiv preprint arXiv:2002.01731* (2020)
75. Evolved universal terrestrial radio access (E-UTRA); LTE physical layer; General description (Release 8), 3GPP TS 36.201, Tech. Rep., Mar 2009
76. Coordinated multi-point operation for LTE physical layer aspects (Release 11), 3GPP TR 36.819, Tech. Rep., Aug 2016
77. Study on network-assisted interference cancellation and suppression (NAIC) for LTE (Release 12), 3GPP TR 36.866, Tech. Rep., Mar 2014
78. Study on downlink multiuser superposition transmission (MUST) for LTE (Release 13), 3GPP TR 36.859, Tech. Rep., Dec 2015