Hindawi
Wireless Communications and Mobile Computing
Volume 2018, Article ID 6187580, 26 pages
https://doi.org/10.1155/2018/6187580

Review Article

Uplink Nonorthogonal Multiple Access Technologies Toward 5G: A Survey

Neng Ye,[1,2] **Hangcheng Han**,[1] **Lu Zhao**,[1] **and Ai-hua Wang**[1]

[1]*School of Information and Electronics, Beijing Institute of Technology, Beijing, China*
[2]*China Electronics Technology Group Corporation (CETC), Key Laboratory of Aerospace Information Applications, China*

Correspondence should be addressed to Hangcheng Han; hanhangcheng@bit.edu.cn

Received 26 January 2018; Accepted 14 May 2018; Published 12 June 2018

Academic Editor: Giovanni Stea

Owing to the superior performance in spectral efficiency, connectivity, and flexibility, nonorthogonal multiple access (NOMA) is recognized as the promising access protocol and is now undergoing the standardization process in 5G. Specifically, dozens of NOMA schemes have been proposed and discussed as the candidate multiple access technologies for the future radio access networks. This paper aims to make a comprehensive overview about the promising NOMA schemes. First of all, we analyze the state-of-the-art NOMA schemes by comparing the operations applied at the transmitter. Typical multiuser detection algorithms corresponding to these NOMA schemes are then introduced. Next, we focus on grant-free NOMA, which incorporates the NOMA techniques with uplink uncoordinated access and is expected to address the massive connectivity requirement of 5G. We present the motivation of applying grant-free NOMA, as well as the typical grant-free NOMA schemes and the detection techniques. In addition, this paper discusses the implementation issues of NOMA for practical deployment. Finally, we envision the future research challenges deduced from the recently proposed NOMA technologies.

1. Introduction

In the past several decades, the wireless communication system has evolved from the first generation, an analog communication network which only transfers voice messages, to LTE networks, which satisfies the great demands on mobile broadband data transmissions. Recently, the development of 5G has raised new challenges with respect to peak data rate, user experience data rate, spectral efficiency (SE), energy efficiency (EE), massive connectivity, low latency, and ultra-reliability, etc.

Nonorthogonal multiple access (NOMA) technologies have been recognized by both industry and academia as one promising tendency and progress, ever since the deployment of orthogonal frequency-division multiple access (OFDMA) in LTE, to meet the wide-ranging requirements for 5G and beyond under the strict constraint of the limited radio resources [1]. The idea of NOMA can trace back to the information-theoretic researches about multiuser information theory [2]. In downlink broadcast channel (BC), superposition coding and successive interference cancelation (SIC) receiving are employed to approach the entire capacity region of BC. Meanwhile, in uplink multiple access channel (MAC), the signals of different transmitters are overlapped and SIC receiver is applied to achieve the corner points of MAC capacity region. In 1990s, multiple access protocols, which exploited the differences between the power levels of the received packets, were proposed and studied by Shimamoto (1992) [3], Pedersen (1996) [4], and Mazzini (1998) [5], respectively. In 2008, Y. Yan and A. Li in [6] proposed a superimposed radio resource sharing (SRRS) scheme which utilizes the near-far effect to enhance the uplink throughput performance. SRRS superimposes different uplink data streams on the same radio resources and applies SIC at the receiver, which can be regarded as a prototype of NOMA,

Despite all the related researches, NOMA is still not commercialized in the past decades due to the concern of high computational complexity of SIC-type receiver. However, the rapid growth of processing power of the microprocessors in these years has provided an opportunity to the standardization and commercialization of NOMA technologies. Recently, downlink nonorthogonal transmissions, featured

Table 1: Summery of existing surveys about NOMA.

Survey	Scope	Contributions
[13]	Power-domain NOMA	A comprehensive survey about power domain-NOMA, as well as the related designs.
[14]	NOMA schemes	Review of power-domain and code-domain NOMA schemes
[15]	NOMA schemes and waveforms	Review of some NOMA schemes and nonorthogonal waveforms.
[16]	NOMA schemes	Review of some NOMA schemes towards 5G, as well as the application scenarios and typical receivers.
[17]	Theoretical analysis of NOMA	Review of the theoretical analysis about power-domain NOMA and cognitive radio inspired NOMA.
[18]	Downlink NOMA	Industrial view about downlink NOMA in 5G.

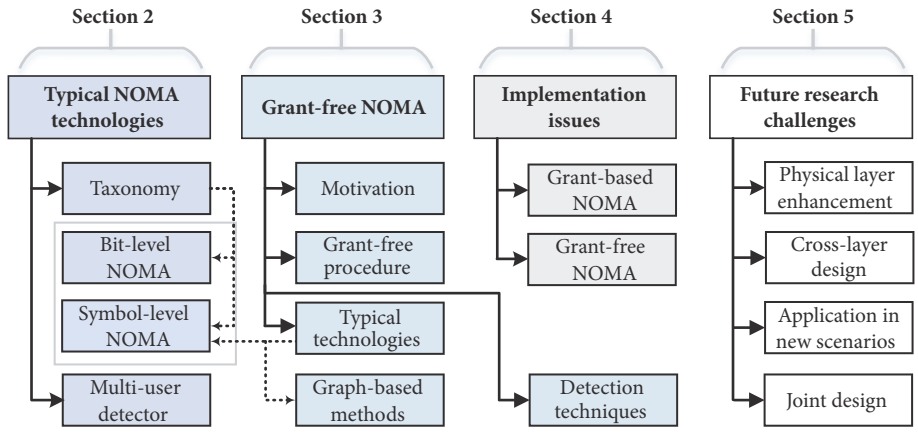


Figure 1: The outline of this paper.

by multiuser sharing technology (MUST), are specified in LTE Release-14 in 2017. Toward the evolution of 5G, the industrial community has proposed dozens of NOMA schemes as candidate multiple access technologies. In the meantime, a study item of NOMA, actuated by its potential advantages, has been approved by 3GPP RAN plenary [7] in March 2017, which promotes the standardization of NOMA in 5G.

The core idea of NOMA is to multiplex different data streams over the same radio resources and employ multiuser detection algorithm at the receiver to recover multiple users' signal streams. The major design target of NOMA is to introduce controllable mutual interference among users to achieve a fine tradeoff between multiplexing gain and detection reliability. According to both theoretical and numerical analysis, NOMA outperforms OMA with respect to SE, EE, and connectivity [8–11]. To grasp the development of NOMA technologies, some published review papers have presented different aspects of NOMA. We summarize the main contributions of these articles in Table 1.

Different from the existing literature, this paper presents a comprehensive review about the recent progress of NOMA proposed in the standardization process of 3GPP toward 5G, including candidate NOMA schemes and multiuser receiving technologies. Meanwhile, we also survey the state-of-the-art grant-free NOMA schemes, which are expected to satisfy the massive connectivity and high EE requirements in massive machine-type communication (mMTC) scenario. Additionally, we discuss the implementation issues about NOMA. The contributions of this survey are summarized in the following four aspects:

(i) It is a comprehensive survey about the candidate NOMA schemes proposed in 3GPP, as well as the promising multiuser detection methods. NOMA schemes are categorized into bit-level and symbol-level schemes for illustrations, according to the agreements in 3GPP [12].

(ii) The motivation and main idea of grant-free NOMA are presented in this survey. In addition to the grant-free procedures, this survey also introduces the typical grant-free NOMA schemes, as well as the detection algorithms.

(iii) The implementation issues about NOMA, especially grant-free NOMA, are discussed, with respect to resource allocation, procedures, and physical layer signals.

(iv) The future research challenges related to NOMA are identified, including physical layer enhancement, cross layer design, applications of NOMA in new scenarios, and the joint design of NOMA with other technologies.

Figure 1 illustrates the broad outline of this review. The rest of this review is organized as follows. Section 2 introduces the typical transmission and reception technologies of

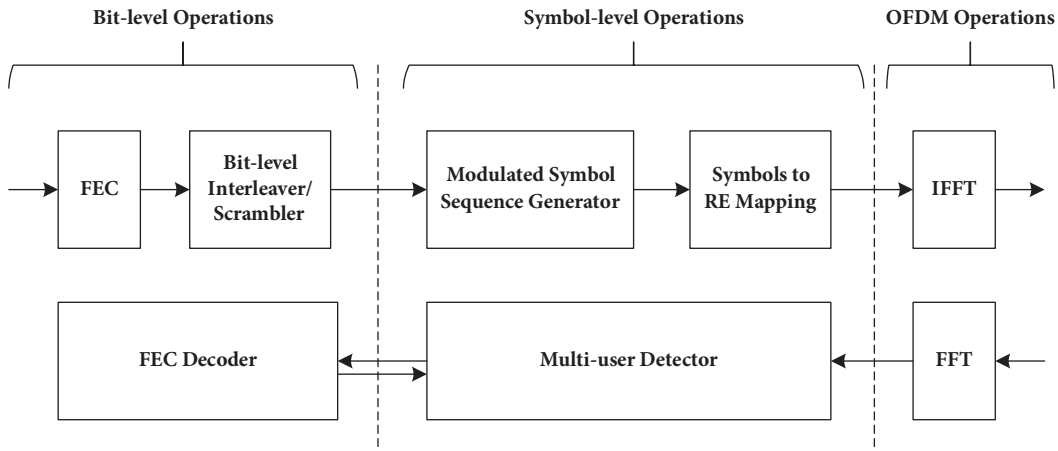


Figure 2: A unified structure of NOMA technologies.

NOMA. Section 3 analyzes the grant-free NOMA, including the motivations, procedures, and typical transceiving schemes. In Section 4, we discuss the implementation issues of NOMA toward 5G. The future research challenges about NOMA are highlighted in Section 5, and Section 6 concludes this paper.

2. Typical NOMA Technologies

The industrial community has proposed plenty of NOMA schemes to meet the diversified requirements toward 5G. Until 3GPP TSG-RAN WG1 (RAN-1) #86, at least 15 candidate NOMA schemes have been proposed for 5G new radio (NR) [19–32]. On RAN-1 #86b, a general framework of NOMA schemes is agreed upon [12], which helps to categorize existing operations in NOMA schemes into bit-level operations and symbol-level operations, as shown in Figure 2. Note that several component operations may be simultaneously adopted in the future 3GPP Release-15 to satisfy the requirements of different application scenarios.

Correspondingly, the proposed NOMA schemes can also be categorized into two classes, namely, bit-level NOMA and symbol-level NOMA, where the bit-level NOMA focuses on the design related to channel coding and bit-level interleaving, while the symbol-level NOMA mainly lays emphasis on symbol spreading and mapping. According to the above classifications, we summarize the state-of-the-art NOMA schemes in Table 2 and then give a comprehensive analysis. Meanwhile, the detection algorithms designed for these NOMA schemes are analyzed afterwards.

2.1. Bit-Level NOMA. Bit-level NOMA schemes exploit the low-rate forward error-correction (FEC) codes to enhance the detection accuracy, and/or take the advantage of user-specific interleaving to whiten the multiuser interference (MUI). In the following, we analyze several typical bit-level NOMA schemes including power-domain NOMA (PD-NOMA), low coding rate spreading (LCRS), low code rate and signature based shared access (LSSA), interleave-division multiple access (IDMA), and interleave-grid multiple access (IGMA).

2.1.1. PD-NOMA. PD-NOMA [27] multiplexes the users in power domain and applies the iteration-based SIC receiver to detect multiple signal streams at the receiver [33, 34]. In each iteration of SIC receiving, the MUI is regarded as thermal noise, which suggests that the user demultiplexing could be implemented by generating a large power difference among the multiplexed users. According to the simulation results, PD-NOMA can improve the resource utilization efficiency in both uplink and downlink [34]. Meanwhile, PD-NOMA can maintain low peak to average power ratio (PAPR) if single-carrier property is kept [35]. In addition, PD-NOMA does not depend on the information of instantaneous channel state information (CSI) of frequency-selective fading. Therefore, no matter the user mobility or CSI feedback latency, a robust performance gain in practical wide area deployments can be expected.

The major design aspect related to PD-NOMA is the resource allocation, including user association, radio resources assignment, and power allocation [36]. However, solving the resource allocation problem in one shot would be nontrivial. Therefore, this problem is usually decoupled into two subproblems, i.e., user scheduling and power allocation, respectively. In PD-NOMA, the users with large channel gain difference (e.g., large path-loss difference) are normally paired to enhance SE performance [37]. However, this simple criterion may cause unfairness in system-level deployment. Proportional fairness (PF) based scheduling [38], which simultaneously optimizes the user fairness and system throughput, is a practical user scheduling technology for PD-NOMA. The PF metric, calculated by dividing the instantaneous signal to interference and noise ratio (SINR) with the average data rate over the past period, is maximized during the user scheduling stage [39]. In uplink, user scheduling should consider the single-carrier frequency-division multiple access (SC-FDMA) where the subcarriers are distributed continuously to overcome the PAPR problem. One low complexity heuristic method based on greedy sub-band widening is proposed in [40] for practical deployment.

In the meantime, there have been abundant literature sources which address the power allocation problem of PD-NOMA [13]. Due to the nonconvexity of the power allocation

Table 2: Summarization of NOMA schemes toward 5G standardization.

NOMA scheme	Key technical point			Main advantage
IDMA	Low coding rate	Low rate FEC code	Bit-level Interleaving	Randomized the mutual interference
IGMA		Low rate FEC code or moderate one with repetition	Bit-level Interleaving (permutation matrix)	Sparse grid Mapping
LSSA		Low rate FEC code or moderate one with repetition	User-specific bit-level interleaving/permutation pattern	Large number of signatures
LCRS		Low rate FEC code and repetition	Bit-level spreading	Large coding gain
SCMA	Short low density spreading	Multidimensional modulation		Signal space diversity gain
PDMA		Irregular LDS		Irregular protection
LDS-SVE		LDS & User signature vector extension (SVE)		Higher diversity
MUSA	Short dense spreading (low cross-correlation sequence)	Short complex spreading sequence		Easy to generate & Large number
NCMA		NCC obtained by Grassmannian line packing problem		Optimal nonorthogonal sequence
NOCA		Zadoff-Chu sequence		Easy to generate, low PAPR
SSMA		Orthogonal or quasi-orthogonal codes		
GOCA	Long spreading/scrambling sequence	Group-based orthogonal/nonorthogonal sequences		Inter-group orthogonality
RDMA		Cyclic shift based time-frequency repetition		Easy implementation
RSMA		Low cross-correlation Sequence scrambling		Fit for asynchronized scenario
RSMA(single tone)	Single carrier (similar to CDMA), low PAPR modulation			Extended coverage and low PAPR for uplink

problem, advanced optimization techniques are usually employed to optimize the system throughput, reliability, and/or connectivity. In [41], the maximization of PF metric is presented. At first, the optimal power allocation of MAC is calculated iteratively. Then the optimization results can be converted to BC based on uplink-downlink duality. Several water-filling based methods are summarized and further studied in [42], where a weighted water-filling method is proposed in presence of user priority. In [43], an iterative suboptimal power allocation algorithm based on difference of convex (DC) programming is presented. The readers may refer to [13] for an extensive review about resource allocation algorithms in PD-NOMA.

Nevertheless, the existing methods may still be complex for system-level deployment. Hence, several power allocation methods are proposed by industrial community to enable efficient and practical applications. When the users have been paired into groups, one option is to apply the predefined power allocation ratios to different users as done in [44]. An alternative method is to choose one option that can maximize the PF metric out of several options; e.g., for two-user NOMA, the options of the power ratio can be $[0.2, 0.8]$ and $[0.3, 0.7]$, which is also termed fixed transmission power allocation (TPA). Since the indexes of TPA can be predefined, TPA can effectively decrease the amount of downlink signaling related to PD-NOMA. Another commonly used method is the fractional transmit power allocation (FTPA) inspired from the transmission power control used in the LTE uplink [39]. In FTPA, the users with poorer channel conditions are allocated with more power to partially compensate the channel loss. In the above FTPA, the related parameters can be optimized via system-level simulation. After the resource allocation stage, a sophisticated design of the constellations, e.g., constellation rotation [45], may provide additional gain in enhancing the detection accuracy.

When multicell or dense-network scenario is considered [46], the uplink PD-NOMA would increase the intercell interference (ICI) because multiple users are allowed to transmit on the shared carriers. Therefore, user association and ICI-aware power allocation should also be studied to control the transmission power and avoid causing severe ICI to the neighboring cells [47, 47].

2.1.2. IDMA. In addition to the power-domain multiplexing, the users can also be distinguished if they have different interleaving patterns, which is exploited in interleave-division multiple access (IDMA). IDMA is initially proposed by P. Li et al. [48] to enhance the performance of asynchronous code division multiple access (CDMA). It has the benefits of preventing the effect of fading and mitigating the ICI as in CDMA [48]. Researchers in [49] expound that IDMA can exhibit some other attractive characteristics such as flexible

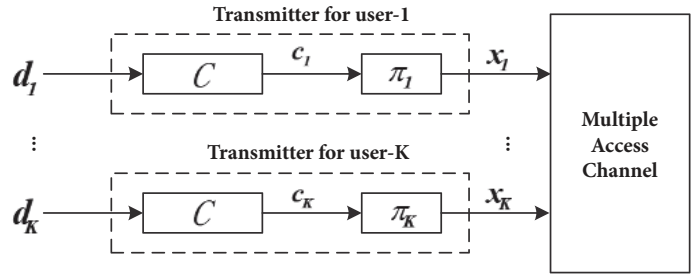


Figure 3: The transmission structure of IDMA with K multiplexed users.

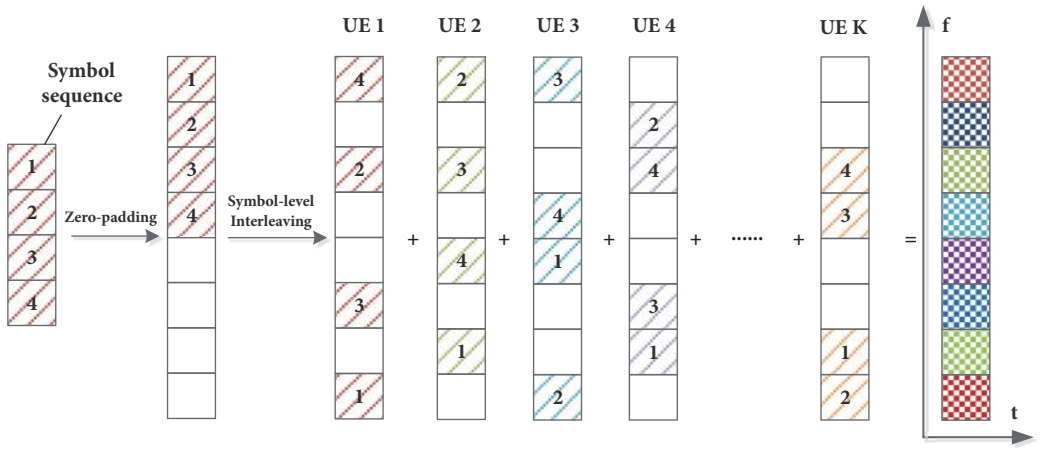


Figure 4: An illustration of the grid mapping procedure of IGMA.

rate adaptation, frequency diversity, and power efficiency. Besides, the theoretical study of IDMA also shows that the interleaved low-rate codes with a simple chip-by-chip iterative decoding strategy could achieve the capacity of a Gaussian MAC [50].

The transmission structure of IDMA is illustrated in Figure 3. The low-rate FEC encoder C is applied to encode the user-k's data bits $\mathbf{d}_k$. The output is referred as coded bits $\mathbf{c}_k$. The coded bits $\mathbf{c}_k$ pass through the interleaver π_k, after which multiple users' signals are multiplexed in the air. The interleaving patterns are generated independently and randomly and vary from each other in order to distinguish the users. Therefore, the design of reasonable interleavers is rather essential. A user-specific interleaver design method is proposed in [51], which can resolve the memory cost problem and reduce the signaling exchanging between the gNB and the users. Besides, to accommodate IDMA in multicarrier transmission, e.g., in OFDM, a multicarrier interleave-division-multiplexing-aided IDMA (MC-IDM-IDMA) is presented in [52].

IDMA has been wildly studied because of its robustness and user overload tolerance [19]. The structure of IDMA in single-path and multipath environments is elaborated in [53]. Besides, a power allocation method is introduced to enhance the performance of IDMA by taking the advantage of the semianalytical technique [48].

2.1.3. IGMA. Interleave-grid multiple access (IGMA) goes one step further than IDMA by introducing the grid mapping patterns [20], which can cooperate with the interleaving patterns to distinguish the signal streams from different users.

The flexibility to choose bit-level interleavers and/or grid mapping pattern for distinguishing the users could be easily supported in IGMA. Meanwhile, the scalability supporting different connection densities would be achieved with the abundant signatures generated by bit-level interleavers and grid mapping patterns.

Hereinafter, we briefly explain the general procedure of IGMA. Firstly, the user's data bits are encoded by the channel encoder to generate the coded bit sequence. The sequence is then interleaved to randomize the order of coded bits based on a preconfigured interleaver. The interleaved bit sequence is then modulated into the symbol sequence. Finally, the grid mapping process is conducted to interleave the symbol sequence as shown in Figure 4. The whole procedure of IGMA can further help in combating frequency selectivity and ICI due to the randomization.

2.1.4. LSSA. The low code rate and signature based shared access (LSSA) scheme is proposed to support asynchronous massive transmission in uplink [23]. LSSA randomizes the MUI among the users by multiplexing the users' data streams with user-specific signature patterns at bit-level, where the signature patterns are usually unknown to others. Besides,

Table 3: Distinction between long and short sequences.

	Long sequence	Short sequence
Level of operation	Bit level/symbol level	Usually symbol level
Generation of sequence	Randomly	Carefully designed
Usage	Disperses the encoded bit sequences so that the adjacent bits are approximately uncorrelated	To facilitate MUD
Receiving technique	Requires iterative detection between symbol-level and bit-level, e.g., ESE-SIC	Symbol level detection, e.g., MPA/EPA
Synchronization requirement	Supports asynchronous transmission when combined with single carrier waveform, e.g., RSMA	Synchronization is usually required, e.g. SCMA
Blind detection	Does not support	Support

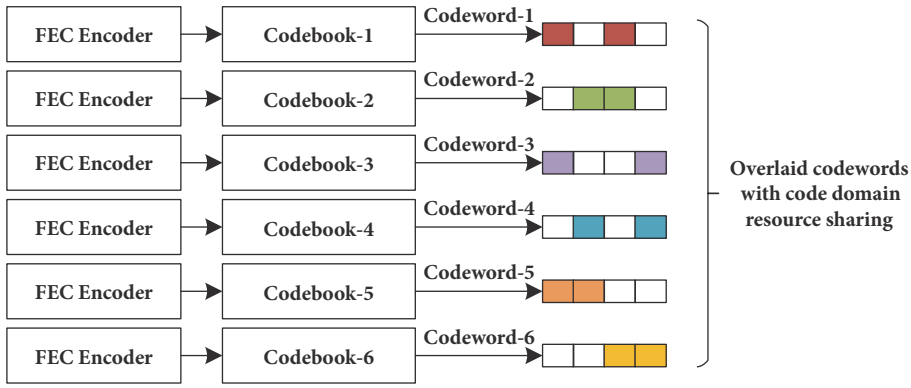


Figure 5: A typical transmission structure of SCMA with six users and four subcarriers.

the channel coding scheme which has very low code rate is adopted to encode each user's information bits in LSSA, which helps to mitigate the effect of the MUI. The low-rate FEC code can also be replaced by employing higher rate FEC code along with spreading. After the channel coding, bit-level multiplexing with user-specific signature would be used. The user-specific signature may relate to the reference signal, the complex/binary sequence, and the permutation pattern of a short length vector. The length of orthogonal spreading codes is a factor that influences the number of simultaneous transmissions. Fortunately, the receiver in LSSA does not depend on orthogonal multiplexing codes to distinguish the target users' signals. Instead, the interference cancelation is exploited, so that high user overloading is well supported. The signature of LSSA can be chosen randomly at the user side or assigned to the user by the gNB. Furthermore, LSSA can also be optionally modified to have a multicarrier variant in order to exploit frequency diversity provided by wider bandwidth and to achieve lower latency.

2.1.5. LCRS. Low code rate spreading (LCRS) is another NOMA scheme which utilizes the bit-level repetition and low-rate coding to spread information bits over the total nonorthogonal transmission area [21]. Therefore, LCRS can achieve the maximum coding gain by combining channel coding and spreading through low-rate codes. Under this circumstance, a user-specific channel interleaver [48] can be further exploited to aid the multiuser signal separation at the receiver.

2.2. Symbol-Level NOMA. Different from bit-level NOMA schemes which focus on the *bits*, symbol-level NOMA schemes play with *symbols* and mainly lay emphasis on the bit-to-symbol mapping. As illustrated in Table 2, a large portion of symbol-level NOMA schemes utilize the short sequence-based spreading to enhance the connectivity. These schemes can be further divided into two subcategories according to the densities of the spreading sequences. Some other symbol-level NOMA schemes make use of long sequence-based scrambling/spreading/permutation, where the receiver exploits the difference between these sequences. Table 3 compares the pros and cons of applying long or short sequences in symbol-level NOMA, which are further illustrated in the following subsections.

Short Sparse Spreading NOMA

(1) SCMA. Sparse code multiple access (SCMA) is a low density spreading-based NOMA scheme, which can achieve high overloading while maintaining high reliability [32, 54, 55]. The core idea of SCMA is to directly map the coded bits to the multidimensional modulation symbols, according to a predefined sparse codebook, instead of sequentially conducting modulation and low density spreading. Therefore, both the resource element mapping and the multidimensional constellation are essential designs in SCMA [56]. The transmission process of SCMA is illustrated in Figure 5, where multiple signal layers are multiplexed on the same radio resources. One major design aspect of SCMA is the sparse

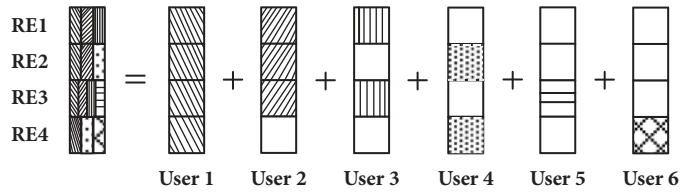


Figure 6: Resource mapping of PDMA, with six users and four subcarriers.

multidimensional codebook [57]. In [57], a SCMA codebook design method based on rotation, shuffling, and permutation is proposed, along with an example SCMA codebook which brings large shaping gain. SCMA can also achieve the signal space diversity by permuting the signal components to paired symbols located in multiple radio resources. Besides, with the sparse structure of SCMA, iterative multiuser detection algorithms, e.g., message passing algorithm (MPA), can be applied to simultaneously detect multiple data streams in symbol-level.

However, one concern of SCMA is that the sparse structure may be violated when single carrier is performed [23]. And MPA receiver may cause large computational burden and processing delay when the number of multiplexed users is large. Hence, a good tradeoff ought to be achieved between complexity and performance in the design of SCMA.

(2) PDMA. Inspired by unequal transmission diversity and sparse coding, pattern division multiple access (PDMA) is proposed as a novel NOMA scheme to enhance the performance of multiuser communication system [28]. Different from SCMA which utilizes regular spreading signatures, PDMA usually employs irregular sparse signatures to facilitate the SIC receiving [58]. Besides, with the irregular sparse spreading signatures, PDMA can have a total number of 2^N signatures where N is the length of spreading.

An example of the code domain pattern matrix of PDMA is shown in (1), which involves six users and four subcarriers. A "1" means that the subcarrier is occupied by a user. According to the spreading patterns, the signals of the six users are illustrated in Figure 6. However, we also see that one drawback of PDMA is that it cannot guarantee to accommodate the strict sparsity constraints as in SCMA; i.e., four users multiplex on the 3rd RE.

$$G_{\text{PDMA}} = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 1 \end{bmatrix} \tag{1}$$

To further enhance the ability of distinguishing the multiplexed users, PDMA allows utilizing multiple domains in the design of the signature matrix, including temporal, spatial, code, power, and interleave domains [59, 60]. For example, PDMA with large-scale antenna array (LSA-PDMA) is proposed where the spreading signatures are designed jointly in beam and power domain to improve the system sum rate and access connectivity, respectively [61]. In addition, an interleaver-based PDMA (IPDMA) scheme is proposed, where the signal separation can be done according to different bit-level interleavers and/or characteristic patterns [62]. With the joint design at the transmitter and the receiver, PDMA can meet the need of higher spectral efficiency in 5G, while ensuring a reasonable receiver complexity.

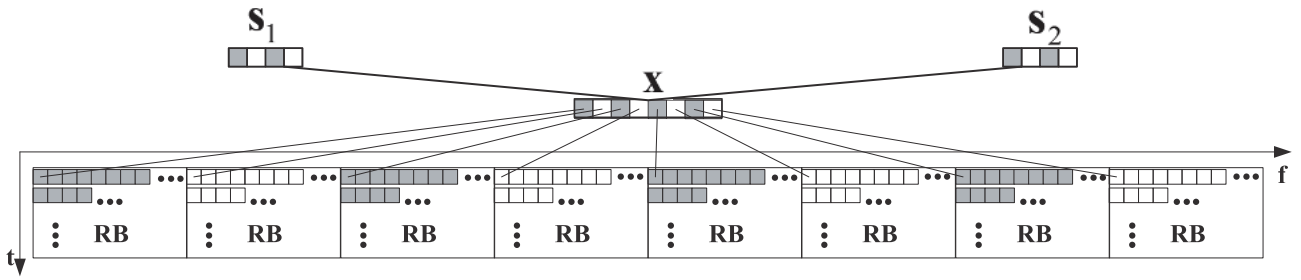
(3) LDS-SVE. Low density signature-signature vector extension (LDS-SVE) is another LDS-based NOMA scheme [22]. The major difference between LDS-SVE and the other LDS-based NOMAs, i.e., SCMA and PDMA, is that the former introduces user-specific signature vector extension, which is performed by transforming and concatenating two element signature vectors into a larger signature vector.

In the following, we show an example of LDS-SVE in Figure 7. The modulated symbols are first divided into two vectors, i.e., $\mathbf{s}_i$, $i = 1, 2$, according to a serial-to-parallel transformation. Define $\mathbf{s}_R$ as a real vector obtained by stacking the real and imaginary parts of signature $\mathbf{s}_1$ and $\mathbf{s}_2$, as shown in Figure 7. The SVE output is defined as a real vector $\mathbf{x}_R$, which is achieved by multiplying $\mathbf{s}_R$ with a transformation matrix $\mathbf{U}$. At last, transmission complex signal $\mathbf{x}$ can be recovered from $\mathbf{x}_R$, i.e., by reconstructing the complex symbols from the real vector $\mathbf{x}_R$. The main advantage of LDS-SVE is that, by multiplying $\mathbf{U}$, the original modulation symbols are spread on more REs, which brings higher order of diversity.

Short Dense Spreading NOMA

(4) MUSA. Multiuser sharing access (MUSA) is a NOMA scheme based on short complex spreading sequence and SIC receiver [24]. In general, the spreading sequences in MUSA do not have sparsity as in SCMA, PDMA, and LDS-SVE [15]. We illustrate the transmission procedure of MUSA in Figure 8. After channel encoding and modulation, as shown in Figure 8, each user's data symbols are spread by a complex sequence, whose elements take values in complex field. Then the spread symbols of each user are transmitted on the shared radio resources. At the receiver, the well-designed spreading sequences are exploited by the multiuser detectors to distinguish different users' data streams. It is worth mentioning that different symbols of the same user may use different spreading sequences, which can average the MUI and improve the system-level performance.

Short sequence-based spreading is the major operation in the MUSA transmitter. Each user can randomly pick one spreading sequence from a sequence pool consisting of multiple spreading sequences. The spreading sequence design



$$\mathbf{s}_R = \left[\text{real}(\mathbf{s}_1) \quad \text{real}(\mathbf{s}_2) \quad \text{imag}(\mathbf{s}_1) \quad \text{imag}(\mathbf{s}_2)\right]^T$$
$$\mathbf{x}_R = \left[\text{real}(\mathbf{x}) \quad \text{imag}(\mathbf{x})\right]^T \Bigg\} \quad \mathbf{x}_R = \mathbf{U}\mathbf{s}_R$$

Figure 7: An illustration of LDS-SVE.

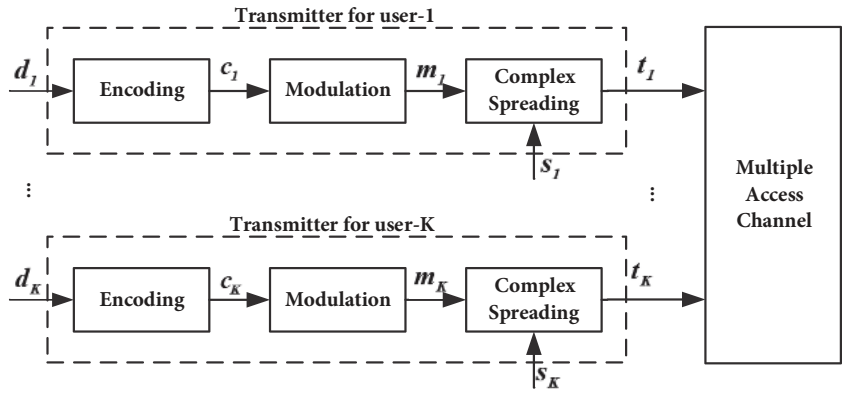


Figure 8: The transmission structure of MUSA with K simultaneous users.

in MUSA follows the guidelines of low cross-correlation, where each element of the sequence is chosen out of a complex scalar set, e.g., $\{\pm 1, 0, \pm i\}$. Due to the utilization of the imaginary part, complex spreading sequences could perform the lower cross-correlation compared to pseudorandom noise (PN), even with a short-spreading length [63]. In addition, with arbitrarily selected complex elements, the pool of the spreading sequences in MUSA can be very large.

(5) NCMA. Similar to MUSA, nonorthogonal coded multiple access (NCMA) also uses nonorthogonal dense spreading sequence to minimize MUI and support high overloading capability [25]. The spreading sequences of NCMA, also named as nonorthogonal cover codes (NCC), are obtained by solving the Grassmannian line packing problem, where the solutions of the problem guarantee the optima nonorthogonal sequences [64]. Due to the design of NCC, the interference level between two users is predictable.

The transmission structure of NCMA is illustrated in Figure 9. In NCMA, each user's data symbol is spread with NCC, and an additional FFT operation can be implemented before IFFT to reduce the PAPR. At the receiver, a simple despreading and parallel interference cancelation (PIC) detector can be implemented to recover the multiplexed signals. We note that applying IFFT on spare spreading-based NOMA schemes, e.g., SCMA and PDMA, would destroy the sparse structure and lead to high computational complexity in detection. To further improve the connectivity and bring additional throughput gain under specific QoS constraints, multistage spreading based on NCC can be applied. However, the correlation properties of the multistage spreading sequences, which are composed by multiplying several NCCs, need to be clarified. Hence, a good tradeoff between the connection density and the decoding performance needs to be further evaluated [23].

(6) NOCA. Nonorthogonal coded access (NOCA) is another spreading-based NOMA scheme. Similar to other symbol-level NOMA schemes, the data symbols in NOCA are spread according to nonorthogonal sequences before transmission [26]. The spreading in NOCA is operated in both time and frequency domain. We demonstrate the transmission structure of NOCA in Figure 10. The serial modulated symbol sequence is first converted to P parallel subsequences by a

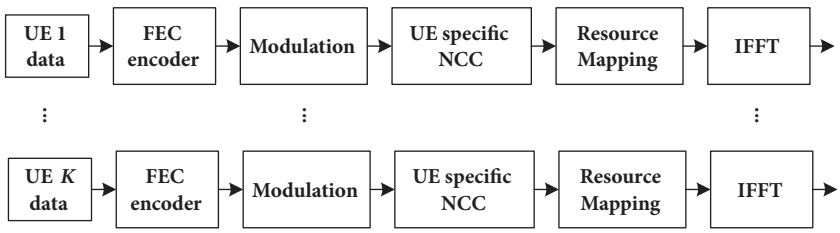


Figure 9: The transmission structure of NCMA with K simultaneous users.

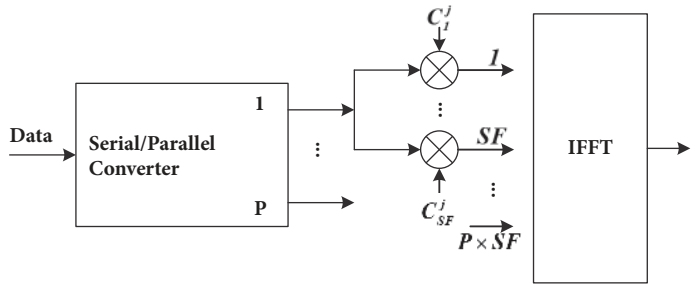


Figure 10: The transmission structure of NOCA.

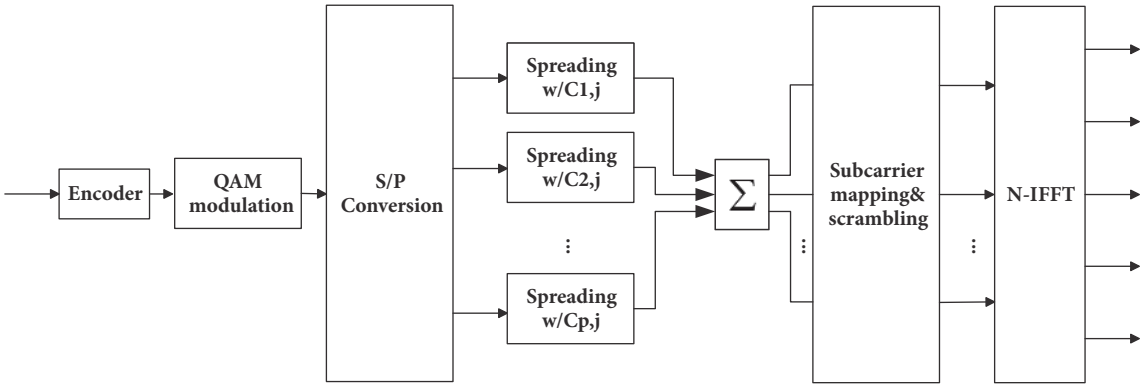


Figure 11: The transmission structures of SSMA.

S/P converter. We denote C_j as the nonorthogonal spreading sequence with length SF, where SF denotes the spreading factor. The jth subsequence is then spread on SF subcarriers according to C_j. Hence, a total number of $P \times SF$ subcarriers are required for NOCA. Besides, to accommodate the single-carrier transmission in uplink, FFT operation can also be applied before IFFT to reduce the PAPR.

To ensure high detection accuracy and high overloading, the spreading sequences used in NOCA should follow some properties, such as good autocorrelation, low cross-correlation, and low storage requirement. Meanwhile, the sequences should have constant modulus to ensure low cubic metric. Besides, multiple spreading factors might be supported for flexible adaptation.

(7) SSMA. Short sequence spreading-based multiple access (SSMA) is another spreading-based NOMA scheme [21], which directly spreads the modulation symbols with multiple orthogonal or quasi-orthogonal codes and transmits the spread symbols in time-frequency resources allocated for nonorthogonal transmission. The transmission structure of SSMA, as illustrated in Figure 11, is similar to NOCA, where user-specific scrambling is applied to average the MUI.

Long Sequence-Based NOMA

(8) RSMA. Resource spread multiple access (RSMA) is a novel NOMA scheme which applies long spreading or scrambling sequence to disperse the users signal over the entire radio resources. In RSMA, each user's codewords can be spread over all available time and frequency resources [24]. Therefore, RSMA can achieve full diversity compared to short-spreading-based NOMA schemes. At the receiver, different spreading/scrambling sequences can be exploited to distinguish different signal streams. Besides, low-rate FEC codes and advanced detection algorithms in RSMA can ensure high transmission reliability. The scrambler can also

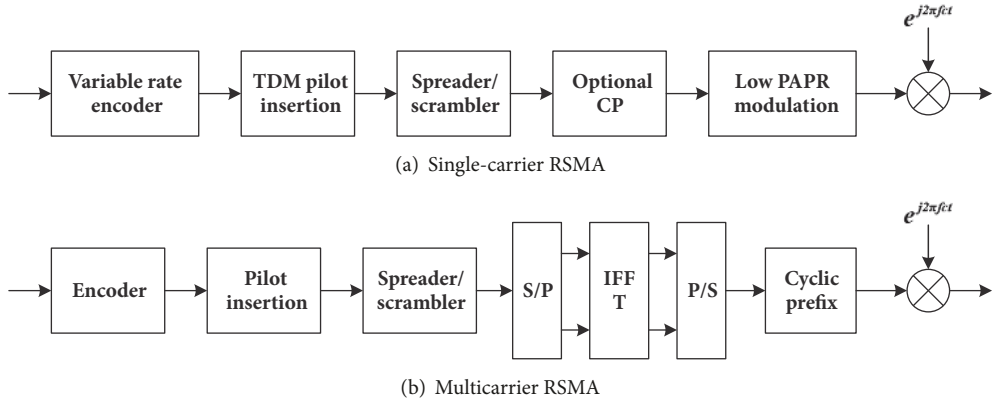


Figure 12: The transmission structures of single-carrier and multicarrier RSMAs.

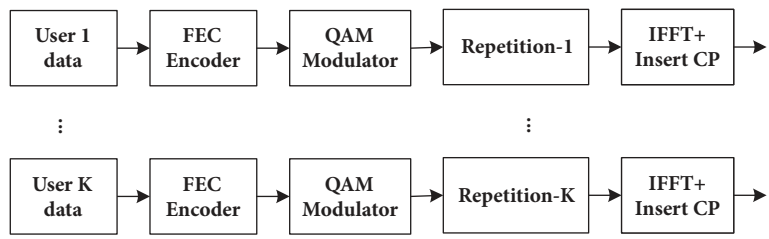


Figure 13: The transmission structure of RDMA with K simultaneous users.

be replaced by different interleavers for the sake of whitening the MUI.

According to different application scenarios, two kinds of RSMA schemes have been proposed, i.e., single-carrier RSMA and multicarrier RSMA [30], as shown in Figure 12. On the one hand, single-carrier RSMA employs the single-carrier waveforms and low PAPR modulations to enhance the performance of battery power consumption and coverage extension for small data transmission. Match-filter (MF) based receiver can be applied to distinguish different signals of single-carrier RSMA with low computational complexity. In addition, single-carrier RSMA does not rely on joint detection, which loses the synchronization requirement and makes it a good candidate for asynchronous access. On the other hand, multicarrier RSMA is studied to lower the latency and to promote the spectral efficiency for legacy users.

(9) RDMA. Repetition division multiple access (RDMA) can be regarded as an interleave-based NOMA scheme [29]. However, instead of deploying bit-level interleaving as in IDMA, RDMA focuses on the symbol-level interleaving which is designed based on simple cyclic-shift repetitions. In RDMA, each user's modulation symbol vector is repeatedly transmitted for several times, where different cyclic-shift indexes are assigned to the repetitions. Besides, different users would have different repetition and cyclic-shift patterns, which enables completely randomized MUI and achieves both time and frequency diversities.

The transmission structure of RDMA with K simultaneous users is illustrated in Figure 13. Compared with IDMA and RSMA, RDMA is simpler and may reduce the signaling overhead, since the user-specific scrambling and interleaving patterns are not needed. Meanwhile, SIC receiver is used in RDMA to provide good tradeoff between receiving complexity and detection performance.

(10) GOCA. Group orthogonal coded access (GOCA) is another long sequence-based NOMA scheme, which can be seen as an enhanced version of RDMA [29]. The major difference between GOCA and RDMA lies in the fact that the former employs the group orthogonal sequences to spread the modulation symbols into shared time and frequency resources after repetitions, as shown in Figure 14. Similar to RDMA, SIC receiver is expected to achieve good detection performance with moderate computational complexity.

The group orthogonal sequences have a two-stage structure, where orthogonal sequences and nonorthogonal sequences are used in first and second stage, respectively. Therefore, as shown in Figure 15, we can divide the GOCA sequences into several nonorthogonal groups according to the nonorthogonal sequences used in the second stage, while the sequences within a group are orthogonal to each other due to the design in the first stage.

2.3. Multiuser Detection Technologies. According to NOMA protocol, different users' signal streams are multiplexed on the same radio resources; therefore the multiuser detection (MUD) technologies are needed to distinguish independent signal streams. In the sequel, we analyze some essential MUD technologies, which are proposed to match the NOMA schemes in the above subsections.

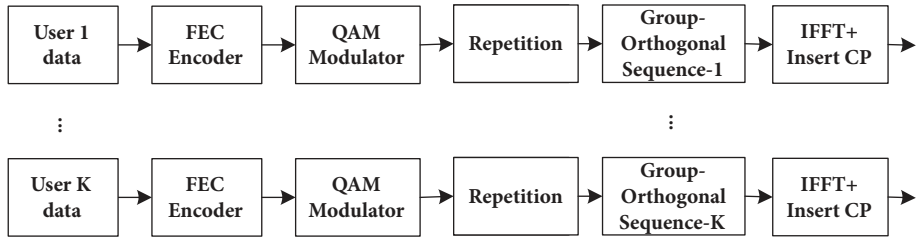


Figure 14: The transmission structure of GOCA with K simultaneous users.

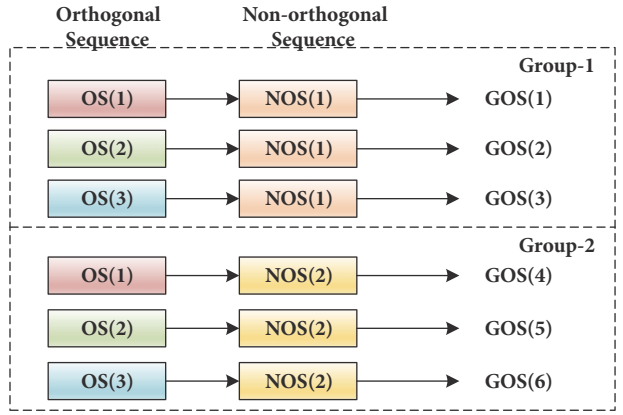


Figure 15: An example of group orthogonal sequences in GOCA, with two groups and a total of six sequences.

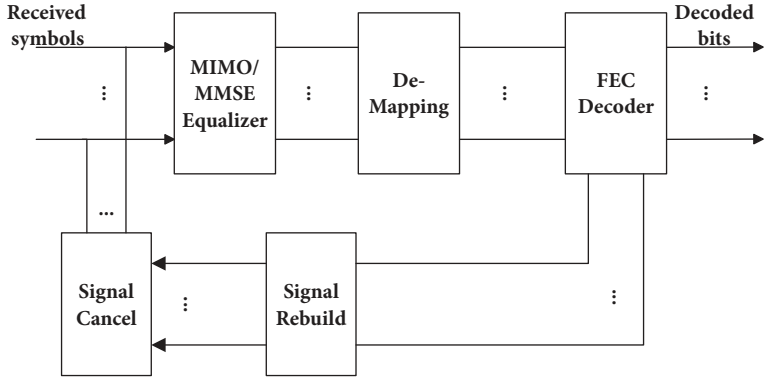


Figure 16: The structure of MMSE-SIC receiver.

2.3.1. MMSE-SIC. The minimum mean square error-successive interference cancelation (MMSE-SIC) receiver is a direct extension of MMSE receiver, as shown in Figure 16. In the first iteration of MMSE-SIC, the signal with largest received SINR is first detected by MMSE receiver by regarding the interference as noise, demapped, and then decoded to obtain the information bits. After that, the signal of this user is reconstructed and canceled from the received signal. The above procedure is repeated in the following iterations until no signal stream can be successfully recovered.

MMSE-SIC receiver suffers from error propagation problem, where the estimation errors in previous signal layers may propagate to the remaining layers. With the aim of mitigating the error propagation, the received SNRs of different data streams shall have large differences to ensure sufficient SINR in each iteration. Therefore, MMSE-SIC is especially suitable for PD-NOMA, as well as other NOMA schemes where users have diversified channel conditions.

2.3.2. MPA. MPA is an iteration-based nonlinear symbol detection algorithm, which can exploit the structure of sparse spreading sequences and achieve near maximum-likelihood (ML) performance. Different from ML receiving which estimates the entire spreading block with full search method, MPA only conducts localized optimal detection on each resource element to acquire the soft information about the transmitted symbols, and then delivers the information to the neighboring resource elements as the extrinsic information

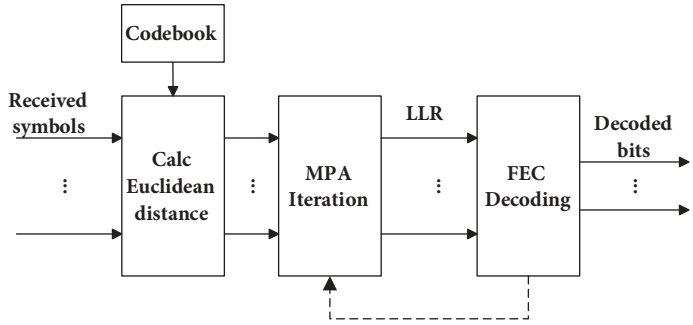


Figure 17: The structure of MPA receiver.

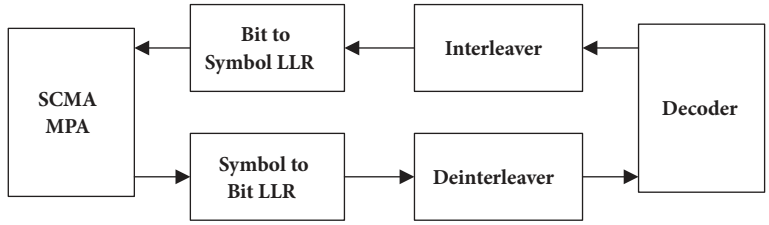


Figure 18: MPA-turbo.

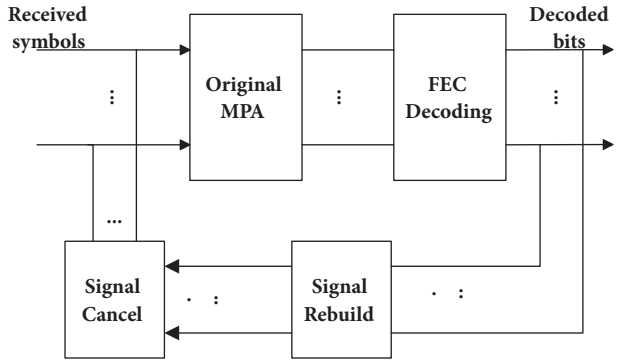


Figure 19: MPA-SIC.

of the next localized optimal detection. We show the MPA receiver in Figure 17.

The original MPA receiver focuses on symbol-level detection and does not exploit the error-correction ability of FECs to separate different signal streams. To resolve this problem, MPA-turbo and MPA-SIC are proposed as shown in Figures 18 and 19, respectively. As revealed in its name, MPA-turbo works just like the turbo decoding, where the FEC decoder processes the soft information provided by MPA and then gives feedback on the processed soft information to MPA module as extrinsic information. Different from MPA-turbo, MPA-SIC directly cancels the recovered signal streams from the received signal to mitigate the MUI.

2.3.3. EPA. Although MPA significantly reduces the computational complexity compared to ML, the complexity still grows exponentially with the number of multiplexed users on each radio resource. Estimation propagation algorithm (EPA) is another graphical-based multiuser detection algorithm, proposed for SCMA, to further reduce the computational complexity order from exponential to linear [65]. The idea of EPA originates from the variational approximate inference method, which is commonly applied in the machine learning era [45]. Different from MPA, EPA employs a Kullback-Leibler divergence based projection in the message update steps to align with the expectation propagation principle. We can directly replace the MPA module with the EPA module in the SCMA receiver, as shown in Figure 20, and generate new variants of EPA such as EPA-turbo and EPA-SIC using the similar approaches in MPA.

2.3.4. ESE-PIC. Elementary signal estimation-parallel interference cancelation (ESE-PIC) receiver is originally proposed in IDMA, which has shown robust performance even when a large number of users are multiplexed. As shown in Figure 21, ESE-PIC first detects transmitted symbols via ESE detection, a linear symbol detector. Then the detected signals are parallelly deinterleaved and decoded to acquire the coding

Table 4: Comparisons of NOMA receivers.

Receiver	Main character	Pros	Cons	Applications
MMSE-SIC	Reuses single-user receiver and SIC	Low complexity	Requires large SNR gaps among users	Almost all schemes, especially PD-NOMA
MPA/EPA	Symbol-level iterative detection	Near-ML symbol detection	Middle/high complexity	Sparse spreading NOMA
MPA-turbo/SIC	Iterative detection between symbol-level and bit-level	Better performance than MPA/EPA	High complexity	Sparse spreading NOMA
ESE-PIC	Iterative detection between symbol-level and bit-level	No requirement on sparsity	High complexity & hardware overhead	Especially bit-level NOMA

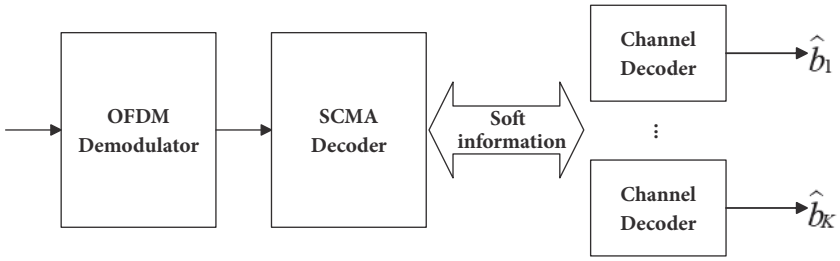


Figure 20: EPA.

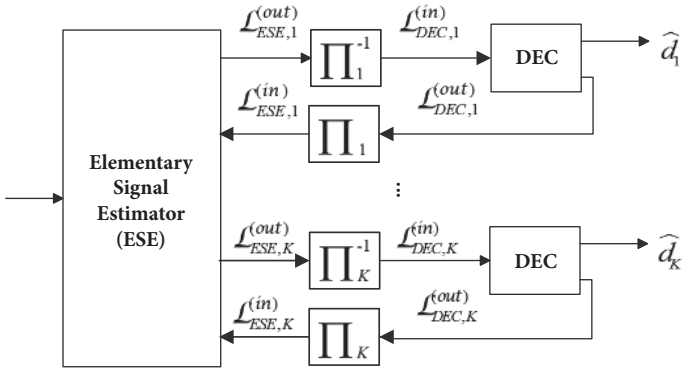


Figure 21: ESE-PIC.

gain. The output information from the decoder is then sent back to ESE module to aid symbol detection.

2.3.5. Comparisons on MUD Receivers. We now compare the pros and cons of the aforementioned receiving technologies, as well as their applicable NOMA schemes, as shown in Table 4. To sum up, the overall structure of MUD receivers consists of two parts, i.e., symbol detector and FEC decoder. Joint symbol detection, i.e., MPA and EPA, achieves better performance than single user detection, i.e., MMSE. However, they only work with short and sparse spreading sequences. Long sequence-based schemes require iterative detection, i.e., ESE-PIC; however, due to parallel message passing, several decoders may work at the same time, which leads to even larger hardware cost than MPA-turbo/SIC. To facilitate the implementation of NOMA and satisfy the diversified requirements of 5G, a good tradeoff between detection accuracy, computational complexity, latency, and hardware requirements should be achieved, which certainly requires further study.

3. Grant-Free NOMA for mMTC

The state-of-the-art NOMA schemes, mentioned in Section 2, are mainly based on centralized scheduling, where spreading sequences, interleaving patterns, and/or transmission powers of different users are scheduled by the gNB. However, the major drawback of the scheduling-based NOMA is that the signaling overhead occupies a large portion of radio resources, which makes the grant-free NOMA inevitable. In the following section, we analyze the motivation and the procedures of grant-free NOMA, as well as the typical transmission and reception technologies.

3.1. Motivation. The conventional human-type communications are normally optimized for mobile broadband (MBB) services [66], with small amount of users, high data rate,

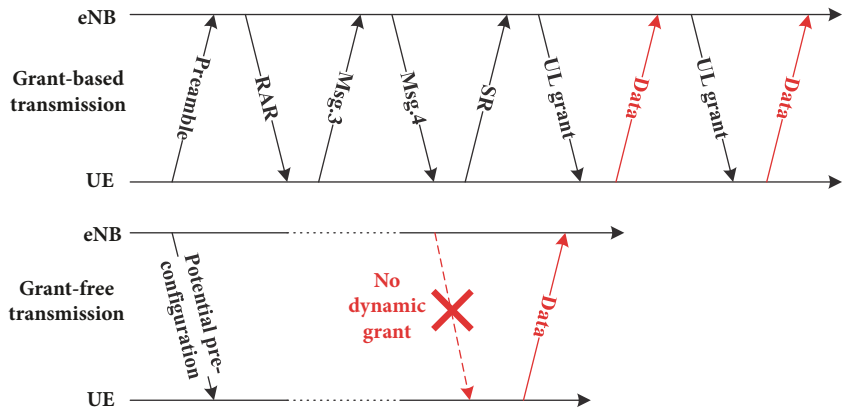


Figure 22: Comparison between the general procedures of grant-based and grant-free transmission.

and large packet size. Compared with the size of data packet in MBB, the control signaling is relatively few. Therefore, the human-type communications are not sensitive to the signaling overhead and usually involve frequent interactions between the gNB and the users to maintain high reliability and high data rate.

Despite the MBB services, 5G also aims at supporting mMTC, where massive connectivity and long battery life cycle are two key requirements. Different from in MBB scenario, the arrival of data packets in mMTC is sporadic and the packet sizes are rather small [67]. Based on [1], mMTC would support more than a million devices per square-kilometer, and this, together with the small packet sizes, makes the control signaling overhead rather significant. Therefore, the simplification on access procedure and the reduction on signaling overhead are both needed to satisfy the massive connectivity requirement of mMTC.

Grant-free NOMA, where multiple users conduct uplink instant transmissions without grant, can significantly reduce the signaling overhead. It is agreed that grant-free NOMA is more suitable for mMTC scenario due to the following concerns [66]:

(i) Energy saving: resource allocation has been well studied to extend the battery times of the devices, however, with a waste of the signaling overhead. Grant-free access can save the energy of the devices; i.e., the devices can decide to turn to active mode when small packets arrive or keep in sleep state if transmission is not needed.

(ii) Low cost devices: grant-free access can trade the computational complexity at the gNB with the hardware cost at the devices.

(iii) Latency and signaling overhead reduction: no additional latency or signaling overhead is induced by the signaling interactions.

Out of the above considerations, 3GPP has agreed that NR should target to support uplink autonomous/grant-free/contention-based transmission at least for mMTC scenario [68].

3.2. Grant-Free Procedure. In Figure 22, we compare the signaling procedures between the scheduling-based transmission in LTE and the grant-free transmission. In LTE, when a user becomes active, it would conduct random access procedure firstly, which includes at least 4 steps, i.e., preamble transmission, random access response (RAR), Message. 3, and Message. 4. After that, a scheduling request (SR) is transmitted to the gNB if the buffer is not empty, and the user does not transmit packets until it receives the uplink grant from the gNB. The above-mentioned procedures may take dozens of millisecond, which impose large signaling overhead on the network, consume more power for the signaling transmission/detection on the device side, and incur large latency for the data transmission. In contrast, grant-free transmission achieves autonomous transmission without explicit dynamic grant. Compared with the scheduling-based transmission, grant-free reduces signaling overhead, as well as control/user plane latency [69]. We note that, due to the decentralized uplink instant transmissions, the signals of users are multiplexed, which naturally leads to nonorthogonal transmissions.

We show the state graph of a grant-free user in Figure 23. If the user has no data in buffer, it stays in a sleep state; otherwise, it would wake up, synchronize according to reference signals, and acquire some necessary system broadcast information and some predefined uplink grant information. Before directly transmitting information block with the grant-free manner, preamble may be transmitted for the uplink synchronization for detection in the receiver side. Furthermore, some multiple access information could be implicitly indicated by the preamble, such as spreading signature, locations of radio resources, and the timing of retransmission. With this information, the collisions can be detected, and the blind detection complexity of the gNB can also be greatly reduced.

According to whether random access channel (RACH) is required, grant-free transmission can be classified into RACH-based grant-free and RACH-less grant-free.

3.2.1. RACH-Based Grant-Free. When all the users have performed RACH, grant-free transmission would occur in a more synchronized manner, i.e., the timing offsets among

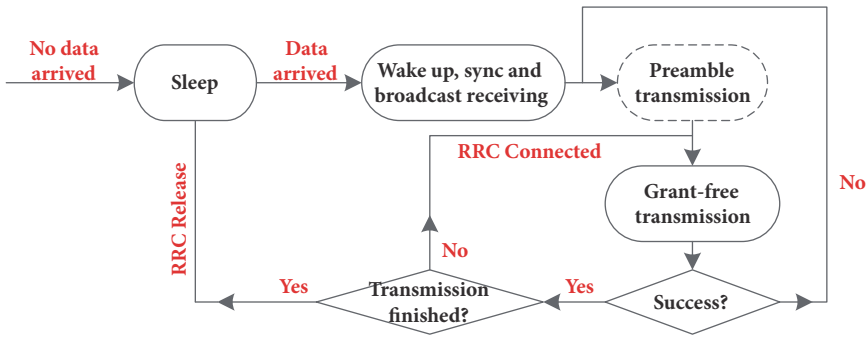


Figure 23: Grant-free uplink transmission illustrations.

users are mostly within the cyclic-prefix (CP) length. Therefore, this type of grant-free transmission is referred to as RACH-based grant-free [70] since RACH procedures have been done before data transmission. RACH-based grant-free could also reduce the overhead of SR and uplink grant, and, at the same time, it is beneficial for signal detection.

3.2.2. RACH-Less Grant-Free. In order to reduce the signaling overhead, the RACH procedure could also be canceled; i.e., data transmission phase starts whenever there are packets arriving. This method can be referred to as RACH-less grant-free [70, 71]. In this way, not only the RACH associated signaling, but also the battery energy can be saved, since the user can go to sleep if there is no data to transmit. However, the absence of RACH may result in the asynchronization among users, which may cause large detection complexity at the receiver.

3.3. Typical Grant-Free NOMA Technologies. In this subsection, we analyze the typical grant-free NOMA technologies, which are categorized into two classes, i.e., grant-free bit/symbol-level NOMA schemes and graph-based access, according to different design principles.

3.3.1. Grant-Free Bit/Symbol-Level NOMA. Grant-free bit/symbol-level NOMA can be obtained by directly incorporating grant-free access protocol with the existing NOMA schemes mentioned in Section 2, especially the short-spreading-based NOMA, e.g., SCMA, PDMA, and MUSA. To enable grant-free access in NOMA, a contention-based unit (CTU) is defined as the basic multiple access resource, as shown in Figure 24, where each CTU may consist of several fields including radio resources, reference signal, and spreading sequence [55]. One CTU may differ from the others in any fields, and these differences can be exploited by the receiver to distinguish different signal streams.

When a user has data in buffer, it randomly selects a CTU and then transmits its data packet accordingly, i.e., spreading the modulation symbols with the given spreading sequence over the given radio resources, as well as the given reference signals. When the number of active users is

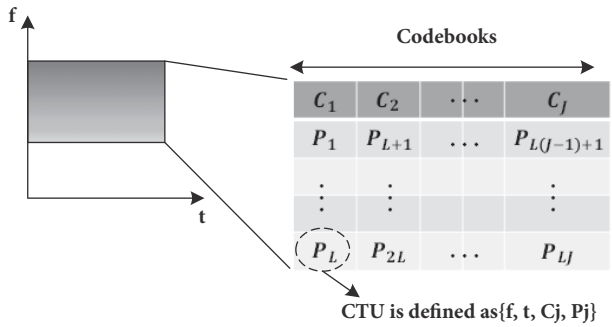


Figure 24: An illustration of CTU.

small, the users may choose different radio resources, and hence they are orthogonal. Otherwise, when the number of active users is large, the signals of the users are overlapped on the radio resources, and other fields in the CTU come into play in MUD. Therefore, grant-free NOMA can be regarded as a generalized orthogonal and nonorthogonal access.

The spreading sequences in CTUs can reuse the sequences designed for SCMA, PDMA, or MUSA. Spreading can also be replaced with sparse repetition, as proposed in [72], where simple inter- and intraslot SIC can be employed to recover multiple signal streams. In the meantime, with sparse spreading sequences, the MUI is mitigated and the receiving complexity also remains low. On the other hand, with dense spreading sequences, more diversity gain can be achieved which may combat the fading of wireless channel.

Due to the uncoordinated transmissions, the collision among users would be a severe problem in grant-free symbol-level NOMA. A hard collision happens when several users choose the same CTU. Under such circumstances, these users may be distinguished and detected only if they have distinctive channel gains. From this perspective, it is important to enlarge the pool of the spreading sequences, as done in MUSA [67]. However, enlarging the pool size may also increase the cross-correlations among the sequences, which may degrade the transmission reliability. Therefore, a good tradeoff between the pool size and the cross-correlations should be achieved to mitigate the collisions while maintaining high reliability.

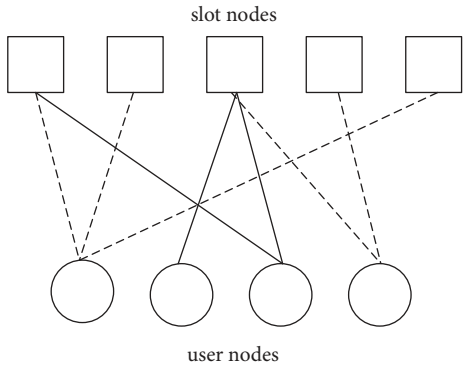


Figure 25: Bipartite graph representation of CSA.

3.3.2. Graph-Based Access Schemes. Slotted ALOHA (SA) is a conventional uncoordinated random access method which was proposed in 1970s. Recently, a class of graphical-based random access schemes has been proposed which introduces the theory of linear coding into ALOHA access [73–82]. The main idea of these schemes is to regard the random access as random coding and to optimize the access probability with coding theory. Interslot SIC receiving is applied to deal with the collisions. Besides, density evolution (DE) algorithm is usually applied to design or evaluate the transmission patterns of these schemes.

In [75], contention resolution diversity slotted ALOHA (CRDSA) is proposed which combines repetition codes with SA, where two replicas of each burst are transmitted randomly in two slots and the collided received bursts are divided by the SIC algorithm. An enhanced scheme of CRDSA, named irregular repetition slotted ALOHA (IRSA), is also introduced in [75] which optimizes the transmission method in CRDSA by bipartite graph and allows more feasible repetition pattern than CRDSA. In [73], a more generic scheme is proposed, named coded slotted ALOHA (CSA), which encodes the bursts via linear block code instead of the replicas and combines the iterative SIC with linear block code decoding to recover the source packets. A frameless ALOHA scheme based on rateless codes is provided in [83] where the transmissions of bursts act as the encoding process of rateless codes. Then, the receiver would send a feedback to the transmitter when its burst is recovered.

Hereinafter, we show a bipartite graph representation of the transmissions of CSA in Figure 25, where 4 users transmit bursts within 5 slots. Each burst node denotes the burst belonging to a user, each slot node denotes a slot, and each edge denotes that the replica of the corresponding burst is transmitted in the corresponding slot. Meanwhile, we also show the SIC process by a bipartite graph in Figure 26. In each iteration of SIC, the bursts occurred in the slots without collisions can be recovered immediately; thus the edges connected to these bursts can be removed. Then the next iteration starts and the iterations continue until no slots can be recovered. The nodes in green denote that the bursts of these users have been already recovered in the previous iterations.

3.4. Detection Techniques. In grant-free NOMA systems, the users randomly select the resources to transmit data without the dynamic scheduling. Therefore, the aforementioned MUD technologies in Section 2.3, where the identities of active users and their selected signatures are known to the receiver, are unreasonable in the practical grant-free NOMA system. Blind detection, where user activation, channel coefficients, and data packets are simultaneously detected, should be studied. Furthermore, since the transmission phase of grant-free NOMA is pretty simple, the complexity is transferred to the receiver side, which makes it rather important to design efficient blind detection algorithms.

We illustrate the general procedure at the grant-free NOMA receiver in Figure 27. The whole receiving process can be divided into two stages, i.e., user activity activation stage and data detection stage. In the first stage, active users are identified out of a potential user list. Then, in the second stage, the channel coefficients are estimated and the data packets are detected.

The idea of compressive sensing (CS) can be incorporated into the first stage due to the fact that the user activation is sparse. This sparsity is utilized in the CS-MPA detector which jointly uses CS and MPA to realize both stages simultaneously [84]. Compared with the conventional MPA without activity detection, it achieves better BLER and throughput. In addition to CS, the user activity detection could be realized by different algorithms and schemes. For example, focal underdetermined system solver (FOCUSS) and expectation maximization (EM) are proposed and analyzed for active pilot detection [85], and they can be combined with the blind data detection method, i.e., joint data and active codebook detection (JMPA), to recover the data in the spreading-based grant-free systems. It is seen that JMPA can achieve scarcely any performance degradation in decoding users' data without prior knowledge of active codebooks. Furthermore, to avoid the redundant pilot overhead, a novel sparsity-inspired sphere decoding (SI-SD) algorithm is proposed by introducing one additional all-zero codeword to achieve the maximum a posteriori (MAP) detection [86]. However, either CS or EM can only get the rough information about the active users. Detection-based group orthogonal matching pursuit (DGOMP) is a user activation detector which is promising to get a more accurate active user set [87]. Meanwhile, an enhanced version of JMPA is proposed in [87], which takes the channel gain and noise power into consideration when calculating the prior information of the zero codeword. The modified JMPA also helps to eliminate the false detection caused by noise, channel fading, and nonorthogonality of pilot sequences.

4. Implementation Issues

Nonorthogonal transmission is completely different from the orthogonal transmission which has been widely implemented in LTE. As a consequence, nonorthogonal transmission raises some implementation issues for practical deployment. In this section, we analyze some important implementation issues related to scheduling-based NOMA schemes and grant-free NOMA, respectively.

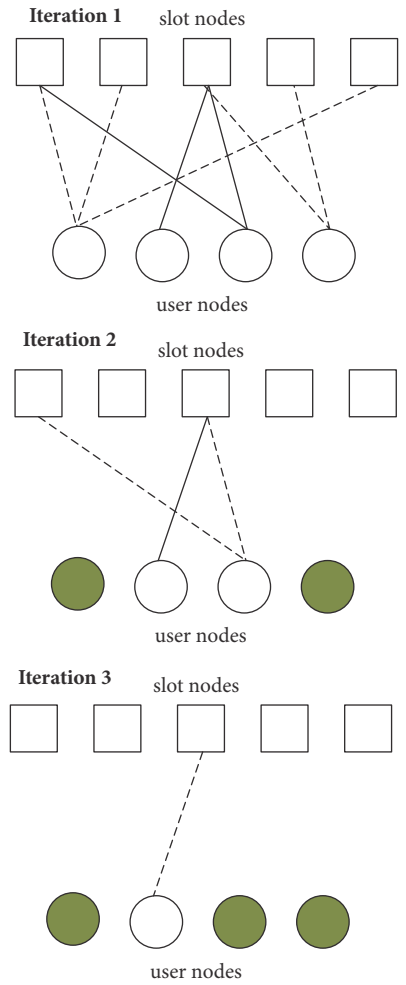


Figure 26: Bipartite graph representation of SIC process.

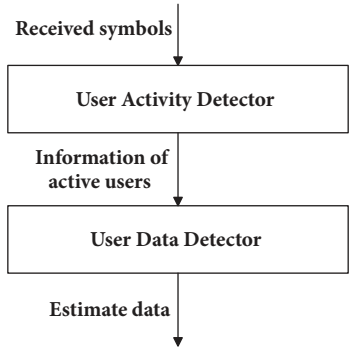


Figure 27: Grant-free NOMA receiver.

4.1. Scheduling-Based NOMA. Recall that the major difference between OMA and scheduling-based NOMA is that the latter allows multiple superimposed transmissions on the same radio resources, while the former only allows orthogonal transmissions. Hence, the resource allocation and demodulation reference signal (DM-RS) should be designed to facilitate scheduling-based NOMA.

4.1.1. Resource Allocation and Scheduling. Resource allocation, where the radio resources are assigned among users via centralized scheduling to meet certain optimization targets, has been extensively exploited in LTE to promote the system-level performance, including peak transmission data rate, average throughput, and user fairness. When multiple scheduling requests are transmitted from the users in LTE, the network orthogonally allocates the limited radio resources to a subset of the candidate users. However, the resource allocation in NOMA would be complex, since the resources in NOMA not only consist of radio resources, but also MA signature resources. Due to the fact that radio resources can be shared among users, the resource allocation problem would be even more complex. Besides, specific resource allocation methods should be designed for different NOMA schemes to match their unique characteristics. For example, PD-NOMA tends to multiplex the cell-center users and cell-edge users, while SCMA is more likely to superimpose the signals of colocated users.

The resource allocation in NOMA should also be designed to mitigate the effect of error propagation, if SIC-based receiver is employed. For example, the users with small

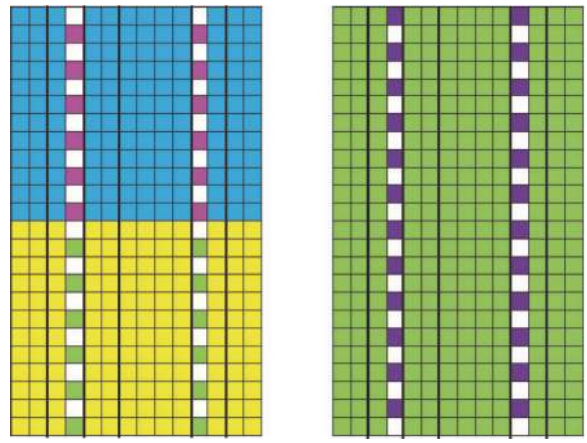


Figure 28: An example of the DM-RS structures with different combs.

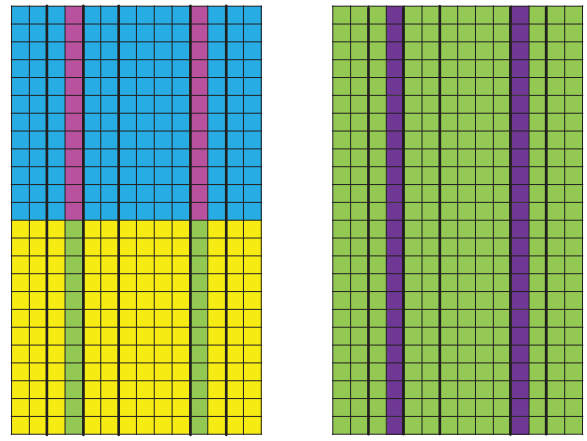


Figure 29: An example of the DM-RS structure with different OCC.

SIC order should be allocated with more radio resources to transmit signals with lower coding rates. When multicell system is considered, resource allocation should be designed to reduce the ICI. As an instance, very low density MA signatures may be allocated to the cell-edge users in uplink NOMA.

4.1.2. DM-RS. In LTE, DM-RSs with different cyclic shifts and orthogonal cover codes (OCCs) can be orthogonally multiplexed when the cyclic shift is longer than the channel delay spread. However, with the increasing demands for DM-RS ports in NOMA, it is unrealistic to add more cyclic shifts, and in the meantime, the OCC resources are also limited. Comb structures may be adopted in the design of DM-RS for NOMA. Figure 28 shows an example of the comb structure of DM-RS, which could increase the number of DM-RS resources without decreasing the accuracy in channel estimation. Compared with previous DM-RS schemes, as shown in Figure 29, the comb structure guarantees the orthogonal property of DM-RS via FDM [88, 89]. To support massive connectivity, nonorthogonal DM-RS may be further introduced to enlarge the number of DM-RS ports, where advanced channel estimation techniques should be exploited to mitigate the effect of nonorthogonality [90].

4.2. Grant-Free NOMA. As discussed in the previous section, data transmission in grant-free NOMA follows an arrive-and-go manner, which is very different from both OMA and scheduling-based NOMA. Therefore, implementing grant-free NOMA would require more efforts. This subsection presents several critical implementation issues related to grant-free NOMA, namely, resource allocation, hybrid automatic repeat request (HARQ), link adaptation, and physical signal design.

4.2.1. Resource Allocation. Similar to scheduling-based NOMA, the resources of grant-free NOMA also consist of radio resources and MA signatures. Two kinds of resource selection methods have been proposed in grant-free NOMA, i.e., random resource selection method and preconfigured method [91].

In random resource selection method, users randomly select the radio resources and the MA signatures and then transmit signals accordingly. In this case, the user activities

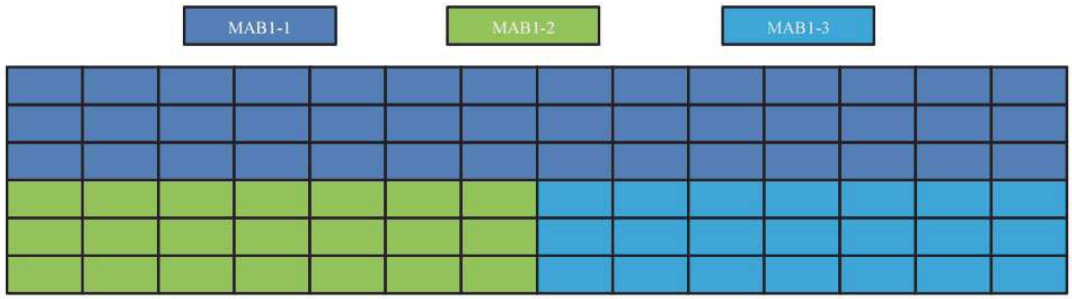


Figure 30: An example configuration of MAB.

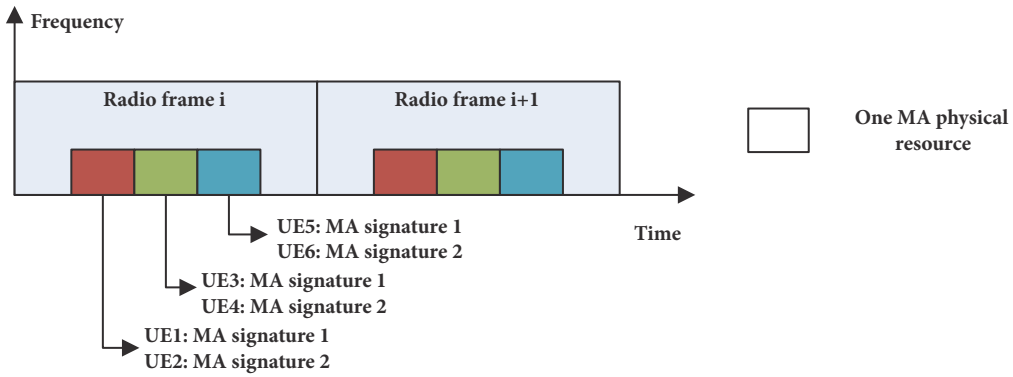


Figure 31: An illustration of predetermined MA resources for grant-free transmission.

are not available at the gNB [92], which may cause the ambiguity. In order to resolve the above problem, the radio resources should be divided into orthogonal multiple access blocks (MABs) in the time and frequency domains, as shown in Figure 30. Different MABs occupy different radio resources and may adopt different transmission settings, such as transmission block sizes (TBSs), modulation and coding schemes (MCSs), and transmission modes (TMs). The configurations of the MABs in a cell can be broadcasted by the gNB as system information. During data transmission phase, each active user first selects one MAB and then selects an MA signature. At the receiver, multiuser detection can be parallelly performed on each MAB. In addition, each MAB may be assigned with a limited number of MA signatures, which can reduce the computational complexity of blind detection at the receiver.

In the preconfigured resource allocation scheme, several users may be allocated the same radio resources along with unique MA signatures [93]. The MA signatures can be used for active user identification and collision reduction. We show an example of preconfigured scheme in Figure 31, where six users are scheduled with three distinctive pieces of MA radio resources, and the multiplexed users are allocated with different MA signatures.

4.2.2. HARQ. When the initial grant-free NOMA transmission is not successfully recovered, there is a need to retransmit the data for one or more times. HARQ is a profitable retransmission scheme which can merge the information of new transmission with previous transmissions in an effective way [94]. Due to the absence of uplink grant, one significant issue of supporting HARQ in uplink grant-free transmission is how gNB identifies the first transmission and the retransmissions for a HARQ process. One potential method is that the gNB can explicitly schedule retransmissions via downlink control signaling. Another method is to divide the MABs into several groups according to the maximum allowed number of retransmissions [95], where different users may select different MAB since they have different retransmission numbers. Another key issue in HARQ is the ACK/NACK indication. As discussed in [95], when collisions happen, gNB can utilize the RAR-style feedback, normally consisting of HARQ-ACK as well as user identification information, from which the collided users may identify whether or not their data are successfully decoded.

4.2.3. Link Adaptation. The link adaptation has been introduced in LTE to adapt to the instantaneous channel condition by adjusting the transmission parameters. Properly design link adaptation not only results in low BLER, but also reduces the retransmission number and collision probability [96]. In addition, the suitable link adaptation could achieve lower latency, which is an important target in NR application scenarios [97].

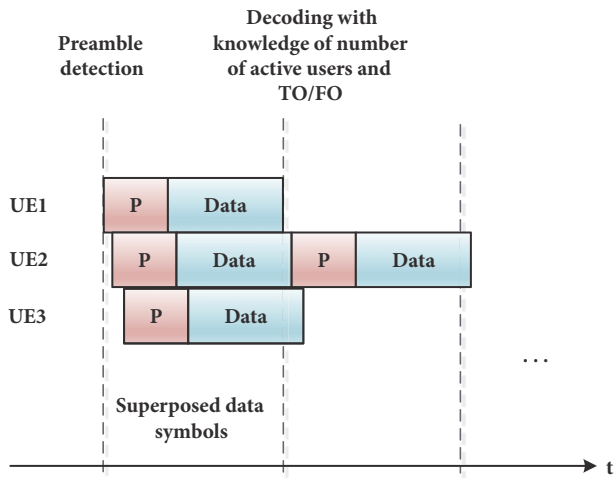


Figure 32: An illustration of RACH-less grant-free NOMA transmissions, where the data symbols are transmitted immediately after preambles.

In general, the link adaptation is realized by obtaining channel state information. However, the intermittent transmissions in uplink grant-free NOMA lead to the fact that the users might not be able to get accurate uplink channel status [98]. One solution is to use the measurements of downlink reference signals to determine the link adaptation parameters for uplink transmission. The link adaptation parameters may include MCS, number of repetitions, size of MA radio resources, and the transmission power during subsequent retransmissions [99].

4.2.4. Physical Signal Design. Physical signals, including preamble and DM-RS, are another important design aspect in grant-free NOMA. Preamble has been used in LTE for random access request [100]. However, with the aim of reducing signaling overhead, the complete random access procedure may be omitted (e.g., RACH-less grant-free). Instead, the preambles are usually directly followed by data symbols, as shown in Figure 32.

In RACH-less grant-free NOMA transmission, the users autonomously choose MA signatures as well as time instant for initial transmission, which are not known by gNB and may lead to asynchronization among received signals. In this case, well-designed preambles could assist the active user identification, MCS indication, timing offset (TO)/frequency offset (FO) estimation, and channel estimation.

Similar to the preambles, DM-RS can be used for channel estimation and user identification in grant-free NOMA [89]. However, due to the uncoordinated transmissions, different users may choose the same DM-RS, which greatly degrades the accuracy of channel estimation. To guarantee low collision probability on DM-RS, sufficient number of orthogonal/semiorthogonal DM-RSs should be provided [88]. Besides, advanced multiuser detection algorithms, such as SIC, could also help to increase the quality of channel estimation [27].

5. Future Research Challenges

Existing NOMA schemes have fully utilized the time, frequency, power, spatial, code, and interleave domains to enhance the connectivity and spectral efficiency. However, NOMA must be further studied and enhanced to satisfy the potential needs beyond NR. In this section, we highlight the future research directions of NOMA, as shown in Figure 33, including physical layer enhancement, cross layer design, joint design with other technologies, and applications of NOMA in new scenarios.

5.1. Physical Layer Enhancement. The existing NOMA schemes focus on either bit-level operations or symbol-level operations, which cannot achieve the global optimal designs. A straightforward idea is to conduct joint design of bit-level and symbol-level operations, e.g., joint design of channel encoding and symbol spreading, where the coding structure is optimized according to NOMA transmission [101, 102]. However, these designs are much too sensitive to certain channel conditions and require further enhancement for practical implementation. Despite the design at the transmitter side, signal detection at the receiver side is another physical layer technology which can be enhanced. To exploit the coding structure, several joint detection methods have been proposed in [103, 104]. However, the existing methods usually require many iterations between symbol detection and channel encoding, which leads to large detection latency. Furthermore, the high complexity and latency of blind detection still constitute an obstacle to the deployment of grant-free transmission. Therefore, simple and uniform design is required to reduce the computational complexity, as well as to maintain high reliability.

5.2. Cross Layer Design. Except for the physical layer enhancement, the cross layer design may also play an important role in the future development of NOMA [105]. For example,

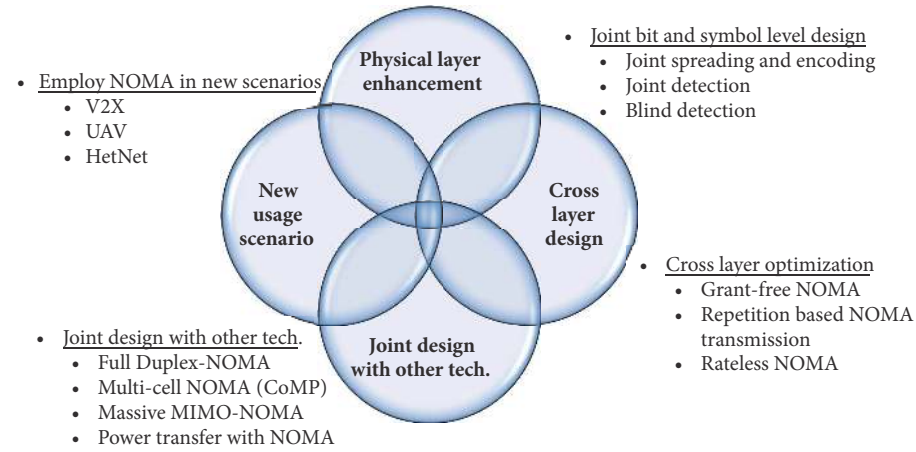


Figure 33: Future research aspects of NOMA.

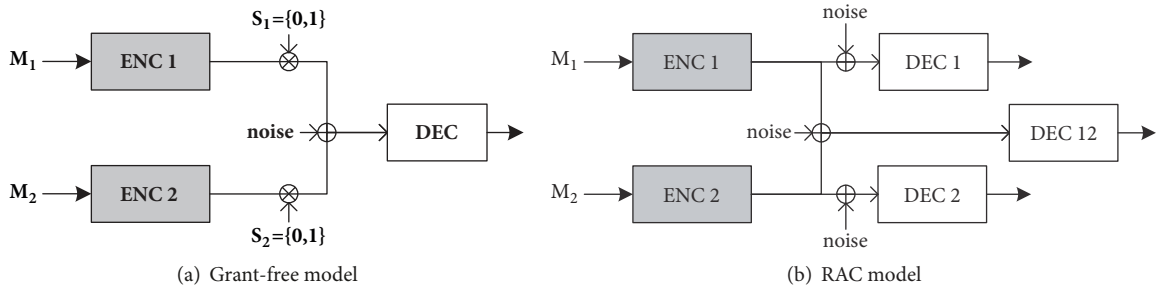


Figure 34: Illustrations of grant-free transmission and RAC models. M_1 and M_2 are the messages to be transmitted.

grant-free NOMA, which integrates the access layer protocol, i.e., grant-free protocol, into NOMA transmission, has been a promising technology for mMTC, as mentioned earlier in this paper. Besides, it is also promising to combine NOMA with other access layer techniques, such as repetition technique or rateless coding, and optimize the repetition number or code structure, respectively.

Unlike the physical layer technology which is always directed by the Shannon information theory, it is nontrivial to derive the potential limits of the channel when access layer is involved; e.g., the achievable channel capacity of grant-free NOMA is still an open problem. One possible solution to analyze the theoretical potential of NOMA with cross layer design is to formulate the Shannon information-theoretic channel model. For example, grant-free NOMA can be formulated as the random access channel (RAC) as shown in Figure 34 [106], which uses the auxiliary receivers to represent different states of user activation in grant-free access. We note that this channel model is similar to the interference channel model, where applying rate splitting can achieve a good capacity region. With this insight, one may naturally consider the deployment of rate splitting into grant-free NOMA. However, elaborate design and optimization are required to enhance the grant-free NOMA with rate splitting, for example, the coding rate and the power allocation coefficient for each splitting layer.

5.3. Joint Design with Other Technologies. Although NOMA, on its own, has met its bottleneck, we can always incorporate NOMA into other cutting edge technologies to see if there are additional advantages. For example, full duplex (FD) technology [107, 108], which is expected to increase the spectrum efficiency by a factor of two, can be jointly designed with NOMA. The conventional FD, which considers a point-to-point communication channel, is modeled by two-way channel (TWC) model. Consider a case where multiple users exist in a cell, and both gNB and the users have FD ability. Consequently, gNB and the users are simultaneously transmitting and receiving, so that the entire system can be regarded as a MAC in uplink, as well as a BC in downlink. Therefore, we may name the channel here as TW-MAC/BC model, which is illustrated in Figure 35. Obviously, NOMA technologies can be directly employed to enhance the throughput in either uplink or downlink, as it does in the conventional MAC and BC. However, there are two major differences between TW-MAC/BC and conventional MAC/BC: the first one is that, in the former case, the uplink and downlink transmissions may interfere each other due to nonideal interference cancelation, which means a good trade-off between mutual interference and sum capacity should be achieved; and the second one is that each user or gNB knows what they are receiving when they are transmitting, which means that the received signal may be exploited to derive

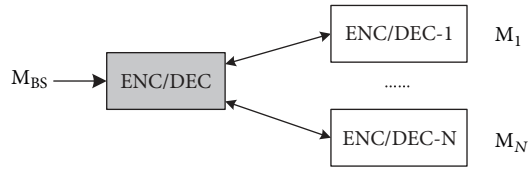


FIGURE 35: TW-MAC/BC. M_{gNB}, M_1, and M_N are the messages to be transmitted by the gNB, user-1, and user-N.

the hidden information about the channel condition and to enhance the transmission reliability.

Furthermore, NOMA can be jointly designed with cooperative multiple point (CoMP) [9, 109, 110] or multiple-antenna technology [41, 111–113]. Also, there have been some initial studies about employing NOMA in wireless power transfer network to increase the access opportunities [114–116]. Although jointly designing NOMA with the other technologies seems straightforward, whether the joint design can produce "a whole greater than the sum of its parts" is still an open question.

5.4. New Usage Scenarios. Despite the application of NOMA in the cellular networks, NOMA is also a promising technology in other new usage scenarios, e.g., vehicle to X (V2X) [117–119], unmanned aerial vehicle (UAV) [120, 121], and heterogeneous network (HetNet) [122], due to its superior performance. Although NOMA can be directly employed into these scenarios, elaborate design is still required to accommodate the channel characteristics of these scenarios and satisfy the diversified performance requirements. For example, due to the mobility of the vehicles in V2X and UAV, the handover is frequent and the cochannel interference is severe, which may degrade the reliability of existing NOMA technologies. Therefore, NOMA should be designed to maintain high reliability, as well as high throughput. As another example, in the HetNet, multiple kinds of cells, which have different transmission SNRs, may colocate together. NOMA should be designed to utilize this effect by paring the signals from different cells.

6. Conclusion

NOMA has been recognized as one of the key enabling technologies to accomplish the diversified requirements of 5G. By enabling multiple users to share the same radio resources and exploiting the advanced MUD algorithms, NOMA exhibits better performance than OMA, especially in SE and connectivity. As demonstrated in this review, the idea of superimposing the users has been carried forward into multiple domains, including power, code, interleave, and scramble, which have motivated many NOMA schemes. Meanwhile, various multiuser receiving technologies also facilitate the application of NOMA in different scenarios. Besides, we also look into grant-free NOMA, which aims to reduce the signaling overhead and increase the access probability for mMTC. Subsequently, the implementation issues and future scope of NOMA are analyzed. We hope that our survey would shed a light on the deployment and development of NOMA technologies.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This work was supported by National Natural Science Foundation of China (NSFC) under Grant 61620106001.

References

[1] 3GPP TR 38.913, "Study on scenarios and requirements for next generation access technologies".

[2] A. E. Gamal and Y. H. Kim, "Lecture notes on network information theory," *Mathematics*, 2010.

[3] S. Shimamoto, Y. Onozato, and Y. Teshigawara, "Performance evaluation of power level division multiple access (PDMA) scheme," in *Proceedings of the [Conference Record] SUPERCOMM/ICC '92 Discovering a New World of Communications*, pp. 1333–1337, Chicago, IL, USA.

[4] K. Pedersen, T. Kolding, I. Seskar, and J. Holtzman, "Practical implementation of successive interference cancellation in DS/CDMA systems," in *Proceedings of the ICUPC - 5th International Conference on Universal Personal Communications*, pp. 321–325, Cambridge, MA, USA.

[5] G. Mazzini, "Power division multiple access," in *Proceedings of the ICUPC '98. IEEE 1998 International Conference on Universal Personal Communications. Conference Proceedings*, pp. 543–546, Florence, Italy.

[6] Y. Yuan, L. Anxin, and H. Kayama, "Superimposed radio resource sharing for improving uplink spectrum efficiency," in *Proceedings of the 2008 14th Asia-Pacific Conference on Communications, APCC 2008*, usa, October 2008.

[7] 3GPP TR 38.812, "Study on non-orthogonal multiple access (NOMA) for NR".

[8] A. Benjebbour, A. Li, K. Saito, Y. Saito, Y. Kishiyama, and T. Nakamura, "NOMA: From concept to standardization," in *Proceedings of the IEEE Conference on Standards for Communications and Networking, CSCN 2015*, pp. 18–23, jpn, October 2015.

[9] W. Shin, M. Vaezi, B. Lee, D. J. Love, J. Lee, and H. V. Poor, "Non-orthogonal multiple access in multi-cell networks: Theory, performance, and practical challenges," *IEEE Communications Magazine*, vol. 55, no. 10, pp. 176–183, 2017.

[10] M. Shirvanimoghaddam, M. Dohler, and S. J. Johnson, "Massive Non-Orthogonal Multiple Access for Cellular IoT: Potentials and Limitations," *IEEE Communications Magazine*, vol. 55, no. 9, pp. 55–61, 2017.

[11] Z. Ding, Y. Liu, J. Choi et al., "Application of Non-Orthogonal Multiple Access in LTE and 5G Networks," *IEEE Communications Magazine*, vol. 55, no. 2, pp. 185–191, 2017.

[12] 3GPP R1-165021, "WF on common features and general framework of MA schemes".

[13] S. M. R. Islam, N. Avazov, O. A. Dobre, and K.-S. Kwak, "Power-Domain Non-Orthogonal Multiple Access (NOMA) in 5G Systems: Potentials and Challenges," *IEEE Communications Surveys & Tutorials*, vol. 19, no. 2, pp. 721–742, 2017.

[14] L. Dai, B. Wang, Y. Yuan, S. Han, C. I, and Z. Wang, "Non-orthogonal multiple access for 5G: solutions, challenges, opportunities, and future research trends," *IEEE Communications Magazine*, vol. 53, no. 9, pp. 74–81, 2015.

[15] Y. Tao, L. Liu, S. Liu, and Z. Zhang, "A survey: Several technologies of non-orthogonal transmission for 5G," *China Communications*, vol. 12, no. 10, pp. 1–15, 2015.

[16] Y. Wang, B. Ren, S. Sun, S. Kang, and X. Yue, "Analysis of non-orthogonal multiple access for 5G," *China Communications*, vol. 13, pp. 52–66, 2016.

[17] Z. DIng, X. Lei, G. K. Karagiannidis, R. Schober, J. Yuan, and V. K. Bhargava, "A Survey on Non-Orthogonal Multiple Access for 5G Networks: Research Challenges and Future Trends," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 10, pp. 2181–2195, 2017.

[18] Z. Wei, J. Yuan, D. W. K. Ng, M. Elkashlan, and Z. Ding, "A survey of downlink non-orthogonal multiple access for 5G wireless communication networks," *ZTE Communications*, vol. 14, no. 4, pp. 17–25, 2016.

[19] 3GPP R1-165021, "Performance of interleave division multiple access (IDMA) in combination with OFDM family waveforms".

[20] 3GPP TSG-RAN WG1-163992, "Non-orthogonal multiple access candidate for NR".

[21] 3GPP R1-162385, "Multiple access schemes for new radio interface".

[22] 3GPP R1-164329, "Initial LLS results for UL non-orthogonal multiple access".

[23] 3GPP R1-164869, "Low code rate and signature based multiple access scheme for NR".

[24] 3GPP R1-162226, "Discussion on multiple access for new radio interface".

[25] 3GPP R1-162517, "Considerations on DL/UL multiple access for NR".

[26] 3GPP R1-165019, "Non-orthogonal multiple access for NR".

[27] 3GPP R1-163111, "Initial views and evaluation results on non-orthogonal multiple access for NR uplink".

[28] 3GPP R1-163383, "Candidate solution for new multiple access".

[29] 3GPP R1-167535, "New uplink non-orthogonal multiple access schemes for NR".

[30] 3GPP R1-163510, "Candidate NR multiple access schemes".

[31] 3GPP R1-164346, "MA for eMBB in mmWave spectrum".

[32] 3GPP R1-162153, "Overview of non-orthogonal multiple access for 5G".

[33] 3GPP RWS-150051, "5G vision for 2020 and beyond".

[34] K. Higuchi and A. Benjebbour, "Non-Orthogonal Multiple Access (NOMA) with successive interference cancellation for future radio access," *IEICE Transactions on Communications*, vol. E98B, no. 3, pp. 403–414, 2015.

[35] A. Li, A. Benjebbour, X. Chen, H. Jiang, and H. Kayama, "Uplink Non-Orthogonal Multiple Access (NOMA) with Single-Carrier Frequency Division Multiple Access (SC-FDMA) for 5G systems," *IEICE Transactions on Communications*, vol. E98B, no. 8, pp. 1426–1435, 2015.

[36] Z. Wei, D. W. K. Ng, J. Yuan, and H.-M. Wang, "Optimal Resource Allocation for Power-Efficient MC-NOMA with Imperfect Channel State Information," *IEEE Transactions on Communications*, vol. 65, no. 9, pp. 3944–3961, 2017.

[37] Y. Saito, A. Benjebbour, Y. Kishiyama, and T. Nakamura, "System-level performance evaluation of downlink non-orthogonal multiple access (NOMA)," in *Proceedings of the 2013 IEEE 24th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications, PIMRC*, pp. 611–615, September 2013.

[38] A. Jalali, R. Padovani, and R. Pankaj, "Data throughput of CDMA-HDR a high efficiency-high data rate personal communication wireless system," in *Proceedings of the 51st Vehicular Technology Conference (VTC '00)*, vol. 3, pp. 1854–1858, IEEE, Tokyo, Japan, May 2000.

[39] N. Otao, Y. Kishiyama, and K. Higuchi, "Performance of non-orthogonal access with SIC in cellular downlink using proportional fair-based resource allocation," in *Proceedings of the 2012 9th International Symposium on Wireless Communication Systems, ISWCS 2012*, pp. 476–480, August 2012.

[40] X. Chen, A. Benjebbour, A. Li, and A. Harada, "Multi-user proportional fair scheduling for uplink non-orthogonal multiple access (NOMA)," in *Proceedings of the 2014 79th IEEE Vehicular Technology Conference, VTC 2014-Spring*, kor, May 2014.

[41] M. Kobayashi and G. Caire, "An iterative water-filling algorithm for maximum weighted sum-rate of Gaussian MIMO-BC," *IEEE Journal on Selected Areas in Communications*, vol. 24, no. 8, pp. 1640–1646, 2006.

[42] M.-R. Hojeij, J. Farah, C. A. Nour, and C. Douillard, "New optimal and suboptimal resource allocation techniques for downlink non-orthogonal multiple access," *Wireless Personal Communications*, vol. 87, no. 3, pp. 837–867, 2016.

[43] P. Parida and S. S. Das, "Power allocation in OFDM based NOMA systems: A DC programming approach," in *Proceedings of the 2014 IEEE Globecom Workshops, GC Wkshps 2014*, pp. 1026–1031, usa, December 2014.

[44] A. Benjebbovu, A. Li, Y. Saito, Y. Kishiyama, A. Harada, and T. Nakamura, "System-level performance of downlink NOMA for future LTE enhancements," in *Proceedings of the IEEE Globecom Workshops (GC '13)*, pp. 66–70, IEEE, Atlanta, Ga, USA, December 2013.

[45] N. Ye, A. Wang, X. Li, W. Liu, X. Hou, and H. Yu, "On Constellation Rotation of NOMA With SIC Receiver," *IEEE Communications Letters*, vol. 22, no. 3, pp. 514–517, 2018.

[46] J. An, K. Yang, J. Wu, N. Ye, S. Guo, and Z. Liao, "Achieving Sustainable Ultra-Dense Heterogeneous Networks for 5G," *IEEE Communications Magazine*, vol. 55, no. 12, pp. 84–90, 2017.

[47] Y. Fu, Y. Chen, and C. W. Sung, "Distributed Power Control for the Downlink of Multi-Cell NOMA Systems," *IEEE Transactions on Wireless Communications*, vol. 16, no. 9, pp. 6207–6220, 2017.

[48] L. Ping, L. Liu, K. Wu, and W. K. Leung, "Interleave-division multiple-access," *IEEE Transactions on Wireless Communications*, vol. 5, no. 4, pp. 938–947, 2006.

[49] L. Ping, Q. Guo, and J. Tong, "The OFDM-IDMA approach to wireless communication systems," *IEEE Wireless Communications Magazine*, vol. 14, no. 3, pp. 18–24, 2007.

[50] L. Ping, L. Liu, K. Y. Wu, and W. K. Leung, "Approaching the Capacity of Multiple Access Channels Using Interleaved Low-Rate Codes," *IEEE Communications Letters*, vol. 8, no. 1, pp. 4–6, 2004.

[51] H. Wu, L. Ping, and A. Perotti, "User-specific chip-level interleaver design for IDMA systems," *IEEE Electronics Letters*, vol. 42, no. 4, pp. 233-234, 2006.

[52] R. Zhang and L. Hanzo, "Three design aspects of multicarrier interleave division multiple access," *IEEE Transactions on Vehicular Technology*, vol. 57, no. 6, pp. 3607–3617, 2008.

[53] . Li Ping, . Lihai Liu, K. Wu, and . Leung WK, "On interleave-division multiple-access," in *Proceedings of the 2004 IEEE International Conference on Communications (IEEE Cat. No.04CH37577)*, pp. 2869–2873 Vol.5, Paris, France, June 2004.

[54] H. Nikopour and H. Baligh, "Sparse code multiple access," in *Proceedings of the IEEE 24th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC '13)*, pp. 332–336, IEEE, London, UK, September 2013.

[55] K. Au, L. Zhang, H. Nikopour et al., "Uplink contention based SCMA for 5G radio access," in *Proceedings of the 2014 IEEE Globecom Workshops, GC Wkshps 2014*, pp. 900–905, usa, December 2014.

[56] H. Yu, Z. Fei, N. Yang, and N. Ye, "Optimal design of resource element mapping for sparse spreading non-orthogonal multiple access," *IEEE Wireless Communications Letters*, vol. PP, no. 99, p. 1, 2018.

[57] M. Taherzadeh, H. Nikopour, A. Bayesteh, and H. Baligh, "SCMA codebook design," in *Proceedings of the 80th IEEE Vehicular Technology Conference, VTC 2014-Fall*, Canada, September 2014.

[58] B. Ren, Y. Wang, X. Dai, K. Niu, and W. Tang, "Pattern matrix design of PDMA for 5G UL applications," *China Communications*, vol. 13, pp. 159–173, 2016.

[59] J. Zeng, B. Li, X. Su, L. Rong, and R. Xing, "Pattern division multiple access (PDMA) for cellular future radio access," in *Proceedings of the International Conference on Wireless Communications and Signal Processing, WCSP 2015*, chn, October 2015.

[60] S. Chen, B. Ren, Q. Gao, S. Kang, S. Sun, and K. Niu, "Pattern division multiple access-a novel nonorthogonal multiple access for fifth-generation radio networks," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 4, pp. 3185–3196, 2017.

[61] P. Li, Y. Jiang, S. Kang, F. Zheng, and X. You, *Pattern division multiple access with large-scale antenna array*, 2017.

[62] J. Zeng, B. Liu, and X. Su, "Interleaver-Based Pattern Division Multiple Access with Iterative Decoding and Detection," in *Proceedings of the 2017 IEEE 85th Vehicular Technology Conference (VTC Spring)*, pp. 1–5, Sydney, NSW, June 2017.

[63] Z. Yuan, G. Yu, W. Li, Y. Yuan, X. Wang, and J. Xu, "Multi-User Shared Access for Internet of Things," in *Proceedings of the 2016 IEEE 83rd Vehicular Technology Conference (VTC Spring)*, pp. 1–5, Nanjing, China, May 2016.

[64] H. Hu and J. Wu, "New constructions of codebooks nearly meeting the Welch bound with equality," *Institute of Electrical and Electronics Engineers Transactions on Information Theory*, vol. 60, no. 2, pp. 1348–1355, 2014.

[65] X. Meng, Y. Wu, Y. Chen, and M. Cheng, "Low complexity receiver for uplink SCMA system via expectation propagation," in *Proceedings of the 2017 IEEE Wireless Communications and Networking Conference, WCNC 2017*, usa, March 2017.

[66] 3GPP R1-164268, "GB and GF MA for mMTC".

[67] 3GPP R1-166403, "Grant-free Multiple Access Schemes for mMTC".

[68] 3GPP R1-165021, "WF on clarification of grant-free transmission for mMTC".

[69] 3GPP R1-1609398, "Uplink grant-free access for 5G mMTC".

[70] 3GPP R1-167392, "Discussion on multiple access for UL mMTC".

[71] 3GPP R1-166405, "Discussion on grant-free concept for UL mMTC".

[72] N. Ye, A. Wang, X. Li, H. Yu, A. Li, and H. Jiang, "A Random Non-Orthogonal Multiple Access Scheme for mMTC," in *Proceedings of the 2017 IEEE 85th Vehicular Technology Conference (VTC Spring)*, pp. 1–6, Sydney, NSW, June 2017.

[73] Z. Sun, Y. Xie, J. Yuan, and T. Yang, "Coded Slotted ALOHA for Erasure Channels: Design and Throughput Analysis," *IEEE Transactions on Communications*, vol. 65, no. 11, pp. 4817–4830, 2017.

[74] E. Paolini, Č. Stefanović, G. Liva, and P. Popovski, "Coded random access: Applying codes on graphs to design random access protocols," *IEEE Communications Magazine*, vol. 53, no. 6, pp. 144–150, 2015.

[75] G. Liva, "Graph-based analysis and optimization of contention resolution diversity slotted ALOHA," *IEEE Transactions on Communications*, vol. 59, no. 2, pp. 477–487, 2011.

[76] L. Toni and P. Frossard, "Prioritized Random MAC Optimization Via Graph-Based Analysis," *IEEE Transactions on Communications*, vol. 63, no. 12, pp. 5002–5013, 2015.

[77] G. Liva, E. Paolini, M. Lentmaier, and M. Chiani, "Spatially-coupled random access on graphs," in *Proceedings of the 2012 IEEE International Symposium on Information Theory, ISIT 2012*, pp. 478–482, usa, July 2012.

[78] S. Kudekar, T. J. Richardson, and R. L. Urbanke, "Threshold saturation via spatial coupling: why convolutional LDPC ensembles perform so well over the BEC," *Institute of Electrical and Electronics Engineers Transactions on Information Theory*, vol. 57, no. 2, pp. 803–834, 2011.

[79] M. Ivanov, F. Brännström, A. GraellAmat, and G. Liva, "Unequal Error Protection in Coded Slotted ALOHA," *IEEE Wireless Communications Letters*, vol. 5, no. 5, pp. 536–539, 2016.

[80] D. Jia, Z. Fei, H. Lin, J. Yuan, and J. Kuang, "Distributed Decoding for Coded Slotted ALOHA," *IEEE Communications Letters*, vol. 21, no. 8, pp. 1715–1718, 2017.

[81] C. Stefanovic and P. Popovski, "Coded slotted ALOHA with varying packet loss rate across users," in *Proceedings of the 2013 1st IEEE Global Conference on Signal and Information Processing, GlobalSIP 2013*, pp. 787–790, usa, December 2013.

[82] Z. Sun, L. Yang, J. Yuan, and M. Chiani, "A novel detection algorithm for random multiple access based on physical-layer network coding," in *Proceedings of the 2016 IEEE International Conference on Communications Workshops, ICC 2016*, pp. 608–613, mys, May 2016.

[83] Č. Stefanović and P. Popovski, "ALOHA random access that operates as a rateless code," *IEEE Transactions on Communications*, vol. 61, no. 11, pp. 4653–4662, 2013.

[84] B. Wang, L. Dai, Y. Yuan, and Z. Wang, "Compressive sensing based multi-user detection for uplink grant-free non-orthogonal multiple access," in *Proceedings of the 82nd IEEE*

Vehicular Technology Conference, VTC Fall 2015, usa, September 2015.

[85] A. Bayesteh, E. Yi, H. Nikopour, and H. Baligh, "Blind detection of SCMA for uplink grant-free multiple-access," in *Proceedings of the 2014 11th International Symposium on Wireless Communications Systems, ISWCS 2014*, pp. 853–857, esp, August 2014.

[86] G. Chen, J. Dai, K. Niu, and C. Dong, "Sparsity-Inspired Sphere Decoding (SI-SD): A Novel Blind Detection Algorithm for Uplink Grant-Free Sparse Code Multiple Access," *IEEE Access*, vol. 5, pp. 19983–19993, 2017.

[87] J. Liu, G. Wu, S. Li, and O. Tirkkonen, "Blind detection of uplink grant-free SCMA with unknown user sparsity," in *Proceedings of the 2017 IEEE International Conference on Communications, ICC 2017*, fra, May 2017.

[88] 3GPP R1-1612573, "Collision analysis of grant-free based multiple access".

[89] 3GPP R1-1608919, "Considerations on pre-configured resource for grant-free based UL non-orthogonal MA".

[90] X. Chen, Z. Zhang, C. Zhong, R. Jia, and D. W. K. Ng, "Fully non-orthogonal communication for massive access," *IEEE Transactions on Communications*, vol. PP, no. 99, p. 1, 2017.

[91] 3GPP R1-1609227, "On MA resource and MA signature configurations".

[92] 3GPP R1-1608917, "Considerations on random resource selection".

[93] 3GPP R1-1609647, "On MA resources for grant-free transmission".

[94] 3GPP R1-1608859, "The retransmission and HARQ schemes for grant-free".

[95] 3GPP R1-1609039, "HARQ operation for grant-free based multiple access".

[96] 3GPP R1-1609649, "Grant-free retransmission with diversity and combining for NR".

[97] 3GPP R1-1609648, "Collision handling for grant-free".

[98] 3GPP R1-1609654, "Link adaptation for grant-free transmissions".

[99] 3GPP R1-1610374, "Support of link adaptation for UL grant-free NOMA schemes".

[100] S. Sesia, I. Toufik, and M. Baker, *LTE-The UMTS Long Term Evolution: Form Theory to Practice*, Wiley, 2011.

[101] J. Dai, K. Niu, Z. Si, C. Dong, and J. Lin, "Polar-coded non-orthogonal multiple access," *IEEE Transactions on Signal Processing*, no. 99, p. 1, 2017.

[102] M. Qiu, Y. Huang, S. Shieh, and J. Yuan, "A Lattice-Partition Framework of Downlink Non-Orthogonal Multiple Access without SIC," in *Proceedings of the 2017 IEEE Global Communications Conference (GLOBECOM 2017)*, pp. 1–6, Singapore, December 2017.

[103] S. Chen, K. Peng, Y. Zhang, and J. Song, "Near capacity LDPC coded MU-BICM-ID for 5G," in *Proceedings of the 11th International Wireless Communications and Mobile Computing Conference, IWCMC 2015*, pp. 1418–1423, hrv, August 2015.

[104] L. Wen, R. Razavi, M. A. Imran, and P. Xiao, "Design of Joint Sparse Graph for OFDM System," *IEEE Transactions on Wireless Communications*, vol. 14, no. 4, pp. 1823–1836, 2015.

[105] G. Liu, Z. Ma, X. Chen, Z. Ding, F. R. Yu, and P. Fan, "Cross-layer power allocation in non-orthogonal multiple access systems for statistical QoS provisioning," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 12, pp. 11–388, Dec 2017.

[106] P. Minero, M. Franceschetti, and D. N. Tse, "Random access: an information-theoretic perspective," *Institute of Electrical and Electronics Engineers Transactions on Information Theory*, vol. 58, no. 2, pp. 909–930, 2012.

[107] M. S. Sim, M. Chung, D. Kim, J. Chung, D. K. Kim, and C.-B. Chae, "Nonlinear Self-Interference Cancellation for Full-Duplex Radios: From Link-Level and System-Level Performance Perspectives," *IEEE Communications Magazine*, vol. 55, no. 9, pp. 158–167, 2017.

[108] A. Kord, D. L. Sounas, and A. Alu, "Achieving Full-Duplex Communication: Magnetless Parametric Circulators for Full-Duplex Communication Systems," *IEEE Microwave Magazine*, vol. 19, no. 1, pp. 84–90, 2018.

[109] Y. Tian, A. R. Nix, and M. Beach, "On the Performance of Opportunistic NOMA in Downlink CoMP Networks," *IEEE Communications Letters*, vol. 20, no. 5, pp. 998–1001, 2016.

[110] Z. Liu, G. Kang, L. Lei, N. Zhang, and S. Zhang, "Power Allocation for Energy Efficiency Maximization in Downlink CoMP Systems with NOMA," in *Proceedings of the 2017 IEEE Wireless Communications and Networking Conference (WCNC)*, pp. 1–6, San Francisco, CA, USA, March 2017.

[111] X. Liu, Y. Liu, X. Wang, and H. Lin, "Highly Efficient 3-D Resource Allocation Techniques in 5G for NOMA-Enabled Massive MIMO and Relaying Systems," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 12, pp. 2785–2797, 2017.

[112] C. Chen, W. Cai, X. Cheng, L. Yang, and Y. Jin, "Low Complexity Beamforming and User Selection Schemes for 5G MIMO-NOMA Systems," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 12, pp. 2708–2722, 2017.

[113] V. Nguyen, H. D. Tuan, T. Q. Duong, H. V. Poor, and O. Shin, "Precoder Design for Signal Superposition in MIMO-NOMA Multicell Networks," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 12, pp. 2681–2695, 2017.

[114] Y. Xu, C. Shen, Z. Ding et al., "Joint beamforming and power-splitting control in downlink cooperative SWIPT NOMA systems," *IEEE Transactions on Signal Processing*, vol. 65, no. 18, pp. 4874–4886, 2017.

[115] Z. Yang, Z. Ding, P. Fan, and N. Al-Dhahir, "The Impact of Power Allocation on Cooperative Non-orthogonal Multiple Access Networks with SWIPT," *IEEE Transactions on Wireless Communications*, vol. 16, no. 7, pp. 4332–4343, 2017.

[116] Y. Liu, Z. Ding, M. Elkashlan, and H. V. Poor, "Cooperative Non-orthogonal Multiple Access with Simultaneous Wireless Information and Power Transfer," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 4, pp. 938–953, 2016.

[117] Y. Chen, L. Wang, Y. Ai, B. Jiao, and L. Hanzo, "Performance Analysis of NOMA-SM in Vehicle-to-Vehicle Massive MIMO Channels," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 12, pp. 2653–2666, 2017.

[118] B. Di, L. Song, Y. Li, and G. Y. Li, "NOMA-Based Low-Latency and High-Reliable Broadcast Communications for 5G V2X Services," in *Proceedings of the GLOBECOM 2017 - 2017 IEEE Global Communications Conference*, pp. 1–6, Singapore, December 2017.

[119] J. Dai, K. Niu, Z. Si, C. Dong, and J. Lin, "Polar-coded non-orthogonal multiple access," *IEEE Transactions on Signal Processing*, vol. 66, no. 5, pp. 1374–1389, 2018.

[120] W. Fawaz, C. Abou-Rjeily, and C. Assi, "UAV-Aided Cooperation for FSO Communication Systems," *IEEE Communications Magazine*, vol. 56, no. 1, pp. 70–75, 2018.

[121] N. H. Motlagh, M. Bagaa, and T. Taleb, "UAV-Based IoT Platform: A Crowd Surveillance Use Case," *IEEE Communications Magazine*, vol. 55, no. 2, pp. 128–134, 2017.

[122] J. Zhao, Y. Liu, K. K. Chai, A. Nallanathan, Y. Chen, and Z. Han, "Resource allocation for non-orthogonal multiple access in heterogeneous networks," in *Proceedings of the ICC 2017 - 2017 IEEE International Conference on Communications*, pp. 1–6, Paris, France, May 2017.